ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΦΙΛΟΣΟΦΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΦΙΛΟΛΟΓΙΑΣ

Ανάλυση ελληνικών σωμάτων κειμένων με τη χρήση τεχνικών μηχανικής μάθησης: Υπολογιστική αναπαράσταση της ιδιολέκτου

Κωνσταντίνος Περήφανος

Ιανουάριος 2019

Στη Μαρία

Περίληψη

Η ιδιόλεκτος, στο πλαίσιο της γλωσσολογίας, αναφέρεται στη μοναδική και ιδιαίτερη χρήση της γλώσσας ενός ατόμου και αποτελεί το αντίστοιχο της κοινωνιολέκτου με σημείο αναφοράς το άτομο. Η έρευνα για την έννοια της ιδιολέκτου στη γλωσσολογία είναι μάλλον περιορισμένη, ειδικά σε ό,τι αφορά την επικύρωσή της με εμπειρικούς τρόπους. Η σχετική έρευνα στη γλωσσολογία σωμάτων κειμένων και στην υφομετρία έχει επίσης περιορισμούς που αφορούν είτε τον αριθμό των υπό εξέταση συγγραφέων (μικρότερος των 10) είτε τον αριθμό των όρων του λεξιλογίου που χρησιμοποιούνται στην εξέταση της ιδιολεκτικής ομοιότητας (περί των 310 λειτουργικών λέξεων).

Η παρούσα διατριβή χρησιμοποιεί κατανεμημένες αναπαραστάσεις (λεξικές ενθέσεις) για την ανάλυση κειμένων χρηστών κοινωνικών δικτύων, τα οποία θεωρείται ότι αντανακλούν το ιδιαίτερο προσωπικό ύφος κάθε συγγραφέα. Τα δεδομένα στα οποία βασίζεται η διατριβή αποτελούνται από ένα σώμα κειμένων του Twitter στα ελληνικά, που προέρχεται από 4.949 χρήστες από το 2009 έως το 2016 (περίπου 325 εκ. λέξεις), και το σώμα κειμένων Blog Authorship Corpus για σύγκριση και επαλήθευση. Με αφετηρία την Κατανεμητική Υπόθεση του Zellig Harris, σύμφωνα με την οποία σημασιολογικά παρόμοιες λέξεις τείνουν να εμφανίζονται σε παρόμοια περικείμενα, η έννοια της λεξικής ένθεσης αποτελεί τη βάση σύνθεσης υφολογικών ενθέσεων, παρέχοντας έτσι τη δυνατότητα να απαντηθεί το ερώτημα της ύπαρξης ιδιολέκτου και παρέχοντας έτσι ένα υφολογικό αποτύπωμα των υπό εξέταση συγγραφέων.

Στη διατριβή εξετάζεται και συγκρίνεται η απόδοση διαφόρων μοντέλων κατασκευής κατανεμητικών αναπαραστάσεων. Πιο συγκεκριμένα, χρησιμοποιούνται λεξικές ενθέσεις που παράγονται από Νευρωνικά Γλωσσικά Μοντέλα (word2vec, doc2vec, fastText), καθώς και μοντέλα που προκύπτουν από παραγοντοποίηση πινάκων συνεμφάνισης όρων (GloVe). Τα επιλεγμένα μοντέλα εφαρμόζονται σε όλο το λεξιλόγιο των υπό εξέταση κειμένων και συνεπώς δεν περιορίζονται σε περιορισμένο λεξιλόγιο και ταυτόχρονα είναι εύκολα επεκτάσιμα σε σώματα κειμένων δεκάδων χιλιάδων συγγραφέων.

Στην παρούσα διατριβή βρέθηκε ότι οι λεξικές ενθέσεις α) μπορούν να χρησιμοποιηθούν ως δομικό συστατικό αναπαράστασης του ατομικού κειμενικού ύφους και β) οι ιδιολεκτικές ενθέσεις παρέχουν τη δυνατότητα συσταδοποίησης ιδιολεκτικής ομοιότητας, δημιουργώντας έτσι ομάδες παρόμοιου ύφους, καθώς επίσης και μέτρα αποτίμησης της σταθερότητας του κειμενικού ύφους στο πέρασμα του χρόνου. Τα ευρήματα αυτά έχουν σημαντικές εφαρμογές σε πεδία όπως η αναγνώριση συγγραφέα, η ανίχνευση λογοκλοπής, η ανίχνευση διαδικτυακής παρενόχλησης και κακοποίησης. Επιπλέον, η παρούσα διατριβή είναι η πρώτη εκτεταμένη μελέτη της ιδιολέκτου στην ελληνική γλώσσα με τη

χρήση τεχνικών μηχανικής μάθησης, γεγονός που υποδηλώνει ότι οι λεξικές ενθέσεις μπορούν να εφαρμοστούν με επιτυχία σε ευρύ φάσμα ερευνητικών περιοχών που αφορούν την ελληνική γλώσσα.

Abstract

Idiolect, as a term in linguistics, refers to the unique and distinctive use of language by an individual and is the individual counterpart of sociolect. Research on idiolect has so far been rather neglected in sociolinguistics, especially as concerns its validation by empirical means. Research on idiolect in corpus linguistics and stylometry has also been limited in terms of either the number of authors examined (typically less that 10 authors) or the number of vocabulary items used in the examination of idiolectal similarity (up to 310 functional words).

This thesis employs learning distributed representations or lexical embeddings to analyse texts by social media users that are considered to reflect their writing style. Data include a Twitter corpus of Greek texts, posted by 4.494 users from 2009 to 2016 (325 million words approx.) and the Blog Authorship Corpus, used for comparison. Based on Zellig Harris' Distributional Hypothesis, according to which semantically similar words tend to appear in the same contexts, the notion of lexical (or word) embeddings can be used to answer the question of idiolect, providing thus a stylistic fingerprint for the authors involved.

The performance of various models of distributed representation are explored and compared; in particular, these involve lexical embeddings produced by Neural Probabilistic Language models (namely, word2vec, fastText and doc2vec) and matrix factorization (namely, GloVe). The selected models are applied to the entire vocabulary of the texts concerned and thus are not limited by corpus vocabulary size and are scalable to thousands of authors.

It is found that idiolect embeddings a) can be used to represent the style of individual authors and b) can provide the means of clustering users in terms of their idiolectal similarity, revealing clusters of the same style, as well as the means of quantifying idiolect stability over time. The findings have considerable applications in areas such as authorship attribution, plagiarism detection, online harassment and abuse. Furthermore, this is the first extended study of idiolect in Greek texts, using machine learning methods, something which suggests that lexical embeddings can be fruitfully employed in further areas of research in this language.

Γλωσσάρι όρων

Ανάκτηση πληροφοριών	Information retrieval
Ανάστροφη διάδοση σφάλματος	Back propagation
Απόδοση συγγραφέα	Authorship Attribution
Αρχιτεκτονικές Κωδικοποίησης-Αποκωδικοποίησης	Encoder-Decoder Architectures
Βαθιά (Μηχανική) Μάθηση	Deep Learning
Γνώση Βασικού Επιπέδου	Base Rate Knowledge
Γράφος κοινωνικού δικτύου	Social graph
Διανυσματικός χώρος	Vector Space
Δίγραμμα	Bigram
Διεντροπία	Cross Entropy
Δικανική γλωσσολογία	Forensic linguistics
Εκμάθηση αναπαραστάσεων	Representation learning
Ενθέσεις λέξεων	Word embeddings
Επαλήθευση συγγραφέα	Authorship verification
Επιβλεπόμενη (Μηχανική) Μάθηση	Supervised Learning
Κατανεμητική σημασιολογία	Distributional semantics
Μεταφορά Μάθησης	Transfer Learning
Μη Επιβλεπόμενη (Μηχανική) Μάθηση	Unsupervised Learning
Μηχανική Μάθηση	Machine Learning
Ν-γράμματα	N-grams
Νευρωνικό δίκτυο	Neural network
Παραγοντοποίηση πίνακα	Matrix factorisation
(Στοχαστική) Κατάβαση Κλίσης	(Stochastic) Gradient Descent
Συνάρτηση κόστους	Loss function, Cost Function
Σωρός λέξεων	Bag of words

Περιεχόμενα

Διαγράμματα

Πίνακες

Κεφάλαιο 1

Εισαγωγή

Στο κεφάλαιο αυτό τίθεται το ερευνητικό ερώτημα της διδακτορικής διατριβής, που αφορά την ανίχνευση της ιδιολέκτου σε κείμενα που δημιουργούνται σε κοινωνικά δίκτυα. Το ερευνητικό ερώτημα της διατριβής θα πρέπει να τοποθετηθεί στο συγκεκριμένο χωροχρονικό πλαίσιο, στο οποίο διαπιστώνεται εκθετικός ρυθμός αύξησης του περιεχομένου στο διαδίκτυο, καθώς επίσης και ραγδαίες εξελίξεις στην έρευνα της Τεχνητής Νοημοσύνης και ειδικότερα της Μηχανικής Μάθησης. Ο μεγάλος όγκος δεδομένων δίνει πλέον τη δυνατότητα εκπαίδευσης περίπλοκων και αποδοτικών συστημάτων Επεξεργασίας Φυσικής Γλώσσας, και αυτή η δυνατότητα αξιοποιείται στο πλαίσιο της παρούσας διατριβής στο ειδικότερο κειμενικό είδος που μελετάται. Στο παρόν κεφάλαιο παρουσιάζεται επίσης η δομή της παρούσας διατριβής όσον αφορά την οργάνωση των κεφαλαίων που ακολουθούν και συνοπτικά η συμβολή της διατριβής στην μελέτη της έννοιας της ιδιολέκτου στο πλαίσιο της Γλωσσολογίας Σωμάτων Κειμένων.

1.1 Ερευνητικό ερώτημα και συμβολή της διατριβής

Στην παρούσα διατριβή μελετώνται από τη σκοπιά της Υπολογιστικής Γλωσσολογίας και της Γλωσσολογίας Σωμάτων Κειμένων οι έννοιες της ιδιολέκτου και της ιδιολεκτικής ομοιότητας, καθώς επίσης και της μεταβολής των χαρακτηριστικών τους στο πέρασμα του χρόνου. Πιο συγκεκριμένα, τα ερευνητικά ερωτήματα που τίθενται είναι τα ακόλουθα:

1. Είναι εφικτή η ανίχνευση της ιδιολέκτου με υπολογιστικές μεθόδους;
2. Μπορεί η υπολογιστική προσέγγιση στην ανίχνευση της ιδιολέκτου να χρησιμοποιηθεί για την ομαδοποίηση συγγραφέων με βάση το ύφος;

3. Διαπιστώνεται και, αν ναι, σε ποιο βαθμό μεταβολή στην ιδιολεκτική συμπεριφορά των συγγραφέων στο πέρασμα του χρόνου

Η μελέτη επικεντρώνεται στο ιδιολεκτικό ύφος από υπολογιστική/μαθηματική σκοπιά με έμφαση στα κείμενα που παράγονται κατά την επικοινωνία χρηστών στα κοινωνικά δίκτυα (Kaplan & Haenlein, 2010, σ. 60)[1]. Απώτερος στόχος είναι να αξιολογηθεί και να ποσοτικοποιηθεί η έννοια της ιδιολέκτου και της ομοιότητας μεμονωμένων συγγραφέων, αλλά και να διερευνηθεί η κατασκευή ομάδων όμοιων ως προς το κειμενικό ύφος σε σώματα κειμένων. Εξετάζονται δεδομένα από δύο σώματα κειμένων: α) ένα προερχόμενο από κοινωνικά δίκτυα, το οποίο καλύπτει ένα χρονικό φάσμα 8 ετών (2008-2016) και συλλέχθηκε για τους σκοπούς αυτής της διατριβής, και β) το σώμα Blog Authorship Corpus των Schler et al. (2006). Στις εξεταζόμενες παραμέτρους περιλαμβάνονται η ταχεία μεταβολή και η προσαρμογή των γλωσσικών προτιμήσεων και συμπεριφορών των χρηστών των κοινωνικών δικτύων, η μηχανική της προσαρμογής των (νέων) ομιλητών μιας γλώσσας σε ένα ανταγωνιστικό περιβάλλον, όπως αυτό των κοινωνικών δικτύων, και η υιοθέτηση κειμενικού ύφους ως συνάρτηση και απεικόνιση της κοινωνικοπολιτικής τους κοσμοθεωρίας. Τέλος, μελετάται η διείσδυση και η μεταβολή των γλωσσικών αυτών χαρακτηριστικών στα παραδοσιακά Μέσα, κυρίως σε θέματα που άπτονται πολιτικής (Alejandro, 2010; Jørgensen et al., 2015).

Το διάστημα των τελευταίων δέκα ετών χαρακτηρίζεται από ραγδαίες αλλαγές στο κοινωνικοπολιτικό κατεστημένο του δυτικού, και όχι μόνο, πολιτισμού. Η διάχυση αυτών των αλλαγών διευρύνεται από την καταλυτική παρουσία της τεχνολογίας ως μέσου μεταφοράς πληροφοριών, καθώς επίσης και από την ευκολία πρόσβασης στις πληροφορίες αυτές. Ο όγκος και η ταχύτητα παραγωγής, αναπαραγωγής και μεταποίησης του περιεχόμενου αυξάνει με εκθετικούς ρυθμούς σε ετήσια βάση, καθώς νέα μέσα και δίαυλοι πρόσβασης στο περιεχόμενο είναι πλέον προσβάσιμα από την πλειονότητα των χρηστών. Το βάρος της ενημέρωσης και της διάδοσης πληροφοριών με τη μορφή της είδησης, και ενίοτε προπαγάνδας, έχει μεταφερθεί στο διαδίκτυο και ως επί το πλείστον, στα κοινωνικά δίκτυα (Shirky, 2011; Woolley & Howard, 2018; Gerbaudo, 2018; Farkas et al., 2018).

Αξιοσημείωτο αποτέλεσμα της μαζικής παραγωγής και κατανάλωσης περιεχομένου, όπως θα διαπιστωθεί στη συνέχεια, είναι η ταχύτητα με την οποία οι χρήστες των κοινωνικών δικτύων και πιο συγκεκριμένα του twitter, στο οποίο επικεντρώνεται η παρούσα μελέτη, υιοθετούν και μεταβάλλουν τις κειμενικές και υφολογικές τους προτιμήσεις στην καθημερινή επικοινωνία τους με άλλους χρήστες. Είναι σημαντικό να παρατηρήσουμε επίσης πως ο περιορισμός στο

[1]Η συγκεκριμένη δημοσίευση αναφέρει ως στατιστικό χρήσης της πλατφόρμας facebook τα 175 εκατομμύρια χρηστών, που είχε το 2010. 8 χρόνια αργότερα ο αριθμός των χρηστών είναι κοντά (ή έχει ξεπεράσει) τα δύο δισεκατομμύρια χρήστες

μήκος των μηνυμάτων που επιβάλλεται από το μέσο, αλλά και οι δημοφιλείς χρήστες-κόμβοι επηρεάζουν δομικά το συνολικό ύφος της γλωσσικής παραγωγής στο συγκεκριμένο μέσο και, κατά μια έννοια, επιβάλλουν τάσεις και ύφος σε σχετικές υποομάδες χρηστών (Bryden et al., 2018; Shapiro et al., 2018).

Η άντληση κειμενικών δεδομένων από κοινωνικά δίκτυα ως βασικό συστατικό έρευνας και εφαρμογών είναι πλέον ιδιαίτερα διαδεδομένη στη βιβλιογραφία και έχει αξιοποιηθεί σε ένα ευρύ φάσμα εφαρμογών, όπως η ανάλυση συναισθημάτων (sentiment analysis), η ταξινόμηση προσωπικότητας (personality classification), η απόδοση συγγραφέα (authorship attribution), η εξόρυξη γνώμης (opinion mining), και πιο πρόσφατα η αξιοπιστία των ειδήσεων και η πιστοποίηση γεγονότων (fake news detection / fact checking) (Bamman et al., 2014; Mikros & Perifanos, 2013; Schwartz et al., 2013; Golbeck et al., 2011; Liu, 2012).

Η έκρηξη στο ρυθμό παραγωγής περιεχόμενου καθώς επίσης και οι πρόσφατες εξελίξεις στο πεδίο της Τεχνητής Νοημοσύνης και πιο συγκεκριμένα της Μηχανικής/Βαθιάς Μάθησης (Machine/Deep Learning), οι οποίες συμβαδίζουν με την πληθώρα υπολογιστικών εργαλείων, που είναι πλέον διαθέσιμα στην ερευνητική κοινότητα, δίνουν τη δυνατότητα για ανάλυση τεράστιων όγκων δεδομένων, που είναι πλέον σχετικά εύκολο να συλλεχθούν, να αποθηκευτούν και να αποτελέσουν αντικείμενο επεξεργασίας (Deng, 2014; Schmidhuber, 2015; Abadi et al., 2016; Paszke et al., 2017). Έτσι, μπορεί να θεωρηθεί εύλογη η επέκταση των τεχνικών της Μηχανικής Μάθησης στο ερώτημα της ιδιολέκτου.

Ένα χρήσιμο εφαλτήριο για τη μελέτη της ιδιολέκτου είναι η παραδοχή ύπαρξης της ιδιολέκτου στο πλαίσιο της Δικανικής Γλωσσολογίας (Forensic Linguistics), όπως ορίζεται και αναπτύσσεται από τον Coulthard (2004). Η δικανική γλωσσολογία έχει ως αντικείμενο την εφαρμογή της γλωσσολογικής ανάλυσης σε νομικό πλαίσιο και εκτείνεται σε εφαρμογές όπως η μελέτη της χρήσης της γλώσσας στη νομοθεσία, η ανάλυση της γλώσσας της νομικής διαδικασίας, καθώς επίσης και η χρήση της γλώσσας ως αποδεικτικού στοιχείου (language as evidence). Στην οπτική της γλώσσας ως αποδεικτικού στοιχείου περιλαμβάνονται εφαρμογές φωνητικής, όπως για παράδειγμα η ταυτοποίηση ομιλητή με βάση τα φωνητικά του χαρακτηριστικά, η διαπίστωση της λογοκλοπής, καθώς και η Απόδοση Συγγραφέα (Authorship Attribution), που αφορά κυρίως γραπτά κείμενα (Coulthard, 2004, σ. 5). Θεμελιώδους σημασίας στην απόδοση συγγραφέα είναι ακριβώς η παραδοχή της ύπαρξης ιδιολέκτου, η υπόθεση δηλαδή ότι ο δημιουργός κάθε κειμένου διαθέτει το προσωπικό του ύφος που συντίθεται από ιδιαίτερα χαρακτηριστικά, τα οποία τον διακρίνουν από κάθε άλλον δημιουργό κειμένου.

Όπως θα συζητηθεί στο επόμενο Κεφάλαιο, αν και η ιδιόλεκτος έχει θεωρηθεί ως σημαντική διάσταση της γλωσσικής ποικιλότητας, δεν έχει μελετηθεί επαρκώς από την κοινωνιογλωσσολογία, αλλά ούτε και από τη γενική γλωσσολογία ή συναφείς κλάδους όπως η (γλωσσολογική) υφολογία. Οι μόνες σχετικές μελέτες έχουν γίνει στο πλαίσιο της υπολογιστικής γλωσσολογίας και της γλωσσολογίας

σωμάτων κειμένων, χωρίς ωστόσο να έχουν λάβει υπόψη τους τις εξελίξεις στον όγκο των δεδομένων από τα κοινωνικά δίκτυα και τις τεχνικές μηχανικής μάθησης, όπως αναφέρθηκαν πιο πάνω. Οι μελέτες αυτές είναι ακόμη λιγότερες σε σώματα κειμένων στα ελληνικά, αν και ο όγκος δεδομένων που είναι διαθέσιμος στη γλώσσα αυτή έχει εξίσου αναπτυχθεί εκθετικά τα τελευταία χρόνια. Το κενό αυτό επιχειρεί να καλύψει η παρούσα διπλωματική διατριβή, αξιοποιώντας τις σύγχρονες τεχνικές μηχανικής μάθησης για να αναπτύξει μια υπολογιστική αναπαράσταση της ιδιολέκτου σε ελληνικά σώματα κειμένων.

Ειδικότερα, στη διατριβή αυτή υποστηρίζεται ότι το σύνολο των λέξεων με πλήρη σημασία (λεξικών τύπων) μπορεί να αξιοποιηθεί για την υπολογιστική ανίχνευση του υφολογικού αποτυπώματος των δημιουργών των κειμένων στο πλαίσιο της γλωσσολογίας σωμάτων κειμένων. Με αυτή την έννοια, μια κύρια συμβολή της διατριβής έγκειται στο ότι εισάγει τη σημασιολογική συνιστώσα ως κεντρικό σημείο αναφοράς στην ανάλυση της ιδιολέκτου.

1.2 Δομή της διατριβής

Η δομή της διατριβής είναι η ακόλουθη. Αρχικά, στο Κεφάλαιο 2 συζητώνται οι έννοιες της κοινωνιολέκτου και της ιδιολέκτου με αναφορά στην ιδιότητας της γλωσσικής ποικιλότητας και γίνεται λεπτομερής ανασκόπηση της βιβλιογραφίας που αφορά την έρευνα για την ιδιόλεκτο, με ειδική αναφορά στη σκοπιά της γλωσσολογίας σωμάτων κειμένων. Με βάση τη συζήτηση της μεθοδολογίας, των ευρημάτων και των περιορισμών της έρευνας αυτής αναπτύσσεται η μεθοδολογία της παρούσας διατριβής, η οποία διαφέρει σημαντικά σε σχέση με τις ως τώρα μελέτες. Τα κομβικά σημεία στα οποία η παρούσα διατριβή διαφοροποιείται αφορούν το μέγεθος του λεξιλογίου που χρησιμοποιείται για τον καθορισμό και την ανάλυση της ιδιολέκτου και η προσέγγιση που ακολουθείται, η οποία αναπτύσσει την Κατανεμητική Υπόθεση του Harris (1954) και εντάσσεται στο πεδίο της Κατανεμητικής Σημασιολογίας.

Στη συνέχεια, στο Κεφάλαιο 4 αναφερόμαστε αρχικά στις αναπαραστάσεις διανυσματικών χώρων, στη βασική έννοια των λεξικών ενθέσεων και στην εκμάθηση αναπαραστάσεων με τη χρήση Μηχανικής Μάθησης. Η προτεινόμενη μεθοδολογία, που εμπλέκει τις έννοιες αυτές, εφαρμόζεται σε δύο σώματα κειμένων, τα οποία περιγράφονται στο Κεφάλαιο 3. Στο Κεφάλαιο 5 περιγράφεται η διαδικασία εκπαίδευσης μοντέλων και εξάγονται τα αποτελέσματα έτσι ώστε να απαντηθούν τα ερευνητικά ερωτήματα που τέθηκαν στην αρχή του παρόντος κεφαλαίου. Τέλος, στο Κεφάλαιο 6 συνοψίζονται τα αποτελέσματα της διατριβής και περιγράφονται εφαρμογές και μελλοντικές κατευθύνσεις της σχετικής έρευνας.

Κεφάλαιο 2

Γλωσσική ποικιλότητα και ιδιόλεκτος

> Recognizing that each person has an idiosyncratic personal dialect, linguists long ago coined the term idiolect. And it's not just vocabulary; it's everything from how we pronounce certain words to how we put them together to what we imagine they mean (Gretchen McCulloch)

Στο κεφάλαιο αυτό παρουσιάζεται το θεωρητικό υπόβαθρο της έννοιας της ιδιολέκτου από τη σκοπιά της γλωσσικής ποικιλότητας. Το κεφάλαιο εκκινεί από τη συζήτηση της ποικιλότητας και ποικιλιών όπως η κοινωνιόλεκτος και η διάλεκτος και επικεντρώνεται στην ιδιόλεκτο, που έχει συζητηθεί στη βιβλιογραφία κυρίως ως πτυχή του γλωσσικού συστήματος. Διαπιστώνεται η έλλειψη εμπειρικής τεκμηρίωσης της έννοιας της ιδιολέκτου, που συμβαδίζει με τη σχετική απουσία εκτεταμένων κοινωνιογλωσσικών ερευνών. Στην ενότητα 2.4 εξετάζονται οι έρευνες των Barlow, Mollin και Φραγκάκη, στα ελληνικά, στην ενότητα 2.5 οι υφομετρικές προσεγγίσεις στην ιδιόλεκτο και στην ενότητα 2.6 οι έρευνες των Hughes et. al. και Bamman et. al., που αποτελούν τις μόνες απόπειρες προσδιορισμού της ιδιολέκτου μέσω σωμάτων κειμένων. Υπογραμμίζεται ότι το πέρασμα από την αντιμετώπιση της ιδιολέκτου ως πτυχής του γλωσσικού συστήματος στη θεώρησή της ως πτυχής της γλωσσικής πραγμάτωσης και ο περαιτέρω προσδιορισμός της ως προβλήματος ανάκτησης πληροφοριών και ως διανυσματικό πρόβλημα μεγάλης κλίμακας θέτει τη βάση για τον πληρέστερο

καθορισμό της ιδιολέκτου με εκτεταμένη εμβέλεια όπως επιχειρείται στην παρούσα διατριβή.

2.1 Γλωσσικές ποικιλίες

Η γλωσσική ποικιλότητα (language variation), δηλαδή η διαρκής και πανταχού παρούσα ύπαρξη διαφορών στον τρόπο που χρησιμοποιείται μια γλώσσα, αποτελεί μια βασική ιδιότητα των ανθρώπινων γλωσσών (βλ.Γούτσος, 2012), αλλά και «ζωτικό μέρος της κανονικής γλωσσικής συμπεριφοράς» (Trask, 2007, σ. 319). Σύμφωνα με τη διατύπωση στο Γούτσος (2012, σ. 101), «όλες οι γλώσσες του κόσμου και κάθε γλωσσικό σύστημα χωριστά δεν υπάρχουν παρά μόνο σε παραλλαγές που διαφέρουν περισσότερο ή λιγότερο μεταξύ τους». Σύμφωνα με τους Evans & Levinson (2009, σ. 446) μάλιστα, η διαφορετικότητα των γλωσσών αποτελεί, από βιολογική σκοπιά, την πιο θαυμαστή γλωσσική ιδιότητα, καθώς δεν υπάρχει άλλο ζώο που το επικοινωνιακό του σύστημα να διαφέρει τόσο στη μορφή όσο και στο περιεχόμενο. Με αυτή την έννοια, η γλωσσική ποικιλότητα αποτελεί γι' αυτούς το μόνο πραγματικό καθολικό χαρακτηριστικό της γλώσσας και το κεντρικό ερώτημα που πρέπει να απαντηθεί από τη θεωρία της ανθρώπινης επικοινωνίας. Ιδιαίτερο ενδιαφέρον έχει και η παρατήρηση της Eckert (2012, σ. 94), που συνοψίζει την παράδοση τεσσάρων δεκαετιών κοινωνιογλωσσολογικής έρευνας, τονίζοντας ότι το θεωρητικό υπόβαθρο του λεγόμενου τρίτου κύματος στην κοινωνιογλωσσολογία υπογραμμίζει ότι η γλωσσική ποικιλότητα δεν αποτελεί ένα συμπτωματικό παρεπόμενο της κοινωνικής διαστρωμάτωσης, αλλά ουσιώδες χαρακτηριστικό της γλώσσας.

Η ποικιλότητα εκφράζεται στη γλώσσα με τη μορφή γλωσσικών ποικιλιών. Ως γλωσσική ποικιλία (language variety) ορίζεται το σύνολο των γλωσσικών χαρακτηριστικών (π.χ. φωνολογικών, μορφολογικών κ.ά.) που χαρακτηρίζουν μια χρήση της γλώσσας (O'Grady et al., 1997, σ. 374). Τα χαρακτηριστικά αυτά μπορεί να εξαρτώνται από γεωγραφικούς, εθνολογικούς, κοινωνικούς ή πολιτισμικούς παράγοντες. Ιδιαίτερα χρήσιμη είναι η διάκριση του Halliday (1978) ανάμεσα σε διαλεκτικές και διατυπικές ποικιλίες, που αντιστοιχεί στη διαφοροποίηση της γλώσσας σύμφωνα με τον χρήστη της (τα ιδιαίτερα χαρακτηριστικά του ομιλητή) και τη χρήση της από κάθε ομιλητή. Η διάκριση αυτή συνοψίζεται στον πίνακα 2.1 (βλ. Γούτσος, 2012, σ. 104 , Αρχάκης & Κονδύλη, 2004, σ. 57):

Πίνακας 2.1: Ποικιλότητα

Χρήστης	Χρήση: Διατυπικές ποικιλίες
Διάλεκτοι Κοινωνιόλεκτοι	Μέσο Τόνος Πεδίο

Ένας εναλλακτικός όρος για τη γλωσσική ποικιλία είναι η λέκτος (lect), που προέρχεται αναγωγικά από τους όρους για επιμέρους ποικιλίες (διάλεκτος, κοινωνιόλεκτος, εθνόλεκτος κ.ά.) και με τον οποίο αποφεύγεται να δηλωθεί αν οι ποικιλίες ανήκουν στην ίδια ή σε διαφορετικές γλώσσες.

Τέλος, μια συναφής έννοια είναι αυτή της γλωσσικής μεταβλητής (language variable), ενός γλωσσικού χαρακτηριστικού (π.χ. φωνολογικού, μορφολογικού κ.λπ.) που έχει δύο τουλάχιστον διακριτές παραλλαγές (variants), σύμφωνα με τον Trask (2007, σ. 3015). Η έννοια της γλωσσικής μεταβλητής είναι σύμφυτη με την εμφάνιση και εξέλιξη του κλάδου της κοινωνιογλωσσολογίας και συνδέει την ποικιλία πραγμάτωσης στη γλωσσική χρήση με ενδογλωσσικούς και εξωγλωσσικούς παράγοντες (Mikros, 2008, 69 κ. εξ).

2.2 Κοινωνιόλεκτος και ιδιόλεκτος

Στην ποικιλότητα σύμφωνα με τον χρήστη της γλώσσας η κύρια αντίθεση είναι αυτή μεταξύ διαλέκτου και κοινωνιολέκτου. Η διαφοροποίηση αυτή αφορά το πλαίσιο στο οποίο ορίζονται και αναφέρονται. Η διάλεκτος περιορίζεται γεωγραφικά, καθώς στη γλωσσική ποικιλία που χρησιμοποιείται σε μια περιοχή διαπιστώνονται συγκεκριμένοι φωνολογικοί, μορφοσυντακτικοί ή λεξικοί κανόνες. Αντίθετα, η κοινωνιόλεκτος εντοπίζεται σε κοινωνικά υποσύνολα που ορίζονται με βάση την ηλικία, την κοινωνική τάξη, το φύλο ή/και την σεξουαλική ταυτότητα ή την εθνικότητα.

Ως κοινωνιόλεκτος ορίζεται η γλωσσική ποικιλία που σχετίζεται με το κοινωνικό υπόβαθρο του ομιλητή σε αντίθεση με το γεωγραφικό (Trudgill, 2000, σσ. 2,7). Η ιδέα της κοινωνιολέκτου απαντά αρχικά στη διαλεκτολογία, τη μελέτη διαφορετικών διαλέκτων, σε αντιδιαστολή με κοινωνικές ομάδες σε χώρες όπως η Αγγλία, αλλά μόνο πρόσφατα μετατράπηκε σε πεδίο αυξημένου ενδιαφέροντος.

Ωστόσο, αντίθετα με τη διάλεκτο, όπως θα δούμε και στη συνέχεια, η ουσία της κοινωνιολέκτου έγκειται στο ότι ο ομιλητής υιοθετεί συγκεκριμένο ύφος ανάλογα ή σύμφωνα με την κοινωνική ομάδα στην οποία ανήκει ή επιθυμεί να ανήκει. Η συμπεριφορά αυτή αποτελεί επίσης συνάρτηση της εθνικότητας, της ηλικίας, του φύλου, των πολιτικών πεποιθήσεων και του χρονικού πλαισίου στο οποίο λαμβάνει χώρα το γεγονός της πραγμάτωσης της έκφρασης του ομιλητή. Σύμφωνα με την κοινωνιογλωσσολογική προσέγγιση του Labov, η κοινωνιόλεκτος είναι αυτό που προγραμματιστήκαμε να μιλάμε με τέτοιο τρόπο που να ταιριάζει με το γενικό περιβάλλον των κοινοτήτων στις οποίες ζούμε (Labov, 1972, σσ. 18,20). Επομένως, μπορούμε να θεωρήσουμε ότι η κοινωνιόλεκτος αφορά την προσαρμογή του ιδιαίτερου τρόπου ομιλίας κάθε ομιλητή στο άμεσο (ή ευρύτερο) κοινωνικό περιβάλλον.

Η Eckert Eckert (2012, σσ. 93-94) θεωρεί ότι το τρίτο κύμα της

κοινωνιογλωσσικής έρευνας σημαίνει τη μετάβαση από τη θεώρηση της κοινωνιολέκτου ως αντανάκλαση κοινωνικών ταυτοτήτων και κατηγοριών στην προσέγγισή της ως ενός συνόλου γλωσσικών πρακτικών μέσω των οποίων οι ομιλητές τοποθετούν τον εαυτό τους στο κοινωνικό πεδίο. Στη διαδικασία αυτή βασική είναι η έννοια του ύφους, καθώς θεωρείται ότι οι ομιλητές υιοθετούν και δημιουργούν ποικίλες υφολογικές πρακτικές από ένα ευρύ φάσμα σημειωτικών πρακτικών που διαθέτουν συγκεκριμένα νοήματα. Έτσι, οι ομιλητές δεν θεωρούνται πλέον παθητικοί και σταθεροί φορείς μιας λέκτου, αλλά ενεργοί υφολογικοί δράστες που συμμετέχουν σε ιδεολογικά φορτισμένες πρακτικές ταύτισης και διαφοροποίησης του εαυτού Eckert (2012, σσ. 97-98). Η γλωσσική ποικιλία, επομένως, κατασκευάζει κοινωνικά νοήματα και δεν τα αντανακλά απλώς και συνεπώς αποτελεί δύναμη κοινωνικής μεταβολής. Όπως θα διαπιστώσουμε στη συνέχεια αυτό το χαρακτηριστικό είναι καθοριστικός παράγοντας στη δημιουργία υποομάδων που διαθέτουν κοινά χαρακτηριστικά ύφους. Σε κάθε περίπτωση, η έμφαση στην κοινωνιογλωσσική έρευνα βρίσκεται στα κοινωνικά νοήματα και τις κατασκευές στις οποίες συμμετέχει η υφολογική πρακτική των επιμέρους ομιλητών, παρά στις ιδιολεκτικές τους πραγματώσεις

Παρομοίως, όπως σημειώνουν οι Bayley και Lucas, η μελέτη της κοινωνιολέκτου έχει επίσης ιδιαίτερη βαρύτητα σε θέματα σχετιζόμενα με τη δικανική γλωσσολογία (Bayley & Lucas, 2007). Τυπικές περιπτώσεις τέτοιων ερευνών συμπεριλαμβάνουν την ταυτοποίηση φωνής, την ταυτοποίηση συγγραφέα, τη λογοκλοπή, αλλά και πιο σύνθετες περιπτώσεις, στις οποίες η κοινωνιόλεκτος χρησιμοποιείται ως αποδεικτό υλικό (Coulthard, 2004), όπως για παράδειγμα την περίπτωση διεκδίκησης της περιοχής Bear Island στον Καναδά από τους Temagami[1], φυλή Ινδιάνων του Καναδά (Chambers, 1990, σ. 186), (Bray & Thomson, 1996, σ. 103). Και σε αυτές τις περιπτώσεις η ιδιόλεκτος των ομιλητών θεωρείται ότι αντανακλά την ευρύτερη κοινωνιόλεκτο στην οποία εντάσσεται η ομιλία τους.

Γενικά, τόσο η παραδοσιακή όσο και η σύγχρονη κοινωνιογλωσσολογία επισημαίνει ότι το κοινωνικό περιβάλλον ενός ατόμου καθορίζει πώς μιλάει και γράφει και συνεπώς τις γλωσσικές πράξεις και τους συσχετισμούς του με άλλα μέλη του περιβάλλοντος αυτού. Αυτό το φαινόμενο είναι άμεσα παρατηρήσιμο

[1]Η υπόθεση αφορά τη διαμάχη μεταξύ της Καναδικής κυβέρνησης και της φυλής Ινδιάνων Temagami, στην περιοχή του Οντάριο, κατά το χρονικό διάστημα 1982-1984. Οι Temagami ισχυρίζονταν κτήση του νησιού Bear Island της λίμνης Οντάριο για περισσότερο από έναν αιώνα, αντιτιθέμενοι στα πλάνα της κυβέρνησης για κατασκευή θερέτρου. Στη δίκη διάρκειας περίπου 100 ημερών κλήθηκαν να καταθέσουν γλωσσολόγοι, ανθρωπολόγοι, ιστορικοί και δικηγόροι. Οι Temagami τελικά έχασαν τη δίκη με τον δικαστή Donald Steele να χαρακτηρίζει τις καταθέσεις των ειδικών ως «νεφελώδεις (nebulus»), (Chambers, 1990, σ. 12). Οι Temagami άσκησαν έφεση και το 1986 η κυβέρνηση πρότεινε ως αποζημίωση το ποσό των 30 εκατομμυρίων δολαρίων ως εξώδικο συμβιβασμό, για να μη καθυστερήσει περαιτέρω η κατασκευή του θερέτρου.

στις υποομάδες των κοινωνικών δικτύων, όπου διαφορετικές ομάδες τείνουν να χρησιμοποιούν ιδιαίτερες εκφράσεις-δείκτες, που χαρακτηρίζουν τις ομάδες χρηστών στο κοινωνικό γράφο (graph), τις συνδεδεμένες υποομάδες δηλαδή που δημιουργούνται από τη δυνατότητα που παρέχεται στους χρήστες να ακολουθούν άλλους χρήστες. Στη βιβλιογραφία έχει μελετηθεί εκτενώς το φαινόμενο της κοινωνιολέκτου στα κοινωνικά δίκτυα (Gianfortoni et al., 2011; Reynolds et al., 2013), κυρίως όσον αφορά τη γεωγραφική διαφοροποίηση των γλωσσικών χαρακτηριστικών των χρηστών, λαμβάνοντας υπόψη, επίσης, ότι το γεωγραφικό φράγμα αποδυναμώνεται στο διαδίκτυο και ο χρήστης έχει ουσιαστικά τη δυνατότητα να επιλέξει την κοινωνική ομάδα που θεωρεί ότι τον εκφράζει καλύτερα.

Η παρατήρηση αυτή είναι κομβική καθώς ορίζει μια τομή μεταξύ διαλέκτου και κοινωνιολέκτου και έχει ως αποτέλεσμα συγκεκριμένα υφολογικά μοτίβα όσον αφορά τις προτιμήσεις των χρηστών, με βάση την ομάδα που έχουν επιλέξει να ακολουθήσουν. Τυπικά παραδείγματα συμπεριφορών, όπως αυτές στα κοινωνικά δίκτυα, είναι τυποποιημένες αστείες εκφράσεις τα μιμίδια (memes), δηλαδή (Shifman, 2013) με σχετικά βραχύ χρόνο ζωής (συνήθως 1 με 6 μήνες), που υιοθετούνται από συγκεκριμένες υποομάδες χρηστών και αποτελούν φορέα αναγνώρισης. Άλλα παραδείγματα αυτής της συμπεριφοράς, με σχετικά μεγαλύτερη διάρκεια ζωής, αφορούν πολιτικά φορτισμένα κείμενα που χρησιμοποιούνται ως μέσα προπαγάνδας ή αντιπροπαγάνδας. Για παράδειγμα, στα δεδομένα του σώματος κειμένων που μελετώνται στη διατριβή συγκεκριμένοι λεξικοί τύποι χρησιμοποιούνται σε ειρωνικό περιβάλλον με σκοπό την αποδυνάμωση και τον χλευασμό αντίπαλων ιδεολογιών.[2]

Στην παρούσα διατριβή δεν λαμβάνονται υπ' όψιν δεδομένα που σχετίζονται και μπορούν να χρησιμοποιηθούν για την κοινωνιογλωσσολογική ανάλυση του ιδιολεκτικού ύφους, καθώς στόχος είναι η μελέτη του ύφους αποσυνδεδεμένου από τέτοιες μεταβλητές. Αντίθετα, εστιάζει στην εκμάθηση της αναπαράστασης του ύφους και, στη συνέχεια, εφόσον υπάρχουν σχετικά δεδομένα, στο να γίνει απόπειρα συσχέτισης ύφους και ευρύτερων κοινωνιογλωσσολογικών χαρακτηριστικών.

[2] Ένα τέτοιο παράδειγμα αποτελεί η κλιμακούμενη αύξηση της χρήσης της λέξης *Σοβιετία* κατά το διάστημα 2010-2016 τόσο στα κοινωνικά δίκτυα (Βλ. γράφημα Ε.6 στο παράρτημα Ε) αλλά και η μεταφορά της σε παραδοσιακά μέσα καθώς επίσης και στον πολιτικό λόγο.

2.3 Ιδιόλεκτος: Από το γλωσσικό σύστημα στην πραγμάτωση

Ο όρος *ιδιόλεκτος* (*η ιδιόλεκτος*, ή εναλλακτικά *το ιδιόλεκτο*) αναφέρεται στην ιδιαίτερη και μοναδική χρήση της γλώσσας, είτε πρόκειται για τον γραπτό είτε τον προφορικό λόγο ενός ομιλητή (Nguyen et al., 2016). Τα χαρακτηριστικά της ιδιαίτερης χρήσης της γλώσσας περιλαμβάνουν, μεταξύ άλλων, το λεξιλόγιο, τη γραμματική και την προφορά. Συνεπώς, η ιδιόλεκτος μπορεί να θεωρηθεί ως το χαρακτηριστικό αποτύπωμα της χρήσης της γλώσσας ενός ατόμου, κατ' αντιστοιχία με την κοινωνιόλεκτο που ορίζεται ως το σύνολο των κοινών γλωσσικών χαρακτηριστικών στο πλαίσιο επικοινωνίας μιας κοινωνικής ομάδας.

Αν ακολουθήσουμε τη διάκριση του Chomsky σε Εσωτερική γλώσσα (internalized language, I-language) και Εξωτερική γλώσσα (externalized language, E-language), μπορούμε να διακρίνουμε ανάμεσα στην εσωτερική, μη παρατηρήσιμη γλωσσολογική γνώση κάθε ατόμου και στην παρατηρήσιμη πραγμάτωση, το προϊόν δηλαδή που παράγεται κατά την διαδικασία της ομιλίας/γραφής, αντίστοιχα. (I/E language distinction, Chomsky, 1986, σσ. 22-44). Οι Maguire & McMahon (2011) αναφερόμενοι στον διαχωρισμό Εξωτερικής και Εσωτερικής γλώσσας επσημαίνουν ότι το μοναδικό πραγματικό αντικείμενο της γλωσσολογίας υπό αυτή την προσέγγιση είναι η ιδιόλεκτος ως προσωπική γραμματική κάθε ομιλητή.

Ταυτόχρονα, το σύνολο της γλώσσας μπορεί να θεωρηθεί ένα άθροισμα ιδιολέκτων. Σύμφωνα με τον Zuckermann (2006), μια γλώσσα είναι περισσότερο ένα αφηρημένο «άθροισμα» ιδιολέκτων, κοινωνιολέκτων και διαλέκτων, παρά μια αυτόνομη οντότητα. Ο Zuckermann παραλληλίζει μάλιστα την γλώσσα ως ζωικό είδος (species) σε αντιδιαστολή με έναν αυτόνομο οργανισμό. Η θεώρηση αυτή έρχεται σε αντίθεση με την παραδοσιακή γλωσσολογική αντιμετώπιση της ατομικής χρήσης της γλώσσας, όπως θα δούμε και στη συνέχεια, αποτελεί όμως την βάση στην οποία στηρίζεται η απόδοση συγγραφέα και το ευρύτερο πεδίο της δικανικής γλωσσολογίας. Βέβαια, η θεώρηση ότι μια γλώσσα αποτελεί το άθροισμα διακριτών, ατομικών ιδιολέκτων πρέπει να λάβει υπόψη το γεγονός ότι τα μέλη μεγάλων γλωσσικών κοινοτήτων ή ακόμα και ομιλητές διαφορετικών διαλέκτων της ίδιας γλώσσας μπορούν να καταλάβουν ο ένας τον άλλον. Η ατομικότητα, επομένως, συμβαδίζει με την κοινωνικότητα της γλώσσας: σύμφωνα με τον Γούτσο (2012, σ. 20), «η γλώσσα διαπλέκεται άρρηκτα με την ατομική μας συνείδηση και μόνο έτσι μπορεί να λειτουργήσει ως κοινωνικό γεγονός»

Στην ιστορικοσυγκριτική γλωσσολογία αποδίδεται η πρώτη σημαντική συζήτηση της έννοιας της γλωσσικής ατομικότητας. Η υπόθεση του Herman Paul ήταν ότι κάθε γλωσσική δημιουργία είναι πάντα το αποτέλεσμα ενός ατόμου

και μόνο: ”every linguistic creation is always the work of one single individual only” (Paul, 1890, σ. xliii). Ο Sapir διαχωρίζει την κοινωνική νόρμα και την ατομική έκφραση, σημειώνοντας ότι «Όλοι έχουμε ξεχωριστό τρόπο και στην γραφή και στην ομιλία» (Sapir, 1927, σσ. 903-904), ενώ ο Bloomfield (1933, σ. 45), σύγχρονος του Sapir, σημειώνει τη μοναδικότητα των ατομικών γλωσσικών επιλογών, υποστηρίζοντας πως, αν παρατηρήσουμε με την αρμόζουσα προσοχή, θα βρούμε ότι δεν υπάρχουν δύο άτομα -ή ίσως ακόμα και το ίδιο άτομο σε διαφορετικά χρονικά σημεία- με πανομοιότυπη χρήση γλώσσας.

Ο όρος *ιδιόλεκτος* (idiolect) προτάθηκε απ’ τον Bloch (1948, σ. 7), με προέλευση από τις ελληνικές ρίζες *ίδιος* και *λέγω* και σημασιολογικά αποδίδει την έννοια της «επιλεγμένης έκφρασης ή λέξης» (Kuhl, 2003, σ. 4). Ο Bloch όρισε την ιδιόλεκτο ως «το σύνολο των πιθανών αρθρώσεων/εκφράσεων ενός ομιλητή σε μια δεδομένη χρονική στιγμή χρησιμοποιώντας τη γλώσσα ως μέσο αλληλεπίδρασης με έναν άλλο ομιλητή». Διευκρινίζει ότι «η ιδιόλεκτος δεν είναι αποκλειστικά αυτό που ένας ομιλητής εκφέρει σε δεδομένη στιγμή στον χρόνο: είναι γενικότερα αυτό που θα μπορούσε δυνητικά να εκφράσει σε μια δεδομένη γλώσσα» (Bloch, 1948, σ. 7).

Κατά τη διάρκεια του εικοστού αιώνα, οι ορισμοί για την ιδιόλεκτο παρέμειναν σχετικά συνεπείς. Για παράδειγμα, σύμφωνα με τον Hockett, η ιδιόλεκτος αποτελεί «το σύνολο των συνηθειών ομιλίας ενός ατόμου σε ένα δεδομένο χρονικό σημείο» (Hockett, 1958, σ. 321). Ομοίως, ο DeCamp (1970, σ. 18) ορίζει την ιδιολεκτική γραμματική ως «ένα συγκεκριμένο πεπερασμένο σύνολο κανόνων της γλωσσικής ικανότητας ενός ομιλητή/ακροατή». Εντός του σύγχρονου πλαισίου της κοινωνιογλωσσολογίας, ο Dittmar (1996) περιγράφει την ιδιόλεκτο ως «γλώσσα ενός ατόμου, η οποία, εξαιτίας των κεκτημένων συνηθειών και των υφολογικών χαρακτηριστικών της ατομικότητας, διαφέρει από εκείνη των άλλων ατόμων» (Dittmar, 1996, σ. 111). Τείνει, λοιπόν, να υπάρχει ομοφωνία σε γλωσσολογικά πλαίσια ότι οι μεμονωμένοι ομιλητές/συγγραφείς διαθέτουν τα δικά τους ιδιαίτερα γλωσσικά δομικά σχήματα και προτιμήσεις.

Ωστόσο, όπως παρατηρεί ο Kredens (2002, σ. 45), ενώ υπάρχει συμφωνία ότι το ατομικό γλωσσολογικό ρεπερτόριο είναι κατά τον έναν ή τον άλλο τρόπο διακριτό και μοναδικό, η θέση αυτή δεν έχει υποστηριχτεί επαρκώς από την εμπειρική έρευνα. Για παράδειγμα, όπως διαπιστώθηκε στην προηγούμενη ενότητα, η ατομικότητα της χρήσης της γλώσσας δεν βρίσκεται στο επίκεντρο της κοινωνιογλωσσολογικής έρευνας, όπως επισημαίνει η Johnstone (1996, σσ. 13-14). Σύμφωνα με την ίδια, ο κύριος λόγος που οι κοινωνιογλωσσολόγοι επικεντρώνονται στο γλωσσικό σύστημα και όχι στον μεμονωμένο ομιλητή βρίσκεται στο ότι η κοινωνιογλωσσολογία ασχολείται ουσιαστικά με τις αφηρημένες δηλαδή συστηματικές αρχές της γλώσσας παρά με την πραγμάτωση τους, με το γλωσσικό σύστημα ή langue παρά με τη γλωσσική πραγμάτωση ή parole, σύμφωνα με τη διάκριση του Saussure. Ενδεικτικό της έλλειψης εμπειρικών

δεδομένων στην ερμηνεία της ιδιολέκτου είναι και το γεγονός ότι η υφολογία, το επιστημονικό πεδίο που ασχολείται κατεξοχήν με το προσωπικό ύφος του συγγραφέα, έχει αναπτυχθεί με κυρίως ποιοτικές και όχι ποσοτικές μεθόδους (βλ. ωστόσο 2.5 πιο κάτω).[3]

Το γεγονός ότι σχεδόν αποκλειστικά το βάρος της έρευνας επικεντρώνεται στο σύνολο και όχι στον μεμονωμένο ομιλητή γίνεται ξεκάθαρο από τον Labov (1972, σ. 277), που ορίζει τη γλώσσα ως εργαλείο που χρησιμοποιείται από τα μέλη μιας κοινότητας για τη μεταξύ τους επικοινωνία. Ιδιοσυγκρασιακές συνήθειες δεν θεωρούνται ότι αποτελούν τμήμα της γλώσσας και κατά συνέπεια ούτε και οι ιδιοσυγκρασιακές διαφοροποιήσεις αποτελούν αντικείμενο της έρευνας, καθώς η γλώσσα ερευνάται ως συνολικό γλωσσικό σύστημα και όχι ως άθροισμα ιδιαίτερων χαρακτηριστικών διαφορετικών ατόμων. Ως εκ τούτου, η κοινωνιογλωσσολογική έρευνα στη γλώσσα ενός ατόμου είναι μάλλον περιορισμένη. Στην περιθωριοποίηση της έρευνας για την ιδιόλεκτο συμβάλλει και η εγγενής δυσκολία συλλογής του όγκου δεδομένων που απαιτούνται ανά άτομο.

Ο Barlow (2013, σ. 1) σημειώνει ότι δεν υπάρχει άλλη περίπτωση στην γλωσσολογία με τόσο μεγάλο χάσμα μεταξύ της εμφανώς αναγνωρίσιμης ύπαρξης της έννοιας, της ιδιολέκτου δηλαδή, και της έλλειψης εμπειρικών δεδομένων για την περιγραφή του φαινομένου. Μια εξαίρεση αποτελεί η δικανική γλωσσολογία, στην οποία, όπως διαπιστώνει ο Coulthard (2004, σ. 431), η υπόθεση της ύπαρξης ιδιόλεκτου είναι η βάση στην οποία εφαρμόζεται η απόδοση και η αυτόματη αναγνώριση συγγραφέα.

Η αφηρημένη φύση της ιδιόλεκτου καθώς και το αν και κατά πόσο αποτελεί θεωρητικό και μόνο κατασκεύασμα, χωρίς εμπειρικά δεδομένα για την υποστήριξη της, οδήγησε στο να τεθούν ερωτήματα για την ωφελιμότητα της στη δικανική γλωσσολογία. Η Turell (2010, σσ. 216-217) αναγνωρίζει πως ατομικές ιδιόλεκτοι μπορούν να εντοπιστούν υπό την προϋπόθεση μιας άπειρης ποσότητας δεδομένων από κάθε άτομο και διατυπώνει τη θέση ότι η έννοια της ιδιολέκτου έχει ίσως γίνει αποδεκτή αξιωματικά και αυτή η αποδοχή της ύπαρξης της είναι πρόωρη. Αντί ιδιόλεκτου η Turell προτείνει το εναλλακτικό «ιδιολεκτικό ύφος» (idiolectal style), για το οποίο ισχυρίζεται ότι «θα μπορούσε να είναι πιο σχετικό στο πλαίσιο εφαρμογής της δικανικής γλωσσολογίας». Το ιδιολεκτικό ύφος θα είχε να κάνει όχι τόσο με το σύστημα της γλώσσας/διαλέκτου ενός ατόμου αλλά

α) με το πώς αυτό το σύστημα που είναι κοινό μεταξύ πολλών ομιλητών χρησιμοποιείται με ξεχωριστό τρόπο από ένα άτομο, και

β) με το πώς το σύστημα παραγωγής ενός ομιλητή/συγγραφέα εμφανίζεται να

[3]Η έννοια του ύφους είναι σαφώς ευρύτερη από αυτή της ιδιολέκτου και σχετίζεται με ποικίλες παραμέτρους (βλ. ενδεικτικά Eckert παραπάνω, Coupland (2007), (Wales, 2014). Στην παρούσα διατριβή, όπου αναφέρεται η έννοια του ύφους, εννοείται το ιδιολεκτικό ύφος.

είναι ατομικό και μοναδικό.

Υπό αυτό το πρίσμα, η Turell μετακινείται από την αφηρημένη έννοια της ιδιολέκτου ως συνόλου όλων των δυνητικών γλωσσολογικών πραγματώσεων ενός ατόμου προς την ιδέα της γλωσσικής πραγμάτωσης ως μορφής εκδήλωσης του τρόπου ομιλίας ή γραφής.

Ο Grant (2012, σ. 473) προχωρά ένα βήμα παραπέρα, διατυπώνοντας την άποψη πως η αντίληψη ότι η αυτόματη αναγνώριση συγγραφέα στηρίζεται σε στιβαρό θεωρητικό υπόβαθρο ύπαρξης ιδιόλεκτου είναι εσφαλμένη, καθώς επίσης ότι η εμπειρική ανίχνευση γλωσσολογικής συνέπειας και διακριτότητας αποτελούν ικανά στοιχεία να θεμελιώσουν τη βάση για πρακτικές εφαρμογές απόδοσης συγγραφέα. Κατ' αντιστοιχία με την Turell και σε μια καθαρά εμπειρική προσέγγιση στο θέμα της αναγνώρισης συγγραφέα, ο Grant επικεντρώνεται στο ζητούμενο γλωσσολογικό αποτέλεσμα, όπως αυτό εκφράζεται σε γραπτά κείμεαο, και απομακρύνεται από την εξάρτηση μιας εξιδανικευμένης αφηρημένης έννοιας της ιδιολέκτου. Ο Grant θέτει ξεκάθαρα τα όρια της έννοιας της διακριτότητας που απαιτείται για αξιόπιστη αυτόματη αναγνώριση συγγραφέα και παίρνει αποστάσεις από την ανάγκη της αφηρημένης θεωρητικής έννοιας της ιδιολέκτου ως θεωρητικής βάσης πιστοποίησης των αποτελεσμάτων. Η διακριτότητα, όπως ορίζεται κατά τον Grant, αναφέρεται στην πρακτική εφαρμογή της απόδοσης συγγραφέα σε συγκρίσεις μικρών υποσυνόλων. Σε τέτοιες εφαρμογές δεν είναι απαραίτητο να δειχθεί η διαφορετικότητα ενός συγγραφέα σε σχέση με όλους τους υπόλοιπους αλλά μόνο συγκριτικά με ένα μικρό υποσύνολο άλλων που είναι σχετικοί με την υπό εξέταση υπόθεση Grant (2012, σ. 474).

Κατά συνέπεια, αν το κειμενικό ύφος ενός ατόμου μπορεί να ειπωθεί ότι διαφέρει σε σύγκριση με ένα σύνολο αναφοράς που αποτελείται από κείμενα άλλων ατόμων/δημιουεγών κειμένων, αυτό θεωρείται ως διαφοροποίηση σε επίπεδο πληθυσμού (population level distinctiveness, Grant, 2005, σ. 515). Στην περίπτωση που το κειμενικό ύφος ενός ατόμου ξεχωρίζει ως μοναδικό μεταξύ ενός μεγάλου πληθυσμού συγγραφέων, τότε γίνεται αντιληπτή η ιδιολεκτική φύση των ιδιαίτερων γλωσσολογικών χαρακτηριστικών και προτιμήσεων αυτού του συγγραφέα.

Ο Grant προτείνει τη σύγκριση ενός ή περισσοτέρων γλωσσολογικών χαρακτηριστικών (δεικτών) ενός ατόμου υπό την υπόθεση ότι είναι καταγεγραμμένη και διαπιστωμένη η συχνότητα εμφάνισης και χρήσης αυτών των χαρακτηριστικών από έναν ευρύτερο αλλά παρόμοιων χαρακτηριστικών πληθυσμό παραγωγής κειμένων. Η πρόταση αυτή είναι όμοια με εκείνη της Turell (2010, σ. 217), η οποία εισάγει την έννοια της «Γνώσης βασικού επιπέδου» (Base Rate Knowledge), όπου η έννοια του βασικού επιπέδου εδώ είναι ανάλογη της εκ των προτέρων πιθανότητας (prior) στη στατιστική και τη Θεωρία πιθανοτήτων. Η Turell όμως ισχυρίζεται ότι η πραγμάτωση της γνώσης βασικού επιπέδου, όσον αφορά τη σπανιότητα ή την πιθανότητα εμφάνισης ενός συγκεκριμένου

γλωσσολογικού χαρακτηριστικού ή δομικού σχήματος, απαιτεί σώματα κειμένων που να απαρτίζονται από όλες τις πιθανές ιδιολεκτικές εκφράσεις όλων των συγγραφέων στο σώμα κειμένων (Turell, 2010, σ. 240).

Πιο πρόσφατα, η αυστηρότητα ενός συστήματος γνώσης βασικού επιπέδου, που περιλαμβάνει όλες τις πραγματώσεις χαρακτηριστικών σε επίπεδο πληθυσμού, έχει χαλαρώσει και η εστίαση μετακινείται στη δημιουργία συγκεκριμένων ή πιο σχετικών μοντέλων. Για παράδειγμα, οι Nini & Grant (2013, σσ. 8-9) χρησιμοποιούν το πολυδιάστατο πλαίσιο του Biber (1991) ως γνώση βασικού επιπέδου για τη συχνότητα συγκεκριμένων γλωσσικών χαρακτηριστικών μεταξύ διαφόρων κειμενικών ειδών. Συγκεκριμένα, συγκρίνοντας τη χρήση ορισμένων χαρακτηριστικών σε σχέση με τη γνώση βασικού επιπέδου, που προκύπτει από τον Biber, ποσοτικοποιούν τον βαθμό στον οποίο είναι διακριτές οι επιλογές του εκάστοτε συγγραφέα συγκριτικά με τον πληθυσμό. Η ίδια η Turell επαναδιατυπώνει τον ορισμό της γνώσης βασικού επιπέδου ως εξής:

«.. η γνώση βασικού επιπέδου υπονοεί τη συλλογή δεδομένων όσον αφορά την γενική χρήση των υπό εξέταση γλωσσικών παραμέτρων από έναν σχετικό πληθυσμό ή ομάδα χρηστών μιας γλώσσας σε μια γλωσσική κοινότητα και κατά συνέπεια η σύγκριση της συμπεριφοράς των ομιλητών ή συγγραφέων μπορεί να είναι εφικτή» (Turell & Gavalda, 2012, σ. 499).

Η έμφαση σε αυτό το σημείο δίνεται στη σημασία της ύπαρξης της γνώσης βασικού επιπέδου όσον αφορά ένα γλωσσικό χαρακτηριστικό ή δομικό σχήμα, το οποίο έχει ληφθεί από έναν «σχετικό πληθυσμό» και από συγγραφείς μέλη της ίδιας γλωσσικής κοινότητας, έτσι ώστε να είναι εφικτή η σύγκριση.

Η εστίαση (specificity) αυτή οδήγησε στην εγκατάλειψη της πρακτικά ανέφικτης προϋπόθεσης να συλλεχθούν δεδομένα ως σημείο αναφοράς από όλους τους συγγραφείς και ομιλητές για μια συγκεκριμένη γλώσσα. Αντίστοιχες προτάσεις έχουν διατυπωθεί από τους Coulthhard et al. (Coulthard et al., 2011, σ. 512), οι οποίοι διατυπώνουν την άποψη ότι η κατασκευή εξειδικευμένων σωμάτων κειμένων για την κάλυψη των στατιστικών χαρακτηριστικών του πληθυσμού είναι μονόδρομος για τη δικανική γλωσσολογία.

Επιπλέον, ο Kredens διατυπώνει την άποψη ότι μια ευνοϊκή εξέλιξη θα ήταν η ανάπτυξη σωμάτων κειμένων αναφοράς, τα οποία θα χρησιμοποιούνται ως κανονιστικές/ρυθμιστικές πηγές δεδομένων. Επιπλέον, θέτει και τον περιορισμό ότι τα σχετικά σώματα κειμένων πρέπει να χαρακτηρίζονται από βιολογικές, κοινωνικές και αλληλεπιδραστικές μεταβλητές, ταυτόσημες με εκείνες των κειμένων υπό αμφισβήτηση, τα οποία εξετάζονται συγκριτικά με τα σώματα κειμένων. Ομοίως, ο Grant (2012, σ. 473) τονίζει τη σημασία του να είναι τα δεδομένα του πληθυσμού σχετικά με το αντικείμενο έρευνας.

Παρομοίως, από τη σκοπιά της Μηχανικής μάθησης, ο βαθμός συνέπειας της συλλογής σε σώματα κειμένων με στόχο την ιδιολεκτική σύγκριση αποτελεί σημαντικό παράγοντα στην αξιοπιστία των σχετικών αποτελεσμάτων που θα

προκύψουν από τη διαδικασία της αναγνώρισης. Η δημιουργία ενός συνεπούς γλωσσικά συγκριτικού σώματος κειμένων, το οποίο να λαμβάνει υπόψη τόσο το κειμενικό υπο-είδος όσο και άλλες πηγές γλωσσικής ποικιλότητας, αποτελεί κομβικό συστατικό (όπως σε όλες της εφαρμογές μηχανικής μάθησης) για τη μεγιστοποίηση της ακρίβειας και της ικανότητας του αλγορίθμου να γενικεύσει με μεγάλη επιτυχία σε νέα, άγνωστα δεδομένα.

Κατά συνέπεια, με την εισαγωγή και την ανάπτυξη της γνώσης βασικού επιπέδου και τη διαχωρισιμότητα σε επίπεδο πληθυσμού, η ατομική γλωσσική ποικιλότητα παύει να είναι ιδεατή έννοια, όπως αρχικά περιγράφηκε από τους Paul, Sapir, Bloomfield και Bloch, και μετατρέπεται σε εμπειρικό φαινόμενο, το οποίο είναι εφικτό να μετρηθεί και να αναλυθεί. Με την προϋπόθεση ότι υπάρχουν δεδομένα που αντανακλούν τις νόρμες της κοινότητας από την οποία προέρχεται ένα σύνολο συγγραφέων, μπορούν να εξαχθούν στοιχειοθετημένα συμπεράσματα για τη φύση και τη μοναδικότητα των γλωσσικών τους προτιμήσεων και με αυτή την έννοια το πρόβλημα της ιδιολέκτου από ζήτημα γλωσσικού συστήματος (langue) μετατρέπεται σε υπόθεση γλωσσικής πραγμάτωσης (parole).

Ο ορισμός της ιδιολέκτου συμβαδίζει εμπειρικά με τη διαφοροποίηση του ύφους των χρηστών σε διάφορα κείμενα όπως στα κείμενα που δημιουργούνται στα μέσα κοινωνικής δικτύωσης αλλά και στα «παραδοσιακά» μέσα μαζικής ενημέρωσης, όπως αυτά εκφράζονται στους διαδικτυακούς τους ιστότοπους με άρθρα γνώμης, κύρια άρθρα κ.λπ. Στη συγκεκριμένη διατριβή επικεντρωνόμαστε σε σώματα κειμένων, που συλλέχτηκε από κοινωνικά δίκτυα, και πιο συγκεκριμένα από το twitter και ιστολόγια. Το σώμα κειμένων του Twitter παρέχει την απαραίτητη κειμενική ποικιλία ενσωματώνοντας έναν μεγάλο αριθμό συγγραφέων (4498) και καλύπτοντας μια χρονική περίοδο 8 ετών (2008 έως το 2016). Η ιδιαιτερότητα του μέσου, που παρέχει πρόσβαση για άμεσο σχολιασμό της επικαιρότητας στους χρήστες, συνομιλία και ανταλλαγή απόψεων, καθώς επίσης και οι περιορισμοί που επιβάλλει (όριο 140 χαρακτήρων ανά δημοσίευση), προσδίδουν ιδιαίτερη σημασία τόσο στις άμεσες και έμμεσες νόρμες, που αναπτύσσονται κατά τη χρήση του, όσο και στις σημαντικές διαφοροποιήσεις που παρατηρούνται στην κειμενική συμπεριφορά των χρηστών από την αναμενόμενη, όπως θα φανεί στη συνέχεια.

2.4 Η ιδιόλεκτος στη γλωσσολογία σωμάτων κειμένων

Το ζήτημα της ποσοτικοποίησης των χαρακτηριστικών της ιδιολέκτου έχει τεθεί από τις -λίγες σχετικά- έρευνες για την ιδιόλεκτο στη γλωσσολογία σωμάτων κειμένων. Συγκεκριμένα, με την ιδιόλεκτο έχουν ασχοληθεί ο ο Barlow (2013), η

Mollin (2009) και στα ελληνικά η Φραγκάκη (2012), που ακολουθεί το πρότυπο ανάλυσης του Barlow (2013).

Στην ανάλυσή του ο Barlow (2013) επικεντρώνεται στην ανάλυση των χαρακτηριστικών των ομιλιών πέντε Γραμματέων Τύπου (κυβερνητικών εκπροσώπων) του Λευκού Οίκου από το 1994 έως το 1998 και το 2001 έως το 2008. Μετά την επεξεργασία των τεσσάρων Γραμματέων, ακολουθεί ποσοτική και ποιοτική ανάλυση των συχνοτήτων των διγραμμάτων λέξεων. Η μεθοδολογία που ακολουθείται είναι απλή. Ορίζεται ως ελάχιστη συχνότητα εμφάνισης ενός διγράμματος οι 7 εμφανίσεις και στη συνέχεια τα διγράμματα ταξινομούνται σε φθίνουσα σειρά με βάση τη συχνότητα. Όπως τονίζεται, τα διγράμματα δεν αντανακλούν γραμματικές ιδιότητες. Παρ' όλα αυτά, μπορούν να θεωρηθούν δείκτες γραμματικών προτύπων που μπορούν να χρησιμοποιηθούν ως «αποτυπώματα» του ξεχωριστού κειμενικού ύφους κάθε συγγραφέα/ομιλητή.

Στην πρώτη ανάλυση, από την ταξινομημένη κατά συχνότητα λίστα των διγραμμάτων επιλέγονται τα 10 συχνότερα δείγματα ανά συγγραφέα και στη συνέχεια συγχωνεύονται σε μια λίστα 15 πιο συχνών διγραμμάτων ανά δείγμα. Χρησιμοποιώντας αυτές τις συχνότητες, ο Barlow δημιουργεί ιστογράμματα ανά συγγραφέα και προχωρά σε ποιοτική ανάλυση, που είναι επικεντρωμένη στην οπτική σύγκριση των ιστογραμμάτων. Παρατηρεί δηλαδή απλά τις διαφορές των ιστογραμμάτων μεταξύ των ομιλητών και σημειώνει ότι αυτό αποτελεί καλή ένδειξη της ύπαρξης προσωπικού ύφους. Επίσης, συγκρίνει διαφορετικά δείγματα από τους ίδιους συγγραφείς, με την ίδια πάντα μεθοδολογία, για να φτάσει στο σχετικά ασφαλές συμπέρασμα (με βάση τη μορφή των υπό μελέτη ιστογραμμάτων) ότι υπάρχει ομοιομορφία και σχετική σταθερότητα στη συχνότητα των πιο συχνών διγραμμάτων ανά ομιλητή.

Σε δεύτερο χρόνο, επαναλαμβάνει το ίδιο πείραμα αυτή τη φορά με τα 46 πιο συχνά διγράμματα. Και σε αυτή την περίπτωση φαίνεται να ισχύουν οι παρατηρήσεις του πρώτου πειράματος. Στη συνέχεια προχωρά σε κοινή ανάλυση συχνοτήτων διγραμμάτων και τριγραμμάτων, συχνοτήτων ν-γραμμάτων μερών του λόγου (POS tag ngrams), καθώς επίσης και σε γραμματικούς ελέγχους σε φαινόμενα όπως η άρνηση, η χρήση των τροπικών ρημάτων *will* και *going to* και η χρήση παθητικής φωνής. Πιο συγκεκριμένα, ο Barlow προχωρά σε ανάλυση πολυδιάστατων συσχετισμών μεταξύ δειγμάτων χρησιμοποιώντας Παραγοντική Ανάλυση Αντιστοιχιών (Correspondence Analysis), βασισμένη στις αποστάσεις του στατιστικού ελέγχου χ^2 μεταξύ δειγμάτων και διγραμμάτων. Καθένα από τα δείγματα ελέγχεται ως προς την ομοιότητά του με τα άλλα δείγματα χρησιμοποιώντας τη συχνότητα των διγραμμάτων. Στη συνέχεια, ο πολυδιάστατος χώρος, που προκύπτει από την παραπάνω διαδικασία, προβάλλεται σε δισδιάστατη απεικόνιση κρατώντας μόνο τους 2 δεσπόζοντες ως προς τη διασπορά παράγοντες. Με αυτό τον τρόπο επιτυγχάνεται μια μορφή ομαδοποίησης (clustering) των διγραμμάτων έτσι ώστε τα διγράμματα που

προέρχονται από τον ίδιο ομιλητή να αποτελούν κοινή ομάδα στη δισδιάστατη αυτή απεικόνιση. Έπειτα, γίνεται ποιοτική ερμηνεία των αποτελεσμάτων χρησιμοποιώντας επιλεγμένα διγράμματα, που εμφανίζονται στην παραπάνω απεικόνιση, καθώς και ανάλυση χρήσης και συχνότητας ιδιοσυγκρασιακών φράσεων μεταξύ ομιλητών. Το συμπέρασμα είναι ότι η διαφορά στην ιδιόλεκτο έγκειται σε διαφοροποιήσεις στη χρήση γραμματικών δομών και σε διαφοροποιήσεις στη συχνότητα χρήσης συνηθισμένων εκφράσεων, όπως *in terms of* και *as a result*. Η Παραγοντική Ανάλυση Αντιστοιχιών σε δεδομένα από διγράμματα μερών του λόγου (part of speech bigrams) επιτυγχάνει επίσης ξεκάθαρη διαφοροποίηση.

Επισημαίνεται ότι, δεδομένου του περιορισμένου όγκου δεδομένων και αριθμού συγγραφέων τα συμπεράσματα που μπορούν να εξαχθούν αφορούν μόνο τους υπό μελέτη ομιλητές. Παρ' όλα αυτά οι διαφοροποιήσεις μεταξύ τους είναι ισχυρές. Στα ενδιαφέροντα σημεία της μελέτης ανήκει και η παρατήρηση ότι στο συγκεκριμένο σώμα φαίνεται πως οι επιμέρους επιλογές των ομιλητών δείχνουν σταθερότητα στο πέρασμα του χρόνου.

Η ανάλυση του Barlow, αν και ενδιαφέρουσα καθώς καλύπτει μεγάλο φάσμα ποιοτικών και γραμματικών φαινομένων, δεν παρέχει κάποιο εργαλείο ποσοτικής σύγκρισης της ομοιότητας (ή αντίστοιχα, της απόστασης) του κειμενικού ύφους μεταξύ των εξεταζόμενων συγγραφέων. Δεν επιχειρεί δηλαδή να απαντήσει σε ερωτήματα της μορφής «ποιος ομιλητής έχει παρόμοιο κειμενικό ύφος με τον Χ» ή, αντίστροφα, «ποιος ομιλητής έχει τη μεγαλύτερη υφολογική διαφοροποίηση από τον Χ;» Όπως θα δούμε και στη συνέχεια, αυτές οι ερωτήσεις μπορούν να απαντηθούν ακόμα και σε απλές αναλύσεις, όπως η παραπάνω, με σχετικά απλές γεωμετρικές μεθόδους ή τεχνικές από τη Θεωρία Πληροφοριών.

Η Mollin (2009) εξετάζει διαφοροποιήσεις στη συχνότητα χρήσης επιτατικών (maximizers) στα κείμενα του Tony Blair σε σύγκριση με το σώμα κειμένων British National Corpus (BNC corpus) για να διατυπώσει την προσέγγισή της στην ανίχνευση της ιδιολέκτου. Πιο συγκεκριμένα, η Mollin εξετάζει τη διαφοροποίηση λεξικών συνάψεων του τύπου *entirely understand* σε σύγκριση με το BNC. Η κεντρική ιδέα είναι ότι διαφοροποιήσεις και προτίμηση σε συγκεκριμένες λεξικές συνάψεις σε σχέση με άλλες παρόμοιου σημασιολογικού περιεχομένου (π.χ. *fully understand*, *completely understand*) προσδίδουν ιδιολεκτική χροιά στον ομιλητή/συγγραφέα.

Η Mollin αιτιολογεί την επιλογή του πρωθυπουργού Blair για τη μελέτη αυτή αναφέροντας την εγγενή δυσκολία συλλογής μεγάλου όγκου κειμένων ενός ατόμου. [4] Η διαδρομή του Tony Blair στην πολιτική σκηνή της Μεγάλης Βρετανίας από το 1994 έως το 2007 και κατά συνέπεια η καταγραφή των δημοσίων λόγων του προσέφεραν ακριβώς τον όγκο των κειμενικών δειγμάτων για τη

[4]Βρισκόμαστε στο 2009 και ακόμα τα social media δεν είναι τόσο διαδεδομένα.

μελέτη αυτή. Βεβαίως, όπως σημειώνει η Mollin, η επιλογή της κατηγορίας του πολιτικού λόγου δεν ήταν ίσως η ιδανική για σύγκριση με το BNC και αυτό παίζει όντως ρόλο στην εξαγωγή των συμπερασμάτων της. Συνολικά, το σώμα κειμένων που δημιουργήθηκε, αποτελείται από 3.119.931 λέξεις με κύριες πηγές το αρχείο δηλώσεων, λόγων και συνεντεύξεων από την ιστοσελίδα του γραφείου του πρωθυπουργού της Βρετανίας[5] και τα πρακτικά του Κοινοβουλίου[6], ενώ ένα πολύ μικρό ποσοστό της τάξης του 1% του συνολικού σώματος κειμένου καταλαμβάνεται από συνεντεύξεις του Tony Blair σε εφημερίδες.

Η Mollin εντοπίζει δύο σημαντικές παραμέτρους, που θεωρητικά θα μπορούσαν να έχουν αλλοιώσει τα αποτελέσματα αυτής της έρευνας: α) Οι ομιλίες των κοινοβουλευτικών ελέγχονται και γραμματικά λάθη διορθώνονται από τους διορθωτές των πρακτικών (Hansard) αλλά αυτό είναι μάλλον απίθανο στην περίπτωση των επιτακτικών και μετριαστικών (maximizers & minimizers) β) Το χρονικό διάστημα που καλύπτει το σώμα κειμένων είναι μεταξύ 1998-2007 με τον μεγαλύτερο όγκο στο διάστημα 2003-2007, στο οποίο ο Blair ήταν ιδιαίτερα ενεργός στη βρετανική πολιτική σκηνή. Αυτό είναι προβληματικό όσον αφορά την σύγκριση με το BNC διότι τα κείμενα του BNC την στιγμή που γινόταν η ανάλυση έφταναν έως το 1994 και ενδεχομένως οι συχνότητες χρήσης κάποιων από τις προς μελέτη λεξικές συνάψεις να είχαν αλλάξει τη στιγμή της σύγκρισης. Αυτό φυσικά αντικατοπτρίζει και το γενικότερο πρόβλημα της δυσκολίας συλλογής δεδομένων, τουλάχιστον ως τα μέσα της προηγούμενης δεκαετίας (2000-2010). Η παραγωγή και η πρόσβαση σε τεράστιους όγκους κειμενικών δεδομένων άλλαξε άρδην με την επικράτηση των κοινωνικών δικτύων.

Η Mollin στηρίζεται στους Quirk (1985), οι οποίοι τοποθετούν τα επιτατικά στην κατηγορία των εμφατικών τύπων και διακρίνουν ανάμεσα σε ενισχυτικά (amplifiers) και υποβαθμιστικά (downtoners). Παραδείγματα στην κατηγορία των ενισχυτικών από την αγγλική γλώσσα είναι τα επιρρήματα *absolutely*, *completely*, *entirely* και *totally*. Οι Quirk et al. παρατηρούν ότι όλα τα ενισχυτικά εμφανίζονται να έχουν περιορισμούς στις συνεμφανίσεις τους, υπό την έννοια ότι συγκεκριμένα ρήματα τείνουν να συνεμφανίζονται με συγκεκριμένα επιτατικά, ακόμα και αν αυτά ανήκουν στην ίδια σημασιολογική ομάδα (Quirk, 1985, σ. 593). Στο συγκεκριμένο παράδειγμα, ενώ δεν υπάρχει ιδιαίτερη σημασιολογική διαφοροποίηση μεταξύ των *I entirely agree* και *I completely agree*, ο δεύτερος τύπος εμφανίζεται με συχνότητα 11 φορές μεγαλύτερη στο BNC. Η ανάλυση της Mollin επικεντρώνεται στους τύπους *absolutely*, *altogether*, *completely*, *entirely*, *extremely*, *fully*, *perfectly*, *thoroughly*, *totally*, *utterly*, *wholly*.

Το πρώτο εύρημα είναι ότι η συχνότητα εμφάνισης των επιτατικών στο σώμα κειμένων του Blair είναι σαφώς μεγαλύτερη από την αντίστοιχη συχνότητα του

[5]http://www.pm.gov.uk

[6]http://publications.parliament.uk

BNC. Αυτό όμως δεν αποτελεί ιδιαίτερο χαρακτηριστικό του ύφους του Blair, αλλά τείνει να είναι κοινή παρατήρηση για τη γλώσσα που χρησιμοποιούν εν γένει οι πολιτικοί (Grattan, 1998). Η παρατήρηση αυτή μπορεί να εξηγηθεί από το εμφατικό και πειστικό ύφος που χρησιμοποιούν οι πολιτικοί σε αντίθεση με τη συνηθισμένη επικοινωνία. Επομένως, το ερώτημα που τίθεται δεν αφορά τη συχνότητα των επιτατικών αλλά τη συχνότητα με την οποία συνδυάζονται με επίθετα, επιρρήματα και ρήματα και αν αυτές οι συχνότητες είναι ικανές να θεωρηθούν ως δείκτες ύφους ενός ομιλητή.

Η Mollin με τη χρήση στατιστικών μεθόδων εντοπίζει αυτό που ονομάζει «ιδιολεκτικές συνάψεις επιτατικών» (idiolectal maximiser collocations). Για τον εντοπισμό αυτών των λεξικών συνάψεων χρησιμοποιεί τρεις διαφορετικές στατιστικές μεθόδους και πιο συγκεκριμένα την κανονικοποιημένη συχνότητα (normalized frequency) [7], την αμοιβαία πληροφορία (mutual information) [8] και τη λογαριθμική πιθανοφάνεια (log-likelihood) [9]. Η παραδοχή που γίνεται είναι ότι οι λεξικές συνάψεις είναι δύσκολο να οριστούν στο πλαίσιο της παραγωγικότητας μιας φυσικής γλώσσας, χαρακτηρίζονται όμως από την τάση των μελών τους να συνεμφανίζονται με μεγαλύτερη συχνότητα από την τυχαία σύμπτωση.

Στη συνέχεια, η μεθοδολογία που ακολουθεί είναι η εξής: όλες οι πραγματώσεις επιτατικών που ακολουθούνται από επίθετα, επιρρήματα ή ρήματα συγκεντρώνονται και για τα δύο σώματα κειμένων, το BNC και το σώμα κειμένων του Blair, συνοδεύονται από μετρήσεις συχνότητας και ισχύος συνάφειας (collocation strength). Προφανώς, λόγω μεγέθους του BNC υπάρχουν ζεύγη στο BNC που δεν εμφανίζονται στο σώμα κειμένων του Blair. Για την ακρίβεια, οι εμφανίσεις του BNC είναι 32 φορές περισσότερες.

Στον πίνακα (2.2) παρουσιάζονται τα αποτελέσματα της ανάλυσης σε τέσσερις κατηγορίες: γενικές συνάψεις, συνάψεις που απαντούν σε εφημερίδες, συνάψεις που απαντούν κυρίως στον κοινοβουλευτικό λόγο και, τέλος, συνάψεις που χαρακτηρίζουν ιδιολεκτικά τον Tony Blair.

Η ανάλυση της Mollin δίνει έμφαση στη διαφορά συχνότητας χρήσης των επιτατικών ως μέτρο διακριτότητας του ιδιολεκτικού ύφους του Tony Blair. Η ανάλυσή της όμως στηρίζεται σε μικρό αριθμό λέξεων και αφορά τη σύγκριση ενός μόνο ατόμου σε σχέση με το BNC. Δεν προχωρά δηλαδή σε σύγκριση ατόμων αλλά αφορά τη σύγκριση ενός ατόμου με ένα σώμα κειμένων. Συνεπώς, όπως αντίστοιχα και ο Barlow, δεν επιχειρεί δηλαδή να απαντήσει σε ερωτήματα ομοιότητας ή διαφοροποίησης ομιλητών μεταξύ τους.

Η μόνη διαθέσιμη έρευνα στα ελληνικά για τη μελέτη της ιδιολέκτου είναι

[7] Αναγωγή της συχνότητας εμφάνισης μιας λέξης σε σώμα κειμένων 1.000.000 λέξεων

[8] $I(X;Y) = \sum_{y \in Y} \sum_{x \in Y} p(x,y) log(\frac{p(x,y)}{p(x)p(y)})$

[9] Αφορά τη μέτρηση της σημαντικότητας σε σώματα κειμένων Lijffijt et al., 2016

Πίνακας 2.2: Αποτελέσματα ανάλυσης επιτατικών

1. Συνάψεις που δεν είναι χαρακτηριστικές του ιδιολεκτικού ύφους του Tony Blair αλλά αφορούν το σύνολο της αγγλικής γλώσσας	*totally agree*, *perfectly entitled*, *perfectly obvious*, *entirely sensible*,*fully supportive*
2. Συνάψεις που δεν είναι χαρακτηριστικές του Tony Blair αλλά συναντώνται σε περιβάλλον άρθρων εφημερίδων	*perfectly simple*, *perfectly fair*
3. Συνάψεις που δεν είναι χαρακτηριστικές του ύφους Tony Blair αλλά συναντώνται σε κοινοβουλευτικό περιβάλλον	*entirely endorse*, *entirely share*
4. Συνάψεις που είναι χαρακτηριστικές του ιδιολεκτικού ύφους του Tony Blair	*utterly absurd*, *entirely accept*, *absolutely blunt*, *absolutely committed*, *completely committed*,*fully consistent*,*thoroughly decent*, *absolutely frank*,*wholly innocent*, *entirely justified*, *wholly new*,*perfectly prepared*, *completely unacceptable*, *entirely understand*,*totally understand*, *completely wrong*

της Φραγκάκη (2012). Η Φραγκάκη προσαρμόζει την ανάλυση του Barlow[10] σε ελληνικά δεδομένα και πιο συγκεκριμένα στην κειμενική ανάλυση ύφους των κυβερνητικών εκπροσώπων των ετών 2008, 2009 και 2010 (Θεόδωρος Ρουσόπουλος, Ευάγγελος Αντώναρος και Γιώργος Πεταλωτής). Δημιουργεί τρία σώματα κειμένων, ένα για κάθε κυβερνητικό εκπρόσωπο, από τα οποία στη συνέχεια αφαιρούνται οι ερωτήσεις των πολιτικών συντακτών ώστε να μελετηθεί η προσωπική συμβολή των εκπροσώπων, που κινείται γύρω στις 22.000 λέξεις για κάθε ομιλητή.

Στην έρευνα εντοπίζονται οι λέξεις κλειδιά με βάση τη συχνότητα εμφάνισής τους σε ζεύγη σωμάτων. Παρατηρεί πως το ζεύγος Ρουσόπουλου-Αντώναρου δεν παράγει μεγάλο αριθμό λέξεων κλειδιών, κάτι που αποδίδεται στο γεγονός ότι ο λόγος των δύο αυτών εκπροσώπων που προέρχονται από το κόμμα της Νέας Δημοκρατίας παρουσιάζει μεγάλες ομοιότητες, σε αντίθεση με εκείνον του Πεταλωτή που αποκλίνει συγκριτικά με τους άλλους δύο. Στη συνέχεια γίνεται ποιοτική και ποσοτική γλωσσολογική ανάλυση μακροδομικών και οργανωτικών στοιχείων, του γραμματικού προσώπου και της τροπικότητας, όπως παρουσιάζεται στη συνέχεια. Η μακροδομική ανάλυση επικεντρώνεται στις

[10]Το σχετικό άρθρο βασίζεται σε παρουσίαση του Barlow πριν τη δημοσίευσή της

αποκλίσεις των εκφράσεων των εκπροσώπων κατά τη διάρκεια των ενημερώσεων συντακτών, οι οποίες εμφανίζουν την εξής δομή:

- άνοιγμα (χαιρετισμός και ευχές)
- κείμενο ενημέρωσης από τον κυβερνητικό εκπρόσωπο (προσχεδιασμένος λόγος)
- μεταβατική φράση
- κλείσιμο (ευχαριστίες ή/και ευχές ή χαιρετισμός)

Όπως παρατηρεί η Φραγκάκη, στο άνοιγμα και στο κλείσιμο της ενημέρωσης συντακτών εμφανίζεται αρκετή ποικιλία, που εκφράζεται στους διαφορετικούς τρόπους χαιρετισμού, που χρησιμοποιούν οι εκπρόσωποι. Στο κλείσιμο επίσης υπάρχει ποικιλία καθώς οι Ρουσόπουλος και Αντώναρος συνηθίζουν να κλείνουν με ευχαριστίες (*Σας ευχαριστώ*), οι οποίες ορισμένες φορές συνδυάζονται με ευχές, ενώ ο Πεταλωτής άλλοτε κλείνει με χαιρετισμό (*Καλό απόγευμα*) και άλλοτε με ευχαριστίες (*Σας ευχαριστώ*), ενώ μία φορά κλείνει μόνο με ευχή (*Καλό Σαββατοκύριακο*). Η μεταβατική φράση, αντίθετα, είναι κατά τη συντριπτική πλειονότητα ίδια.

Ως προς τα οργανωτικά στοιχεία του λόγου των κυβερνητικών εκπροσώπων ενδιαφέρον παρουσιάζουν οι δείκτες λόγου και οι αναφορικές εκφράσεις και πώς διαφέρουν στη χρήση τους μεταξύ των εκπροσώπων. Τύποι ή λεξικά συμπλέγματα με διαφορετική συχνότητα χρήσης είναι τα *επίσης* και *άλλωστε*, *άρα, λοιπόν* ή το λεξικό σύμπλεγμα *σε κάθε περίπτωση*. Επίσης, ο Πεταλωτής χρησιμοποιεί συστηματικά τα λεξικά συμπλέγματα *Από εκεί και πέρα*, *Για/γι' αυτό ακριβώς*, *Πρώτα απ'/από όλα*, τα οποία δεν εμφανίζονται στα κείμενα των άλλων. Συνολικά, ο Πεταλωτής εμφανίζεται να αποκλίνει περισσότερο συγκρινόμενος με τους άλλους δύο κυβερνητικούς εκπροσώπους, οι οποίοι δεν εμφανίζουν τόσο έντονες διαφορές.

Επιπλέον, οι κυβερνητικοί εκπρόσωποι χρησιμοποιούν δύο βασικές αναφορικές εκφράσεις (*Σε ό,τι αφορά* + άρθρο ή *στο/στη* και *Όσον αφορά* + άρθρο ή *στο/στη*), με διαφορετική όμως συχνότητα. Διαφοροποιήσεις παρατηρούνται επίσης στις αναφορικές φράσεις/δομές της μορφής *Σε ό,τι αφορά* + άρθρο, την οποία φαίνεται να προτιμά ο Ρουσόπουλος, ενώ οι άλλοι δύο τη χρησιμοποιούν μόνο μία φορά ο καθένας. Από την άλλη, οι Αντώναρος και Πεταλωτής χρησιμοποιούν συχνότερα τη δομή *Όσον αφορά* + άρθρο ή *στο/στη*. Οι αναφορικές φράσεις είναι ιδιαίτερα σημαντικές στο λόγο των κυβερνητικών εκπροσώπων διότι αποτελούν βασικά στοιχεία δομής των απαντήσεών τους στους πολιτικούς συντάκτες. Ο ρόλος αυτός φαίνεται από λεξικά συμπλέγματα που συνδυαζόμενα δημιουργούν φράσεις όπως: *Όσον αφορά (σ)το πρώτο/δεύτερο*

σκέλος (της ερώτησής σας). Επομένως, ως προς τις αναφορικές φράσεις και οι τρεις εκπρόσωποι έχουν σχετικά ξεκάθαρο στίγμα, εκείνος όμως που εμφανίζεται να αποκλίνει, αντίθετα με ό,τι παρατηρήθηκε σε σχέση με τους δείκτες λόγου, είναι ο Ρουσόπουλος.

Μια δεύτερη διαφοροποίηση στο ύφος των τριών εκπροσώπων αφορά τη χρήση του προσώπου και πιο συγκεκριμένα τους διαφορετικούς τρόπους χρήσης πρώτου του πληθυντικού (*εμείς*) με αναφορά στην κυβέρνηση. Η χρήση της αντωνυμίας αυτής είναι πολύ μικρότερη από τους εκπροσώπους της Νέας Δημοκρατίας σε σχέση με τον Πεταλωτή. Το ίδιο συμβαίνει με τον τύπο *μας*, που χρησιμοποιείται από τον Πεταλωτή και πάλι με αναφορά στην κυβέρνηση και το κυβερνητικό έργο ως επί το πλείστον, ενώ αντίθετα οι Ρουσόπουλος και Αντώναρος το χρησιμοποιούν για να αναφερθούν σε όλους τους Έλληνες. Αντίστοιχη χρήση με έμφαση στην κυβέρνηση και διαφοροποίηση από τους άλλους δύο εκπροσώπους έχουν τα ρήματα *κάνω* και *γνωρίζω*» στο α' και β' πληθυντικό.

Τέλος, η Φραγκάκη παρατηρεί σημαντική η διαφοροποίηση στη χρήση του τροπικού επιρρήματος *φυσικά* από τον Πεταλωτή, το οποίο εμφανίζεται ως λέξη-κλειδί με υψηλό βαθμό σπουδαιότητας. Ανάλογη λειτουργία έχει και το λεξικό σύμπλεγμα *είναι γνωστό ότι*.

Σε αντίθεση με τους Barlow και Mollin, η Φραγκάκη προβαίνει και σε ανάλυση γλωσσικών πράξεων που θεωρούνται σημαντικές για τη διαφοροποίηση της ιδιολέκτου των ομιλητών. Με βάση τα δεδομένα, η γλωσσική πράξη της υπενθύμισης με στόχο την απόκρουση παρουσιάζεται στο λόγο και των τριών εκπροσώπων, η γλωσσική πράξη της άρνησης απάντησης εμφανίζεται μόνο στα κείμενα των Ρουσόπουλου και Αντώναρου και οι γλωσσικές πράξεις της τοποθέτησης και της διαφωνίας εμφανίζονται μόνο στα κείμενα του Αντώναρου και του Πεταλωτή αντίστοιχα. Ενδιαφέρον εδώ έχουν οι συχνότερες δομές που χρησιμοποιούνται για τη δήλωση γλωσσικών πράξεων όπως *θα ήθελα επίσης να*, *δεν έχω να*, *δεν θα/μπορώ να/πρόκειται να πω (κάτι/τίποτε/τίποτα περισσότερο*), *Σας θυμίζω ότι, γνωρίζετε πολύ καλά ότι* κ.ά.

Η μελέτη των σωμάτων κειμένων δείχνει ότι οι κυβερνητικοί εκπρόσωποι εμφανίζουν συστηματικές προτιμήσεις στο λόγο τους, οι οποίες στις περισσότερες περιπτώσεις διαφέρουν μεταξύ τους. Ομοιότητες εμφανίζονται σε στοιχεία που σχετίζονται με το κειμενικό είδος, όπως η χρήση της στερεότυπης μεταβατικής έκφρασης *Παρακαλώ τις ερωτήσεις σας*. Επίσης, παρότι ακολουθούν με συνέπεια τις συμβάσεις του κειμενικού είδους, παρουσιάζουν διαφοροποιήσεις σε επιμέρους επιλογές. Ο λόγος των εκπροσώπων της κυβέρνησης της Νέας Δημοκρατίας παρουσιάζει συγκριτικά περισσότερες ομοιότητες, όπως φαίνεται και από τον ιδιαίτερα μικρό αριθμό λέξεων-κλειδιών, που είναι πιθανό να προέρχονται από την κοινή κυβερνητική γραμμή την οποία υπηρετούν.

Η Φραγκάκη καταλήγει ότι το σημαντικότερο εύρημα της ανάλυσης είναι

ότι οι τρεις κυβερνητικοί εκπρόσωποι εμφανίζουν ιδιολεκτικές προτιμήσεις στη χρήση λεξιλογίου και λεξικών συμπλεγμάτων. Η ανάλυση της Φραγκάκη επιβεβαιώνει με ποσοτικό και ποιοτικό τρόπο την ύπαρξη ιδιολεκτικών προτιμήσεων των κυβερνητικών εκπροσώπων, χρησιμοποιώντας στατιστική ανάλυση σε λεξικά συμπλέγματα με παρόμοιο τρόπο όπως οι Barlow και Mollin. Προχωρά επίσης στην ποσοτικοποίηση της ομοιότητας ύφους και επίσης δίνει και ερμηνείες για το παρόμοιο ύφος των εκπροσώπων που προέρχονται από το κόμμα της Νέας Δημοκρατίας.

Το σώμα κειμένων και ο αριθμός συγγραφέων που χρησιμοποιεί είναι μικρό από τη φύση του κειμενικού είδους. Η ανάλυση επίσης επικεντρώνεται σε συχνότητες και χρησιμοποιείται στατιστικός έλεγχος της συχνότητας χρήσης λέξεων και συμπλεγμάτων. Δεν ορίζεται όμως η έννοια του ιδιολεκτικού αποτυπώματος και δεν χρησιμοποιούνται πλήρεις τύποι για την ανίχνευση της ιδιολεκτικής συμπεριφοράς. Παρ'όλα αυτά, η Φραγκάκη, σε αντίθεση με τους Barlow και Mollin, παρατηρεί την ομοιότητα στο λόγο δύο ομιλητών όσον αφορά τη χρήση του προσώπου και τις γλωσσικές πράξεις που επιτελούν, το οποίο αποδίδεται στην κυβερνητική γραμμή. Αυτό υποδηλώνει σαφώς σημασιολογική επιρροή, καθώς η συνειδητή επιλογή αυτή των εκπροσώπων της Νέας Δημοκρατίας αντανακλά επικοινωνιακή στρατηγική και συνεπώς είναι σημασιολογικά φορτισμένη.

Συνολικά, οι έρευνες της ιδιολέκτου με τη χρήση σωμάτων κειμένων είναι σημαντικές γιατί επισημαίνουν λέξεις, συνάψεις και συμπλέγματα που είναι πιθανόν να διαφοροποιούν το ύφος ενός ομιλητή από εκείνο άλλων. Υπογραμμίζουν επίσης τη σημασία ποσοτικοποίησης των δεδομένων για την ανάλυση της ιδιολέκτου. Ωστόσο, έχουν αναλύσει μικρό συγκριτικά αριθμό λέξεων (από 20.000 έως 3 εκατομμύρια λέξεις ανά ομιλητή) και, κυρίως, δεν προσφέρουν έναν σαφώς προσδιορισμένο μαθηματικό μέτρο για την ομοιότητα και τη διαφοροποίηση ενός ομιλητή από έναν άλλο. Τέλος, όλες οι έρευνες επικεντρώνονται κυρίως σε γραμματικούς τύπους και φράσεις, κάτι που είναι σημαντικό γιατί οι τύποι αυτοί δεν υπόκεινται στο συνειδητό έλεγχο των ομιλητών. Από την άλλη, οι γραμματικοί τύποι (ή τύποι χωρίς πλήρες σημασιολογικό περιεχόμενο) αποτελούν ένα κλειστό σύνολο (βλ. Carter, 2012 και για τα ελληνικά Μπακάκου-Ορφανού, 2005) και επομένως η χρησιμότητά τους είναι περιορισμένη αν αυξήσουμε σημαντικά τον αριθμό των ομιλητών που εξετάζονται. Επιπλέον, η απουσία πλήρων τύπων στην ανίχνευση της ιδιολέκτου μειώνει τη συμβολή των συνειδητών επιλογών των συγγραφέων και περιορίζει τη δυνατότητα ανίχνευσης σημασιολογικών επιρροών.

2.5 Η ιδιόλεκτος ως υφολογικό αποτύπωμα

Όπως αναφέρθηκε πιο πάνω, η απόδοση συγγραφέα (authorship attribution), η διαδικασία δηλαδή κατηγοριοποίησης κειμένων με βάση το κειμενικό ύφος του συγγραφέα, στηρίζεται στην έννοια της ιδιολέκτου ως γλωσσικής πραγμάτωσης. Η απόδοση συγγραφέα αποτελεί τον πυρήνα της υφομετρικής έρευνας (stylometry) και βρίσκει εφαρμογές τόσο στην ανάλυση λογοτεχνίας (Juola, 1998; Juola et al., 2008) όσο και στην ανάλυση άλλων κειμενικών ειδών, όπως κειμένων που προέρχονται από κοινωνικά δίκτυα (Mikros & Perifanos, 2013). Το ερευνητικό αυτό πεδίο ασχολείται με προβλήματα που αφορούν την ταύτιση του συγγραφέα, τον χαρακτηρισμό του συγγραφέα και τη σύγκριση ομοιότητας. Οι διαδικασίες αυτές ανήκουν στα παραδοσιακά ενδιαφέροντα του επιστημονικού κλάδου της υφολογίας, της γλωσσολογικής μελέτης δηλαδή των λογοτεχνικών κειμένων. Η ποσοτική και στατιστική γλωσσολογική μελέτη της λογοτεχνίας αντλεί τόσο από την παράδοση του Saussure όσο και την υφολογία του Jakobson (βλ. Fialho & Zyngier, 2017).

Η απόδοση συγγραφέα περιλαμβάνει μεθόδους καταγραφής του υφολογικού αποτυπώματος ενός πομπού. Μας ενδιαφέρει όχι μόνο η στιγμιαία αποτύπωση ή ο μέσος όρος της αποτύπωσης αλλά η μελέτη του τρόπου με τον οποίο μεταβάλλονται τα χαρακτηριστικά του αποτυπώματος ως συνάρτηση του χρόνου. Στόχος είναι να βρεθεί, αν υπάρχει, το αποτύπωμα εκείνο που χαρακτηρίζει έναν συγγραφέα και ταυτόχρονα παρέχει τη δυνατότητα σύγκρισης με άλλους συγγραφείς, ώστε να ποσοτικοποιηθεί η υφολογική ομοιότητα (ή η διαφορετικότητά) τους. Ο Mikros (2015, σσ. 12-15) ορίζει αυτά τα αποτυπώματα ως «υφομετρικά αποτυπώματα» ή «υφομετρικά γονιδιώματα».

Ο αυτόματος εντοπισμός συγγραφέα, μέσα από την οπτική της Ανάκτησης πληροφοριών, μετατρέπεται σε αλγοριθμικό πρόβλημα, το οποίο επικεντρώνεται όχι στη θεματική αλλά στη μετακειμενική ανάκτηση πληροφοριών. Παράλληλα,τόσο η Επεξεργασία Φυσικής Γλώσσας όσο και η Μηχανική μάθηση έχουν κάνει σημαντικά βήματα προόδου και οι ερευνητές πλέον διαθέτουν ακριβή εργαλεία για να επισημειώσουν τα κείμενα και στη συνέχεια να αναλύσουν αποτελεσματικά τα κειμενικά διανύσματα που παράγονται. Στο πλαίσιο αυτής της ερευνητικής προσέγγισης, ο αυτόματος εντοπισμός συγγραφέα έχει κάνει την τελευταία δεκαετία σημαντική πρόοδο τόσο ως προς την αξιοπιστία των μεθόδων όσο και ως προς την αποτελεσματικότητα και την ευαισθησία των τεχνικών που έχουν αναπτυχθεί. Σήμερα, ο αυτόματος εντοπισμός συγγραφέα ασχολείται κυρίως με τέσσερα διαφορετικά προβλήματα συγγραφικής απόδοσης, τα οποία ταξινομούνται σύμφωνα με τον Mikros (2015, σ. 8) ως εξής:

- **Κλειστά προβλήματα**: Σε αυτά προσπαθούμε να εντοπίσουμε την πατρότητα ενός ή περισσότερων ανώνυμων κειμένων μέσα από μια

συγκεκριμένη (κλειστή) λίστα υποψήφιων συγγραφέων, των οποίων έχουμε διαθέσιμο δείγμα γραφής. Απαντάμε, επομένως, στην ερώτηση «ποιος από τους Α, Β, Γ... έγραψε το κείμενο Χ»

- **Ανοιχτά προβλήματα**: Σε αυτά προσπαθούμε να διερευνήσουμε την πατρότητα ενός ή περισσότερων ανώνυμων κειμένων, έχοντας στην κατοχή μας δείγματα γραφής από έναν ή περισσότερους συγγραφείς για τους οποίους ωστόσο, δεν γνωρίζουμε αν είναι υποψήφιοι συγγραφείς. Στην περίπτωση αυτή ο πραγματικός συγγραφέας του κειμένου μπορεί να είναι οποιοσδήποτε και, επομένως, η ερώτηση που απαντάμε είναι: «έγραψε ο Α συγγραφέας το κείμενο Χ»;

- **Προβλήματα εντοπισμού χαρακτηριστικών του συγγραφέα**: Ο στόχος σε αυτήν την περίπτωση δεν είναι η ταυτοποίηση ενός κειμένου με το πρόσωπο που το έγραψε αλλά με δημογραφικά, ψυχολογικά κ.ά. χαρακτηριστικά του συγγραφέα, π.χ. ο προσδιορισμός του φύλου του συγγραφέα, της ηλικίας του ή ακόμα και της ψυχολογικής του κατάστασης.

- **Προβλήματα υφομετρικής ομοιογένειας**: Σε αυτή την κατηγορία ερευνητικών προβλημάτων εντάσσεται ο αυτόματος εντοπισμός της λογοκλοπής και της κακόβουλης τροποποίησης του περιεχομένου σελίδων συνεργατικών ψηφιακών μέσων (π.χ. αλλοίωση ή καταστροφή των σελίδων της Wikipedia). Ο ερευνητικός στόχος σε αυτήν την περίπτωση είναι η μελέτη της κανονικότητας του υφομετρικού προφίλ ενός κειμένου και η χρήση ποσοτικών μεθόδων για την αξιολόγησή του.

Όλες οι παραπάνω περιπτώσεις ανάγονται σε προβλήματα επιβλεπόμενης μηχανικής μάθησης. Συγκεκριμένα, δεδομένου ενός συνόλου ζευγών (κείμενο, ετικέτα), τα δεδομένα εκπαίδευσης δίνονται ως είσοδος σε έναν αλγόριθμο/μοντέλο μηχανικής μάθησης και το ζητούμενο είναι να βρεθεί το μοντέλο που έχει την ικανότητα να γενικεύει με μεγάλη ακρίβεια σε νέα δεδομένα εκτός του συνόλου εκπαίδευσης.

Σε τυπικές εφαρμογές της απόδοσης συγγραφέα το ζητούμενο είναι να εκπαιδευτεί ένα μοντέλο μηχανικής μάθησης με είσοδο τα κείμενα διαφορετικών συγγραφέων κατά τέτοιο τρόπο ώστε να είναι σε θέση να προβλέψει με μεγάλη ακρίβεια τον συγγραφέα στον οποίο ανήκει ένα νέο, άγνωστο κείμενο. Τυπικά το πρόβλημα ανήκει στην κατηγορία της επιβλεπόμενης μάθησης (supervised learning), στην οποία καθένα από τα κείμενα κωδικοποιείται σε διανυσματικές απεικονίσεις και το ζεύγος (διανυσματική απεικόνιση κείμενου, συγγραφέας) αποτελεί ένα στιγμιότυπο εκπαίδευσης (training instance) για τον αλγόριθμο μηχανικής μάθησης. Το ζητούμενο, λοιπόν, είναι να δημιουργηθεί το σύνολο ζευγών, που θα δοθεί ως είσοδος σε έναν αλγόριθμο μηχανικής μάθησης, και

με βάση αυτή την απεικόνιση ο αλγόριθμος να «μάθει» εμμέσως το υφολογικό αποτύπωμα (signature, footprint) του συγγραφέα.

Όμως, όπως χαρακτηριστικά αναφέρει ο Mikros (2015), η έννοια του αποτυπώματος ύφους είναι μεταφορική και η χρήση της είναι σαφώς επηρεασμένη από βιομετρικές μεθόδους ταυτοποίησης. Οι βιομετρικές όμως μέθοδοι, λόγω της ιδιαίτερης φύσης και της ιδιαιτερότητας των βιομετρικών χαρακτηριστικών, είναι σαφώς πιο ακριβείς. Η πιθανότητα σφάλματος στην ταυτοποίηση ατόμου με βάση το DNA του είναι περίπου μια στο ένα δισεκατομμύριο. Τέτοια υψηλά ποσοστά ακρίβειας είναι μάλλον απίθανο να επιτευχθούν στην υφομετρική ανάλυση, καθώς η παραγωγή κειμένων από ένα άτομο, αν και είναι οπωσδήποτε μια βιολογική διαδικασία, δέχεται πιέσεις από εξωγενείς παράγοντες, κοινωνικούς και ψυχολογικούς ως επί το πλείστον, που καθορίζουν σε μεγάλο βαθμό το παραγόμενο κοινωνικό αποτέλεσμα. Είναι σαφές ότι το βιολογικό αποτύπωμα είναι ντετερμινιστικά καθορισμένο, ενώ, αντίθετα, η χρήση λέξεων και συντακτικών σχημάτων επηρεάζεται από το περιβάλλον του ατόμου, το κοινωνικοπολιτικό περιβάλλον, το επίπεδο εκπαίδευσης και τις επιρροές του (Mikros, 2015, σ. 12). Αυτό βέβαια δεν σημαίνει ότι η μέτρηση και η ποσοτικοποίηση του ατομικού ύφους δεν είναι εφικτή. Τα πειραματικά και εμπειρικά αποτελέσματα δείχνουν το αντίθετο. Ως μαθηματική/στοχαστική διαδικασία όμως είναι εξαιρετικά πιο περίπλοκη και με μεγάλο θόρυβο, ο οποίος πρέπει να ληφθεί υπόψη κατά τη μοντελοποίηση και την αξιολόγηση των αποτελεσμάτων.

Οι απεικονίσεις των κειμένων υλοποιούνται ως διάνυσμα συχνοτήτων των συστατικών που χρησιμοποιούνται για την αποτύπωση της κειμενικής ιδιαιτερότητας. Τυπικά υποψήφια χαρακτηριστικά είναι οι κατανομές λέξεων, διγραμμάτων και τριγραμμάτων λέξεων, καθώς επίσης και οι κατανομές χαρακτήρων, διγραμμάτων και τριγραμμάτων χαρακτήρων και μερών του λόγου. Η υποκείμενη υπόθεση είναι ότι κάθε μονάδα παραγωγής κειμένων έχει διαφορετικές προτιμήσεις και συνήθειες στην επιλογή λέξεων, συντακτικών σχημάτων και σημείων στίξης και ότι αυτή η διαφορετικότητα αντανακλάται στις συχνότητες χρήσης των χαρακτηριστικών, που αναφέρθηκαν παραπάνω. Οι ιδιαίτερες προτιμήσεις των συγγραφέων αντανακλούν εμμέσως το μορφωτικό τους επίπεδο, το οποίο με τη σειρά του επηρεάζει το βάθος λεξιλογίου, τις συντακτικές και γραμματικές διαφοροποιήσεις και τη χρήση σημείων στίξης.

Οι διανυσματικές απεικονίσεις που δημιουργούνται σε υποσύνολο των παραπάνω χαρακτηριστικών, παραδείγματος χάριν τα 1000 πιο συχνά ν-γράμματα (n-grams) από κάθε κατηγορία (Mikros & Perifanos, 2013, σ. 2), δίνονται ως είσοδος σε αλγόριθμους μηχανικής μάθησης. Αλγόριθμοι, που χρησιμοποιούνται στην πράξη, είναι η λογιστική παλινδρόμηση (Logistic Regression), μηχανές διανυσμάτων υποστήριξης (Support Vector Machines), δέντρα απόφασης (Decision trees, Random Forests) ή συνδυασμός των παραπάνω (Stamatatos, 2009).

Το θεωρητικό υπόβαθρο στο οποίο βασίζονται οι παραπάνω εφαρμογές είναι η παραδοχή της ύπαρξης της ιδιολέκτου σύμφωνα με τον Coulthard (2004). Όμως, οι αλγόριθμοι αυτοί δεν δίνουν την άμεση δυνατότητα σύγκρισης της ομοιότητας ύφους των διαφορετικών συγγραφέων, εφόσον δεν μοντελοποιούν άμεσα την ιδιόλεκτο. Αντίθετα, στόχος είναι η μεγιστοποίηση της ακρίβειας του μοντέλου όσον αφορά την πρόβλεψη του σωστού συγγραφέα. Δεδομένης μιας συλλογής κειμένων, που προέρχονται από διάφορους συγγραφείς, για τη μέτρηση της ομοιότητας ύφους μεταξύ των συγγραφέων χρησιμοποιούνται συνήθως τεχνικές ποσοτικοποίησης και απεικόνισης όλων των κειμένων ανά συγγραφέα σε διανυσματική αναπαράσταση. Στη συνέχεια, εφαρμόζονται τεχνικές μέτρησης ομοιότητας μεταξύ των παραχθέντων διανυσμάτων. Τα διανύσματα αυτά κατασκευάζονται με τέτοιο τρόπο ώστε να αντικατοπτρίζουν λεξιλογικές επιλογές των συγγραφέων συνήθως ως συχνότητες χρήσης συχνών λέξεων, αλλά και συντακτικές προτιμήσεις των συγγραφέων, όπως αυτές εκφράζονται ως ακολουθίες μερών του λόγου. Για τη σύγκριση της απόστασης/ομοιότητας μεταξύ αυτών των διανυσμάτων-αποτυπωμάτων, τυπικά μέτρα που χρησιμοποιούνται είναι η απόσταση συνημιτόνου, που ποσοτικοποιεί την ομοιότητα δύο διανυσμάτων ως το εσωτερικό τους γινόμενο, καθώς και μέτρα που προέρχονται από την θεωρία πληροφοριών, όπως η σχετική εντροπία (KL Divergence, Stamatatos, 2009).

Επομένως, στην προσέγγιση της ιδιολέκτου ως υφολογικού αποτυπώματος ποσοτικοποιείται με την εφαρμογή διανυσμάτων η διαφορά μεταξύ του ύφους συγγραφέων, οι οποίοι ωστόσο είναι εκ των προτέρων γνωστοί και λίγοι σχετικά σε αριθμό. Στην παρούσα διατριβή το ζητούμενο είναι η αναπαράσταση των συγγραφέων ως προς το ύφος τους ως μαθηματικές οντότητες, με τέτοιο τρόπο ώστε να είναι συγκρίσιμες ποσοτικά ως προς την ομοιότητά τους. Υπό αυτή την οπτική, τα μοντέλα μηχανικής μάθησης που χρησιμοποιούνται ανήκουν περισσότερο στην κατηγορία της μη επιβλεπόμενης μάθησης (unsupervised learning), καθώς δεν υπάρχει εκ των προτέρων γνωστή μαθηματική αναπαράσταση του ύφους. Αντίθετα, αυτή δημιουργείται έμμεσα ως υποπροϊόν αλγόριθμων επιβλεπόμενης μάθησης ή ως αποτέλεσμα μαθηματικής παραγοντοποίησης.

2.6 Η ιδιόλεκτος ως διανυσματικό πρόβλημα μεγάλης κλίμακας

Η πρώτη απόπειρα ανάλυσης κειμενικού ύφους σε μεγάλη κλίμακα γίνεται από τους Hughes et al. (2012). Στην ανάλυση αυτή εξετάζονται 7.733 κείμενα από 537 συγγραφείς από το Project Gutenberg[11], την τεράστια βάση δεδομένων, κυρίως

[11] https://www.gutenberg.org/

λογοτεχνικών, κειμένων του παρελθόντος. Η ανάλυση αυτή επικεντρώνεται στην χρήση των 307 πιο συχνών λειτουργικών λέξεων. Πιο συγκεκριμένα, η επιλογή των συγγραφέων από τους Hughes et. al γίνεται με τα εξής κριτήρια: επιλέγονται συγγραφείς με γλώσσα γραφής την αγγλική, από το έτος 1550 και έπειτα με τουλάχιστον 5 κείμενα στην συλλογή του Project Gutenberg και για τους οποίους είναι καταγεγραμμένη η ημερομηνία γέννησης και θανάτου.

Κατά τη διάρκεια της προεπεξεργασίας, δημιουργούνται διανύσματα για τον κάθε συγγραφέα με τις συχνότητες εμφάνισης των 307 λειτουργικών λέξεων και στη συνέχεια τα διανύσματα αυτά μετατρέπονται σε πιθανοτικά ώστε το άθροισμα των συχνοτήτων να αθροίζει στην μονάδα. Για τον υπολογισμό της ομοιότητας μεταξύ δύο διανυσμάτων χρησιμοποιείται η συμμετρική σχετική εντροπία (Relative entropy, KL Divergence)[12].

Εξίσου μεγάλης κλίμακας είναι η ανάλυση των Bamman et al. (2014), οι οποίοι εξετάζουν την περίπτωση της λεξικής ποικιλίας σε κοινωνικά δίκτυα, και πιο συγκεκριμένα στο Twitter, σε σώμα κειμένων προερχόμενο από 14.464 χρήστες και 9,2 εκατομμυρίων λέξεων, που συλλέχθηκε έως το 2012. Στο σώμα κειμένων εφαρμόστηκαν διάφορα στάδια επεξεργασίας (αφαίρεση χρηστών εκτός ΗΠΑ, περιορισμός χρηστών με βάση τις αλληλεπιδράσεις, κανονικοποίηση κατά φύλο με βάση δημογραφικά στοιχεία). Στόχος τους είναι η ταξινόμηση φύλου και η προσέγγιση τους βασίζεται σε τεχνικές μηχανικής μάθησης. Χρησιμοποιώντας λογιστική παλινδρόμηση, με αναπαράσταση οικογένειας λέξεων στις 10.000 πιο συχνές λέξεις του σώματος κειμένων, εκπαιδεύουν έναν ταξινομητή και αναζητούν τους λεξικούς δείκτες που διαφοροποιούν τη χρήση του λεξιλογίου μεταξύ φύλων ή διαφορετικά τις λέξεις εκείνες που λειτουργούν ως δείκτες. Η επιλογή λέξεων γίνεται με στατιστικό έλεγχο στις 10.000 λέξεις που δίδονται ως είσοδος στο μοντέλο. Στη συνέχεια προχωρούν σε εφαρμογή τεχνικών συσταδοποίησης (clustering), η οποία έχει ως στόχο την ομαδοποίηση χρηστών με βάση την ομοιότητα λεξικών επιλογών, στο σύνολο των 10.000 λέξεων και αντιστρέφουν την ροή της ανάλυσης εξετάζοντας ποιοτικά τις συστάδες (clusters) των χρηστών, καθώς και τον λόγο μεταξύ αντρών/γυναικών σε κάθε συστάδα.

Η ποιοτική ανάλυση δίνει ενδιαφέροντα αποτελέσματα και καταλήγει με το συμπέρασμα στο οποίο βασίζεται η παρούσα διατριβή. Η ασάφεια όσον αφορά τη χρήση της γλώσσας δεν αποτελεί στατιστικό θόρυβο αλλά ενσυνείδητες και ασυνείδητες επιλογές των χρηστών, οι οποίες απορρέουν από τις στάσεις, την ταυτότητα και τις κοινωνικές συσχετίσεις που έχει επιλέξει ο κάθε χρήστης. Οι χρήστες μεταβάλλουν το ύφος τους ώστε να ταιριάξει με εκείνο των συνομιλητών τους στο πλαίσιο της κοινωνικής δικτύωσης. Η επιλογή συνομιλητών και κοινωνικής ομάδας εμπεριέχει προφανώς προφανή ή λανθάνουσα τοποθέτηση και πληροφορίες όσον αφορά το κοινωνικό και πολιτικό καθεστώς, το θέμα συζήτησης

[12]Για τις διακριτές κατανομές P,Q η σχετική εντροπία ορίζεται ως $D_{KL}(P||Q) = \sum_i P_i log \frac{P_i}{Q_i}$

(topic), την ηλικιακή ομάδα και την ομάδα του φύλου.

Οι δύο αυτές αναλύσεις των Hughes et al. (2012) και Bamman et al. (2014) διαφοροποιούνται από τις προηγούμενες των Mollin (2009), Barlow (2013) και Φραγκάκη (2012) αφού δεν εξετάζουν τη διαφοροποίηση της χρήσης λέξεων ή συμπλεγμάτων. Αντίθετα, προχωρούν σε αναπαράσταση διανυσματικού χώρου χρησιμοποιώντας προκαθορισμένο αριθμό λέξεων και στη συνέχεια εφαρμόζουν τεχνικές από την Ανάκτηση πληροφοριών και από τη Μηχανική μάθηση για την εξαγωγή συμπερασμάτων. Διαφέρει επίσης από τις προσεγγίσεις στην ιδιόλεκτο ως υφολογικό αποτύπωμα, εφόσον δεν περιορίζονται σε μικρό αριθμό συγγραφέων και επομένως αντιμετωπίζουν την ανάλυση της ιδιολέκτου ως ανοικτό πρόβλημα. Ωστόσο, ο αριθμός λέξεων που χρησιμοποιείται και σε αυτές τις προσεγγίσεις για τον καθορισμό της ιδιολεκτικής συμπεριφοράς των δημιουργών κειμένων είναι περιορισμένος. Ειδικότερα στην περίπτωση των Hughes et al. (2012), η επιλογή μόνο 307 λέξεων ως αναπαράσταση του ιδιολεκτικού διανύσματος του χρήστη περιορίζεται ως επί το πλείστον σε λειτουργικές λέξεις και επομένως εμπίπτει στα προβλήματα που επισημάνθηκαν πιο πάνω (2.4) για τη χρήση γραμματικών τύπων.

2.7 Συμπεράσματα

Η ανασκόπηση της βιβλιογραφίας έδειξε ότι η ιδιόλεκτος αποτελεί μια αναγνωρισμένη γλωσσική ποικιλία, η οποία, σε αντίθεση με την κοινωνιόλεκτο, δεν έχει αναλυθεί με εμπειρικά δεδομένα. Μια βασική αιτία γι' αυτό είναι η αντιμετώπιση της ιδιολέκτου ως της ατομικής πτυχής του γλωσσικού συστήματος. Επιπλέον, η παραδοσιακή κοινωνιογλωσσολογία εξ ορισμού δίνει έμφαση σε κοινωνικές και όχι ατομικές ποικιλίες, ενώ η γλωσσολογική υφολογία λαμβάνει ως δεδομένη την ιδιόλεκτο του λογοτέχνη και δεν προχωρεί στη θεμελίωση της ομοιότητας ή της διαφοροποίησης του ύφους του από εκείνο άλλων.

Οι αναλύσεις που σχετίζονται με τη γλωσσολογία σωμάτων κειμένων είναι οι πρώτες απόπειρες για αντιμετώπιση της ιδιολέκτου ως πτυχή της γλωσσικής πραγμάτωσης. Ταυτόχρονα, οι έρευνες που σχετίζονται με την υφομετρία και την απόδοση συγγραφέα στο πλαίσιο της δικανικής γλωσσολογίας έχουν ορίσει την ιδιόλεκτο ως το υφολογικό αποτύπωμα ενός συγγραφέα και έτσι έχουν εξειδικεύσει την ανίχνευση της ιδιολέκτου ως μαθηματικό πρόβλημα, ενώ ορισμένες πρόσφατες έρευνες έχουν αντιμετωπίσει το πρόβλημα ως διανυσματικό εξετάζοντας μεγάλης κλίμακας δεδομένα.

Οι έως τώρα εμπειρικές απόπειρες για τον καθορισμό της ιδιολέκτου στη βιβλιογραφία επικεντρώνονται στην εξαγωγή συμπερασμάτων από τη συχνότητα χρήσης λέξεων (κυρίως γραμματικών τύπων) και διγραμμάτων, σε σχετικά μικρό όγκο κειμένων και σε μικρό αριθμό συγγραφέων. Η μελέτη των (Hughes et al.,

2012) είναι η πρώτη μεγάλης κλίμακας αλλά και σε αυτή εξετάζεται η ιδιολεκτική ομοιότητα σε περιορισμένο αριθμό υψίσυχνων λέξεων, ενώ οι Bamman et al. (2014) επικεντρώνονται στην κειμενική ταξινόμηση του φύλου, και όχι στην ιδιόλεκτο, χρησιμοποιώντας παρόμοιες τεχνικές.

Επιπλέον, οι προσεγγίσεις των Barlow, Mollin και Φραγκάκη στηρίζονται στις διαφοροποιήσεις μεταξύ των ατόμων στη συχνότητα χρήσης υψίσυχνων λέξεων, μερών του λόγου και συνάψεων. Τα αποτελέσματα που παρουσιάζουν είναι αναμενόμενα τηρουμένου του αριθμού ατόμων που εξετάζονται (5 στην περίπτωση Barlow, 1 στην περίπτωση της Mollin, 3 στην περίπτωση της Φραγκάκη). Όπως έχει επισημανθεί, όμως, στο πλαίσιο της δικανικής γλωσσολογίας, η χρήση της συχνότητας συνάψεων ως δείκτη ιδιολέκτου πρέπει να αντιμετωπίζεται με προσοχή και πάντα σε σχέση με το σώμα κειμένων αναφοράς (Goutsos, 1995). Ενώ οι Huges et. al. και οι Bamman et. al., επεκτείνουν το σώμα κειμένων αυξάνοντας κατά πολύ τον αριθμό των συγγραφέων/χρηστών, κρατούν τη χρήση του λεξιλογίου και των συχνών λέξεων σε σχετικά χαμηλά επίπεδα ή απλοποιούν κατά πολύ τον έλεγχο, καθώς επικεντρώνονται μόνο σε δυαδικές διαφοροποιήσεις ύφους που αφορούν στοιχεία όπως το φύλο (gender).

Οι μελέτες αυτές, ενώ αποτελούν εξαιρετικά καλή βάση για περαιτέρω έρευνα μεγάλης κλίμακας, δεν απαντούν σε ερωτήματα μεγιστοποίησης της κλίμακας: τι συμβαίνει στην περίπτωση χιλιάδων συγγραφέων ή μερικών εκατομμυρίων λέξεων; Κατά πόσο οι μεθοδολογίες που είναι βασισμένες σε συχνότητες μπορούν να επεκταθούν σε μεγαλύτερη κλίμακα; Υπάρχει κάποιο ποσοτικό μέτρο με το οποίο μπορούμε να κωδικοποιήσουμε τον βαθμό ομοιότητας των κειμένων ενός ατόμου Α με ένα άτομο Β; Υπάρχει ομαδοποίηση ή κοινά χαρακτηριστικά τα οποία συνιστούν μια ομάδα συγγραφέων και ορίζουν -κατά μια έννοια- μια γλωσσική ποικιλία ή λέκτο μη γεωγραφικά ή κοινωνικά προσδιορισμένης;

Επιπλέον, τα χαρακτηριστικά που χρησιμοποιήθηκαν για την ποσοτική και ποιοτική ανάλυση του κειμενικού ύφους στις υπάρχουσες εμπειρικές έρευνες έχουν εστιάσει σε συχνότητες χρήσης λέξεων και επαναλαμβανόμενα σχήματα μερών του λόγου που αφορούν τις πιο συχνά εμφανιζόμενες λέξεις, που αποτελούν γραμματικούς τύπους. Στην πραγματικότητα, όμως, η κειμενική συνάφεια φέρει και χαρακτηριστικά σημασίας, που εντοπίζονται σε λεξικούς τύπους ή λέξεις πλήρους σημασιολογικού περιεχομένου, οι οποίες δεν περιλαμβάνονται στις παραπάνω μεθοδολογίες.

Αν και η χρήση υψίσυχνων λέξεων συνδέεται με τις ασυνείδητες προτιμήσεις των συγγραφέων, είναι αυτονόητο ότι η παρόμοια χρήση λεξιλογίου είναι ένα από τα βασικά χαρακτηριστικά που αντικατοπτρίζουν την συνάφεια. Θα πρέπει επομένως να αναπροσανατολιστεί η ομοιότητα και το ιδιαίτερο ύφος ενός ατόμου σε σημασιολογικά στοιχεία και βαθύτερες και πολυπλοκότερες δομές από τη συχνότητα χρήσης υψίσυχνων λέξεων και μερών του λόγου.

Επίσης, η χρήση συχνοτήτων λέξεων και μόνο ως μέτρο ιδιολεκτικής

συνάφειας είναι επιρρεπής σε εσφαλμένες κατηγοριοποιήσεις (false positives) καθώς δεν είναι απίθανο σε μεγάλη κλίμακα δύο ή περισσότεροι χρήστες να έχουν παρόμοιο αποτύπωμα και τα κείμενα τους να είναι εντελώς ξένα μεταξύ τους. Η πιθανότητα εσφαλμένης κατηγοριοποίησης αυξάνει προφανώς όσο αυξάνει ο αριθμός των συγγραφέων στο σώμα κειμένων, ενώ ο αριθμός των γραμματικών τύπων παραμένει σταθερός σε κάθε γλώσσα, εφόσον αποτελούν κλειστό σύνολο.

Το προφανές επόμενο βήμα, λοιπόν, είναι να διερευνηθεί εάν και κατά πόσον είναι εφικτή η χρήση σημασιολογικών διαφοροποιήσεων που περιλαμβάνουν και μη λειτουργικές λέξεις. Το ερώτημα αυτό μπορεί να επιλυθεί στο πλαίσιο της Κατανεμητικής Σημασιολογίας, καθώς η βασική της αρχή είναι ότι παρόμοιες σημασιολογικά λέξεις τείνουν να εμφανίζονται στα ίδια περικείμενα (Harris, 1954). Εφόσον η ανίχνευση της ιδιολέκτου βασίζεται στη διαφοροποίηση συχνοτήτων χρήσης μεταξύ λέξεων, η επέκταση της έρευνας σε όλο το λεξιλόγιο ενός σώματος κειμένων θα συμπεριλαμβάνει τις διαφορές μεταξύ των συγγραφέων στη χρήση λέξεων σε διαφορετικά περικείμενα. Συνεπώς, με αυτόν τον τρόπο μπορούν να καταγραφούν οι υφολογικές διαφοροποιήσεις των συγγραφέων σε βαθύτερο, σημασιολογικό επίπεδο, πέρα από την επιφανειακή χρήση των λειτουργικών λέξεων. Στο επόμενο κεφάλαιο αναλύονται οι αρχές της Κατανεμητικής Σημασιολογίας και οι εφαρμογές της στα ερωτήματα της παρούσας διατριβής.

Τέλος, μια επιπλέον παράμετρος που σχετίζεται με την ιδιόλεκτο είναι αυτή της διακειμενικότητας ως διαμόρφωσης ενός κειμένου ως σύνθεσης άλλων κειμένων (De Beaugrande & Dressler, 1981). Τυπικά σχήματα διακειμενικότητας είναι ο υπαινιγμός, η αναφορά, το μεταφραστικό δάνειο, η λογοκλοπή, η μεταφορά, η μίμηση και η παρωδία, ενώ μπορούμε να θεωρήσουμε ότι η διακειμενικότητα αφορά τη μηχανική της δημιουργίας σχέσεων (interelationships) μεταξύ κειμένων. Η διακειμενικότητα είναι τεχνική που χρησιμοποιείται σε διάφορα κειμενικά είδη, αλλά και σε μορφές τέχνης (Gadavanij, 2002, σ. 3), (Ivanič, 1998, σσ. 84,85), ενώ αποτελεί συνηθισμένο φαινόμενο στα κοινωνικά δίκτυα, όπου χρήστες υιοθετούν εκφράσεις άλλων χρηστών είτε επειδή γίνονται viral, είτε για λόγους αναγνωρισιμότητας και τοποθέτησης σε συγκεκριμένο σύνολο χρηστών. Στην παρούσα διατριβή ο εντοπισμός της ιδιολέκτου με υπολογιστικές μεθόδους σε πλατφόρμες κοινωνικών δικτύων και ιστολογίων επεκτείνεται και στην έννοια της διακειμενικότητας για τη μελέτη της διάχυσης του κειμενικού ύφους μεταξύ ομιλητών και την ομαδοποίηση κατά κειμενικό ύφος.

Κεφάλαιο 3

Δεδομένα της διατριβής

Στο κεφάλαιο αυτό επικεντρωνόμαστε στα χαρακτηριστικά των σωμάτων κειμένων που χρησιμοποιούνται σε αυτή τη διατριβή. (Για τη θεωρία και τη μεθοδολογία της γλωσσολογίας σωμάτων κειμένων, βλ. Γούτσος και Φραγκάκη 2015). Το πρώτο συλλέχθηκε από την πλατφόρμα Twitter και αποτελείται από κείμενα ως επί το πλείστον στα ελληνικά, ενώ το δεύτερο είναι το Blog Authorship Corpus των Schler et al. (2006). Παρουσιάζονται τα στατιστικά στοιχεία των σωμάτων κειμένων, οι πιο συχνές λέξεις του πρώτου συνολικά και κάθε έτους συλλογής των δεδομένων και ακολουθεί σύγκριση με το Σώμα Ελληνικών Κειμένων (ΣΕΚ), καθώς και ποιοτική ανάλυση των διαφορών που εντοπίζονται στα ποσοτικά χαρακτηριστικά μεταξύ του σώματος κειμένων Twitter και του ΣΕΚ. Στη συνέχεια περιγράφεται το δεύτερο σώμα κειμένων και αναπτύσσονται οι λόγοι για τους οποίους χρησιμοποιείται στην παρούσα διατριβή.

3.1 Σώμα κειμένων Twitter

Δεδομένου ότι σκοπός της παρούσας έρευνας είναι η μελέτη της ιδιόλεκτου και της κειμενικής ομοιότητας σε μεγάλη κλίμακα, φαίνεται αυτονόητη η συλλογή δεδομένων από κοινωνικά δίκτυα και πιο συγκεκριμένα από το Twitter. Οι λόγοι που συνέβαλαν στην επιλογή αυτή αφορούν κυρίως την ποικιλία κειμενικών δεδομένων, τον όγκο και το χρονικό διάστημα που καλύπτει το σώμα κειμένων. Πιο συγκεκριμένα, συλλέχθηκαν 26 εκατομμύρια αναρτήσεις (tweets) από 4.494 χρήστες, που καλύπτουν ένα χρονικό διάστημα από το 2009 έως το 2016. Το σώμα κειμένων από μόνο του παρουσιάζει εξαιρετικά ενδιαφέροντα ποσοτικά και ποιοτικά χαρακτηριστικά ως προς τις συχνότητες των πιο συχνών λέξεων, τα οποία θα παρουσιαστούν στη συνέχεια λεπτομερώς.

Το Twitter (http://www.twitter.com) είναι ένα κοινωνικό δίκτυο που επιτρέπει στους χρήστες να δημοσιεύουν και αλληλεπιδρούν μεταξύ τους με σύντομα

μηνύματα (microblogging) τα οποία πρέπει να περιλαμβάνουν περιορισμένο αριθμό χαρακτήρων. Καθώς η συλλογή κειμένων αφορά την περίοδο έως τα μέσα του 2016, οι αναρτήσεις έχουν μέγιστο μήκος 140 χαρακτήρων[1]. Οι χρήστες μπορούν να ακολουθήσουν άλλους χρήστες ή να αποκτήσουν ακόλουθους (followers), σχηματίζοντας έτσι ένα κοινωνικό δίκτυο αλληλοσυνδεδεμένων χρηστών. Στην πλειονότητα τους οι αναρτήσεις των χρηστών είναι δημόσιες, δηλαδή οι χρήστες μπορούν να διαβάσουν και σχολιάσουν τις αναρτήσεις όλων των άλλων χρηστών, εκτός από τους χρήστες εκείνους που έχουν απενεργοποιήσει την λειτουργία αυτή και επιτρέπουν αλληλεπίδραση μόνο στους ακολούθους τους.

Επιπλέον, το Twitter έχει ιδιαίτερες συμβάσεις όπως τη χρήση των ετικετών @tags και #tags (τα λεγόμενα hashtags) για να συνδεθεί μια ανάρτηση με ορισμένους χρήστες ή ορισμένα θέματα, αντίστοιχα, καθώς και τη δυνατότητα επανάληψης μια ανάρτησης άλλου χρήστη (retweet), αλλά και ανάρτησης φωτογραφιών, ιστολογίων, βίντεο κ.λπ.

Το μεγαλύτερο μέρος της δημοσιευμένης έρευνας για το Twitter αφορά τις συμβάσεις αυτές (π.χ. Honey & Herring, 2009; Boyd et al., 2010; Huang et al., 2010; Browning, 2017) και ιδίως τη συμβολή τους στις διαπροσωπικές σχέσεις των συμμετεχόντων (Zappavigna, 2012), ενώ δεδομένα από το Twitter έχουν αναλυθεί και με μεθόδους μηχανικής μάθησης (Zappavigna, 2012; Kalyanam, 2017), όπως επιχειρείται και στην παρούσα διατριβή.

3.1.1 Στατιστικά του Σώματος Κειμένων Twitter

Το σώμα κειμένων αποτελείται από 26.103.963 αναρτήσεις 4.494 χρηστών και συνολικά 325.243.302 λέξεις (white space tokenized). Το μέσο μήκος ανάρτησης είναι 12,45 λέξεις, η διάμεσος 12 λέξεις και η τυπική απόκλιση 6,22. Η κατανομή των λέξεων ανά δημοσίευση φαίνεται στο διάγραμμα Ε.2 και τα στατιστικά συνοψίζονται στον πίνακα 3.1.

Αριθμός χρηστών	4,494
Αριθμός αναρτήσεων	26,103,963
Αριθμός λέξεων	325μ243,302
Μέσος αριθμός λέξεων/tweet	12.45
Τυπική απόκλιση λέξεων/tweet	6.22

Πίνακας 3.1: Στατιστικά Σώματος twitter

Στο διάγραμμα Ε.3 βλέπουμε το ιστόγραμμα του αριθμού κοινοποιήσεων των χρηστών, το οποίο υποδηλώνει ότι η κατανομή παραγωγικότητας ακολουθεί κατά τα αναμενόμενα τον νόμο του Zipf.

[1] Το μέγιστο μήκος των χαρακτήρων άλλαξε τον Νοέμβρη του 2017 σε 280 χαρακτήρες

Στη συνέχεια, στο διάγραμμα Ε.4 στο παράρτημα Ε βλέπουμε την κατανομή των 25 πιο συχνών λέξεων στο σώμα κειμένων και ακολούθως την ίδια κατανομή ανά έτος καθώς επίσης και τον αριθμό των κειμένων ανά έτος.

Από τα διαγράμματα των κατανομών των πιο συχνών λέξεων προκύπτουν δύο εξαιρετικά ενδιαφέροντα χαρακτηριστικά. Κατ' αρχήν, η κατάταξη των πιο συχνών λέξεων δεν είναι σταθερή στο πέρασμα του χρόνου και αποκλίνει από την αναμενόμενη κατανομή των πιο συχνών λέξεων του Σώματος Ελληνικών Κειμένων (Goutsos, 2010).

Επίσης, το σχήμα της κατανομής των πιο συχνών λέξεων στο σώμα κειμένων έχει σημαντικές αποκλίσεις από την αναμενόμενη κατανομή Zipf. Από τα διαγράμματα φαίνεται ότι υπάρχουν 3 ή 4 υποομάδες λέξεων, κάθε μια από τις οποίες έχει διαφορετική συμπεριφορά συχνότητας/κλίσης. Ως συνολικό συμπέρασμα εξάγεται ότι η κατανομή δεν είναι, συγκριτικά πάντα με το ΣΕΚ, τόσο ομαλή όσο θα ανέμενε κανείς για κατανομή συχνότητας λέξεων.

3.1.2 Συχνότητα λέξεων στο σώμα κειμένων Twitter

Για την περίπτωση της κατάταξης των πιο συχνών λέξεων, μια πιθανή εξήγηση μπορεί να θεωρηθεί πως το όριο των 140 χαρακτήρων επιβάλλει στους χρήστες οικονομία και προσεκτικότερη διαχείριση με αποτέλεσμα την κωδικοποίηση συνδέσμων, όπως το *και* με ισοδύναμους ή σχεδόν ισοδύναμους χαρακτήρες όπως το *κ* ή το *&* καθώς επίσης και το σύμβολο + . Με άλλα λόγια, μπορούμε να πούμε ότι οι περιορισμοί που επιβάλλει το μέσο αναγκάζουν τους χρήστες να αλλάξουν γλωσσική συμπεριφορά και να προσαρμοστούν σε αυτό. Μια επιπλέον παρατήρηση στην κατανομή των πιο συχνών λέξεων είναι ότι δεν υπάρχει ο αναμενόμενος αριθμός ρημάτων. Φαίνεται πως ο περιορισμός των χαρακτήρων έχει ως αποτέλεσμα την διαταραχή στη συντακτική ανάπτυξη των προτάσεων αποκόπτοντας τον πυρήνα των λέξεων, δηλαδή το ρήμα. Στο ίδιο πλαίσιο της οικονομίας αποφεύγεται η χρήση κύριων ονομάτων και τη θέση τους παίρνουν αντωνυμίες και ουσιαστικά που μπορούν να αναφερθούν στον ίδιο όρο χωρίς να απαιτείται ονομαστική φράση που θα περιλαμβάνει άρθρο και ουσιαστικό. Από την υψηλή συχνότητα εμφάνισης του *δε(ν)* σε συνδυασμό με την απουσία εμφάνισης του δείκτη *μη(ν)*, μπορούμε να συμπεράνουμε ότι η πλειονότητα των αναρτήσεων είναι προτάσεις κρίσεως και όχι επιθυμίας. Τέλος, η εμφάνιση του δείκτη *τι* από το 2013 και μετέπειτα στη λίστα των υψίσυχνων λέξεων υποδηλώνει ενδεχομένως ότι αυξάνεται το ποσοστό αναρτήσεων που θέτουν ερωτήσεις. Αυτό επαληθεύεται και από το σώμα κειμένων, αν μετρηθεί ο λόγος των εμφανίσεων της λέξης ανά έτος, όπως φαίνεται και στο διάγραμμα Ε.1.

Φαίνεται δηλαδή πως ο συνδυασμός οικονομίας και του φαινομένου ανάδρασης (feedback loop), που προκαλείται από δημοφιλείς εκφράσεις και χρήστες με μεγάλη επιρροή, μεταβάλλουν δυναμικά τα στατιστικά της χρήσης

του γραπτού λόγου στο μέσο.

3.2 Το σώμα κειμένων ιστολογίων Blog Authorship Corpus

Το δεύτερο σώμα κειμένων, στο οποίο εφαρμόστηκε η μέθοδος ανίχνευσης ιδιολέκτου, που αναπτύσσεται στην παρούσα διατριβή, είναι το Σώμα ιστολογίων Blog Authorship Corpus (BAC) των Schler et al. (2006). Πρόκειται για ένα σώμα κειμένων που έχει χρησιμοποιηθεί εκτενώς στη βιβλιογραφία (βλ. Nguyen et al., 2016) και χρησιμοποιείται στην παρούσα διατριβή πρώτον για επαλήθευση των αποτελεσμάτων σε κείμενα με διαφορετική κειμενική συμπεριφορά των χρηστών, δηλαδή χωρίς τους περιορισμούς μήκους ανά δημοσίευση που θέτει το Twitter και, δεύτερον, για τη γενίκευση των αποτελεσμάτων σε κείμενα στην αγγλική γλώσσα. Το Σώμα κειμένων ιστολογίων BAC αποτελείται από 681.288 καταχωρήσεις 19.320 συγγραφέων και περίπου 137 εκατομμύρια λέξεις, δηλαδή 35 καταχωρήσεις και 7.250 λέξεις ανά συγγραφέα. Τα κείμενα καλύπτουν μια περίοδο 7 ετών, από το 1999 έως το 2006.

Η κατανομή των λέξεων στο σώμα κειμένων ανά συγγραφέα φαίνεται στο γράφημα Ε.5 του Παραρτήματος Ε.

Αριθμός bloggers	19,320
Αριθμός αναρτήσεων	681,288
Αριθμός λέξεων	136,818,878
Μέσος αριθμός λέξεων/ανάρτηση	200
Τυπική απόκλιση λέξεων/ανάρτηση	415.21

Πίνακας 3.2: Στατιστικά Σώματος κειμένων BAC

3.3 Συμπεράσματα

Στο κεφάλαιο αυτό περιγράφηκαν τα σώματα κειμένων που χρησιμοποιούνται σε αυτή τη διατριβή. Γίνεται λεπτομερής περιγραφή του Σώματος Κειμένων του Twitter, το οποίο συλλέχθηκε στο πλαίσιο της παρούσας διατριβής. Σημειώνονται επίσης οι διαφοροποιήσεις στις συχνότητες των λέξεων σε σχέση με το Σώμα Ελληνικών Κειμένων. Η ερμηνεία για την αιτία αυτών των διαφοροποιήσεων εντοπίζεται στον περιορισμό των 140 χαρακτήρων, που επιβάλλει η πλατφόρμα του Twitter.

Επιπλέον, αναφέρονται τα στατιστικά του Σώματος κειμένων Blog Authorship Corpus και οι διαφορές με το Σώμα κειμένων Twitter. Οι κύριες διαφορές

σε σχέση με αυτό εντοπίζονται στον αριθμό συγγραφέων, που είναι σχεδόν πενταπλάσιος από το Σώμα κειμένων Twitter, και στο μέσο αριθμό λέξεων ανά συγγραφέα, που είναι πολύ μικρότερος από εκείνον του Σώματος κειμένων Twitter. Βασική διαφορά είναι ότι το Σώμα κειμένων Twitter είναι ως επί το πλείστον στην ελληνική γλώσσα, ενώ το Σώμα κειμένων BAC περιλαμβάνει κείμενα στην αγγλική γλώσσα.

Κεφάλαιο 4

Διανυσματικές αναπαραστάσεις και λεξικές ενθέσεις

You can't learn too much linear algebra. Benedict Gross

You shall know a word by the company it keeps (Firth, 1957)

Στο Κεφάλαιο 2 η βιβλιογραφική ανασκόπηση της έννοιας της ιδιολέκτου κατέληξε στον ορισμό της ως διανυσματικό πρόβλημα μεγάλης κλίμακας. Στο κεφάλαιο αυτό συζητείται η μεθοδολογία που έχει αναπτυχθεί για την αναπαράσταση διανυσματικών προβλημάτων με υπολογιστικό τρόπο, και ειδικότερα εκείνα τα μοντέλα που επιτρέπουν την αναπαράσταση της σημασίας. Πρόκειται για τα Μοντέλα διανυσματικών χώρων (vector space models, VSM), που αποτελούν τη βάση της μαθηματικής και στατιστικής αναπαράστασης της σημασιολογικής συνάφειας των λέξεων ενός σώματος κειμένων, η οποία, σύμφωνα με την κατανεμητικής υπόθεση του Harris (1954), συνδέεται άρρηκτα με την ίδια τη σημασία των λέξεων.

Τα μοντέλα διανυσματικών χώρων παρέχουν τα απαραίτητα εργαλεία για την μαθηματική περιγραφή του προβλήματος και αποτελούν την πρώτη, βασική μορφή αφηρημένης αναπαράστασης, που αποτελεί δομικό στοιχείο της μηχανικής μάθησης. Η αναλυτική περιγραφή της δημιουργίας τους στις πέντε πρώτες ενότητες ακολουθείται από τη συζήτηση του ρόλου της Κατανεμητικής σημασιολογίας και της μετάβασης σε πιο σύνθετα και περίπλοκα μοντέλα

νευρωνικών δικτύων, που αντικαθιστούν πλέον τις παραδοσιακές τεχνικές, εισάγοντας την έννοια των λεξικών ενθέσεων στην έκτη ενότητα. Έπειτα από την αναλυτική παρουσίαση των μοντέλων που εκπαιδεύονται στην παρούσα διατριβή, υποστηρίζεται ότι οι λεξικές ενθέσεις μπορούν να αναλύσουν με ιδανικό τρόπο την ιδιόλεκτο, καθώς μπορούν να χρησιμοποιηθούν για την περιγραφή και ανάλυση ατομικών προτύπων και ιδιαιτεροτήτων που συνθέτουν το ιδιολεκτικό ύφος, με τη συμπερίληψη λεξικών τύπων με πλήρες σημασιολογικό περιεχόμενο και κάλυψη του συνόλου του λεξιλογίου ενός σώματος κειμένων.

4.1 Μοντέλα Διανυσματικών Χώρων

Στόχος των σημασιολογικών μοντέλων διανυσματικών χώρων είναι η αναπαράσταση των κειμένων ενός σώματος κειμένων σε ένα d-διάστατο διανυσματικό χώρο. Με άλλα λόγια, κάθε κείμενο ενός σώματος κειμένων μπορεί να αναπαρασταθεί από ένα d-διάστατο διάνυσμα, συνήθως στο R^D, με τέτοιο τρόπο ώστε σημεία που βρίσκονται γεωμετρικά κοντά στον διανυσματικό χώρο να αναφέρονται σε σημασιολογικά σχετιζόμενα κείμενα και σημεία με μεγάλη απόσταση σε κείμενα που δεν έχουν σημασιολογική συνάφεια μεταξύ τους.

Η εισαγωγή διανυσματικών μοντέλων εφαρμόστηκε για πρώτη φορά στο πεδίο της Ανάκτησης πληροφοριών, και πιο συγκεκριμένα στο σύστημα ανάκτησης πληροφορίας SMART (Salton et al., 1975). Η επιτυχία της εφαρμογής διανυσματικών μοντέλων σε συστήματα ανάκτησης πληροφοριών οδήγησε στην εφαρμογή τους και σε περιβάλλοντα Επεξεργασίας Φυσικής Γλώσσας.

Οι διανυσματικές αναπαραστάσεις, βέβαια, είναι δημοφιλείς στην Τεχνητή Νοημοσύνη και τη γνωσιακή επιστήμη και προϋπάρχουν των τεχνικών VSM. Η καινοτομία των VSM έγκειται στη χρήση συχνοτήτων όρων στην αναπαράσταση. Στη μηχανική μάθηση, για παράδειγμα, οι περισσότεροι αλγόριθμοι έχουν ως είσοδο ένα διάνυσμα συγκεκριμένων διαστάσεων και έξοδο είτε έναν πραγματικό αριθμό (παλινδρόμηση, regression), είτε μια κατανομή πιθανότητας, ένα διάνυσμα αριθμών δηλαδή στο διάστημα $[0, 1]$, το οποίο αθροίζει στη μονάδα.

Κοινό χαρακτηριστικό των αναπαραστάσεων VSM είναι η δημιουργία διανυσματικών μοντέλων/αναπαραστάσεων που δίνουν τη δυνατότητα αλγεβρικής και στατιστικής επεξεργασίας της σημασίας, αναπτύσσοντας τη λεγόμενη Στατιστική Σημασιολογία (Statistical Semantics, Turney & Pantel, 2010, σ. 146). Διανυσματικές αναπαραστάσεις χρησιμοποιούνται επίσης σε Συστήματα Συστάσεως Περιεχομένου (Recommender Systems), καθώς και στη γνωσιακή επιστήμη μέσω της τεχνικής Λανθάνουσας Σημασιολογικής Ανάλυσης (Latent Semantic Analysis, LSA) (Deerwester et al., 1989). Η τεχνική LSA εισάγει την έννοια ενός λανθάνοντα σημασιολογικού χώρου, στον οποίο

κείμενα και λέξεις αναπαρίστανται ως διανύσματα με σαφείς προεκτάσεις στη σημασιολογική θεωρία. Στη συνέχεια θα αναλύσουμε τις τεχνικές σύμφωνα με τις οποίες διαπιστώνεται η ομοιότητα κειμένων, λέξεων και σχέσεων με διανυσματικές αναπαραστάσεις για να καταλήξουμε σε έναν ορισμό της ομοιότητας.

4.1.1 Ομοιότητα κειμένων: Πίνακας όρων - κειμένων

Δεδομένης μιας συλλογής κειμένων, όπου κάθε κείμενο της συλλογής αναπαρίσταται με διάνυσμα, μια τυπική απεικόνιση είναι αυτή στην οποία κατασκευάζεται πίνακας του οποίου οι γραμμές αντιστοιχούν στους όρους της συλλογής και οι στήλες στα κείμενά της. Οι όροι (τύποι) είναι συνήθως οι λέξεις της συλλογής και η επιλογή του τι θεωρείται όρος/τύπος εξαρτάται συνήθως από τη φύση της συλλογής και το πρόβλημα προς μοντελοποίηση. Για παράδειγμα, μπορεί να ενδιαφερόμαστε να εφαρμόσουμε λημματοποίηση ή να διαφοροποιήσουμε λέξεις με πεζούς και κεφαλαίους χαρακτήρες (βλ. ενότητα 4.2.2).

Ο πίνακας αυτός ονομάζεται Πίνακας Όρων-Κειμένων (term-document matrix). Τα στοχεία x_{ij} του πίνακα X χτίζονται με βάση την κωδικοποίηση σωρού λέξεων (bag of words). Σύμφωνα με αυτή την κωδικοποίηση, ένας σωρός (bag, multiset) είναι το ανάλογο της μαθηματικής έννοιας του συνόλου, με τη διαφορά ότι ένας σωρός μπορεί να έχει περισσότερα του ενός όμοια μέλη. Για παράδειγμα, ο σωρός {α, β, β, γ, δ, δ} περιλαμβάνει τα στοιχεία {α, β, γ, δ}. Τα στοιχεία ενός σωρού δεν είναι διατεταγμένα και η σειρά δεν έχει σημασία.

Η διανυσματική αναπαράσταση ενός σωρού γίνεται με ένα διάνυσμα στο οποίο η διάσταση i αντιστοιχεί στον διακριτό όρο i του σωρού. Για παράδειγμα, στον σωρό {α, β, β, γ, δ, δ} αντιστοιχεί το διάνυσμα $[1, 2, 1, 2]$, όπου στη θέση 1 του διανύσματος περιέχεται η συχνότητα εμφάνισης του όρου «α», στη θέση 2 η συχνότητα εμφάνισης του όρου «β» κ.ο.κ.

Κατ' επέκταση μπορεί λοιπόν ένα σύνολο σωρών να αναπαρασταθεί από έναν πίνακα , στο οποίο κάθε στήλη αντιστοιχεί σε ένα σωρό και κάθε γραμμή σε ένα μοναδικό στοιχείο του σωρού, και στην θέση x_{ij} περιέχεται η συχνότητα αυτού του i-οστού όρου στο j-οστό κείμενο.

Στην περίπτωση του πίνακα όρων-κειμένων, ένα κείμενο αναπαρίσταται με το διάνυσμα στήλη των όρων του κειμένου. Σύμφωνα με την κωδικοποίηση σωρού λέξεων (bag of words), η συνάφεια δύο κειμένων μπορεί να υπολογιστεί με βάση τη γεωμετρική απόσταση των διανυσματικών τους αναπαραστάσεων, δηλαδή κείμενα τα οποία έχουν ίδιες συχνότητες όρων (που δεν είναι απαραίτητα στην ίδια διάταξη), ή αλλιώς, κείμενα, που έχουν κοινό λεξιλόγιο, τείνουν να είναι όμοια. Συνεπώς, οι διανυσματικές αναπαραστάσεις κειμένων με βάση την κωδικοποίηση σωρού λέξεων καταγράφουν σε ικανοποιητικό βαθμό το θέμα στο

οποίο αναφέρεται ένα κείμενο.

Έστω ένας πίνακας όρων-κειμένων, που έχει κατασκευαστεί από συλλογή n κειμένων m στα οποία εμφανίζονται m διακριτοί όροι. Ο πίνακας θα έχει m γραμμές, που αντιστοιχούν στους όρους και n στήλες, που αντιστοιχούν στα κείμενα. Το στοιχείο x_{ij} του πίνακα που αντιστοιχεί στον όρο i του κειμένου j, θα περιέχει τη συχνότητα εμφάνισης αυτού του όρου στο συγκεκριμένο κείμενο.

Σε τυπικές εφαρμογές φυσικής γλώσσας, τα περισσότερα στοιχεία του διανύσματος του κειμένου j $x_{:,j}$ θα είναι μηδενικά, δεδομένου του γεγονότος ότι μόνο ένα μικρό υποσύνολο λέξεων θα χρησιμοποιείται σε κάθε κείμενο j κατά συνέπεια το διάνυσμα θα είναι «αραιό».

Το διάνυσμα της γραμμής i που αντιστοιχεί στον όρο i δίνει ένα αποτύπωμα της λέξης και αντίστοιχα το διάνυσμα της στήλης j, που αντιστοιχεί στο κείμενο j, δίνει ένα αποτύπωμα του κειμένου j. Είναι σημαντικό να παρατηρήσουμε εδώ ότι ενώ δεν υπάρχει συντακτική πληροφορία σε αυτή την αναπαράσταση καθώς επίσης και η σειρά των λέξεων δεν διατηρείται, η αναπαράσταση σωρού λέξεων δίνει πολύ καλά αποτελέσματα για την σημασιολογική συνάφεια των κειμένων. Κείμενα που τείνουν να αποτελούνται από ίδιες λέξεις θα τείνουν να έχουν σημασιολογική συνάφεια. Γενικεύοντας, αν δυο κείμενα αναφέρονται σε παρόμοιες θεματικές ενότητες ή θέματα, τα διανύσματά τους θα είναι παρόμοια. Αυτή η ιδέα μοντελοποιείται ευθέως στις προσεγγίσεις pLSA (Hofmann, 1999) και LDA (Blei et al., 2003).

4.1.2 Ομοιότητα λέξεων: Πίνακας περικειμένου λέξεων

Οι Salton et al. (1975), στο πλαίσιο της ανάκτησης πληροφοριών, επικεντρώνονται στην ποσοτικοποίηση και τη μέτρηση της κειμενικής ομοιότητας, θεωρώντας πως ένα ερώτημα αποτελεί ένα κείμενο και το ζητούμενο είναι να βρεθούν τα κείμενα στη συλλογή με τη μεγαλύτερη σημασιολογική συνάφεια ως προς το κείμενο του ερωτήματος. Η συνάφεια στη συνέχεια μετριέται ως διανυσματική ομοιότητα, όπου ομοιότητα σε αυτό το πλαίσιο σημαίνει (συνήθως) μικρή γεωμετρική απόσταση στον διανυσματικό χώρο.

Οι Deerwester et al. (1990) σημειώνουν ότι μετακινώντας το βάρος προσοχής από τις στήλες του πίνακα X στις γραμμές του αντί για την κειμενική ομοιότητα είναι εφικτό να ποσοτικοποιηθεί η ομοιότητα όρων (λέξεων).

Μια κομβική παρατήρηση σε αυτό το σημείο είναι ότι ο πίνακας όρων-κειμένων δεν είναι ο ιδανικός τρόπος για την μέτρηση ομοιότητας λέξεων. Μια πιο γενική λύση είναι ο Πίνακας Περικειμένου (co-occurence matrix), ο οποίος μπορεί να κατασκευαστεί ανά περίπτωση θεωρώντας ως περικείμενο λέξεις, φράσεις, προτάσεις, παραγράφους κ.ο.κ. Έχουμε ήδη αναφερθεί στην κατανεμητική υπόθεση (distributional hypothesis, Harris, 1954, σ. 158), σύμφωνα με την οποία λέξεις που εμφανίζονται στο ίδιο περικείμενο τείνουν να έχουν την ίδια σημασία.

Η υπόθεση αυτή παρέχει τη θεωρητική αιτιολόγηση της εφαρμογής μοντέλων διανυσματικών χώρων στη μέτρηση ομοιότητας λέξεων.

Υπάρχουν διάφοροι τρόποι για τη διανυσματική αναπαράσταση μιας λέξης, οι οποίοι προκύπτουν από το τι θεωρείται περικείμενο. Το περικείμενο μπορεί να πειρλαμβάνει το κειμενικό περικείμενο C λέξεων δεδομένης μιας λέξης w (Lund & Burgess, 1996), γραμματικές συσχετίσεις (Mitchell & Lapata, 2008), καθώς και πιο σύνθετες περιπτώσεις περικειμένων σε συνδέσμους εξάρτησης (dependency links) και θέσεις κατηγορημάτων (Erk & Padó, 2008).

4.1.3 Ομοιότητα σχέσεων: Πίνακας ζεύγους-δομικού σχήματος

Σε έναν Πίνακα Ζεύγους-Δομικού σχήματος, οι γραμμές αντιστοιχούν σε ζεύγη λέξεων, για παράδειγμα *ψηφοφόρος:κόμμα* ή *ξυλουργός:ξύλο* και οι στήλες στα κείμενα στα οποία εμφανίζονται αυτά τα ζεύγη: για παράδειγμα ο «Χ ψηφίζει το Υ κόμμα» και «Ο Χ δουλεύει με Υ». Οι Lin & Pantel (2001) εισήγαγαν αυτόν τον πίνακα για την μέτρηση της ομοιότητας μεταξύ τέτοιου τύπου δομικών σχημάτων, εκφραζόμενης ως ομοιότητας των στηλών του πίνακα X. Δεδομένου ενός δομικού σχήματος, όπως για παράδειγμα ”X solves Y”, ο αλγόριθμός τους έχει τη δυνατότητα να βρει παρόμοια δομικά σχήματα, όπως λόγου χάρη ”Y is solved by X”, ”Y is resolved in Y” και ”X resolves Y” (Lin & Pantel, 2001, σ. 11).

Οι ίδιοι εισάγουν την εκτεταμένη κατανεμητική υπόθεση ως εξής *Δομικά σχήματα που τείνουν να συνεμφανίζονται με παρόμοια ζεύγη τείνουν να έχουν παρόμοια σημασία. Τα δομικά σχήματα ”X solves Y” και ”Y is solved by Y” τείνουν να συνεμφανίζονται με παρόμοια ζεύγη X:Y, κάτι που υπονοεί ότι αυτά τα δομικά σχήματα έχουν παρόμοιες σημασιολογικές λειτουργίες.* (Lin & Pantel, 2001, σ. 8).

Η ομοιότητα αυτών των δομικών σχημάτων μπορεί να χρησιμοποιηθεί ως εργαλείο συμπερασματολογίας σε ερωτήματα τύπου «είναι η πρόταση Α παράφραση της πρότασης Β;» (Lin & Pantel, 2001, σ. 16).

4.1.4 Ορισμός της ομοιότητας

Οι πίνακες ζευγών-δομικών σχημάτων χρησιμοποιούνται για την ποσοτικοποίηση και μέτρηση της ομοιότητας σημασιολογικών σχέσεων (relational similarity). Αντίστοιχα, πίνακες όρων-περικειμένου χρησιμοποιούνται για τη μέτρηση ομοιότητας χαρακτηριστικών/χρήσης λέξεων (attributional similarity), έννοιες που περιγράφονται από την Gentner (1983, σσ. 11-13).

Η ομοιότητα χαρακτηριστικών/χρήσης, sim_a, μεταξύ δύο λέξεων α και β, $sim_a(\alpha, \beta) \in R$ εξαρτάται από τον βαθμό αντιστοιχίας μεταξύ των χαρακτηριστικών των α, β. Όσο μεγαλύτερος ο βαθμός αντιστοιχίας των

χαρακτηριστικών αυτών, τόσο μεγαλύτερη είναι η ομοιότητα.

Η ομοιότητα σημασιολογικών σχέσεων sim_r μεταξύ δύο ζευγών λέξεων α:β και γ:δ, $sim_r(\alpha : \beta, \gamma : \delta) \in R$, εξαρτάται από τον βαθμό αντιστοιχίας μεταξύ των σχέσεων των ζευγών α:β και γ:δ. Κατ' αναλογία με τα προηγούμενα, όσο μεγαλύτερος ο βαθμός αντιστοιχίας των σχέσεων αυτών τόσο μεγαλύτερη είναι η ομοιότητα. Στο παράδειγμα που δίνει ο Turney (2006, σ. 382), οι λέξεις *σκύλος* και *λύκος* έχουν σχετικά υψηλή ομοιότητα χαρακτηριστικών ενώ τα ζεύγη *σκύλος:γάβγισμα* και *γάτα:νιαούρισμα* έχουν σχετικά υψηλό βαθμό σχεσιακής σημασιολογικής ομοιότητας.

Στις επόμενες ενότητες θα αναφερθούμε αναλυτικά στις διαδικασίες που απαιτούνται για την κατασκευή πινάκων ομοιότητας για γλωσσικούς όρους.

4.2 Γλωσσική προεπεξεργασία

Η ενότητα αυτή επικεντρώνεται στην επεξεργασία ενός σώματος κειμένων που απαιτείται για να κατασκευαστούν πίνακες όρων-κειμένων, όρων-περικειμένου και ζευγών-δομικού σχήματος. Ανάλογα με τη φύση του σώματος και το υπό μελέτη πρόβλημα, διαφορετικοί τύποι επεξεργασίας είναι συνήθως απαραίτητοι. Οι τύποι προεπεξεργασίας μπορούν να διακριθούν σε τρεις κατηγορίες:

- διαχωρισμός όρων (tokenisation)
- κανονικοποίηση (normalisation)
- επισημείωση

4.2.1 Διαχωρισμός όρων

Ο διαχωρισμός όρων ενός κειμένου είναι η διαδικασία με την οποία το αρχικό ακατέργαστο κείμενο απεικονίζεται με διακριτούς όρους. Η διαδικασία αυτή είναι συνήθως αρκετά πιο περίπλοκη από τον διαχωρισμό ορίων όρων με βάση το κενό και εξαρτάται επίσης και από τη γλώσσα των κειμένων (Μαρκόπουλος, 2006, pp. 97 κ.εξ.). Ειδική μέριμνα συνήθως λαμβάνεται για σημεία στίξης, αρκτικόλεξα ή περιπτώσεις πολυλεκτικών συμπλόκων, π.χ. τοπωνύμια (*Άγιος Νικόλαος*, *San Fransisco*). Ανάλογα με την περίπτωση ή την εφαρμογή είναι θεμιτό να αφαιρεθούν οι υψίσυχνοι όροι και οι λειτουργικές λέξεις (stop words) του σώματος κειμένων, καθώς δεν μεταφέρουν σημασιολογικό περιεχόμενο και η μη χρήση τους δεν μεταβάλλει την ποιότητα του αποτελέσματος, μειώνοντας ταυτόχρονα την διαστασιμότητα και συνεπώς και την υπολογιστική πολυπλοκότητα.

4.2.2 Κανονικοποίηση

Η κανονικοποίηση είναι η επεξεργασία κατά την οποία διαφορετικές εκφάνσεις των ίδιων όρων ομαδοποιούνται σε έναν κοινό όρο, όπως η μετατροπή όλων των όρων/λέξεων σε πεζούς χαρακτήρες, για παράδειγμα *Γάτα*, *γάτα*. Στην κατηγορία της κανονικοποίησης ανήκει και η λημματοποίηση η οποία ορίζεται από τον Καρασίμο ως

«η διαδικασία της ομαδοποίησης όλων των κλιτών μορφών μιας λέξης τοποθετημένες κάτω από τη βασική μορφή της λέξης ή η ανάκτηση της βασικής μορφής της λέξης από τις κλιτές μορφές της, π.χ. η δυνατότητα ομαδοποίησης των κλιτών μορφών *άνθρωπος, ανθρώπου, άνθρωπο, άνθρωπε, άνθρωποι, ανθρώπων, ανθρώπους* κάτω από τη βασική μορφή *άνθρωπος* (Μαρκόπουλος, 2006, σσ. 111-113; Karasimos et al., 2015, σ. 45).

Μια άλλη πρακτική, που χρησιμοποιείται συχνά, είναι αυτή της μορφολογικής ανάλυσης των όρων και την εξαγωγή του θέματός τους (stem, stemming), συνήθως για την κανονικοποίηση τύπων πληθυντικού αριθμού ή παρελθοντικών χρόνων ρημάτων, ιδίως για γλώσσες όπως τα αγγλικά.

4.2.3 Επισημείωση

Η επισημείωση είναι κατά μία έννοια το αντίστροφο της κανονικοποίησης: όροι της συλλογής κειμένων επισημειώνονται/επαυξάνονται με συντακτικές ή σημασιολογικές πληροφορίες. Στην κατηγορία αυτή ανήκει η επισημείωση μερών του λόγου (Part of Speech Tagging, POS), κατά την οποία οι λέξεις του σώματος επισημειώνονται με τα μέρη του λόγου, όπως αυτά πραγματώνονται στα κείμενα, η σημασιολογική ανάλυση (word sense disambiguation) και η γραμματική ανάλυση (parsing) (Pantel & Lin, 2002; Μαρκόπουλος, 2006, σ. 109).

Σύμφωνα με τους Pantel & Lin et.al, η επισημείωση μπορεί να βελτιώσει τα αποτελέσματα μέτρησης της σημασιολογικής ομοιότητας λέξεων.

4.3 Μαθηματική επεξεργασία πινάκων

Αμέσως μετά τα βήματα της προεπεξεργασίας, που περιγράφηκαν στην προηγουμενη ενότητα, ακολουθεί η κατασκευή του πίνακα συχνοτήτων. Σε αυτόν τον πίνακα είναι επιθυμητό να εφαρμοστούν μετασχηματισμοί, όπως λ.χ. στάθμιση συχνοτήτων, καθώς υψίσυχνες λειτουργικές λέξεις δεν προσφέρουν σημασιολογική πληροφορία. Εφαρμόζεται επίσης ομαλοποίηση και αφαίρεση θορύβου και συμπύκνωση του αραιού συνήθως πίνακα σε πιο συμπαγή (dense) μορφή. Τέλος, υπάρχουν διάφοροι τρόποι για τη μέτρηση της ομοιότητας των

διανυσμάτων, που αντιστοιχούν, όπως είδαμε στην πρώτη ενότητα του κεφαλαίου, στις υπό εξέταση οντότητες.

Μια επισκόπηση των μεθόδων μαθηματικής επεξεργασίας πινάκων όρων-περικειμένου δίνεται από τον Lowe (2001). Η κατασκευή των πινάκων χωρίζεται σε τέσσερα διακριτά βήματα: υπολογισμός συχνοτήτων, μετασχηματισμός/στάθμιση των συχνοτήτων, ομαλοποίηση και μείωση διαστασιμότητας (smoothing, dimensionality reduction) και υπολογισμός ομοιότητας.

4.3.1 Μέτρηση

Σε κάθε θέση X_{ij} ενός πίνακα συχνοτήτων αντιστοιχεί ένα γεγονός: ένα συγκεκριμένο στοιχείο του σώματος (όρος, λέξη, ζεύγος) εμφανίστηκε σε μια συγκεκριμένη θέση (κείμενο, περικείμενο, δομικό σχήμα) με έναν συγκεκριμένο αριθμό εμφανίσεων. Η κατασκευή του πίνακα είναι συνεπώς θέμα μέτρησης τέτοιων γεγονότων. Η μέτρηση όμως τέτοιων γεγονότων σε μεγάλα σώματα κειμένων μπορεί να είναι προβληματική δεδομένων των περιορισμών μνήμης και της αλγοριθμικής πολυπλοκότητας. Σε αυτές τις περιπτώσεις και ανάλογα με τον όγκο του σώματος:

- Το σώμα κειμένων είναι μεγάλο αλλά μπορεί να γίνει επεξεργασία του από έναν μόνο υπολογιστή. Σε αυτή την περίπτωση η δημιουργία του πίνακα μπορεί να διακριθεί σε δύο διαφορετικές διακριτές εργασίες: αρχικά τη σάρωση του σώματος και τη μέτρηση των γεγονότων αυτών και σε δεύτερο χρόνο τη δημιουργία του πίνακα.

- Ο όγκος του σώματος κειμένων είναι τέτοιος που η μνήμη ενός υπολογιστή δεν είναι αρκετή. Σε αυτή την περίπτωση γίνεται χρήση κατανεμημένων συστημάτων αποθήκευσης και παράλληλης επεξεργασίας, όπως το Apache Hadoop (Hadoop, 2009) ή το Apache Spark (Spark, 2016) και η υπολογιστική τεχνική MapReduce (Dean & Ghemawat, 2008).

4.3.2 Μετασχηματισμός και στάθμιση

Η κεντρική ιδέα για τη στάθμιση των στοιχείων X_{ij} είναι να αποδοθεί μεγαλύτερο βάρος σε γεγονότα μικρής αναμενόμενης πιθανότητας εμφάνισης και μικρότερο βάρος σε γεγονότα με μεγάλη αναμενόμενη πιθανότητα εμφάνισης. Αυτό είναι επιθυμητό διότι γεγονότα με υψηλή αναμενόμενη πιθανότητα εμφάνισης (π.χ. λειτουργικές λέξεις) δεν παρέχουν ουσιώδη πληροφορία στο μοντέλο. Στην θεωρία πληροφοριών ένα μη αναμενόμενο γεγονός έχει περισσότερο πληροφοριακό περιεχόμενο από ένα αναμενόμενο γεγονός (Shannon, 1948).

Ο πιο δημοφιλής τρόπος στάθμισης είναι η οικογένεια τεχνικών Συχνότητας Όρου - Ανάστροφης Συχνότητας Κειμένου (Term Frequency - Inverse Document Frequency, Sparck Jones, 1972). Ένα στοιχείο, σύμφωνα με αυτή την τεχνική, παίρνει μεγαλύτερο βάρος όταν είναι συχνό σε ένα κείμενο, αλλά σπάνιο σε εμφανίσεις σε άλλα κείμενα του σώματος. Οι Salton & Buckley (1988) όρισαν έναν μεγάλο αριθμό συναρτήσεων tf-idf και διερεύνησαν την απόδοσή τους σε συστήματα ανάκτησης πληροφορίας, αποδεικνύοντας ότι η στάθμιση tf-idf επιφέρει σημαντικές βελτιώσεις στην απόδοση σε σχέση με τη χρήση απλών συχνοτήτων εμφάνισης. Η τεχνική tf-idf είναι η πλέον συνηθισμένη σε εφαρμογές μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας, καθώς επίσης και η τυπική τεχνική σε συστήματα ανάκτησης περιεχομένου.

Μια άλλη δημοφιλής τεχνική, που συχνά συνδυάζεται με το tf-idf, είναι η κανονικοποίηση μήκους (length normalisation) (Singhal et. al, 1996): οι διανυσματικές αναπαραστάσεις κειμένων διαιρούνται δια το μήκος των κειμένων, συνήθως χρησιμοποιώντας ως μέτρο μήκους την ευκλείδεια νόρμα $||x|| = \sqrt{\sum_i x_i})$, απεικονίζοντας έτσι τα διανύσματα στη μοναδιαία σφαίρα.

Τέλος, μια εναλλακτική τεχνική στάθμισης, που προέρχεται από τη θεωρία πληροφοριών, είναι η Σημειακή Αμοιβαία Πληροφορία (Pointwise Mutual Information, PMI, Church & Hanks, 1990; Turney, 2001), η οποία δίνει καλά αποτελέσματα τόσο σε πίνακες περικειμένου όσο και σε πίνακες όρων-κειμένων (Pantel & Lin, 2002, 2002).

Έστω πίνακας όρων-περικειμένου διαστάσεων mxn, με m γραμμές και n στήλες, με $w_j = X_{i,:}$ την i-οστή γραμμή και $c_j = X_{:,j}$ την j-οστή στήλη. Το διάνυσμα v_{w_j} αντιστοιχεί στην λέξη w_i του σώματος και το διάνυσμα v_{c_j} αντιστοιχεί στο περικείμενο c_j. Η θέση x_{ij} περιέχει τον αριθμό που η λέξη w_i εμφανίζεται στο περικείμενο w_j. Ο πίνακας X_{PMI} της Σημειακής Αμοιβαίας Πληροφορίας θα είναι πίνακας ίδιων διαστάσεων με τον X και κατασκευάζεται ως εξής:

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \tag{4.1}$$

$$p_{i*} = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \tag{4.2}$$

$$p_{*j} = \frac{\sum_{i=1}^{m} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \tag{4.3}$$

$$pmi_{ij} = log(\frac{p_{ij}}{p_{i*}p_{*j}}) \tag{4.4}$$

Μια παραλλαγή του PMI, που επίσης χρησιμοποιείται συχνά και τείνει να δίνει καλύτερα αποτελέσματα, είναι η Θετική Σημειακή Αμοιβαία Πληροφορία (Positive Pointwise Mutual Information, PPMI, Fano, 1961), στην οποία τα στοιχεία του πίνακα X, για τα οποία $x_{ij} < 0$ αντικαθίστανται με 0, δηλαδή:

$$PPMI = \begin{cases} pmi_{ij} & pmi_{ij} \geq 0 \\ 0 & \text{διαφορετικά} \end{cases} \tag{4.5}$$

Η αντικατάσταση των αρνητικών όρων με μηδέν, όπως γίνεται στο PPMI, έχει στην πράξη καλύτερα αποτελέσματα από διάφορες άλλες μεθόδους στάθμισης σε εφαρμογές σημασιολογικής ομοιότητας με πίνακες όρων-λέξεων (Bullinaria & Levy, 2007).

Στα παραπάνω, p_{ij} είναι η πιθανότητα εμφάνισης της λέξης w_i στο περικείμενο, w_j, p_{i*} είναι η πιθανότητα εμφάνισης της λέξης w_i και p_{*j} η πιθανότητα του περικειμένου c_j.

Αν τα w_i, c_j είναι στατιστικά ανεξάρτητα, τότε από τον ορισμό της στατιστικής ανεξαρτησίας προκύπτει ότι $p_{i*}p_{*j} = p_{ij}$, δηλαδή $\frac{p_{ij}}{p_{i*}p_{*j}} = 1$ και συνεπώς $pmi_{ij} = 0$ αφού $log(1) = 0$. Εάν η λέξη w_i και το περικείμενο c_j δεν είναι στατιστικά ανεξάρτητα, δηλαδή υπάρχει μία ιδιαίτερη σχέση μεταξύ τους, τότε η πιθανότητα p_{ij} θα είναι μεγαλύτερη από το γινόμενο $p_{i*}p_{*j}$ και σε αυτή την περίπτωση το pmi_{ij} θα είναι θετικό. Εάν η λέξη w_i δεν σχετίζεται με το περικείμενο c_j, το pmi_{ij} ενδέχεται να πάρει αρνητικές τιμές, που όπως είδαμε στην εξίσωση 4.5, αντικαθίστανται με μηδέν στο PPMI.

Ένα πρόβλημα του PMI είναι ότι ο υπολογισμός τείνει να ευνοεί χαμηλής συχνότητας συνεμφανίσεις. Πιο συγκεκριμένα, αν τα w_i, c_j είναι στατιστικά εξαρτημένα με μέγιστη συσχέτιση, τότε $p_{ij} = p_{i*} * p_{*j}$ και η εξίσωση 4.4 γίνεται $log(1/p_{i*})$ και το PMI αυξάνεται όσο η πιθανότητα εμφάνισης της λέξης w_i μικραίνει. Ένας τυπικός τρόπος για να αντιμετωπιστεί αυτό το πρόβλημα είναι να σταθμιστεί το pmi_{ij} σύμφωνα με την εξίσωση:

$$\delta_{ij} = \frac{x_{ij}}{x_{ij} + 1} \tag{4.6}$$

$$new\_pmi_{ij} = \delta_{ij} \cdot pmi_{ij} \tag{4.7}$$

(Pantel & Lin, 2002).

Μια άλλη μέθοδος χειρισμού γεγονότων χαμηλών συχνοτήτων είναι η εξομάλυνση Laplace, η οποία εφαρμόζεται στις εκτιμήσεις των πιθανοτήτων p_{ij}, p_{i*} και p_{*j}. Σύμφωνα με την τεχνική αυτή στις συχνότητες x_{ij} προστίθεται μια θετική σταθερά k και στις εξισώσεις 4.1,4.2, 4.3 το x_{ij} αντικαθίσταται από $x_{ij} + k$, όπου k θετικός πραγματικός αριθμός.

Μεγαλύτερες τιμές του k έχουν ως αποτέλεσμα ισχυρότερη εξομάλυνση. Η εφαρμογή της εξομάλυνσης Laplace έχει ως αποτέλεσμα τη μετακίνηση των τιμών pmi_{ij} προς το μηδέν, με το μέγεθος της μετακίνησης να εξαρτάται από την αρχική συχνότητα x_{ij}. Εάν η συχνότητα είναι μικρή, η μετακίνηση είναι μεγάλη και αντίστροφα αν η συχνότητα είναι μεγάλη, η μετακίνηση είναι μικρή. Συνεπώς αντιμετωπίζεται το πρόβλημα με εμφανίσεις χαμηλής συχνότητας (Turney & Littman, 2003).

4.4 Μείωση διαστασιμότητας

Η ενότητα αυτή επικεντρώνεται σε διαδικασίες εξομάλυνσης που αφορούν τη μείωση της διαστασιμότητας. Οι διαδικασίες αυτές είναι σημαντικές για δύο λόγους: πρώτον για την αφαίρεση θορύβου από το μοντέλο και δεύτερον για λόγους υπολογιστικής απόδοσης όσον αφορά τη σύγκριση διανυσμάτων

Η μείωση της διαστασιμότητας μπορεί να επιτευχθεί αρχικά περιορίζοντας τον αριθμό των στοιχείων του πίνακα στο μοντέλο, για παράδειγμα αφαιρώντας σπάνιες λέξεις και κρατώντας μόνο υψίσυχνες. Οι υψίσυχνες λειτουργικές λέξεις, όμως, όπως π.χ. άρθρα και σύνδεσμοι, δεν εμπεριέχουν σημασιολογική πληροφορία και συνεπώς η μέθοδος αυτή δεν είναι ιδανική. Η εφαρμογή μεθόδων στάθμισης, όπως το tf-idf, μπορεί να βοηθήσει στην επιλογή σημαντικών λέξεων από το σώμα κειμένων και συνεπώς στη μείωση του θορύβου και τη βελτίωση των αποτελεσμάτων του μοντέλου, καθώς επίσης και τη βελτίωση της υπολογιστικής απόδοσης.

Ο υπολογισμός των ομοιοτήτων μεταξύ όλων των διανυσμάτων σε έναν πίνακα είναι υπολογιστικά επίπονη διαδικασία (πολυπλοκότητα $\mathcal{O}(n^2)$ δεδομένων n διανυσμάτων). Στην πράξη η διαδικασία υπολογισμού των ομοιοτήτων μπορεί να επιταχυνθεί συγκρίνοντας διανύσματα τα οποία είναι υποψήφια να έχουν μεγάλη ομοιότητα και αποκλείοντας από τη σύγκριση διανύσματα τα οποία είναι γνωστό εκ των προτέρων ότι θα είναι ανόμοια. Για παράδειγμα, αν σε ένα από τα δυο διανύσματα μια από τις διαστάσεις είναι μηδενική, τότε τα διανύσματα αυτά θα είναι ανόμοια, καθώς δεν έχουν κοινό το χαρακτηριστικό που να αναλογεί στη συγκεκριμένη διάσταση (θέση) του διανύσματος. Υψίσυχνες λέξεις όπως το *και* τείνουν να εμφανίζονται σε όλα τα διανύσματα και συνεπώς δεν προσφέρουν σημασιολογική πληροφορία. Για παράδειγμα, στο σχήμα στάθμισης PMI διαστάσεις οι οποίες έχουν υψηλή στάθμιση αντιστοιχούν σε συνδυασμούς όρων-περικειμένου με υψηλό πληροφοριακό περιεχόμενο. Με βάση αυτή την παρατήρηση, ο Lin (1998), κρατώντας τις τιμές εκείνες του πίνακα X_{ij} που είναι μεγαλύτερες από ένα κατώφλι k, μειώνει δραστικά τον αριθμό συγκρίσεων, αυξάνοντας έτσι σημαντικά την υπολογιστική απόδοση, ενώ παράλληλα η απώλεια ακρίβειας του συστήματος είναι αμελητέα για τις 200 πιο

όμοιες λέξεις, για κάθε λέξη του σώματος κειμένων.

Κατά τη διαδικασία ομαλοποίησης υπολογίζεται ένας ανεστραμμένος κατάλογος μη μηδενικών συντεταγμένων. Με βάση αυτόν τον κατάλογο για τη σύγκριση δύο διανυσμάτων εξετάζονται μόνο εκείνα τα διανύσματα τα οποία δεν έχουν μη μηδενικές συντεταγμένες.

4.4.1 Λανθάνουσα Σημασιολογική Ανάλυση

Οι Deerwester et al. (1990) προτείνουν έναν μαθηματικό μετασχηματισμό από τη Γραμμική Άλγεβρα για τη βελτίωση των αποτελεσμάτων ομοιότητας σε πίνακες όρων-περικειμένου. Ο μετασχηματισμός αυτός είναι η Περικεκομμένη Παραγοντοποίηση Ιδιαζουσών Τιμών (Truncated Singular Value Decomposition, Truncated SVD). Οι Deerwester et al. (1990) εφαρμόζουν την τεχνική με στόχο την ομοιότητα κειμένων. Οι Landauer & Dumais (1997) επικεντρώνονται στην ομοιότητα λέξεων και την εφαρμόζουν σε ερωτήσεις πολλαπλής επιλογής σε τεστ της αγγλικής ως ξένης γλώσσας (TOEFL, Test of English as a Foreign Language), επιτυγχάνοντας αποτελέσματα συγκρίσιμα με ανθρώπους.

Η τεχνική SVD συνήθως απαντά με το όνομα Λανθάνουσα Σημασιολογική Δεικτοδότηση (Latent Semantic Indexing, LSI) όταν εφαρμόζεται για την μέτρηση ομοιότητας κειμένων, ενώ σε εφαρμογές ομοιότητας λέξεων/όρων συνήθως αναφέρεται ως Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis, LSA). Στη συνέχεια παρουσιάζεται η μαθηματική περιγραφή της τεχνικής και ερμηνεύεται η λειτουργία της υπό διαφορετικά οπτικά πρίσματα.

Η τεχνική SVD χρησιμοποιείται για την παραγοντοποίηση ενός πίνακα X σε γινόμενο 3 πινάκων, $U\Sigma V^T$, όπου U, V ορθοκανονικοί πίνακες ($U^TU = VV^T = I$) και Σ διαγώνιος πίνακας με τη διαγώνιο να περιέχει τις ιδιοτιμές του πίνακα X σε φθίνουσα σειρά.

Αν ο πίνακας X είναι τάξης r τότε και ο Σ είναι τάξης r. Έστω $\Sigma_k, k < r$ ο διαγώνιος πίνακας των k-πρώτων ιδιοτιμών του X και U_k, V_k, που προκύπτουν, αν ληφθούν οι πρώτες k στήλες από τους πίνακες U και V.

Τότε, το γινόμενο $\hat{X} = U_k\Sigma_k V_k^T$ είναι πίνακας τάξης r, ο οποίος ελαχιστοποιεί το σφάλμα προσέγγισης $||\hat{X} - X||_F$ για όλους τους πίνακες τάξης r, και $||...||_F$ η νόρμα Frobenious.

Διαφορετικά, η περικεκομμένη παραγοντοποίηση SVD ενός πίνακα αντιστοιχεί στην προσέγγιση του πίνακα X, η οποία ελαχιστοποιεί το σφάλμα προσέγγισης υπό την νόρμα Frobenious.

Η παραγοντοποίηση αυτή εφαρμοσμένη σε προβλήματα διανυσματικών χώρων έχει διάφορες ερμηνείες. Οι Deerwester et al. (1989), Landauer & Dumais (1997) περιγράφουν το SVD ως μέθοδο εύρεσης λανθάνουσας σημασιολογίας (latent meaning). Πιο συγκεκριμένα, το αποτέλεσμα της εφαρμογής της τεχνικής SVD σε έναν πίνακα X για κάποιον αριθμό k είναι μια απεικόνιση από τον

διανυσματικό χώρο που ορίζουν οι γραμμές και οι στήλες του X στον νέο μικρότερης διαστασιμότητας διανυσματικό χώρο, που παράγεται από το SVD. Αυτός ο διανυσματικός χώρος μπορεί να θεωρηθεί ότι αντικατοπτρίζει μια απεικόνιση από τις λέξεις (γραμμές του X) και το περικείμενο (στήλες του) στον νέο αυτό χώρο. Η απεικόνιση αυτή μπορεί να θεωρηθεί ότι «συμπιέζει» με τέτοιο τρόπο λέξεις και περικείμενο ώστε να συνυπάρχουν σε μικρότερο αριθμό διαστάσεων καταγράφοντας έτσι λανθάνοντα χαρακτηριστικά σημασιολογίας.

4.4.2 Μείωση θορύβου

Ο Rapp (2003) περιγράφει την εφαρμογή SVD σε πίνακες όρων-περικειμένου σαν τεχνική αφαίρεσης θορύβου. Υπό αυτή την προοπτική, ο πίνακας $\hat{X} = U_k \Sigma_k V_k^T$ προκύπτει από εξομάλυνση του αρχικού πίνακα X.

Ο πίνακας U_k απεικονίζει τον διανυσματικό χώρο που ορίζεται από τις γραμμές του X σε ένα μικρότερο k-διάστατο χώρο και ο πίνακας V_k απεικονίζει τον διανυσματικό χώρο που ορίζεται από τις στήλες του πίνακα X στον ίδιο διανυσματικό χώρο. Ο διαγώνιος πίνακας $\Sigma_k j$ εκφράζει τα βάρη της ποικιλότητας τιμών κάθε διάστασης στον νέο k-διάστατο χώρο με τιμές σ_{ii} και με τις ιδιοτιμές του πίνακα σε φθίνουσα σειρά.

Αν υποθέσουμε ότι ο πίνακας X είναι μείγμα πληροφορίας και θορύβου, τότε ο πίνακας $U_k \Sigma_k V_k^T$ θα περιέχει τις πρώτες k διαστάσεις που περιέχουν την μεγαλύτερη ποικιλότητα, ενώ οι υπόλοιπες διαστάσεις θα είναι αυτές με τον μεγαλύτερο θόρυβο.

4.5 Υπολογισμός ομοιότητας και σύγκριση διανυσμάτων

Η πιο δημοφιλής μέθοδος υπολογισμού ομοιότητας δύο διανυσμάτων είναι ο υπολογισμός του συνημιτόνου της γωνίας που σχηματίζουν. Εάν x, y είναι διανύσματα διαστάσεως n με στοιχεία $x = [x_1, x_2, ..., x_n]$ και $y = [y_1, y_2, ..y_n]$, τότε το συνημίτονο της γωνίας θ που σχηματίζουν τα δύο διανύσματα είναι

$$cos(x,y) = \frac{x}{||x||} \frac{y}{||y||} \tag{4.8}$$

Το συνημίτονο της γωνίας, που σχηματίζουν τα δύο διανύσματα, συνεπώς, είναι το εσωτερικό γινόμενο των κανονικοποιημένων κατά νόρμα L_2 διανυσμάτων αυτών.

Στην περίπτωση που τα διανύσματα αντιστοιχούν σε συχνότητες λέξεων, μια υψίσυχνη λέξη θα έχει μεγάλο διάνυσμα, ενώ μια σπάνια λέξη θα έχει μικρότερο διάνυσμα. Άρα, το μήκος των διανυσμάτων δεν είναι κατάλληλο

μέτρο ομοιότητας. Κάτι τέτοιο είναι η γωνία που σχηματίζουν τα διανύσματα αυτά. Διανύσματα που αντιστοιχούν σε σημασιολογικά συναφείς όρους/λέξεις ενδεχομένως να έχουν διαφορετικά μεγέθη, αλλά η γωνία που σχηματίζουν θα είναι πολύ μικρή.

Η συνάρτηση συνημιτόνου έχει πεδίο τιμών $[-1,+1]$, με τιμή -1 όταν τα διανύσματα δείχνουν σε αντίθετη κατεύθυνση, (η γωνία θ που σχηματίζουν είναι 180 μοίρες), +1 όταν τα διανύσματα δείχνουν προς την ίδια κατεύθυνση (η γωνία θ που σχηματίζουν είναι 0 μοίρες) και 0 όταν τα διανύσματα είναι κάθετα μεταξύ τους ($\theta = 90$).

Στα διανύσματα συχνοτήτων τα οποία δεν έχουν αρνητικά στοιχεία το συνημίτονο δεν μπορεί να είναι αρνητικό, αλλά με την εφαρμογή μεθόδων στάθμισης και εξομάλυνσης τα διανύσματα ενδέχεται να έχουν και αρνητικά στοιχεία.

Ένα μέτρο απόστασης, όπως η απόσταση συνημιτόνου, μπορεί εύκολα να μετατραπεί σε μέτρο ομοιότητας με αντιστροφή ή αφαίρεση:

$$sim(x,y) = 1 - dist(x,y) \tag{4.9}$$

$$sim(x,y) = 1/dist(x,y) \tag{4.10}$$

Ένα μεγάλο εύρος μέτρων ομοιότητας έχει προταθεί στη βιβλιογραφία τόσο στην ανάκτηση πληροφοριών όσο και στη λεξική σημασιολογία (Jones & Furnas, 1987; Lin, 1998; Dagan et al., 1999; Lee, 1999; Weeds et al., 2004). Η επικρατούσα αίσθηση στην ερευνητική κοινότητα είναι ότι οι διαφορές στην απόδοση από τη χρήση διαφορετικών μέτρων ομοιότητας είναι αμελητέες. Είναι, επίσης, συνήθης πρακτική κάποιας μορφής κανονικοποίηση πριν την εφαρμογή μέτρων ομοιότητας.

Δημοφιλή γεωμετρικά μέτρα μέτρησης απόστασης διανυσμάτων είναι η Ευκλείδεια απόσταση[1] και η απόσταση Manhattan[2]. Μέτρα, που προέρχονται από την Θεωρία Πληροφορίας, είναι η απόκλιση Kullback-Leibler, η απόσταση Hellinger[3] και η απόσταση Bhattacharya[4]. Άλλα δημοφιλή μέτρα είναι η απόσταση Dice[5] και η απόσταση Jaccard[6].

[1] Ευκλείδεια απόσταση $d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_j)^2}$

[2] Απόσταση Manhattan: $d_m(p,q) = ||p-q|| = \sum_{i=1}^{n}|p_i - q_i|$

[3] Απόσταση Hellinger: $D_H(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}(\sqrt{p},\sqrt{q})^2}$ με p, q διανύσματα πιθανοτήτων

[4] Απόσταση Bhattacharya: $D_B(p,q) = -ln(\sum_{x\in X}\sqrt{(p(x)q(x))}$

[5] Απόσταση Dice: $DSC(X,Y) = \frac{2|X\cap Y|}{|X\cup Y|}$

[6] Απόσταση Jaccard: $J(X,Y) = \frac{|X\cup Y|}{|X\cap Y|}$

Η Lee σημειώνει ότι στην περίπτωση ομοιότητας λέξεων μέτρα που αποδίδουν μεγαλύτερο βάρος σε κοινές συντεταγμένες και λιγότερο βάρος στη σημασία μη κοινών χαρακτηριστικών μεταξύ λέξεων τείνουν να δίνουν καλύτερα αποτελέσματα. Στα πειράματα της Lee φαίνεται ότι οι ομοιότητες Jaccard, Jensen-Shannon και L1 τείνουν να αποδίδουν καλύτερα (Lee, 1999, σ. 7).

Οι Weeds et al. (2004, σσ. 6-7) κατηγοριοποιούν τα μέτρα σε τρεις κατηγορίες:

- Μέτρα ευαίσθητα σε υψίσυχνες λέξεις (συνημίτονο),
- Μέτρα ευαίσθητα σε λέξεις χαμηλών συχνοτήτων
- Μέτρα ευαίσθητα σε όρους παρόμοιων συχνοτήτων.

Έστω μια λέξη w_j. Εάν το μέτρο που θα χρησιμοποιήσουμε για τη μέτρηση της ομοιότητας με άλλες λέξεις είναι ευαίσθητο σε υψίσυχνες λέξεις, λέξεις με υψηλή συχνότητα στο σώμα θα τείνουν να έχουν μεγαλύτερο βαθμό ομοιότητας σε σχέση με χαμηλόσυχνες λέξεις. Αντίθετα, αν χρησιμοποιηθεί μέτρο ευαίσθητο σε χαμηλόσυχνες λέξεις, θα υπάρχει τάση για μεροληψία προς χαμηλόσυχνες λέξεις, ενώ, τέλος, σε περιπτώσεις που το μέτρο τείνει να είναι ευαίσθητο σε λέξεις με ίδια συχνότητα, θα τείνουν να ευνοηθούν και να παρουσιαστούν ως περισσότερο όμοιες λέξεις με παρόμοιες συχνότητες, χωρίς απαραίτητα να έχουν ισχυρή σημασιολογική συνάφεια.

Η επιλογή του κατάλληλου μέτρου εξαρτάται από τη φύση του προβλήματος, τη φύση των δεδομένων και το πόσο αραιά (sparse) είναι τα προς σύγκριση διανύσματα, αλλά και από την κατανομή των ως προς σύγκριση στοιχείων και από τη μέθοδο εξομάλυνσης.

Η κατασκευή ενός μοντέλου διανυσματικών χώρων από ένα σώμα κειμένων σε συνδυασμό με κάποιο μέτρο ομοιότητας μπορεί να χρησιμοποιηθεί για προβλήματα ταξινόμησης (classification) ή συσταδοποίησης (clustering), για απλές εφαρμογές ανάκτησης πληροφορίας ή ταξινόμησης. Για παράδειγμα, μπορεί να χρησιμοποιηθεί για την απάντηση ερωτημάτων του τύπου «Να βρεθούν τα k πιο όμοια κείμενα δεδομένου του κειμένου q» ή ακόμα και σε μη παραμετρικά μοντέλα μηχανικής μάθησης, όπως μοντέλα γειτνίασης (kNN, k nearest neighbors), όπου π.χ. το είδος ενός άγνωστου κειμένου αποφασίζεται από την πλειοψηφία των k «γειτόνων» του στο διανυσματικό χώρο.

Τα παραγόμενα διανύσματα όμως μπορούν να χρησιμοποιηθούν ως είσοδος σε αλγόριθμους μηχανικής μάθησης χωρίς την άμεση χρήση κάποιου μέτρου ομοιότητας. Για παράδειγμα, σε εφαρμογές ανίχνευσης συναισθήματος μπορεί να δοθεί ως είσοδος σε αλγόριθμο μηχανικής (όπως Logistic Regression, Support Vector Machine, Neural Net) το διάνυσμα του κειμένου ως χαρακτηριστικό και ως ετικέτα η κλάση του κειμένου (θετικό ή αρνητικό).

4.6 Κατανεμητική Σημασιολογία και μοντέλα λεξικών ενθέσεων

Στην ενότητα αυτή παρουσιάζονται τα μοντέλα που εκπαιδεύτηκαν στην παρούσα διατριβή για την ανίχνευση της ιδιολέκτου, με έμφαση στο μοντέλο word2vec και αναφορά στο μοντέλο fastText (4.6.4), το οποίο μπορεί να θεωρηθεί ως βελτίωση και επέκταση του word2vec, το μοντέλο GloVe (4.6.5), που ακολουθεί διαφορετική μεθοδολογία αλλά έχει ευθέως συγκρίσιμα αποτελέσματα σε σχέση με τα προηγούμενα και, τέλος, το μοντέλο doc2vec, που αποτελεί φυσική επέκταση του word2vec σε παραγράφους ή ολόκληρα κείμενα. Προηγείται η θεωρητική παρουσίαση της κατανεμητικής σημασιολογίας και της έννοιας των λεξικών ενθέσεων, στα οποία βασίστηκε η ανάπτυξη των μοντέλων αυτών.

4.6.1 Κατανεμητική σημασιολογία

Όπως επισημαίνει η Φλώρου (2017, σ. 32), στην οποία βασίζεται η συζήτηση που ακολουθεί, η ιστορία της κατανεμητικής υπόθεσης ξεκινά έξω από το πεδίο της σημασιολογίας και ανάγεται στις ευρετικές διαδικασίες γλωσσολογικής ανάλυσης του αμερικανικού δομισμού. Σύμφωνα με τον Harris (1954, σ. 7), τόσο στις φωνολογικές όσο και στις μορφολογικές αναλύσεις ο γλωσσολόγος πρώτα αντιμετωπίζει το πρόβλημα προσδιορισμού των σχετικών στοιχείων. Για να είναι τα στοιχεία σχετικά πρέπει να αναλυθούν σε μια κατανεμητική βάση: τα χαρακτηριστικά χ και ψ περιλαμβάνονται στο στοιχείο Α εάν η κατανομή του χ, που σχετίζεται με άλλα στοιχεία Β και Γ, είναι κατά κάποιο τρόπο ίδια με την κατανομή του ψ. Από τα ανωτέρω προκύπτει ότι τα στοιχεία Β και Γ αναγνωρίζονται από τη στιγμή που προσδιορίζεται το στοιχείο Α. Έτσι, τα στοιχεία προσδιορίζονται αμοιβαία και επί των μεταξύ τους κατανεμητικών σχέσεων.

Ο Harris ισχυρίζεται ότι η ομοιότητα στις κατανομές μεταξύ όρων είναι σε θέση να αξιοποιηθεί ως ερμηνεία της ίδιας της σημασίας των όρων και έτσι είναι δυνατόν να σχηματίζονται παραδειγματικές τάξεις σημασιολογικά όμοιων κατανεμητικών γλωσσικών εκφράσεων. Η ιδέα της κατανεμητικής ανάλυσης των γλωσσικών περικειμένων αποτελεί το κλειδί για την κατανόηση της σημασίας μιας λέξης στην παραδοσιακή γλωσσολογία του Firth, όπως συνοψίζεται στη φράση του «γνωρίζουμε μια λέξη από τη συντροφιά της» (1957, σ. 11) και απέκτησε την πλήρη της δυναμική στο πλαίσιο της γλωσσολογίας σωμάτων κειμένων (Γούτσος & Φραγκάκη, 2015) . Αντίθετα, αρκετοί εκπρόσωποι της γνωσιακής προσέγγισης (Lewis, 1976; Langacker, 2017) διαφωνούν με την κατανεμητική προσέγγιση της σημασίας, καθώς θεωρούν ότι η σημασία ενός όρου δεν είναι δυνατόν να αποδοθεί μέσα από εσωτερικές κατανομές των λέξεων, αλλά πρέπει να αγκιστρωθεί σε

εξωγλωσσικές οντότητες, οι οποίες σχετίζονται είτε με αντικείμενα του αισθητού κόσμου είτε με εννοιακές αναπαραστάσεις στο μυαλό του ομιλητή.

Σε αντίθεση με τις παραδεδομένες σημασιολογικές προσεγγίσεις της αναλυτικής φιλοσοφίας, οι κατανεμημένες αναπαραστάσεις δεν βασίζονται σε αφηρημένες λογικές μορφές (παραδείγματος χάριν το «αυτοκίνητο» ως κατηγόρημα για την εκπροσώπηση της έννοιας του αυτοκινήτου). Αντίθετα, στο πλαίσιο της κατανεμητικής σημασιολογίας, η εσωτερική δομή της σημασίας μιας λέξης εξάγεται όχι από τη μεμονωμένη λέξη, αλλά από τις λέξεις που απαρτίζουν το άμεσο περικείμενό της. Γι' αυτόν ακριβώς τον λόγο, οι διανυσματικές απεικονίσεις των λέξεων προέρχονται από έναν κατανεμητικό αλγόριθμο που επεξεργάζεται τα δεδομένα του σώματος κειμένων. Αυτές οι διανυσματικές απεικονίσεις μπορούν να δώσουν μια γεωμετρική ερμηνεία της σημασιολογικής εγγύτητας των όρων (Widdows & Widdows, 2004, σ. 111), καθώς οι λέξεις (τόσο οι λέξεις-στόχοι όσο και οι λέξεις που απαρτίζουν το περικείμενο) αποδίδονται ως σημεία σε ένα γεωμετρικό χώρο με τρόπο που να μπορεί να υπολογιστεί αριθμητικά η εγγύτητά τους.

Η κατανεμητική υπόθεση (distributional hypothesis, Harris, 1954) έχει ως πηγή αυτή τη σημασιολογική θεωρία και εμφανίζεται στην βιβλιογραφία με πληθώρα ορισμών, όπως:

«λέξεις που είναι παρόμοιες σε σημασία τείνουν να εμφανίζονται σε παρόμοια περικείμενα» (Rubenstein & Goodenough, 1965, σ. 627)

«λέξεις που χρησιμοποιούνται και εμφανίζονται σε ίδια περικείμενα, τείνουν να έχουν την ίδια σημασία» (Pantel, 2005, σ. 126)

«Λέξεις με παρόμοια σημασία θα εμφανιστούν σε παρόμοια περιβάλλοντα με την προϋπόθεσή ότι υπάρχει αρκετός κειμενικός όγκος» (Schütze & Pedersen, 1995, σ. 10), καθώς και

«μια αναπαράσταση που ερμηνεύει πώς χρησιμοποιούνται λέξεις σε φυσικό περικείμενο θα ερμηνεύσει το τι εννοούμε ως σημασιολογία» (Landauer & Dumais, 1997, σ. 218)

Στη συνέχεια της διατριβής επικεντρωνόμαστε στους δύο τελευταίους ορισμούς και αναπτύσσουμε μια κατανεμημένη σημασιολογική θεωρία για την ιδιόλεκτο, καθώς τα υπό εξέταση μοντέλα έχουν ως κοινό χαρακτηριστικό τη συνεμφάνιση όρων σε προκαθορισμένο μήκος περικειμένου, εκ των οποίων προκύπτουν στη συνέχεια οι λεξικές ενθέσεις.

Η έννοια των λεξικών ενθέσεων ουσιαστικά καλύπτει τις αδυναμίες της διανυσματικής αναπαράστασης που στηρίζεται στην απεικόνιση σωρού λέξεων

Πίνακας 4.1: Παράδειγμα κωδικοποίησης 1-hot

the	[1,0,0,0]
quick	[0,1,0,1]
brown	[0,0,1,0]
fox	[0,0,0,1]

(bag of words). Σε αυτήν ένα κείμενο αναπαρίσταται ως διάνυσμα, του οποίου η θέση i αντιστοιχεί στην i-οστή λέξη του λεξιλογίου V. Η αναπαράσταση μπορεί να είναι δυαδική (1 αν υπάρχει η λέξη στην πρόταση προς μοντελοποίηση, 0 αν δεν υπάρχει), να λαμβάνει υπόψη της τη συχνότητα εμφάνισης (στην θέση i καταγράφεται το πλήθος εμφανίσεων της λέξης w_i στην ιοστή θέση του λεξιλογίου) ή κάποια κανονικοποίηση, π.χ. το σχήμα TF-IDF[7]. Η θεώρηση αυτή, αν και υπεραπλουστεύμενη, καθώς δεν αντανακλά τη σειρά των λέξεων σε ένα κείμενο, λειτούργησε εξαιρετικά καλά σε μεγάλη γκάμα εφαρμογών επεξεργασίας φυσικής γλώσσας και συνεχίζει να ανήκει ακόμα στις βασικές αναπαραστάσεις, που χρησιμοποιούνται σε σύγχρονα συστήματα ανίχνευσης συγγραφέα (Stamatatos, 2009; Mikros & Perifanos, 2013).

Μια άλλη δημοφιλής αναπαράσταση είναι η one-hot, όπου η λέξη w_i, που αντιστοιχεί στη θέση i του λεξικού, αναπαρίσταται με ένα $|V|$-διάστατο διάνυσμα, του οποίου όλα τα στοιχεία είναι μηδενικά, εκτός από τη θέση i που είναι 1. Η αναπαράσταση αυτή είναι συνηθισμένη σε εφαρμογές νευρωνικών δικτύων και έχει το προτέρημα ότι μπορεί να χρησιμοποιηθεί με τέτοιον τρόπο ώστε να αντικατοπτρίζει τη σειρά των λέξεων σε περικείμενο n λέξεων. Αυτό επιτυγχάνεται συνενώνοντας τα διανύσματα των n λέξεων του περικειμένου σε ένα διάνυσμα διάστασης $|V| * n$.

Ένα παράδειγμα πιθανής αναπαράστασης one hot για τις λέξεις *the quick brown fox* σε διανύσματα one-hot 4 θέσεων παρουσιάζεται στον πίνακα 4.1.

Ανάλογα, στο προηγούμενο παράδειγμα, αν θεωρήσουμε την απεικόνιση 4.2 κωδικοποίηση σωρού λέξεων της φράσης *the quick* θα ήταν το διάνυσμα

$$v = [1,1,0,0]$$

ενώ για τις λέξεις *the fox* το διάνυσμα

$$v = [1,0,0,1]$$

Η εφαρμογή της κατανεμητικής θεωρίας με τη χρήση νευρωνικών δικτύων και λεξικών ενθέσεων υπερβαίνει τις προσεγγίσεις αυτές καθώς μπορεί να αναφέρεται και σε πλήρη σημασιολογικά στοιχεία και να καλύπτει το σύνολο του λεξιλογίου ενός σώματος κειμένων.

[7]https://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

Πίνακας 4.2: Απεικόνιση bag-of-words

the	1
quick	2
brown	3
fox	4

4.6.2 Λεξικές ενθέσεις

Από τη δεκαετία του '90 και μετέπειτα τα μοντέλα διανυσματικής απεικόνισης κυριαρχούσαν στην περιοχή της κατανεμητικής σημασιολογίας. Στο διάστημα αυτό αναπτύχθηκε πληθώρα μοντέλων για τον υπολογισμό και την εκμάθηση συνεχών αναπαραστάσεων λέξεων. Τα πιο διαδεδομένα και γνωστά μοντέλα είναι η Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis,LSA, Latent Semantic Indexing, LSI, Deerwester et al., 1989), η Πιθανοθεωρητική Λανθάνουσα Σημασιολογική Ανάλυση, (Probabilistic Latent Semantic Analysis, pLSA,Hofmann, 1999), και η Λανθάνουσα Κατανομή Dirichlet, (Blei et al., 2003).

Τα μοντέλα αυτά αποτελούν τον πυρήνα της κατανεμητικής σημασιολογίας και βασίζονται στην κατανεμητική ιδιότητα της γλώσσας, σύμφωνα με την οποία παρόμοιες λέξεις τείνουν να εμφανίζονται σε παρόμοια περικείμενα. Η μετεξέλιξη αυτών των τεχνικών έγκειται στο συνδυασμό και την ανάπτυξη υπολογιστικών και μαθηματικών μεθόδων, οι οποίες εκμεταλλεύονται αυτή την ιδιότητα, και σε συνδυασμό με την πιθανοτική προσέγγιση κατασκευής γενικευμένων γλωσσικών μοντέλων, οδήγησε στις λεξικές ενθέσεις. Οι λεξικές ενθέσεις, που προκύπτουν με μεθόδους μη επιβλεπόμενης μηχανικής μάθησης (νευρωνικά δίκτυα, word2vec, GloVe), σημειώνουν τεράστια επιτυχία σχεδόν σε όλο το φάσμα των εφαρμογών της επεξεργασίας φυσικής γλώσσας σε τέτοιο βαθμό που πλέον θεωρούνται πανάκεια. Η αποτελεσματικότητά τους είναι τόσο μεγάλη, που σε παραδοσιακές αρχιτεκτονικές επεξεργασίας φυσικής γλώσσας έχουν αντικαταστήσει παραδοσιακά κατανεμητικά μοντέλα, όπως π.χ. τα παραγόμενα χαρακτηριστικά από Brown Clusters[8] ή την LSA[9].

Οι Λεξικές ενθέσεις (word embeddings) είναι ο όρος που έχει επικρατήσει στη διεθνή βιβλιογραφία για την περιγραφή της αναπαράστασης λέξεων ως διανύσματα σε σχετικά χαμηλών διαστάσεων διανυσματικούς χώρους. Ισοδύναμοι όροι μπορούν να θεωρηθούν οι όροι «διανύσματα λέξεων» (word vectors) και «κατανεμημένες αναπαραστάσεις» (distributed representations).

Τα μοντέλα λεξικών ενθέσεων, όπως θα δούμε και στη συνέχεια στη λεπτομερή περιγραφή του μοντέλου word2vec, είναι άμεσα συνδεδεμένα με τα γλωσσικά μοντέλα και αποτελούν κατά μια έννοια υποπροϊόν της γενίκευσης

[8]Brown et al., 1992.

[9]Deerwester et al., 1990.

της προσέγγισης (approximation) γλωσσικών μοντέλων. Οι σημασιολογικές συσχετίσεις, που προκύπτουν μεταξύ των λεξικών ενθέσεων με απλές αλγεβρικές πράξεις όπως πρόσθεση και αφαίρεση, είναι εντυπωσιακές. Ειδικά το περίφημο παράδειγμα των Mikolov et al. (2013):

$$king - man + woman \approx queen$$

δεσπόζει σε κάθε παρουσίαση ή δημοσίευση σχετική με επεξεργασία φυσικής γλώσσας και Βαθιών αρχιτεκτονικών. Το παράδειγμα αυτό είναι εξαιρετικά σημαντικό καθώς δείχνει ότι το μοντέλο είναι ικανό να μάθει, χωρίς επίβλεψη, γραμμικές σημασιολογικές συσχετίσεις του τύπου «Η λέξη Βασιλιάς είναι για τη λέξη Βασίλισσα ότι η λέξη Άντρας για τη λέξη Γυναίκα», δηλαδή σημασιολογικές αναλογίες (word analogies) [10] σε επίπεδο λέξεων. Τα αποτελέσματα αυτά ισχύουν για οποιαδήποτε γλώσσα, δεδομένου ικανού όγκου κειμένων και επιβεβαιώνονται για τα ελληνικά με εκπαίδευση αντίστοιχων μοντέλων στο Σώμα Ελληνικών Κειμένων (ΣΕΚ, Goutsos, 2010) [11].

Στο σημείο αυτό θα προχωρήσουμε σε εκτεταμένη επισκόπηση των λεξικών ενθέσεων, καθώς η εφαρμογή τους σε εφαρμογές Επεξεργασίας Φυσικής Γλώσσας θεωρείται τεχνική αιχμής. Ο όρος *word embeddings* προτάθηκε από τους Bengio et al. (2003) ως το αποτέλεσμα εκπαίδευσης ενός νευρωνικού δικτύου, που είχε στόχο την εκμάθηση γλωσσικού μοντέλου [Βλ. Παράρτημα Β]. Στόχος δηλαδή του νευρωνικού δικτύου είναι η εκμάθηση μιας μαθηματικής συνάρτησης (το νευρωνικό δίκτυο), η οποία να είναι σε θέση να υπολογίζει την πιθανότητα μια λέξη w να ακολουθεί n-1 λέξεις w_1, w_2, ..., w_{n-1}. Αυτή η ριζοσπαστική ιδέα είχε ως κίνητρο την άρση του περιορισμού των παραδοσιακών μοντέλων φυσικής γλώσσας, που είναι βασισμένα στο στατιστικό υπολογισμό λεξικών συνεμφανίσεων. Αν στο σώμα κειμένων, στο οποίο εκπαιδεύεται το γλωσσικό μοντέλο, δεν υπάρχει η ακολουθία $w1$, $w2$, η πιθανότητα $P(w_2|w_1)$ συνεπώς είναι μηδέν, το οποίο προφανώς δεν ευσταθεί για καμία φυσική γλώσσα. Το πρόβλημα αυτό επιλύεται στην πράξη με κάποιου τύπου ομαλοποίηση (smoothing), συνήθως με κάποιας μορφής γραμμική παρεμβολή ή ανάλογες τεχνικές. Η εισαγωγή νευρωνικών δικτύων είχε ως σκοπό την αντιμετώπιση αυτής ακριβώς της αδυναμίας και αντί της ομαλοποίησης στόχος ήταν το νευρωνικό δίκτυο να μπορεί να γενικεύει σε περιπτώσεις λέξεων για τις οποίες δεν είχε εκπαιδευτεί ρητά.

Οι λεξικές ενθέσεις είναι υποπροϊόν της διαδικασίας εκπαίδευσης και, στην περίπτωση του word2vec, αποτελούν αναπαραστάσεις των διανυσμάτων των λέξεων, κωδικοποίησης 1-hot, σε πυκνά διανύσματα πολύ μικρότερης

[10] https://aclweb.org/aclwiki/Analogy_(State_of_the_art)

[11] Παραδείγματα ενθέσεων και σημασιολογικής ομοιότητας από την εκπαίδευση στο ΣΕΚ μπορούν να βρεθούν στο παράρτημα ΣΕΚ

διαστασιμότητας. Τα διανύσματα κωδικοποίησης 1-hot είναι εξαιρετικά μεγάλης διαστασιμότητας, ίσης με τον αριθμό των λέξεων του λεξιλογίου. Ο αριθμός αυτός σε μεγάλα σώματα κειμένων μπορεί να ξεπεράσει τις εκατοντάδες χιλιάδες λέξεις. Αντίθετα, τα διανύσματα λεξικών ενθέσεων, όπως θα δούμε, τυπικά έχουν 200-500 διαστάσεις. Ο αριθμός των διαστάσεων συνήθως καθορίζεται από τη συμπεριφορά και την απόδοση των ενθέσεων στις εφαρμογές, που χρησιμοποιούνται, μπορούν δηλαδή να θεωρηθούν ως υπερπαράμετροι (hyper-parameters) του γενικότερου μοντέλου ή εφαρμογής, του οποίου αποτελούν τμήμα οι ενθέσεις. Ως υπερπαραμέτρους ορίζουμε εκείνες τις παραμέτρους ενός μοντέλου που καθορίζουν τη συμπεριφορά του χωρίς να αποτελούν τμήμα της διαδικασίας εκπαίδευσης, δηλαδή οι τιμές τους δε προκύπτουν κατά την διαδικασία εκπαίδευσης ενός μοντέλου. Η ρύθμιση των υπαραμέτρων γίνεται συνήθως με επαναληπτικές διαδικασίες εκπαίδευσης και ελέγχοντας το τελικό αποτέλεσμα της διαδικασίας. Για παράδειγμα, οι υπερπαράμετροι που ελέγχουν την ακρίβεια σε ένα μοντέλο ανίχνευσης συναισθήματος με νευρωνικά δίκτυα αποτελούν τον αριθμό των κρυφών επιπέδων και τις διαστάσεις των επιπέδων αυτών.

Τα μοντέλα που χρησιμοποιούν λεξικές ενθέσεις έχουν την ιδιότητα να απεικονίζουν λέξεις σε διανύσματα συνδυάζοντας την κατανεμητική υπόθεση και τη μοντελοποίηση της φυσικής γλώσσας κατά τη διαδικασία εκμάθησης ενός νευρωνικού δικτύου. Με αφορμή αυτή την παρατήρηση, τα μοντέλα που προκύπτουν από την εκπαίδευση νευρωνικών δικτύων και εκμεταλλεύονται την κατανεμητική φύση της γλώσσας μπορούν να χαρακτηριστούν ως *νευροκατανεμητικά* (neuro-distributional models). Με αυτή την έννοια μπορούμε να κάνουμε λόγο για Νευροκατανεμητική Σημασιολογία ως εφαρμογή της κατανεμητικής θεωρίας με τη χρήση νευρωνικών δικτύων και λεξικών ενθέσεων/αναπαραστάσεων.

Οι λεξικές ενθέσεις, όπως είδαμε, είναι αναπαραστάσεις λέξεων ως d-διάστατα διανύσματα. Φυσικό επακόλουθο είναι, λοιπόν, να χρησιμοποιηθούν τα διανύσματα αυτά ως αναπαραστάσεις σε εφαρμογές επεξεργασίας φυσικής γλώσσας. Τα πρώτα παραδείγματα εφαρμογών, που βασίζονται σε λεξικές ενθέσεις, δίνονται από τους Collobert et al. (2011). Οι Collobert & Weston αναδεικνύουν τη σημασία των ενθέσεων ως βασικού συστατικού αρχιτεκτονικών επεξεργασίας φυσικής γλώσσας και περιγράφουν την αρχιτεκτονική ενός νευρωνικού δικτύου, που θέτει τα θεμέλια για πολλές σύγχρονες προσεγγίσεις.

Η έκρηξη, όμως, έρχεται το 2013 με τη δημοσίευση του εργαλείου word2vec (Mikolov et al., 2013), με το οποίο έγινε εφικτή η ταχεία εκμάθηση μοντέλων λεξικών ενθέσεων σε σώματα κειμένων εκατοντάδων εκατομμυρίων λέξεων. Η ευελιξία εκμάθησης, καθώς και τα παραδείγματα που συνόδευαν τη συγκεκριμένη δημοσίευση, δημιούργησαν τεράστιο αντίκτυπο στην ερευνητική κοινότητα. Στη συνέχεια, οι Pennington et al. (2014) παρουσίασαν το GloVe επιβεβαιώνοντας

ουσιαστικά ότι οι λεξικές ενθέσεις θεωρούνται πλέον τεχνολογία αιχμής.

Οι λεξικές ενθέσεις αποτελούν ένα εξαιρετικά επιτυχημένο αποτέλεσμα τεχνικών μη επιβλεπόμενης μηχανικής μάθησης. Το βασικό πρακτικό όφελος, πέρα από την απόδοσή τους σε εφαρμογές, είναι ότι δεν χρειάζεται επισημείωση σωμάτων κειμένων. Αντιθέτως, μπορούν να παραχθούν από μεγάλου όγκου σώματα κειμένων χωρίς καμία επισημείωση. Το σημείο αυτό είναι θεμελιώδους σημασίας όσον αφορά την χρήση προ-εκπαιδευμένων ενθέσεων ως εισόδου με τη μορφή χαρακτηριστικών (features) σε άλλα μοντέλα επεξεργασίας φυσικής γλώσσας, όχι μόνο όσον αφορά την αποτελεσματικότητά τους συγκριτικά με παραδοσιακές τεχνικές, αλλά και ως προς τον χρόνο και το κόστος που αποσοβείται από την επίπονη επισημείωση.

4.6.3 Word2vec

Η συνήθης πρακτική όσον αφορά αρχιτεκτονικές νευρωνικών δικτύων για τις λεξικές ενθέσεις είναι η χρήση δικτύων πρόσθιας τροφοδότησης (feed forward neural networks), τα οποία παίρνουν ως είσοδο λέξεις από ένα λεξικό και στα οποία η εκπαίδευση γίνεται με τη μέθοδο της οπίσθιας διάδοσης (back propagation). Οι λέξεις απεικονίζονται σε μια μικρής (συγκριτικά με τον αριθμό διακριτών λέξεων του λεξιλογίου) διάστασης και το επίπεδο αυτό αναφέρεται ως Επίπεδο ένθεσης (Embedding Layer).

Οι κύριες διαφορές των παραδοσιακών μοντέλων λεξικών ενθέσεων (Bengio et al., 2003) με νέες σύγχρονες προσεγγίσεις όπως το word2vec εντοπίζονται στα εξής σημεία. Στις παραδοσιακές προσεγγίσεις οι λεξικές ενθέσεις παράγονται ως δευτερεύον αποτέλεσμα της διαδικασίας εκμάθησης του δικτύου, ενώ αντίθετα στα νευροκατανεμητικά μοντέλα, όπως το word2vec και το fastText, οι ενθέσεις είναι συνήθως ο στόχος της διαδικασίας εκπαίδευσης. (Στην περίπτωση του GloVe ακολουθείται διαφορετική στρατηγική, καθώς αποτελεί μοντέλο μη επιβλεπόμενης μάθησης και το ζητούμενο είναι η παραγοντοποίηση του πίνακα συνεμφανίσεων). Η πλέον σημαντική όμως και ίσως κρίσιμη διαφορά είναι η σχέση χρόνου εκπαίδευσης σε σχέση με την ποιότητα των παραγόμενων ενθέσεων. Ο ιδιαίτερα μεγάλος χρόνος εκπαίδευσης ήταν ο κύριος λόγος που συναφή μοντέλα ενθέσεων δεν ήταν δημοφιλή, τουλάχιστον έως την εμφάνιση του word2vec, καθώς η εκπαίδευσή τους αποτελούσε εξαιρετικά επίπονη και αργή διαδικασία.

Με το word2vec έγινε εφικτή η εκπαίδευση μοντέλων σε σώματα κειμένων εκατοντάδων εκατομμυρίων λέξεων με χρόνο εκτέλεσης τυπικά μικρότερο των 60 λεπτών, με τη χρήση υπολογιστών μέσης ή και μικρής υπολογιστικής ισχύος και φυσικά, αυτό, σε συνδυασμό με την εξαιρετική βελτίωση που παρατηρήθηκε σε σχεδόν όλα τα μοντέλα επεξεργασίας φυσικής γλώσσας, επέβαλλαν τα μοντέλα λεξικών ενθέσεων ως το βασικό πλέον εργαλείο εκμάθησης λεξικών

αναπαραστάσεων. Πλέον μοντέλα λεξικών ενθέσεων μπορούν να παραχθούν σε σώματα κειμένων εκατοντάδων εκατομμυρίων λέξεων σε λίγες ώρες. Γα παράδειγμα, στο πλαίσιο των πειραμάτων της παρούσας διατριβής οι τυπικοί χρόνοι εκπαίδευσης μοντέλων κυμαίνονται μεταξύ 45 λεπτών (word2vec, GloVe) και 3-4 ωρών (fastText, doc2vec).

Μια άλλη διαφορά θα μπορούσε να θεωρηθεί ότι είναι η ποιοτική διαφοροποίηση των παραγόμενων ενθέσεων. Στο word2vec και το GloVe η έμφαση δίνεται σε παραγωγή ενθέσεων που κωδικοποιούν σημασιολογικές ιδιότητες και σχέσεις. Αυτή η κωδικοποίηση αποδεικνύεται εξαιρετικά σημαντική στην επίλυση πληθώρας προβλημάτων επεξεργασίας φυσικής γλώσσας, στα οποία η κωδικοποίηση σημασιολογικών σχέσεων αποδεικνύεται κεντρικής σημασίας. Αντίθετα, οι λεξικές ενθέσεις που παράγονται από παραδοσιακά μοντέλα νευρωνικών δικτύων τείνουν να είναι χρήσιμες μόνο στο πλαίσιο της εφαρμογής που εκπαιδεύτηκαν.

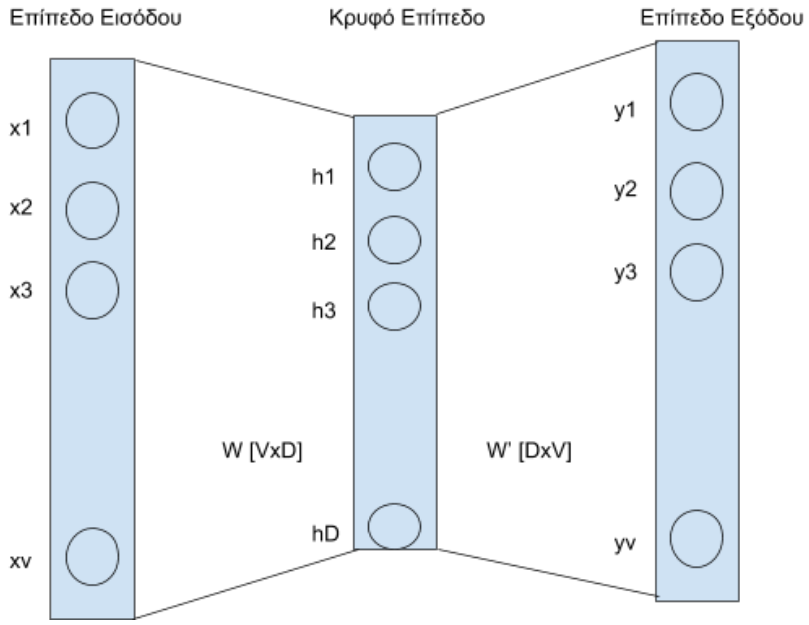
Αξίζει να σημειωθεί σε αυτό το σημείο ότι όλα τα μοντέλα ενθέσεων μπορούν να θεωρηθούν ως ιδανικές πραγματώσεις μεταφοράς μάθησης (transfer learning), όπου τα αποτελέσματα ενός μοντέλου που έχει εκπαιδευτεί για συγκεκριμένη εργασία μεταφέρονται αυτούσια ως είσοδος σε άλλα, διαφορετικά μοντέλα και εφαρμογές. Ανάλογες πρακτικές είναι εξαιρετικά διαδεδομένες σε εφαρμογές αναγνώρισης εικόνας, όπου τα παραγόμενα βάρη των επιπέδων των δικτύων μεταφέρονται και χρησιμοποιούνται ως είσοδος σε διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων (Simonyan & Zisserman, 2014; Krizhevsky et al., 2012).

Έστω λοιπόν ένα σώμα κειμένων, το οποίο μπορεί να θεωρηθεί ως μια ακολουθία T, το πλήθος λέξεων $w_1, w_2, w_3, ..., w_T$, οι οποίες ανήκουν σε ένα λεξικό V του οποίου το μέγεθος δηλώνεται ως $|V|$. Τα μοντέλα συνήθως λαμβάνουν ως είσοδο μια λέξη του σώματος κειμένων καθώς και το περικείμενο n λέξεων της λέξης αυτής. Κάθε όρος του λεξιλογίου V, w_i, απεικονίζεται σε μια d-διάστατη ένθεση. Τέλος, η εκπαίδευση επιτυγχάνεται βελτιστοποιώντας μια συνάρτηση-στόχο (objective function) $J(\theta)$ ως προς τις παραμέτρους θ του μοντέλου. Στη γενική περίπτωση, είσοδος στα μοντέλα είναι ένα διάνυσμα x λέξεων/περικειμένου και έξοδος κάποια συνάρτηση $f(x)$, η οποία ορίζεται με διαφορετικό τρόπο σε κάθε μοντέλο.

Αρχιτεκτονικές του word2vec: CBOW και SkipGram

Το μοντέλο word2vec υλοποιεί δύο αρχιτεκτονικές, τη «Συνεχή αναπαράσταση σωρού λέξεων», (Continuous Bag-Of-Word Model, CBOW) και το μοντέλο Παράλειψης ν-γραμμάτων (Skip-gram model)

CBOW Στην αρχιτεκτονική CBOW το μοντέλο έχει ως στόχο να υπολογίσει την πιθανότητα της εμφάνισης της λέξης w_i όταν εμφανίζεται σε κάποιο περικείμενο



Διάγραμμα 4.1: Αρχιτεκτονική μοντέλου CBOW, $C = 1$

$w_1, w_2, ..., w_{n1}$. Οι λέξεις αναπαρίστανται χρησιμοποιώντας την κωδικοποίηση μιας θέσης (one hot encoding), κατά την οποία σε κάθε λέξη του λεξιλογίου V αντιστοιχεί ένα διάνυσμα v με όλα τα στοιχεία του διανύσματος μηδενικά, εκτός από τη θέση που αντιστοιχεί στην συγκεκριμένη λέξη, που λαμβάνει την τιμή 1.

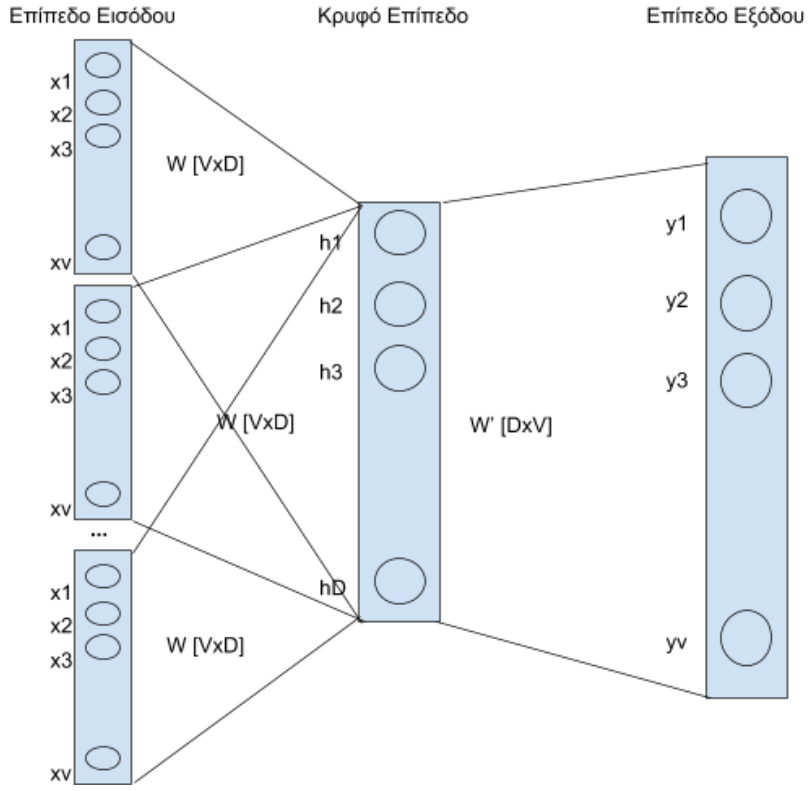
Στην απλούστερη περίπτωση $n = 2$, στην οποία η είσοδος είναι μόνο μια λέξη και προσπαθούμε να προβλέψουμε την πιθανότητα εμφάνισης της επόμενης, το δίκτυο έχει την εξής μορφή. Στην αρχιτεκτονική αυτή το δίκτυο αποτελείται από τρία επίπεδα, το επίπεδο εισόδου, το κρυφό επίπεδο και το επίπεδο εξόδου. Τα επίπεδα εισόδου και εξόδου έχουν την ίδια διάσταση, τον αριθμό των λέξεων του λεξιλογίου, |V|. Το κρυφό επίπεδο έχει διάσταση D, συνήθως πολύ μικρότερη από το $|V|$ με τυπικές τιμές σε εφαρμογές μεταξύ 50 και 2000.

Τα βάρη $W_{VxD} = w_{ij}$ μεταξύ εισόδου και κρυφού επιπέδου αναπαρίστανται με έναν πίνακα διαστάσεων VxD. Με αυτή την αρχιτεκτονική, κάθε λέξη του λεξιλογίου αντιστοιχεί σε ένα D-διάστατο διάνυσμα, στην αντίστοιχη γραμμή του πίνακα W. Το διάνυσμα x της k-οστής λέξης του λεξιλογίου κατά την κωδικοποίηση one-hot, θα έχει τιμή 1 στη θέση k και 0 σε όλες τις υπόλοιπες, $x_k = 1$ και $x_{k'} = 0, k' \neq k$

Η συνάρτηση ενεργοποίησης στο κρυφό επίπεδο είναι η γραμμική συνάρτηση

$$h = W^T x = W^T_{k,:} = v^T_{wi}$$

στην οποία απλά η έξοδος της συνάρτησης h είναι η k-οστή γραμμή του πίνακα



Διάγραμμα 4.2: Αρχιτεκτονική μοντέλου CBOW, $C > 1$

W.

Τα βάρη μεταξύ του κρυφού επιπέδου και του επιπέδου εξόδου, $W' = w'_{ij}$ αναπαρίστανται από έναν πίνακα W' διαστάσεων DxV. Χρησιμοποιώντας τα βάρη αυτά μπορούμε να υπολογίσουμε το αποτέλεσμα της διαδικασίας για κάθε θέση j στο διάνυσμα εξόδου,

$$u_j = v'^T_{wj} h$$

με v'_{wj} την j-οστή στήλη του πίνακα W'.

Τέλος, για τη μετατροπή αυτών των τιμών σε πιθανότητες χρησιμοποιείται η συνάρτηση softmax, η οποία δίνει

$$p(w_j|w_I) = y_j = \frac{e^{u_j}}{\sum_{k=1}^{|V|} e^{u_k}}$$

Συνδυάζοντας τις παραπάνω εξισώσεις, παίρνουμε

$$p(w_j|w_I) = \frac{e^{v^T_{w_j} v_{wI}}}{\sum_{k=1}^{|V|} e^{v^T_{wI} v_{wI}}}$$

Τα διανύσματα v_w και v_w είναι αναπαραστάσεις της λέξης w, με το πρώτο να έρχεται από τον πίνακα W και το δεύτερο από τον πίνακα W

Κατά τη διαδικασία εκπαίδευσης του μοντέλου στόχος είναι η μεγιστοποίηση της υπό συνθήκη πιθανότητας $p(w_O|w)$ να παρατηρηθεί η σωστή λέξη στόχος w_O. Εάν η θέση της λέξης-στόχου στο διάνυσμα εξόδου είναι j^*, τότε

$$maxp(w_O|w_I) = max(y_{j^*}) = maxlog(max(y_{j^*})) = u_{j^*} log \sum_{k=1}^{|V|} e^{u_k} := E$$

και E η συνάρτηση κόστους της εκπαίδευσης του μοντέλου.

Η κλίση της συνάρτησης κόστους ως προς την έξοδο είναι

$$\frac{\partial E}{\partial u_j} = y_j t_j := e_j$$

και

$$t_j = \mathbf{1}(j = j^*)$$

Στη συνέχεια, η κλίση της συνάρτησης κόστους ως προς τα βάρη W_{ij} είναι

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial w_{ij}} = e_j \cdot h_i$$

Ανάλογα, η κλίση της συνάρτησης κόστους ως προς τα βάρη W_{ij} είναι

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{|V|} \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial h_i} = \sum_{j=1}^{|V|} e_j \cdot w_{ij} := EH_i$$

$$h_i = \sum_{k=1}^{|V|} x_k \cdot w_{ki}$$

και, τέλος, από εδώ

$$\frac{\partial E}{\partial w_{ki}} = \frac{\partial E}{\partial h_i} \cdot \frac{\partial h_i}{\partial w_{ki}} = EH_i \cdot x_k$$

Από τις εξισώσεις αυτές προκύπτουν οι ενημερώσεις των βαρών για τον αλγόριθμο κατάβασης κλίσης για τα βάρη του πίνακα W' που συνδέει το κρυφό επίπεδο με το επίπεδο εξόδου

$$w_{ij} \leftarrow w_{ij}\eta \cdot e_j \cdot h_i$$

ή

$$v_{w_j} \leftarrow v_{w_j}\eta \cdot e_j \cdot \mathbf{h}$$

για

$$j = 1, 2, ..., |V|$$

Τέλος, για τα βάρη

$$w_{ij}$$

που συνδέουν το επίπεδο εισόδου με το κρυφό επίπεδο, η ενημέρωση είναι

$$v_{w_I} \leftarrow v_{w_I}\eta \cdot EH^T$$

όπου η ο ρυθμός εκπαίδευσης.

Οι κανόνες αυτοί είναι τα βήματα ενημέρωσης των βαρών των πινάκων W, $W^{'}$ στην περίπτωση που το περικείμενο είναι μια λέξη, όπως είδαμε (n=2). Στην γενική περίπτωση, όταν n>2 και με $C = n - 1$ τον αριθμό των λέξεων του περικειμένου, η συνάρτηση ενεργοποίησης στο κρυφό επίπεδο γίνεται:

$$h = \frac{1}{C} \cdot W^T(x_1 + x_2 + ... + x_C) = \frac{1}{C}(v_{w_1} + v_{w_2} + ...v_{w_c})$$

δηλαδή ο μέσος όρος των διανυσμάτων του πίνακα W που αντιστοιχούν στις λέξεις εισόδου. Η συνάρτηση κόστους έχει μορφή

$$E = logp(w_O|w_{I_1}, w_{I_2}, ..., w_{I_C}) = u_j^* + log \sum_{k=1}^{|V|} e^{u_k} = \mathbf{v}_{w_O}^T \cdot \mathbf{h} + log \sum_{k=1}^{|V|} e^{v_{w_k}^T \cdot \mathbf{h}}$$

Οι ενημερώσεις των βαρών W'_{ij} παραμένουν ίδιες με παραπάνω στην περίπτωση που C=1

$$v_{w_j} \leftarrow v_{w_j} \eta \cdot e_j \cdot \mathbf{h}$$

για $j = 1, 2, ..., |V|$, ενώ για τα βάρη του επίπεδου εισόδου με το κρυφό επίπεδο ισχύει

$$v_{w_I,c} \leftarrow v_{w_I,c} \frac{1}{C} \eta \cdot EH^T$$

για $c = 1, 2, ..., C$ όπου $v_{w_I,c}$ το διάνυσμα της c-οστής λέξης εισόδου και η ο ρυθμός εκπαίδευσης (learning rate).

Skip-Gram Η αρχιτεκτονική του μοντέλου skip-gram είναι αντίστροφη από αυτή του CBOW. Στην αρχιτεκτονική αυτή είσοδος στο μοντέλο είναι μία μόνο λέξη και στόχος του μοντέλου είναι να προβλέψει την πιθανότητα εμφάνισης των λέξεων του περικειμένου. Η αρχιτεκτονική του επιπέδου εισόδου προς το κρυφό επίπεδο συνεπώς είναι η ίδια με την περίπτωση του CBOW για C=1. Το διάνυσμα της λέξης εισόδου συμβολίζεται με v_{wI}, όπως και στην περίπτωση του CBOW, και η συνάρτηση ενεργοποίησης στο κρυφό επίπεδο h παραμένει η ίδια

$$h = W_{k,:}^T := v_{wI}^T$$

Στο επίπεδο εξόδου όμως αντί μιας πολυωνυμικής κατανομής το μοντέλο παράγει διαφορετικές C πολυωνυμικές κατανομές, καθεμιά με διάσταση $|V|$. Η πιθανότητα της λέξης j στην κατανομή c δίνεται από τη σχέση

$$p(w_{c,j} = w_{O,c}|w_I) = y_{c,j} = \frac{exp(u_{c,j})}{\sum = 1^{|V|} exp(u_k)}$$

με

$$c = 1, 2, ..., C, j = 1, 2, ..., |V|$$

και $w_{O,c}$ η c-οστή λέξη στόχος και $u_{c,j}$ η j-οστή τιμή της c-οστής κατανομής από το κρυφό επίπεδο προς το επίπεδο εξόδου. Επειδή οι κατανομές εξόδου έχουν κοινό πίνακα βαρών W', $u_{c,j} = u_j = \mathbf{v}_{w_j}^T \cdot \mathbf{h}$, για $c = \{1, 2, ..., C\}$,

όπου

$$\mathbf{v}_{w_j}^T$$

η έξοδος της λέξης j του λεξικού, w_j από τον πίνακα W.

Η συνάρτηση κόστους για το μοντέλο skip-gram είναι της μορφής

$$E = logp(w_{O,1}, w_{O,2}, ..., w_{O,C}|w_I) = log \prod_{c=1}^{C} \frac{exp(u_{c,j_c^*})}{\sum_{k=1}^{|V|} exp(u_k)} = \sum_{c=1}^{C} u_{j_c^*} + C \cdot log \sum_{k=1}^{|} V|exp(u_k)$$

όπου $u_{j_c^*}$ είναι ο δείκτης της λέξης στόχου στην c-οστή κατανομή εξόδου. Όπως και στα προηγούμενα,

$$\frac{\partial E}{\partial u_{c,j}} = y_{c_j} t_{c,j} := e_{c,j}$$

δηλαδή το σφάλμα ανά λέξη j και περικείμενο c.

Αν τώρα θέσουμε

$$EI_j = \sum_{i=1}^{C} e_{c,j}$$

το άθροισμα των σφαλμάτων για την λέξη j σε όλα τα επίπεδα C, παίρνουμε το διάνυσμα

$$EI = \{EI_1, EI_2, ..., EI_{|V|}\}$$

Στη συνέχεια για τον κανόνα ενημέρωσης βαρών του πίνακα W' ,

$$\frac{\partial E}{\partial w'_{i,j}} = \sum_{c=1}^{C} \frac{\partial E}{\partial u_{c,j}} \frac{\partial u_{c,j}}{\partial w'_{i,j}} = EI_j \cdot h_i$$

και συνεπώς

$$\mathbf{v}_{w_j} \leftarrow \mathbf{v}'_{w_j} \eta \cdot EI_j \cdot \mathbf{h}$$

Τέλος, αν ορίσουμε το D-διάστατο διάνυσμα

$$EH = \{EH_1, EH_2, ..., EH_D\}$$

με

$$EH_i = \sum_{k=1}^{|V|} EI_k \cdot w'_{i,k}$$

η εξίσωση ενημέρωσης των $w_{i,j}$ είναι παρόμοια με αυτή του CBOW για $C = 1$ με τη διαφορά ότι το σφάλμα e_j αντικαθίσταται με το σφάλμα EI_j και είναι

$$v_{w_I} \leftarrow v_{w_I} - \eta \cdot EH^T$$

όπου η ο ρυθμός εκπαίδευσης του μοντέλου.

Βελτιστοποίηση

Στο μοντέλο εκπαίδευσης, όπως περιγράφηκε παραπάνω, υπολογίζεται η πιθανότητα κάθε λέξης στην έξοδο του μοντέλου για κάθε λέξη του λεξιλογίου. Η διαδικασία αυτή έχει μεγάλο υπολογιστικό κόστος και είναι σχεδόν απαγορευτική για πολύ μεγάλα σώματα κειμένων με λεξικό αρκετών δεκάδων ή εκατοντάδων εκατομμυρίων λέξεων. Για να αποφευχθεί αυτό το πρόβλημα στην πράξη χρησιμοποιούνται δύο τεχνικές, το ιεραρχικό softmax (hierarchical softmax) (Morin & Bengio, 2005) και η αρνητική δειγματοληψία (negative sampling) (Mikolov et al., 2013).

Στην περίπτωση του ιεραρχικού softmax κατασκευάζεται ένα δέντρο Huffmanß[12] για τις λέξεις του λεξιλογίου, βασισμένο στην συχνότητα εμφάνισης των λέξεων. Κάθε λέξη είναι τερματικός κόμβος αυτού του δέντρου και είναι προσβάσιμη από την ρίζα του δέντρου. Κάθε επίπεδο του δέντρου είναι κανονικοποιημένο (οι πιθανότητες διακλάδωσης αθροίζουν στην μονάδα), οπότε η κανονικοποίηση αντί να εφαρμοστεί σε $O(V)$ βήματα χρειάζεται μόνο $O(logV)$ βήματα, όσα χρειάζονται για να φτάσουμε στη λέξη-στόχο.

Στην περίπτωση της αρνητικής δειγματοληψίας αντί να υπολογιστεί η συνάρτηση softmax σε όλο το λεξιλόγιο V, γίνεται δειγματοληψία σε K λέξεις από το V, συμπεριλαμβανομένης και της λέξης στόχου, με $K << V$. Το όνομα «αρνητική δειγματοληψία» (negative sampling) προέρχεται ακριβώς από την επιλογή δείγματος λέξεων από την «αρνητική» κλάση, δηλαδή όλες τις λέξεις του λεξιλογίου εξαιρουμένης της λέξης στόχου.

Η ενημέρωση των βαρών W χρειάζεται να υπολογιστεί μόνο για K λέξεις, δηλαδή στο σύνολο που αποτελείται από το δείγμα των λέξεων στην αρνητική δειγματοληψία και τη λέξη στόχο και όχι σε όλο το λεξικό και, συνεπώς, ο υπολογισμός είναι εξαιρετικά πιο αποδοτικός.

Τέλος, αξίζει να σημειωθεί ότι οι Levy & Goldberg (2014) αναλύοντας τη συνάρτηση-στόχο του μοντέλου word2vec με αρνητική δειγματοληψία απέδειξαν ότι το μοντέλο αυτό πραγματοποιεί στην ουσία παραγοντοποίηση του πίνακα Σημειακής αμοιβαίας πληροφορίας (Pointwise mutual information, PMI) των λέξεων του σώματος κειμένων.

[12] https://www.siggraph.org/education/materials/HyperGraph/video/mpeg/mpegfaq/huffman_tutorial.html

4.6.4 Το μοντέλο fastText

Το μοντέλο fastText (Joulin et al., 2016) είναι κατά μία έννοια επέκταση του word2vec. Όπως είδαμε, στο word2vec (και γενικότερα σε νευρωνικά γλωσσικά μοντέλα) οι λέξεις απεικονίζονται σε d-διάστατο διανυσματικό χώρο με βάση την πιθανότητα εμφάνισης της λέξης w_{n+1} μετά από μια ακολουθία λέξεων $w_1, w_2, ..., w_n$. Οι λέξεις συνήθως προκύπτουν από το μεγάλο σε όγκο, ως προς τον αριθμό λέξεων, σώμα κειμένων μετά την εφαρμογή κάποιου τύπου λεξικής ανάλυσης. Συνήθως η λεξική ανάλυση υλοποιείται με εφαρμογή απλών κανόνων. Για παράδειγμα, τα όρια των λέξεων ορίζονται από μη κενούς (non-whitespace) χαρακτήρες (white space tokenization).

Κατά την προσέγγιση αυτή δεν υπάρχει άμεσος τρόπος χρήσης μορφολογικών πληροφοριών της γλώσσας στην οποία εκπαιδεύεται το μοντέλο. Έτσι, σπάνιοι μορφολογικοί τύποι ή λέξεις με τυπογραφικά λάθη ενδεχομένως να αγνοηθούν από το μοντέλο. Το fastText λύνει αυτό το πρόβλημα μαθαίνοντας d-διάστατες αναπαραστάσεις n-γραμμάτων χαρακτήρων. Κάθε λέξη αντιστοιχεί σε ένα σύνολο σωρού ν-γραμμάτων με δύο επιπλέον ειδικούς χαρακτήρες "<" και ">", που υποδηλώνουν την αρχή και το τέλος κάθε λέξης. Κάθε ν-γραμμα απεικονίζεται με τη σειρά του σε d-διάστατο διάνυσμα και στο τέλος οι λεξικές ενθέσεις υπολογίζονται ως το άθροισμα των διανυσμάτων των ν-γραμμάτων που τις απαρτίζουν (Bojanowski et al., 2016).

Με την κωδικοποίηση αυτή, περιπτώσεις συχνών ορθογραφικών λαθών, όπως για παράδειγμα ο τύπος *ορκομωσία* (περίπου 65.000 αποτελέσματα στη μηχανή αναζήτησης Google) αντί του *ορκωμοσία* (650.000 αποτελέσματα στο google) που διαφέρουν μόνο κατά δύο χαρακτήρες,[13], θα βρίσκονται πάρα πολύ κοντά και ως διανυσματικές ενθέσεις.

Το ίδιο προφανώς συμβαίνει και για τους τύπους που προκύπτουν από την κλίση λέξεων και τους μορφολογικούς κανόνες της γλώσσας, κατά συνέπεια οι λέξεις *ήχος*, *ήχου*, *ήχοι* θα βρίσκονται γεωμετρικά πάρα πολύ κοντά ως διανυσματικές ενθέσεις.

4.6.5 Το μοντέλο GloVe (Global Vectors)

Στα μοντέλα word2vec και fastText οι ενθέσεις είναι αποτέλεσμα της εκπαίδευσης ενός νευρωνικού δικτύου που στόχο έχει την εκμάθηση γλωσσικού μοντέλου. Το μοντέλο GloVe (Global Vectors, Pennington et al., 2014) ταξινομείται στην κατηγορία των τεχνικών παραγοντοποίησης πινάκων (matrix factorization). Οι τεχνικές παραγοντοποίησης πινάκων έχουν τις ρίζες τους στη Λανθάνουσα

[13] Levenshtein distance 2 και jacard similarity 1 μιας και έχουν ακριβώς τους ίδιους χαρακτήρες (με μία αντιμετάθεση των ω, ο)

σημασιολογική ανάλυση (LSA), στην οποία οι πίνακες έχουν την μορφή όρος-κείμενο, δηλαδή οι γραμμές των πινάκων αντιστοιχούν στις λέξεις ενός κειμένου και οι στήλες των πινάκων αντιστοιχούν στα κείμενα, που απαρτίζουν το σώμα κειμένων. Στο μοντέλο HAL (Hyperspace Analogue to Language, Lund & Burgess, 1996), οι πίνακες έχουν την μορφή όρος-όρος. Οι γραμμές και οι στήλες του πίνακα αντιστοιχούν σε λέξεις και κάθε στοιχείο (i,j) του πίνακα περιέχει τον αριθμό των εμφανίσεων της λέξης, που αντιστοιχεί στην γραμμή i στο περικείμενο της λέξης, που αντιστοιχεί στην στήλη j.

Το βασικό πρόβλημα με το μοντέλο HAL είναι ότι οι υψίσυχνες λέξεις προσθέτουν θόρυβο στο μοντέλο και αλλοιώνουν τα μέτρα της ομοιότητας. Για την αντιμετώπιση αυτών των προβλημάτων αναπτύχθηκαν διάφορες μέθοδοι, όπως π.χ. η μέθοδος COALS (Rohde et al., 2006), στην οποία εφαρμόζεται κανονικοποίηση με βάση την εντροπία ή συντελεστή συσχέτισης στον πίνακα συνεμφανίσεων.

Το μοντέλο GloVe ακολουθεί διαφορετική προσέγγιση για να λύσει αυτό το πρόβλημα. Πιο συγκεκριμένα, έστω X τετραγωνικός πίνακας με στοιχεία X_{ij} τον αριθμό εμφανίσεων της λέξης j στο περικείμενο της λέξης i για περικείμενο C λέξεων, $X_i = \sum_k X_{ik}$ ο συνολικός αριθμός συνεμφανίσεων όλων των λέξεων με την λέξη i και $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ η πιθανότητα η λέξη j να εμφανιστεί στο περικείμενο της λέξης i. Αν θεωρήσουμε δύο λέξεις i,j και τη σχέση τους με μια τρίτη λέξη k, ο λόγος $\frac{P_{ik}}{P_{jk}}$ μπορεί να θεωρηθεί ως μέτρο του πόσο σχετική είναι η λέξη k με την i ή την j αντίστοιχα. Για παράδειγμα, όπως αναφέρουν οι Pennington et al. (2014, σ. 2) αν i = *πάγος*, j = *ατμός*, για λέξεις που σχετίζονται με τον πάγο αλλά όχι με τον ατμό, π.χ. *στερεός*, ο λόγος P_{ik}/P_{jk} θα είναι μεγάλος, ενώ αντίστοιχα αν k=*αέριος*, το P_{ik}/P_{jk} θα τείνει να είναι μικρό. Για λέξεις μη σχετικές με το i, j, ο λόγος θα είναι περίπου 1.

Με βάση αυτή την παρατήρηση αναζητείται μια συνάρτηση F τέτοια ώστε

$$F(w_i, w_j, \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$$

.

Οι παράμετροι w_i, w_j είναι διανύσματα λέξεων διαστάσεως D και $\tilde{w_k}$ τα διανύσματα των λέξεων περικειμένου. Για να περιοριστεί το σύνολο των πιθανών συναρτήσεων που ικανοποιούν τη συνθήκη εφαρμόζονται περιορισμοί. Αρχικά, τα w είναι διανύσματα και ο λόγος πραγματικός αριθμός, οπότε μας ενδιαφέρει γραμμική σχέση μεταξύ των w και η συνάρτηση F με ορίσματα διανύσματα να δίνει ως αποτέλεσμα πραγματικό αριθμό. Το εσωτερικό γινόμενο διανυσμάτων ικανοποιεί αυτή την απαίτηση, οπότε η F μπορεί να περιοριστεί σε συναρτήσεις της μορφής $F((w_i w_j)^T \tilde{w_k}) = \frac{P_{ik}}{P_{jk}}$. Συνεχίζοντας, ο ορισμός διανυσμάτων λέξεων και περικειμένου είναι αυθαίρετος και συνεπώς τα διανύσματα λέξεων

συνεμφανίσεων και λέξεων περικειμένου πρέπει να μπορούν να εναλλάσσονται, απαιτούμε η F να είναι ομομορφισμός (homomorphism)[14] μεταξύ των ομάδων $(R,+)$ και$(R+,*)$, δηλαδή

$$F((w_i - w_j)^T \tilde{w_k}) = \frac{F(w_i^T \tilde{w_k})}{F(w_j^T \tilde{w_k})} \tag{4.11}$$

Η λύση της εξίσωσης 4.11 είναι η εκθετική συνάρτηση, δηλαδή

$$w_i^T \tilde{w_k} = log(P_{ik} = log(X_{ik})log(X_i)$$

.Τέλος, προστίθενται οι όροι b_i και $\tilde{b_k}$ για λόγους συμμετρίας και η ζητούμενη εξίσωση γίνεται

$$w_i^T \tilde{w_k} + b_i + \tilde{b_k} = log(X_{ik}) \tag{4.12}$$

Για την εύρεση των παραμέτρων w η εξίσωση 4.12 μετατρέπεται σε πρόβλημα ελαχίστων τετραγώνων της μορφής

$$J = \sum_{i,j=1}^{|V|} f(X_{ij})(w_i^T \cdot \tilde{w}_j + b_i + \tilde{b}_j - log(X_{ij}))^2 \tag{4.13}$$

με J τη συνάρτηση κόστους προς ελαχιστοποίηση, $|V|$ το μέγεθος του λεξικού και f συνάρτηση στάθμισης, η οποία επιλέγεται ώστε να έχει τις εξής ιδιότητες:

- $f(0) = 0$
- $f(x)$ μη φθίνουσα
- Η $f(x)$ να έχει σχετικά μικρές τιμές για μεγάλες τιμές του x, ώστε να ομαλοποιεί συχνές συνεμφανίσεις.

Μια συνάρτηση που ικανοποιεί τις παραπάνω ιδιότητες και δίνει καλά εμπειρικά αποτελέσματα στο μοντέλο είναι η

$$f(x) = \begin{cases} (x/x_{max})^a, x < x_{max} \\ 1, \quad x \geq x_{max} \end{cases}$$

Τυπικές τιμές για τα a, x_{max} είναι $a = 3/4$, $x_{max} = 100$.

Οι παράμετροι w, δηλαδή τα διανύσματα των λέξεων προκύπτουν από την ελαχιστοποίηση της συνάρτηση κόστους J με κάποια παραλλαγή αλγόριθμου στοχαστικής κατάβασης κλίσης. Οι Pennington et al. (2014) χρησιμοποιούν στην υλοποίησή τους τον αλγόριθμο AdaGrad (Duchi et al., 2011).

[14]Εστω V, W διανυσματικοί χώροι. Μια απεικόνιση $f : V \to W$ καλείται ομομορφισμός όταν ισχύει i) $f(v_1 + v_2) = f(v_1) + f(v_2)$ και ii) $f(\lambda v) = \lambda f(v)$

4.6.6 Το μοντέλο doc2vec (paragraph vectors)

Το μοντέλο doc2vec αποτελεί φυσική επέκταση του μοντέλου word2vec. Το μοντέλο αυτό μαθαίνει κατανεμημένες αναπαραστάσεις τμημάτων κειμένου μεταβλητού μήκους. Τα τμήματα αυτά μπορεί να είναι παράγραφοι ή ολόκληρα κείμενα. Η αρχιτεκτονική και η εκπαίδευση του μοντέλου γίνεται με τέτοιο τρόπο ώστε το μοντέλο να μπορεί να προβλέπει τις λέξεις ενός τμήματος κειμένου. Η εκπαίδευση γίνεται τυπικά με τη μέθοδο στοχαστικής κατάβασης κλίσης (Stochastic Gradient Descent,SGD, Παράρτημα 1). Πιο συγκεκριμένα, στη διαδικασία της εισόδου στο μοντέλο διανύσματα λέξεων από την παράγραφο (ή το κείμενο) μαζί με το διάνυσμα της παραγράφου χρησιμοποιούνται ως είσοδος και το ζητούμενο είναι να προβλεφθεί η σειρά λέξεων, που ακολουθεί αυτή την είσοδο δεδομένης της παραγράφου. Τα διανύσματα των λέξεων είναι κοινά για όλες τις παραγράφους και τα διανύσματα των παραγράφων μοναδικά (ένα διάνυσμα για κάθε παράγραφο).

Στη συνέχεια, με παρόμοιο τρόπο με το word2vec, δεδομένης μιας λέξης του κειμένου το περικείμενο σταθερού μήκους k και ταυτόχρονα και το διάνυσμα της παραγράφου δίνεται ως είσοδος στο νευρωνικό δίκτυο και στόχος παραμένει η πρόβλεψη της λέξης w, δεδομένου του περικειμένου αλλά και του διανύσματος της παραγράφου.

Τα διανύσματα των λέξεων και των παραγράφων αρχικοποιούνται με τυχαίες τιμές στο διάστημα $[-1,-1]$ είτε από ομοιόμορφη κατανομή ή από περικεκομμένη κανονική κατανομή (truncated gaussian). Όταν τελειώσει η διαδικασία της εκπαίδευσης, τα διανύσματα των παραγράφων μπορούν να χρησιμοποιηθούν ως είσοδος σε μοντέλα μηχανικής μάθησης, επιβλεπόμενης (Logistic Regression, Support Vector Machines) ή μη επιβλεπόμενης (K-means)

Ο αλγόριθμος λειτουργεί σε δύο διακριτά στάδια: 1) το στάδιο της εκπαίδευσης 2) το στάδιο της εξαγωγής συμπερασμάτων, όπου το ζητούμενο είναι ο υπολογισμός του διανύσματος μιας νέας, άγνωστης παραγράφου. Τα παραγόμενα διανύσματα κληρονομούν τα χαρακτηριστικά των ανάλογων διανυσμάτων, που παράγονται από το word2vec, δηλαδή παρόμοιες σημασιολογικά λέξεις βρίσκονται κοντά στο διανυσματικό χώρο. Επίσης, λαμβάνεται υπόψη η σειρά των λέξεων σε αντίθεση με την αναπαράσταση σωρού λέξεων (bag of words) και αυτό έχει ως αποτέλεσμα τα παραγόμενα διανύσματα να έχουν καλύτερα αποτελέσματα, όταν χρησιμοποιούνται σε εφαρμογές επεξεργασίας φυσικής γλώσσας. Επίσης, επειδή τυπικά η διαστασιμότητα του χώρου είναι μικρή (τυπικές τιμές 200-500), τα παραγόμενα διανύσματα είναι αποδοτικότερα σε σχέση με προσεγγίσεις ν-γραμμάτων, τα οποία, αν και καταγράφουν τη σειρά των λέξεων, έχουν τεράστια διαστασιμότητα με αποτέλεσμα να έχουν μεγάλο υπολογιστικό κόστος και επίσης να έχουν περιορισμένη ικανότητα γενίκευσης.

Το μοντέλο doc2vec συνεπώς μπορεί να χρησιμοποιηθεί για την καταγραφή των υφολογικών χαρακτηριστικών των χρηστών. Συγκεντρώνοντας όλα τα κείμενα των συγγραφέων ενός σώματος κειμένου σε ένα υπερ-κείμενο το ζητούμενο είναι να εκπαιδευτεί ένα μοντέλο doc2vec ώστε να μάθει τα διανύσματα των κειμένων των διαφορετικών αυτών συγγραφέων. Με βάση αυτά τα διανύσματα μπορούμε πλέον να συγκρίνουμε δύο διαφορετικούς συγγραφείς όσον αφορά την ομοιότητά τους.

Η ιδέα αυτή έρχεται σε συμφωνία με την κατανεμητική σημασιολογία, αν υποθέσουμε ότι συγγραφείς που κάνουν παρόμοια χρήση λεξιλογίου και έχουν παρόμοιες προτιμήσεις στη σειρά με την οποία επιλέγουν και συνθέτουν κείμενα θα τείνουν να βρίσκονται πολύ κοντά στον διανυσματικό χώρο. Η χαρακτηριστική ιδιότητα του μοντέλου να διατηρεί τη σειρά των λέξεων στην είσοδο ενισχύει την υπόθεση, καθώς υψίσυχνες λέξεις και συνάψεις θα είναι διαμοιρασμένες μεταξύ όλων των συγγραφέων και οι όποιες διαφοροποιήσεις θα ενισχύουν σημαντικά τη διακριτική ικανότητα των παραγόμενων διανυσμάτων.

Συνοψίζοντας, ένα νευρωνικό δίκτυο εκπαιδεύεται με τέτοιο τρόπο ώστε να μαθαίνει α) διανυσματικές αναπαραστάσεις λέξεων με τέτοιο τρόπο ώστε σημασιολογικά κοντινές λέξεις να είναι κοντά γεωμετρικά, και β) με βάση τις σημασιολογικές αναπαραστάσεις των λέξεων να μαθαίνει αντίστοιχες διανυσματικές αναπαραστάσεις των συγγραφέων του σώματος κειμένων με βάση τα σχήματα συχνότητας χρήσης λέξεων, αλλά και σημασιολογικών επιλογών των συγγραφέων, παράγοντας με αυτόν τον τρόπο μια κατανεμημένη αναπαράσταση της ιδιολέκτου. Αυτή η σύνθεση κατανεμητικής υπόθεσης με βάση τη χρήση νευρωνικών δικτύων αιτιολογεί και την χρήση του όρου *νευροκατανεμητικά σημασιολογικά μοντέλα*.

4.7 Λεξικές ενθέσεις και ιδιόλεκτος

Στην ενότητα αυτή τίθεται το πρόβλημα της αποτύπωσης ύφους ως αποτέλεσμα εκπαίδευσης μοντέλων λεξικών ενθέσεων. Στην ενότητα 4.7.1 αναπτύσσονται προσεγγίσεις κατασκευής ενθέσεων ύφους είτε με χρήση των ενθέσεων σε επίπεδο λέξης (για τα μοντέλα word2vec, GloVe, fastText) είτε απευθείας (από το μοντέλο doc2vec).

Στην περίπτωση της κατασκευής διανυσμάτων ύφους με χρήση ενθέσεων περιγράφονται οι δύο βασικές τεχνικές που χρησιμοποιούνται στην πράξη καθώς και στην παρούσα διατριβή, η διανυσματική προσέγγιση και η τεχνική Ανάστροφης Συχνότητας, στην ενότητα 4.7.3. Σε επόμενη παράγραφο (4.7) θα δούμε πώς η παραγωγή ενθέσεων σε επίπεδο λέξης μπορεί να γενικευθεί και σε επίπεδο συγγραφέα και να ερμηνευθεί ως «αποτύπωμα» ύφους.

4.7.1 Λεξικές ενθέσεις ύφους

Όπως είδαμε μέχρι στιγμής, οι λεξικές ενθέσεις έχουν αποδειχτεί στην πράξη μια εξαιρετικά αποδοτική τεχνική όσον αφορά την επίλυση προβλημάτων επεξεργασίας φυσικής γλώσσας. Στην περίπτωσή μας το ζητούμενο είναι η εύρεση μιας μεθοδολογίας που θα απαντά στα εξής ερωτήματα:

- Δεδομένου ενός σώματος κειμένων, που περιέχει μεγάλο όγκο κειμένων από μεγάλο πλήθος συγγραφέων/ομιλητών, υπάρχει τρόπος να κωδικοποιηθεί το ιδιαίτερο ύφος καθενός από αυτούς, ώστε να μας επιτρέπει να ποσοτικοποιήσουμε την ομοιότητα μεταξύ τους;

- Είναι εφικτό να χρησιμοποιήσουμε το υφολογικό αυτό αποτύπωμα ως εργαλείο ομαδοποίησης χρηστών με παρόμοιο ύφος;

Στην προηγούμενη ενότητα είδαμε πως λεξικές ενθέσεις που παράγονται ως αποτέλεσμα της εκπαίδευσης ενός από τα μοντέλα ενθέσεων αφορούν δυνητικά όλο το λεξιλόγιο του σώματος κειμένων και δεν περιορίζονται στη χρήση των πιο υψίσυχνων λέξεων. Περιορισμοί που εφαρμόζονται συνήθως στην πράξη κυρίως για τη διατήρηση των μοντέλων σε λογικά υπολογιστικά πλαίσια είναι η αφαίρεση θορύβου. Δηλαδή, λέξεις με συχνότητα εμφάνισης μικρότερη από μια σταθερά (π.χ. 2, 10) που συνήθως παράγονται είτε από ορθογραφικά λάθη ή, στην περίπτωση των ελληνικών για παράδειγμα, τη μίξη ελληνικών και λατινικών χαρακτήρων.

Επίσης, όπως φαίνεται στην πράξη, τα σημασιολογικά χαρακτηριστικά που συλλαμβάνουν οι παραγόμενες ενθέσεις μπορούν στην πράξη να αντικαταστήσουν χαρακτηριστικά συντακτικής φύσης και να δώσουν εξαιρετικά καλά αποτελέσματα σε προβλήματα σχετιζόμενα με σχήματα λόγου, όπως η μεταφορά (Φλώρου, 2017) [15].

Όπως είδαμε, οι λεξικές ενθέσεις ως απεικονίσεις λέξεων σε ν-διάστατο διανυσματικό χώρο προσφέρουν τη δυνατότητα ποσοτικοποίησης και απαντήσεων σε ερωτήματα σημασιολογικής ομοιότητας σε επίπεδο τύπων με εφαρμογή απλών αλγεβρικών πράξεων. Ωστόσο, το ζητούμενο εδώ είναι η ποσοτικοποίηση και η σύγκριση οντοτήτων που παράγουν κείμενα και όχι η κατά σημείο (τύπο) σύγκριση.

Το πρόβλημα λοιπόν μεταφέρεται στην εύρεση ενός τρόπου με τον οποίο οι λεξικές ενθέσεις που αντιστοιχούν στις ιδιολεκτικές προτιμήσεις ενός συγγραφέα, εκφρασμένες ως συχνότητες λέξεων, συντακτικά σχήματα και σημασιολογικές

[15] Είναι εξαιρετικά ενδιαφέρουσα η παρατήρηση ότι επαληθεύεται εμπειρικά η θεωρία της Gattis για τη συσχέτιση αφηρημένων εννοιών και της απεικόνισης τους σε χωρικά σχήματα (Gattis, 2003, σ. 5)

επιλογές, θα συνδυαστούν με τέτοιο τρόπο ώστε να χαρακτηρίζουν μονοσήμαντα τον ίδιο συγγραφέα.

Ένας διαφορετικός τρόπος προσέγγισης του προβλήματος είναι η τροποποίηση ενός από τα μοντέλα με τέτοιον τρόπο ώστε να διατηρεί τις ιδιότητες των λεξικών ενθέσεων αλλά να παρέχει τρόπο κωδικοποίησης των χαρακτηριστικών των οντοτήτων που παράγουν τα κείμενα.

Δεδομένου ότι η ιδιότητα αλγεβρικών πράξεων σε διανυσματικό χώρο ισχύει για τις λεξικές ενθέσεις, μια προσέγγιση θα ήταν να εκπαιδευτεί ένα μοντέλο ενθέσεων σε όλα τα διαθέσιμα κείμενα και να χρησιμοποιηθούν τα διανύσματα των λεξικών ενθέσεων ανά συγγραφέα για τις N πιο συχνές λέξεις του λεξιλογίου και οι συγκρίσεις να μεταφερθούν σε αυτόν το *Nxd*-διάστατο χώρο. Αυτή η προσέγγιση, όμως, δεν προσφέρει καμία βελτίωση σε σχέση με τη σύγκριση των N πιο συχνών λέξεων των χρηστών υπό την έννοια ότι παραλείπει λέξεις μέσης, χαμηλής ή πολύ χαμηλής συχνότητας. Επίσης, το γενικό μοντέλο εκπαιδεύεται σε όλο το σώμα κειμένων, ακυρώνοντας έτσι τις μορφολογικές, συντακτικές και σημασιολογικές ιδιαιτερότητες που διαφοροποιούν υφολογικά τους συγγραφείς.

Μια εναλλακτική προσέγγιση για την παραγωγή του ιδιολεκτικού διανύσματος είναι η εφαρμογή μαθηματικών συναρτήσεων και μετασχηματισμών στις ενθέσεις που αντιστοιχούν στις λέξεις που παράγει ένας συγγραφέας. Στη θεώρηση αυτή οι ενθέσεις που αντιστοιχούν στις λέξεις που χρησιμοποιεί ο συγγραφέας σχηματίζουν έναν πίνακα, στον οποίο εφαρμόζονται στη συνέχεια μαθηματικοί μετασχηματισμοί με στόχο την παραγωγή ενός τελικού διανύσματος το οποίο: α) περιγράφει μονοσήμαντα τον χρήστη, β) αντικατοπτρίζει το ιδιολεκτικό του ύφος και γ) προσφέρει τη δυνατότητα σύγκρισης ιδιολεκτικής ομοιότητας μεταξύ των χρηστών - οντοτήτων που παράγουν τα κείμενα.

Τέλος, μια άλλη προσέγγιση είναι να χρησιμοποιηθεί ένα παρόμοιο μοντέλο σαν αυτό του word2vec (είτε με προσέγγιση CBOW είτε με Skip-gram) αλλά να ληφθεί υπόψη η πηγή που παράγει τα κείμενα. Πιο συγκεκριμένα, δίνοντας ως είσοδο στο μοντέλο περικείμενο n λέξεων να συμπεριλαμβάνεται με κάποιο τρόπο ένα χαρακτηριστικό - ετικέτα που να παρέχει στο μοντέλο την πληροφορία ότι η συγκεκριμένη ακολουθία λέξεων προέρχεται από έναν συγκεκριμένο συγγραφέα. Αυτή η τροποποίηση ακριβώς προτείνεται στο μοντέλο Paragraph Vectors (εναλλακτικά Document Vectors ή doc2vec, Le & Mikolov (2014), Dai et al. (2015)).

4.7.2 Σύνδεση με τη θεωρία ανάκτησης πληροφοριών

Θεωρώντας ότι έχουν κατασκευαστεί υφολογικές ενθέσεις μπορούμε να χρησιμοποιήσουμε κάποιο από τα μέτρα που αναφέρονται στην ενότητα 4.5 για να συγκρίνουμε συγγραφείς ως προς το ύφος αλλά και ταξινομήσουμε τα αποτελέσματα. Αυτή η πρακτική απαντά στη θεωρία ανάκτησης πληροφοριών

και πιο συγκεκριμένα στα πιθανοτικά μοντέλα ανάκτησης πληροφορίας. Η αρχή αυτή είναι γνωστή ως Αρχή της Ταξινόμησης κατά Πιθανότητα (Probability Ranking Principle, Van Rijsbergen, 1979) και χρησιμοποιείται για την κατάταξη και ταξινόμηση κατά συνάφεια των εγγράφων d_i δεδομένου ενός ερωτήματος q.

Συγκεκριμένα,

«Αν η απόκριση ενός συστήματος αναφοράς Ανάκτησης [Πληροφοριών] στα ερωτήματα των χρηστών είναι μια ταξινόμηση των κειμένων της συλλογής αυτής με φθίνουσα πιθανότητα σχετικότητας ως προς το ερώτημα, όπου οι πιθανότητες υπολογίζονται με τη μέγιστη δυνατή ακρίβεια από τα δεδομένα που είναι διαθέσιμα στο σύστημα για την επιτέλεση αυτής της λειτουργίας, η συνολική αποτελεσματικότητα του συστήματος ανάκτησης θα είναι η καλύτερη δυνατή που μπορεί να αποκτηθεί με βάση τα συγκεκριμένα δεδομένα» (Van Rijsbergen, 1979, σσ. 113-114).

Κατ' αναλογία, μπορούμε να θεωρήσουμε την υφολογική ομοιότητα ως ερώτημα ανάκτησης πληροφοριών: δεδομένου ενός συγγραφέα $author_1$ μπορούν να ταξινομηθούν κατά φθίνουσα σειρά ομοιότητας με αυτόν οι συγγραφείς του σώματος κειμένων.

4.7.3 Ενθέσεις ύφους μέσω μοντέλων λεξικών ενθέσεων

Το μοντέλο doc2vec έχει την ιδιότητα να μαθαίνει ενθέσεις ύφους ως κομμάτι της αρχιτεκτονικής του. Στην περίπτωση των word2vec, fastText και GloVe το ζητούμενο είναι να βρεθεί τρόπος να κατασκευαστεί αναπαράσταση για καθέναν από τους συγγραφείς με βάση τις ενθέσεις που αντιστοιχούν στις λέξεις των κειμένων τους.

Ένας απλός τρόπος κατασκευής τέτοιων αναπαραστάσεων είναι το άθροισμα ή ο μέσος όρος των διανυσμάτων λέξεων σε μικρό περικείμενο, όπως δείχνουν οι Mikolov et al. (2013). Η εφαρμογή αυτής της τεχνικής ως αναπαράσταση σε επίπεδο πρότασης δίνει εξαιρετικά καλά αποτελέσματα στην ανίχνευση μεταφοράς στα ελληνικά (Φλώρου, 2017).

Μια άλλη μέθοδος, που έχει εξαιρετικά καλά πρακτικά αποτελέσματα, είναι να θεωρηθεί ως διάνυσμα παραγράφου ο μέσος όρος των κανονικοποιημένων σε μοναδιαία νόρμα διανυσμάτων των λέξεων της συλλογής. Κάθε διάνυσμα - ένθεση της συλλογής κανονικοποιείται ώστε να έχει μοναδιαία Ευκλείδεια νόρμα $u \leftarrow u/||u||_2$ και στη συνέχεια υπολογίζεται ο μέσος όρος τους. Η μέθοδος είναι εξαιρετικά δημοφιλής και χρησιμοποιείται στο fastText.

4.7.4 Ανάστροφη Συχνότητα (SIF)

Η μέθοδος της Ομαλής Ανάστροφης Συχνότητας (Smooth Inverse Frequency, SIF) (Arora et al., 2017) είναι μια εναλλακτική μέθοδος παραγωγής διανύσματος

ένθεσης σε επίπεδο συγγραφέα. Αρχικά υπολογίζεται διάνυσμα u_s για κάθε συγγραφέα ως ο σταθμικός μέσος όρος των διανυσμάτων ενθέσεων των λέξεων, που αποτελούν τα κείμενα του συγγραφέα, με S το σύνολο των συγγραφέων, u_w οι ενθέσεις των λέξεων από το λεξικό V και a παράμετρος.

$$u_s \leftarrow \frac{1}{N} \sum_{w \in S} \frac{a}{a + p(w)} u_w$$

Τα διανύσματα των συγγραφέων ορίζουν έναν πίνακα S, ο οποίος παραγοντοποιείται με τη μέθοδο SVD και από την παραγοντοποίηση αυτή λαμβάνεται το πρώτο ιδιόμορφο διάνυσμα (singular vector) u. Στη συνέχεια, η ένθεση του συγγραφέα υπολογίζεται αφαιρώντας από κάθε διάνυσμα u_s την προβολή του στο ιδιόμορφο διάνυσμα:

$$u_s \leftarrow u_s - uu^T u_s$$

4.8 Συμπεράσματα

Στο παρόν κεφάλαιο παρουσιάστηκαν με λεπτομέρεια οι κύριες τεχνικές κατασκευής ενθέσεων μέσω μοντέλων διανυσματικών χώρων. Στην αρχή παρουσιάστηκε η έννοια της διανυσματικής αναπαράστασης όσον αφορά γλωσσικούς όρους και στη συνέχεια οι απαραίτητες διαδικασίες γλωσσικής και μαθηματικής επεξεργασίας για τη δημιουργία των μοντέλων. Στην επόμενη ενότητα παρουσιάστηκε η θεωρητική βάση της κατανεμητικής σημασιολογίας και η έννοια των λεξικών ενθέσεων, στις οποίες βασίζονται τα μοντέλα διανυσματικής αναπαράστασης που χρησιμοποιούνται στην παρούσα διατριβή. Μπορούμε να διακρίνουμε τα μοντέλα αυτά σε δύο κατηγορίες: νευροκατανεμητικά μοντέλα (word2vec, fastText, doc2vec) και παραγοντοποίηση πινάκων συνεμφανίσεων λέξεων (GloVe). Οι μέθοδοι αυτές έχουν ως κύριο χαρακτηριστικό το γεγονός ότι οι παραγόμενες ενθέσεις καταγράφουν σημασιολογικά χαρακτηριστικά των λέξεων ως γραμμικές πράξεις στον διανυσματικό χώρο, στον οποίο απεικονίζονται ως ενθέσεις οι λέξεις του σώματος κειμένων με το οποίο εκπαιδεύονται τα μοντέλα αυτά.

Το ιδιαίτερα σημαντικό χαρακτηριστικό για τους σκοπούς της παρούσας διατριβής είναι ότι οι παραγόμενες ενθέσεις λέξεων μετασχηματίζονται σε ενθέσεις συγγραφέων, με τέτοιο τρόπο ώστε οι νέες παραγόμενες ενθέσεις να διατηρούν τα χαρακτηριστικά που μας ενδιαφέρουν, δηλαδή της ιδιολεκτικής ομοιότητας μεταξύ συγγραφέων. Η υπόθεση, η οποία αιτιολογεί τη στροφή στην κατανεμητική σημασιολογία για την ανίχνευση του ιδιολεκτικού ύφους, είναι το γεγονός ότι το ύφος, εκτός από το λεξικό και συντακτικό συστατικό, φέρει και σημασιολογικά στοιχεία, που δεν έχουν ληφθεί υπόψη στις ως τώρα μελέτες με

τη χρήση μικρού αριθμού συχνών λέξεων και συχνότητας χρήσης μερών του λόγου. Η προτεινόμενη μέθοδος έχει το ιδιαίτερο χαρακτηριστικό ότι ακόμα και στην περίπτωση που οι συγγραφείς χρησιμοποιούν ελαφρώς διαφορετικό αλλά σημασιολογικά παρόμοιο λεξιλόγιο και συντακτικές προτιμήσεις, αυτά τα πρότυπα και οι ατομικές ιδιαιτερότητες θα εντοπιστούν με τη χρήση ενός κατανεμητικού μοντέλου. Στο επόμενο κεφάλαιο θα παρουσιαστούν τα αποτελέσματα από τις εφαρμογές των μοντέλων λεξικών ενθέσεων στα δεδομένα της διατριβής και θα εξαχθούν τα σχετικά αποτελέσματα.

Κεφάλαιο 5

Εκπαίδευση μοντέλων

Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα της εφαρμογής των μοντέλων word2vec, fastText, GloVe και doc2vec στην ανίχνευση της ιδιολέκτου στα δυο σώματα κειμένων, το Σώμα κειμένων Twitter, που συλλέχθηκε στο πλαίσιο αυτής της διατριβής, και στο Σώμα κειμένων ιστολογίων BAC των Schler et al., 2006. Στην ενότητα 5.1 γίνεται λεπτομερής αναφορά στην προεπεξεργασία, που προηγήθηκε της εκπαίδευσης. Στη συνέχεια, στην ενότητα 5.2 αναπτύσσεται η διαδικασία εκπαίδευσης των μοντέλων. Στην ενότητα 5.3 γίνεται εκτενής παρουσίαση του πειράματος που εκτελέστηκε για την επαλήθευση σταθερότητας των μοντέλων και στην ενότητα 5.4 αναλύονται λεπτομερώς τα αποτελέσματα των μοντέλων και για τα δύο σώματα κειμένων.

Στην ενότητα 5.6.1 διερευνάται η συνάφεια των ιδιολεκτικών ενθέσεων και αν και κατά πόσο είναι εφικτή η εξαγωγή συστάδων συγγραφέων με παρόμοιο ιδιολεκτικό ύφος. Για την απάντηση στο ερώτημα αυτό χρησιμοποιούμε τεχνικές Εξόρυξης δεδομένων και Μη επιβλεπόμενης μηχανικής μάθησης. Στόχος είναι η δημιουργία υποομάδων των συγγραφέων με χαρακτηριστικό την ομοιότητα κειμενικού ύφους.

5.1 Προεπεξεργασία

5.1.1 Σώμα κειμένων Twitter

Πριν την εισαγωγή και την εκπαίδευση ήταν απαραίτητη η επεξεργασία του σώματος κειμένων. Σκοπός της προεπεξεργασίας ήταν να διασφαλιστεί το γεγονός ότι το μοντέλο μαθαίνει μόνο από λεξική πληροφορία και αγνοεί πλήρως το υπονοούμενο κοινωνικό δίκτυο (social graph) μεταξύ των αλληλεπιδράσεων των χρηστών. Πιο συγκεκριμένα, στα κείμενα του Σώματος κειμένων Twitter, όπως είδαμε, περιέχονται αναρτήσεις χρηστών αλλά και αλληλεπιδράσεις με

τη μορφή απαντήσεων από άλλους χρήστες. Οι χρήστες ταυτοποιούνται με το σύμβολο @ πριν το όνομα χρήστη και έτσι εάν ο χρήστης @userB συνομιλεί με τον χρήστη @userA στο κείμενο απάντησης περιέχεται η συμβολοσειρά @userA υποδηλώνοντας ότι ο δεύτερος χρήστης απαντά και αλληλεπιδρά με τον πρώτο χρήστη.

Αν αυτές οι συμβολοσειρές που χρησιμοποιούνται ως δεικτοδότηση (indexing) των ονομάτων των χρηστών, συμπεριληφθούν ως λέξεις του λεξιλογίου στο προς εκπαίδευση μοντέλο, τότε στο σώμα εμπεριέχεται η συχνότητα αλληλεπίδρασης ενός χρήστη με άλλους χρήστες. Αυτό λειτουργεί ως λανθάνουσα κωδικοποίηση της γειτνίασης των χρηστών στο κοινωνικό δίκτυο. Αυτό προφανώς είναι ένα γενικότερα επιθυμητό χαρακτηριστικό, ειδικότερα σε εφαρμογές μοντελοποίησης συμπεριφοράς χρηστών σε κοινωνικά δίκτυα και καταγραφής της δυναμικής απόψεων (opinion mining, opinion shifting). Στην παρούσα όμως μελέτη στόχος είναι να ελαχιστοποιηθούν στο δυνατότερο εφικτό βαθμό εξωγλωσσικά δεδομένα ώστε το εκάστοτε μοντέλο να προσπαθήσει να μάθει το ιδιαίτερο ύφος κάθε χρήστη από καθαρό κείμενο και μόνο. Για τον λόγο αυτό, όλες οι συμβολοσειρές που ταυτοποιούν χρήστες, οι οποίες, όπως αναφέρθηκε, έχουν ως πρώτο χαρακτήρα το σύμβολο "@", αντικαταστάθηκαν κατά την προεπεξεργασία με το σύμβολο _USER_ αφαιρώντας έτσι την σύνδεση μεταξύ χρηστών και την πληροφορία για τις ιδιαίτερες προτιμήσεις τους όσον αφορά την αλληλεπίδραση με άλλους χρήστες.

Ένα άλλο σημείο που επίσης εφαρμόστηκε προεπεξεργασία στο Σώμα κειμένων Twitter αφορά τους συνδέσμους στις αναρτήσεις. Σε διάφορα κείμενα οι χρήστες αναρτούν συνδέσμους ιστοσελίδων, συνήθως ενημερωτικών άρθρων, που αφορούν την επικαιρότητα της εποχής ή άλλους γενικότερου ενδιαφέροντος, καθώς και φωτογραφίες. Οι σύνδεσμοι αυτοί αντικαταστάθηκαν με μια κοινή συμβολοσειρά, με το σκεπτικό ότι το κειμενικό περιεχόμενο εξωτερικών συνδέσμων ή η τάση των χρηστών να αναρτούν φωτογραφίες δεν επηρεάζει την ιδιολεκτική τους συμπεριφορά.

Και σε αυτή την περίπτωση, οι ιδιαίτερες προτιμήσεις των χρηστών όσον αφορά τον διαμοιρασμό συνδέσμων μπορούν να χρησιμοποιηθούν ως χαρακτηριστικό σε εφαρμογές, που έχουν στόχο καταγραφή αναλυτικών στοιχείων και τάσεων καθώς και τη δημιουργία προσωπικών και εξατομικευμένων προφίλ χρηστών χωρίς να συμβάλλουν όμως στην καταγραφή προσωπικού κειμενικού ύφους, αλλά μόνο στη γενικότερη καταγραφή συμπεριφοράς χρηστών. Για αυτό τον λόγο οι σύνδεσμοι ιστοσελίδων, που τυπικά ξεκινούν με τους χαρακτήρες http://, επίσης αντικαταστάθηκαν με τη συμβολοσειρά _URL_.

Το σώμα κειμένων περιέχει τα κείμενα των χρηστών ταξινομημένα αλφαβητικά ανά όνομα χρήστη και ώρα ανάρτησης. Στην περίπτωση των μοντέλων word2vec και doc2vec, ο αλγόριθμος εκπαίδευσης μοντέλων λεξικών ενθέσεων είναι τυπικά ο αλγόριθμος Στοχαστικής Κατάβασης κατά Μέγιστη

Κλίση (Stochastic Gradient Descent, SGD). Το αποτέλεσμα της εκπαίδευσης είναι εξαιρετικά ευαίσθητο στη σειρά εισαγωγής δεδομένων στο μοντέλο. Εάν διατηρηθεί η σειρά των δεδομένων εισόδου ταξινομημένη κατά χρήστη και χρόνο ανάρτησης κειμένου, τα μοντέλα χάνουν τη δυνατότητα να γενικεύσουν και παράγουν εξαιρετικά φτωχά αποτελέσματα. Για τον λόγο αυτό, πριν την εισαγωγή και την εκπαίδευση των μοντέλων word2vec, fastText και doc2vec, το σώμα κειμένων τροποποιήθηκε ώστε η σειρά των κειμένων να είναι τυχαία, με στόχο να αποφευχθεί το ενδεχόμενο δύο διαδοχικά κείμενα να ανήκουν στον ίδιο χρήστη ή να ακολουθούν κάποια συγκεκριμένη χρονική σειρά. Το μοντέλο GloVe δεν επηρεάζεται από τη σειρά των κειμένων, καθώς στηρίζεται στην κατασκευή πίνακα συνεμφάνισης λέξων και η σειρά των κειμένων δεν έχει σημασία.

Τέλος, τα κείμενα μετατράπηκαν σε πεζούς χαρακτήρες για να περιοριστεί το εύρος του λεξιλογίου. Αυτό προφανώς έχει ως αποτέλεσμα απώλεια πληροφοριών, ειδικά σε περιβάλλοντα κοινωνικών δικτύων, όπου οι χρήστες είναι εξαιρετικά ευρηματικοί χρησιμοποιώντας αρκετά συχνά κεφαλαίους χαρακτήρες κυρίως για απόδοση συναισθήματος (έμφαση, οργή, έκπληξη, φόβο, απορία). Παρόλα αυτά, και σε αυτή την περίπτωση έγινε συμβιβασμός για να διατηρηθεί το μοντέλο σε υπολογιστικά εφικτά επίπεδα από πλευράς μεγέθους λεξιλογίου. Σε αυτό συνέβαλε και η παρατήρηση ότι χρήστες με συγκεκριμένες προτιμήσεις όσον αφορά τη γραφή και μόνο, όπως αποκλειστική χρήση κεφαλαίων, greeklish ή ιδιαίτερη χρήση σημείων στίξης λ.χ. πολλαπλά θαυμαστικά στο τέλος μιας πρότασης για εκδήλωση έκπληξης (πχ. !!!!!!!), επαναλαμβανόμενα ερωτηματικά (π.χ. ;;;;;), επαναλαμβανόμενες τελείες (π.χ .....) ως αποσιωπητικά, είναι εξαιρετικά ευκολότερο να εντοπιστούν και να κατηγοριοποιηθούν υφολογικά, χρησιμοποιώντας απλές τεχνικές εξόρυξης πληροφοριών από κείμενα.

Εδώ είναι σημαντικό να παρατηρηθεί ότι η συμπεριφορά σε κοινωνικά δίκτυα ακολουθεί συγκεκριμένες συβάσεις, οι οποίες τείνουν να αποτελούν νόρμα και ετικέτα και αφορούν λ.χ. τη μη χρήση κεφαλαίων χαρακτήρων ή greeklish. Σε αρκετές περιπτώσεις, όμως, ομάδες χρηστών -κυρίως μεγαλύτερης ηλικίας ή χαμηλότερου επιπέδου εκπαίδευσης και με χαμηλότερη εξοικείωση με τα κοινωνικά δίκτυα- χρησιμοποιούν αποκλειστικά κεφαλαίους χαρακτήρες στις αναρτήσεις τους. Ωστόσο, σε αυτή την περίπτωση δεν έχουμε άμεση ή έστω σημαντική απώλεια υφολογικής πληροφορίας καθώς η χρήση κεφαλαίων χαρακτήρων δεν γίνεται επιτηδευμένα και συνεπώς δεν μεταφέρει κάποιο επιπλέον υφολογικό επίπεδο επικοινωνίας. Με λίγα λόγια, το υφολογικό περιεχόμενο των αναρτήσεων αυτών διατηρείται με μετατροπή κεφαλαίων χαρακτήρων σε πεζούς χωρίς ιδιαίτερη απώλεια πληροφοριών. Περαιτέρω ανάλυση των περιπτώσεων αυτών θα ήταν ενδιαφέρουσα στο πλαίσιο της κοινωνιογλωσσολογίας υπό την προϋπόθεση ύπαρξης των αντίστοιχων δεδομένων.

5.1.2 Σώμα κειμένων ιστολογίων Blog Authorship Corpus

Στο Σώμα κειμένων ιστολογίων BAC (Schler et al., 2006) η προεπεξεργασία περιορίστηκε στην μετατροπή των χαρακτήρων σε πεζούς χωρίς άλλη ιδιαίτερη επεξεργασία.

5.2 Εκπαίδευση στα σώματα κειμένων

Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκε η βιβλιοθήκη gensim της γλώσσας Python (Rehurek & Sojka, 2011), καθώς και τα εργαλεία Glove[1] και fastText.[2] Η βιβλιοθήκη gensim[3] παρέχει εξαιρετικά αποτελεσματικές από υπολογιστική σκοπιά υλοποιήσεις των word2vec και doc2vec, εκμεταλλευόμενη την παράλληλη χρήση όλων των επεξεργαστών ενός υπολογιστή.

Έπειτα από την προεπεξεργασία, τα κείμενα των δύο σωμάτων κειμένων δόθηκαν ως είσοδος σε κάθε μοντέλο για εκπαίδευση, με διαφορετικές παραμέτρους για την διαδικασία εκπαίδευσης. Πιο συγκεκριμένα, εκπαιδεύτηκαν μοντέλα διαστάσεων στο διάστημα [100, 650] με βήμα 150, δηλαδή $D = 100, 150, 200, 250, ..., 650$ και χρησιμοποιήθηκε σταθερό πλαίσιο περικειμένου μήκους 5 λέξεων ($w = 5$). Η συγκεκριμένη τιμή είναι τυπική σε εφαρμογές περικειμένου, από τις πρώτες μάλιστα βιβλιογραφικές συζητήσεις του μήκους του περικειμένου (Sinclair, 1974). Ο χρόνος εκπαίδευσης κάθε μοντέλου είναι συνάρτηση των παραμέτρων του μοντέλου, των παραμέτρων και της πολυπλοκότητας που τις συνοδεύει. Για παράδειγμα, μεγαλύτερος αριθμός διαστάσεων ένθεσης D έχει ως αποτέλεσμα πιο πολύπλοκο μοντέλο και συνεπώς μεγαλύτερο χρόνο εκπαίδευσης.

Ο τυπικός χρόνος εκπαίδευσης ενός μοντέλου κυμαίνεται από 2 ως 4 ώρες σε υπολογιστή με 8 πυρήνες, 32 Gb μνήμη RAM. Για τον έλεγχο της υπόθεσης ότι τα μοντέλα υπό έλεγχο έχουν την ιδιότητα να καταγράφουν την ομοιότητα μεταξύ συγγραφέων χρησιμοποιήθηκαν αρχικά λογαριασμοί χρηστών από το Σώμα κειμένων Twitter που ήταν εκ των προτέρων γνωστό ότι ανήκουν στο ίδιο άτομο και συνεπώς η υπόθεση θα έπρεπε να ισχύει για τα κείμενα τους.

Δηλαδή, εφόσον τα κείμενα των αναρτήσεων των χρηστών @userA και @userB προέρχονται από το ίδιο άτομο, η υπόθεση ήταν ότι τα παραγόμενα διανύσματα ύφους για αυτούς τους χρήστες θα έχουν πολύ μικρή γεωμετρική απόσταση, ιδανικά τη μικρότερη δυνατή μεταξύ όλων των διανυσμάτων όλων των άλλων χρηστών. Η υπόθεση αυτή επαληθεύτηκε επαναληπτικά σε όλους τους συνδυασμούς μοντέλων και παραμέτρων που αναφέρθηκαν νωρίτερα για τους

[1] https://nlp.stanford.edu/projects/glove/

[2] https://github.com/facebookresearch/fastText

[3] https://radimrehurek.com/gensim/

συγκεκριμένους χρήστες. Συγκρίνοντας τα παραγόμενα διανύσματα των @userA, @userB με τους συνολικά 4.804 διαφορετικούς χρήστες του σώματος κειμένων, το διάνυσμα του χρήστη @userA είχε την μικρότερη απόσταση από το διάνυσμα του @userB σε σχέση με τα διανύσματα όλων των άλλων χρηστών.

Ένα εξαιρετικά ενδιαφέρον αποτέλεσμα της σύγκρισης απόστασης στο διανυσματικό αυτό χώρο είναι ότι τα παραγόμενα μοντέλα έχουν την ιδιότητα να καταγράφουν την σημασιολογική και υφολογική συνάφεια μεταξύ του τρόπου γραφής των χρηστών ως συνάρτηση του κειμενικού περιεχομένου, αλλά και του τρόπου επιλογής λέξεων. Φαίνεται δηλαδή ότι επαληθεύεται και η δεύτερη υπόθεση για την καταγραφή υφολογικής συνάφειας και κατ' επέκταση την ανίχνευση κειμενικής ομοιότητας μεταξύ συγγραφέων, όπως θα δούμε και στη συνέχεια στην ενότητα 5.3.

5.3 Επαλήθευση

Για τη συνολική επαλήθευση των αποτελεσμάτων όσον αφορά τη σταθερότητα των μοντέλων σχεδιάστηκε και εκτελέστηκε το ακόλουθο πείραμα. Δεδομένης της υπόθεσης ότι υπάρχει και μπορεί να καταγραφεί από ένα νευροκατανεμητικό μοντέλο το κειμενικό ύφος- αποτύπωμα με την μορφή *d*-διάστατου διανύσματος, αυτό το αποτύπωμα θα έπρεπε να είναι συνεπές και να είναι ταυτόσημο ή σχεδόν ταυτόσημο σε χρήστες με περισσότερους του ενός λογαριασμούς (όπως είδαμε στην προηγούμενη παράγραφο η υπόθεση αυτή αρχικά φαίνεται να ισχύει). Η επιβεβαίωση αυτής της υπόθεσης έδωσε έναν γρήγορο έλεγχο για το αν και κατά πόσο η μεθοδολογία που προτείνεται είναι εφικτή. Η μη επιβεβαίωση αυτής της υπόθεσης θα ήταν αποτρεπτική για τη συνέχεια εκτέλεσης των πειραμάτων.

Εφόσον η υπόθεση αυτή επιβεβαιώθηκε σε δεδομένα ελέγχου, το επόμενο βήμα ήταν να επιβεβαιωθεί σε ευρεία κλίμακα για όλους τους συγγραφείς των σωμάτων κειμένων. Για τον έλεγχο της υπόθεσης σε ευρεία κλίμακα οι χρήστες και στα δύο σώματα κειμένων ταξινομήθηκαν όσον αφορά την παραγωγικότητα τους, η οποία ορίστηκε ως ο αριθμός κειμένων ανά χρήστη.

Αρχικά τα κείμενα των *C* πιο παραγωγικών χρηστών χωρίστηκαν σε δύο κατηγορίες για κάθε έναν χρήστη @user, θεωρώντας ότι παράγονται από δύο διαφορετικές οντότητες, @userA και @userA_control. Κάθε κατηγορία περιέχει το 50% των κειμένων κάθε χρήστη περίπου. Η επιλογή των κειμένων ανά κατηγορία έγινε με τυχαίο τρόπο και δεν αντανακλά χρονολογική σειρά. Η επιλογή του αριθμού *C* έγινε με τέτοιο τρόπο ώστε να αντικατοπτρίζει με ρεαλιστικό τρόπο την απόδοση των μοντέλων και να διατηρεί τον υπολογιστικό χρόνο εκτέλεσης των πειραμάτων σε λογικά πλαίσια.

Στη συνέχεια, για καθέναν από τους αρχικούς και τους *C* «νέους» συγγραφείς παράχθηκαν οι αντίστοιχες ενθέσεις από όλα τα μοντέλα. Τα διανύσματα *C*

χρησιμοποιούνται για τον έλεγχο σταθερότητας και απόδοσης του μοντέλου. Πιο συγκεκριμένα, ως έλεγχος ορθότητας ορίστηκε η ακρίβεια (accuracy) του μοντέλου όσον αφορά την ταξινόμηση του πιο κοντινού διανύσματος ενός χρήστη από το αρχικό σώμα σε σχέση με τους υπόλοιπους συμπεριλαμβανομένων και των *C* επιπλέον χρηστών ελέγχου. Δηλαδή, εάν ο πιο κοντινός ως προς την γεωμετρική απόσταση στον d-διαστατο διανυσματικό χώρο συγγραφέας στο χρήστη @userA είναι ο @userA_control θεωρούμε επιτυχή την ταξινόμηση. Διαφορετικά την θεωρούμε ανεπιτυχή. Ως συνολικό δείκτη απόδοσης του μοντέλου θεωρούμε το ποσοστό ακρίβειας στο σύνολο ελέγχου. Αυτό το μέτρο ακρίβειας είναι ανάλογο με το ”accuracy @1”, σφάλμα που ακολουθείται στο διαγωνισμό τεχνητής όρασης ImageNet[4].

Αν τώρα θεωρήσουμε επιτυχές αποτέλεσμα το να βρεθεί ο συγγραφέας @userA_control στα 5 πρώτα αποτελέσματα, τότε το μοντέλο αυτό σφάλματος είναι το αντίστοιχο ”accuracy @5” του ImageNet και κατ’ επέκταση, το μοντέλο ”accuracy @k” είναι αυτό στο οποίο θεωρούμε επιτυχή ταξινόμηση εκείνη, όπου ο συγγραφέας ελέγχου βρεθεί στα πρώτα k αποτελέσματα.

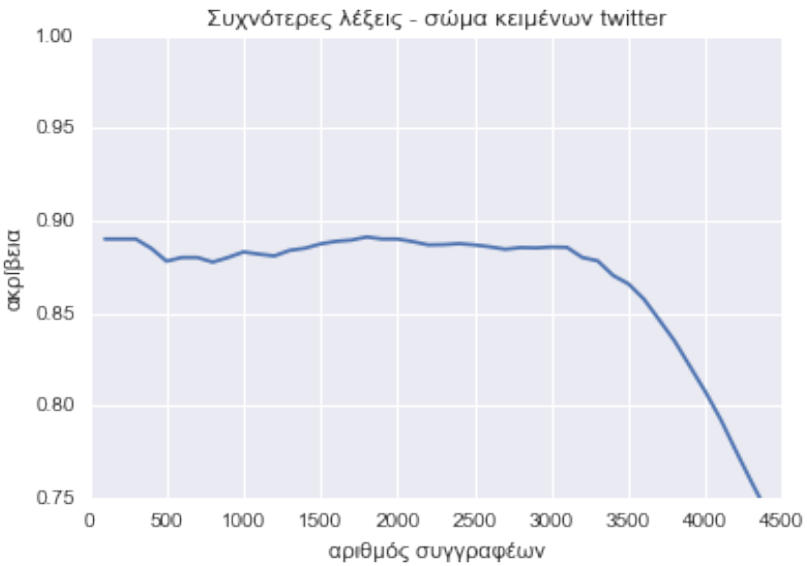
Όπως είδαμε και στο κεφάλαιο 3, η μεθοδολογία που περιγράφεται παραπάνω είναι ασφαλής όσον αφορά τα ποσοτικά χαρακτηριστικά των σωμάτων κειμένων υπό την έννοια ότι υπάρχουν αρκετά κείμενα για την πλειονότητα των συγγραφέων προκειμένου να ελεγχθεί η υπόθεση σταθερότητας των μοντέλων.

Ως σημείο αναφοράς και σύγκρισης των αποτελεσμάτων της παραπάνω μεθόδου, χρησιμοποιήσαμε την μεθοδολογία των Hughes et al. (2012) όσον αφορά την εξαγωγή διανυσμάτων. Για τα κείμενα καθενός από τους @userA και @@userA_control δημιουργήθηκε διάνυσμα με τις 307 πιο συχνές λέξεις και στη συνέχεια ακολουθούμε τη μεθοδολογία που περιγράφεται παραπάνω. Η μέθοδος αυτή, αν και δίνει καλά αποτελέσματα ως προς την ακρίβεια, έχει ένα βασικό μειονέκτημα. Καθώς οι πιο συχνές λέξεις ενός σώματος κειμένων δεν έχουν σημασιολογικό περιεχόμενο (λειτουργικές λέξεις), η μέθοδος αυτή δεν μπορεί να χρησιμοποιηθεί για υφολογική σύγκριση, με την έννοια ότι η σύγκριση των διανυσμάτων αυτών αναφέρεται μόνο στη χρήση των συχνότερων λέξεων και δεν παρέχεται καμία άλλη πληροφορία για τις προτιμήσεις ενός συγγραφέα από το υπόλοιπο λεξιλόγιο.

Συνεπώς, η σύγκριση διανυσμάτων παραγόμενων από συχνές λέξεις, αν και συλλαμβάνουν τη λανθάνουσα χρήση συχνών λέξεων, δεν είναι καλό μέτρο ομοιότητας, γιατί ενδέχεται χρήστες με εντελώς διαφορετική χρήση λεξιλογίου να έχουν παρόμοια χρήση συχνών λέξεων. Η χρήση λοιπόν συχνών λέξεων ως ιδιολεκτικό αποτύπωμα μπορεί να χρησιμοποιηθεί κυρίως ως μέσο απόδοσης συγγραφέα, με την υποσημείωση ότι είναι ενδεχομένως ευαίσθητη σε λάθος ταξινομήσεις (false positives).

[4]http://image-net.org/index

Τα αποτελέσματα της μεθόδου αυτής φαίνονται στα γραφήματα 5.1 και 5.2.



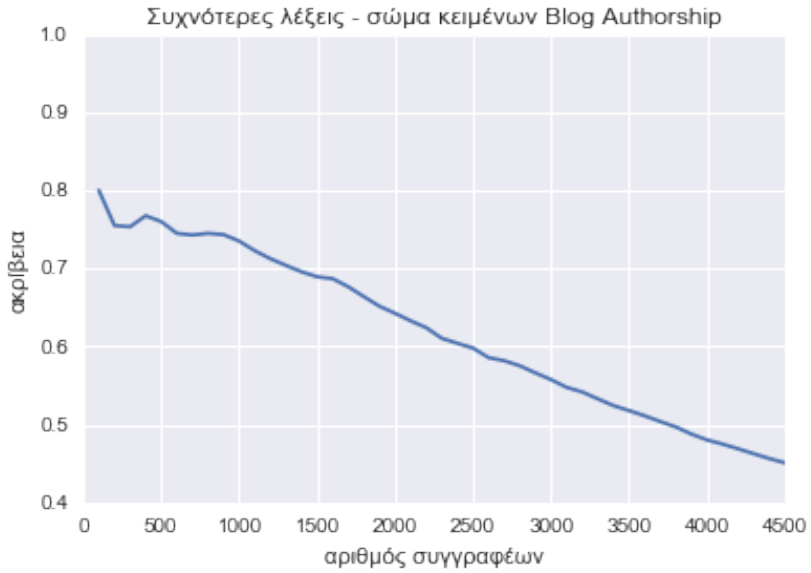
Διάγραμμα 5.1: Ακρίβεια με τη χρήση των 307 πιο συχνών λέξεων - ΣΚ Twitter

Το πείραμα επαναλήφθηκε για τα τέσσερα μοντέλα, word2vec, doc2vec, GloVe και fastText και για τιμές του C στο διάστημα $[100, 4500]$ με βήμα 100, δηλαδή $[100, 200, 300, ..., 4500]$ και για τιμές διαστασιμότητας μεταξύ 100 και 650 με βήμα 50. Τα αποτελέσματα του πειράματος παρουσιάζονται στο παράρτημα Ε.

5.4 Αποτελέσματα

Από τα αποτελέσματα του πειράματος που περιγράφηκε στην προηγούμενη ενότητα μπορούμε να εξάγουμε σημαντικά συμπεράσματα. Αρχικά, παρατηρούμε ότι στην περίπτωση παραγωγικών χρηστών το αποτέλεσμα της ταυτοποίησης είναι εξαιρετικά υψηλό σε όλα τα μοντέλα και σταθερό ως προς την καταγραφή ύφους μεταξύ χρηστών: το διάνυσμα του χρήστη @userX είχε ως πιο κοντινό διάνυσμα το διάνυσμα του χρήστη @userX_control. Με άλλα λόγια, όλα τα μοντέλα αποδίδουν εξαιρετικά σε περιπτώσεις που υπάρχει αρκετό κείμενο, κάτι που αποδεικνύει ότι η απεικόνιση λέξεων με τη μορφή ενθέσεων μπορεί να χρησιμοποιηθεί ως μέσο για την παραγωγή υφολογικού διανύσματος.

Προχωρώντας σε ποιοτική ανάλυση του διαγράμματος στο όριο των 4000 συγγραφέων η ακρίβεια της ταξινόμησης πέφτει στο 85%. Αυτό οφείλεται κυρίως στο γεγονός ότι εφόσον η κατανομή του αριθμού των κειμένων, που έχουν



Διάγραμμα 5.2: Ακρίβεια με τη χρήση των 307 πιο συχνών λέξεων - ΣΚ BAC

παραχθεί από τους χρήστες, ακολουθεί κατανομή Νόμου Δύναμης (Zipf's law, Zipfian distribution) [5] ο αριθμός των κειμένων των χρηστών, που είναι κοντά στις θέσεις 3500-4000, είναι σημαντικά μικρότερος συγκριτικά με τον αριθμό των κειμένων των χρηστών στις πρώτες 2000-2500 θέσεις.

Και τα 4 μοντέλα που εφαρμόστηκαν στο σώμα (GloVe, word2vec, fastText, doc2vec) δίνουν εξαιρετικά καλά και παρόμοια αποτελέσματα και στα 2 σώματα κειμένων που εφαρμόστηκαν. Συνολικά, το μοντέλο doc2vec φαίνεται να είναι το πιο σταθερό και ακολουθεί το word2vec. Και στα δύο σώματα κειμένων, η διαστασιμότητα D που δίνει την καλύτερη ακρίβεια φαίνεται πειραματικά ότι βρίσκεται στην περιοχή $D \in N : D \in [400, 550]$, αποτέλεσμα που βρίσκεται σε συμφωνία με εμπειρικά αποτελέσματα στη βιβλιογραφία (Arora et al. (2015)). Η ακρίβεια στο σώμα κειμένων BAC είναι χαμηλότερη, γεγονός το οποίο ερμηνεύεται από τον μικρότερο αριθμό λέξεων του σώματος καθώς και τον κατά πολύ μεγαλύτερο αριθμό συγγραφέων. Επίσης, φαίνεται ότι το μοντέλο GloVe υστερεί ελαφρώς σε σχέση με τα υπόλοιπα τρία, τα οποία βασίζονται σε νευρωνικές αρχιτεκτονικές, και αυτό ενδεχομένως να οφείλεται σε διαφορές σημασιολογικών χαρακτηριστικών που καταγράφουν τα νευρωνικά μοντέλα σε σύγκριση με το μοντέλο παραγοντοποίησης GloVe.

Παραδείγματα ιδιολεκτικής ομοιότητας από το μοντέλο doc2vec, $D = 500$ δίνονται στο παράρτημα Z Z, για δίαφορες κατηγορίες χρηστών από το σώμα

[5] Στη στατιστική, ένας νόμος δύναμης είναι μια συναρτησιακή σχέση μεταξύ δύο ποσοτήτων, όπου μια ποσότητα μεταβάλλεται ως δύναμη της άλλης.

κειμένων twitter (πολιτική, media κλπ). Αξίζει να σημειωθεί εδώ ότι στην κατηγορία της πολιτικής, φαίνεται να επαληθεύονται τα συμπεράσματα της Φραγκάκη όσον αφορά την ιδιολεκτική ομοιότητα πολιτικών που προέρχονται από το ίδιο κόμμα.

Αν επιχειρηθεί βαθύτερη ανάλυση του αποτελέσματος της ταξινόμησης, βλέπουμε ότι το σφάλμα ταξινόμησης είναι 70% για χρήστες με συνολικό αριθμό λέξεων μικρότερο των 500-600 και μειώνεται στο 10% για χρήστες, οι οποίοι έχουν γράψει κείμενα με περισσότερες από 500-600 λέξεις (Διαγράμματα 5.3).

Η παρατήρηση αυτή είναι κομβική καθώς δίνει μια αίσθηση του ελάχιστου αριθμού λέξεων που απαιτούνται για την εξαγωγή συμπερασμάτων. Για την επαλήθευση αυτής της παρατήρησης εκτελέστηκε το ακόλουθο πείραμα. Από τις λέξεις όλων των κειμένων ενός χρήστη γίνεται δειγματοληψία k λέξεων με τυχαίο τρόπο. Η δειγματοληψία επαναλαμβάνεται 2 φορές και για καθένα από τα δύο σύνολα λέξεων εξάγεται ένα υφολογικό διάνυσμα. Στη συνέχεια υπολογίζεται η ομοιότητα των δύο αυτών διανυσμάτων. Η διαδικασία επαναλαμβάνεται για αυξάνουσες τιμές του k και για μεγάλο αριθμό συγγραφέων.

Το αποτέλεσμα αυτής της διαδικασίας παρουσιάζεται στο διάγραμμα 5.6. Όπως βλέπουμε, η ακρίβεια είναι εν γένει αύξουσα συνάρτηση του αριθμού δείγματος k και συγκλίνει σε ακρίβεια 1.0 μονάδα για τιμές $k \geq 500$.

Η παρατήρηση αυτή φαίνεται πως είναι συνεπής και στο Σώμα κειμένων BAC, καθώς, όπως φαίνεται στο διάγραμμα 5.7, τα αποτελέσματα είναι παρόμοια.

Στη συνέχεια υπολογίζεται ο πρώτος αριθμός δείγματος για τον οποίο η ακρίβεια είναι μεγαλύτερη του 0.995 και ο μέσος όρος για όλους τους συγγραφείς είναι κοντά στις 600 λέξεις με τυπική απόκλιση 100.

Από το διάγραμμα φαίνεται πως για αριθμό λέξεων $x \geq 800$ (μέσος συν 2 τυπικές αποκλίσεις) η ακρίβεια σταθεροποιείται σε τιμές ≥ 0.997.

Προχωρώντας ένα βήμα παραπέρα, αν χαλαρώσουμε τον ορισμό της ακρίβειας ταξινόμησης και θεωρήσουμε το μέτρο acc_K = Σωστή ταξινόμηση χρήστη εάν το διάνυσμα ελέγχου εντοπιστεί στα k πιο κοντινά διανύσματα, βλέπουμε ότι το μοντέλο παράγει συνεπή αποτελέσματα δίνοντας ακρίβεια της τάξης του 95% για k = 10 και πλησιάζει τάχιστα στο 100% για $k \rightarrow 100$. Η σταθερότητα του μοντέλου στις περιπτώσεις k>1 οφείλεται στο γεγονός ότι είναι εξαιρετικά χρήσιμη ιδιότητα σε πρακτικές εφαρμογές: η επιλογή του k περιορίζει δραστικά τον χώρο αναζήτησης σε περιπτώσεις εφαρμογών εύρεσης διπλών λογαριασμών χρηστών, δηλαδή λογαριασμών που έχει ο ίδιος χρήστης (doppelganger effect) (Afroz et. al., 2014).

Ένα παράδειγμα εφαρμογής της παραπάνω μεθοδολογίας σε περιβάλλον δικανικής γλωσσολογίας είναι το ακόλουθο. Έστω ότι έχουμε έναν χρήστη με λογαριασμό @userA, για τον οποίο θα θέλαμε να βρούμε αν διατηρεί άλλο account @userB. Με αυτόν τον τρόπο και υποθέτοντας ότι ο χρήστης δεν έχει συνειδητά αλλάξει εντελώς το ύφος του μεταξύ δυο διαφορετικών λογαριασμών μπορούμε

να μειώσουμε την αναζήτηση διπλών λογαριασμών στους k=10 λ.χ. πιο όμοιους του χρήστη @userA.

5.5 Η επίδραση της διαστασιμότητας στα μοντέλα

Το πείραμα επαναλήφθηκε για διάφορες τιμές του D για όλα τα μοντέλα. Όπως βλέπουμε, για τιμές του D μεταξύ 100-900 η απόδοση των μοντέλων παραμένει περίπου σταθερή με μέγιστη ακρίβεια κοντά στο 0.916 για $D = 500$.

5.6 Ομάδες ιδιολεκτικού ύφους

Στη συγκεκριμένη ενότητα περιλαμβάνονται τα αποτελέσματα της εφαρμογής των αλγορίθμων που αφορούν την πιθανή ομαδοποίηση του ιδιολεκτικού ύφους των συγγραφέων.

5.6.1 Πίνακες ενθέσεων συγγραφέων και ομοιότητας

Όπως είδαμε στις ενότητες 5.1 και 5.3, χρησιμοποιώντας ως πρώτη ύλη τις λεξικές ενθέσεις είναι εφικτό να κατασκευαστούν ιδιολεκτικές ενθέσεις που καταγράφουν με εξαιρετικά υψηλή ακρίβεια το ιδιαίτερο ιδιολεκτικό ύφος των συγγραφέων των σωμάτων κειμένων. Δεδομένου ενός σώματος κειμένων με N συγγραφείς, κατασκευάζονται N διανύσματα διαστασιμότητας D το καθένα, ένα για κάθε συγγραφέα. Έστω $\mathbf{X}$ ο NxD πίνακας όλων των διανυσμάτων όλων των συγγραφέων μπορεί να χρησιμοποιηθεί ως είσοδος σε αλγόριθμους συσταδοποίησης με σκοπό την εύρεση συστάδων με παρόμοιο ύφος. Τέτοιοι αλγόριθμοι συσταδοποίησης είναι οι αλγόριθμοι DBSCAN, KMeans, Hierarchical Clustering.

Από τον πίνακα $\mathbf{X}$, χρησιμοποιώντας το κατάλληλο μέτρο απόστασης ή ομοιότητας, μπορούμε να κατασκευάσουμε έναν τετραγωνικό συμμετρικό πίνακα $\mathbf{S}$, διαστάσεων NxN, στον οποίο κάθε στοιχείο S_{ij} ορίζει αντίστοιχα την ομοιότητα του συγγραφέα i με τον j ή την απόσταση του συγγραφέα i από τον j, ανάλογα με το εφαρμοζόμενο μέτρο απόστασης ή ομοιότητας. Ο πίνακας $\mathbf{S}$ μπορεί να θεωρηθεί επίσης ως το μη κατευθυνόμενο γράφημα ομοιότητας/ανομοιότητας μεταξύ συγγραφέων.

Στη συνέχεια, είτε ο πίνακας $\mathbf{X}$ είτε ο πίνακας $\mathbf{S}$ δίνονται ως είσοδοι στους ανάλογους αλγορίθμους ομαδοποίησης.

Ο πίνακας $\mathbf{S}$ είναι πυκνός (dense) πίνακας, επειδή περιέχει τις ομοιότητες μεταξύ όλων των στοιχείων του πίνακα. Αυτό δεν είναι πάντα μια επιθυμητή ιδιότητα, υπό την έννοια ότι μικρές ομοιότητες μεταξύ συγγραφέων δεν δίνουν

κάποια επιπλέον πληροφορία. Μια λύση στο ζήτημα αυτό είναι να μετατραπεί το πλήρως συνδεδεμένο γράφημα που ορίζει ο πίνακας **S** σε αραιό γράφημα μηδενίζοντας τα βάρη από κάποια τιμή ομοιότητας και κάτω.

Μια εναλλακτική προσέγγιση δίνεται από τους Hughes et al., 2012, κατά την οποία, αντί αυτής της μεθόδου, εφαρμόζεται ανάλυση στατιστικής σημαντικότητας στους κόμβους του γραφήματος με ζητούμενο να βρεθεί η «ραχοκοκαλιά» (backbone) του γραφήματος.

Στην παρούσα διατριβή η ομαδοποίηση των χρηστών/συγγραφέων έχει την ερμηνεία της ιδιολεκτικής κοινότητας σε χρήστες κοινωνικών δικτύων. Επιπλέον, η ανάλυση της δομής του γραφήματος/δικτύου ομοιότητας παρέχει πληροφορίες για τους ισχυρούς κόμβους του γραφήματος, δηλαδή τους χρήστες/συγγραφείς που φαίνεται να έχουν τάση να έχουν ισχυρή ομοιότητα με πολλούς άλλους συγγραφείς.

5.6.2 Ανάλυση δικτύου

Δεδομένου ενός γραφήματος, ένα ερώτημα που ακολουθεί είναι το ποιος είναι ο πιο σημαντικός κόμβος του γραφήματος αυτού. Η έννοια της σημαντικότητας δεν είναι μονοσήμαντη: ένας κόμβος μπορεί να είναι σημαντικός αν έχει πολλές εισόδους/εξόδους ή αν έχει σημαντικούς γείτονες. Επιπλέον μπορεί να είναι σημαντικός γιατί ενώνει διαφορετικά τμήματα του δικτύου.

Η σημαντικότητα ενός κόμβου (node) όσον αφορά τις εισόδους/εξόδους του ονομάζεται *βαθμός* του γραφήματος και τυπικά ορίζεται ως ο αριθμός των ακμών που συνδέονται με αυτόν. Στην προκειμένη περίπτωση τα παραγόμενα γραφήματα περιέχουν τις ομοιότητες των συγγραφέων/χρηστών ως βάρη και συνεπώς θα μπορούσαμε να χρησιμοποιήσουμε και αυτή την πληροφορία ως μέτρο βαθμού ενός κόμβου. Στην περίπτωση αυτή ο βαθμός του γραφήματος είναι απλά το άθροισμα των βαρών των ακμών που οδηγούν σε αυτόν τον κόμβο.

5.6.3 Αλγόριθμος DBSCAN

Ο αλγόριθμος DBSCAN (Ester et. al., 1996) είναι από τους πλέον διαδεδομένους αλγόριθμους ομαδοποίησης (συσταδοποίησης). Όπως όλοι οι αλγόριθμοι συσταδοποίησης είναι αλγόριθμος μη επιβλεπόμενης μάθησης (unsupervised learning). Είσοδός του είναι ο πίνακας ομοιότητας και έξοδος οι ομάδες που ανιχνεύτηκαν στα δεδομένα εισόδου. Ο αλγόριθμος DBSCAN έχει το χαρακτηριστικό ότι ανιχνεύει αυτόματα τον αριθμό των ομάδων στα δεδομένα εισόδου, σε αντίθεση με άλλους αλγόριθμους ομαδοποίησης, όπως π.χ. ο αλγόριθμος K-Μέσων (K-means, Murphy, 2012). Η ιδιότητά του αυτή είναι εξαιρετικά χρήσιμη καθώς στις περισσότερες περιπτώσεις δεν είναι εκ των προτέρων γνωστός ο αριθμός των ομάδων σε ένα σύνολο δεδομένων.

5.6.4 Αποτελέσματα συσταδοποίησης

Τα αποτελέσματα συσταδοποίησης των πιο σημαντικών κόμβων παρουσιάζονται στο παράρτημα Ε.3. Για την εγκυρότητα των συστάδων στο Σώμα κειμένων Twitter χρησιμοποιήθηκαν λογαριασμοί Ελλήνων πολιτικών και μέσων μαζικής ενημέρωσης. Για καθένα από τα μοντέλα παράγεται ο πίνακας ομοιότητας **S** και στη συνέχεια δίνεται ως είσοδος στον αλγόριθμο DBSCAN. Οι παραγόμενες συστάδες ταξινομούνται σε φθίνουσα σειρά ως προς τη σημαντικότητά τους (node centrality).

Μια εξαιρετικά ενδιαφέρουσα παρατήρηση είναι ότι οι αλγόριθμοι ανάλυσης δικτύου φαίνεται να λειτουργούν εξαιρετικά καλά ως μηχανισμός ανίχνευσης διασποράς ειδήσεων, προπαγάνδας και διάχυσης παρόμοιου υφολογικού περιεχομένου. Πιο συγκεκριμένα, οι αλγόριθμοι ανίχνευσαν μικρές υποομάδες χρηστών με παρόμοιο ύφος, με πολύ κλειστό ιδεολογικό πυρήνα.

Οι αλγόριθμοι ανάλυσης δικτύου μπορούν να χρησιμοποιηθούν ως ανίχνευση της κεντρικής οντότητας, που -είτε εσκεμμένα είτε ακούσια- τείνουν να αντιγράφουν υφολογικά μέλη της ομάδας αυτής.

5.7 Χρονική σταθερότητα ιδιολεκτικού αποτυπώματος

Ένα επιπλέον ερώτημα που μας απασχολεί σε αυτό το κεφάλαιο είναι αν και κατά πόσο μεταβάλλεται το υφολογικό αποτύπωμα ως συνάρτηση του χρόνου. Στην ενότητα 5.7.1 ορίζονται τα μέτρα σταθερότητας και στη συνέχεια γίνεται εφαρμογή στο Σώμα κειμένων Twitter και στο Σώμα κειμένων Blog Authorship Corpus (5.7.4) .

5.7.1 Μέτρα σταθερότητας

Όπως είδαμε μέχρι στιγμής, είναι εφικτή η αναπαράσταση του κειμενικού ύφους ως d-διάστατο διάνυσμα. Στην περίπτωση μελέτης σταθερότητας (ή, από διαφορετική σκοπιά, της ποσοτικοποίησης της μεταβολής του κειμενικού ύφους ενός συγγραφέα στο πέρασμα του χρόνου) το ζητούμενο είναι να κατασκευαστούν οι αντίστοιχες ενθέσεις για κάθε συγγραφέα στο υπό εξέταση χρονικό διάστημα και στη συνέχεια να ποσοτικοποιηθεί η διαφορά των διανυσμάτων αυτών ως συνάρτηση του χρόνου. Για παράδειγμα, υποθέτοντας ότι είναι διαθέσιμος ικανός όγκος κειμενικών δεδομένων μπορούν να κατασκευαστούν ενθέσεις για τον συγγραφέα σε επίπεδο δεκαετίας, έτους, μήνα κ.ο.κ.

Έτσι, για την περίοδο $t \in \{1, 2, ..., T\}$ στους N συγγραφείς αυτής της περιόδου αντιστοιχούν τα διανύσματα $v_{t_1}, v_{t_2}, ..., v_{t_N}$

5.7.2 Μέση απόσταση συνημιτόνου

Θεωρούμε ότι για ένα συγγραφέα σε ένα σώμα κειμένων με N διαφορετικούς συγγραφείς έχουν δημιουργηθεί ιδιολεκτικά διανύσματα για T χρονικές περιόδους $t_1, t_2, ..., t_T$. Οι χρονικές περίοδοι αντιστοιχούν σε έτη, μήνες ή κάποιο άλλο, συνήθως σταθερό, χρονικό διάστημα.

Ένα αριθμητικό μέτρο, που μπορεί να θεωρηθεί ως εκτίμηση σταθερότητας, είναι ο μέσος όρος των αποστάσεων των ιδιολεκτικών διανυσμάτων του συγγραφέα αυτού στα ζεύγη διαδοχικών χρονικών διαστημάτων.

Αν θεωρήσουμε δηλαδή την απόσταση συνημιτόνου $d(v_i, v_j)$ μεταξύ δύο διανυσμάτων v_i, v_j, η μέση απόσταση μεταξύ όλων ζευγών διανυσμάτων που αντιστοιχούν σε διαδοχικές χρονικές περιόδους t_i, t_{i+1}

$$AvDistance(\mathbf{u}) = \frac{1}{T-1} \sum_{t=1}^{T-1} d(u_t, u_{t+1}) \tag{5.1}$$

Το μέτρο παρέχει μια ποσοτική αίσθηση της σταθερότητας ιδιολεκτικού διανύσματος στο πέρασμα του χρόνου και δίνει επίσης τη δυνατότητα ιεράρχησης των συγγραφέων του σώματος κειμένων με βάση τη συνολική σταθερότητά τους, κατ' αναλογία με την Αρχή της Ταξινόμησης κατά Πιθανότητα των Van Rijsbergen (1979), όπως είδαμε στο κεφάλαιο 4.

Το μέτρο μέσης απόστασης συνημιτόνου λειτουργεί εξαιρετικά καλά σε περιπτώσεις που οι χρονικές περίοδοι είναι καλά ορισμένες, όταν δηλαδή υπάρχει αρκετό κειμενικό υλικό για κάθε συγγραφέα σε περίοδο t_i. Σε περιπτώσεις που το κειμενικό υλικό δεν είναι αρκετό, τυπικά τουλάχιστον 500-600 λέξεις, όπως προκύπτει από την ανάλυση της παραγράφου 5.4, το ιδιολεκτικό διάνυσμα που αντιστοιχεί στην περίοδο αυτή ενδέχεται να έχει σημαντική απόσταση σε σχέση με προηγούμενες ή και επόμενες χρονικές περιόδους και αυτό επηρεάζει αρνητικά την αξιοπιστία του.

Η αδυναμία αυτή μπορεί να διορθωθεί αν εξαιρεθούν από το μέτρο τα διανύσματα εκείνα που αντιστοιχούν σε χρονικές περιόδους με περιορισμένο κειμενικό υλικό ή, εναλλακτικά, αν αντί του μέσου όρου χρησιμοποιηθεί η διάμεσος τιμή των αποστάσεων αυτών, καθώς η διάμεσος τιμή δεν είναι ευαίσθητη στην ύπαρξη ακραίων τιμών.

5.7.3 Μέση ακρίβεια @k

Το μέτρο της μέσης απόστασης που παρουσιάστηκε στην προηγούμενη παράγραφο, αν και δίνει μια συνολικά καλή εικόνα όσον αφορά την ιεράρχηση και την ταξινόμηση της σταθερότητας του ύφους, δεν δίνει συνολική αίσθηση του πόσο καλά λειτουργεί η κατάτμηση των κειμένων των συγγραφέων σχετικά με την ευαισθησία του μέτρου σε περιπτώσεις σφάλματος ταξινόμησης.

Στις περιπτώσεις εκείνες που ο κειμενικός όγκος των συγγραφέων του σώματος κειμένων είναι μεγαλύτερος από 500-600 λέξεις, για το διάστημα της χρονικής περιόδου που επιλέχθηκε μπορεί να επαναληφθεί η διαδικασία που ορίστηκε στην ενότητα 5.4, διασπώντας τα κείμενα του συγγραφέα σε δύο τυχαία υποσύνολα με τον ίδιο αριθμό λέξεων κατά προσέγγιση.

Στη συνέχεια, υπολογίζεται η ακρίβεια ταξινόμησης @k, όπως ορίστηκε στην ενότητα 5.4, για κάθε μία από τις χρονικές περιόδους t_i.

Αν δηλαδή στη χρονική περίοδο t_i αντιστοιχούν N συγγραφείς, παράγονται $2N$ συγγραφείς με τα μισά κείμενα και ελέγχεται αν το διάνυσμα του συγγραφέα $A\_Control_j$ ανήκει στο σύνολο των k-κοντινότερων διανυσμάτων του συγγραφέα A_j. Σε περίπτωση που ανήκει, αυτό θεωρείται θετική ταξινόμηση. Σε αντίθετη περίπτωση θεωρείται αρνητική. Τέλος, υπολογίζεται η ακρίβεια ταξινόμησης acc_{t_i} για αυτή τη χρονική περίοδο t_i και στη συνέχεια η μέση ακρίβεια για τις χρονικές περιόδους $t_i, , i \in \{1,2,...T\}$.

$$AvAccuracy = \frac{1}{T}\sum_{t=1}^{T} acc_{t_i} \tag{5.2}$$

5.7.4 Εφαρμογή στα σώματα κειμένων

Όπως είδαμε στην ενότητα 5.4, διαιρώντας τα κείμενα ενός συγγραφέα σε δύο ισομεγέθη τυχαία υποσύνολα, διαπιστώνουμε ότι οι ενθέσεις ιδιολέκτου στα υποσύνολα αυτά βρίσκονται πάρα πολύ κοντά στο γεωμετρικό χώρο για κείμενα με περισσότερες από 500-600 περίπου λέξεις. Από τα σώματα κειμένων που εξετάζουμε μπορούμε να χρησιμοποιήσουμε το έτος κάθε ανάρτησης και συνεπώς να επαναλάβουμε το πείραμα για τα έτη που κάθε χρήστης είναι ενεργός. Μπορούμε, επίσης, χρησιμοποιώντας τα μέτρα που ορίστηκαν στην ενότητα 5.7.1, να κατατάξουμε τους χρήστες ως την υφολογική τους σταθερότητα.

Πιο συγκεκριμένα, για κάθε χρήστη $\mathbf{u}$ υπολογίζεται το μέτρο $AvDistance(\mathbf{u})$ και στη συνέχεια οι χρήστες ταξινομούνται σε φθίνουσα σειρά μέσης ανά χρονικό διάστημα απόστασης. Η ταξινόμηση αυτή μας δίνει μια αίσθηση της συνολικής ιδιολεκτικής σταθερότητας ενός χρήστη.

Τα αποτελέσματα της εφαρμογής της μεθοδολογίας που αναπτύσσουμε στο παρόν κεφάλαιο δίνουν ιδιαίτερα ενθαρρυντικά αποτελέσματα ως προς το μέτρο ιδιολεκτικής σταθερότητας, σε αντιστοιχία με τα αποτελέσματα σταθερότητας υφολογικού διανύσματος. Ωστόσο, επειδή η μέτρηση γίνεται σε μικρότερο αριθμό κειμένων και η άμεση σύγκριση γίνεται με υποσύνολα των κειμένων των συγγραφέων, υπάρχει διακύμανση στην ακρίβεια των αποτελεσμάτων.

Η εφαρμογή αυτών των μετρικών για την ποσοτικοποίηση της μεταβολής ύφους παρέχει σημαντικά εργαλεία που μπορούν να χρησιμοποιηθούν σε πιο σύνθετα προβλήματα εφαρμογών δικανικής γλωσσολογίας, καθώς επίσης και σε

προβλήματα συσχέτισης κειμενικού ύφους και προσωπικότητας. Στο κεφάλαιο 6 γίνεται εκτενής αναφορά σε πιθανές εφαρμογές της μεθοδολογίας αυτής σε περιβάλλον κοινωνικών δικτύων.

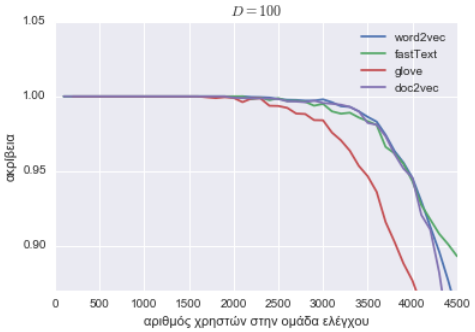
5.8 Συμπεράσματα

Στο κεφάλαιο αυτό προχωρήσαμε στην εκμάθηση ενθέσεων ύφους συγγραφέων στα δύο σώματα κειμένων. Τα αποτελέσματα δείχνουν ότι είναι εφικτή η αναπαράσταση ύφους ως διανυσματική αναπαράσταση. Οι αναπαραστάσεις αυτές δίνουν τη δυνατότητα σύγκρισης υφολογικού αποτυπώματος και συνεπώς η υφολογική ομοιότητα μπορεί να αντιμετωπιστεί ως πρόβλημα Ανάκτησης πληροφοριών.

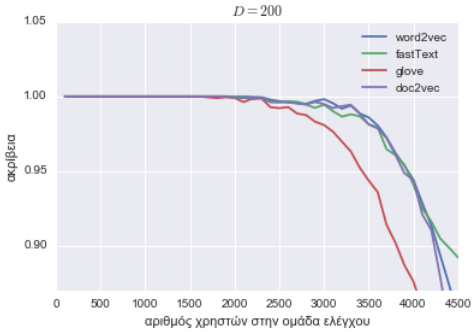
Τα μοντέλα που δοκιμάστηκαν δίνουν παρόμοια χαρακτηριστικά, με το μοντέλο doc2vec να υπερτερεί ως προς την σταθερότητα στην ακρίβεια πρόβλεψης.

Στη συνέχεια, διερευνήθηκε η σταθερότητα της αναπαράστασης του κειμενικού ύφους ως συνάρτηση του χρόνου, ορίστηκαν τα ανάλογα μέτρα σταθερότητας και εφαρμόστηκαν στα σώματα κειμένων. Από την εφαρμογή αυτή προκύπτουν χρήσιμα συμπεράσματα, καθώς και πρακτικές εφαρμογές, όπως, για παράδειγμα, η ταξινόμηση των συγγραφέων κατά σταθερότητα ύφους ως συνάρτηση του χρόνου.

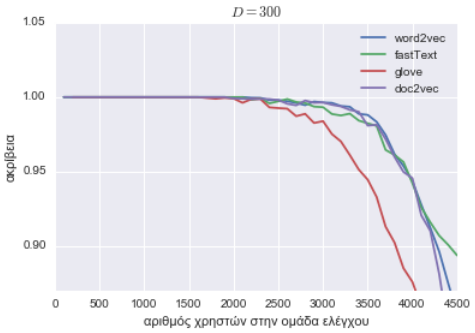
Επιπλέον, καθώς οι παραγόμενες ενθέσεις αντιστοιχούν σε μοντέλο διανυσματικών χώρων και καθώς έχει οριστεί μέτρο απόστασης μεταξύ δύο αναπαραστάσεων (απόσταση συνημιτόνου) είναι εφικτή η εφαρμογή αλγορίθμων συσταδοποίησης για την εύρεση ομάδων συγγραφέων με παρόμοιο ύφος, αλλά και εφαρμογές ανάλυσης δικτύων, όπως, για παράδειγμα, η εύρεση των πιο σημαντικών οντοτήτων (συγγραφέων) στο παραγόμενο γράφημα απόστασης.



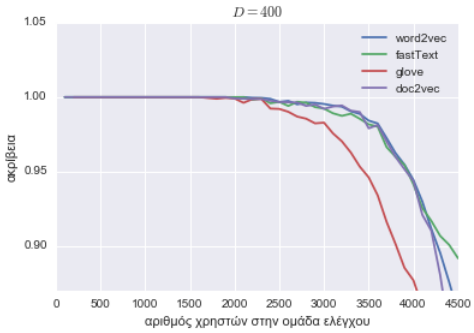
(a) $D = 100$



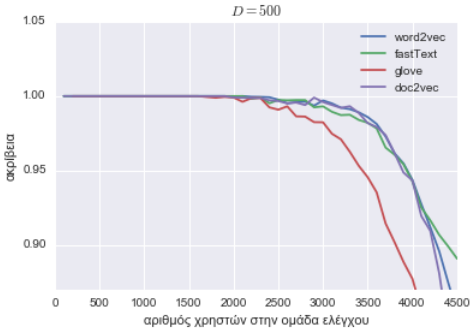
(b) $D = 200$



(c) $D = 300$

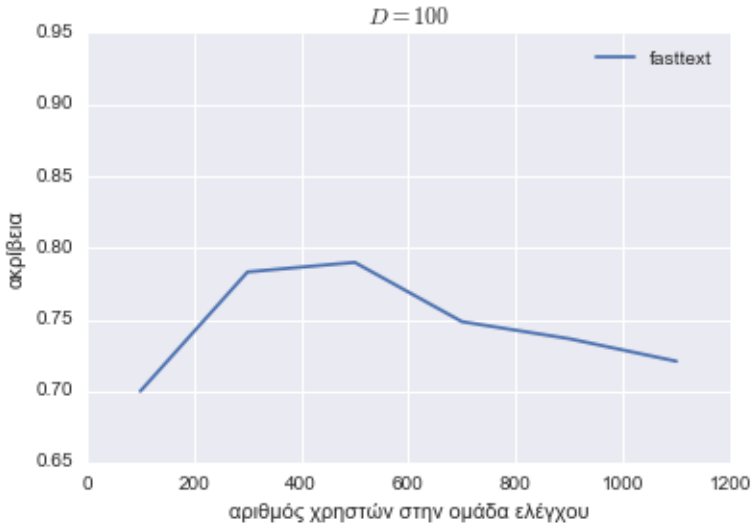


(d) $D = 400$

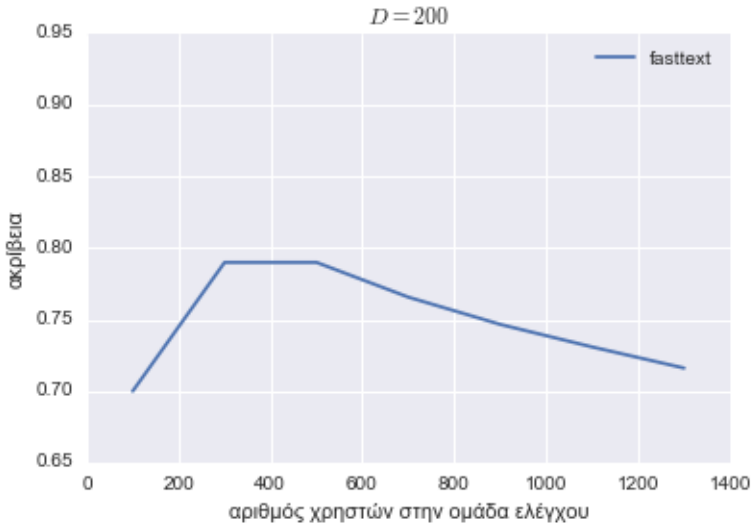


(e) $D = 500$

Διάγραμμα 5.3: Απόδοση Μοντέλων στο Σώμα Twitter

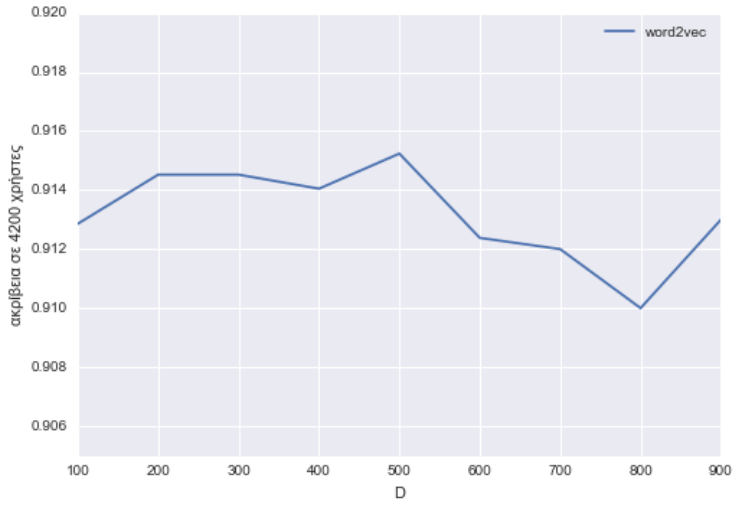


(a) $D = 100$



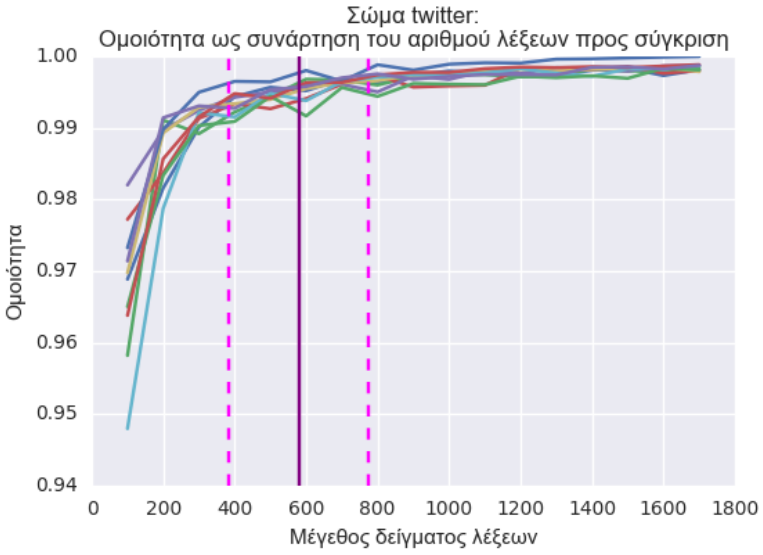
(b) $D = 200$

Διάγραμμα 5.4: Απόδοση Μοντέλων στο Σώμα κειμένων BAC

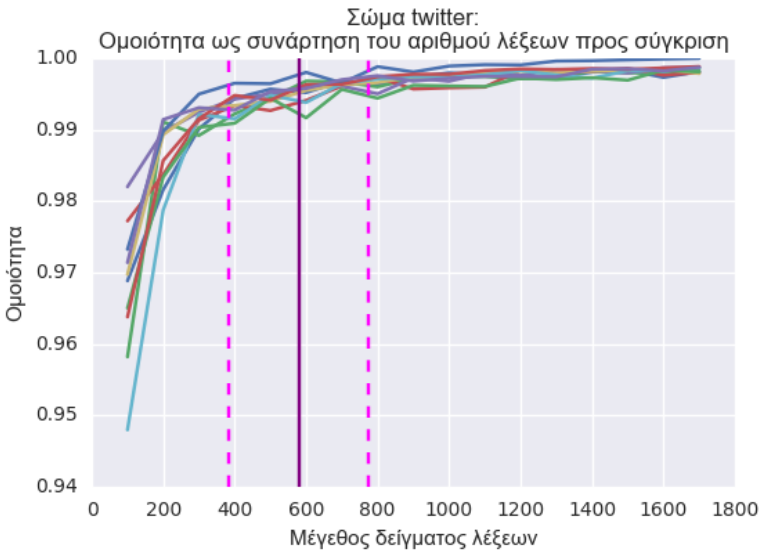


(a) $D = 100$

Διάγραμμα 5.5: Συμπεριφορά μοντέλου για διάφορες τιμές του D



Διάγραμμα 5.6: Ακρίβεια ως συνάρτηση αριθμού λέξεων - Σώμα κειμένων Twitter



Διάγραμμα 5.7: Ακρίβεια ως συνάρτηση αριθμού λέξεων - Σώμα κειμένων BAC

Κεφάλαιο 6

Συμπεράσματα και προεκτάσεις

Στο κεφάλαιο αυτό συνοψίζουμε τα αποτελέσματα της παρούσας διατριβής ως προς τα ερευνητικά ερωτήματα της διατριβής, που αφορούν την ανίχνευση ύφους (6.1.1), τη σταθερότητα ύφους στο χρόνο (6.1.3) και την ομαδοποίηση κατά παρόμοιο ύφος (6.1.2). Στη συνέχεια αναφέρονται πιθανές εφαρμογές της διατριβής σε πρακτικά προβλήματα (6.2). Επίσης, παρουσιάζονται κατευθύνσεις για μελλοντική έρευνα (6.3), που βασίζονται στους αναγκαστικούς περιορισμούς αυτής της διατριβής.

6.1 Συμπεράσματα

Στην ενότητα αυτή συνοψίζονται τα αποτελέσματα της παρούσας διατριβής που αφορούν τα ερευνητικά ερωτήματα τα οποία τέθηκαν στο Κεφάλαιο 1:

1. Είναι εφικτή η ανίχνευση της ιδιολέκτου με υπολογιστικές μεθόδους;
2. Μπορεί η υπολογιστική προσέγγιση στην ανίχνευση της ιδιολέκτου να χρησιμοποιηθεί για την ομαδοποίηση συγγραφέων με βάση το ύφος;
3. Διαπιστώνεται και, αν ναι, σε ποιο βαθμό μεταβολή στην ιδιολεκτική συμπεριφορά των συγγραφέων στο πέρασμα του χρόνου;

6.1.1 Ενθέσεις Ιδιολέκτου

Το βασικό ερευνητικό ερώτημα που τέθηκε στην παρούσα διατριβή ήταν αν και κατά πόσο είναι εφικτή η υπολογιστική αναπαράσταση της ιδιολέκτου. Σε αντίθεση με τις έως τώρα μελέτες, η παρούσα διατριβή διαφέρει τόσο στη βάση της μεθοδολογίας (λεξικές ενθέσεις αντί συχνοτήτων λέξεων και ν-γραμμάτων μερών του λόγου) όσο και στον αριθμό των λέξεων του λεξιλογίου

που χρησιμοποιείται ως βάση για την ποσοτικοποίηση της ιδιολέκτου (όλες οι λέξεις του σώματος αντί μερικών εκατοντάδων υψίσυχνων λέξεων).

Η χρήση εκτεταμένου λεξιλογίου ως μέτρου της ποσοτικοποίησης της ιδιολέκτου και η άμεση σχέση των νευροκατανεμητικών μοντέλων με την κατανεμητική θεωρία επιβεβαιώνουν πειραματικά την υπόθεση ότι η ιδιόλεκτος φέρει σημασιολογικό φορτίο, δηλαδή συνδέεται με λέξεις περιεχομένου. Οι (ενσυνείδητες ή υποσυνείδητες) επιλογές ενός συγγραφέα όσον αφορά το γενικότερο λεξιλόγιό του χαρακτηρίζουν μονοσήμαντα το ύφος του, το οποίο είναι δυνατόν να ποσοτικοποιηθεί με τη μορφή διανυσματικής αναπαράστασης.

Χρησιμοποιώντας την μεθοδολογία αυτή παράγονται ιδιολεκτικές ενθέσεις για το ύφος των χρηστών κοινωνικών δικτύων. Οι λεξικές αυτές ενθέσεις είναι εξαιρετικά σταθερές όσον αφορά τη διαφοροποίηση ύφους μεταξύ χρηστών και μπορούν να χρησιμοποιηθούν και για την ομαδοποίηση χρηστών με βάση το κειμενικό τους ύφος.

Από την ανάλυση της απόδοσης των μοντέλων προκύπτει ότι το κομβικό σημείο για την αξιοπιστία της απεικόνισης του ιδιολεκτικού ύφους ενός συγγραφέα είναι η ύπαρξη κειμένων του με περισσότερες από 500-600 περίπου λέξεις.

Συνοψίζοντας, η μεθοδολογία που προτείνεται στην παρούσα διατριβή καλύπτει τα ερωτήματα που τέθηκαν όσον αφορά τη χρήση του λεξιλογίου, την ομαδοποίηση κατά ύφος και την ιδιολεκτική σταθερότητα στο χρόνο. Τα παραγόμενα αποτελέσματα μελετούν το ύφος χιλιάδων συγγραφέων σε σώματα κειμένων αποτελούμενων από εκατοντάδες εκατομμύρια λέξεις, που προέρχονται από λεξικά εκατοντάδων χιλιάδων λέξεων.

Δείξαμε ότι είναι εφικτό να καταγραφεί η ομοιότητα του ιδιολεκτικού διανύσματος με τη χρήση ποσοτικών μέτρων, της γωνίας δηλαδή των διανυσμάτων ιδιολεκτικών ενθέσεων. Χρησιμοποιώντας αυτόν τον ορισμό απόστασης, προχωρήσαμε σε εύρεση συστάδων συγγραφέων, δηλαδή σε εντοπισμό κοινών χαρακτηριστικών των συγγραφέων ως προς το ύφος, τα οποία συνιστούν «ομάδα».

Τέλος, έγινε ποσοτικοποίηση της ιδιολεκτικής σταθερότητας, που αναφέρεται στο πόσο σταθερό είναι το ιδιολεκτικό ύφος ενός συγγραφέα στην πάροδο του χρόνο.

Συνοψίζοντας, η βασική συνεισφορά της παρούσας διατριβής έγκειται στην ανάπτυξη της έννοιας των ιδιολεκτικών ενθέσεων και την εφαρμογή τους στην ανάλυση της ιδιολέκτου σε ελληνικά δεδομένα. Κατ' αντιστοιχία με τις λεξικές ενθέσεις, οι οποίες αποτελούν αφηρημένες μαθηματικές αναπαραστάσεις των κειμένων σε επίπεδο λέξεων που εμπεριέχουν μετρήσιμα σημασιολογικά χαρακτηριστικά, οι παραγόμενες ιδιολεκτικές ενθέσεις μπορούν να θεωρηθούν ως εξατομικευμένη αφηρημένη αναπαράσταση της υφολογικής περιοχής στην οποία αναμένεται να κινηθεί ένας συγγραφέας με βάση το κειμενικό του

ιστορικό και σε συνάρτηση με τη σημασιολογική περιοχή από την οποία επιλέγει να αντλεί λέξεις, διαμορφώνοντας το υπό εξέταση κειμενικό ύφος. Η διαδικασία επιλογής σημασιολογικής περιοχής καθορίζεται από παραμέτρους σχετιζόμενες με το κειμενικό είδος, το χρονικό περιβάλλον και την προθετικότητα του δημιουργού. Το παραγόμενο ιδιολεκτικό ύφος με τη μορφή υφολογικής ένθεσης λοιπόν είναι η μαθηματική αναπαράσταση του λεξιλογίου του ως συνάρτηση των παραμέτρων που καθορίζουν την διαδικασία επιλογής λέξεων από κάθε συγγραφέα, του ατομικού ιδιολεκτικού του γονιδιώματος, για να χρησιμοποιήσουμε μια μεταφορά.

6.1.2 Συστάδες ύφους

Όπως είδαμε, το αποτέλεσμα της εκπαίδευσης μοντέλων ύφους από μαθηματική σκοπιά αντιστοιχεί σε έναν πίνακα διαστάσεων *NxD*, με τον αριθμό των συγγραφέων σε ένα σώμα κειμένων και *D* τον αριθμό των διαστάσεων. Αυτός ο πίνακας, στη συνέχεια, μπορεί να αποτελέσει τη βάση για εφαρμογή αλγορίθμων ομαδοποίησης μη επιβλεπόμενης μηχανικής μάθησης και τεχνικών εξόρυξης δεδομένων, συνήθως με στόχο την κατασκευή υποσυνόλων-υποομάδων με μεγάλο βαθμό ομοιότητας.

6.1.3 Κειμενική σταθερότητα

Στην περίπτωση της κειμενικής σταθερότητας, η υπόθεση εργασίας είναι ότι το ύφος ενός συγγραφέα παραμένει σταθερό στο πέρασμα του χρόνου. Εφαρμόζοντας την τεχνική που αναπτύξαμε, η υπόθεση φαίνεται να επιβεβαιώνεται. Δεδομένου ότι η εκπαίδευση μοντέλων ύφους με παράμετρο το χρόνο έχει ως αποτέλεσμα τη μείωση του απόλυτου αριθμού για κάθε ζεύγος (συγγραφέα, έτος), η εμπειρική αυτή παρατήρηση χρειάζεται να επαληθευτεί σε μεγαλύτερο όγκο δεδομένων και αριθμό συγγραφέων. Η ανάλυση εδώ έγινε μετρώντας την υφομετρική συνάφεια με μονάδα μέτρησης το ημερολογιακό έτος. Στην περίπτωση περισσότερων κειμένων και συγγραφέων, το επόμενο βήμα ανάλυσης είναι η ποσοτικοποίηση και διερεύνηση της υφολογικής σταθερότητας σε μικρότερα χρονικά διαστήματα (υποθέτοντας ότι υπάρχει συνεχής και ικανή σε αριθμό ροή κειμένων από τον συγγραφέα κατά τα συγκεκριμένα διαστήματα).

Πέρα από το υπολογιστικό κομμάτι, ένα ερώτημα που χρήζει περαιτέρω διερεύνησης είναι οι λόγοι που ωθούν έναν συγγραφέα σε αλλαγή ύφους ή τρόπου γραφής. Για το ερώτημα αυτό μια πιθανή ερευνητική κατεύθυνση θα ήταν η απόπειρα συσχετισμού υφολογικού αποτυπώματος με την ταξινόμηση προσωπικότητας και τη μεταβολή τους στο χρόνο. Σχετική έρευνα στο πλαίσιο της δικανικής γλωσσολογίας έχει γίνει από την Luyckx (Luyckx & Daelemans, 2008).

6.2 Εφαρμογές

Η αξιοποίηση μοντέλων υφολογικής ομοιότητας αφορά διάφορες δυνητικές εφαρμογές ανάλογα με το προς εξέταση σώμα κειμένων, καθώς επίσης και με το περιβάλλον στο οποίο ανήκει το υπό εξέταση κειμενικό σώμα. Θεωρώντας ένα ζεύγος (μοντέλο λεξικών ενθέσεων, σώμα κειμένων) ως σύστημα ανάκτησης πληροφοριών, μπορούμε να κωδικοποιήσουμε ερωτήματα ως ενθέσεις και στη συνέχεια να κατατάξουμε τα κείμενα του σώματος με φθίνουσα σειρά ομοιότητας. Οι αναπαραστάσεις συγγραφέων/κειμένων μπορούν να δοθούν ως είσοδος σε αλγόριθμους μηχανικής μάθησης σε περιπτώσεις επιβλεπόμενης μάθησης και σε αλγόριθμους μη επιβλεπόμενης μάθησης, όπως διαπιστώσαμε για περιπτώσεις εύρεσης συστάδων (clustering). Οι τεχνικές αυτές είναι εξαιρετικά αποτελεσματικές όσον αφορά την ανίχνευση υποομάδων με παρόμοιο κειμενικό ύφος και αποτελούν μια σημαντική δομική συνιστώσα για πληθώρα εφαρμογών που έχουν ως άξονα το ύφος.

6.2.1 Εφαρμογές μεγάλης κλίμακας σε λογοτεχνικά κείμενα

Από φιλολογική σκοπιά, μια εξαιρετικά ενδιαφέρουσα εφαρμογή της παρούσας διατριβής θα ήταν στη λογοτεχνία, σε μεγάλη κλίμακα και εύρος συγγραφέων. Μια τέτοια εφαρμογή τυπικά είναι δυνητικά εφικτό να συμπεριλάβει, για παράδειγμα, το σύνολο της ελληνικής λογοτεχνίας τους τελευταίους τρεις ή τέσσερις αιώνες, με την προϋπόθεση ότι τα λογοτεχνικά κείμενα διατίθενται σε ψηφιακή μορφή.

Μια τέτοια εφαρμογή μεγάλης κλίμακας και θεωρώντας ότι υπάρχουν τα απαραίτητα μεταδεδομένα (συγγραφέας, χρονολογία, κειμενικό είδος) θα μπορούσε κάλλιστα να χρησιμοποιηθεί ως εργαλείο αναζήτησης ύφους, καθώς επίσης να απαντήσει σε ερωτήματα σχετικά με την πατρότητα λογοτεχνικών έργων. Στα ελληνικά μια πρώτη τέτοια απόπειρα έχει γίνει στο Πολίτου-Μαρμαρινού κ.α. (2013, σ. 361-392) και αφορά το ύφος του Παπαδιαμάντη.

Έχει ιδιαίτερο ενδιαφέρον ότι οι τρεις περιοχές που που προτείνει ο Salgaro (2018) για τις εφαρμογές της ποσοτικής έρευνας στη λογοτεχνία συνάδουν με τα τρία ερευνητικά ερωτήματα της παρούσας διατριβής. Συγκεκριμένα, ο Salgaro συζητά τη χρήση της ποσοτικής έρευνας για την απόδοση συγγραφέα (σε αμφισβητούμενα δοκίμια του Musil), για τη διερεύνηση του ύφους μιας σχολής (όσον αφορά το λεγόμενο ”ύστερο ύφος” σε Γερμανούς συγγραφείς) και για τη μελέτη της λογοτεχνικής ιστορίας. Συνοψίζει επίσης τα προτερήματα της ποσοτικής έρευνας σε αντιδιαστολή με την ποιοτική, υπογραμμίζοντας, μεταξύ άλλων, τον επαγωγικό της χαρακτήρα, τον λειτουργικό προσδιορισμό (operationalization) των ερωτημάτων της και τη γενικευσιμότητα των συμπερασμάτων της.

Ειδικότερα για τη μελέτη της λογοτεχνικής ιστορίας αξίζει να αναφερθεί το θεμελιώδες έργο του Moretti (λ.χ. Moretti, 2005), που άνοιξε δρόμους στη μελέτη της λογοτεχνίας με την χρήση αφηρημένων μοντέλων. Η έρευνα αυτή συνεχίζεται στο εργαστήριο Literary Lab του Πανεπιστημίου Stanford.[1] Όπως αναφέρθηκε πιο πριν, απαραίτητη προϋπόθεση για τέτοιους είδους μελέτες

Salgaro, Massimo 2018. The digital humanities as a toolkit for literary theory: Three case studies of the operationalization of the concepts of "late style," "authorship attribution," and "literary movement". iperstoria 12, 50-60.

6.2.2 Εφαρμογές ταξινόμησης κειμένων

Μια συνηθισμένη πρακτική όσον αφορά εφαρμογές επεξεργασίας φυσικής γλώσσας είναι η χρήση λεξικών ενθέσεων ως χαρακτηριστικών σε σύνθετα μοντέλα μηχανικής μάθησης ή ανάκτησης πληροφοριών. Οι ενθέσεις αυτές μπορούν είτε να παραχθούν από κάποιο υπάρχον σώμα κειμένων σχετικό με την εφαρμογή, είτε να χρησιμοποιηθεί κάποιο προ-εκπαιδευμένο μοντέλο, ως παράδειγμα μεταφοράς μάθησης. Η Φλώρου (Φλώρου, 2017), για παράδειγμα, χρησιμοποιεί το άθροισμα των λεξικών ενθέσεων του περικειμένου του ρήματος ως χαρακτηριστικό στο μοντέλο ανίχνευσης μεταφοράς.

Η πρακτική αυτή μπορεί να χρησιμοποιηθεί σε επίπεδο συγγραφέα θεωρώντας το ιδιολεκτικό ύφος ως αναπαράσταση/χαρακτηριστικό και την τροφοδότηση αυτών των αναπαραστάσεων σε μοντέλα μηχανικής μάθησης με σκοπό την ταξινόμηση χαρακτηριστικών ενός νέου συγγραφέα, π.χ. ως προς το κειμενικό ύφος.

Η τεχνική αυτή μπορεί να γενικευτεί και σε άλλα περιβάλλοντα πέραν της λογοτεχνίας με σκοπό την ταξινόμηση ενός ατόμου σε κάποια κατηγορία με βάση τα κείμενα που παράγει. Για παράδειγμα, οι Luyckx & Daelemans (2008) και Komisin & Guinn (2012) εφαρμόζουν ταξινόμηση προσωπικότητας με βάση κειμενικά δεδομένα. Ανάλογα, οι Mikros & Perifanos (2013) εφαρμόζουν μοντέλα μηχανικής μάθησης για την ταξινόμηση φύλου. Στις μελέτες αυτές μπορεί να χρησιμοποιηθεί το ιδιολεκτικό ύφος ως χαρακτηριστικό εισόδου στα μοντέλα μηχανικής μάθησης.

Για τον σκοπό αυτό διατίθενται προ-εκπαιδευμένα μοντέλα ενθέσεων κατασκευασμένα με τις τεχνικές των word2vec, fastText και GloVE και εκπαιδευμένα στο ΣΕΚ για διάφορες τιμές διαστάσεων D. Τα μοντέλα αυτά είναι διαθέσιμα για μεταφόρτωση στην ιστοσελίδα του ΣΕΚ, http://sek.edu.gr

[1] https://litlab.stanford.edu/

6.2.3 Εφαρμογές σε κοινωνικά δίκτυα και ανίχνευση ψευδών ειδήσεων

Η εκθετική ανάπτυξη των κοινωνικών δικτύων και η άμεση επιρροή που ασκούν όχι μόνο στον κοινωνικό ιστό, αλλά πλέον και στην παγκόσμια γεωπολιτική σκηνή, έχει ως αποτέλεσμα τη διεύρυνση φαινομένων από μικρή-τοπική σε παγκόσμια κλίμακα. Η παγκόσμια δικτύωση και η μεταφορά πληροφοριών σε πραγματικό χρόνο ενίσχυσαν φαινόμενα όπως η παραγωγή και μεταφορά προπαγάνδας, ψευδών ειδήσεων και θεωριών συνωμοσίας σε βαθμό που τείνουν να γίνουν επικρατούσα τάση.

Επιπλέον, φαινόμενα όπως η παρενόχληση, ο εκφοβισμός, η ρητορική μίσους στα κοινωνικά δίκτυα είναι πλέον κοινός τόπος ενώ δεν υπάρχουν ακόμα επαρκείς μηχανισμοί για την αντιμετώπιση τέτοιων ακραίων συμπεριφορών. Σχετικές εφαρμογές και σώματα κειμένων μπορούν, για παράδειγμα, να βρεθούν σε συνέδρια όπως το PAN [2], σε σχέση με θέματα, όπως η ανίχνευση σεξουαλικού θύτη (sexual predator identification) (Morris & Hirst, 2012; Bogdanova & Rosso, 2012) κ.ά.

Οι τεχνικές που παρουσιάζονται σε αυτή τη διατριβή μπορούν να χρησιμοποιηθούν σε ανάλογες περιπτώσεις και εφαρμογές. Δεδομένου ενός επισημειωμένου σώματος μπορούμε να κατασκευάσουμε τις ανάλογες κειμενικές αναπαραστάσεις, οι οποίες στη συνέχεια θα δοθούν ως είσοδος σε αλγόριθμους επιβλεπόμενης μηχανικής μάθησης με σκοπό την κατασκευή μοντέλων πρόβλεψης και ανίχνευσης ακραίων συμπεριφορών.

Ειδικότερα, για τη χρονική περίοδο 2015-2018 σημαντικό μέρος της έρευνας της επεξεργασίας φυσικής γλώσσας και της τεχνητής νοημοσύνης έχει επικεντρωθεί στην Ανίχνευση ψευδών ειδήσεων (Conroy et al., 2015; V. L. Rubin et al., 2015; V. Rubin et al., 2016; Wang, 2017; Shu et al., 2017; Tacchini et al., 2017; Potthast et al., 2017). Ο συνδυασμός μοντέλων μηχανικής μάθησης και κατάταξης περιεχομένου αποτελεί βασικό συστατικό στην προσπάθεια αυτή. Οι τεχνικές που αναπτύσσονται στην παρούσα διατριβή μπορούν να χρησιμοποιηθούν ως συστατικό αλγορίθμων μηχανικής μάθησης σε συστήματα ανίχνευσης ψευδών ειδήσεων, κυρίως όσον αφορά την υφολογική ομοιότητα πηγών που τείνουν να παράγουν ψευδείς ειδήσεις. Η υφολογική συνάφεια των πηγών αυτών μπορεί να συνδυαστεί με πληροφορία προερχόμενη από την κοινωνική δικτύωση και να χρησιμοποιηθεί ως χαρακτηριστικό (ή αναπαράσταση) της πηγής με βάση την οποία ένα κείμενο υπόκειται σε έλεγχο αξιοπιστίας.

[2] http://pan.webis.de

6.2.4 Εφαρμογές σε μηχανές σύστασης περιεχομένου και μηχανές αναζήτησης

Σε ένα περιβάλλον παραγωγής κειμένων, όπως λ.χ. σε μια ηλεκτρονική εφημερίδα ή σε κοινωνικά δίκτυα, ένα σημαντικό χαρακτηριστικό είναι η σύσταση περιεχομένου σε χρήστες με βάση τις προτιμήσεις τους. Μια μηχανή πρότασης περιεχομένου σε ειδησεογραφικό περιβάλλον μπορεί να λαμβάνει υπόψη τη συμπεριφορά του χρήστη ως προς τις κατηγορίες περιεχομένου για τις οποίες ενδιαφέρεται ο χρήστης (λ.χ. Πολιτική, Οικονομία, Αθλητικά κ.λπ).

Στη λογική αυτή, μια μηχανή σύστασης περιεχομένου μπορεί να λαμβάνει αναπαραστάσεις υφομετρικής ομοιότητας, δηλαδή να προτείνει κείμενα υφολογικά όμοια με αυτά που προτιμά ο αναγνώστης του ηλεκτρονικού εντύπου. Ανάλογα, αναπαραστάσεις ύφους μπορούν να χρησιμοποιηθούν σε κοινωνικά δίκτυα, όπως το Twitter, όπου το κείμενο είναι το βασικό συστατικό πληροφορίας με το οποίο έρχεται σε επαφή ο χρήστης. Η αρχή εδώ είναι ότι, αν ένας χρήστης τείνει να προτιμά περιεχόμενο συγκεκριμένου ύφους, ίσως να ενδιαφέρεται για χρήστες/πηγές που παράγουν περιεχόμενο ανάλογου ύφους.

Σε περιβάλλοντα αναζήτησης και ανάκτησης πληροφοριών, όπως έχουμε ήδη αναφέρει στην ενότητα 4.7.2, είναι εφικτό να γίνει ανάκτηση περιεχομένου με βάση την υφολογική ομοιότητα μεταξύ χρηστών.

6.2.5 Εφαρμογές σε πηγαίο κώδικα

Τεχνικές αναγνώρισης συγγραφέα έχουν εφαρμοστεί με επιτυχία σε αποθήκες (repositories) πηγαίου κώδικα, όπως λ.χ. github[3], bitbucket[4]. Frantzeskou et al. (2007), Burrows & Tahaghoghi (2007), Yang et al. (2017). Στην περίπτωση αυτή κατ' αναλογία με την απόδοση συγγραφέα, το ζητούμενο είναι η απόδοση πατρότητας λογισμικού. Ακολουθώντας το μοτίβο που παρουσιάζεται στις προηγούμενες περιπτώσεις, μπορούμε να κατασκευάσουμε αναπαραστάσεις χρηστών σε αποθήκες κώδικα και να χρησιμοποιήσουμε αυτές τις αναπαραστάσεις ως αποτυπώματα του ιδιαίτερου ύφους κάθε προγραμματιστή. Επιπλέον, αναπαραστάσεις μπορούν να κατασκευαστούν σε επίπεδο έργου ανοιχτού λογισμικού (λ.χ. Apache Hadoop) ή ακόμα και σε επίπεδο οργανισμού. Στη συνέχεια, μπορούμε να χρησιμοποιήσουμε τις αναπαραστάσεις είτε ως επαύξηση σε μηχανές αναζήτησης κώδικα ή ακόμα και σε εφαρμογές που άπτονται πνευματικών δικαιωμάτων.

[3] http://github.com

[4] http://bitbucket.com

6.3 Μελλοντική έρευνα

Στην ενότητα αυτή παρουσιάζονται οι προεκτάσεις της παρούσας διατριβής σε θεωρητικό επίπεδο και με αναφορά στους περιορισμούς της έρευνας.

6.3.1 Θεωρητικές προεκτάσεις

Η παρούσα διατριβή βασίζεται σε αρχές που προέρχονται από τη γλωσσολογία και ειδικότερα την κοινωνιογλωσσολογία, την ανάκτηση πληροφοριών, τη μηχανική μάθηση και την τεχνητή νοημοσύνη. Ο συνδετικός κρίκος αυτών των ερευνητικών περιοχών είναι η Εκμάθηση αναπαραστάσεων (Representation Learning) (Bengio et al., 2013, Goodfellow et al., 2016). Ο κλάδος αυτός της μηχανικής μάθησης ασχολείται με την εκμάθηση αναπαραστάσεων. Ένας αλγόριθμος μηχανικής μάθησης εκπαιδεύεται με τέτοιο τρόπο ώστε να κατασκευάζει αναπαραστάσεις των δεδομένων, οι οποίες χρησιμοποιούνται στη συνέχεια ως είσοδος σε παραδοσιακούς αλγόριθμους μηχανικής μάθησης. Το ζητούμενο είναι η κατασκευή αναπαραστάσεων που να βελτιώνουν τα αποτελέσματα των εργασιών στις οποίες εφαρμόζεται μηχανική μάθηση.

Δεδομένων διάφορων τεχνικών εκμάθησης αναπαραστάσεων, οι τεχνικές που αναφέρθηκαν στην παρούσα διατριβή προέρχονται από μια οικογένεια αναπαραστάσεων, οι οποίες αποτελούν σύνθεση πολλαπλών μη γραμμικών μετασχηματισμών και στοχεύουν στην κατασκευή αφηρημένων και χρήσιμων αναπαραστάσεων. Κάνοντας λόγο για χρησιμότητα εννοούμε, όπως αναφέρθηκε παραπάνω, την απόδοση της χρήσης αυτών των τεχνικών σε προβλήματα μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας. Από μαθηματική αλλά και πρακτική σκοπιά, ένα σύστημα εκμάθησης αναπαραστάσεων πρέπει να έχει τη δυνατότητα να παράγει αναπαραστάσεις συμπαγείς και με εκφραστικότητα, δηλαδή ένας σχετικά μικρός αριθμός διαστάσεων να είναι ικανός να συλλαμβάνει ένα ευρύ φάσμα ιδιοτήτων των δεδομένων εισόδου.

Η εκμάθηση αναπαραστάσεων εφαρμόζεται σε όλο το φάσμα προβλημάτων που άπτονται της τεχνητής νοημοσύνης: αναγνώριση φωνής και επεξεργασία σήματος, αναγνώριση αντικειμένων και μηχανική όραση, επεξεργασία φυσικής γλώσσας, καθώς και σε σύνθετες περιπτώσεις συνδυασμού τροπικοτήτων και μεταφοράς γνώσης (Multi-Task and Transfer Learning).

Κατά δεύτερο λόγο, είναι ενδιαφέρον να συζητηθεί το ερώτημα που επικρατεί στην ερευνητική κοινότητα και δεν έχει ακόμα απαντηθεί σε ικανοποιητικό βαθμό, γιατί οι λεξικές ενθέσεις είναι τόσο αποδοτικές στην πράξη. Η πιο στοιχειοθετημένη άποψη προέρχεται από τους Arora et al. (2015), οι οποίοι θέτουν δύο σημαντικά υπο-ερωτήματα:

1. ποια είναι εκείνα τα χαρακτηριστικά της γλώσσας τα οποία συλλαμβάνει

ένα μοντέλο ενθέσεων;

2. γιατί σχετικά μικρές διαστάσεις (της τάξης του $D = 100$ έως $D = 500$) τείνουν να λειτουργούν καλύτερα από μοντέλα μεγαλύτερης διαστασιμότητας;

Οι Arora et al. (2015) δίνουν ως πιθανή απάντηση για τις ερωτήσεις αυτές την υπόθεση που διατυπώνουν οι Pennington et al. (2014), και πιο συγκεκριμένα ό,τι ισχύει κατά προσέγγιση για λέξη X του σώματος κειμένων.

$$\frac{p(X|king}{p(X|queen} \approx \frac{p(X|man)}{p(X|woman}$$

με $p(X|king)$ η πιθανότητα εμφάνισης της λέξης X σε περικείμενο λίγων λέξεων που να περιέχει τη λέξη king. Η υπόθεση αυτή πηγάζει άμεσα από την κατανεμητική υπόθεση, σύμφωνα με την οποία λέξεις που τείνουν να συνεμφανίζονται στα ίδια περικείμενα τείνουν να έχουν παρόμοια σημασία.

Οι Levy & Goldberg (2014) δείχνουν εμπειρικά ότι για το μοντέλο word2vec με αρχιτεκτονική skipgram ισχύει $< u^w, u^{w'} > \approx PMI(w, w')$ για $u^w, u^{w'}$ διανύσματα ενθέσεων.

Οι Arora et al. (2015) σημειώνουν ότι ο λόγος που η σχέση αυτή φαίνεται να ισχύει και να δίνει καλύτερα αποτελέσματα σε ενθέσεις με χαμηλή διαστασιμότητα ενδέχεται να είναι το γεγονός ότι τα διανύσματα σε αυτόν τον αριθμό διαστάσεων είναι ευκολότερο να είναι χωρικά ισοτροπικά (spatially isotropic), δηλαδή ομοιόμορφα κατανεμημένα σε όλες τις διαστάσεις του διανυσματικού χώρου, κάτι που είναι δυσκολότερο να επιτευχθεί σε μεγαλύτερες διαστάσεις.

Το βάρος της έρευνας, λοιπόν, μετακινείται στην ανάπτυξη ενός ισχυρού μαθηματικού φορμαλισμού, ικανού να τεκμηριώσει την επιτυχία των μοντέλων ενθέσεων και των διανυσματικών αναπαραστάσεων γενικότερα.

Τέλος, η απάντηση στο τελευταίο ερώτημα έχει σαφείς προεκτάσεις σε επιστημονικά πεδία της γλωσσολογίας όπως η σημασιολογία, η κοινωνιογλωσσολογία κ.λπ. Τα ζητήματα που τίθενται εδώ έχουν να κάνουν με τον χαρακτήρα και τον ρόλο των λεξικών ενθέσεων, που φαίνεται να υπερβαίνουν τις παραδοσιακές έννοιες της λεξικής και γραμματικής σύναψης, έννοιες όπως τα δομοσυναπτικά σχήματα (collostruction, Gries & Stefanowitsch, 2003), αλλά των σημασιολογικών σχέσεων όπως των σχέσεων λέξεων σε ένα σημασιολογικό πεδίο. Θα είχε ιδιαίτερο ενδιαφέρον να διερευνηθεί η σχέση αυτών των εννοιών με τις διανυσματικές αναπαραστάσεις λέξεων και να βρεθούν πιθανές αντιστοιχίσεις με οντολογίες σημασιολογικής αναπαράστασης όπως το Wordnet (Fellbaum, 2005) κ.ά.

Με παρόμοιο τρόπο, η μαθηματική/υπολογιστική αναπαράσταση της ιδιολέκτου φαίνεται να υπερβαίνει τον παραδοσιακό ορισμό των λέκτων στην κοινωνιογλωσσολογία, επιτυγχάνοντας μια αυτόματη συσχέτιση κοινωνικών παραγόντων και γλωσσικών χαρακτηριστικών. Το ερώτημα της συνειδητότητας των γλωσσικών επιλογών ή της προθετικότητας, που έχει απασχολήσει τη γλωσσολογική υφολογία με κορυφαία την ανάλυση του Jakobson (1985, σσ. 69–), μετατίθεται πλέον σε άλλο επίπεδο, καθώς αφορά τη συσχέτιση του συνόλου του λέξεων που ανήκουν στο λεξιλόγιο ενός συγγραφέα μεταξύ τους, και όχι μεμονωμένες συνάψεις ή δομικά σχήματα.

6.3.2 Εσκεμμένη αλλαγή ύφους

Η υπόθεση εργασίας στη συγκεκριμένη διατριβή είναι ότι η παραγωγή κειμένων γίνεται χωρίς εσκεμμένη προσπάθεια από την πλευρά των συγγραφέων να αλλάξουν το ύφος τους, υποδυόμενοι κάποιον άλλον. Η υπόθεση αυτή ισχύει στην πλειονότητα των περιπτώσεων. Ωστόσο, υπάρχουν περιπτώσεις στις οποίες οι χρήστες αλλάζουν το ύφος τους ανάλογα με το περιβάλλον ή/και τον ρόλο που υποδύονται. Στη συγκεκριμένη διατριβή αυτό το θέμα δεν εξετάζεται. Από τη σκοπιά της ανάκτησης πληροφοριών, στην περίπτωση που ένας χρήστης αλλάζει εσκεμμένα με τρόπο που να αντιγράφει ή μιμείται το ύφος κάποιου άλλου χρήστη, η συγκεκριμένη τεχνική που περιγράφεται στην παρούσα διατριβή θα ανιχνεύσει την ομοιότητα, θα αποτύχει όμως να εντοπίσει το πραγματικό ιδιολεκτικό ύφος του συγγραφέα.

6.3.3 Βαθιά μοντέλα και αναπαραστάσεις

Τα νευροκατανεμητικά μοντέλα που χρησιμοποιήθηκαν στη διατριβή (δηλαδή τα word2vec,fastText και doc2vec) ανήκουν στην κατηγορία των ρηχών μοντέλων. Ένα ενδιαφέρον θεωρητικό ερώτημα είναι αν και κατά πόσο μοντέλα με πολλά επίπεδα, που είναι τυπικά σε εφαρμογές βαθιάς μάθησης, μπορούν επίσης να χρησιμοποιηθούν για την καταγραφή και παραγωγή αναπαραστάσεων κειμενικού ύφους. Ένα μειονέκτημα αυτών των αρχιτεκτονικών είναι προφανώς ο χρόνος εκπαίδευσης, που αυξάνει μαζί με την πολυπλοκότητα των μοντέλων.[5]

Μεγάλο βάρος της έρευνας τα τελευταία χρόνια έχει μετατοπιστεί σε αρχιτεκτονικές βαθιάς μάθησης (Deep Learning) και πιο συγκεκριμένα σε αρχιτεκτονικές κωδικοποίησης-αποκωδικοποίησης (encoder-decoder). Στις

[5]Εναλλακτικά μοντέλα εκτός αυτών που χρησιμοποιήθηκαν στην διατριβή είναι επίσης τα σιαμαία μοντέλα (Siamese Neural Models) (Zheng et al., 2018), καθώς επίσης και μοντέλα τρίπλευρου σφάλματος (triplet loss models) όπως για παράδειγμα το Starspace (L. Y. Wu et al., 2018) που για λόγους οικονομίας δεν χρησιμοποιήθηκαν στη διατριβή. Ευχαριστώ τον κ. Α. Πικράκη για την επισήμανση.

αρχιτεκτονικές αυτές ο κωδικοποιητής μετασχηματίζει το διάνυσμα εισόδου σε αναπαράσταση και στη συνέχεια ο αποκωδικοποιητής παίρνει αυτή την αναπαράσταση και την απεικονίζει στο ζητούμενο διάνυσμα εξόδου. Τυπικό παράδειγμα τέτοιων αρχιτεκτονικών είναι τα συστήματα Μηχανικής μετάφρασης (machine translation), στα οποία είσοδος είναι μια πρόταση από τη γλώσσα-πηγή και έξοδος η ισοδύναμη φράση στη γλώσσα-στόχο. Ο κωδικοποιητής, λοιπόν, παίρνει την φράση εισόδου και την απεικονίζει σε μια εσωτερική αναπαράσταση και έπειτα ο αποκωδικοποιητής παίρνει τη φράση αυτή και την απεικονίζει στη φράση εξόδου στη γλώσσα-στόχο. Η διαδικασία εκπαίδευσης γίνεται με τέτοιο τρόπο ώστε να ελαχιστοποιηθεί το σφάλμα στην παραγόμενη από το νευρωνικό δίκτυο φράση και την φράση-στόχο (Sutskever et al., 2014).

Οι αρχιτεκτονικές αυτές έχουν δυο σημεία ενδιαφέροντος: πρώτον, τη χρήση αναδρομικών νευρωνικών δικτύων και, δεύτερον, τη γενίκευση της έννοιας της εκμάθησης αναπαραστάσεων των δεδομένων εισόδου. Τα μοντέλα που έχουν εφαρμοστεί σε αυτή τη διατριβή δεν χρησιμοποιούν άμεσα τη δομή και τη σειρά των κειμένων. Αντίθετα, η βάση τους στην πράξη είναι η αναπαράσταση σωρού λέξεων (bag of words). Αν και αυτή η αναπαράσταση δίνει εξαιρετικά καλά αποτελέσματα, δεν χρησιμοποιεί άμεσα την δομή της πρότασης, ειδικά σε μοντέλα όπως το GloVe.

6.3.4 Αναπαραστάσεις και νευροκατανεμητικά μοντέλα σημασιολογίας

Η εφαρμογή τεχνικών εκμάθησης αναπαραστάσεων συνδυάζεται με τις αρχές της κατανεμητικής και νευροκατανεμητικής σημασιολογίας. Με τη χρήση νευροκατανεμητικών μοντέλων σημασιολογίας κατασκευάζουμε αναπαραστάσεις συμβόλων (λέξεων) και, στη συνέχεια, από τις αναπαραστάσεις των συμβόλων πρώτου επιπέδου παράγονται αναπαραστάσεις εννοιών (concepts), οι οποίες συντίθενται από σύμβολα του πρώτου επιπέδου. Για παράδειγμα, στην περίπτωση αυτής της διατριβής, η αναπαράσταση των ιδιολεκτικών επιλογών των συγγραφέων θεωρείται συνάρτηση των αναπαραστάσεων των λέξεων που τείνουν να χρησιμοποιούν.

Η θεωρητική υπόθεση στην οποία βασίζεται αυτή η διατριβή είναι ότι το ιδιολεκτικό ύφος είναι αποτέλεσμα συνειδητών και ασυνείδητων επιλογών και εκφράζεται ως άθροισμα κοινωνικών, γεωγραφικών και γνωσιακών παραγόντων. Οι μέχρι τώρα ποσοτικές μελέτες της ιδιολέκτου σε σώματα κειμένων δεν χρησιμοποιούν καθόλου τη συνιστώσα των λεξικών τύπων και, επομένως, του πλήρους σημασιολογικού περιεχομένου ως παράμετρο της ιδιολέκτου. Η παρούσα διατριβή ενεργοποιεί αυτή τη σημασιολογική συνιστώσα της έννοιας της ιδιολέκτου με τη χρήση λεξικών αναπαραστάσεων και επεκτείνοντας το

υποσύνολο των λέξεων του λεξιλογίου το οποίο χρησιμοποιόταν στις έως σήμερα μελέτες καταγραφής ιδιολέκτου κατά τουλάχιστον μια τάξη μεγέθους. Αυτό γίνεται επίσης σε συνδυασμό με τις σημασιολογικές ιδιότητες που εμπεριέχουν οι λεξικές ενθέσεις. Σημειώνεται ότι οι τεχνικές που περιγράφονται στην παρούσα διατριβή εφαρμόζονται σε αυστηρά καθορισμένο κειμενικό είδος και δεν επιχειρείται μεταφορά υφολογικής αναπαράστασης σε άλλα είδη. Όπως αποδείχτηκε, οι τεχνικές αυτές στα συγκεκριμένα σώματα κειμένων υπερτερούν όσον αφορά την ταυτοποίηση ατόμου σε σχέση με τις παραδοσιακές υφομετρικές τεχνικές (συχνότητα λειτουργικών λέξεων), και συνεπώς το ερευνητικό προς απάντηση ερώτημα είναι καλά ορισμένο. Επομένως, το ερώτημα για το τι είναι ιδιόλεκτος δεν μπορεί να απαντηθεί εκτός των παραδοχών των οποίων έχει θέσει η παρούσα διατριβή. Με άλλα λόγια, μπορούμε να διαχωρίσουμε το ιδιολεκτικό ύφος κάθε συγγραφέα αλλά δεν μπορούμε να καθορίσουμε αυστηρά αν αυτό αποτελείται από θεματικά στοιχεία, γραμματικά στοιχεία, λεξικούς τύπους ή άλλα γλωσσικά στοιχεία που δεν έχουν αναλυθεί στην έρευνά μας, αλλά μη γραμμική σύνθεση όλων των παραπάνω.

Αν θεωρηθεί ότι ένα κείμενο αποτελεί δειγματοληψία λέξεων από τις λέξεις του λεξιλογίου που είναι διαθέσιμες σε έναν συγγραφέα, οι ενθέσεις ύφους του συγγραφέα μπορούν να θεωρηθούν ως δείγμα από την κατανομή πιθανότητας στο διανυσματικό χώρο των αφηρημένων απεικονίσεων που ορίζουν οι λεξικές ενθέσεις ενός σώματος κειμένων. Υπό αυτή την θεώρηση, το ιδιολεκτικό ύφος ενός συγγραφέα αποτελεί αφηρημένη προσέγγιση της αναπαράστασης του langue του ατόμου, μιας αναπαράστασης που κατασκευάστηκε από το ιστορικό του parole του. Δηλαδή, η υφολογική αναπαράσταση μπορεί να θεωρηθεί ως προσέγγιση της *υφολογικής περιοχής*, στην οποία είναι αναμενόμενο να κινηθεί ένα άτομο, με βάση την πραγμάτωση των παρελθοντικών υφολογικών επιλογών του. Με αυτό τον τρόπο, γεφυρώνεται η διαφορά μεταξύ γλωσσικού συστήματος και γλωσσικής πραγμάτωσης στο πεδίο της ιδιολέκτου.

Βιβλιογραφία

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., …, Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *Osdi* (Vol. 16, σσ. 265–283).

Afroz, S., Islam, A. C., Stolerman, A., Greenstadt, R., & McCoy, D. (2014). Doppelgänger finder: Taking stylometry to the underground. In *Security and privacy (sp), 2014 ieee symposium on* (σσ. 212–226). IEEE.

Alejandro, J. (2010). Journalism in the age of social media. *Reuters Institute Fellowship Paper, University of Oxford*, 2009–2010.

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2015). Rand-walk: A latent variable model approach to word embeddings. *arXiv preprint arXiv:1502.03520*.

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160.

Barlow, M. (2013). Individual usage: A corpus-based study of idiolects. *Linguistics*, *18*(4).

Bayley, R. & Lucas, C. (2007). Sociolinguistic variation. *Cambridge: Cambridge*.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137–1155.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bloch, B. (1948). A set of postulates for phonemic analysis. *Language*, *24*(1), 3–46.

Bloomfield, L. (1933). *Language history: From language (1933 ed.)*. Holt, Rinehart and Winston.

Bloomfield, L. (1965). *Language new york*. Taylor & Francis.

Bogdanova, D. & Rosso, P. (2012). Submission to the 1st International Competition on Sexual Predator Identification. http://www.uni-weimar.de/medien/webis/events/pan-12. From the Heilongjiang Institute of Technology.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System sciences (hicss), 2010 43rd hawaii international conference on* (σσ. 1–10). IEEE.

Bray, M. & Thomson, A. (1996). *Temagami: A debate on wilderness*. Dundurn.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4), 467–479.

Browning, D. N. (2017). *# twitterdiscoursemarkers: A corpora based study of the pragmatic functions of hashtags* (Doctoral dissertation).

Bryden, J., Wright, S. P., & Jansen, V. A. (2018). How humans transmit language: Horizontal transmission matches word frequencies among peers on twitter. *Journal of The Royal Society Interface*, *15*(139), 20170738.

Bullinaria, J. A. & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510–526.

Burrows, S. & Tahaghoghi, S. M. (2007). Source code authorship attribution using n-grams. In *Proceedings of the twelth australasian document computing symposium, melbourne, australia, rmit university* (σσ. 32–39). Citeseer.

Carter, R. (2012). *Vocabulary: Applied linguistic perspectives*. Routledge.

Chambers, J. K. (1990). Forensic dialectology and the bear island land claim. *Annals of the New York Academy of Sciences*, *606*(1), 19–31.

Chen, S. F. & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359–394.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22–29.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*(Aug), 2493–2537.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th asis&t annual meeting: Information science with impact: Research in and for the community* (σ. 82). American Society for Information Science.

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied linguistics*, *25*(4), 431–447.
Coulthard, M. & Johnson, A. (2010). *The routledge handbook of forensic linguistics*. Routledge.
Coulthard, M., Grant, T., & Kredens, K. (2011). Forensic linguistics.
Coupland, N. (2007). *Style: Language variation and identity*. Cambridge University Press.
Cox, T. (2017). Who's blogging now?: Linguistic features and authorship analysis in sports blogs. *ProQuest LLC*.
Dagan, I., Lee, L., & Pereira, F. C. (1999). Similarity-based models of word cooccurrence probabilities. *Machine learning*, *34*(1-3), 43–69.
Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.
De Beaugrande, R. (1981). Introduction to text linguistics.
De Beaugrande, R. & Dressler, W. U. (1981). *Einführung in die textlinguistik*. Niemeyer Tübingen.
Dean, J. & Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107–113.
DeCamp, D. (1970). Is a sociolinguistic theory possible. *JE Alatis (éd.), Report of the Twentieth Annual Round Table Moeting on Linguistics and Language Studies. Georgetown*, 157–173.
Deerwester, S. C., Dumais, S. T., Furnas, G. W., Harshman, R. A., Landauer, T. K., Lochbaum, K. E., & Streeter, L. A. (1989). Computer information retrieval using latent semantic structure. US Patent 4,839,853. Google Patents.
Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391.
Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, *3*.
Dittmar, N. (1996). Explorations in'idiolects'. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 109–128.
Dubin, D. (2004). The most influential paper gerard salton never wrote.
Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121–2159.
Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, *41*, 87–100.
Erk, K. & Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the conference on empirical methods in natu-*

ral language processing (σσ. 897–906). Association for Computational Linguistics.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, *34*, σσ. 226–231).

Evans, N. & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, *32*(5), 429–448.

Fano, R. M. (1961). Transmission of information: A statistical theory of communications.

Farkas, J., Schou, J., & Neumayer, C. (2018). Cloaked facebook pages: Exploring fake islamist propaganda in social media. *New Media & Society*, *20*(5), 1850–1867.

Fellbaum, C. (2005). Wordnet and wordnets.

Fialho, O. & Zyngier, S. (2017). Quantitative methodological approaches to stylistics. *The Routledge handbook of stylistics*, 329.

Firth, J. (1957). Papers in linguistics, 0xford: Oxford university press.

Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C. E., & Howald, B. S. (2007). Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *International Journal of Digital Evidence*, *6*(1), 1–18.

Gadavanij, S. (2002). Intertextuality as discourse strategy: The case of no-confidence debates in thailand. *Leeds working papers in linguistics and phonetics*, *9*, 35–55.

Gattis, M. (2003). *Spatial schemas and abstract thought*. MIT press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155–170.

Gerbaudo, P. (2018). *Tweets and the streets: Social media and contemporary activism*. Pluto Press.

Gianfortoni, P., Adamson, D., & Rosé, C. P. (2011). Modeling of stylistic variation in social media with stretchy patterns. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties* (σσ. 49–59). Association for Computational Linguistics.

Gleitman, L. R. (1993). A human universal: The capacity to learn a language. *Modern philology*, *90*, S13–S33.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (σσ. 315–323).

Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. In *Chi'11 extended abstracts on human factors in computing systems* (σσ. 253–262). ACM.

Goldberg, Y. & Levy, O. (2014). Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Goodenow, M., Huet, T., Saurin, W., Kwok, S., Sninsky, J., & Wain-Hobson, S. (1989). Hiv-1 isolates are rapidly evolving quasispecies: Evidence for viral mixtures and preferred nucleotide substitutions. *Journal of acquired immune deficiency syndromes*, *2*(4), 344–352.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. http://www.deeplearningbook.org. MIT Press.

Goutsos, D. (1995). Review article: Forensic stylistics. *Forensic Linguistics*, *2*(1), 99–113.

Goutsos, D. (2010). The corpus of Greek texts: A reference corpus for modern greek. *Corpora*, *5*(1), 29–44.

Grant, T. (2005). *Authorship attribution in a forensic context* (Doctoral dissertation, The University of Birmingham).

Grant, T. (2012). Txt 4n6: Method, consistency, and distinctiveness in the analysis of sms text messages. *JL & Pol'y*, *21*, 467.

Grattan, M. (1998). The politics of spin. *Australian Studies in Journalism*, (7), 32–45.

Gries, S. T. & Stefanowitsch. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243.

Gutmann, M. & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (σσ. 297–304).

Hadoop, A. (2009). Hadoop.

Halliday, M. (1978). K (1978). language as social semiotic: The social interpretation of language and meaning. *London: Arnold*.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Hockett, C. F. (1958). *A course in modern linguistics*. Oxford and Ibh Publishing Co.; New Delhi.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (σσ. 289–296). Morgan Kaufmann Publishers Inc.

Honey, C. & Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via twitter. In *System sciences, 2009. hicss'09. 42nd hawaii international conference on* (σσ. 1–10). Ieee.

Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proceedings of the 21st acm conference on hypertext and hypermedia* (σσ. 173–178). ACM.

Hughes, J. M., Foti, N. J., Krakauer, D. C., & Rockmore, D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*, *109*(20), 7682–7686.

Ivanič, R. (1998). *Writing and identity*. John Benjamins Publishing Company.

Jakobson, R. (1985). *Verbal art, verbal sign, verbal time*. U of Minnesota Press.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., …, Corrado, G., et al. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Johnstone, B. (1996). *The linguistic individual: Self-expression in language and linguistics*. Oxford University Press.

Jones, W. P. & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American society for information science*, *38*(6), 420.

Jørgensen, A., Hovy, D., & Søgaard, A. (2015). Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text* (σσ. 9–18).

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213.

Juola, P. et al. (2008). Authorship attribution. *Foundations and Trends® in Information Retrieval*, *1*(3), 233–334.

Kalyanam, J. (2017). *Machine learning and applications on social media data*. University of California, San Diego.

Kaplan, A. M. & Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, *53*(1), 59–68.

Karasimos, A., Markopoulos, G., Sgarbas, K., & Christofidou, A. (2015). Ορολογία υπολογιστικής γλωσσολογίας, 245–327.

Komisin, M. & Guinn, C. (2012). Identifying personality types.

Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval* (σσ. 659–660). ACM.

Kredens, K. (2002). Towards a corpus-based methodology of forensic authorship attribution: A comparative study of two idiolects. In *Palc* (Vol. 1, σσ. 405–437).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (σσ. 1097–1105).

Kuhl, J. W. (2003). *The idiolect, chaos, and language custom far from equilibrium: Conversations in morocco* (Doctoral dissertation, University of Georgia).

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

Landauer, T. K. & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Langacker, R. W. (2017). Conceptualization, symbolization, and grammar. In *The new psychology of language* (σσ. 1–39). Routledge.

Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (σσ. 1188–1196).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics* (σσ. 25–32). Association for Computational Linguistics.

Levy, O. & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (σσ. 2177–2185).

Lewis, D. (1976). General semantics. In *Montague grammar* (σσ. 1–50). Elsevier.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Literary and Linguistic Computing*, *31*(2), 374–397.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on computational linguistics-volume 2* (σσ. 768–774). Association for Computational Linguistics.

Lin, D. & Pantel, P. (2001). Discovery of inference rules for question-answering. *Natural Language Engineering*, *7*(4), 343–360.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1–167.

Lopez Long, H. (2012). Tweeting to the choir: Religious identity construction and negotiating positions of authority on twitter.

Lowe, W. (2001). Towards a theory of semantic space. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 23, *23*).

Lund, K. & Burgess, C. (1996). Hyperspace analogue to language (hal): A general model semantic representation. In *Brain and cognition* (Vol. 30, *3*, σσ. 5–5). ACADEMIC PRESS INC JNL-COMP SUBSCRIPTIONS 525 B ST, STE 1900, SAN DIEGO, CA 92101-4495.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, *28*(2), 203–208.

Luyckx, K. & Daelemans, W. (2008). Personae: A corpus for author and personality prediction from text. In *Lrec*.

Maguire, W. & McMahon, A. (2011). *Analysing variation in english*. Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (σσ. 3111–3119). Curran Associates, Inc. Διαθέσιμο στο http://papers.nips.cc/paper/5021 - distributed - representations - of - words - and - phrases - and - their - compositionality.pdf

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (σσ. 3111–3119).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikros, G. K. (2008). Η ποσοτική ανάλυση της κοινωνιογλωσσολογικής ποικιλίας: Θεωρητικές και μεθοδολογικές προσεγγίσεις. Αθήνα: Μεταίχμιο.

Mikros, G. K. (2015). Υπολογιστική υφολογία.

Mikros, G. K. & Perifanos, K. (2013). Authorship attribution in greek tweets using author's multilevel n-gram profiles.

Mitchell, J. & Lapata, M. (2008). Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, 236–244.

Mollin, S. (2009). "i entirely understand" is a blairism. *International Journal of Corpus Linguistics*, *14*(3), 367–392.

Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. Verso.

Morin, F. & Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *Aistats* (Vol. 5, σσ. 246–252). Citeseer.

Morris, C. & Hirst, G. (2012). Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features—Notebook for PAN at CLEF 2012. In P. Forner, J. Karlgren, & C. Womser-Hacker (Eds.), *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy*. Διαθέσιμο στο http://www.clef-initiative.eu/publication/working-notes

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, *42*(3), 537–593.

Nini, A. & Grant, T. (2013). Bridging the gap between stylistic and cognitive approaches to authorship analysis using systemic functional linguistics and multidimensional analysis. *International Journal of Speech, Language & the Law*, *20*(2).

Ogden, C. K., Richards, I. A., Malinowski, B., & Crookshank, F. G. (1923). *The meaning of meaning*. Kegan Paul London.

O'Grady, W., Dobrovolsky, M., & Katamba, F. (1997). *Contemporary linguistics*. St. Martin's.

Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (σσ. 125–132). Association for Computational Linguistics.

Pantel, P. & Lin, D. (2002). Discovering word senses from text. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (σσ. 613–619). ACM.

Pantel, P. & Lin, D. (2002). Document clustering with committees. In *Proceedings of the 25th annual international acm sigir conference on research and development in information retrieval* (σσ. 199–206). ACM.

Paszke, A., Gross, S., Chintala, S., & Chanan, G. (2017). Pytorch.

Paul, H. (1890). *Principles of the history of language*. Рипол Классик.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (σσ. 1532–1543).

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Quirk, R. (1985). *A comprehensive grammar of the english language*. Pearson Education India.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the ninth machine translation summit* (σσ. 315–322).

Rehurek, R. & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Reynolds, W. N., Salter, W. J., Farber, R. M., Corley, C., Dowling, C. P., Beeman, W. O., … Choi, J. N. (2013). Sociolect-based community detection. In *Intelligence and security informatics (isi), 2013 ieee international conference on* (σσ. 221–226). IEEE.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, *8*(627-633), 116.

Rubenstein, H. & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: Three types of fakes. In *Proceedings of the 78th asis&t annual meeting: Information science with impact: Research in and for the community* (σ. 83). American Society for Information Science.

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection* (σσ. 7–17).

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Sapir, E. (1927). Speech as a personality trait. *American journal of sociology*, *32*(6), 892–905.

Schaal, M., O'Donovan, J., & Smyth, B. (2012). An analysis of topical proximity in the twitter social graph. In *International conference on social informatics* (σσ. 232–245). Springer.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *Aaai spring symposium: Computational approaches to analyzing weblogs* (Vol. 6, σσ. 199–205).

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85–117.

Schütze, H. & Pedersen, J. O. (1995). Information retrieval based on word senses.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., …, Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

Serrano, M. Á., Boguná, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, *106*(16), 6483–6488.

Shannon, C. E. (1948). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3–55.

Shapiro, M. A., Hemphill, L., & Otterbacher, J. (2018). Updates to congressional speech acts on twitter.

Shifman, L. (2013). Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication*, *18*(3), 362–377.

Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign affairs*, 28–41.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
Sinclair, J. M. (1974). English lexical collocations: A study in computational linguistics. *Cahiers de lexicologie*.
Singhal, A., Salton, G., Mitra, M., & Buckley, C. (1996). Document length normalization. *Information Processing & Management*, *32*(5), 619–633.
Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11–21.
Spark, A. (2016). Apache spark: Lightning-fast cluster computing. *URL http://spark. apache. org*.
Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, *60*(3), 538–556.
Subrahmonia, J. (2000). Similarity measures for writer clustering. In *Proceedings of the 7th international workshop on frontiers in handwriting recognition* (σσ. 541–546).
Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (σσ. 3104–3112).
Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.
Trask, R. L. (2007). *Language and linguistics: The key concepts*. Routledge.
Trudgill, P. (2000). *Sociolinguistics: An introduction to language and society*. Penguin UK.
Turell, M. T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech, Language & the Law*, *17*(2).
Turell, M. T. & Gavalda, N. (2012). Towards an index of idiolectal similitude (or distance) in forensic authorship analysis. *JL & Pol'y*, *21*, 495.
Turney, P. D. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *European conference on machine learning* (σσ. 491–502). Springer.
Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379–416.
Turney, P. D. & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, *21*(4), 315–346.
Turney, P. D. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141–188.
Van Rijsbergen, C. (1979). Information retrieval. dept. of computer science, university of glasgow. *URL: citeseer. ist. psu. edu/vanrijsbergen79information. html*, *14*.

Wales, K. (2014). *A dictionary of stylistics*. Routledge.
Wang, W. Y. (2017). ” liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
Weeds, J., Weir, D., & McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on computational linguistics* (σ. 1015). Association for Computational Linguistics.
Widdows, D. & Widdows, D. (2004). *Geometry and meaning*. Citeseer.
Woolley, S. C. & Howard, P. N. (2018). *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press.
Wu, L. Y., Fisch, A., Chopra, S., Adams, K., Bordes, A., & Weston, J. (2018). Starspace: Embed all the things! In *Thirty-second aaai conference on artificial intelligence*.
Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., & Weston, J. (2017). Starspace: Embed all the things! *arXiv preprint arXiv:1709.03856*.
Yang, X., Xu, G., Li, Q., Guo, Y., & Zhang, M. (2017). Authorship attribution of source code by using back propagation neural network based on particle swarm optimization. *PloS one*, *12*(11), e0187204.
Zappavigna, M. (2012). *Discourse of twitter and social media: How we use language to create affiliation on the web*. A&C Black.
Zhang, L., Zhao, J., & Xu, K. (2015). Who creates trends in online social media: The crowd or opinion leaders? *Journal of Computer-Mediated Communication*, *21*(1), 1–16.
Zheng, W., Yang, L., Genco, R. J., Wactawski-Wende, J., Buck, M., & Sun, Y. (2018). Sense: Siamese neural network for sequence embedding and alignment-free comparison. *Bioinformatics*.
Zuckermann, G. (2006). A new vision for israeli hebrew: Theoretical and practical implications of analyzing israel’s main language as a semi-engineered semito-european hybrid language. *Journal of Modern Jewish Studies*, *5*(1), 57–71.
Αρχάκης, Α. & Κονδύλη, Μ. (2004). Εισαγωγή σε ζητήματα κοινωνιογλωσσολογίας. Αθήνα: νήσος.
Γούτσος, Δ. (2012). Γλώσσα. κείμενο, ποικιλία, σύστημα. Αθήνα: Κριτική.
Γούτσος, Δ. & Φραγκάκη, Γ. (2015). Εισαγωγή στη γλωσσολογία σωμάτων κειμένων. *Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών*. Διαθέσιμο στο http://hdl.handle.net/11419/1930
Μαρκόπουλος, Γ. (2006). Ζητήματα υπολογιστικής γλωσσολογίας: Prolog και μορφολογική ανάλυση. *Περιοδικό ”Παρουσία”*, *Παράρτημα 69*, 209–243.
Μπακάκου-Ορφανού, Α. (2005). *Η λέξη της νέας ελληνικής στο γλωσσικό σύστημα και στο κείμενο*. Αθήνα: Περιοδικό ”Παρουσία” - Παράρτημα αρ. 65.

Πολίτου-Μαρμαρινού, Ε., Μικρός, Γ., & Δημουλά, Τ. (2013). Εφαρμογή υφομετρικών τεχνικών στην αναγνώριση πατρότητας κειμένου: Πρωτότυπα έργα και μεταφράσεις του παπαδιαμάντη. In *Πρακτικά του γ' διεθνούς συνεδρίου για τον αλ. παπαδιαμάντη* (Vol. Β, *3*, σσ. 361–392). Γ' Διεθνές Συνεδρίο για τον Αλ. Παπαδιαμάντη.

Φλώρου, Ε. (2017). *A neural model for metaphor detection* (Doctoral dissertation, National and Kapodestrian University of Athens).

Φραγκάκη, Γ. (2012). Η ιδιόλεκτος των κυβερνητικών εκπροσώπων.

Παράρτημα Α

Συχνότητες λέξεων

Α.1 ΣΕΚ

Α.2 Σώμα κειμένων Twitter

Πίνακας Α.1: οι 25 πιο συχνές λέξεις του ΣΕΚ

και	918730
του	617267
να	617267
της	545486
το	521219
την	423768
που	376707
με	352984
η	343697
από	318877
των	315778
για	287791
τα	255977
είναι	229042
ο	219317
σε	215353
τους	195337
τη	190935
θα	190319
στο	177673
οι	173868
στην	173775
τον	169218
δεν	164148
ότι	164148

Πίνακας Α.2: Συχνότερες λέξεις του Σώματος κειμένων Twitter, 2009

rt	8299
the	8232
to	5755
of	4574
in	4145
a	4061
&	3655
on	2437
το	2373
is	2228
for	2226
να	2197
i	2037
και	1662
για	1524
it	1383
that	1234
η	1193
την	1154
from	1118

Πίνακας Α.3: Συχνότερες λέξεις του Σώματος κειμένων Twitter, 2010

το	13706
να	12762
the	12634
και	12149
to	10927
rt	10282
για	7208
a	6909
in	6572
of	6494
δεν	6279
με	6092
i	6048
θα	5868
η	5604
που	5343
την	5220
ο	5080
τα	5063
είναι	5058
του	5044
στο	4570
for	3979
is	3715
on	3703
από	3649
&	3618
and	3598
σε	3526
της	3437

Πίνακας Α.4: Συχνότερες λέξεις του Σώματος κειμένων Twitter, 2011

το	56200
να	51310
και	49416
για	30925
η	26557
με	25746
του	23835
θα	23539
ο	23121
δεν	23067
την	23006
τα	22825
που	21583
στο	19450
της	18194
είναι	17489
σε	15972
από	13898
τον	13878
οι	13078
μου	11617
στην	11595
τη	9764
a	9689
τους	9341

Πίνακας Α.5: Συχνότερες λέξεις του Σώματος κειμένων Twitter, 2012

το	56200
να	51310
και	49416
για	30925
η	26557
με	25746
του	23835
θα	23539
ο	23121
δεν	23067
την	23006
τα	22825
που	21583
στο	19450
της	18194
είναι	17489
σε	15972
από	13898
τον	13878
οι	13078
μου	11617
στην	11595
τη	9764
a	9689
τους	9341

Πίνακας Α.6: Συχνότερες λέξεις του Σώματος κειμένων Twitter, 2013

να	165988
το	165327
και	145220
για	86267
η	82921
με	82312
ο	78287
που	74333
δεν	73334
την	73056
θα	72441
του	72035
τα	71981
της	55051
στο	52264
σε	50914
τον	46792
είναι	45893
οι	40990
από	36831
στην	35809
μου	35087
κ	30292
the	30104
τους	29625
τη	27610
μας	23984

Πίνακας Α.7: Συχνότερες λέξεις του Σώματος κειμένων Twitter, 2014

να	178734
το	174279
και	157081
με	92298
για	92150
ο	89799
η	87199
δεν	78460
τα	78077
που	77307
θα	76917
του	76234
την	73317
στο	57229
της	55249
σε	53930
τον	53694
είναι	44964
οι	42166
μου	42096
από	38053
κ	37139
στην	35424
τους	30289
τη	29821
τι	26597

Πίνακας Α.8: Συχνότερες λέξεις του Σώματος κειμένων Twitter, 2015

να	240824
το	236586
και	201451
ο	133975
η	128120
για	127682
με	120718
δεν	114286
θα	113625
του	110329
τα	107272
την	104704
που	103720
της	79361
τον	76229
στο	74278
σε	72171
οι	62147
είναι	61106
κ	55159
από	51434
στην	47921
μου	47643
τους	46297
τη	40551
μας	37316
τι	37233

Πίνακας Α.9: Συχνότερες λέξεις του Σώματος κειμένων Twitter, 2016

το	71858
να	71155
και	60730
για	41521
ο	40999
η	40236
του	38705
με	37900
τα	33025
δεν	32929
την	31981
που	31746
θα	31186
της	26167
στο	24191
τον	23796
σε	22552
οι	20492
είναι	19236
κ	18616
από	17489
στην	16058
τους	15401
μου	14750
τη	12754
τι	11317
μας	10722
των	10625

Πίνακας Α.10: Σώμα κειμένων Twitter

να	865087
το	861768
και	754445
για	459032
η	438764
με	433866
ο	431078
δεν	393482
θα	389232
του	384361
τα	380830
που	375887
την	370750
της	280366
στο	277348
σε	261145
τον	254742
είναι	235052
οι	215502
από	193483
μου	185682
στην	176856
κ	172170
τους	158842
τη	146269
μας	126278

Παράρτημα Β

Στατιστικά Γλωσσικά μοντέλα

Ένα στατιστικό γλωσσικό μοντέλο είναι η κατανομή της πιθανότητας ακολουθίας λέξεων. Δεδομένης μιας ακολουθίας λέξεων μήκους m $(w_1, w_2, ..., w_m)$, το γλωσσικό μοντέλο δίνει την πιθανότητα $P(w_1, w_2, ..., w_m)$ της ακολουθίας. Τα στατιστικά γλωσσικά μοντέλα αποτελούν βασικό συστατικό σε πολλές εφαρμογές Επεξεργασίας Φυσικής Γλώσσας, όπως για παράδειγμα σε εφαρμογές Αναγνώρισης φωνής, Μηχανικής μετάφρασης, Ανάκτησης περιεχομένου κ.ο.κ.

Στην γενικευμένη του μορφή, ένα γλωσσικό μοντέλο ορίζεται ως

$$p(w_1, w_2, ..., w_m) = p(w_1) * p(w_2|w_1) * p(w_3|w_1, w_2).....p(w_m|w_1, w_2, ...w_k)$$

.

Δεδομένης της φύσης των φυσικών γλωσσών, ένα γλωσσικό μοντέλο το οποίο το εκπαιδεύουμε σε ένα σώμα κειμένων δεν θα περιέχει την πλειονότητα των προτάσεων που δυνητικά μπορεί να σχηματιστούν, και φυσικά, ένα γλωσσικό μοντέλο δεν μπορεί να αποδώσει μηδενική πιθανότητα σε μια ακολουθία λέξεων. Για τον λόγο αυτό, εφαρμόζονται δύο τεχνικές: πρώτον, η υπόθεση ότι η πιθανότητα εμφάνισης μιας λέξης εξαρτάται μόνο από τις προηγούμενες n λέξεις που προηγούνται (τα μοντέλα αυτά είναι γνωστά ως μοντέλα n-γραμμάτων (n-gram models), με $n = 1, 2, 3, ...$) και, δεύτερον, ένα γλωσσικό μοντέλο πρέπει να μπορεί να χειρίζεται νέες λέξεις, λέξεις δηλαδή που δεν υπήρχαν στο σώμα κειμένων όταν έγινε η εκπαίδευσή του, αλλά, προφανώς, η πιθανότητα εμφάνισης τους δεν είναι μηδενική. Για την αντιμετώπιση αυτού του προβλήματος εφαρμόζονται τεχνικές εξομάλυνσης (smoothing), οι οποίες έχουν στόχο να εξασφαλίσουν ότι το μοντέλο θα δώσει μικρή αλλά όχι μηδενική πιθανότητα σε λέξεις που δεν υπήρχαν στο σώμα εκπαίδευσης.

Έτσι, στην περίπτωση 1-γραμμάτων, η πιθανότητα εμφάνισης των λέξεων w1,w2,w3 απλοποιείται σε

$$P(w_1, w_2, w_3) = P(w_1)P(w_2|w_1)P(w_3|w_2, w_3) = P(w_1)P(w_2)P(w_3)$$

.

Στο μοντέλο αυτό, συνεπώς, η πιθανότητα κάθε λέξης είναι η πιθανότητα (συχνότητα) εμφάνισής της στο σώμα κειμένων και το άθροισμα των πιθανοτήτων όλων των λέξεων του σώματος κειμένων είναι 1.

Στην περίπτωση των μοντέλων διγραμμάτων, η πιθανότητα εμφάνισης 4 λέξεων, για παράδειγμα, μεταφράζεται ως

$$p(w_1, w_2, w_3, w_4) = P(w_1, < start >)P(w_2|w_1)P(w_3|w_2)P(w_4| < end >)$$

,

όπου <start>, <end> είναι βοηθητικοί χαρακτήρες, που δηλώνουν την αρχή και το τέλος μιας ακολουθίας λέξεων.

Για την εξομάλυνση, τυπικές τεχνικές, που εφαρμόζονται στην πράξη, είναι η τεχνική Good-Turing, η τεχνική Kneser-Ney κλπ. (Chen & Goodman, 1999).

Παράρτημα Γ

Μέγιστη Κατάβαση Κλίσης-Στοχαστική Κατάβαση Κλίσης (Gradient Descent/Stochastic Gradient Descent)

Η μέγιστη κατάβαση κλίσης είναι ένας επαναληπτικός αλγόριθμος βελτιστοποίησης πρώτης τάξης. Στόχος είναι να βρεθεί ένα τοπικό ελάχιστο μιας συνάρτησης και αυτό επιτυγχάνεται κάνοντας σε κάθε επανάληψη του αλγόριθμου ένα βήμα προς την αντίθετη κατεύθυνση της μέγιστης κλίσης στο συγκεκριμένο σημείο. Στην περίπτωση της Μηχανικής μάθησης η συνάρτηση προς ελαχιστοποίηση (objective function), J(θ), είναι η μέση τιμή της συνάρτησης κόστους (ή ζημίας) (loss or cost function), η οποία χρησιμοποιείται για να ποσοτικοποιήσει το μέγεθος της διαφοράς μεταξύ της πραγματικής τιμής ενός φαινομένου και της τιμής που εκτιμά το μοντέλο. Τυπικά, σε εφαρμογές ταξινόμησης (classification) με νευρωνικά δίκτυα, η συνάρτηση κόστους που χρησιμοποιείται είναι η διεντροπία (cross entropy):

$$H(p,q) = E_p[logq] = H(p) + D_{KL}(p||q) = \sum_x p(x)logq(x)$$

,

όπου $H(p) = xp(x)logp(x)$ η εντροπία της κατανομής p και $D_{KL}(p||q) = \sum_x p(x)logq(x)$ η απόκλιση Kullback-Leibler (σχετική εντροπία της κατανομής p ως προς την κατανομή q). Η ελαχιστοποίηση της συνάρτησης $J(\theta)$ έχει ως στόχο την συνολική ελαχιστοποίηση του σφάλματος του μοντέλου, να βρεθούν δηλαδή οι παράμετροι θ που ελαχιστοποιούν το κόστος. Η μέγιστη κατάβαση κλίσης

ελαχιστοποιεί την $J(\theta)$ επαναληπτικά, όπου σε κάθε επανάληψη οι παράμετροι θ υπολογίζονται ως εξής:

$$\theta \leftarrow \theta - \eta \nabla J(\theta) = w - \eta \sum_{i=1}^{n} J_i(\theta)/n$$

,

όπου $J_i(\theta)$ είναι η συνάρτηση κόστους της i-οστής παρατήρησης, n ο αριθμός των παρατηρήσεων και η παράμετρος η ορίζει το μέγεθος του βήματος ενημέρωσης του θ, γνωστή στη βιβλιογραφία και ως ρυθμός μάθησης (learning rate).

Όπως παρατηρούμε, η ενημέρωση των παραμέτρων θ γίνεται αφού πρώτα υπολογιστεί η μέση τιμή της κλίσης της συνάρτησης κόστους για όλα τα σημεία εκπαίδευσης. Εάν ο αριθμός των σημείων n είναι ιδιαίτερα μεγάλος, ο υπολογισμός αυτός μπορεί να γίνει ιδιαίτερα χρονοβόρος.

Η Στοχαστική Κατάβαση Κλίσης είναι στοχαστική προσεγγιστική μέθοδος της Μέγιστης Κατάβασης Κλίσης, κατά την οποία αντί να υπολογιστεί η κλίση της συνάρτησης στόχου σε όλο το σύνολο σημείων εκπαίδευσης, επιλέγεται ως προσέγγιση της κλίσης ένα τυχαίο υποσύνολο m σημείων με m « n. Αυτό έχει ως αποτέλεσμα ο αλγόριθμος ελαχιστοποίησης να γίνει εξαιρετικά αποδοτικός όταν ο αριθμός των σημείων εκπαίδευσης n είναι τεράστιος.

Στην περίπτωση που $m = 1$, η μέθοδος καλείται online stochastic gradient και ως συνάρτηση $J(\theta)$ χρησιμοποιείται προσεγγιστικά το κόστος ενός σημείου i, $J_i(\theta)$ για κάθε επανάληψη του αλγόριθμου. Ο αλγόριθμος δηλαδή γίνεται:

$$\theta \leftarrow w - \eta J_i(\theta)$$

Στην πράξη χρησιμοποιείται $m > 1$ αλλά πολύ μικρότερο του n, οπότε ο αλγόριθμος υπολογίζει την νέα τιμή του θ έχοντας υπολογίσει προσεγγιστικά την κλίση σε m σημεία. Η εκδοχή αυτή του αλγόριθμου καλείται Κατάβαση Κλίσης Δέσμης (mini-batch gradient descent).

Ο αλγόριθμος στοχαστικής κατάβασης κλίσης είναι ο πλέον διαδεδομένος αλγόριθμος εκπαίδευσης νευρωνικών δικτύων.

Παράρτημα Δ

Νευρωνικά Δίκτυα και ο Αλγόριθμος ανάστροφης διάδοσης σφάλματος (Backpropagation)

Ένα πολυεπίπεδο νευρωνικό δίκτυο (multilayer neural network, multilayer perceptron, MLP) είναι μία συνάρτηση

$$f : R^D \rightarrow R^L$$

, με D η διάσταση του διανύσματος εισόδου x και L η διάσταση του διανύσματος εξόδου f(x) ορισμένη ως:

$$f(x) = G(b^{(2)} + W^{(2)} * (s(b^{(1)} + W^{(1)}x)))$$

Με διανύσματα κατωφλίου b(1), b(2), πίνακες βαρών W(1), W(2) και συναρτήσεις ενεργοποίησης G, s.

Τυπικές επιλογές συναρτήσεων για την συνάρτηση s είναι η συνάρτηση υπερβολικής εφαπτομένης (tanh), η σιγμοειδής συνάρτηση

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

ή πιο πρόσφατα, γραμμικοί ανορθωτές (rectifier linear units, RELUs) (Glorot et al., 2011), ενώ για τη συνάρτηση G σε προβλήματα ταξινόμησης η πλέον διαδεδομένη συνάρτηση είναι η συνάρτηση softmax

$$softmax(x_i) = \frac{e^{s_i}}{\sum_{c=1}^{n} e^{s_c}}$$

Ο αλγόριθμος ανάστροφης διάδοσης σφάλματος (backpropagation algorithm) είναι αλγόριθμος εκπαίδευσης νευρωνικών δικτύων. Αναζητά το ελάχιστο μιας

συνάρτησης σφάλματος J(θ) στον παραμετρικό χώρο, που ορίζεται από το θ, με τη χρήση της μεθόδου Μέγιστης Κατάβασης Κλίσης. Μια τιμή της παραμέτρου θ που ελαχιστοποιεί την συνάρτηση σφάλματος θεωρείται ως λύση της διαδικασίας εκμάθησης.

Για την περίπτωση νευρωνικών δικτύων ταξινόμησης σε προβλήματα δύο κλάσεων, η τυπική αρχιτεκτονική που χρησιμοποιείται είναι νευρωνικό δίκτυο με μία μονάδα εξόδου, με συνάρτηση ενεργοποίησης την σιγμοειδή (sigmoid, logistic) συνάρτηση και συνάρτηση κόστους την διεντροπία. Η αρχιτεκτονική αυτή δίνει την ευχέρεια πιθανοθεωρητικής ερμηνείας των αποτελεσμάτων, καθώς οι τιμές της εξόδου μπορούν να θεωρηθούν ως η πιθανότητα της κλάσης δεδομένης της εισόδου.

Η διεντροπία για μια είσοδο με n εξόδους έχει τη μορφή:

$$E = -\sum_{i=1}^{n}(t_i * log(x_i) + (1-t_i) * log(1-x_i))$$

όπου t (target) η τιμή, x η έξοδος του νευρωνικού στη θέση i. Η έξοδος x_i είναι η συνάρτηση ενεργοποίησης

$$x_i = \frac{1}{1+e^{s_i}}$$

και

$$s_i = \sum_{j=1} x_j w_{ji}$$

.

Η εκπαίδευση του νευρωνικού δικτύου έγκειται στην εύρεση των παραμέτρων (βαρών) w_{ji} για τις οποίες η διεντροπία είναι ελάχιστη. Εδώ πρέπει να σημειωθεί πως η διεντροπία δεν είναι κυρτή (convex) συνάρτηση των παραμέτρων w_{ji} και, κατά συνέπεια, το ελάχιστο που θα βρεθεί τυπικά θα είναι τοπικό και όχι ολικό ελάχιστο.

Ο αλγόριθμος ανάστροφης διάδοσης σφάλματος είναι η εφαρμογή του αλυσιδωτού κανόνα (chain rule) στους νευρώνες κάθε επιπέδου.

Το πρώτο βήμα είναι να υπολογιστεί η κλίση του σφάλματος ως προς τα βάρη του επιπέδου εξόδου:

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial x_i} * \frac{\partial x_i}{\partial s_i} * \frac{\partial s_i}{\partial w_{ji}}$$

Η εξίσωση αυτή μας δίνει την κλίση των βαρών του τελευταίου επιπέδου του δικτύου.

Στη συνέχεια, χρειαζόμαστε την κλίση της συνάρτησης κόστους στα προηγούμενα επίπεδα του δικτύου:

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial s_j} * \frac{\partial s_j}{\partial w_k j}$$

από την οποία προκύπτει

$$\frac{\partial E}{\partial w_{kj}} = \sum_{i=1}^{n} (x_i - t_i) w_{ij} x_j (1 - x_j) x_k$$

.

Συνεπώς, για να υπολογιστεί η κλίση του σφάλματος σε κάθε επίπεδο

$$\frac{\partial E}{\partial w_i j}$$

χρειάζεται να υπολογιστεί αναδρομικά το

$$\frac{\partial E}{\partial s_j}$$

και τέλος να πολλαπλασιαστεί με

$$\frac{\partial s_j}{\partial w_k j} = x_k$$

.

Για προβλήματα ταξινόμησης με περισσότερες από δύο κλάσεις, χρησιμοποιούμε την επέκταση της σιγμοειδούς συνάρτησης, γνωστή και ως softmax. Σε αυτή την περίπτωση:

$$x_i = \frac{e^{s_i}}{\sum_{c=1}^{n} e^{s_c}}$$

η συνάρτηση κόστους είναι:

$$E = \sum_{i=1}^{n} t_i * log(x_i))$$

.

Από εδώ προκύπτει ότι η κλίση στο επίπεδο εξόδου είναι:

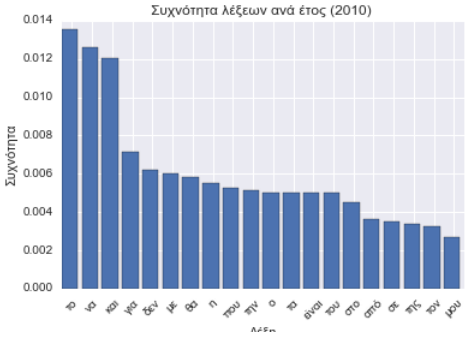
$$\frac{\partial E}{\partial w_j i} = \sum_i \frac{\partial E}{\partial si} \frac{\partial s_i}{\partial w_j k} = (x_i t_i) x_j$$

και για το αμέσως προηγούμενο επίπεδο j:

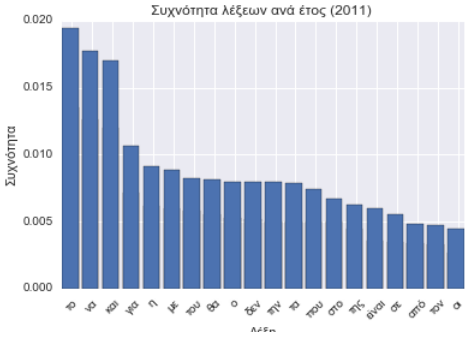
$$\frac{\partial E}{\partial s_j} = \sum_{i=1}^{n} (x_i)(w_{ji}) x_j (1 - x_j)$$

Παράρτημα Ε

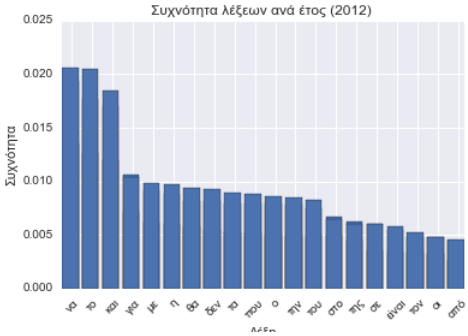
Διαγράμματα



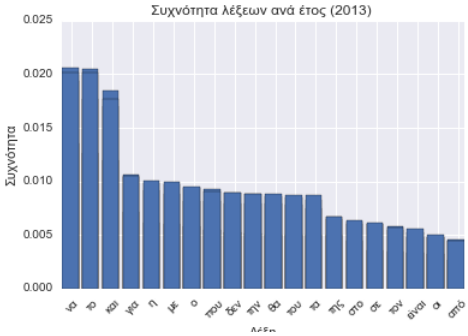
(a) 2010



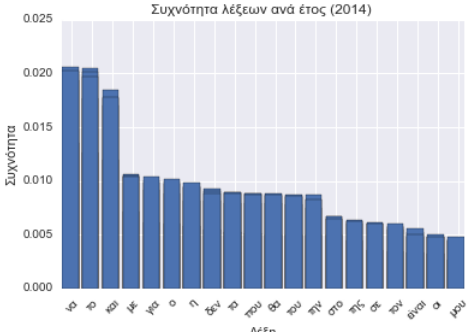
(b) 2011



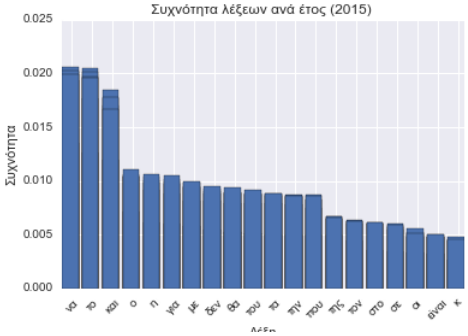
(c) 2012



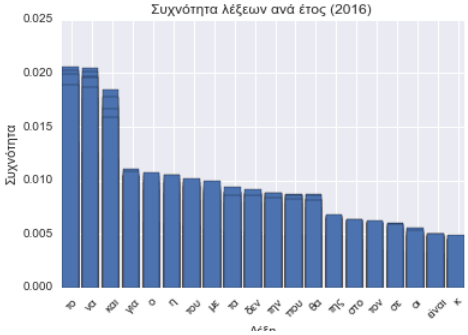
(d) 2013



(e) 2014

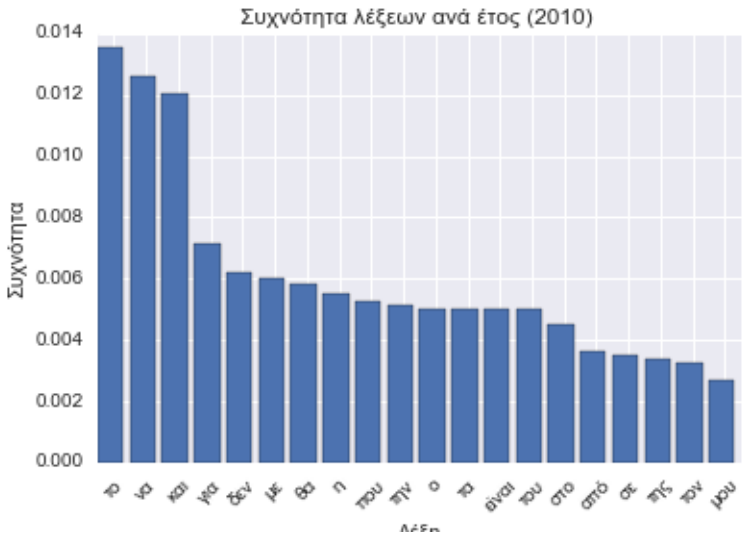


(f) 2015

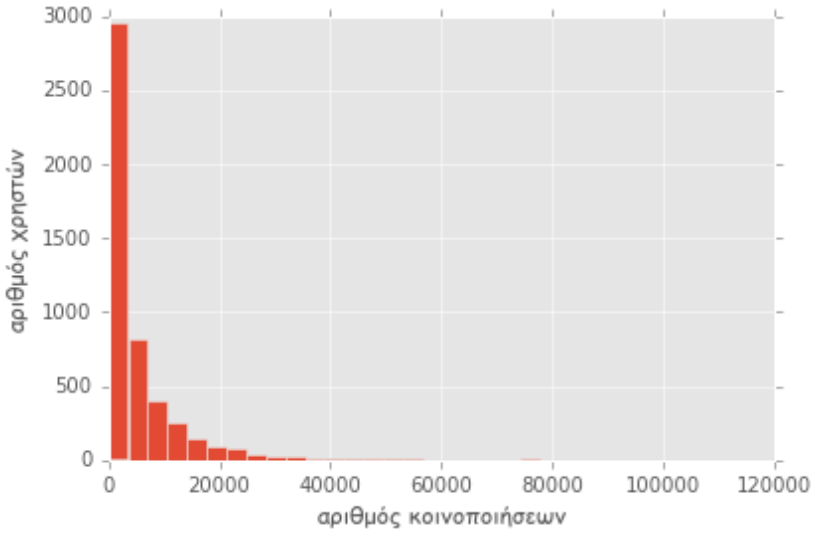


(g) 2016

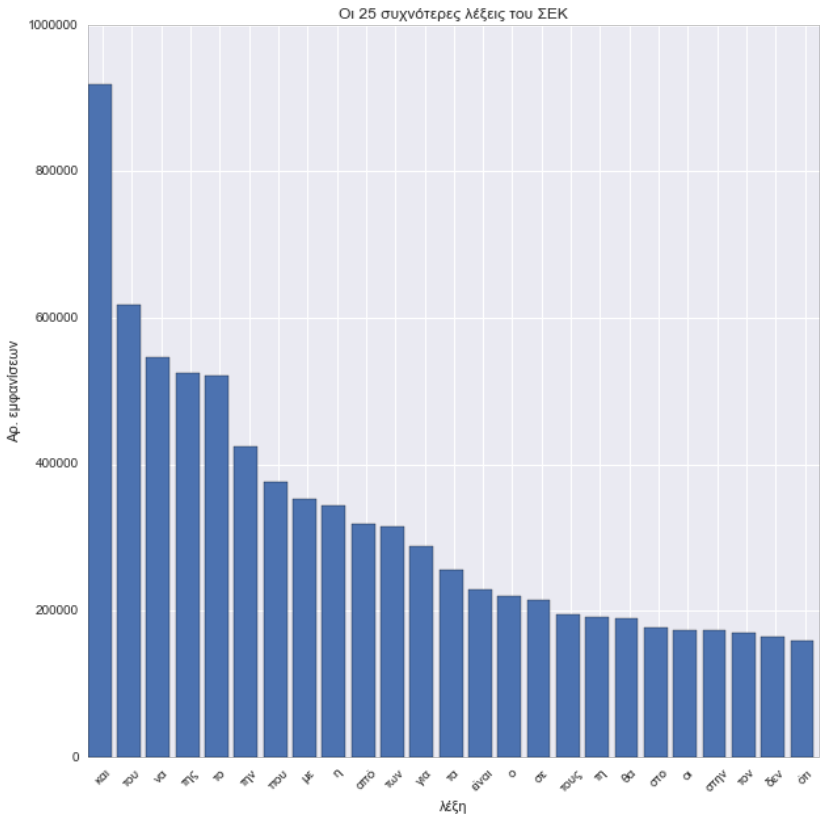
Διάγραμμα Ε.1: Κατανομή των 20 πιο συχνών λέξεων ανά έτος



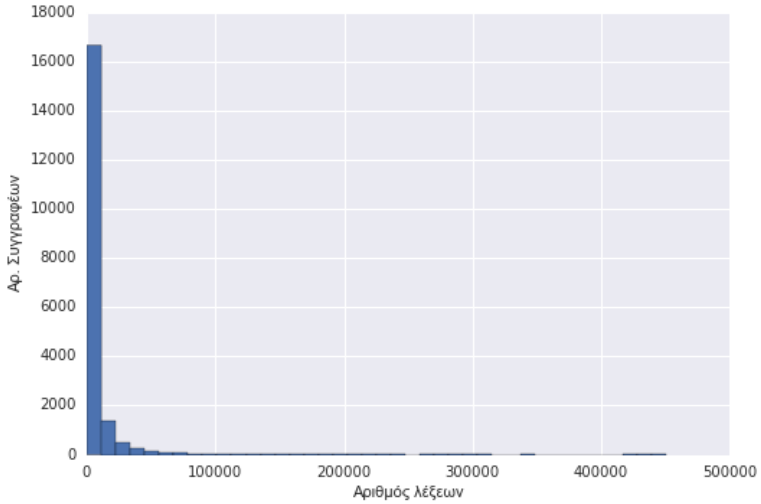
Διάγραμμα Ε.2: Κατανομή λέξεων ανά ανάρτηση 1



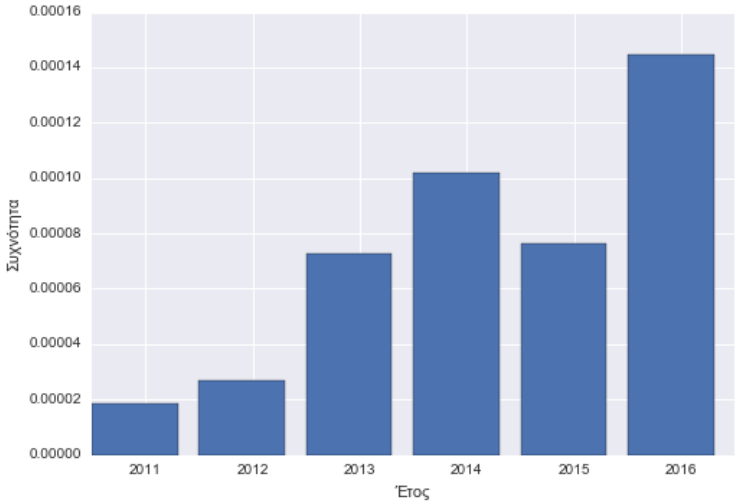
Διάγραμμα Ε.3: Ιστόγραμμα συχνότητας αναρτήσεων



Διάγραμμα Ε.4: 25 Συχνότερες λέξεις του ΣΕΚ



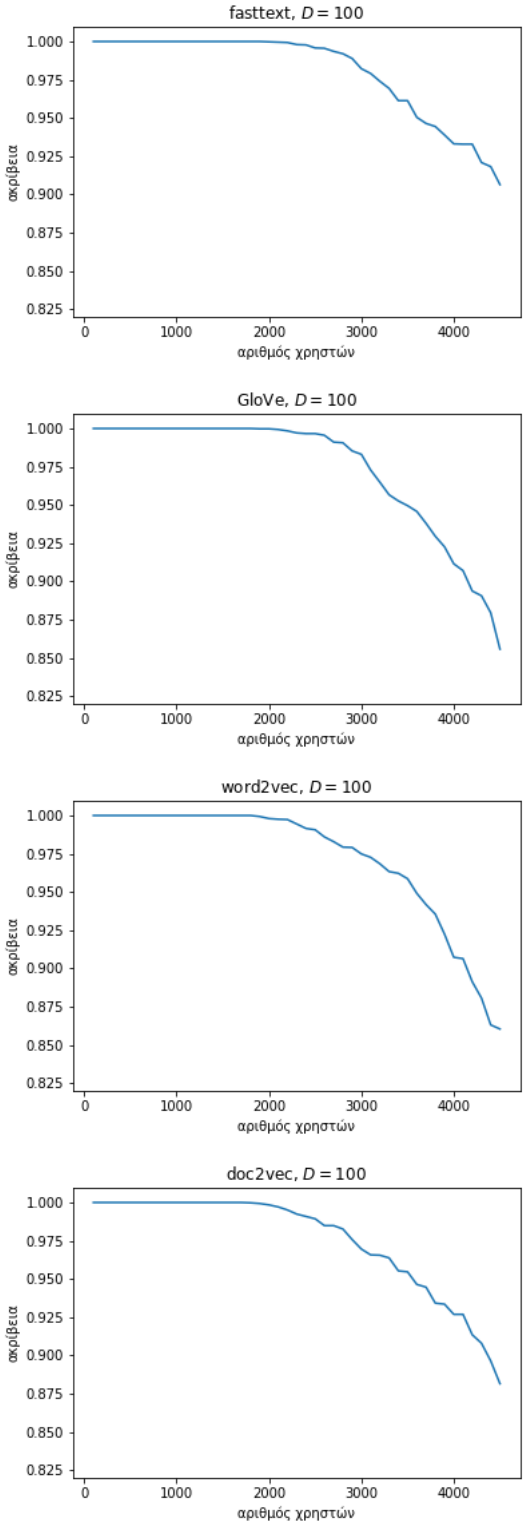
Διάγραμμα Ε.5: Κατανομή λέξεων/συγγραφέα στο Blog Authorship Corpus



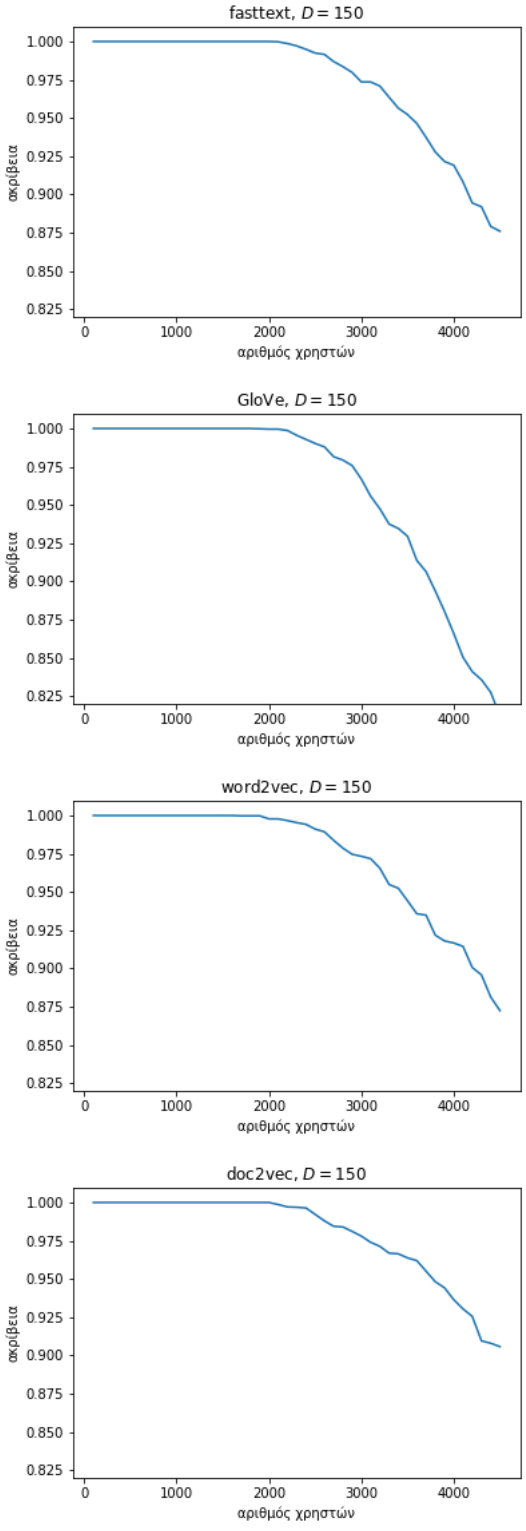
Διάγραμμα Ε.6: Συχνότητα εμφάνισης της λέξης «Σοβιετία» στο Σώμα κειμένων Twitter

Ε.1 Διαγράμματα σώματος κειμένων Twitter

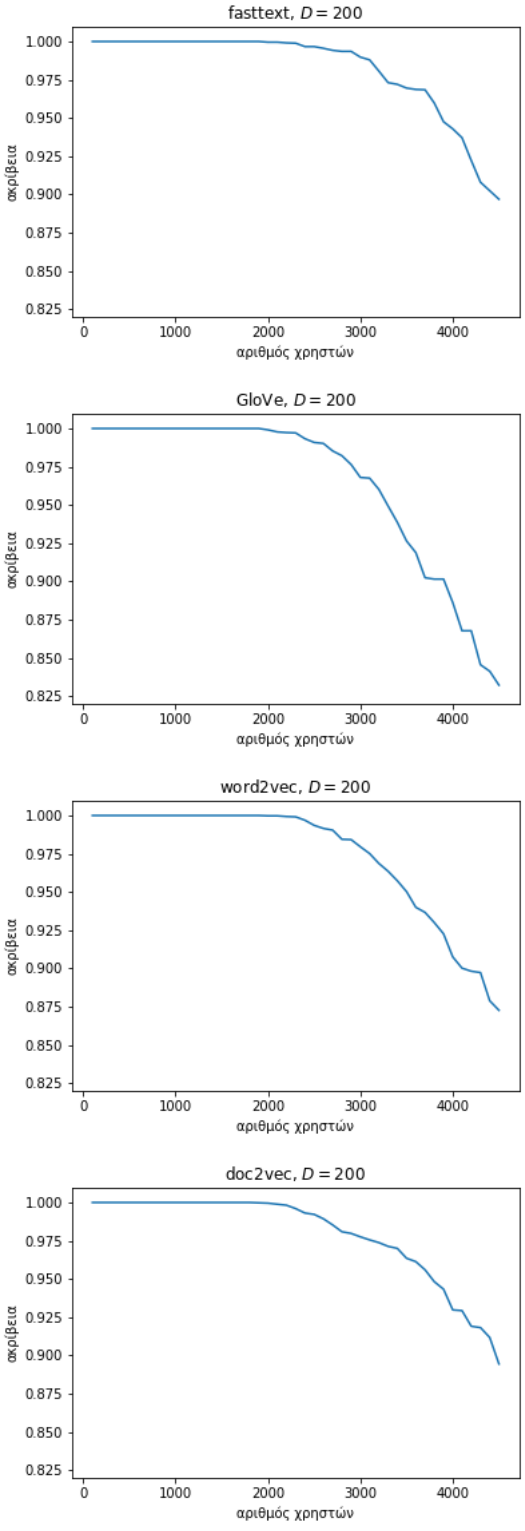
Διάγραμμα Ε.7: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=100



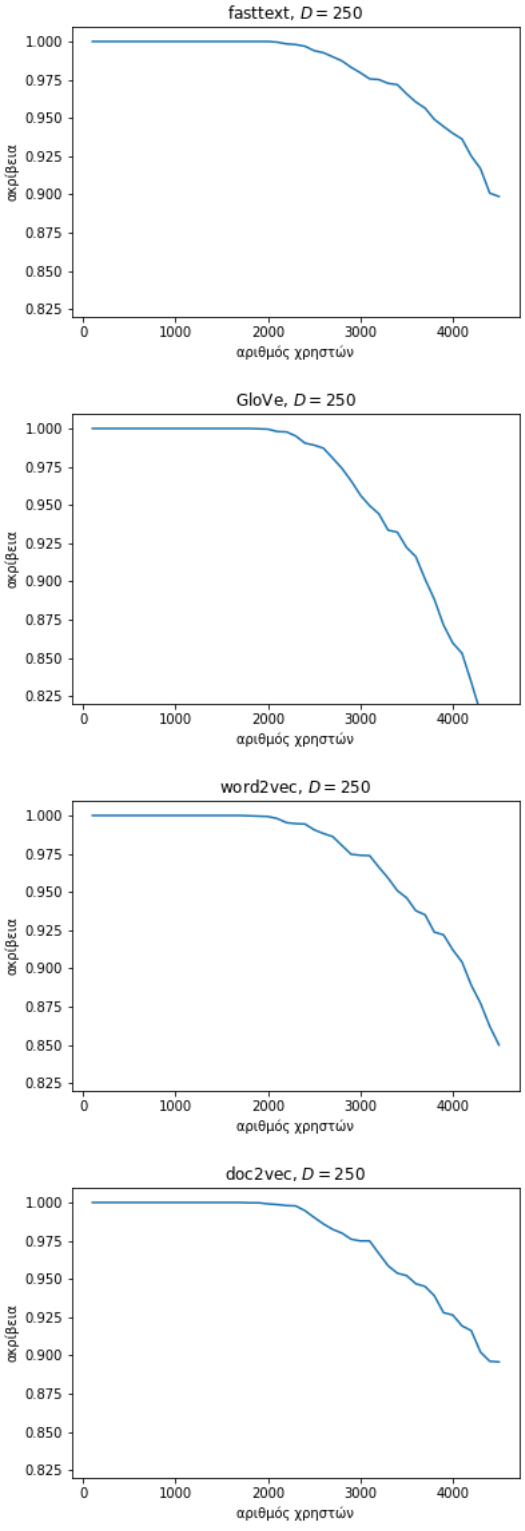
Διάγραμμα Ε.8: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=150



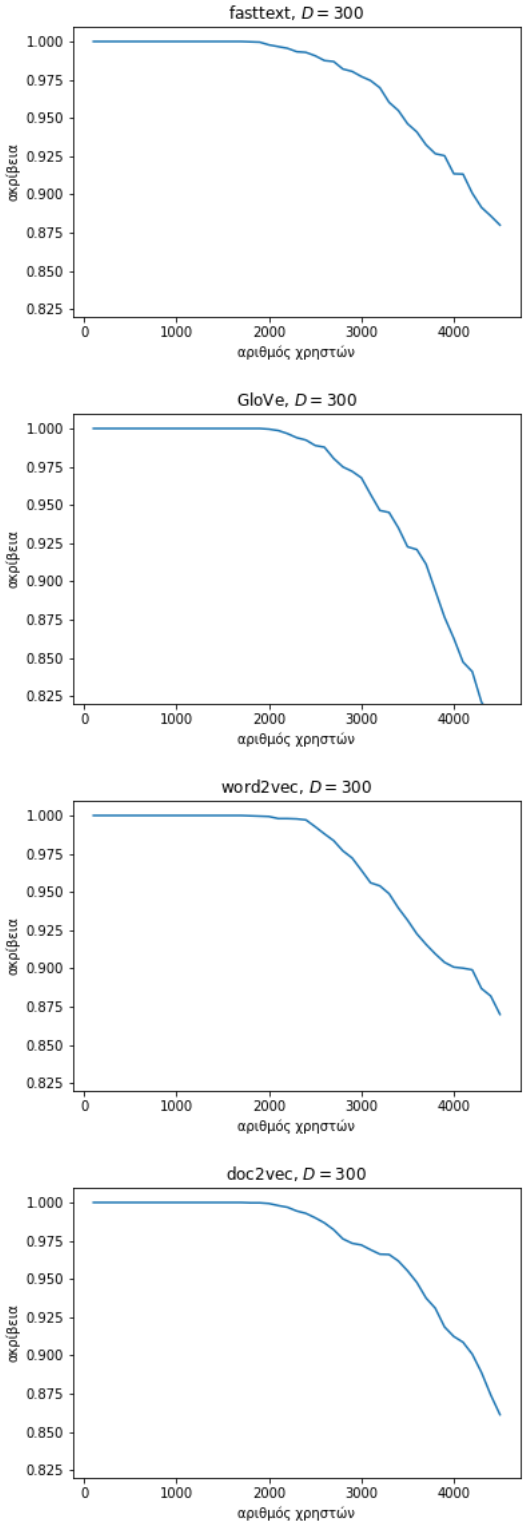
Διάγραμμα Ε.9: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=200



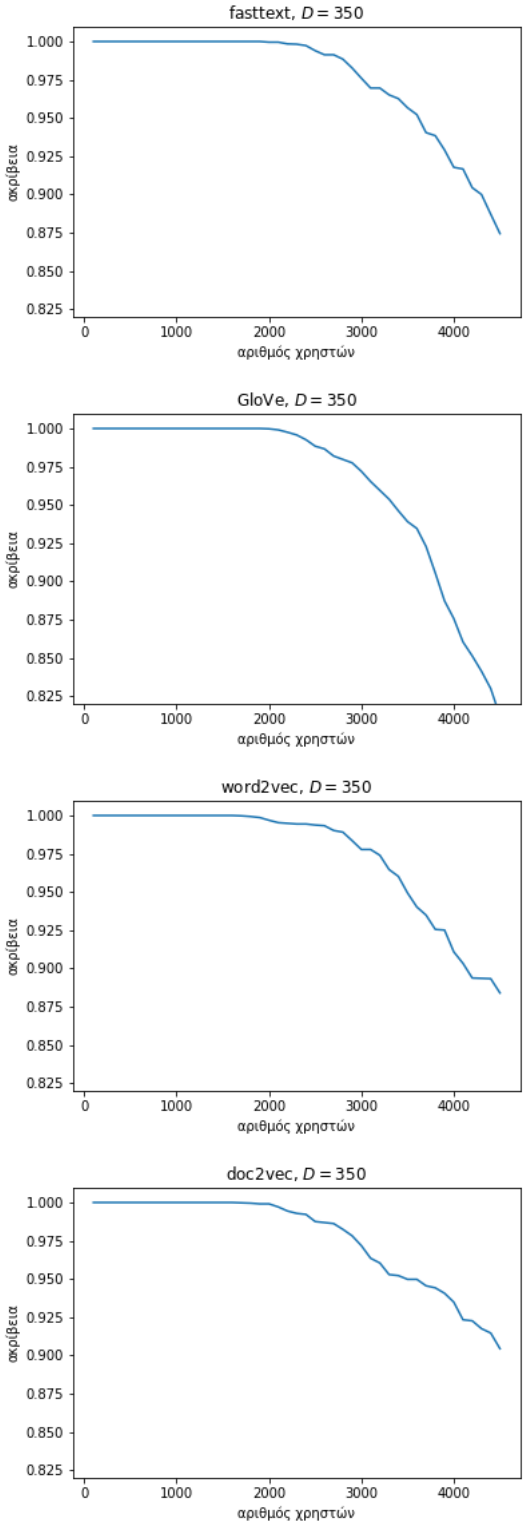
Διάγραμμα Ε.10: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=250



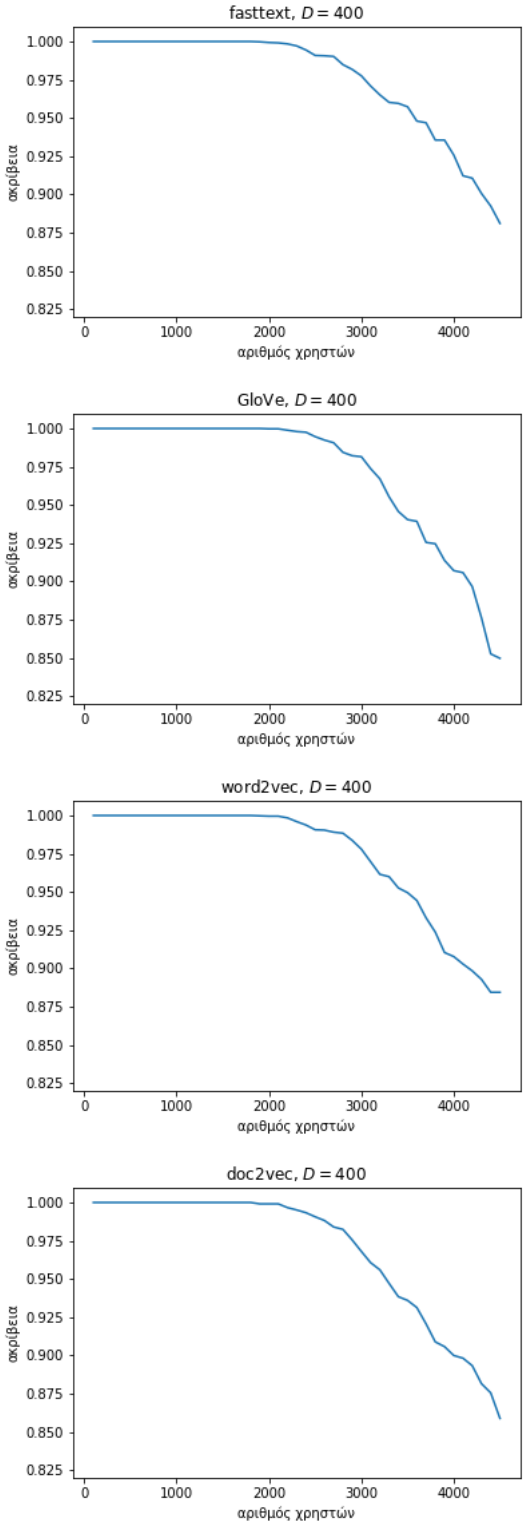
Διάγραμμα Ε.11: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=300



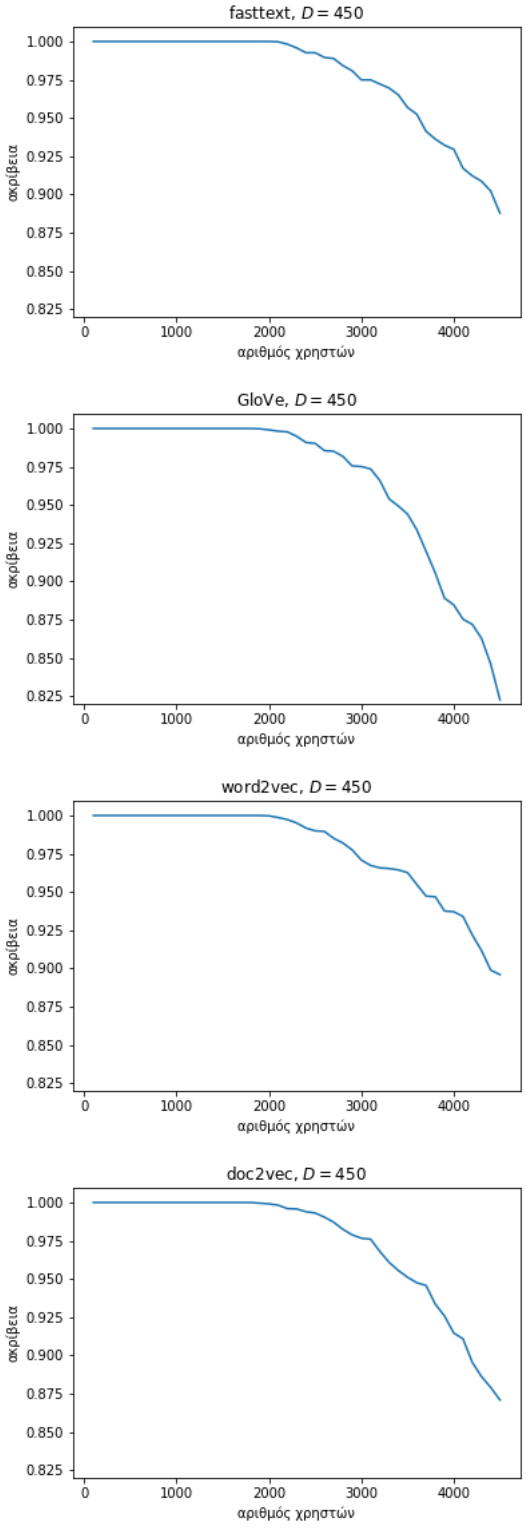
Διάγραμμα Ε.12: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=350



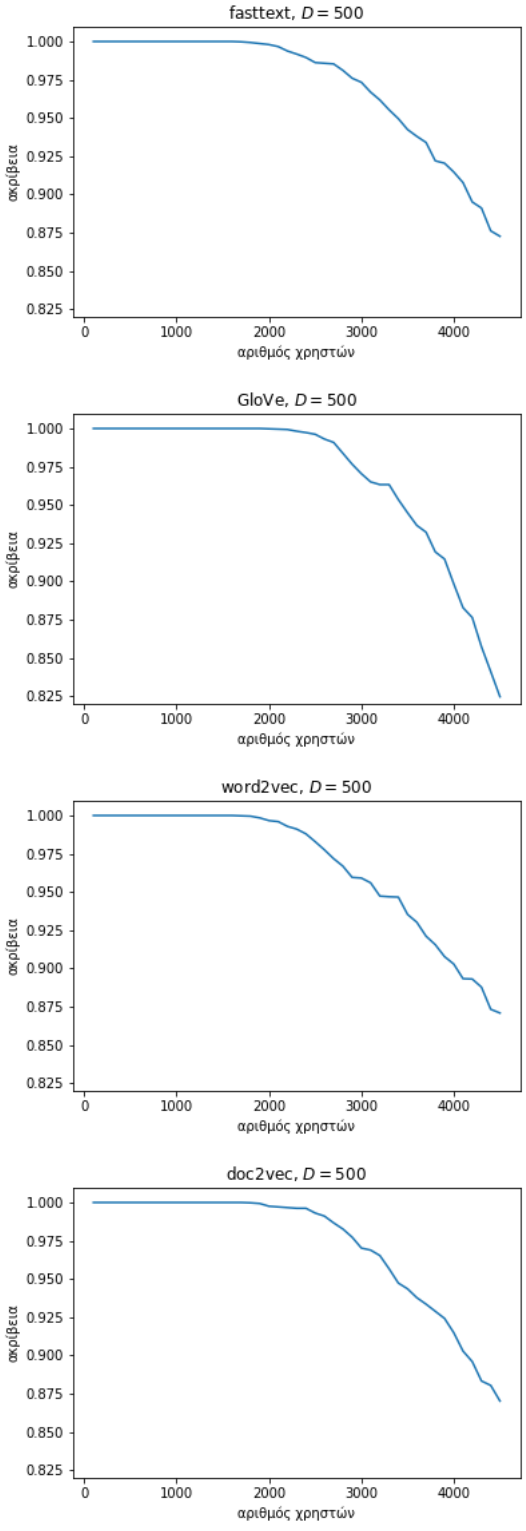
Διάγραμμα Ε.13: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=400



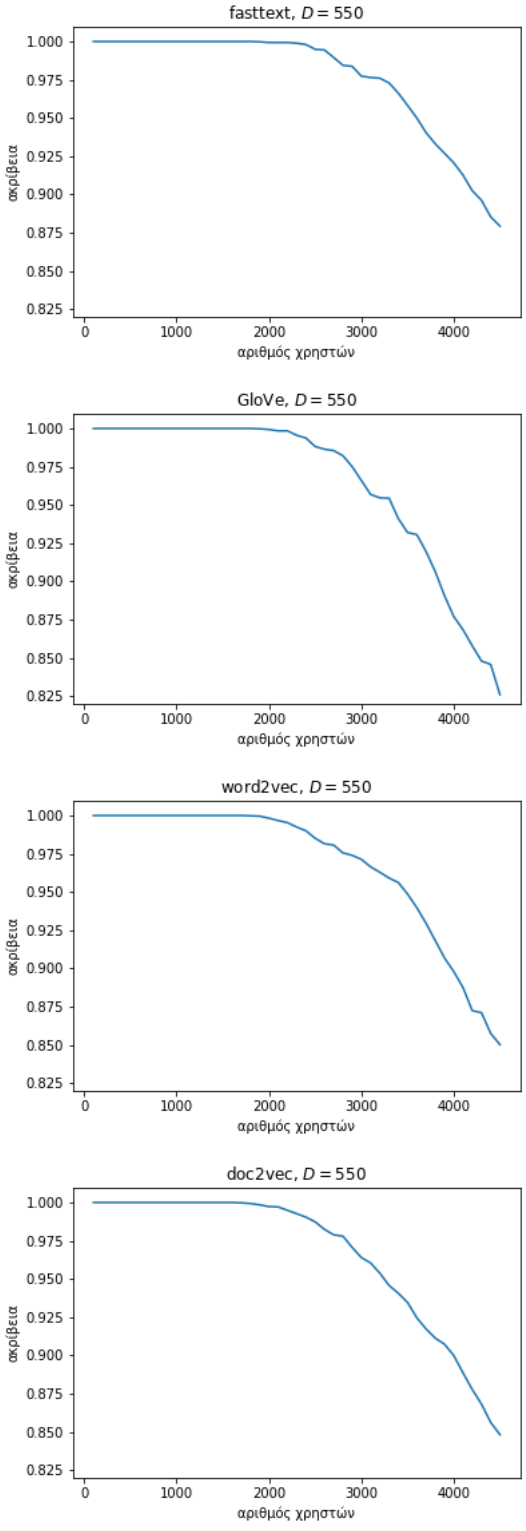
Διάγραμμα Ε.14: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=450



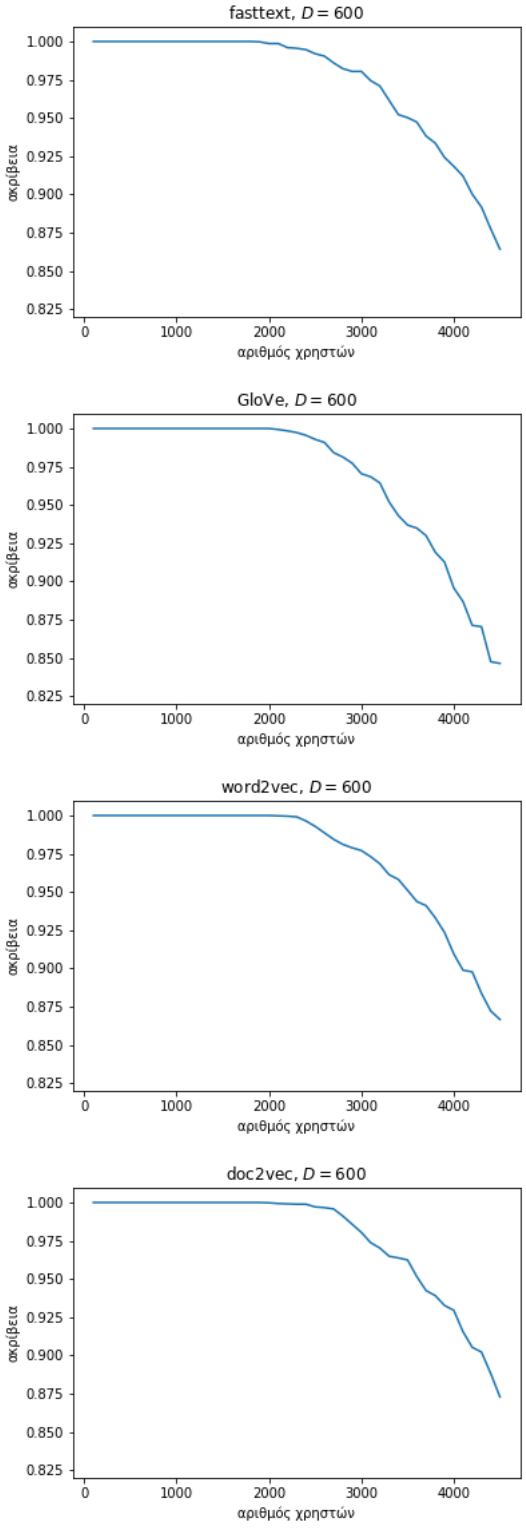
Διάγραμμα Ε.15: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=500



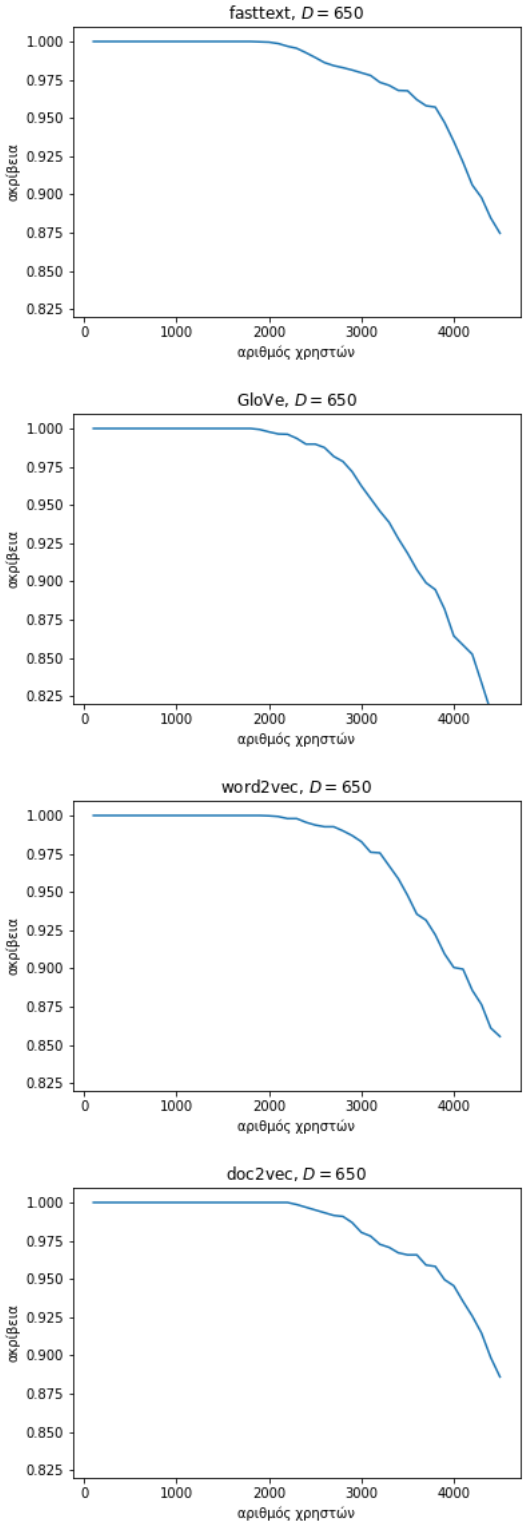
Διάγραμμα Ε.16: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=550



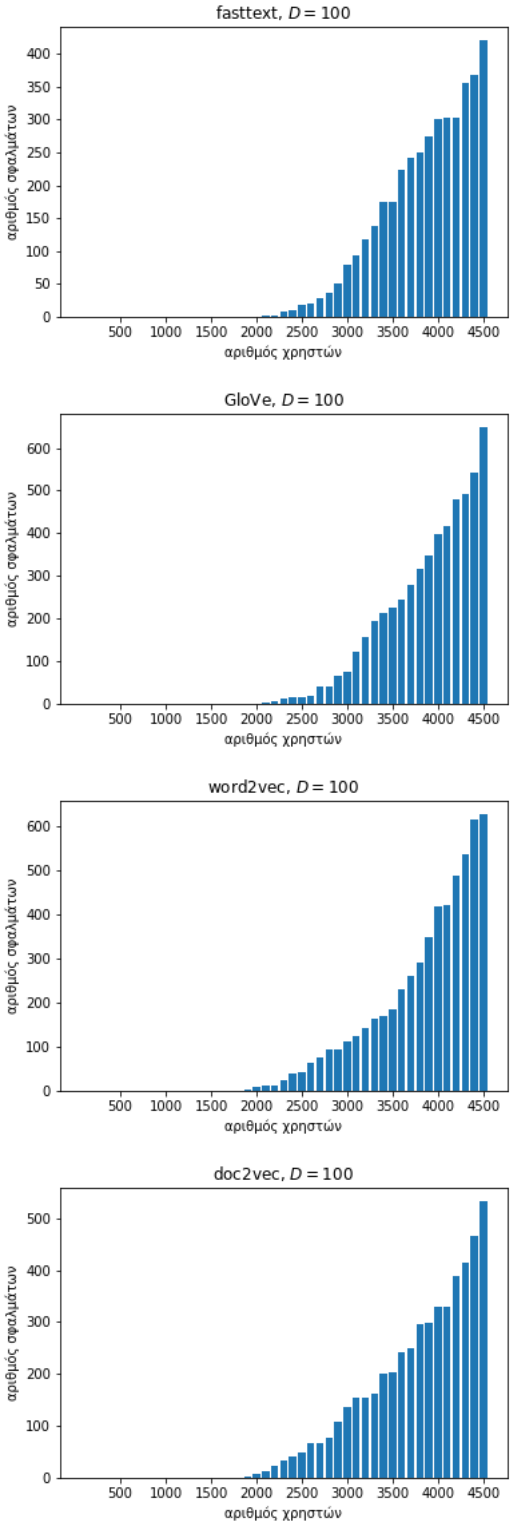
Διάγραμμα Ε.17: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=600



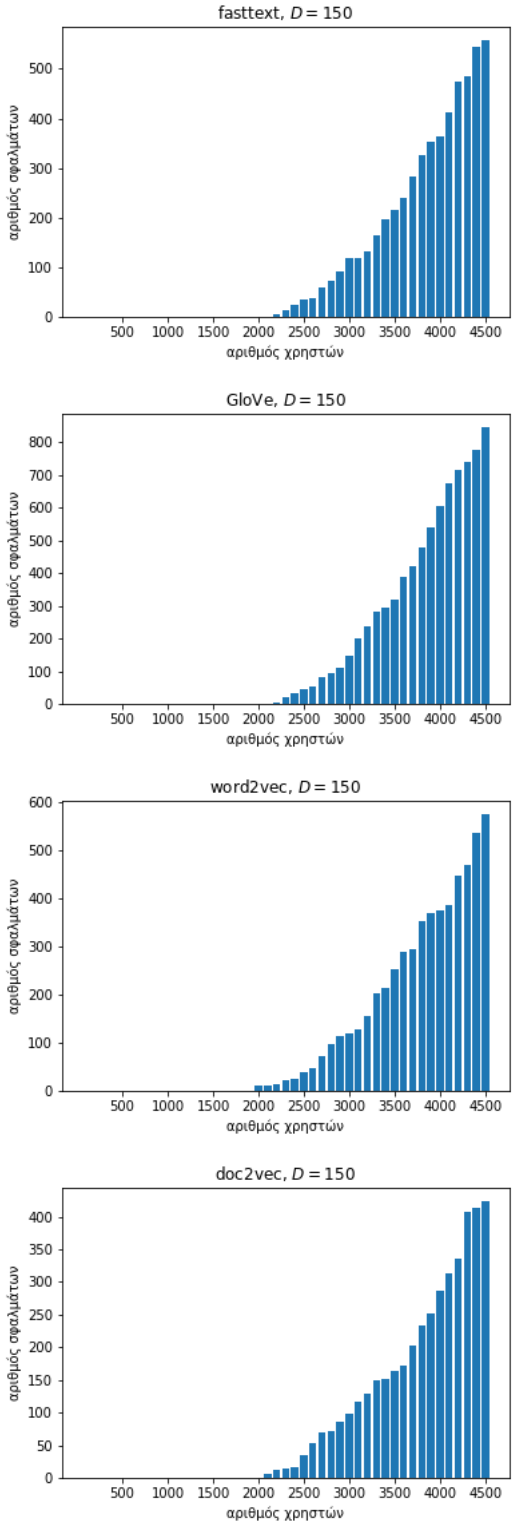
Διάγραμμα Ε.18: Ακρίβεια μοντέλων στο Σώμα κειμένων Twitter, D=650



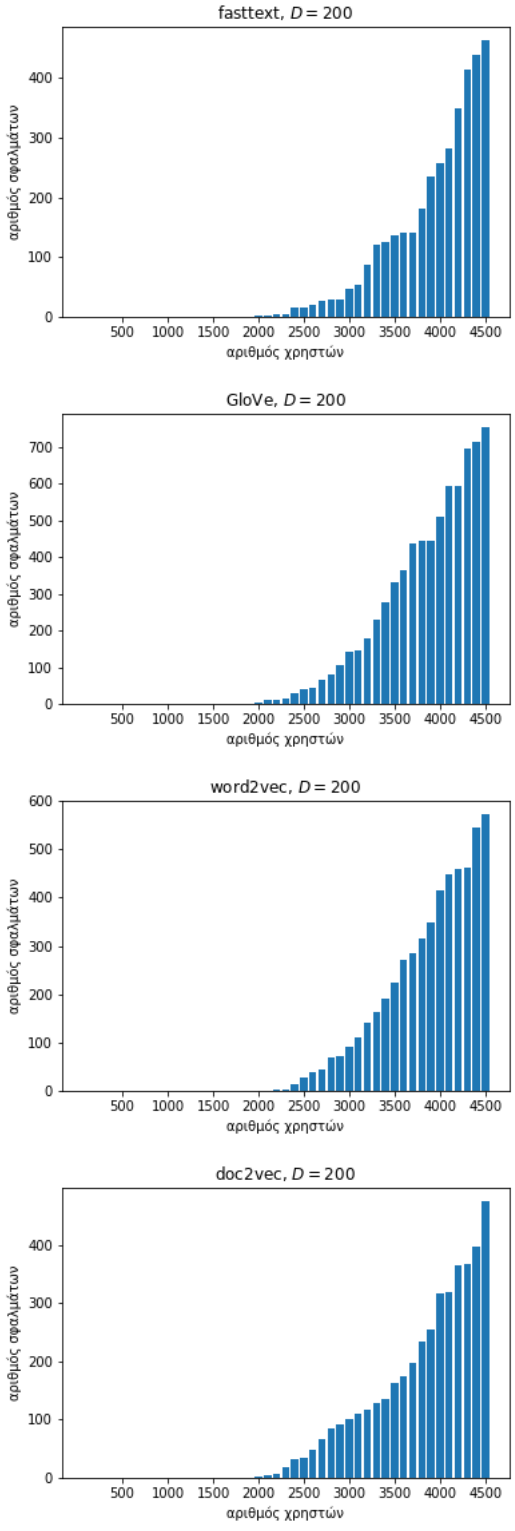
Διάγραμμα Ε.19: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=100



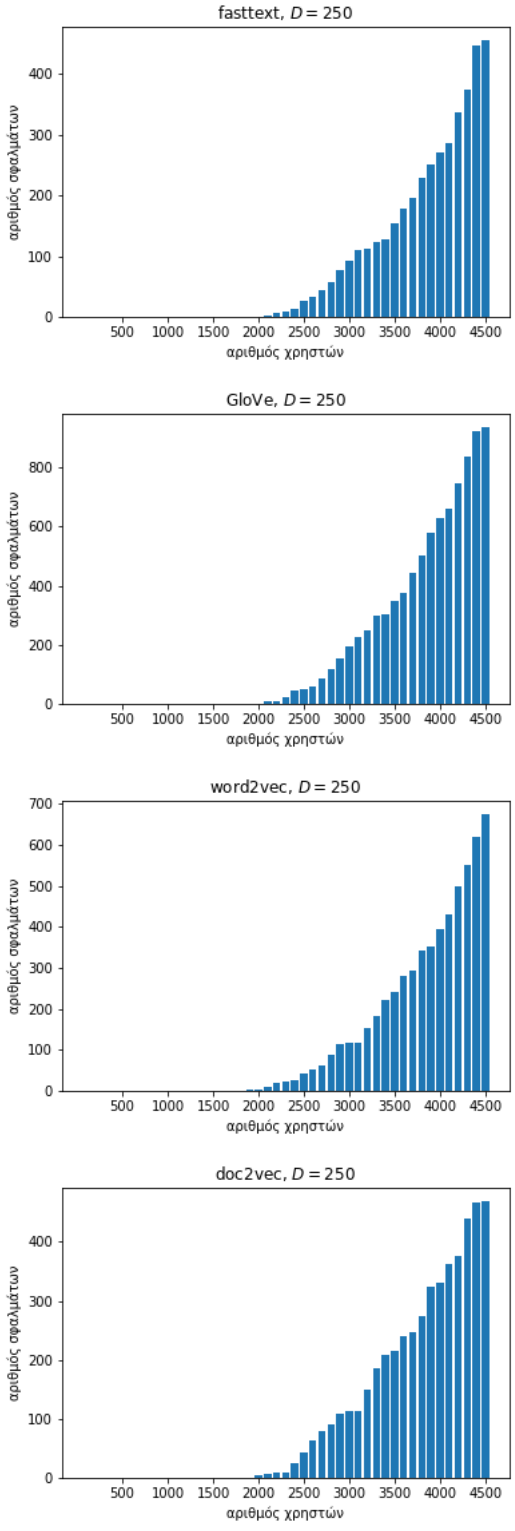
Διάγραμμα Ε.20: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=150



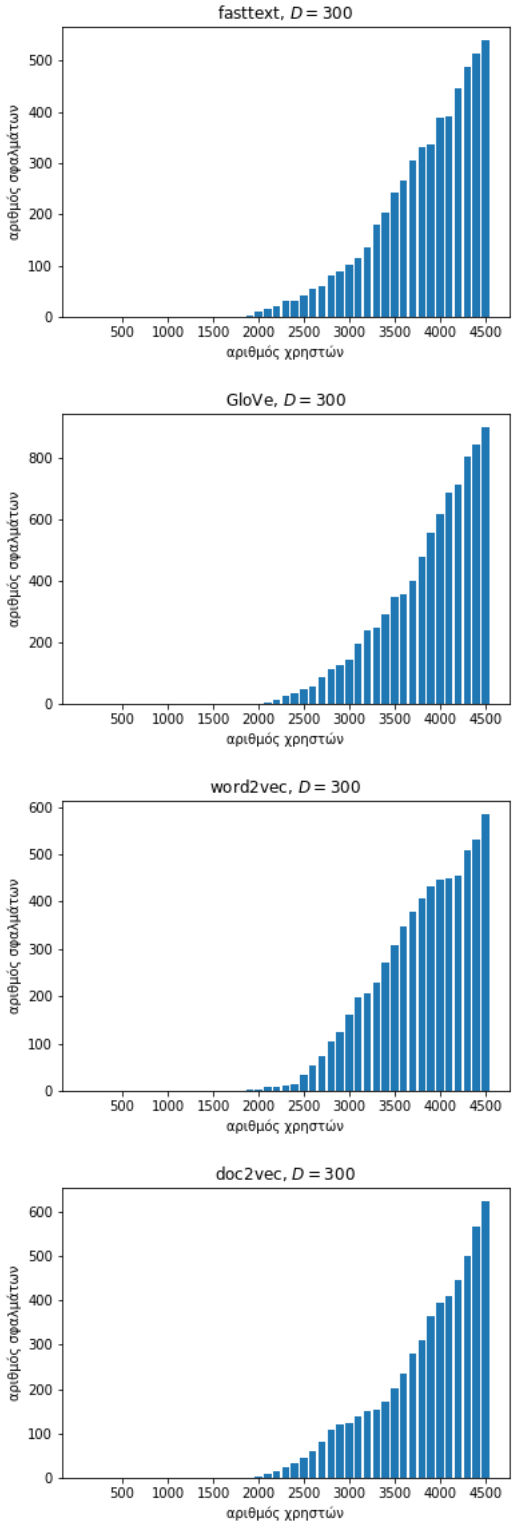
Διάγραμμα Ε.21: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=200



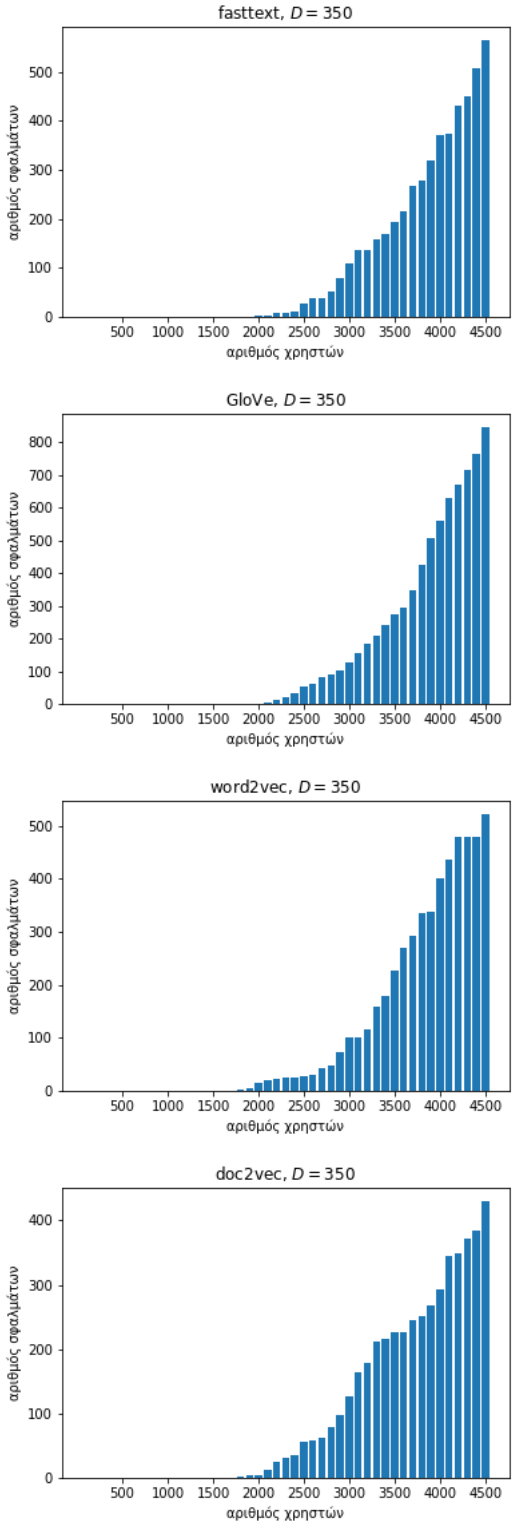
Διάγραμμα Ε.22: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=250



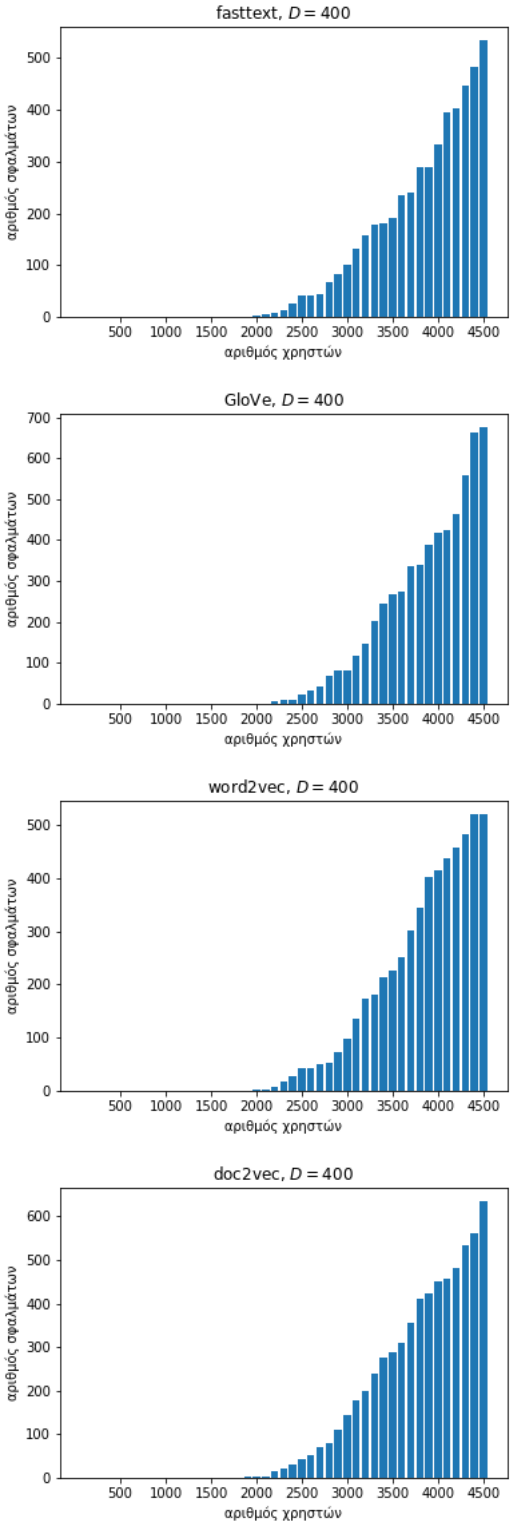
Διάγραμμα Ε.23: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=300



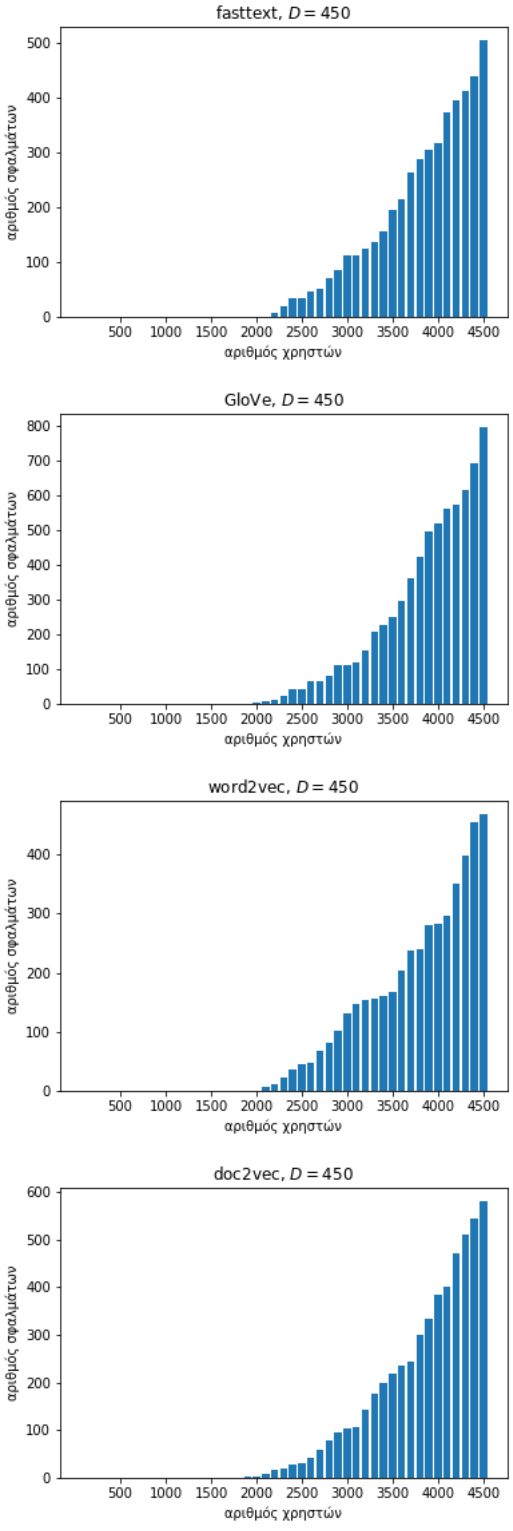
Διάγραμμα Ε.24: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=350



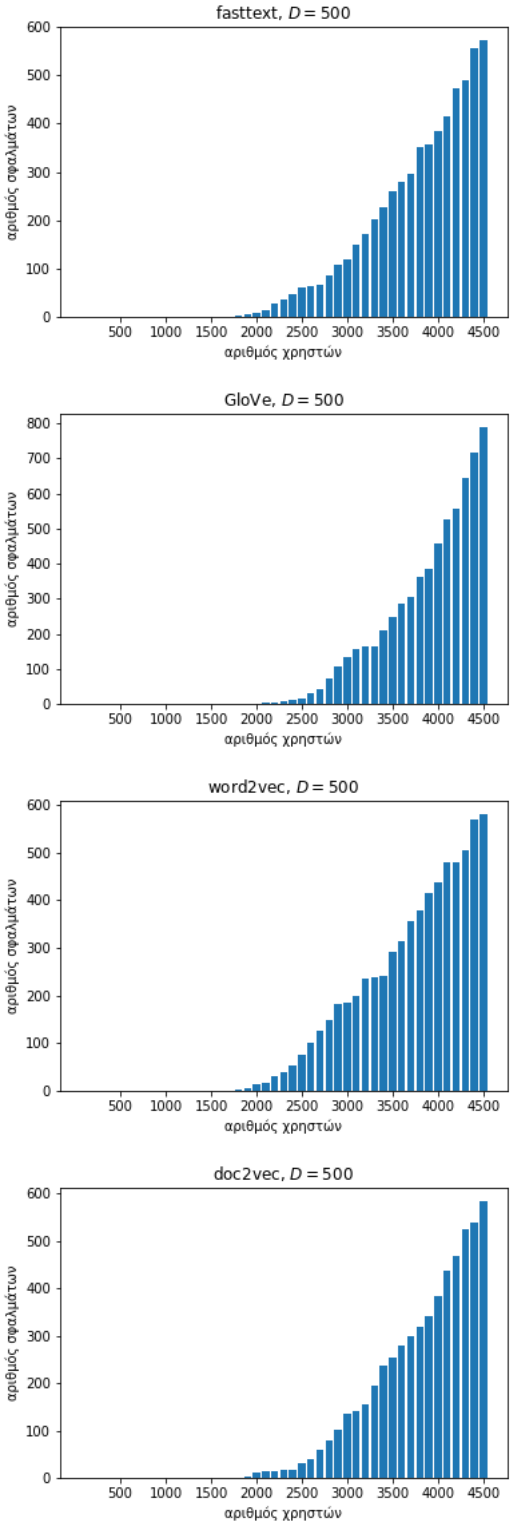
Διάγραμμα Ε.25: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=400



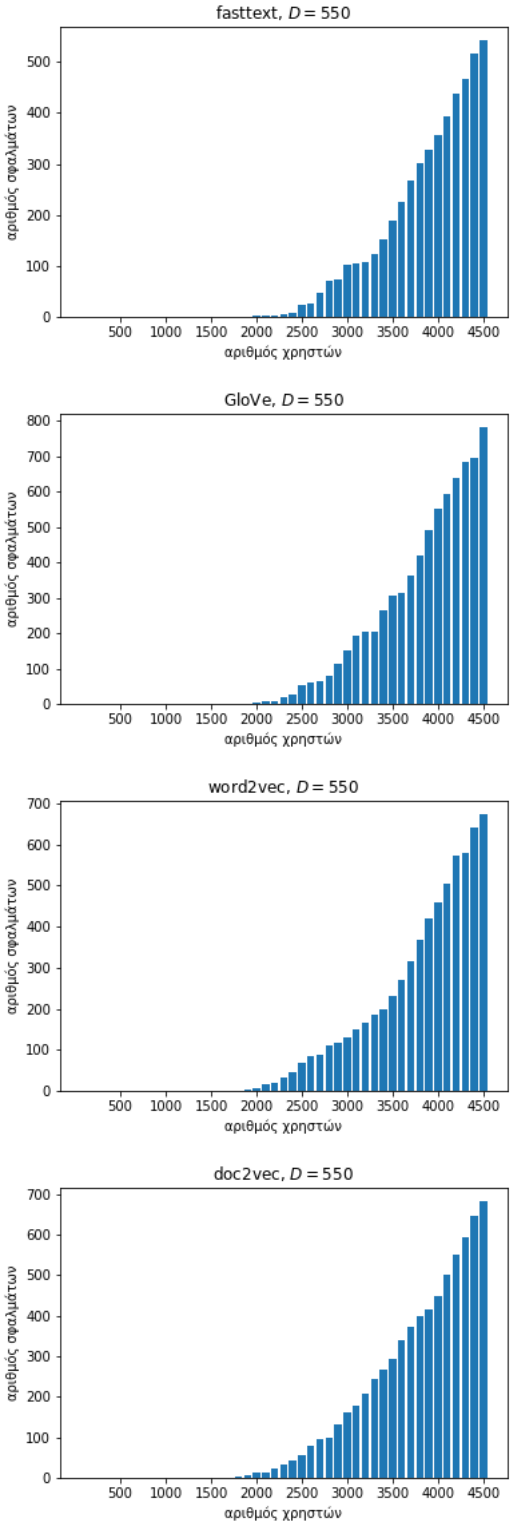
Διάγραμμα Ε.26: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=450



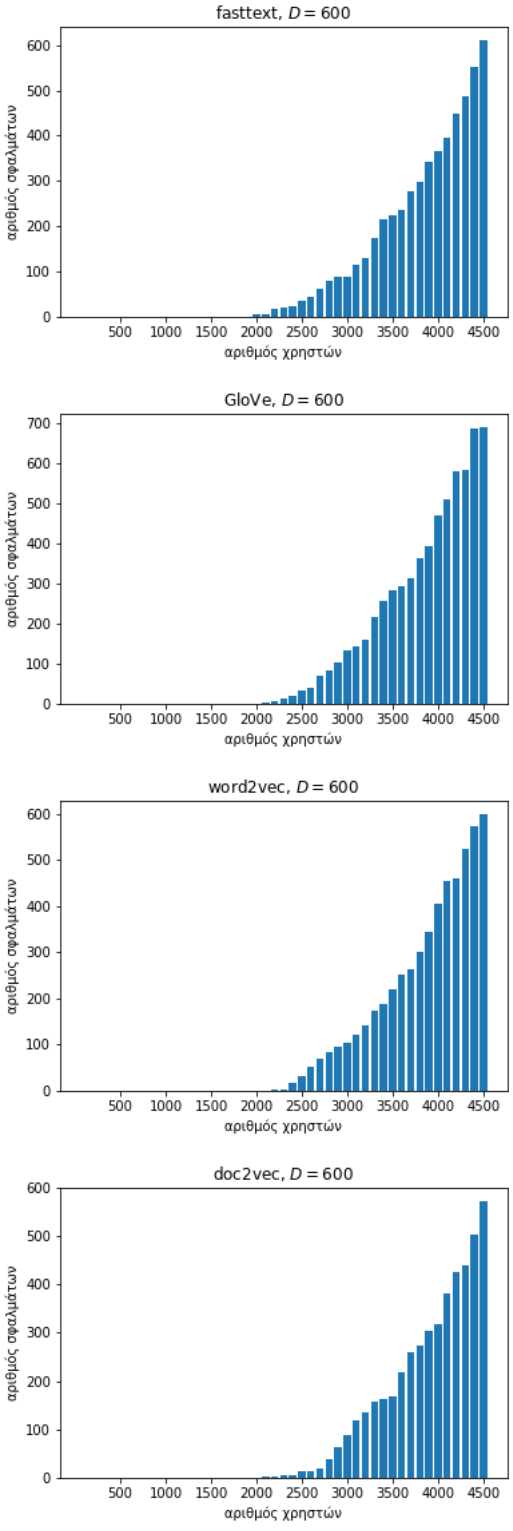
Διάγραμμα Ε.27: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=500



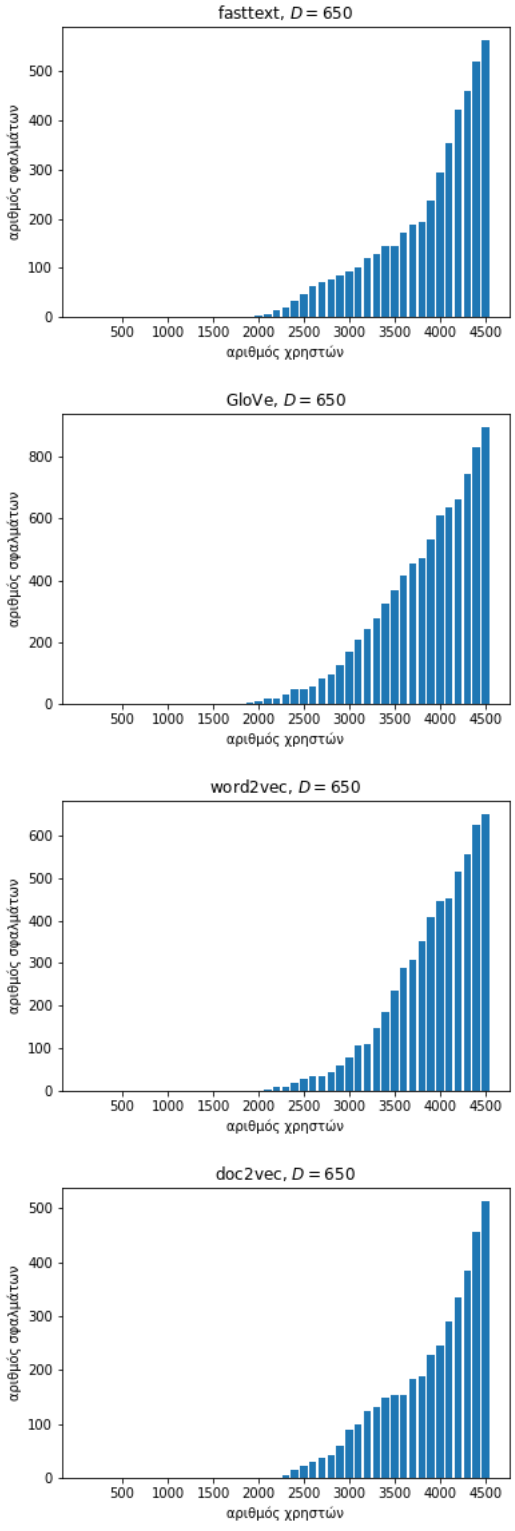
Διάγραμμα Ε.28: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=550



Διάγραμμα Ε.29: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=600

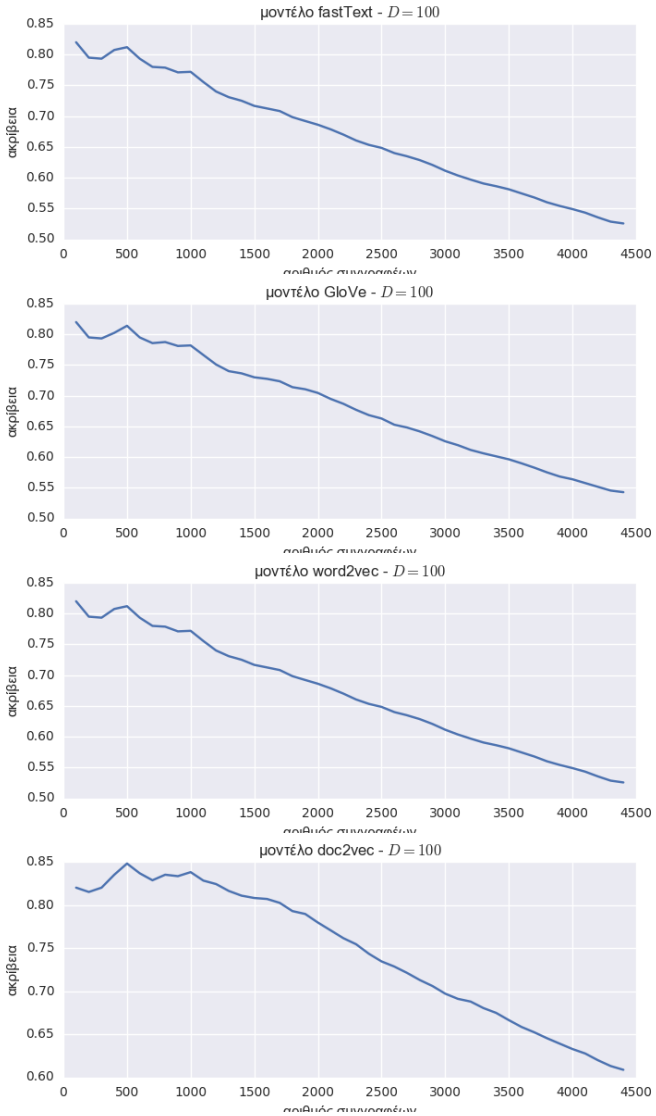


Διάγραμμα Ε.30: Σφάλμα μοντέλων στο Σώμα κειμένων Twitter, D=650

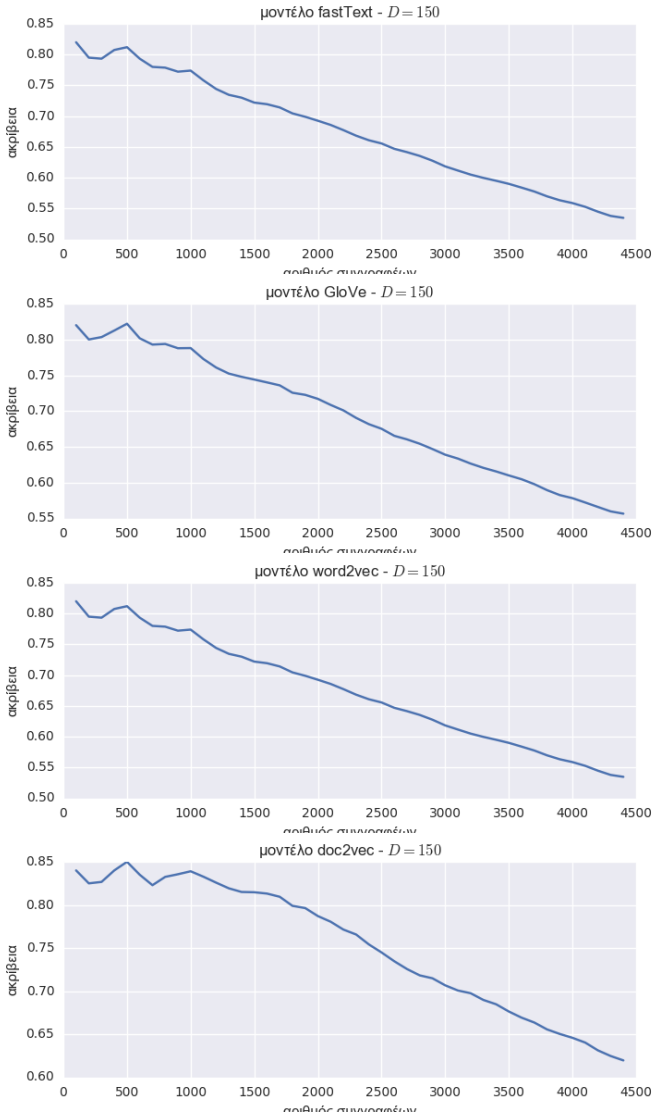


Ε.2 Διαγράμματα Σώματος κειμένων ιστολογίων BAC

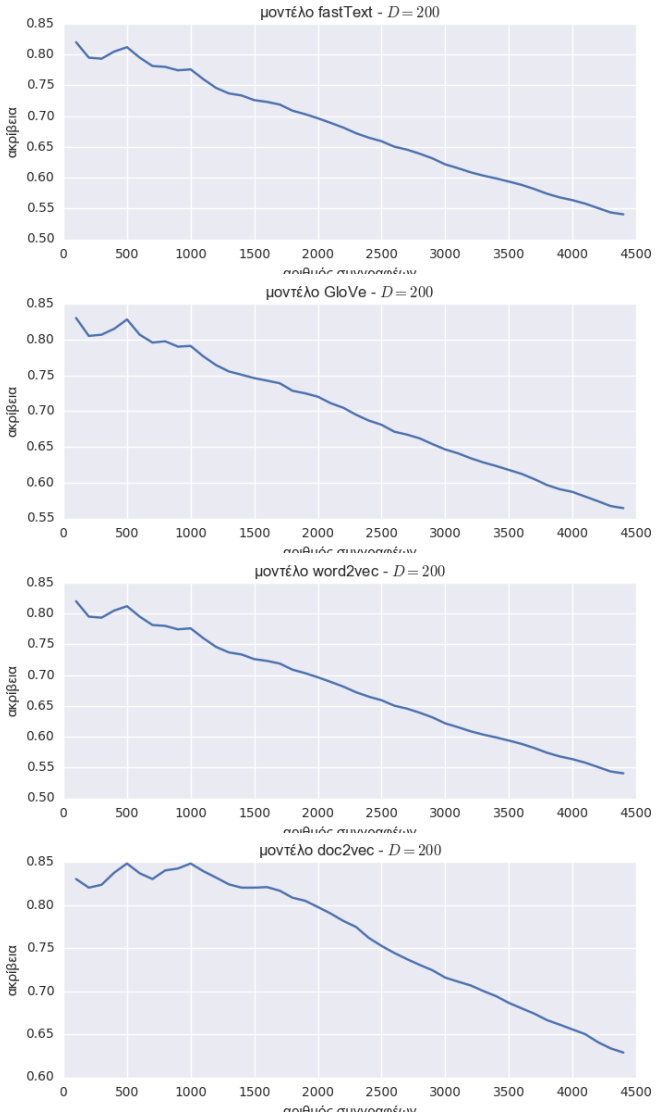
Διάγραμμα Ε.31: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=100



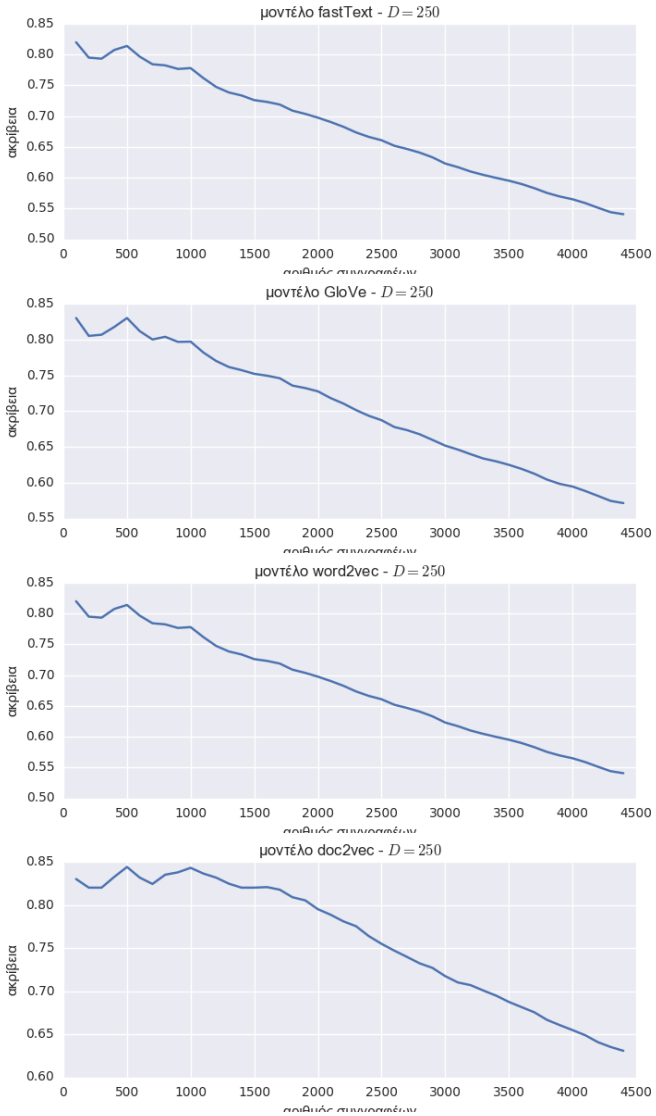
Διάγραμμα Ε.32: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=150



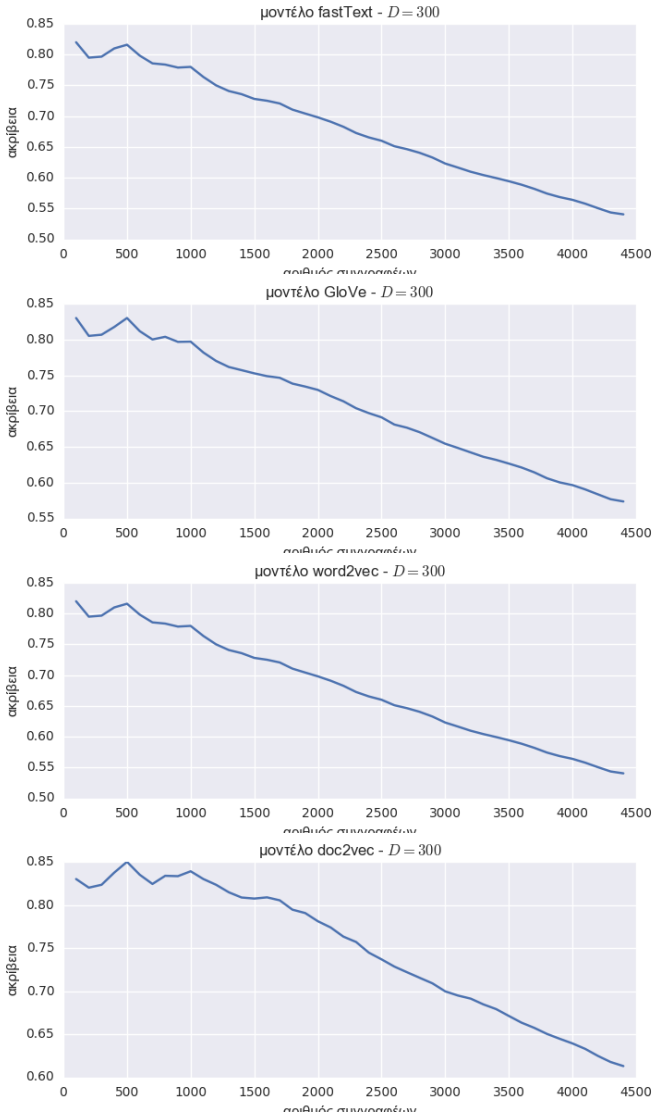
Διάγραμμα Ε.33: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=200



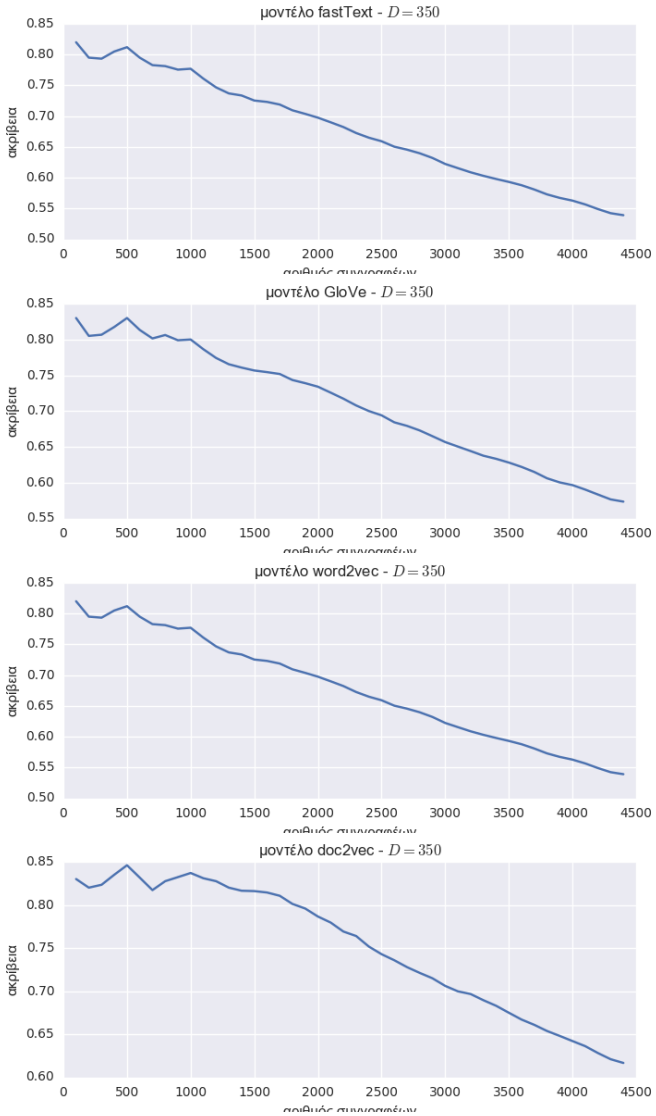
Διάγραμμα Ε.34: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=250



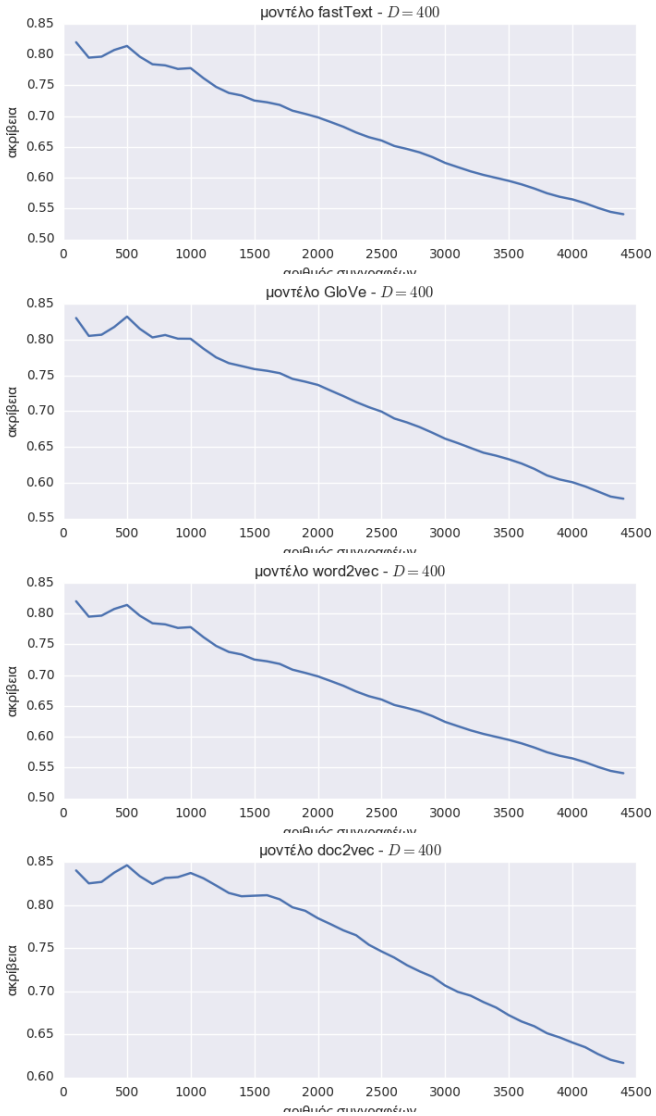
Διάγραμμα Ε.35: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=300



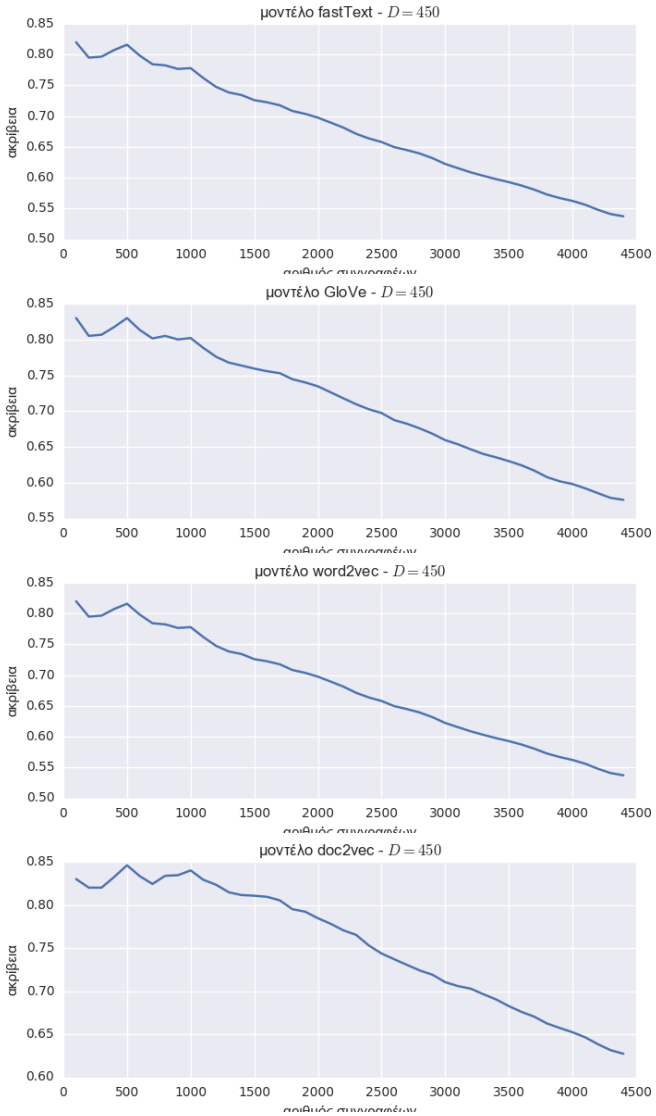
Διάγραμμα Ε.36: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=350



Διάγραμμα Ε.37: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=400

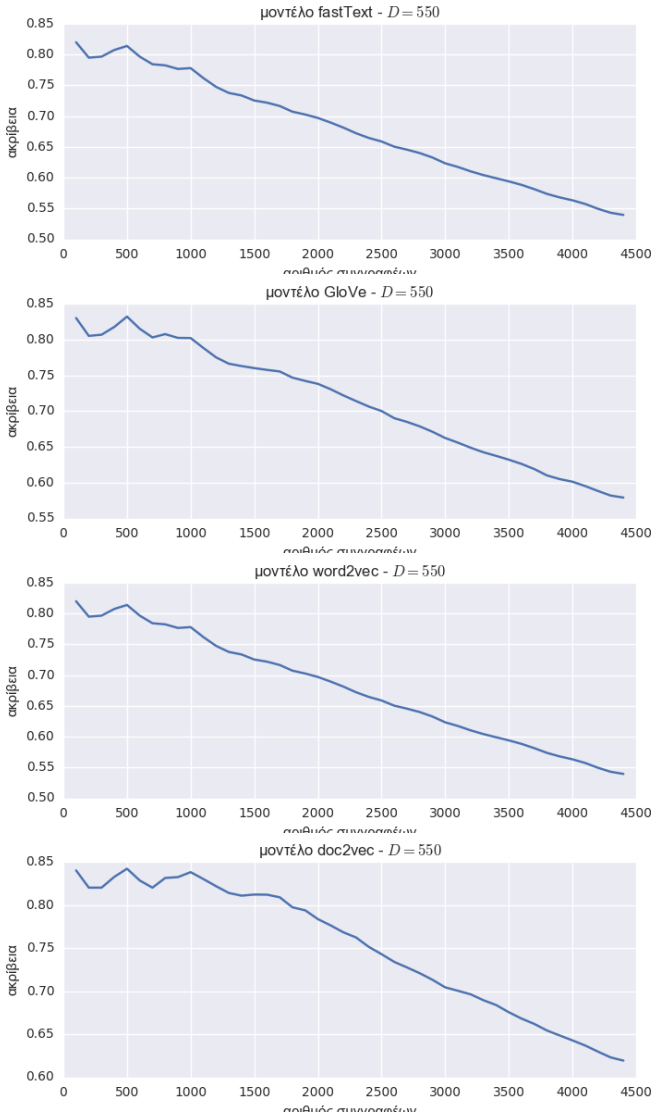


Διάγραμμα Ε.38: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=450

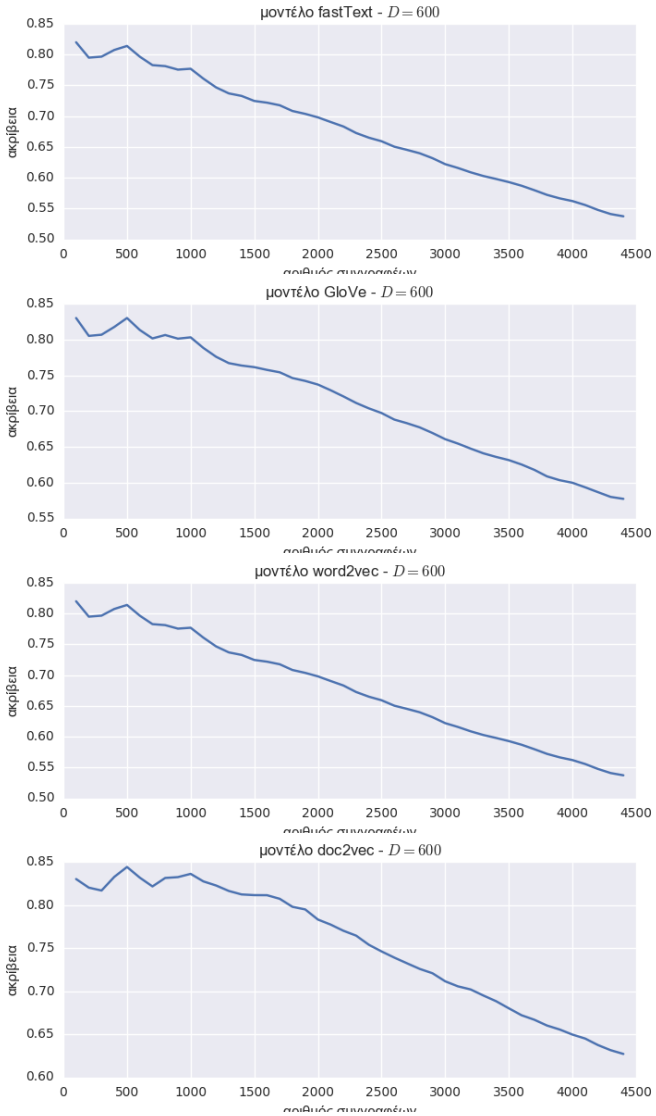


Διάγραμμα Ε.39: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=500

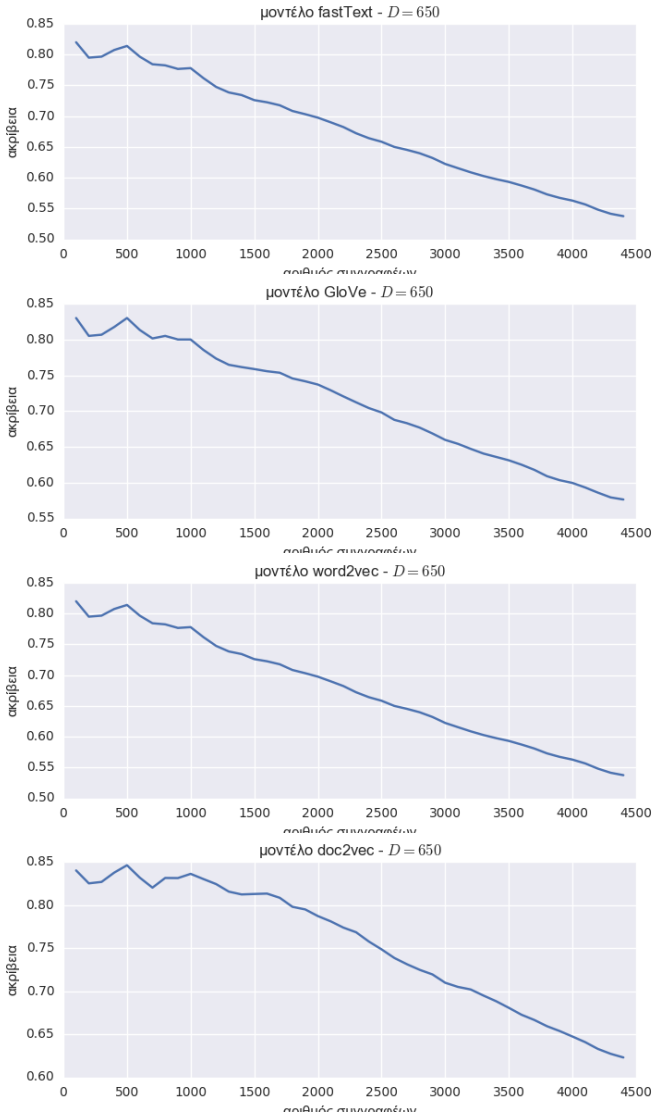
Διάγραμμα Ε.40: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=550



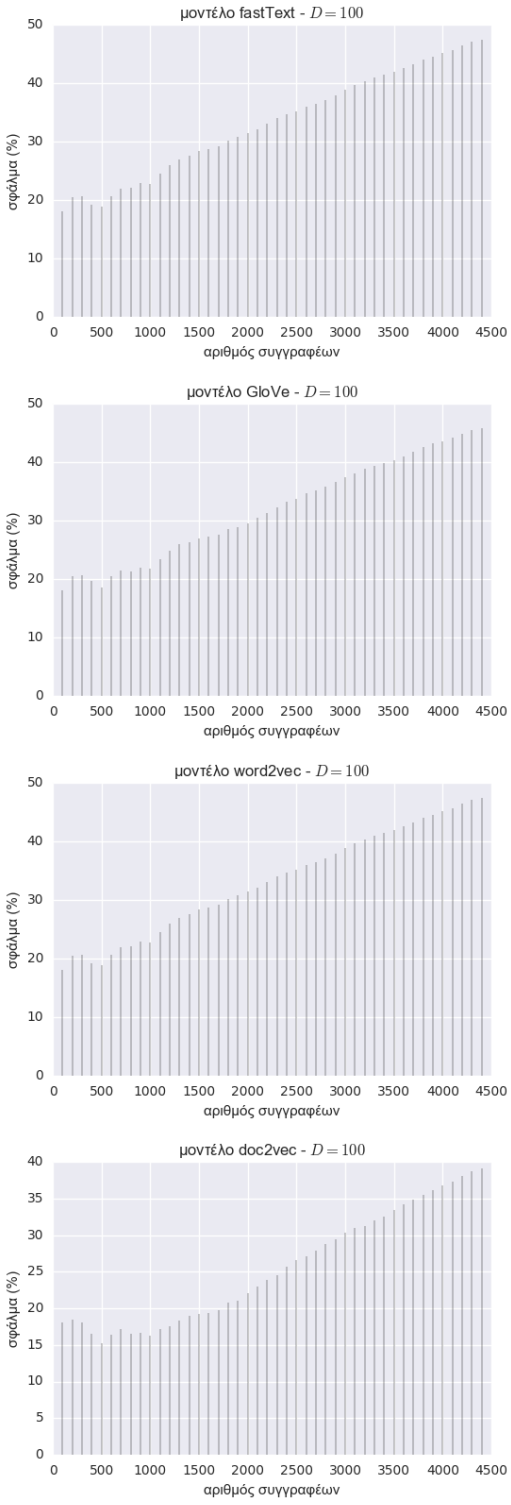
Διάγραμμα Ε.41: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=600



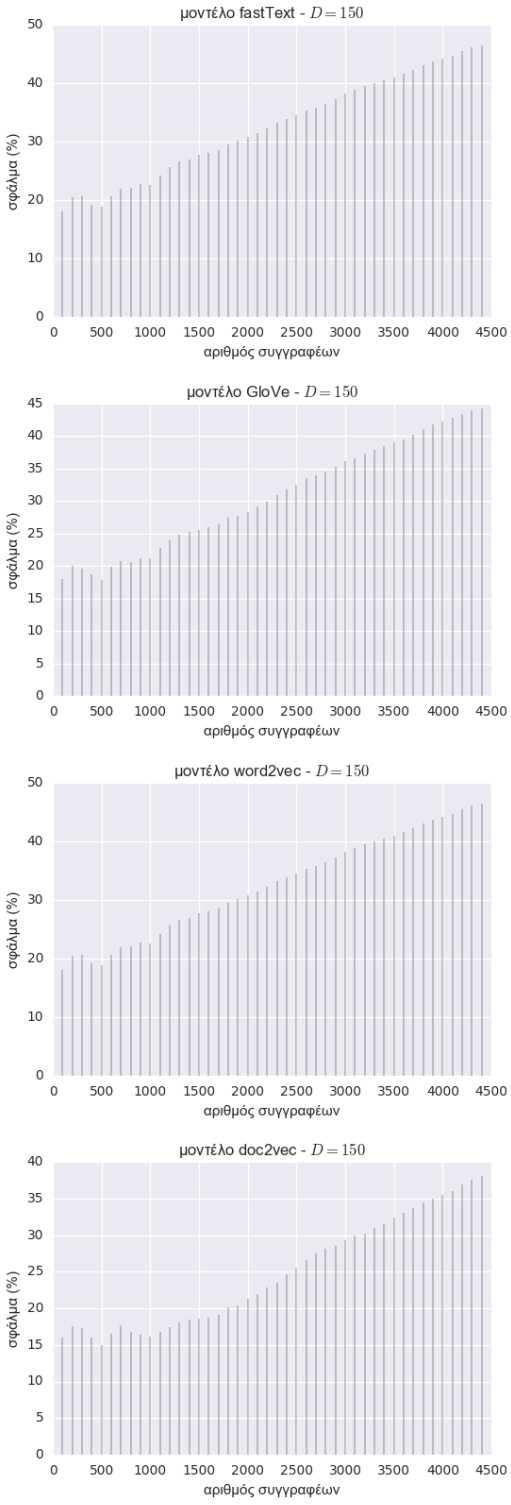
Διάγραμμα Ε.42: Ακρίβεια μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=650



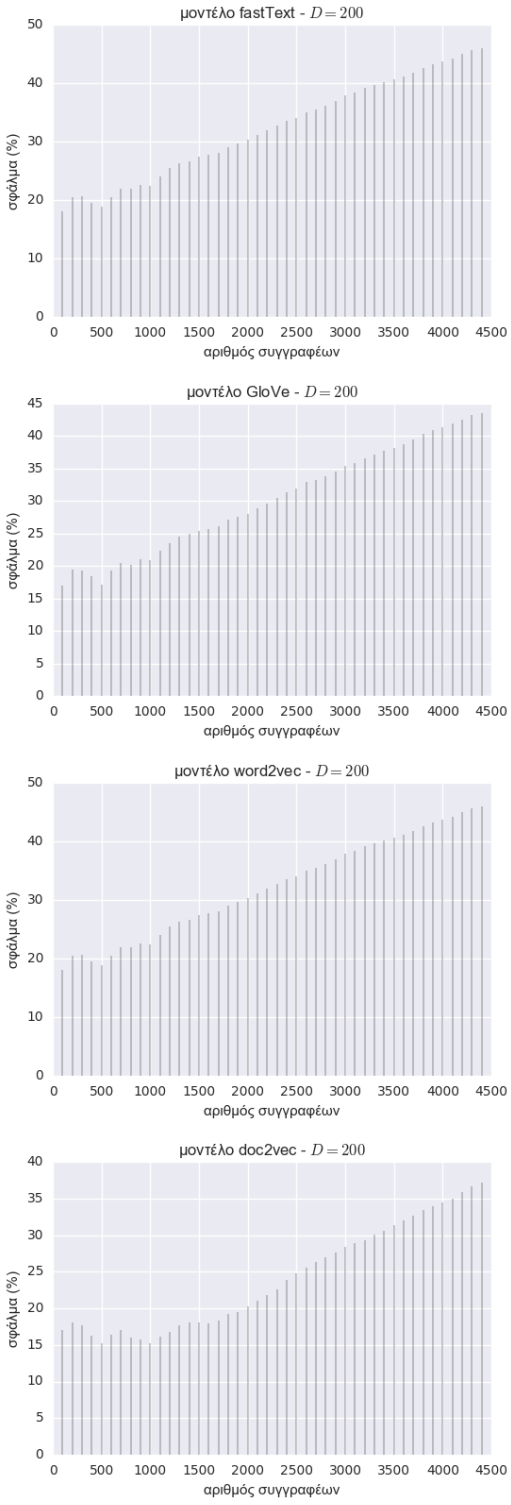
Διάγραμμα Ε.43: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=100



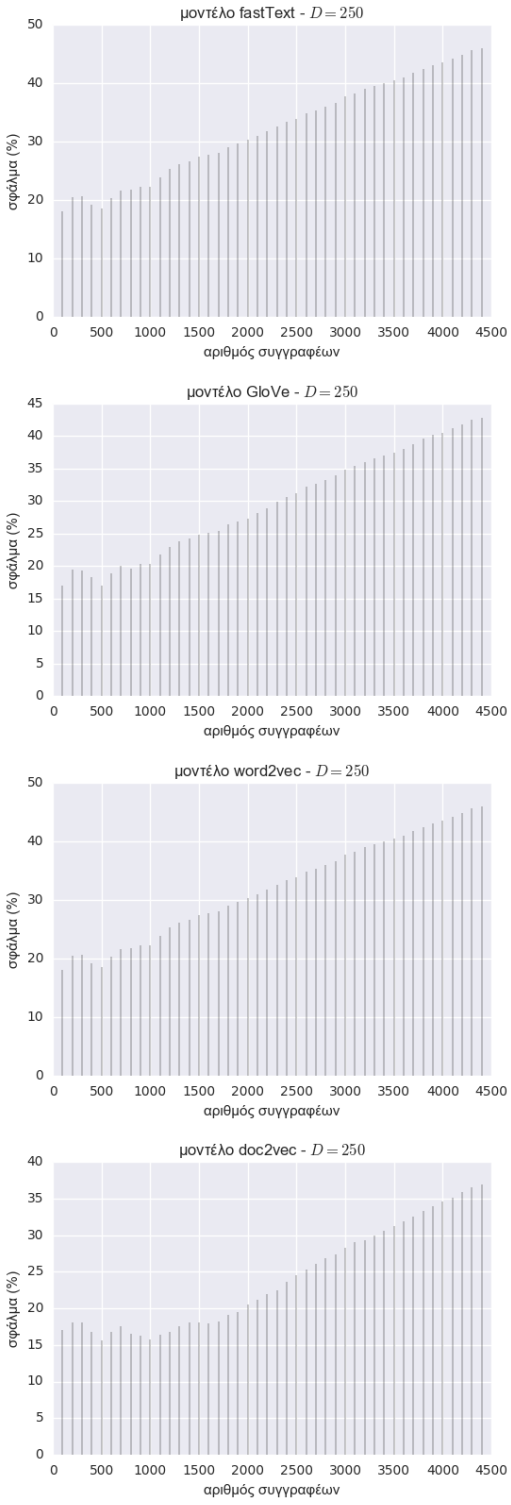
Διάγραμμα Ε.44: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=150



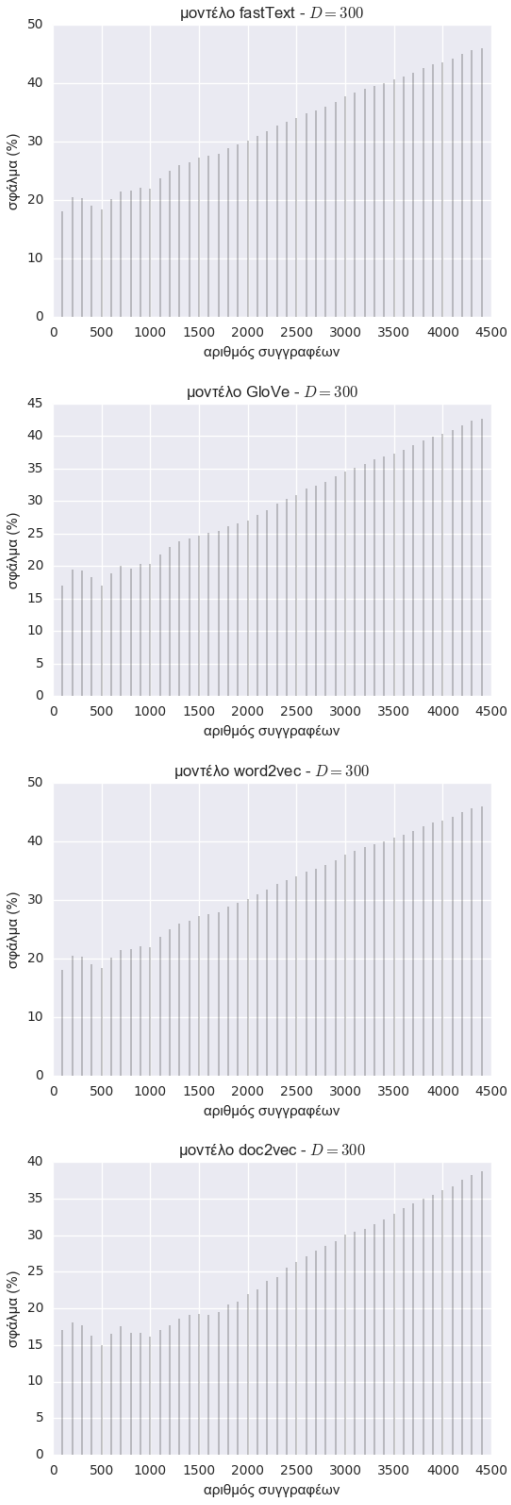
Διάγραμμα Ε.45: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=200



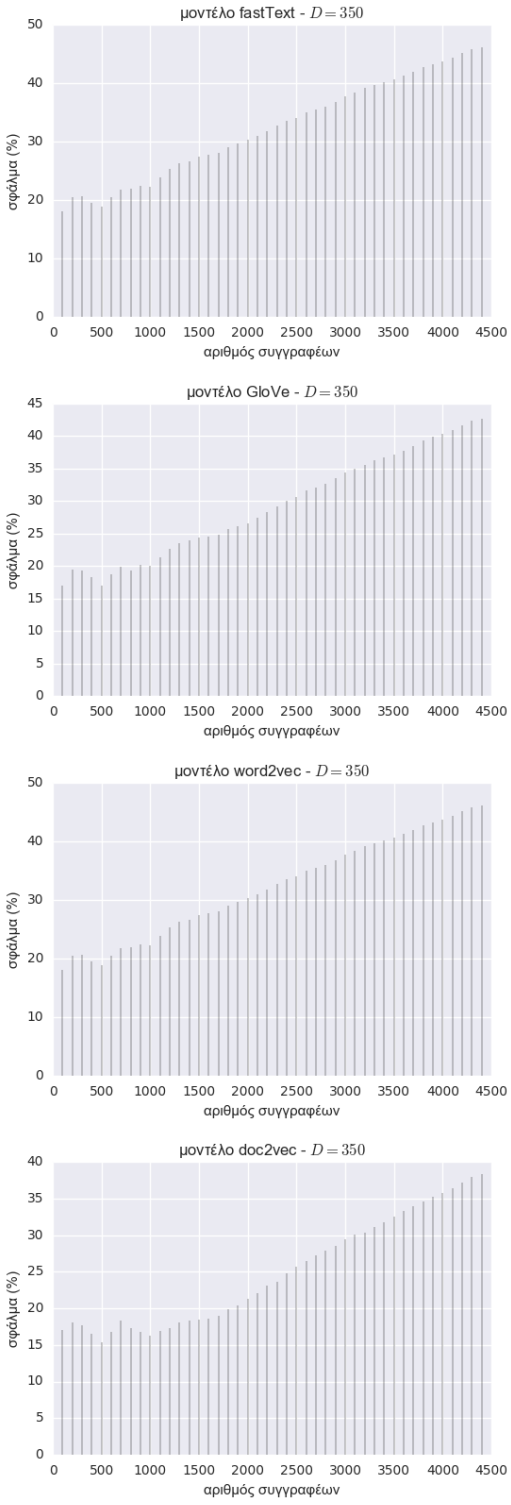
Διάγραμμα Ε.46: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=250



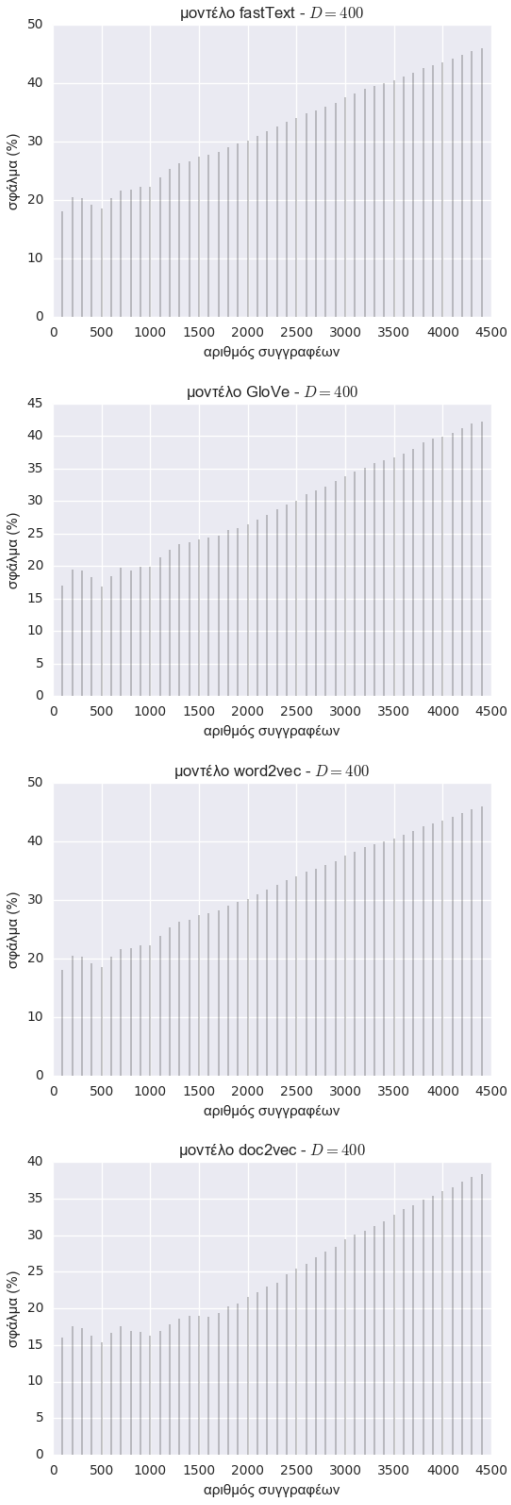
Διάγραμμα Ε.47: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=300



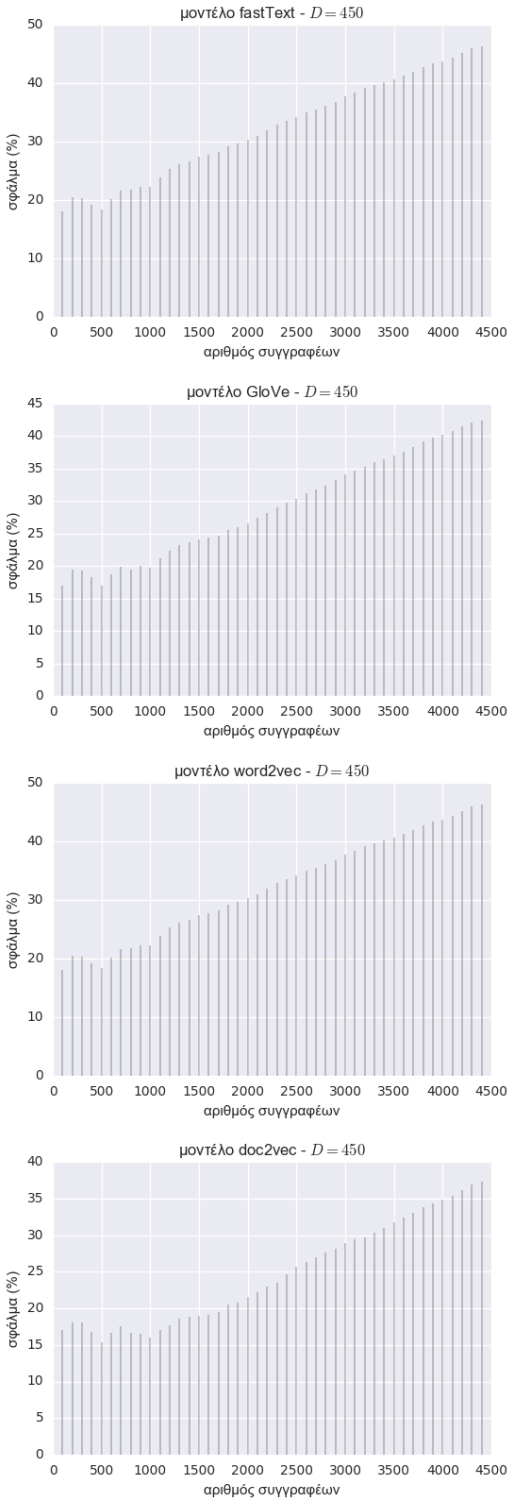
Διάγραμμα Ε.48: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=350



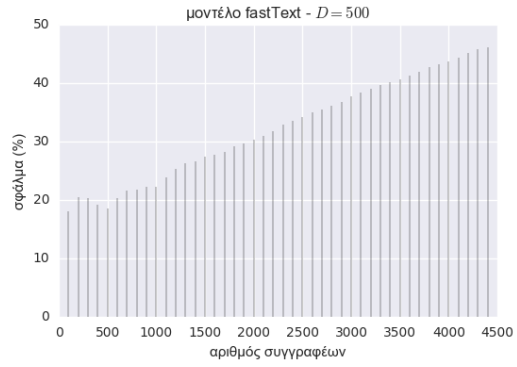
Διάγραμμα Ε.49: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=400

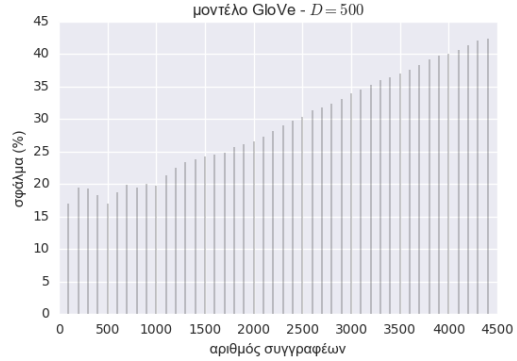


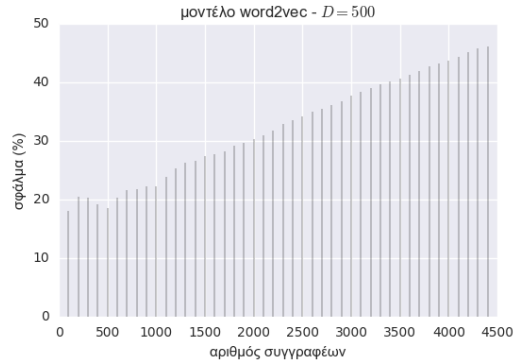
Διάγραμμα Ε.50: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=450

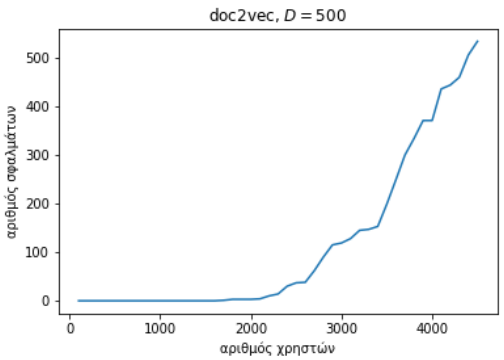


Διάγραμμα Ε.51: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=500

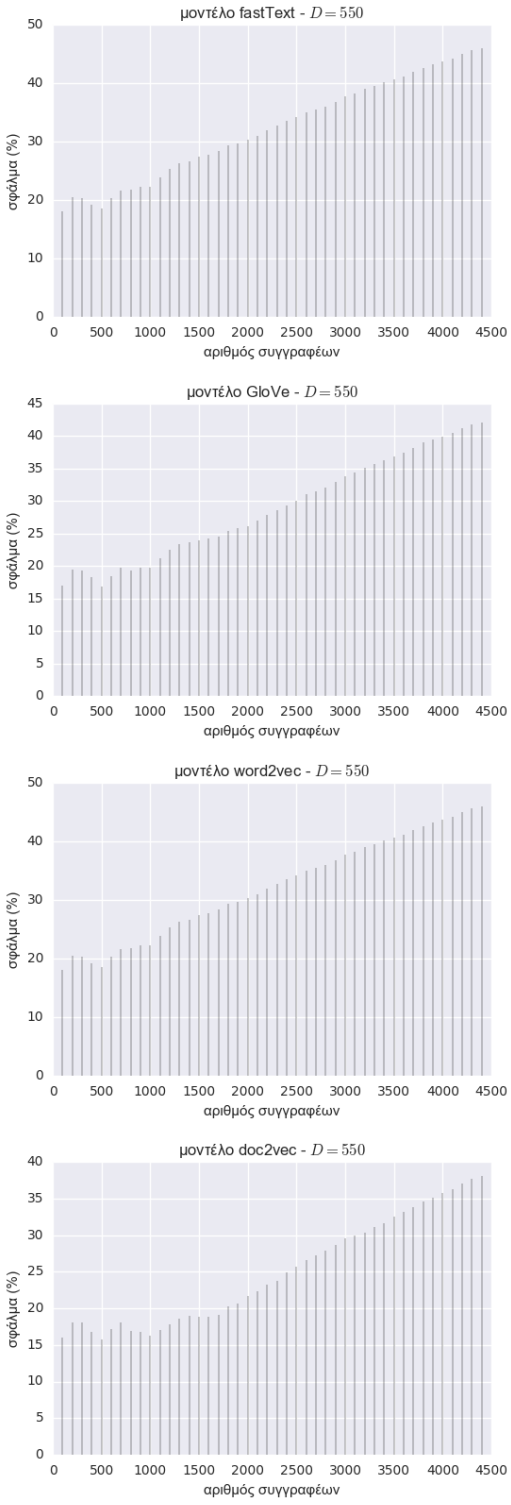




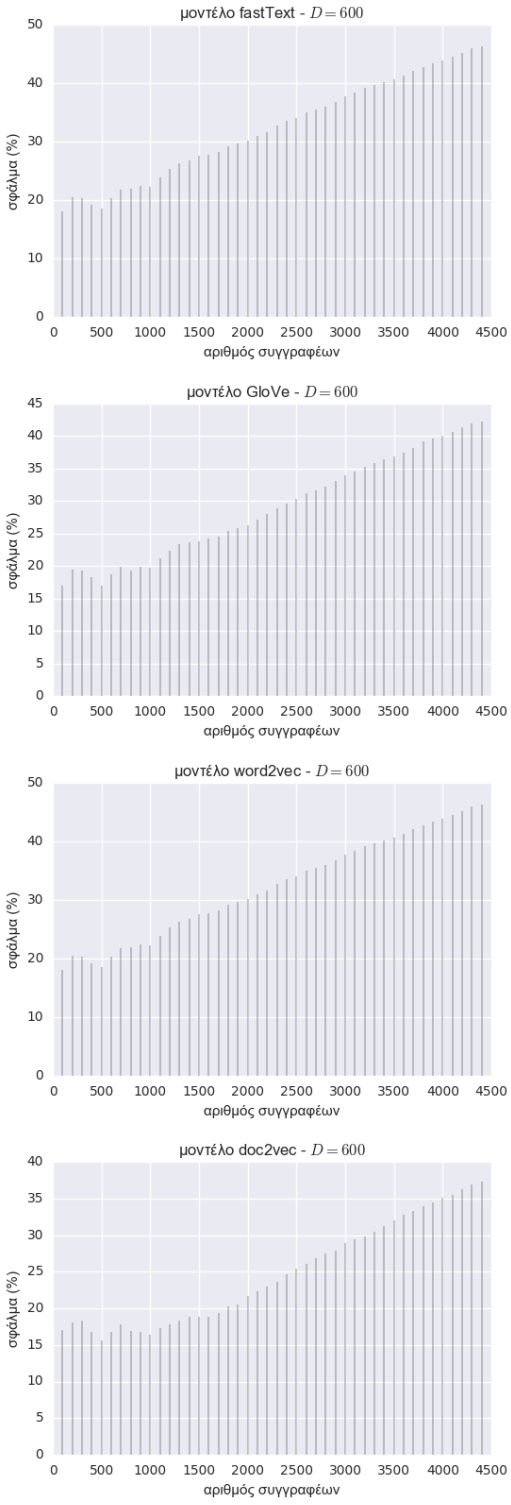




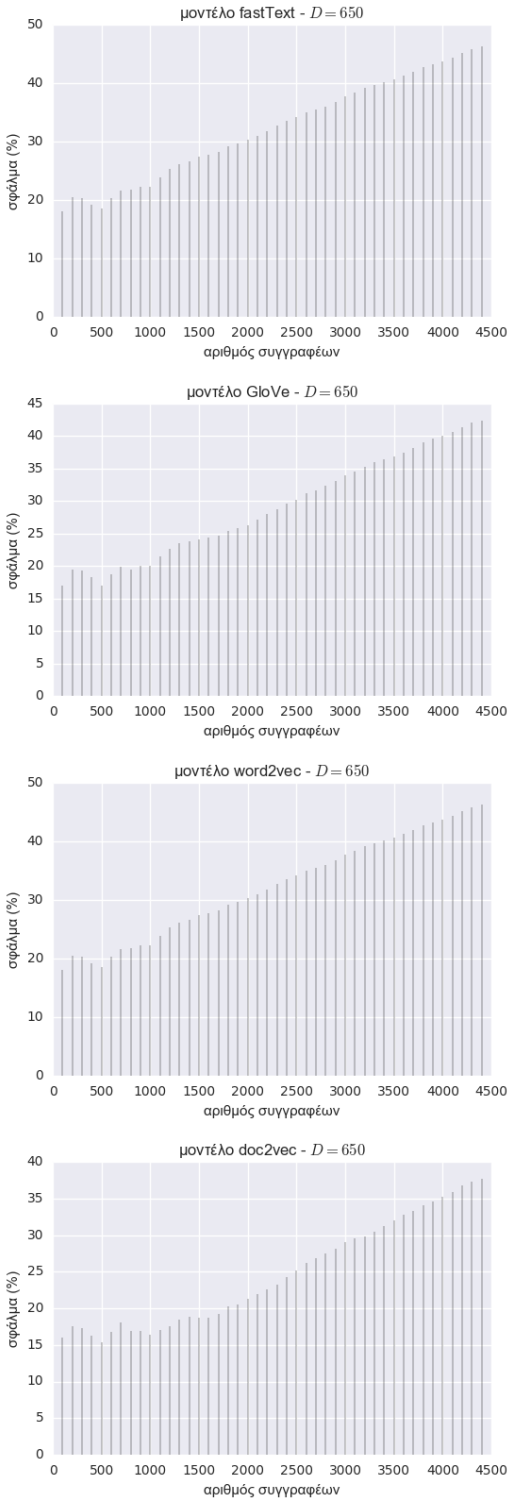
Διάγραμμα Ε.52: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=550



Διάγραμμα Ε.53: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=600



Διάγραμμα Ε.54: Σφάλμα μοντέλων στο Σώμα κειμένων ιστολογίων BAC D=650



Πίνακας Ε.1: Συστάδα ”Έλληνες Πολιτικοί”

ChristosSenekis
egrammes
geoterzis
FofiGennimata
patrianakou
DKoureas
NikosDendias
kmitsotakis
Skylakakis
loverdos
nkaklamanis
GKoumoutsakos
mihaliskatrinis

Ε.3 Αποτελέσματα συσταδοποίησης

Πίνακας Ε.2: Συστάδα ”Politics - in English”

katmouts	loukia_g
paulmasonnews	JunckerEU
VassiliouEU	pdragoumis
damomac	MarevaGrabowski
SpiegelPeter	iGlinavos
alexapostolides	TheEconomist
MariosZachariad	daphnenews
Greek_QB	tsipras_eu
helena_chari	argyris
Lagarde	IrateGreek
GuyVerhofstadt	Phil_Kimby
KatKanelidou	Heleppolis_68
Prodactiv	GreekAnalyst
BcnFox	VDombrovskis
NickMalkoutzis	WSJ
MatinaStevis	cocodinos
apas	MargSchinas
BlueKyp	KaterinaSokou
MartinSchulz	Hugodixon
KazimirPeter	MartinSelmayr
teacherdude	Vlattas_C
YanniKouts	business
eucopresident	Simon_Nixon
CNN	J_Panaretos
panostsapanidis	EP_President

Παράρτημα ΣΤ

Παραδείγματα ενθέσεων από το Σώμα Ελληνικών Κειμένων

Στους παρακάτω πίνακες παρουσιάζονται παραδείγματα ενθέσεων στο Σώμα Ελληνικών Κειμένων. Πιο συγκεκριμένα, παρουσιάζονται παραδείγματα από 3 μοντέλα που εκπαιδεύτηκαν στον ΣΕΚ, GloVe, fastText και word2vec και τα τρία με διάσταση $D = 100$. Τα μοντέλα word2vec και fastText έχουν εκπαιδευτεί με την αρχιτεκτονική SkipGram. Για την εκπαίδευση χρησιμοποιούνται λέξεις με αφαίρεση τόνων και σε πεζούς χαρακτήρες, με συχνότητα εμφάνισης μεγαλύτερη από 10 λέξεις, το οποίο μεταφράζεται σε 95489 τύπους.

ΣΤ.1 Ενθέσεις GloVe, D=100

Πίνακας ΣΤ.1: GloVe, ερώτημα ***ελλαδα***

Λέξη	**Απόσταση**
ευρωπη	0.873917
χωρα	0.867070
χωρες	0.734388
κυπρο	0.716526
ελλαδα	0.715083
ευρωπαικη	0.713202
αμερικη	0.712597
τουρκια	0.709791
ιταλια	0.706087
ελληνικη	0.704340

Πίνακας ΣΤ.2: GloVe, ερώτημα: ***υπουργος***

Λέξη	**Απόσταση**
υφυπουργος	0.810412
πρωθυπουργος	0.728271
υπουργειο	0.726952
υπουργειου	0.709713
υπουργο	0.702854
λαλιωτης	0.690091
ανεφερε	0.688136
δηλωσε	0.685690
παπαντωνιου	0.673536

Πίνακας ΣΤ.3: Glove, ερώτημα: ***τεχνη***

Λέξη	**Απόσταση**
συγχρονη	0.689257
επιστημη	0.653254
γλωσσα	0.652297
ιστορια	0.650613
ποιηση	0.650097
θρησκεια	0.649671
λογοτεχνια	0.646323
φιλοσοφια	0.641130
ζωγραφικη	0.640943
τεχνης	0.636624

Πίνακας ΣΤ.4: Glove, ερώτημα: ***θεωρητικα***

Λέξη	**Απόσταση**
λογικα	0.6483901739120483
διαφορετικα	0.6463874578475952
μοντελα	0.6046797633171082
επιστημονικα	0.5630309581756592
πολιτικα	0.5552231073379517
τυπικα	0.5520616769790649
ορισμενα	0.551082968711853
γνωστικα	0.5450225472450256
διαφορα	0.5369558334350586
ελαχιστα	0.532689094543457

Πίνακας ΣΤ.5: Glove, ερώτημα: ***προβλημα***

Λέξη	**Απόσταση**
σοβαρο	0.8023325204849243
θεμα	0.7905410528182983
ζητημα	0.7856243252754211
προβληματα	0.7409477233886719
αυτο	0.7097760438919067
τετοιο	0.7086629867553711
υπαρχει	0.7075026631355286
αντιμετωπιζει	0.6899952292442322
κανενα	0.6892690658569336
γεγονος	0.6872876882553101

Πίνακας ΣΤ.6: Glove, ερώτημα: ***γνωριζω***

Λέξη	**Απόσταση**
ξερω	0.7782631516456604
ξερετε	0.71067214012146
εγω	0.6975244283676147
νομιζω	0.6837608814239502
λεω	0.6799349784851074
ηξερα	0.6718450784683228
πιστευω	0.6716136932373047
καταλαβαινω	0.6666320562362671
πω	0.6645398139953613
ειπα	0.6579594612121582

ΣΤ.2 Ενθέσεις word2vec, skipgram, D=100

Πίνακας ΣΤ.7: word2vec, ερώτημα: ***ελλαδα***

Λέξη	**Απόσταση**
χώρα	0.648565
κύπρο	0.637494
ιταλία	0.636556
αμερική	0.578446
αλβανία	0.556006
αγγλία	0.547622
τουρκία	0.546276
ισπανία	0.534198
κύπρος	0.528702
αθήνα	0.509910

Πίνακας ΣΤ.8: word2vec, ερώτημα: ***υπουργος***

Λέξη	**Απόσταση**
υφυπουργος	0.809430
υπουργειο	0.740858
υπουργο	0.694222
γπουργος	0.680804
υπουργου	0.659633
υπουργειου	0.632333
υπουργοι	0.605352
υπουργους	0.602762
υφυπουργου	0.588899
γφυπουργος	0.579382
υπερυπουργος	0.570509
υφυπουργο	0.559465

ΣΤ.3 Ενθέσεις fastText, skipgram, D=100

Πίνακας ΣΤ.9: word2vec, ερώτημα: ***τεχνη***

Λέξη	**Απόσταση**
ζωγραφικη	0.649726
γλυπτικη	0.621072
λογοτεχνια	0.600939
ποιηση	0.593968
τεχνης	0.586396
καλλιτεχνης	0.586310
θρησκευτικοτητα	0.557746
rodin	0.553346
στιχουργικη	0.527143
επιστημη	0.527022
θυμοσοφια	0.523850
καλλιτεχνικη	0.522464

Πίνακας ΣΤ.10: fastText, ερώτημα: ***ελλαδα***

Λέξη	**Απόσταση**
ευρωπη	0.802091
χωρα	0.780383
ελλαδα!	0.741905
κυπρο	0.731024
αλβανια	0.716028
αμερικη	0.713282
ιταλια	0.703496
αυστραλια	0.695904
τουρκια	0.695199
βουλγαρια	0.694449

Πίνακας ΣΤ.11: fastTet, ερώτημα: ***υπουργος***

Λέξη	**Απόσταση**
υφυπουργος	0.895928
υπερυπουργος	0.881617
yφυπουργος	0.878041
<υφυπουργος	0.85251
πρωθυπουργος	0.766725
yπουργος	0.762733
υπουργειο	0.759061
υπερυπουργειο	0.755404
eξωτερικων	0.748338
εξωτερικων	0.745396

Πίνακας ΣΤ.12: fastText, ερώτημα: ***τεχνη***

Λέξη	**Απόσταση**
τεχνης	0.789282
ζωγραφικη	0.767000
λογοτεχνια	0.753931
καλλιτεχνικη	0.753822
καλλιτεχνις	0.730354
καλλιτεχνη	0.711031
ανθρωπολογια	0.710516
μυθοπλαστικη	0.709454
ζωγραφια	0.708254
λογοτεχνικη	0.706272

Παράρτημα Z

Παραδείγματα ομοιότητας ιδιολεκτικού ύφους

Στους παρακάτω πίνακες εμφανίζονται αποτελέσματα ιδιολεκτικής ομοιότητας με βάση ενθέσεις ύφους από το μοντέλο doc2vec, με $D = 500$

Πίνακας Z.1: ***atsipras***

Λέξη	**Απόσταση**
PrimeministerGR	0.852194
YDragasakis	0.822092
syriza_gr	0.819151
NikosDendias	0.808965
PESXanthi	0.794611
Dora_Bakoyannis	0.793161
FofiGennimata	0.791793
olgakef	0.789790
loukakatseli	0.785013
tzitzikostas	0.783398

Πίνακας Ζ.2: ***MVarvitsiotis***

Λέξη	**Απόσταση**
NikosDendias	0.902862
olgakef	0.898828
tzitzikostas	0.892651
loukakatseli	0.890041
Dora_Bakoyannis	0.883282
Vkikilias	0.879885
e_stylianidis	0.876080
a_loverdos	0.875115
nmitarakis	0.872683
androulakisnick	0.872498

Πίνακας Ζ.3: ***madtv***

Λέξη	**Απόσταση**
madVMA	0.813422
dailymotion_GR	0.807777
PanikRecords	0.791497
MADradio	0.780598
Onirama_band	0.758772
VEGAS_group	0.751666
RecBand	0.746090
ninofanss	0.740550
HelenaPanayi	0.740456
MelinaMakris	0.740406

Πίνακας Ζ.4: ***ANT1TV***

Λέξη	**Απόσταση**
ToPrwino	0.872924
ProinoSouKou	0.845405
thevoice_gr	0.827676
MegaTvOfficial	0.780637
OlaKalaMak	0.771973
Klelia_pantazi	0.768865
mikevasileiadis	0.764359
themismallis	0.760022
madVMA	0.759754
Martakis_Kostas	0.758839