Μαθηματικά μοντέλα στατιστικής ανάλυσης, κατηγοριοποίησης και γραφικής απεικόνισης συναισθήματος στα Μέσα Κοινωνικής Δικτύωσης

Ιερεμίας Ιωαννίδης

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Τμήμα Μαθηματικών

Πρόγραμμα Μεταπτυχιακών Σπουδών

Ειδίκευση στα Εφαρμοσμένα Μαθηματικά

Αθήνα, Ιούλιος 2021

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Επιβλέποντες

Μπάρδης Νικόλαος, Καθηγητής (επιβλέπων)

Κότα-Αθανασιάδου Ευαγγελία, Επικ. Καθηγήτρια

Δρακόπουλος Μιχαήλ, Επικ. Καθηγητής

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κο Νικόλαο Μπάρδη για την σημαντική βοήθεια του σε αυτήν την δύσκολη χρονιά, για τις συμβουλές του και τις γνώσεις που μου μετέδωσε. Επίσης θα ήθελα να ευχαριστήσω όλους τους καθηγητές που είχα ανά τα χρόνια. Τέλος χρωστάω ένα μεγάλο ευχαριστώ στην οικογένεια μου που με στήριξε και ήταν πάντα δίπλα μου.

Στον παππού μου

Μαθηματικά μοντέλα στατιστικής ανάλυσης, κατηγοριοποίησης και γραφικής απεικόνισης συναισθήματος στα Μέσα Κοινωνικής Δικτύωσης

Περίληψη

Ο όγκος των δεδομένων που υπάρχουν διαθέσιμα για συλλογή στο διαδίκτυο τα τελευταία χρόνια έχει αυξηθεί ραγδαία. Ταυτόχρονα η επεξεργασία, ταξινόμηση και παρουσίαση αυτών γίνεται με πιο αποδοτικό και αναλυτικό τρόπο. Επίσης οι μέθοδοι για την ανάλυση και την απεικόνιση των δεδομένων έχουν ισχυρό μαθηματικό υπόβαθρο. Τα Μέσα Κοινωνικής Δικτύωσης αποτελούν μια πηγή διαθέσιμων δεδομένων προς επεξεργασία και στατιστική ανάλυση τα οποία βρίσκονται σε μορφή κειμένων.

Η παρούσα εργασία στοχεύει στην ανάλυση της φυσικής γλώσσας των Tweets του Twitter για την εξαγωγή συμπερασμάτων, σχετικά με το συναίσθημα που πηγάζει από αυτά. Το χρονικό πλαίσιο στο οποίο θα πραγματοποιηθεί η ανάλυση είναι το πρώτο μισό του 2021. Παράλληλα θα μελετηθούν οι διάφοροι αλγόριθμοι ταιριάσματος νημάτων που χρησιμοποιήθηκαν καθώς και το μαθηματικό υπόβαθρό τους.

Ακόμα θα αναφερθούμε στην ασφαλή επικοινωνία μεταξύ της πλατφόρμας του Twitter και του χρήστη, αλλά και στην ασφαλή εξαγωγή των Tweets χάρη στην ταυτοποίηση OAuth 2.0.

Τέλος, θα εξαχθούν και θα παρουσιαστούν τα αποτελέσματα της ανάλυσης, με χρήση διαφόρων γραφημάτων, η οποία έγινε με την βοήθεια της γλώσσας προγραμματισμού Python.

Mathematical models of statistical analysis, categorization and graphic representation of sentiment on Social Media

Abstract

The volume of data available for online collection in recent years has increased rapidly. At the same time processing, classification and presentation is done in a more efficient manner. Social Media are a source of available data for statistical analysis that are in text form. Also, the methods for analyzing and displaying the data have a powerful mathematical background. For this project we will use the Twitter network.

This project aims at analyzing the natural language of Tweets, written in Greek, to draw conclusions on the sentiment that stems from them, within the timeframe of the first half of 2021. At the same time the various string-matching algorithms used as well as their mathematical background will be studied.

We will also refer to the secure communication between the Twitter page and the user, as well as the secure export of Tweets thanks to OAuth 2.0 identification.

Finally, the results of the analysis will be extracted and presented, using various chats. To help us achieve our purpose we will need the help of programming language Python.

Περιεχόμενα

Κατάλογος Εικόνων:

Κατάλογος Διαγραμμάτων:

Εισαγωγή

Η σύγχρονη κοινωνία επικοινωνεί και εκφράζεται μέσω των κοινωνικών δικτύων. Αναγκαία λοιπόν κρίνεται η ανάλυση των δεδομένων που πηγάζουν μέσα από τις online επικοινωνίες . Παράλληλα τα δεδομένα αποτελούν μια μεγάλη πηγή ανταγωνισμού. Οι εταιρίες χρησιμοποιούν τα κοινωνικά δίκτυα με σκοπό να διαφημίσουν και να προωθήσουν τα προϊόντα τους καθώς οι επαγγελματίες στοχεύουν στην δικτύωση με περισσότερους πελάτες.

Η παρούσα εργασία επικεντρώνεται στην ανάλυση των κειμένων του Twitter, μια εκ των κορυφαίων υπηρεσιών κοινωνικής δικτύωσης. Ιδιαίτερος σκοπός της είναι η εύρεση της συχνότητας εμφάνισης των λέξεων και η ερμηνεία των αποτελεσμάτων στα πλαίσια της ελληνικής κοινωνίας και της καθημερινότητας που ζήσαμε το πρώτο μισό του 2021.

Αναλυτικότερα τα βήματα που θα ακολουθήσουμε είναι:

1. Σύνδεση με την πλατφόρμα του Twitter
2. Εξόρυξη και αποθήκευση των δεδομένων
3. Επεξεργασία
4. Οπτικοποίηση των αποτελεσμάτων
5. Ερμηνεία

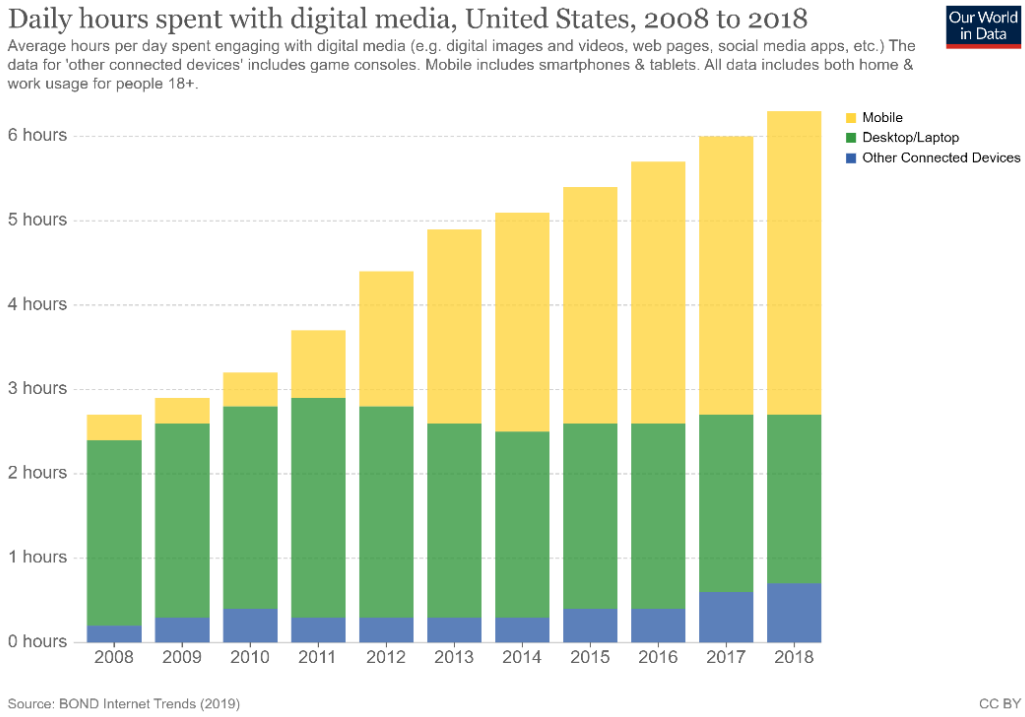
Γενικότερα στο πρώτο κεφάλαιο θα αναφερθούμε στο Twitter, την επικοινωνία με την πλατφόρμα προγραμματιστή και την ασφαλή χρήση της. Στο δεύτερο κεφάλαιο, θα μιλήσουμε για την ανάλυση δεδομένων, τις υποκατηγορίες της και συγκεκριμένα την Επεξεργασία φυσικής γλώσσας. Στο τρίτο κεφάλαιο θα αναλύσουμε διεξωδικά τα βήματα του αλγορίθμου καθώς θα μιλήσουμε και για την γλώσσα προγραμματισμού που θα χρειαστούμε για την ανάλυση μας. Τέλος στο τέταρτο και τελευταίο κεφάλαιο προσπαθήσουμε να οπτικοποιήσουμε και να ερμηνεύσουμε τα αποτελέσματα μας.

Κεφάλαιο 1

1.1 Κοινωνικά Δίκτυα, Twitter

Η ανάπτυξη του ίντερνετ σε συνδυασμό με την τεχνολογική πρόοδο έχουν οδηγήσει σε τεράστια αύξηση την επικοινωνία μέσω του διαδικτύου. Συγκεκριμένα η είσοδος των έξυπνων τηλεφώνων (smartphones) – τα οποία έχουν άμεση πρόσβαση στο διαδίκτυο – στην αγορά, καθιστά πολύ εύκολη την πρόσβαση στις online εφαρμογές. Παράλληλα, ο χρόνος χρήσης αυτών των εφαρμογών έχει διπλασιαστεί σε σχέση με μια δεκαετία πριν.

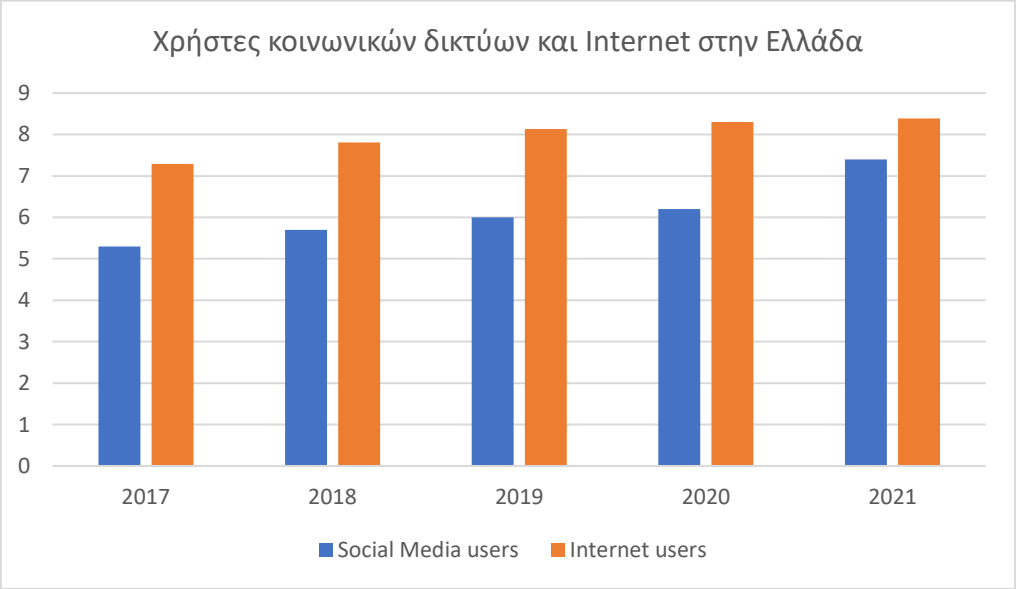
Διάγραμμα 1.1: Ώρες Χρήσης Διαδικτύου



Πηγή: https://ourworldindata.org/grapher/daily-hours-spent-with-digital-media-per-adult-user?country=~USA

Συγκεκριμένα στην Ελλάδα τον Ιανουάριο του 2021 έχουμε 8,39 εκατομμύρια χρήστες του διαδικτύου, κάτι το οποίο αποτελεί το 80,7% του συνολικού πληθυσμού της χώρας μας. Από αυτούς τα 7,4 εκατομμύρια περίπου είναι και χρήστες κοινωνικών δικτύων. Ο αριθμός αυτός έχει αυξηθεί σε σχέση με το 2020 κατά 650 χιλιάδες χρήστες, ένα ποσοστό κοντά στο 10%. Παρακάτω βλέπουμε ένα γράφημα με την εξέλιξη της χρήσης των κοινωνικών δικτύων και του ίντερνετ γενικότερα στη χώρα μας.

Διάγραμμα 1.2: Ελλάδα 2017-2021



Πηγή: https://datareportal.com/reports/digital-2021-greece

Οι διάφορες εφαρμογές κοινωνικών δικτύων που δημιουργήθηκαν στα μέσα της δεκαετίας του 2000 έχουν πλέον γιγαντωθεί και έχουν εδραιωθεί στον χώρο της online επικοινωνίας. Τα μέσα αυτά χωρίζονται σε πολλές κατηγορίες και διαφοροποιούνται σε πολλούς τομείς. Βλέπουμε μερικές από τις βασικότερες κατηγορίες παρακάτω καθώς και μερικά βασικά παραδείγματα:

- Social Networks (Facebook, **Twitter,** LinkedIn)
- Messaging & Voice call Sites (Messenger, Viber)
- Online forums & Social Blogging (Reddit, Quora)

- Photo sharing platforms (Instagram, Snapchat, Pinterest)
- Video sharing platforms (YouTube, Vimeo)

Στην εργασία αυτή θα μιλήσουμε για ένα από τα λεγόμενα Social Networks. Η συγκεκριμένη ομάδα κοινωνικών δικτύων έχει το μεγαλύτερο εύρος επιλογών καθώς παρέχουν δυνατότητες και άλλων κατηγοριών όπως μηνύματα και μοίρασμα φωτογραφιών. Το κυριότερο όμως χαρακτηριστικό τους είναι η δυνατότητα που δίνουν στον χρήστη να εκφράσει δημόσια την άποψη του πάνω σε οποιοδήποτε θέμα. Αυτές τις ιδιότητες παρέχει στους χρήστες του και το Twitter που είναι η εφαρμογή στην οποία θα αναφερθούμε στην παρούσα εργασία.

Συγκεκριμένα λοιπόν το Twitter δίνει την δυνατότητα στους χρήστες του να συντάξουν ένα μικρό κείμενο 280 χαρακτήρων στο οποίο μπορούν να προσθέσουν κάποιον σύνδεσμο (link) ή μια φωτογραφία. Η ανάλυση μας, σχετίζεται με το περιεχόμενο του κειμένου μου συντάσσουν οι χρήστες.

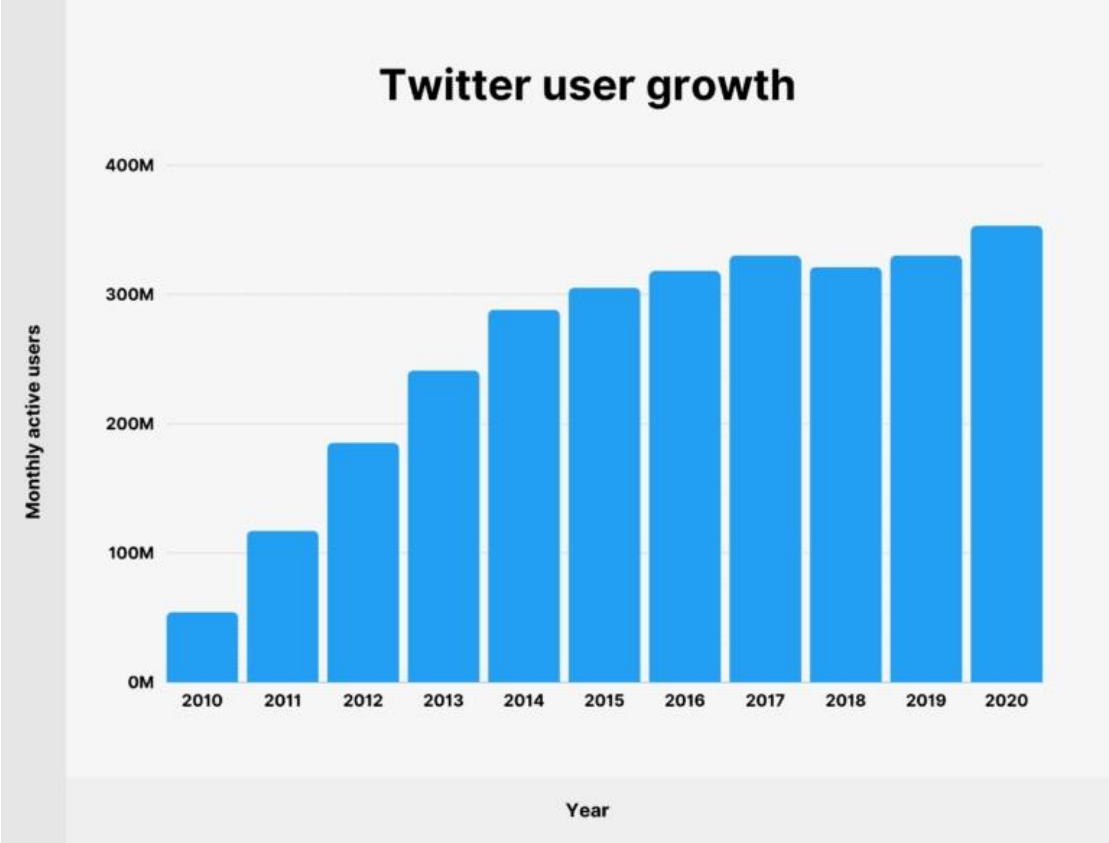
1.2 Το Twitter

Το Twitter δημιουργήθηκε από τους Jack Dorsey, ο οποίος παραμένει CEO της εταιρίας, Noah Glass, Biz Stone και Evan Williams, τον Ιούλιο του 2006. Βασικός σκοπός του ήταν η αποστολή και λήψη SMS μέσω του internet. Το 2013 απέκτησε περισσότερους από 100 εκατομμύρια χρήστες οι οποίοι ανέβαζαν περιεχόμενο 340 εκατομμύρια φορές την ημέρα. Μέσα στα επόμενα χρόνια ο αριθμός των χρηστών τριπλασιάστηκε και έτσι στις αρχές του 2019 το Twitter είναι 330 εκατομμύρια ενεργούν μηνιαίους χρήστες.

Εικόνα 1.1: Λογότυπο Twitter

Πηγή: https://en.wikipedia.org/wiki/Twitter#2007%E2%80%932010

Διάγραμμα 1.3: Πλήθος χρηστών 2010-2020



Πηγή: https://backlinko.com/twitter-users

Παράλληλα προστέθηκαν πολλές ακόμα λειτουργίες όπως η λειτουργία Follow , όπου πρακτικά μπορείς να παρακολουθείς το περιεχόμενο που ανεβάζει ο χρήστης

που επιλέγεις να ακολουθείς. Η λειτουργία αυτή παράλληλα δείχνει και την δημοτικότητα των χρηστών. Παρακάτω βλέπουμε μια λίστα με τους λογαριασμούς που ακολουθούνται από τους περισσότερους χρήστες.

Θέση	**Όνομα**	**Followers (σε εκατ.)**	**Δραστηριότητα**
1	Barack Obama	129	44ος Πρόεδρος ΗΠΑ
2	Justin Bieber	113	Μουσικός
3	Katy Perry	108	Μουσικός
4	Rihanna	102	Μουσικός
5	Cristiano Ronaldo	92	Ποδοσφαιριστής
6	Taylor Swift	88	Μουσικός
7	Lady Gaga	83	Μουσικός
8	Ariana Grande	83	Μουσικός
9	Ellen DeGeneres	78	Παρουσιάστρια
10	YouTube	72	Πλατφόρμα Online Video

Πηγή: https://en.wikipedia.org/wiki/Twitter#2007%E2%80%932010

1.3 Εξαγωγή δεδομένων και API

Για να εξάγουμε τα δεδομένα από το Twitter δεν αρκεί να έχουμε έναν απλό λογαριασμό χρήση. Χρειάζεται να κάνουμε μία αίτηση στο Twitter για να δημιουργήσουμε developer account (λογαριασμό προγραμματιστή). Κάνοντας την αίτηση αυτή αποκτούμε πρόσβαση στο Twitter API (Application Programming Interface). Το API πρακτικά είναι ένας μεσολαβητής ανάμεσα στην πλατφόρμα του Twitter και την εφαρμογή που εμείς θα επιλέξουμε να χρησιμοποιήσουμε για την συλλογή των Tweets. Εμείς μπορούμε να αποστείλουμε διάφορα αιτήματα μέσω

της δικής μας εφαρμογής τα οποία θα περάσουν μέσα από τον server του API και θα επιστέψουν τις κατάλληλες απαντήσεις κάνοντας την αντίθετη διαδρομή.

Το Twitter API έχει πολλές χρήσεις και μπορεί να βοηθήσεις όχι μόνο σε επίπεδο ακαδημαϊκό αλλά και σε τομείς των επιχειρήσεων. Κάποια παραδείγματα είναι τα εξής:

- Δημιουργία εφαρμογής που οι χρήστες μπορούν να έχουν πρόσβαση σε συγκεκριμένα Tweets, όπως Tweets που σχετίζονται μόνο με ένα γεγονός (πχ COVID-19, Euro 2020)
- Δημιουργία εφαρμογής που οι χρήστες μπορούν να έχουν πρόσβαση σε Tweets από συγκεκριμένη τοποθεσία.
- Δημιουργία προσωπικού project για ανάλυση των Tweets.

Το API επίσης παρέχει μια λίστα εντολών που μπορούν να εκτελεστούν για την επικοινωνία μας με την πλατφόρμα. Οι πιο βασικές είναι:

- Get (ανάκτηση δεδομένων)
- Post (δημιουργία δεδομένων)
- Put (ενημέρωση δεδομένων)
- Delete (αφαίρεση δεδομένων)

Σκοπός μας είναι η ανάκτηση δηλαδή την συλλογή δεδομένων και για τον λόγο αυτό η εντολή που θα χρειαστούμε είναι η GET. Συγκεκριμένα αυτή η εντολή μας επιστρέφει τα πιο πρόσφατα tweets που έγιναν σε ένα χρονικό πλαίσιο το οποίο εμείς ορίζουμε. Η εντολή αυτή έχει πολλές παραλλαγές και σε αυτήν που θα χρειαστούμε εμείς συλλέγουμε 18.000 Tweets μίας ημέρας. Η συλλογή ξεκινάει από το τελευταίο Tweet που έγινε την επιλεγμένη ημέρα και καταλήγει στο 1ο ή σταματάει όταν φτάσουμε σε πλήθος 18.000 Tweets.

Κατά την δημιουργία του developer λογαριασμού μας, για να συνδεθούμε με το API του Twitter θα μας δοθούν κάποιοι κωδικοί. Αυτοί είναι:

- API Key
- API Secret Key
- Access Token
- Access Token Secret Key

Αυτοί οι κωδικοί έχουν να κάνουν με την ασφαλή επικοινωνία μεταξύ της πλατφόρμας και του χρήστη. Θα αναφερθούμε στην ασφάλεια αναλυτικά στην παρακάτω ενότητα.

1.4 Ασφάλεια και OAuth2.0

Ασφάλεια χρήστη

Όπως είναι γνωστό η χρήση των κοινωνικών δικτύων επιφέρει κάποιους κινδύνους. Εμείς εδώ θα αναφερθούμε στους κινδύνους που σχετίζονται με το τεχνικό κομμάτι της χρήσης τους και όχι με την το κοινωνικό ή ψυχολογικό αντίκτυπο των social media.

Οι χρήστες έρχονται συχνά αντιμέτωποι με διαδικτυακές επιθέσεις τύπου phishing. Το phishing είναι πρακτικά το ΄΄Ηλεκτρονικό ψάρεμα΄΄. Ο θύτης της συγκεκριμένης επίθεσης υποδύεται μια αξιόπιστη οντότητα και εκμεταλλεύεται την άγνοια του θύματος-χρήστη με σκοπό την αθέμιτη απόκτηση προσωπικών δεδομένων όπως οι κωδικοί πρόσβασης σε μία εφαρμογή, το email, το ονοματεπώνυμο του χρήστη, την διεύθυνση IP (internet protocol) καθώς ακόμα και διάφορα τραπεζικά στοιχεία. Το phishing στηρίζεται στην παραπλάνηση που δημιουργεί ο θύτης προς το θύμα και έτσι τον ΄΄πείθει΄΄. Παρακάτω αναγράφουμε κάποιες βασικές μορφές Phishing με ένα αντίστοιχο παράδειγμα.

Παραπλανητικό κείμενο.

Το κείμενο αυτό είναι συνήθως κάποιος σύνδεσμος που έχει διαφορετική σύνταξη ή ορθογραφία με σκοπό να μπερδέψει τον χρήστη (π.χ. www.twetter.com αντί για www.twitter.com που είναι το σωστό)

Παραπλανητικές εικόνες

Ο θύτης χρησιμοποιεί ίδιες εικόνες δηλαδή ίδια λογότυπα με τα λογότυπα των κανονικών ιστοσελίδων στις οποίες ο χρήσης νομίζει ότι εισέρχεται και σε συνδυασμό με το παραπλανητικό κείμενο κάνει είσοδο σε κάποιο άλλο site.

Παραπλανητικό design

Με την βοήθεια του παραπλανητικού κειμένου και των παραπλανητικών εικόνων ο θύτης μπορεί να δημιουργήσει μια ιστοσελίδα η οποία να έχει την ίδια εμφάνιση (interface) με την αυθεντική ιστοσελίδα στην οποία εμείς θέλουμε όντως να συνδεθούμε.

Απειλητικό μήνυμα

Μια ακόμα τεχνική που μπορεί να χρησιμοποιήσει ο θύτης είναι η απειλή μέσω μηνυμάτων με σκοπό να μας οδηγήσει στον παραπλανητικό ιστότοπο που έχει στήσει. (π.χ. «αν δεν ακολουθήσετε τον σύνδεσμο, ο λογαριασμός σας θα κλειδωθεί» ή «Έγινε συναλλαγή 500€ από τον λογαριασμό σας, για την ακύρωσή της πατήστε εδώ»)

Τρόποι αντιμετώπισης

Καθώς το πρόβλημα του phishing δείχνει πως δεν θα σταματήσει, αναγκαίο είναι να υπάρξουν τρόποι να αποφευχθεί καθώς αν συμβεί να υπάρξουν κυρώσεις στον θύτη. Η αντιμετώπιση λοιπόν περιλαμβάνει τρεις τομείς. Την νομοθεσία, την εκπαίδευση των χρηστών και τα τεχνικά μέτρα ασφαλείας.

Στο κομμάτι της νομοθεσίας οι διάφορες κυβερνήσεις έχουν θεσπίσει τα κατάλληλα νομικά πλαίσια. Το 2006 στο Ηνωμένο Βασίλειο θεσπίστηκε το «Fraud Act» που ορίζει την ηλεκτρονική απάτη ως αδίκημα το οποίο τιμωρείται με ποινή

φυλάκισης έως και 10 ετών ενώ παράλληλα απαγορεύει την κατοχή εργαλείων ηλεκτρονικού ψαρέματος. Αντίστοιχα στις ΗΠΑ το 2006 θεσπίστηκε το «Anti-Phishing Act» το οποίο καταδικάζει την κλοπή ταυτότητας μέσω παραποιημένων ιστοσελίδων με ποινή φυλάκισης 5 ετών. Σχετικά με την Ελληνική νομοθεσία συμπεραίνεται ότι το «phishing» είναι απάτη κατά το άρθρο 386 του Ποινικού Κώδικα.

Παράλληλα η ενημέρωση των χρηστών είναι απαραίτητη διότι η οδηγεί στην πρόληψη του προβλήματος. Υπάρχουν διάφορες ιστοσελίδες που παρέχουν ενημέρωση του κοινού, καταγράφουν διάφορα περιστατικά και στατιστικά στοιχεία σχετικά με επιθέσεις phishing. Κάποια μάλιστα παρέχουν μαζί με antivirus και διάφορες ακόμα εφαρμογές που αναγνωρίζουν τα phishing sites. Μία από αυτές είναι η παρακάτω: https://fraudwatchinternational.com/

Τέλος στο κομμάτι της τεχνικής αντιμετώπισης μπορούμε να χρησιμοποιήσουμε πολλά από τα παρακάτω:

- Λήψη antivirus που παρέχεται μαζί safe browsing advisor (σύμβουλος ασφαλούς περιήγησης)
- Λήψη προγραμμάτων anti-spam για την προστασία των email
- Λήψη προγραμμάτων περιήγησης που αναγνωρίζουν τους παραπλανητικούς ιστοτόπους και τα παραπλανητικά μηνύματα μέσω διαφορετικού URL

OAuth2.0

Όπως αναφέραμε και στην προηγούμενη ενότητα η επικοινωνία και η εξαγωγή των δεδομένων θα γίνει μέσω του Twitter API. Για να επικοινωνήσουμε με ασφάλεια με το API θα χρειαστούμε την εξουσιοδότηση OAuth2.0 (Open Authorization 2.0).

Τι είναι το OAuth 2

Γενικά το OAuth2.0 είναι ένα πλαίσιο εξουσιοδότησης που επιτρέπει στις εφαρμογές να αποκτούν περιορισμένη πρόσβαση σε λογαριασμούς χρηστών σε υπηρεσίες HTTP (Hypertext Transfer Protocol). Το HTTP είναι πρακτικά το πρωτόκολλο επικοινωνίας που χρησιμοποιούν οι browsers (φυλλομετρητές) για την μεταφορά δεδομένων ανάμεσα στον διακομιστή (server στην περίπτωσή μας το API του Twitter) και έναν πελάτη (client). Συνεπώς αυτό που κάνει το OAuth 2 είναι να μεταβιβάζει τον έλεγχο της ταυτοποίησης του χρήστη στην υπηρεσία που φιλοξενεί τον λογαριασμό του και να εξουσιοδοτεί ¨τρίτες¨ εφαρμογές με σκοπό να έχουν πρόσβαση στον λογαριασμό του χρήστη.

Εικόνα 1.2 Λογότυπο OAuth

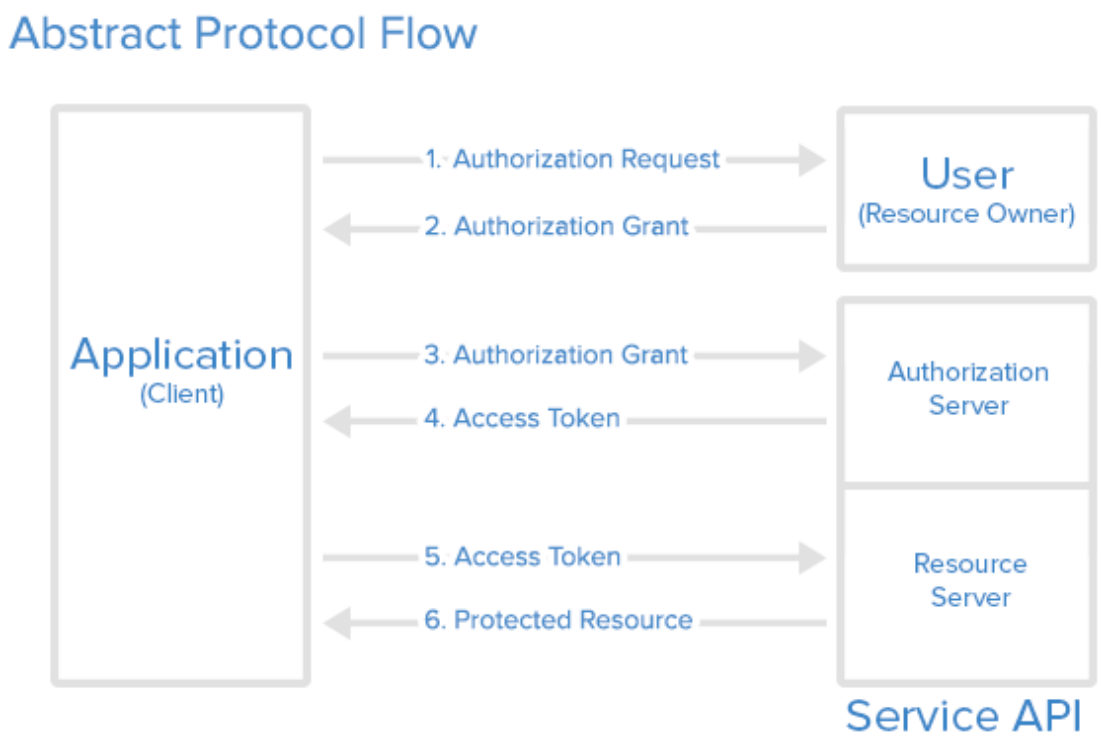
Πηγή: https://en.wikipedia.org/wiki/Twitter

Πως λειτουργεί το OAuth 2

Οι ¨τρίτες¨ εφαρμογές που μπορούν να εξουσιοδοτηθούν είναι δυνατόν να βρίσκονται στον δικό μας υπολογιστή κάτι το οποίο ισχύει στην δική μας περίπτωση. Η εφαρμογή που χρησιμοποιούμε ονομάζεται PyCharm και ουσιαστικά

είναι ένας code editor της γλώσσα προγραμματισμού Python μέσω της οποίας δουλεύουμε. Στο παρακάτω διάγραμμα βλέπουμε τα βασικά βήματα που ακολουθούμε μέσω του OAuth 2 για την εξαγωγή των πόρων που θέλουμε.

Διάγραμμα 1.4: Τρόπος λειτουργίας του OAuth 2



Πηγή: https://www.digitalocean.com/community/tutorials/an-introduction-to-oauth-2

Ας δούμε αναλυτικά τα βήματα που βλέπουμε στο παραπάνω διάγραμμα.

1. Η εφαρμογή που χρησιμοποιούμε ζητά εξουσιοδότηση για πρόσβαση σε πόρους υπηρεσίας από τον χρήστη.
2. Αν ο χρήστης εξουσιοδότησε το αίτημα της εφαρμογής τότε αυτή λαμβάνει μια έγκριση της εξουσιοδότησης αυτής.
3. Έπειτα η εφαρμογή ζητά ένα διακριτικό πρόσβασης (access token) από τον server δηλαδή το API παρουσιάζοντας τον δικό της έλεγχο ταυτότητας και την δική της έγκριση εξουσιοδότησης.

4. Αν η ταυτότητα της εφαρμογής πιστοποιηθεί και η έγκριση εξουσιοδότησης που έχει λάβει είναι έγκυρη τότε ο server εκδίδει το διακριτικό πρόσβασης που ζητήθηκε.
5. Η εφαρμογή ζητά διάφορους πόρους από το API παρουσιάζοντας το access token για έλεγχο ταυτότητας.
6. Αν το access token είναι έγκυρο τότε το API εξυπηρετεί το αίτημα της εφαρμογής επιστρέφοντας τους πόρους που ζητήθηκαν.

Η χρήση του OAuth 2 δεν είναι απλά αναγκαία αλλά υποχρεωτική για την ορθή και ασφαλή επικοινωνία με την πλατφόρμα του Twitter.

Κεφάλαιο 2

2.1 Στόχοι της εργασίας

Αναγνωρίζοντας την ολοένα και ταχύτερη εξέλιξη του διαδικτύου και των κοινωνικών δικτύων μπορούμε να αντιληφθούμε πως η ανάλυση των δεδομένων που παράγονται είναι απαραίτητη. Όπως έχουμε αναφέρει λοιπόν η εξόρυξη των δεδομένων θα γίνει με βάση το Twitter μέσω του API και με την ασφάλεια του OAuth 2.0. Αφού εξηγήσαμε τους όρους Twitter API και το πρωτόκολλο OAuth 2.0 παρακάτω αναφέρουμε τα βήματα που η εργασία στοχεύει να αναλύσει και να εξηγήσει:

- Τι είναι ανάλυση δεδομένων;
- Τι είναι Επεξεργασία Φυσικής Γλώσσας;
- Με ποια μέθοδο θα γίνει το ταίριασμα των λέξεων;
- Ποιο προγραμματιστικό περιβάλλον και ποια γλώσσα προγραμματισμού θα χρησιμοποιηθεί για την ανάλυση των δεδομένων;
- Ποιος είναι ο τρόπος συλλογής των δεδομένων;
- Πως αποθηκεύσαμε τα δεδομένα;
- Ποια φίλτρα χρησιμοποιήσαμε κατά την επεξεργασία των δεδομένων;
- Με ποια κριτήρια θα εξάγουμε τα αποτελέσματα μας;
- Πως θα οπτικοποιηθούν αυτά;
- Μελλοντική έρευνα πάνω στο θέμα.

Τα παραπάνω ερωτήματα θα απαντηθούν στις επόμενες ενότητες. Παράλληλα ο τελικός στόχος της εργασίας είναι η ερμηνεία των αποτελεσμάτων. Το κατά πόσο ανταποκρίνονται στην ελληνική πραγματικότητα και στην καθημερινότητα των χρηστών.

2.2 Ανάλυση δεδομένων

Κάθε post που κάνουμε, κάθε αρχείο που ανεβάζουμε δημόσια σε μία πλατφόρμα, κάθε τι που παρακολουθούμε στο διαδίκτυο υπάρχει διαθέσιμο για την αντίστοιχη εταιρία που κατέχει την πλατφόρμα για ανάλυση. Η διαδικασία που ακολουθεί μετά λέγεται Ανάλυση Δεδομένων.

Η ανάλυση δεδομένων είναι ένα κομμάτι της επιστήμης δεδομένων. Τι είναι όμως η επιστήμη δεδομένων; Η επιστήμη των δεδομένων κάνει εκτεταμένη χρήση τεχνικών και θεωριών από διάφορους τομείς όπως τα μαθηματικά, την επιστήμη της πληροφορίας και η επιστήμη των υπολογιστών. Στην πρακτική προσέγγιση περιλαμβάνει την ανάλυση σημάτων, τα προγνωστικά μοντέλα, τη μηχανική μάθηση (machine learning), τη στατιστική, την εξόρυξη δεδομένων, τις βάσεις δεδομένων (databases), τον προγραμματισμό αλλά, και τέλος, την τεχνητή νοημοσύνη. Οι μέθοδοι διαχείρισης των μεγάλων δεδομένων (big data) έχουν πιθανώς το μεγαλύτερο ενδιαφέρον της συγκεκριμένης επιστήμης, παρόλο που οι μέθοδοι που χρησιμοποιούνται στην επιστήμη δεδομένων δεν αφορούν αποκλειστικά μεγάλους όγκους δεδομένων.

Συγκεκριμένα λοιπόν η ανάλυση δεδομένων χωρίζεται σε πολλά στάδια. Τα πιο βασικά από αυτά είναι:

- Επιλογή των κατάλληλων δεδομένων
- Συλλογή δεδομένων
- Επεξεργασία δεδομένων
- Καθάρισμα δεδομένων
- Διερευνητική ανάλυση
- Οπτικοποίηση δεδομένων

Ας αναφερθούμε διεξοδικά σε κάθε στάδιο της ανάλυσης.

Επιλογή των κατάλληλων δεδομένων

Για να ξεκινήσει η διαδικασία πρέπει να γνωρίζουμε τι είδους δεδομένα αναζητούμε. Αυτό το στάδιο περιέχει τους περιορισμούς που θα βάλουμε εμείς στον τρόπο συλλογής. Στην παρούσα εργασία εμείς αναζητούμε Tweets που περιέχουν κείμενο οπότε αφαιρούμε τα Tweets που έχουν φωτογραφίες καθώς και αυτά που περιέχουν links. Ακόμα θα υπάρξει και περιορισμός ως προς την περιοχή για την οποία αναφερόμαστε. Εμείς θέλουμε να συλλέξουμε Tweets στην περιοχή της Αθήνας οπότε δίνοντας την γεωγραφική θέση της περιοχής μας μπορούμε να ζητήσουμε η συλλογή να γίνει μόνο από τον συγκεκριμένο χώρο.

Συλλογή δεδομένων

Η συλλογή είναι το στάδιο του κατεβάσματος (download) των δεδομένων μέσω του Twitter API από τους servers του Twitter στο δικό μας αρχείο. Η διαδικασία αυτή γίνεται μέσα από την Python στην οποία έχουμε μεταφέρει τους κωδικούς ασφαλείας του API. Η συλλογή είναι ημερήσια καθώς τα tweets που συλλέγονται έχουν παράλληλα και την ¨σφραγίδα¨ (timestamp) της ημέρας που γράφτηκε και ανέβηκε το tweet στο διαδίκτυο. Τέλος η αποθήκευση τους στον δικό μας υπολογιστή έγινε σε αρχεία μορφής csv (comma separated values).

Επεξεργασία δεδομένων

Η επεξεργασία των δεδομένων είναι το κομμάτι της οργάνωσης τους με σκοπό την καλύτερη ανάλυση. Για παράδειγμα η τοποθέτηση των δεδομένων σε σειρές και στήλες σε ένα υπολογιστικό φύλλο ή σε κάποιο στατιστικό πρόγραμμα κάνει τα επόμενα βήματα ακόμα πιο απλά και εύκολα.

Καθάρισμα δεδομένων

Τα δεδομένα μας τώρα έχουν οργανωθεί όμως ίσως υπάρχουν περιπτώσεις διπλών μετρήσεων η λαθών κατά τις μετρήσεις. Το κομμάτι του καθαρισμού των

δεδομένων σχετίζεται ακριβώς με αυτό και διευκολύνει την διερευνητική ανάλυση που ακολουθεί. Εμείς στο κομμάτι αυτό αφαιρέσαμε από τα κείμενά μας κάποιος συνηθισμένες λέξεις του ελληνικού λεξιλογίου που δεν ωφελούν στην ανάλυση. Για παράδειγμα η λέξη ¨και¨ είναι μια πολύ συχνή λέξη που όμως σαν συνδετικό που είναι, δεν προσφέρει κάποια ιδιαίτερη βοήθεια στον σκοπό μας.

Διερευνητική ανάλυση

Το κομμάτι της διερευνητικής ανάλυσης είναι το κομμάτι που πλέον μπορούμε να κάνουμε ότι θέλουμε με τα δεδομένα που έχουμε διαθέσιμα χρησιμοποιώντας διάφορες τεχνικές. Σε αυτό το σημείο εμείς χωρίσαμε τις λέξεις μεταξύ τους για να μπορέσουμε να βρούμε την συχνότητα εμφάνισης της κάθε λέξης. Έπειτα η διαδικασία αυτή επαναλήφθηκε για όλα τα tweets της ημέρες και για όλες τις ημέρες και έτσι προκύπτουν οι συχνότητες των λέξεων σε όλη τη διάρκεια του εξαμήνου.

Οπτικοποίηση δεδομένων

Το τελευταίο στάδιο είναι η οπτικοποίηση (visualization) των συμπερασμάτων. Σε αυτό το κομμάτι δημιουργήσαμε διάφορα γραφήματα που παρουσιάζουν αναλυτικά τα αποτελέσματα της έρευνας.

Πλέον αφού μιλήσαμε για την ανάλυση δεδομένων πρέπει να αναφερθούμε και στον τύπο δεδομένων που εμείς θα συλλέξουμε. Τα Tweets όπως ξέρουμε είναι γραμμένα σε φυσική γλώσσα, δηλαδή στα Ελληνικά τις περισσότερες φορές μια και η περιοχή που εξετάζουμε είναι η Αθήνα, αλλά προφανώς υπάρχει και ένα μεγάλο κομμάτι αυτών που περιέχει Αγγλικό λεξιλόγιο. Η διαδικασία της ανάλυσης της φυσικής γλώσσας των Tweets ονομάζεται Επεξεργασία Φυσικής Γλώσσας (ΕΦΓ) ή αλλιώς Natural Language Processing (NLP)

2.3 Επεξεργασία Φυσικής Γλώσσας (ΕΦΓ)

Τι είναι ΕΦΓ

Ο όρος Επεξεργασία Φυσικής Γλώσσας χρησιμοποιείται για να περιγράψει την διαδικασία που ακολουθεί ένα λογισμικό πρόγραμμα σε ένα υπολογιστικό σύστημα με σκοπό να αναλύσει ή να συνθέσει γλώσσα σε γραπτή ή προφορική μορφή. Η λέξη ¨Φυσικής¨ Γλώσσας υπάρχει για να γίνει ξεκάθαρος διαχωρισμός ανάμεσα στην ανθρώπινη γλώσσα επικοινωνίας που υπάρχει σε προφορική και γραπτή μορφή σε σχέση με τις επίσημες γλώσσες προγραμματισμού που υπάρχουν όπως η Python ή η Java. Αν θέλουμε να είμαστε πιο αυστηροί στον ορισμό μας ο σκοπός της ΕΦΓ είναι η κατανόηση της φυσικής γλώσσας (Natural Language Understanding).

Προφανώς η κατανόηση έχει να κάνει με το πόσο ένα υπολογιστικό σύστημα μπορεί να αντιλαμβάνεται και να χρησιμοποιεί την φυσική γλώσσα (γραπτή ή προφορική) στον ίδιο βαθμό που το κάνουμε και εμείς. Όπως γνωρίζουμε τα υπολογιστικά συστήματα μπορούν να προγραμματιστούν ώστε να αναγνωρίζουν έναν κώδικα που είναι γραμμένος στην γλώσσα Python για παράδειγμα. Επίσης μέσω του ανάλογου κώδικα μπορούν να επιλύσουν σύνθετα προβλήματα μαθηματικού αλλά και λογικού χαρακτήρα. Όταν όμως φτάνουμε στο σημείο της ανάλυσης κειμένου εμφανίζονται πολλά προβλήματα που παραμένουν άλυτα.

Σκοπός της ΕΦΓ

Η λύση των παραπάνω προβλημάτων δεν θα ήταν απαραίτητη αν δεν υπήρχε τόσο μεγάλη ανάγκη για την απόκτηση των πληροφοριών που υπάρχουν πίσω από τα γραπτά κείμενα που είναι διαθέσιμα στο διαδίκτυο. Αυτό καθιστά τον κλάδο της ΕΦΓ ταχύτατα αναπτυσσόμενο. Οι μεγάλες τεχνολογικές εταιρίες, οι οποίες κατέχουν μεγάλο μέρος των πληροφοριών είναι οι πρωτοπόροι σε αυτόν τον κλάδο. Πλέον λοιπόν το ζήτημα δεν έχει να κάνει με την έλλειψη πληροφοριών αλλά με τον

τρόπο αξιοποίησής τους. Συνεπώς όπως είναι προφανές η ΕΦΓ έχει σαν στόχο την αξιοποίηση των διαθέσιμων πηγών που υπάρχουν στο διαδίκτυο με σκοπό την εξαγωγή συμπερασμάτων. Τα πεδία στα οποία έχει αναπτυχθεί η ΕΦΓ είναι αρκετά. Σε πολλά από αυτά η πρόοδος που έχει γίνει τα τελευταία χρόνια είναι πολύ μεγάλη αλλά σε άλλα υπάρχει ακόμα χώρος για πρόοδο. Ας δούμε κάποια από τα πιο σημαντικά πεδία παρακάτω.

Αυτόματη αναγνώριση ομιλίας

Το κομμάτι αυτό σχετίζεται με την μετατροπή του προφορικού λόγου σε κείμενο από το υπολογιστικό σύστημα. Αυτό το πεδίο έχει αναπτυχθεί ιδιαίτερα καθώς πλέον πολλές online πλατφόρμες έχουν συμπεριλάβει προφορική αναζήτηση στις υπηρεσίες τους.

Μηχανική μετάφραση

Το πεδίο αυτό έχει είδη αναπτυχθεί αρκετά και σκοπό έχει την μετάφραση ενός κειμένου από μια ανθρώπινη γλώσσα σε μία άλλη. Επίσης η μετάφραση αυτή μπορεί να γίνει πλέον μέσω προφορικού λόγου. Παρά την ανάπτυξη του πεδίου, αυτό που παραμένει δύσκολο είναι η ορθή σύνταξη του κειμένου που έχει μεταφραστεί.

Σύνθεση ομιλίας

Ο τομέας αυτό αφορά στην τεχνητή αναπαραγωγή ανθρώπινου λόγου από τους υπολογιστές. Αποτελεί πλέον έναν κλάδο αρκετά ανεπτυγμένο, καθώς τα σύγχρονα κινητά τηλέφωνα αλλά και οι υπολογιστές, έχουν ψηφιακούς ¨βοηθούς¨ που μπορούν να αλληλοεπιδράσουν προφορικά με την χρήστη (π.χ. Siri, Google Assistant, Alexa)

Εύρεση μέρους του λόγου

Στο πεδίο αυτό γίνεται διαχωρισμός όλων των λέξεων ενός κειμένου με σκοπό να βρεθεί τι μέρος του λόγου είναι η κάθε μια από τις λέξεις. Ο βαθμός δυσκολίας του συγκεκριμένου ζητήματος εξαρτάται από την γλώσσα που έχουμε επιλέξει. Για παράδειγμα η αγγλική καθώς και η ελληνική γλώσσα μπορούν πιο εύκολα να αναλυθούν σε σχέση με τα μανδαρινικά (βασικότερη κινεζική διάλεκτος).

Ορθογραφικός έλεγχος

Έλεγχος της ορθογραφίας του κειμένου. Βαθιά ανεπτυγμένο πεδίο.

Συντακτικός έλεγχος

Ο έλεγχος της γλωσσικής ορθότητας του κειμένου. Ο κλάδος αυτός δεν έχει τελειοποιηθεί ακόμα αλλά αναπτύσσεται.

Διαχωρισμός κειμένου

Η διαδικασία αυτή προσπαθεί να χωρίσει με ορθό τρόπο το κείμενο μας σε παραγράφους και σε προτάσεις.

Συχνότητα εμφάνισης λέξεων

Χωρίζοντας το κείμενο μας σε μεμονωμένες λέξεις μπορούμε να βρούμε αυτές που ταιριάζουν απόλυτα μεταξύ τους και να ελέγξουμε το πόσο συχνά εμφανίζονται αυτές. Ο τρόπος ταιριάσματος δεν είναι μοναδικός και εξαρτάται από το πόσο ακριβείς θέλουμε να είμαστε στην ομοιότητα των λέξεων μας.

Πηγή λέξεων

Πεδίο στο οποίο ο σκοπός είναι να βρεθούν λέξεις με παρόμοιο θέμα και διαφορετική κατάληξη. Για παράδειγμα η λέξεις ¨κλειστό¨ και ¨κλειστά¨ διαφέρουν μόνο στην κατάληξή τους καθώς ο αρχικός φθόγγος είναι ίδιος. Έτσι θα μπορούσαμε να πούμε πως το νόημα των λέξεων αυτών είναι παρόμοιο, όχι όμως ίδιο. Προφανώς, σημαντικό ρόλο παίζει το πλαίσιο μέσα στο οποίο χρησιμοποιούμε την κάθε λέξη. Το συγκεκριμένο πεδίο βοηθάει στο να βγει κάποιο κοινό νόημα μέσα στο κείμενο μας.

Συναισθηματική ανάλυση

Η εξαγωγή υποκειμενικών πληροφοριών σχετικά με το είδος του περιεχομένου του κειμένου μας. Για παράδειγμα η πρόταση «Χρόνια πολλά, καλή χρονιά σε όλους» έχει προφανώς θετικό χαρακτήρα σε αντίθεση με την φράση «Απέτυχε το κυβερνητικό πλάνο σχετικά με τους εμβολιασμούς» η οποία έχει ξεκάθαρα αρνητικό χαρακτήρα. Το πως αναγνωρίζεται το συναίσθημα που πηγάζει από κάθε λέξει-φράση σχετίζεται με το λεξικό με το οποίο έχουμε εμείς τροφοδοτήσει το υπολογιστικό σύστημα. Συνεπώς η επιλογή του συναισθήματος εξαρτάται απόλυτα από το τι έχουμε διαλέξει.

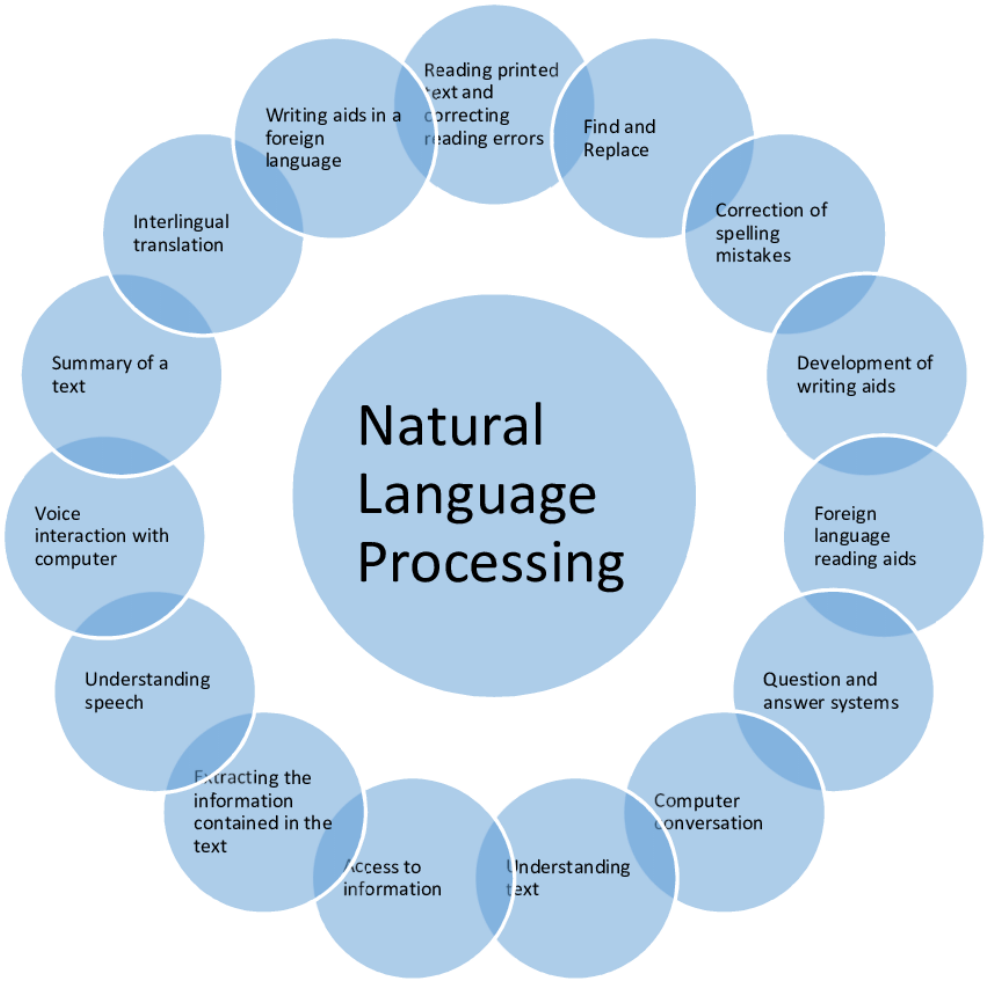
Δημιουργία περίληψης

Τροφοδοτώντας ένα υπολογιστικό σύστημα με ένα κείμενο σκοπός του πεδίου αυτού είναι η εξαγωγή περίληψης. Αυτό το κομμάτι είναι ακόμα υπό ανάπτυξη.

Δημιουργία φυσικής γλώσσας

Το πεδίο αυτό αφορά στην δημιουργία αυτούσιου νέου κειμένου από το υπολογιστικό σύστημα.

Εικόνα 2.1: Τομείς της Επεξεργασίας Φυσικής Γλώσσας



Πηγή: https://www.researchgate.net/figure/Natural-Language-Processing-Topics-Adali-2013-4_fig1_337773927

Πολλοί από τους παραπάνω τομείς έχουν αναπτυχθεί αρκετά ενώ άλλοι βρίσκονται ακόμα σε αρχικά στάδια. Το σίγουρο είναι πως η ΕΦΓ έχει σημαντικά τεχνολογικά οφέλη.

Χρήσεις της ΕΦΓ

Η πιο διαδεδομένη εφαρμογή που κάνει χρήση της Επεξεργασίας Φυσικής Γλώσσας είναι το Google Translate. Είναι η γνωστότερη εφαρμογή μηχανικής

μετάφρασης κειμένου καθώς χρησιμοποιείται από 500 εκατομμύρια χρήστες καθημερινά και σε αυτή μεταφράζονται πάνω από 100 δισεκατομμύρια λέξεις ημερησίως.

Το Facebook επίσης χρησιμοποιεί την μηχανική μετάφραση. Σε αναρτήσεις και σχόλια χρηστών υπάρχει η δυνατότητα μετάφρασης με σκοπό την πιο εύκολη επικοινωνία μεταξύ των χρηστών από διαφορετικές χώρες.

Τέλος το eBay ακολουθεί αντίστοιχα μονοπάτια με σκοπό να επεκτείνει τις συναλλαγές του πραγματοποιούνται από τους χρήστες του, πέρα από τα όρια των συνόρων και να ενώσει των αγοραστή με τον πωλητή όπου κι αν βρίσκονται στον πλανήτη.

2.4 Τρόπος ταιριάσματος λέξεων

Η ανάλυση που θα κάνουμε εμείς σχετίζεται με την συχνότητα εμφάνισης των λέξεων. Συνεπώς απαραίτητο είναι να αναφέρουμε τον τρόπο ταιριάσματος των νημάτων δηλαδή των λέξεων. Η μέθοδος που θα χρησιμοποιήσουμε ονομάζεται Απόσταση Hamming.

Τι είναι η Απόσταση Humming

Απόσταση Humming ονομάζεται το πλήθος των θέσεων που διαφέρουν σε δύο νήματα (λέξεις) ίδιου μήκους. Ουσιαστικά παρουσιάζει το ελάχιστο πλήθος μετατροπών που πρέπει να κάνουμε στο ένα νήμα ώστε να γίνει ολόιδιο με το άλλο. Ας φέρουμε μερικά παραδείγματα:

- «Διαβά**ζ**ω» και «Διαβά**σ**ω» διαφέρουν μόνο κατά ένα γράμμα κι έτσι η Απόσταση Hamming ανάμεσα σε αυτές τις δύο λέξεις είναι 1.
- 3674 και 3474 η απόσταση είναι 2.

Εμείς στην συγκεκριμένη εργασία για να ελαχιστοποιήσουμε τα λάθη θα επιλέξουμε λέξεις με Απόσταση Hamming Μηδέν.

Πλεονέκτημα

Η αντιστοίχιση είναι απόλυτα ακριβής.

Μειονέκτημα

Οι λέξεις με μικρά λάθη συντακτικά ή ορθογραφικά δεν μπόρεσαν να ταιριάξουν απόλυτα και έτσι δεν προστέθηκαν στο σύνολο. Επίσης λέξεις με παρόμοια σημασία, που θα μπορούσαν να μπουν στην ίδια κατηγορία όπως αυτές στο παραπάνω παράδειγμα που φέραμε έμειναν χωρισμένες.

Η απόσταση Hamming πήρε το όνομα της από τον Richard Hamming που την εισήγαγε το 1950 στην πληροφορική επιστήμη με σκοπό την διόρθωση σφαλμάτων (error correcting). Οι χρήσης της επεκτείνεται πέρα από το ταίριασμα νημάτων και την πληροφορική επιστήμη, στους τομείς της κρυπτογραφίας και των τηλεπικοινωνιών.

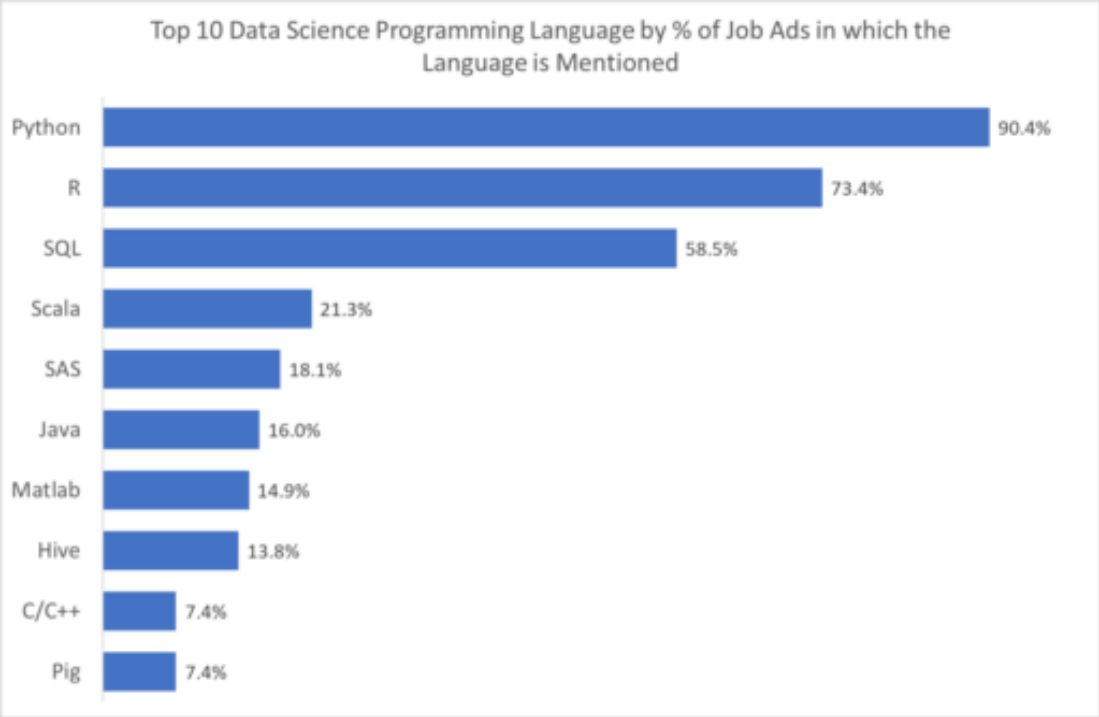
Κεφάλαιο 3

3.1 Python

Χρήση της Python

Ξεκινώντας την διαδικασία συλλογής των δεδομένων πρέπει να αναφέρουμε την γλώσσα προγραμματισμού που θα χρησιμοποιήσουμε. Η Python είναι μια αντικειμενοστραφής και δυναμική γλώσσα προγραμματισμού που σαν κύριο στόχο έχει την εύκολη αναγνωσιμότητα του κώδικα της. Το γεγονός αυτό, σε συνδυασμό με τις έτοιμες βιβλιοθήκες (libraries) που δημιουργούνται από την κοινότητα των χρηστών της Python καθιστούν την γλώσσα πολύ προσιτή σε άτομα που δεν έχουν μεγάλη τριβή με το αντικείμενο του προγραμματισμού, όπως εμείς οι μαθηματικοί. Αρκετές εργασίες που ίσως χρειαζόμασταν πολύ χρόνο για να πραγματοποιήσουμε μπορούν να εκτελεσθούν εύκολα και γρήγορα χάρη στην πληθώρα βιβλιοθηκών που διαθέτει η γλώσσα.

Διάγραμμα 3.1: Δημοτικότητα της Python στην αγορά εργασίας



Πηγή: https://towardsdatascience.com/

Για παράδειγμα αν θέλαμε να συντάξουμε έναν κώδικα που βρίσκει τον μέγιστο κοινό διαιρέτη δύο αριθμών θα σπαταλούσαμε πολύ περισσότερο χρόνο από την εναλλακτική λύση η οποία είναι να καλέσουμε την κατάλληλη βιβλιοθήκη και να απαραίτητη συνάρτηση της βιβλιοθήκης αυτής. Στην προκειμένη περίπτωση ο κώδικας θα είχε την παρακάτω μορφή:

```
import math

k=math.gdc(a,b)

print(k)
```

Η εντολή **import** είναι αυτή που εισάγει την βιβλιοθήκη στον κώδικα μας. Στη συνέχεια για να βρούμε τον μέγιστο κοινό διαιρέτη χρησιμοποιούμε την εντολή **gdc,** όμως επειδή αυτή υπάγεται στην βιβλιοθήκη μας γράφουμε μπροστά της το όνομα της βιβλιοθήκης αυτής. Παράλληλα γράφοντας k=… κάνουμε ανάθεση στην μεταβλητή k τον μέγιστο κοινό διαιρέτη των αριθμών a και b. Τέλος μέσω της εντολής **print** εμφανίζουμε το αποτέλεσμά μας.

Τύποι δεδομένων στην Python

Τα δεδομένα στην Python χωρίζονται σε δύο βασικές κατηγορίες. Τα αρχέγονα (primitive) δεδομένα δηλαδή δεδομένα που δίνονται σε μονάδες και τις δομές δεδομένων που πρακτικά είναι μεγαλύτερες ομάδες δεδομένων. Καθώς όπως είπαμε η Python αποτελεί μια δυναμική γλώσσα, δεν θα χρειαστεί να δηλώνουμε τον τύπο των μεταβλητών πριν χρησιμοποιηθούν. Ας αναφέρουμε κάποιους τύπους δεδομένων που θα συναντήσουμε εμείς στην ανάλυσή μας.

Αρχέγονα δεδομένα

- int Ακέραιος αριθμός οσοδήποτε μεγάλος
- float Αριθμός κινητής υποδιαστολής
- str Συμβολοσειρά, δηλαδή ακολουθία χαρακτήρων που περικλείονται μέσα σε μονά ή διπλά εισαγωγικά και μπορούν να περιέχουν πολλά είδη συμβόλων.
- bytes Αντικείμενο κωδικοποιημένο σε μορφή bytes

Δομές δεδομένων

- list Λίστα που θυμίζει την γνωστή δομή των πινάκων αλλά μπορεί να περιέχει ακόμα κι άλλες λίστες. Η λίστα ορίζεται με ένα ζευγάρι αγκυλών [ ] και ο διαχωρισμός των στοιχείων στο εσωτερικό της γίνεται με κόμμα. Εμείς θα χρειαστούμε την λίστα για να χωρίζουμε το κάθε Tweet σε μεμονωμένες λέξεις μέσω της λίστας.
- Tuple
- Dictionary Λεξικό που ορίζεται από ένα κλειδί και μία τιμή που αντιστοιχεί ακριβώς σε αυτό. Ορίζεται με ένα ζεύγος άγκιστρων {}

Προφανώς οι τύποι των δεδομένων δεν είναι μόνο αυτοί που αναφέραμε παραπάνω αλλά αποτελούν τους βασικότερους εντός της Python.

Τέλος η Python είναι μία δωρεάν γλώσσα και για την ανάλυσή μας χρησιμοποιήσαμε την έκδοση 3.9.1

Βασικές Βιβλιοθήκες

Όπως αναφέραμε οι βιβλιοθήκες είναι ίσως το σημαντικότερο χαρακτηριστικό της Python. Εδώ θα αναφέρουμε τις πιο βασικές βιβλιοθήκες που χρησιμοποιήσαμε στην ανάλυση μας.

tweepy

Βιβλιοθήκη που μας βοηθά στο να επικοινωνήσουμε με το Twitter API.

csv

Επιτρέπει την γραφή των δεδομένων σε αρχεία μορφής csv (coma separated values)

ast

Η βιβλιοθήκη αυτή περιέχει εντολές που βοηθούν στην κωδικοποίηση και αποκωδικοποίηση των αρχείων.

collections

Βιβλιοθήκη μέσω της οποίας θα ομαδοποιήσουμε τα δεδομένα.

re

Βιβλιοθήκη που μέσω της εντολής sub επιτρέπει την αντικατάσταση ενός κομματιού κειμένου με ένα άλλο.

nltk

Μεγάλη και πολύ σύνθετη βιβλιοθήκη που μας δίνει τις Ελληνικές λέξεις διακοπής (stop words) και μας βοηθά στον διαχωρισμό των φράσεων σε κομμάτια.

itertools

Μέσω της βιβλιοθήκης αυτής μπορούμε να μετρήσουμε την συχνότητα εμφάνισης των λέξεων.

pandas

Τεράστια βιβλιοθήκη με πάρα πολλές χρήσεις. Εμείς θα χρησιμοποιήσουμε το DataFrame για να έχουμε μια πιο όμορφη απεικόνιση των δεδομένων.

Εικόνα 3.1: Λογότυπο Python

Πηγή: https://www.python.org/

Προγραμματιστικό Περιβάλλον PyCharm

Για να μπορέσουμε να γράψουμε τον κώδικα μας θα χρειαστούμε ένα ολοκληρωμένο προγραμματιστικό περιβάλλον (Integrated Development Environment) ή αλλιώς IDE. Τα IDEs είναι εφαρμογές λογισμικού που παρέχονται από διάφορες εταιρίες με σκοπό την πιο βολική και άνετη χρήση των γλωσσών προγραμματισμού.

Το IDE που θα χρησιμοποιήσουμε εμείς ονομάζεται PyCharm. Δημιουργήθηκε από την Τσέχικη εταιρία JetBrains το 2011 και σκοπό έχει την διευκόλυνση των όσων προγραμματίζουν συγκεκριμένα στην Python. Η εφαρμογή παρέχει πολλά εργαλεία που διευκολύνουν την διαδικασία γραφής του κώδικα. Μερικά από αυτά είναι:

- Αυτόματη συμπλήρωση κώδικα.
- Αναζήτηση κώδικα
- Έλεγχος συντακτικών λαθών.
- Συντακτική επισήμανση.
- Γρήγορη διόρθωση σφαλμάτων.
- Έλεγχος εκδόσεων και αλλαγών.
- Debugger για την εξέταση και αποσφαλμάτωση λαθών.

Εικόνα 3.2: Λογότυπο PyCharm

Πηγή: https://www.jetbrains.com/pycharm/

Αφού αναφέραμε το περιβάλλον που θα μας βοηθήσει στο να πραγματοποιήσουμε την ανάλυση, μένει τώρα να μιλήσουμε για τον σχεδιασμό του κώδικα.

3.2 Σχεδιασμός

Τρόπος συλλογής δεδομένων

Αρχικά πρέπει να ξεκινήσουμε γράφοντας τον κώδικα που θα επικοινωνεί με το API και θα συλλέγει τα Tweets. Συνεπώς οι πρώτες εντολές που γράφουμε αφορούν στην επικοινωνία της εφαρμογής μας με το Twitter API δίνοντας του κωδικούς (API Key ,API Secret Key, Access Token, Access Token Secret Key) που είχαμε αναφέρει παραπάνω.

Αφού λοιπόν συνδεθούμε στο API με ασφάλεια μπορεί πλέον να ξεκινήσει η εξόρυξη των δεδομένων. Ας αναφέρουμε λοιπόν τον τρόπο συλλογής των ημερήσιων Tweets και τις παραμέτρους επιλογής.

- **Συχνότητα:** Ημερήσια συλλογή

- **Πλήθος Tweets:** 18.000
- **Γλώσσα:** Ελληνικά (οι Αγγλικοί χαρακτήρες δεν απαγορεύονται)
- **Γεωγραφική τοποθεσία:** Αθήνα με ακτίνα 25 χιλιομέτρων από το κέντρο

Πέρα από τις παραπάνω παραμέτρους υπάρχει και η δυνατότητα να φιλτράρουμε τα Retweets που πρακτικά είναι οι απαντήσεις σε ένα Tweet κάτι το οποίο επιλέξαμε να κάνουμε για να έχουμε μεγαλύτερη ποικιλία περιεχομένου.

Παραθέτοντας τα παραπάνω κριτήρια εντός του κώδικα μας ξεκινά η εξόρυξη των Tweets δίνοντας παράλληλα την κατάλληλη ημερομηνία αναζήτησης. Δεν θα εξαχθούν όλα τα Tweets κατευθείαν για λόγους ταχύτητας δικτύου οπότε τα 18.000 Tweets χωρίζονται σε εκατοντάδες και αποστέλλονται προς τον χρήστη. Τα στοιχεία που μπορούμε να συλλέξουμε είναι αρκετά αλλά εμείς έχουμε επιλέξει να κρατήσουμε τα εξής:

- Ακριβής Ώρα και Ημερομηνία που το Tweet γράφτηκε
- Περιεχόμενο του Tweet
- Όνομα χρήστη

Το παρακάτω παράδειγμα δείχνει τον την βασικότερη εντολή που χρησιμοποιούμε:

```
tweepy.Cursor(api.search, q=new_search, result_type='recent',
count=100, lang="el", since_id=0).items()
```

Αποθήκευση των δεδομένων

Αφού γίνει η ημερήσια συλλογή των Tweets αυτά πρέπει να αποθηκευτούν σε κάποιο αρχείο. Εμείς για την αποθήκευση τους επιλέξαμε την μορφή αρχείων με κατάληξη .csv . Αυτά είναι αρχεία τιμών που διαχωρίζονται με κόμμα (comma separated values). Καθώς η συλλογή μας αποτελείται από 18.000 Tweets αυτό σημαίνει πως το αρχείο μας θα έχει 18.000 γραμμές. Κάθε γραμμή του αρχείου csv θα περιέχει 3 τιμές. Η πρώτη είναι η ακριβής ημερομηνία και ώρα που

δημοσιεύτηκε το Tweet από τον χρήστη, η δεύτερη περιέχει το ίδιο το κείμενο που συντάχθηκε κωδικοποιημένο σε bytes και η τρίτη περιέχει το όνομα του χρήστη.

Το παρακάτω παράδειγμα δείχνει την μορφή των τιμών που έχουμε μέσα σε αυτά τα αρχεία:

2021-01-09 23:58:53,b'@SfChania \xce\xb1\xcf\x85\xcf\x84\xcf\x8c \xce\xbc \xce\xb1\xcf\x81\xce\xad\xcf\x83\xce\xb5\xce\xb9',b'lenapolitaki'

Παρατηρούμε πως η πρώτη τιμή είναι η ημερομηνία και η ώρα δημοσίευσης. Έπειτα, μετά το κόμμα, βλέπουμε ένα b΄ το οποίο μας δείχνει πως ότι είναι γραμμένο ανάμεσα στα εισαγωγικά είναι κωδικοποιημένο σε μορφή bytes με κωδικοποίηση UTF-8. Το ίδιο ισχύει για την τρίτη τιμή που είναι το όνομα χρήστη που συνέταξε το Tweet. Προφανώς τα δεδομένα που είναι κωδικοποιημένα δεν μπορούν να επεξεργαστούν όντας σε αυτήν την μορφή συνεπώς πρέπει να κάνουμε αποκωδικοποίηση η οποία θα αποτελέσει το πρώτο κομμάτι της επεξεργασίας των δεδομένων.

Επεξεργασία δεδομένων

Αποκωδικοποίηση

Ξεκινώντας την επεξεργασία των δεδομένων της ημέρας, το βασικότερο βήμα είναι η αποκωδικοποίηση τους από bytes σε strings έτσι ώστε να μπορούν να διαβαστούν. Η αποκωδικοποίηση γίνεται μέσα από την βιβλιοθήκη **ast** και συγκεκριμένα την εντολή **literal_eval** που μας επιτρέπει να αναγνωρίσουμε τον τύπο δεδομένων που έχουμε. Στη συνέχεια μέσω της εντολής **decode** αποκωδικοποιούμε το περιεχόμενο των Tweets. Επαναλαμβάνουμε αυτή τη διαδικασία για κάθε γραμμή του αρχείου μας και έτσι έχουμε τα δεδομένα σε μορφή που μπορούν να διαβαστούν.

Φιλτράρισμα

Το επόμενο κομμάτι είναι το φιλτράρισμα των δεδομένων. Τα Tweets εκτός από το κείμενο που μας είναι χρήσιμα, μπορεί να περιέχουν και στοιχεία που δεν αφορούν στην ανάλυσή μας. Αυτά μπορεί να είναι εξωτερικοί σύνδεσμοι (url) ή emojis, δηλαδή δεδομένα που δεν σκοπεύουμε να επεξεργαστούμε. Μέσω της βιβλιοθήκης **re** και ειδικότερα της εντολής **sub** θα αντικαταστήσουμε τα url και τα emojis κάθε μορφής με το **κενό**. Παρακάτω βλέπουμε τον κώδικα αντικατάστασης που χρησιμοποιήσαμε:

```
re.sub("(\w+:\/\/\S+)", "", txt)  #αντικατάσταση url

re.sub(emoji, '', data)  #αντικατάσταση emojis
```

Μεγάλο κομμάτι του φιλτραρίσματος επίσης αποτελούν οι λέξεις διακοπής (Greek Stop Words). Αυτές είναι διάφορες πολύ κοινές λέξεις του ελληνικού λεξιλογίου όπως άρθρα και συνδετικά και αφαιρούνται από την ανάλυση μας καθώς είναι πολύ συχνές και δεν έχουν κάποιο ξεκάθαρο νόημα. Μέσω της βιβλιοθήκης **nltk** και της εντολής **words** εισάγουμε στο πρόγραμμά μας ένα αρχείο με τίτλο «stop_words» που περιέχει όλες τις ελληνικές λέξεις διακοπής. Έπειτα ελέγχουμε αν στο κείμενο μας υπάρχουν λέξεις κοινές με αυτές του αρχείου και αν συμβαίνει αυτό τότε τις προσπερνάμε. Έτσι καταφέρνουμε να μειώσουμε σε μεγάλο βαθμό το πλήθος τον λέξεων που έχουμε συνολικά αλλά παράλληλα δεν έχουμε απώλεια ιδιαίτερα σημαντικών λέξεων.

Αντικαταστάσεις

Το φιλτράρισμα όμως δεν σταματάει εδώ. Πολλοί είναι οι χρήστες που είτε από επιλογή είτε κατά λάθος ξεχνούν να βάλουν τόνους όσο γράφουν. Έτσι σύμφωνα με την απόσταση Hamming, για παράδειγμα οι λέξεις: ¨πολύ¨ και ¨πολυ¨ θεωρούνται διαφορετικές μεταξύ τους κι έτσι η μία δεν προσμετράτε στο μέτρημα των εμφανίσεων της άλλης. Για την αποφυγή αυτού του προβλήματος επιλέγουμε να αντικαταστήσουμε όλα τα φωνήεντα γράμματα που περιέχουν τόνο με τα αντίστοιχα γράμματα χωρίς τόνο.

Συναντάμε το ίδιο πρόβλημα και σε περιπτώσεις που σύμβολα είναι ενωμένα με λέξεις χωρίς να υπάρχει κενό μεταξύ τους. Για παράδειγμα οι λέξεις: ¨συμφωνία¨ και ¨συμφωνία.¨ πάλι δεν θεωρούνται ίδιες λόγω της τελείας που υπάρχει στο τέλος. Πάλι λοιπόν επιλέγουμε να αντικαταστήσουμε κάποια από τα γνωστά σύμβολα που χρησιμοποιούμε στην γραφή. Αυτά είναι τα παρακάτω:

¨**;**¨ , ¨**!**¨, ¨**:**¨, ¨**.**¨, ¨**,**¨ , ¨**-**¨, ¨**»**¨, ¨**«**¨

Τέλος το πρόβλημα της αντιστοίχισης συνεχίζεται ανάμεσα στα κεφαλαία και τα πεζά γράμματα. Για παράδειγμα οι λέξεις: ¨Τράπεζα¨ και ¨τράπεζα¨ δεν αντιστοιχούν μεταξύ τους λόγω του κεφαλαίου ταυ στην αρχή της λέξεις. Αντικαθιστώντας λοιπόν όλα τα κεφαλαία γράμματα με τα αντίστοιχα πεζά προσπερνάμε αυτό το ζήτημα.

Μετρήσεις

Δημιουργία λίστας

Αφού φιλτράραμε τα δεδομένα το επόμενο μας κομμάτι είναι η καταμέτρηση της συχνότητας εμφάνισης των λέξεων. Κάθε Tweet στη συνέχεια ενώνεται μαζί με όλα το προηγούμενό του σε μια διαδικασία που έχει 18.000 βήματα, όσα και τα κείμενα που έχουμε. Τώρα πλέον αντί για 18.000 Tweets έχουμε ένα μεγάλο κείμενο φιλτραρισμένο και απλοποιημένο το οποίο θα το μετατρέψουμε σε μία τεράστια λίστα με λέξεις μέσω της βιβλιοθήκης **nltk**. Συγκεκριμένα η εντολή **word_tokenize** θα χωρίσει το κείμενο μας σε κομμάτια όπου κάθε κομμάτι αποτελεί αυτούσια μια λέξη όπως στο παρακάτω:

Είσοδος:

```
import nltk
word="χαίρομαι πολύ που είμαι εδώ"
wordlist=nltk.word_tokenize(word)
print(wordlist)
```

Έξοδος:

['χαίρομαι', 'πολύ', 'που', 'είμαι', 'εδώ']

Όπως βλέπουμε ο κώδικας αυτός χωρίζει μια πρόταση σε λέξεις οι οποίες τοποθετούνται σε μία λίστα. Το ίδιο ισχύει για το ενοποιημένο κείμενο των Tweets που έχουμε επεξεργαστεί.

Συχνότητα

Έχοντας πλέον μία λίστα με όλες τις λέξεις έφτασε η ώρα να δούμε ποιες από αυτές είναι ίδιες και πόσες φορές εμφανίστηκαν. Για να γίνει αυτό, έχουμε καλέσει τις βιβλιοθήκες **collections** και **itertools**. Η εντολή **chain** δημιουργεί μία νέα λίστα που ενώνει όλα τα στοιχεία της αρχικής λίστας που είναι ίδια μεταξύ τους. Συνεπώς όλες τις λέξεις που είναι ολόιδιες ενώνονται δίπλα δίπλα. Έπειτα η εντολή **Counter** μετρά το πλήθος εμφάνισης αυτών των λέξεων και δημιουργεί ένα dictionary που περιέχει την λέξη και δίπλα το πλήθος εμφανίσεων της. Ας δούμε το παρακάτω απλό παράδειγμα:

Είσοδος:

```
import nltk
import itertools
import collections
word="It was the best of times, it was the worst of times"
wordlist=nltk.word_tokenize(word)
final=list(itertools.chain(wordlist))
counts=collections.Counter(wordlist)
print(counts)
```

Έξοδος:

Counter({'was': 2, 'the': 2, 'of': 2, 'times': 2, 'It': 1, 'best': 1, ',': 1, 'it': 1, 'worst': 1})

Τέλος αφού γνωρίζουμε πλέον τις λέξεις που εμφανίζονται συχνότερα, επιλέγουμε να κρατήσουμε για το υπόλοιπο της ανάλυσης μας, τις 20 λέξεις που είχαν το μέγιστο πλήθος εμφανίσεων.

Αποθήκευση

Το μόνο που απομένει για να ολοκληρωθεί η ημερήσια ανάλυση είναι η αποθήκευση των νέων δεδομένων. Αυτή θα γίνει της βιβλιοθήκης **pandas** και της εντολής **DataFrame** που μας βοηθά να αποκτήσουμε μια καλύτερη οπτική απεικόνιση των δεδομένων, γράφοντας τα σε μορφή πίνακα. Επιλέξαμε η αποθήκευση αυτού του πίνακα να γίνει σε αρχεία μορφής excel δηλαδή με κατάληξη .xlsx . Αυτό μπορεί να γίνει μέσω της εντολής **to_excel.**

Η διαδικασία αυτή ήταν καθημερινή από την αρχή μέχρι και το τέλος της ημερήσιας συλλογής δεδομένων.

Τελικά αποτελέσματα

Για την εξαγωγή των τελικών αποτελεσμάτων και την δημιουργία των γραφημάτων θα χρησιμοποιήσουμε εργαλεία του Microsoft Excel. Αρχικά θα ενώσουμε τα αρχεία κατά μήνα. Έτσι θα έχουμε συγκεντρώσει σε ένα αρχείο τις λέξεις με την μεγαλύτερη εμφάνιση για τον συγκεκριμένο μήνα. Έπειτα μέσω του **συγκεντρωτικού πίνακα** που μπορούμε να φτιάξουμε στο excel θα μπορέσουμε να βρούμε συγκεκριμένα πόσες φορές αυτές οι λέξεις εμφανίστηκα στο σύνολο του μήνα. Μέσω αυτών των δεδομένων θα καταφέρουμε να σχεδιάσουμε τα γραφήματά μας.

Κεφάλαιο 4

4.1 Χρονικό πλαίσιο

Η ημερήσια συλλογή των Tweets ξεκίνησε στις 9 Ιανουαρίου και τελείωσε στις 9 Ιουλίου του ίδιου έτους. Συνεπώς το χρονικό πλαίσιο στο οποίο έγινε η συλλογή είναι το πρώτο μισό του 2021 και πιο συγκεκριμένα 182 ημέρες. Παρακάτω θα δούμε γραφήματα συνολικά για το διάστημα αυτό αλλά παράλληλα θα χωρίσουμε τα δεδομένα μας σε μήνες ώστε να δούμε την κατάσταση που επικρατεί κάθε μήνα.

4.2 Είδη γραφημάτων

Αυτό που μένει να κάνουμε λοιπόν είναι να παραθέσουμε τα αποτελέσματα. Αρχικά να τονίσουμε πως σε κάθε γράφημα εμφανίζονται οι 15 πρώτες λέξεις τις αντίστοιχης κατηγορίας. Ακόμα πρέπει αναφέρουμε πως τα παρακάτω γραφήματα χωρίζονται σε δύο κατηγορίες. Η πρώτη είναι το άθροισμα όλων τον εμφανίσεων της κάθε λέξης μέσα στο αντίστοιχο χρονικό πλαίσιο. Η δεύτερη είναι ο Μέσος Όρος των εμφανίσεων της κάθε λέξης ανά ημέρα εμφάνισης της δηλ.

$$M.O. = \frac{Άθροισμα\ Εμφανίσεων\ Λέξης}{Πλήθος\ ημερών\ εμφάνισης}$$

Το πρώτο είδος γραφήματος θα μας παρουσιάσει τις λέξεις που ήταν οι πιο κοινές και είχαν μεγαλύτερη εμφάνιση εντός του χρονικού πλαισίου. Σε αντίθεση με αυτό, το δεύτερο είδος γραφημάτων θα περιέχει λέξεις που μπορεί να κέντρισαν το ενδιαφέρον για πολύ λίγες ημέρες αλλά είχαν μεγάλο πλήθος εμφανίσεων.

Ένα παράδειγμα σχετικά με αυτήν την διαφοροποίηση είναι το εξής. Στις 12 Ιουνίου, στον ποδοσφαιρικό αγώνα που διεξήχθη στα πλαίσια του Euro 2020 ο

Δανός ποδοσφαιριστής Christian Eriksen υπέστη καρδιακή προσβολή στο 42ο λεπτό του αγώνα ενάντια στην Φινλανδία. Ο αγώνας σταμάτησε αλλά ευτυχώς τα χειρότερα αποφεύχθηκαν για τον 29χρονο ποδοσφαιριστή. Το γεγονός αυτό τάραξε ολόκληρο τον κόσμο και έτσι το όνομα του Eriksen έφτασε να αναφέρεται πολύ συχνά στα Tweets της 12ης Ιουνίου καθώς η λέξη ¨ερικσεν¨ εμφανίστηκε πάνω από 1200 φορές την ημέρα αυτή. Αυτό δείχνει πως η λέξη ¨ ερικσεν¨ δεν έχει συχνή εμφάνιση στο γράφημα του συνολικού πλήθους αλλά θα εμφανιστεί σίγουρα στο γράφημα του μέσου όρου καθώς το πλήθος είναι ιδιαίτερα υψηλό. Επίσης, στα γραφήματα του Μέσου Όρου, εκτός από την γραμμή που τη ύψος της δείχνει τον μέσο όρο της αντίστοιχης λέξης υπάρχει και μια δεύτερη (μπλε). Η μπλε γραμμή αντιστοιχεί στον δεξιό άξονα και δείχνει το πλήθος ημερών εμφάνισης της αντίστοιχης λέξης.

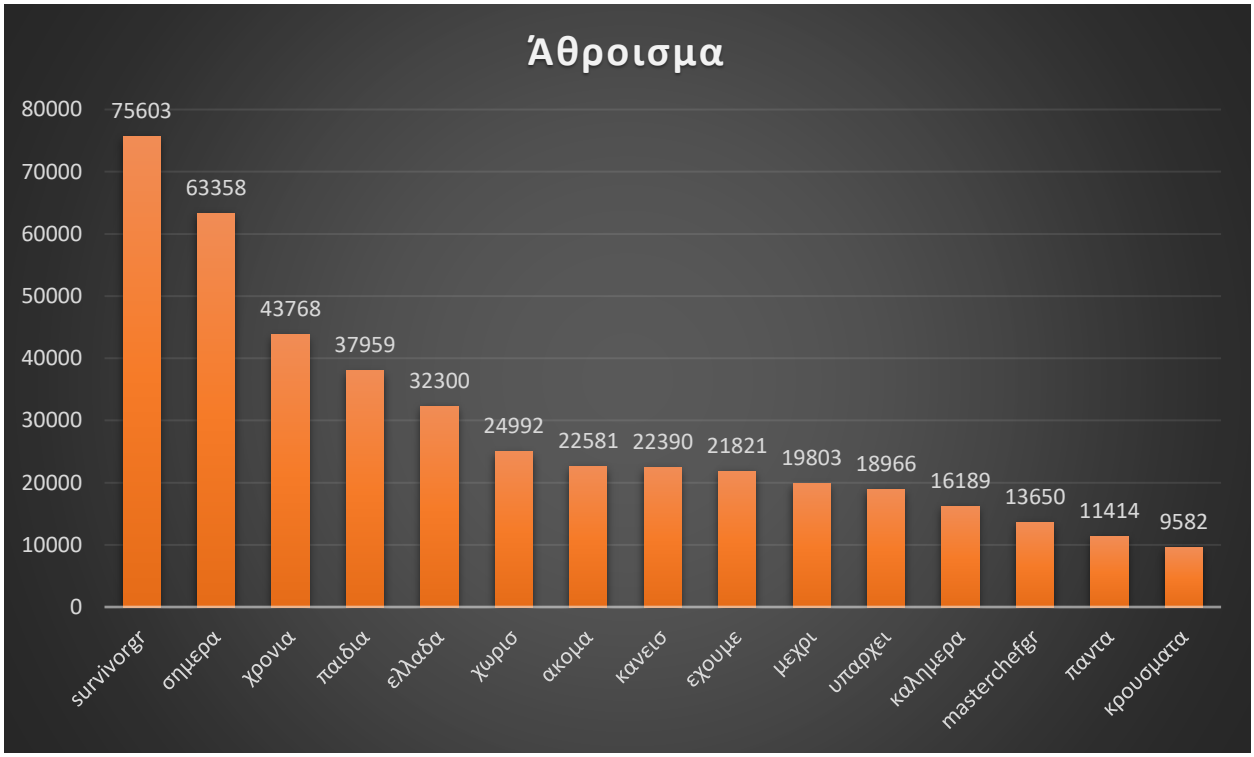
Τέλος ένα ακόμα γράφημα που θα δούμε είναι το γράφημα των λέξεων που είχαν το μεγαλύτερο πλήθος ημερών εμφάνισης. Το συγκεκριμένο γράφημα θα το δούμε μόνο στο συνολικό χρονικό διάστημα.

4.3 Παρουσίαση και Αξιολόγηση των αποτελεσμάτων

Συνολικό διάστημα

Αρχικά ξεκινάμε με τα συνολικά γραφήματα.

Διάγραμμα 4.1: Συνολικό άθροισμα

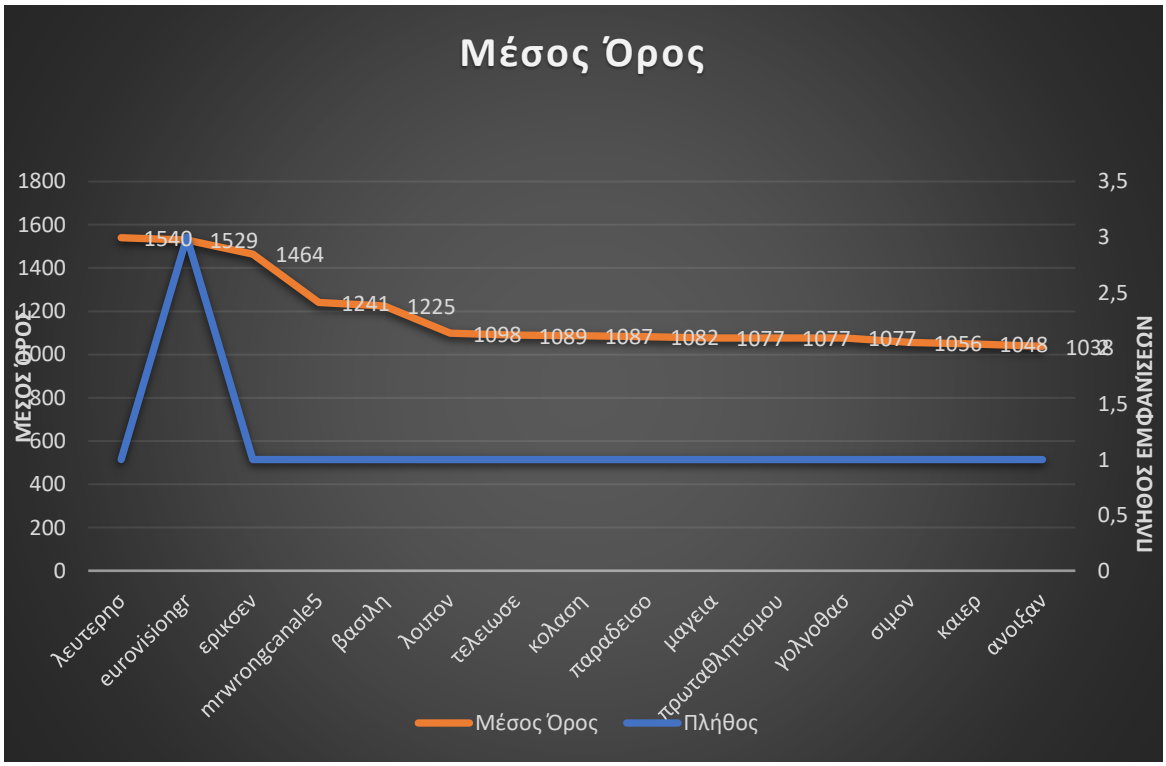


Όπως παρατηρούμε η λέξη ¨survivorgr¨ είναι αυτή που μονοπώλησε το ενδιαφέρον των χρηστών. Αυτό το μοτίβο θα συνεχιστεί και στα υπόλοιπα γραφήματα των άλλων μηνών καθώς το reality αυτό διήρκησε μέχρι τις 7 Ιουλιου. Πέραν από αυτήν όλες σχεδόν οι λέξεις που συναντάμε είναι αρκετά κοινές και χρησιμοποιούνται καθημερινά. Το γεγονός αυτό στο ευρύ χρονικό πλαίσιο των 182 ημερών θεωρείται λογικό.

Διάγραμμα 4.2: Συνολικό πλήθος

Για το συνολικό πλήθος βλέπουμε πως η λέξη ¨σημερα¨ είναι αυτή με τις περισσότερες εμφανίσεις καθώς μόνο έξι ημέρες δεν βρέθηκε στην πρώτη εικοσάδα της ημερήσιας συλλογής μας. Μία ακόμα παρατήρηση εδώ είναι πως οι λέξεις αυτές είναι σχεδόν όλες ίδιες με αυτές του προηγούμενου διαγράμματος.

Διάγραμμα 4.3: Συνολικός Μέσος Όρος



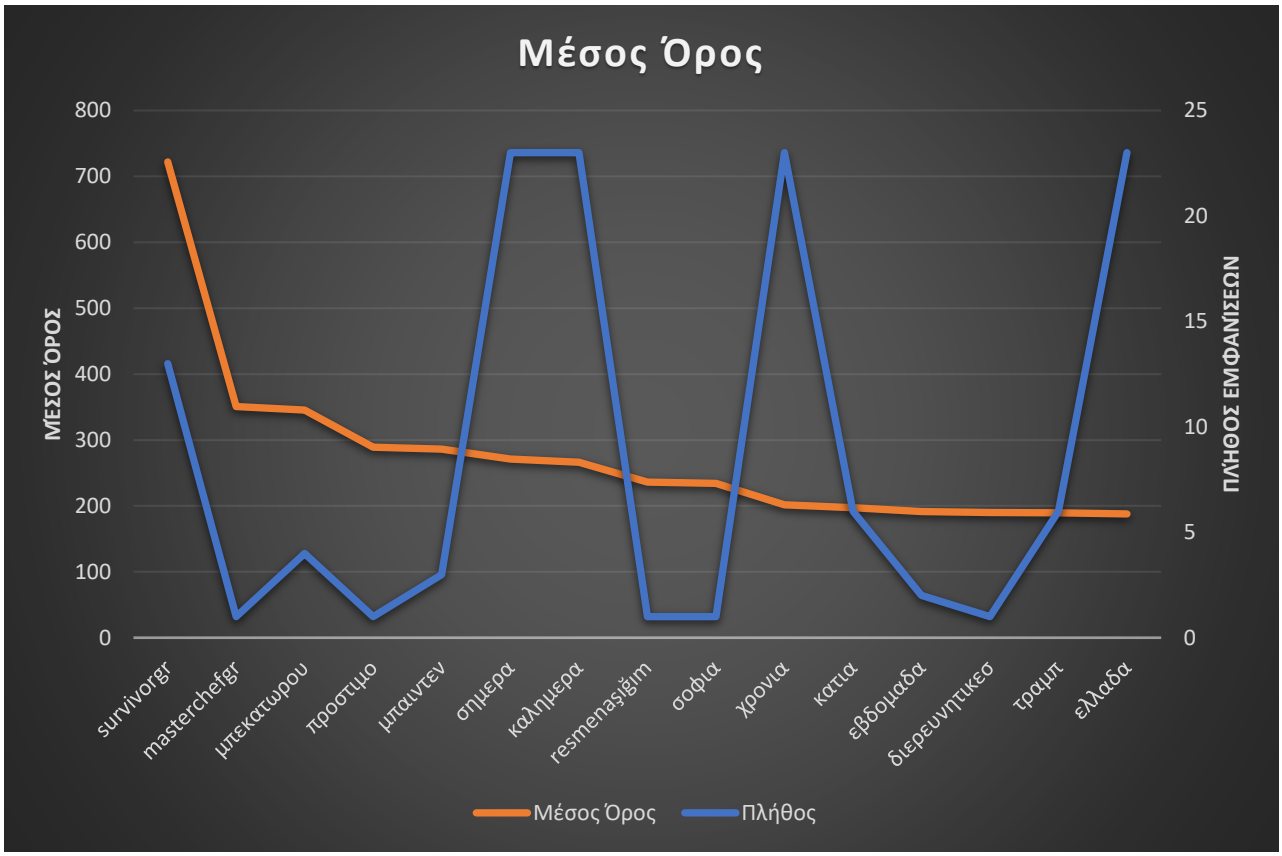
Παρατηρούμε πως οι λέξεις αυτές εμφανίστηκαν συνολικά ελάχιστες ημέρες όμως είχαν πολύ υψηλή συχνότητα τις ημέρες αυτές. Μπορούμε να διακρίνουμε τις λέξεις ¨eurovisiongr¨ και ¨ερικσεν¨ να είναι πολύ ψηλά στην λίστα αυτή.

Ιανουάριος

Διάγραμμα 4.4: Άθροισμα Ιανουαρίου

Τον Ιανουάριο οι μετρήσεις μας ξεκίνησαν από την 9η του μηνός. Όντας στην καρδιά της καραντίνας η λέξη ¨κρουσματα¨ είναι ψηλά στην λίστα. Επίσης και η λέξη ¨μπεκατωρου¨ δείχνει πως οι χρήστες ενδιαφέρθηκαν για το ζήτημα που έφερε στα φώτα της δημοσιότητας η Σοφία Μπεκατώρου σχετικά καταγγελίες για σεξουαλική παρενόχληση που είχε υποστεί στο παρελθόν.

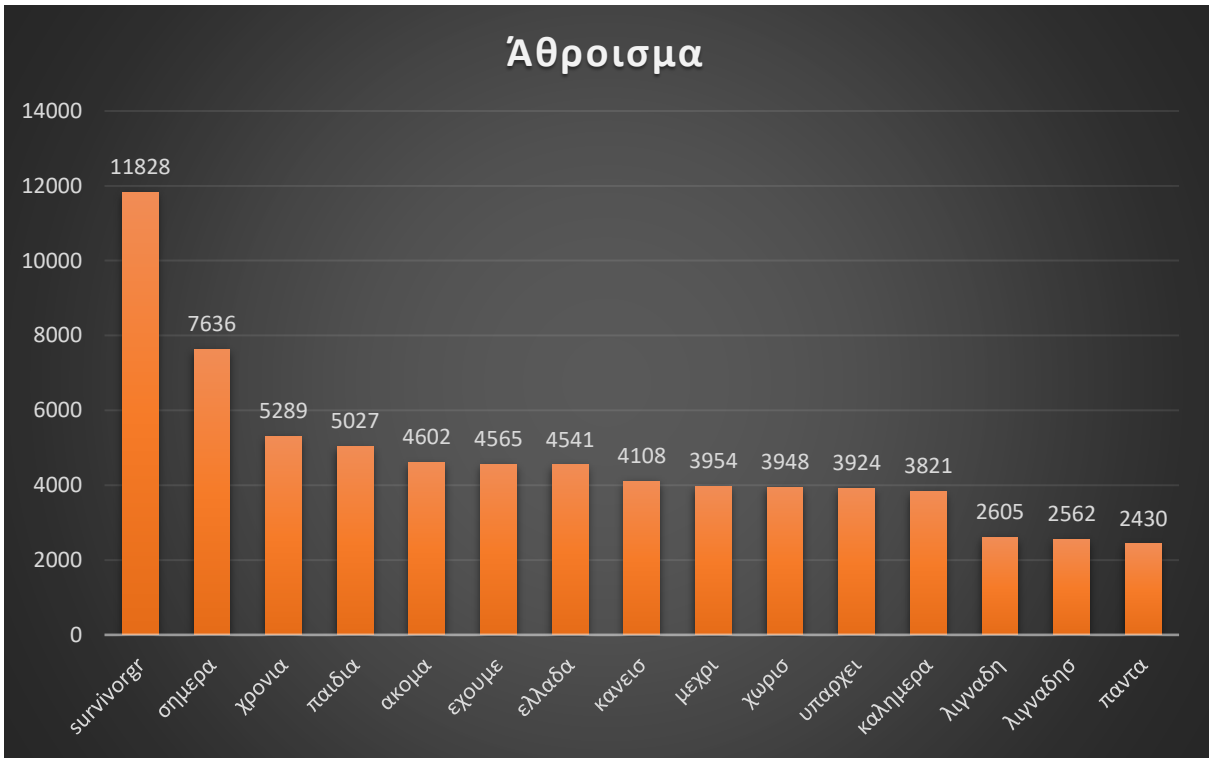
Διάγραμμα 4.5: Μέσος Όρος Ιανουαρίου



Κάποιες λέξεις που μπορούμε να διακρίνουμε εδώ είναι οι λέξεις ¨μπαιντεν¨ και ¨τραμπ¨. Το να εμφανιστούν οι λέξεις αυτές εδώ είναι αρκετά φυσιολογικό. Λίγες μέρες πριν ξεκινήσει η καταμέτρηση μας υπήρξε το περιστατικό με την είσοδο διαδηλωτών υποστηρικτών του Τραμπ στο Καπιτώλιο της Ουάσιγκτον καθώς επίσης η 21η Ιανουαρίου ήταν η ήμερα ορκωμοσίας του Τζο Μπάιντεν ως 47ος πρόεδρος των ΗΠΑ. Επίσης η λέξη ¨διερευνητικεσ¨ αναφέρεται στην έναρξη των διερευνητικών επαφών Ελλάδας και Τουρκίας που έγινε στις 25 Ιανουαρίου.

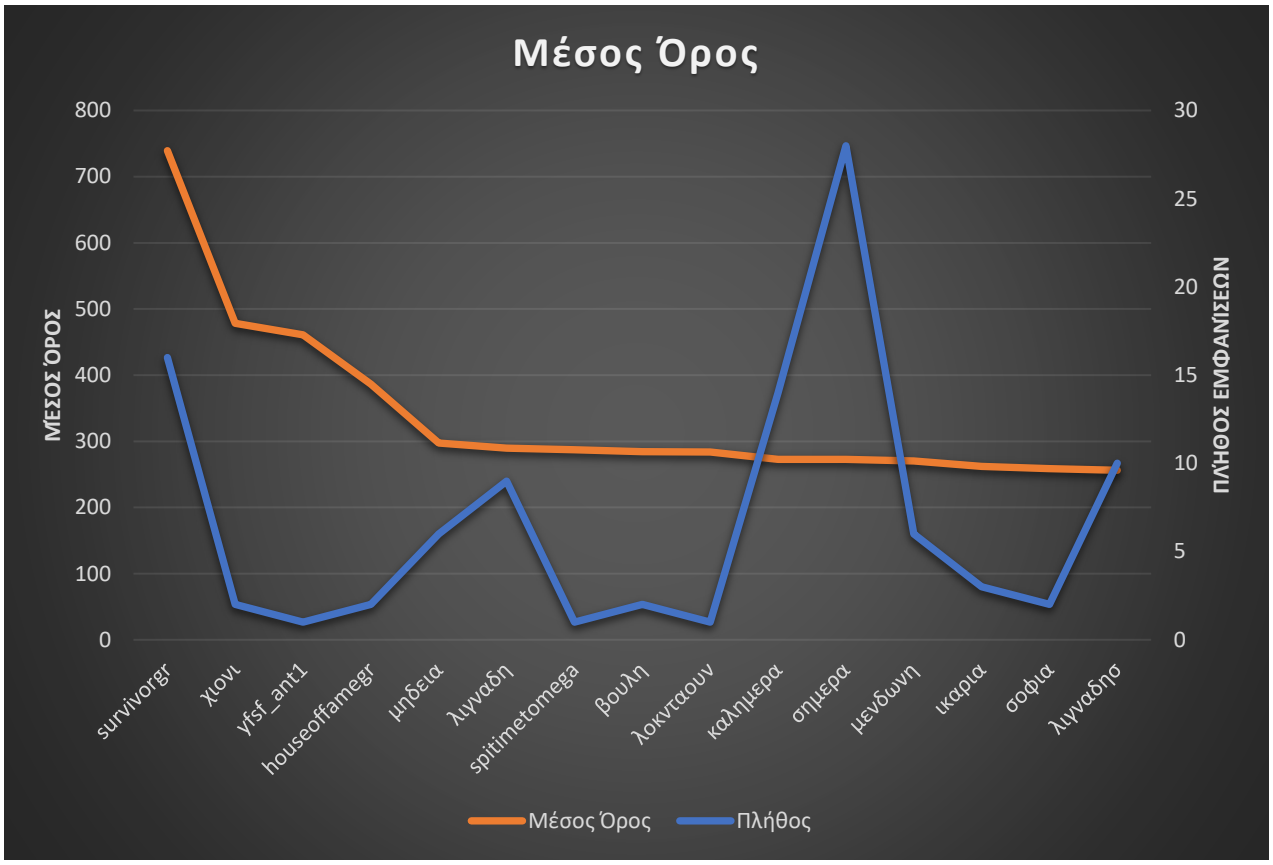
Φεβρουάριος

Διάγραμμα 4.6: Άθροισμα Φεβρουαρίου



Τον Φεβρουάριο βλέπουμε πως πέρα από τις κοινές λέξεις που έχουμε ήδη συναντήσει υπάρχουν και οι λέξεις ¨λιγναδη¨ και ¨λιγναδης¨ που προφανώς σχετίζονται με τις κατηγορίες που δημοσιεύτηκαν εκείνο το χρονικό διάστημα.

Διάγραμμα 4.7: Μέσος Όρος Φεβρουαρίου



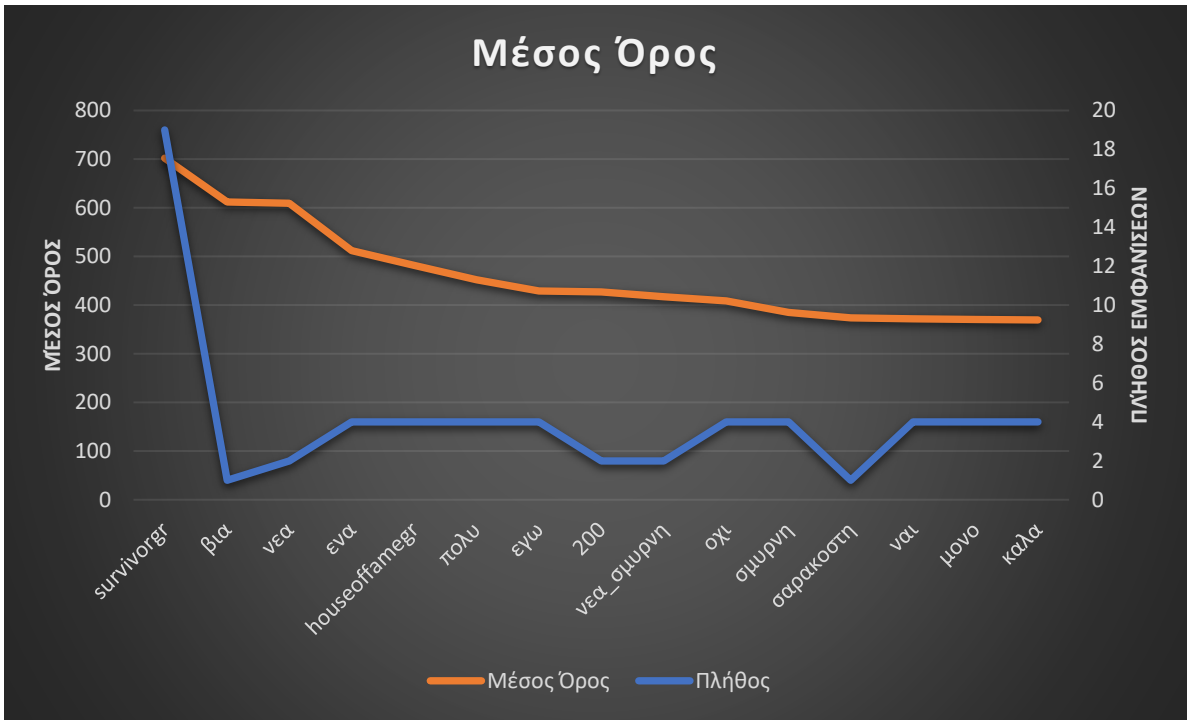
Στο γράφημα του Μέσου Όρου του Μαρτίου βλέπουμε την λέξη ¨χιονι¨ σε ιδιαίτερα υψηλή θέση με πολύ λίγες ημέρες εμφάνισης καθώς επίσης και την λέξη ¨μηδεια¨ που ήταν το όνομα που δόθηκε στην κακοκαιρία που χτύπησε την Αττική στις 15 και 16 Φεβρουαρίου.

Μάρτιος

Διάγραμμα 4.8: Άθροισμα Μαρτίου

Για τον Μάρτιο παρατηρούμε πως η λέξη ¨χρονια¨ που γενικά εμφανίζεται με συχνότητα περίπου 4000-4500 ανά μήνα εδώ έχει εμφανιστεί πάνω από 7800 φορές. Η ερμηνεία εδώ είναι πως ο Μάρτιος είναι ο μήνας της επετείου των διακοσίων χρόνων από την ελληνική επανάσταση και τις ημέρες εκείνες η φράση ¨Χρόνια πολλά Ελλάδα¨ (ή παρόμοιες) γράφτηκε πολύ συχνά από τους χρήστες.

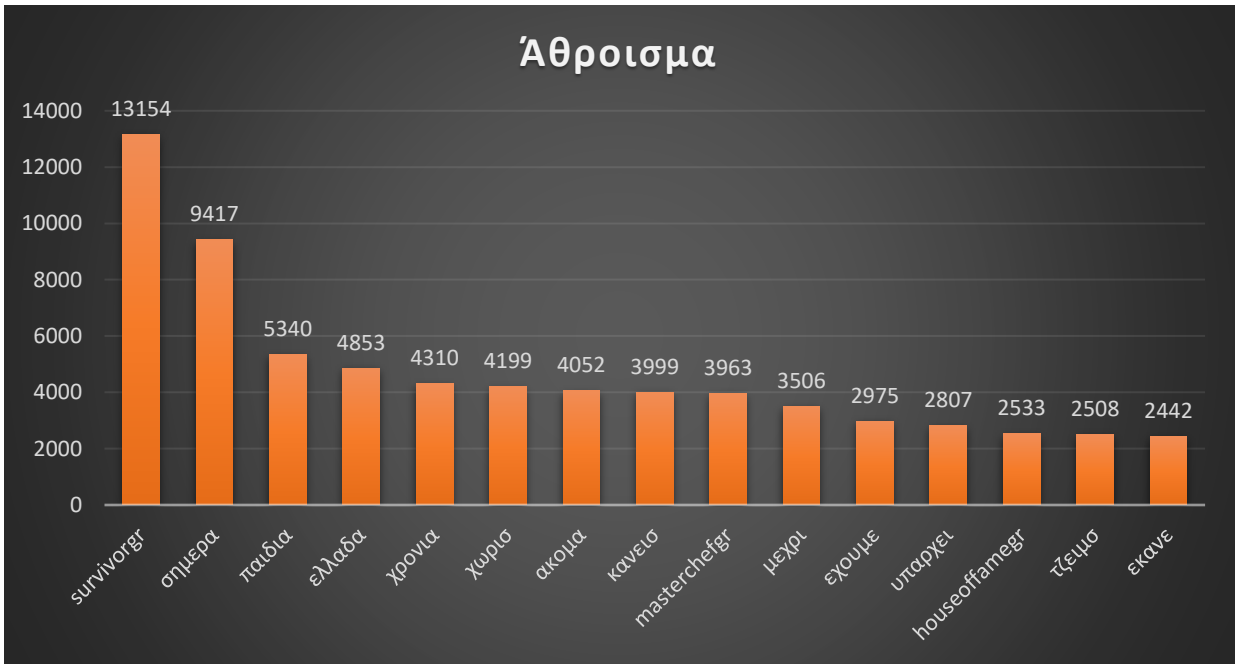
Διάγραμμα 4.9: Μέσος Όρος Μαρτίου



Αντίστοιχα στο γράφημα του Μέσου Όρου βλέπουμε την λέξη ¨200¨ να κάνει την εμφάνισης της, για τον ίδιο λόγο που αναφέραμε παραπάνω. Ακόμα βλέπουμε λέξεις όπως ¨νεα¨ , ¨νεα_σμυρνη¨ και ¨σμυρνη¨ να έχουν μεγάλη συχνότητα για μικρό χρονικό διάστημα κάτι το οποίο οφείλεται στα άσχημα γεγονότα που έλαβαν χώρα στην συγκεκριμένη περιοχή από 7 έως τις 10 Μαρτίου.

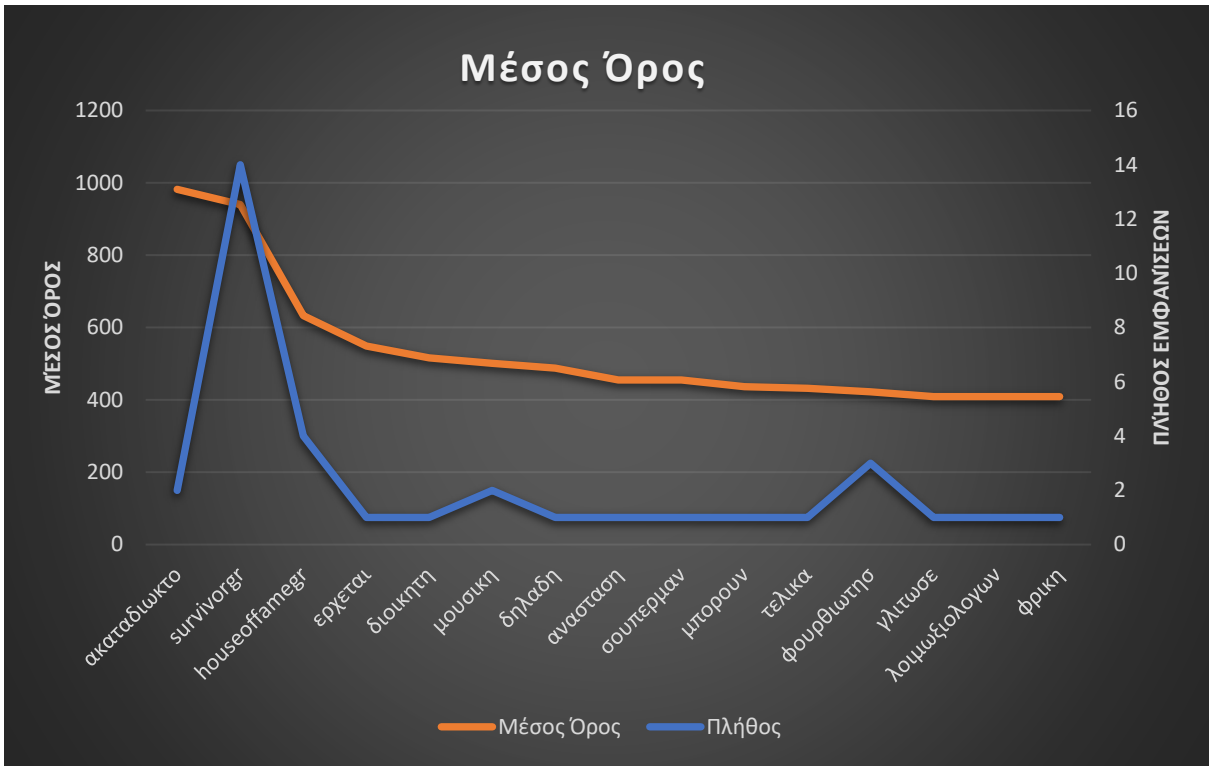
Απρίλιος

Διάγραμμα 4.10: Άθροισμα Απριλίου



Οι λέξεις που είχαν την μεγαλύτερη συχνότητα για τον Απρίλιο είναι πολύ κοντά στις λέξεις που συναντήσαμε στο συνολικό γράφημα.

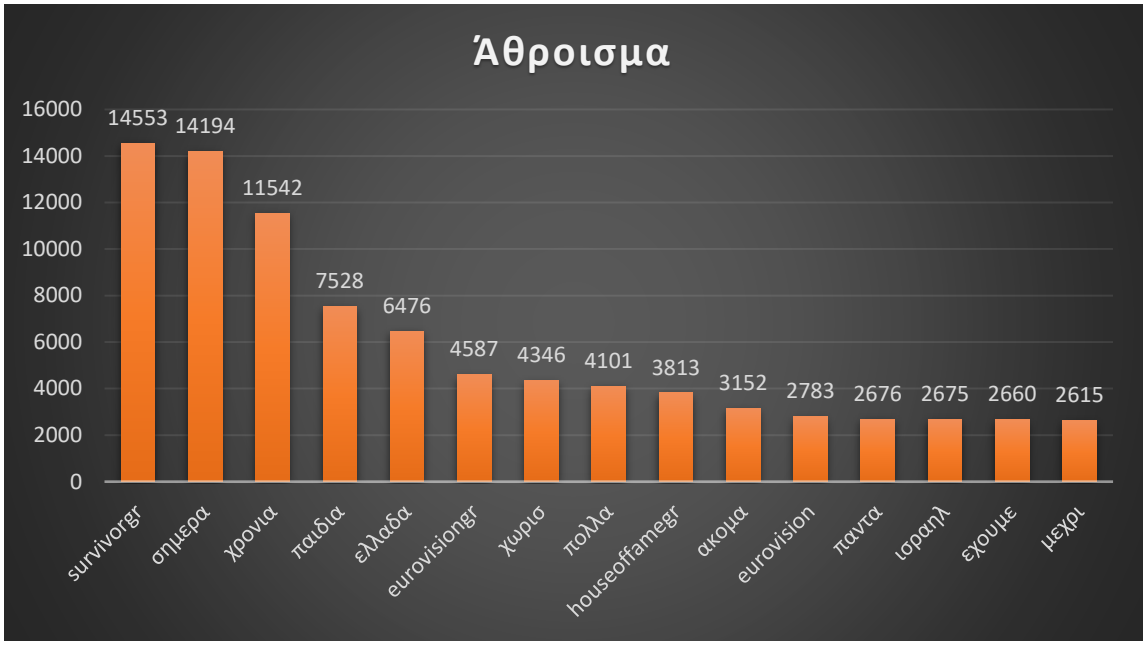
Διάγραμμα 4.11: Μέσος Όρος Απριλίου



Στο κομμάτι του Μέσου Όρου βλέπουμε λέξεις όπως ¨ακαταδιωκτο¨ και ¨λοιμωξιολογων¨ καθώς στις 22 Απριλίου εγκρίθηκε η αντίστοιχη τροπολογία.

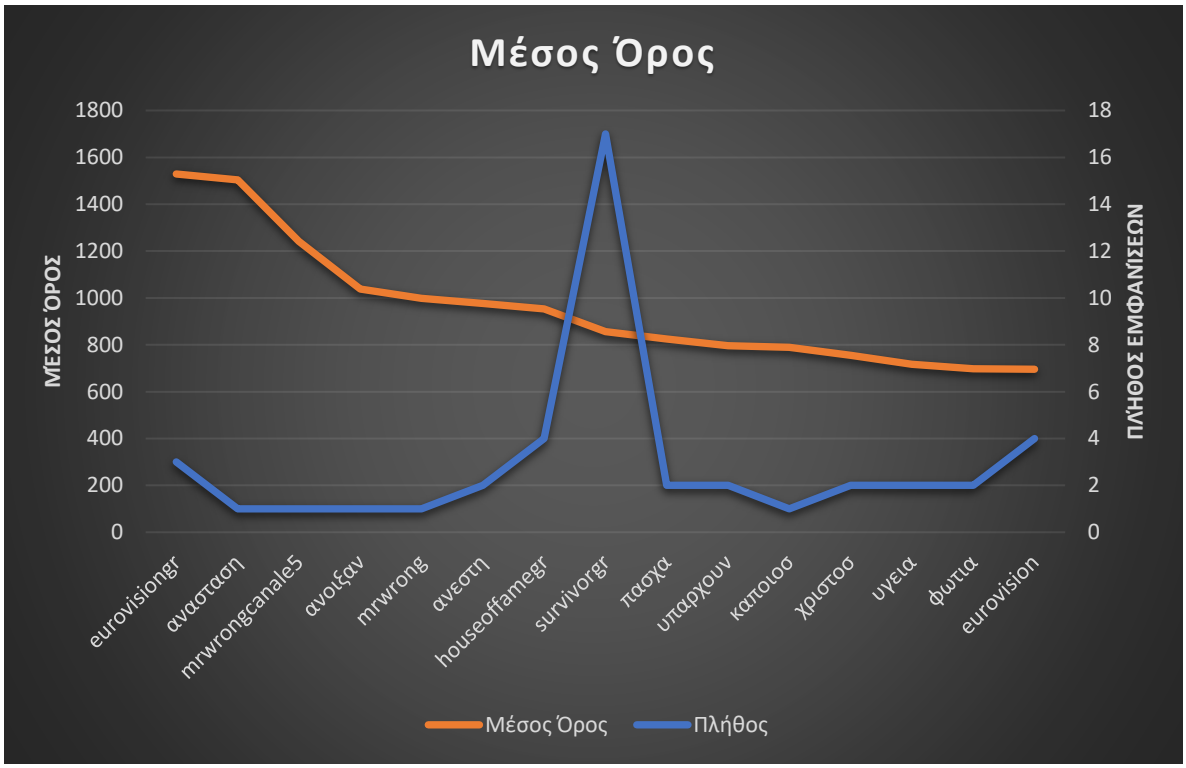
Μάιος

Διάγραμμα 4.12: Άθροισμα Μαΐου



Τον Μάιο βλέπουμε πάλι τις λέξεις ¨χρονια¨ και ¨πολλα¨ να έχουμε υψηλή συχνότητα. Αυτό οφείλεται στο ότι η γιορτή του Πάσχα ήταν στις 2 Μαΐου. Επίσης βλέπουμε την λέξη ¨ισραηλ¨ να εμφανίζεται. Ο λόγος είναι η νέα διαμάχη που ξέσπασε μεταξύ του Ισραήλ και της Παλαιστίνης στα μέσα του μήνα.

Διάγραμμα 4.13: Μέσος Όρος Μαΐου



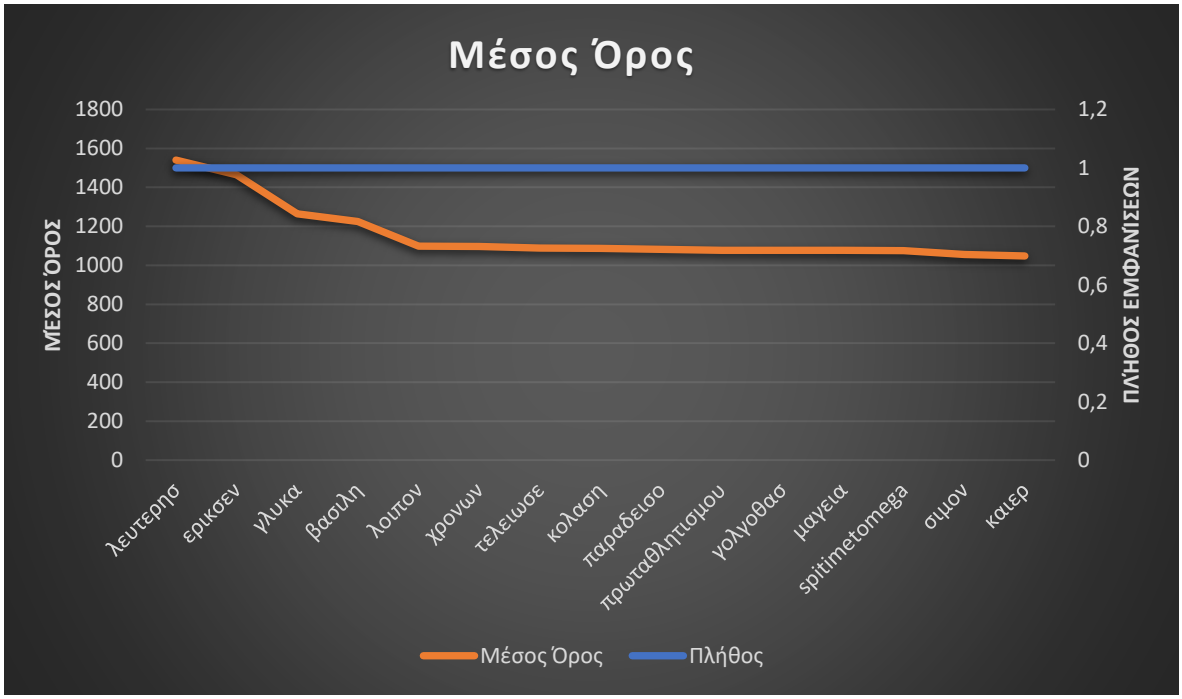
Εδώ παρατηρούμε ότι έχουμε μεγάλη εμφάνιση των λέξεων ¨eurovisiongr¨και ¨eurovision¨ καθώς ο διαγωνισμός έλαβε χώρα στις 22 Μαΐου. Επίσης οι λέξεις που σχετίζονται με το Πάσχα εμφανίζονται και πάλι. Τέλος η λέξη ¨φωτια¨ υπάρχει εδώ διότι στις 20 Μαΐου είχε ξεσπάσει ισχυρή πυρκαγιά στην περιοχή του Σχίνου.

Ιούνιος

Διάγραμμα 4.14: Άθροισμα Ιουνίου

Οι περισσότερες λέξεις του Ιουνίου είναι αρκετά συχνές. Εξαίρεση βέβαια αποτελεί η λέξη ¨καρολαιν¨ η οποία σχετίζεται με τα τραγικά γεγονότα που έλαβαν χώρα στα Γλυκά Νερά καθώς και τις εξελίξεις που ακολούθησαν.

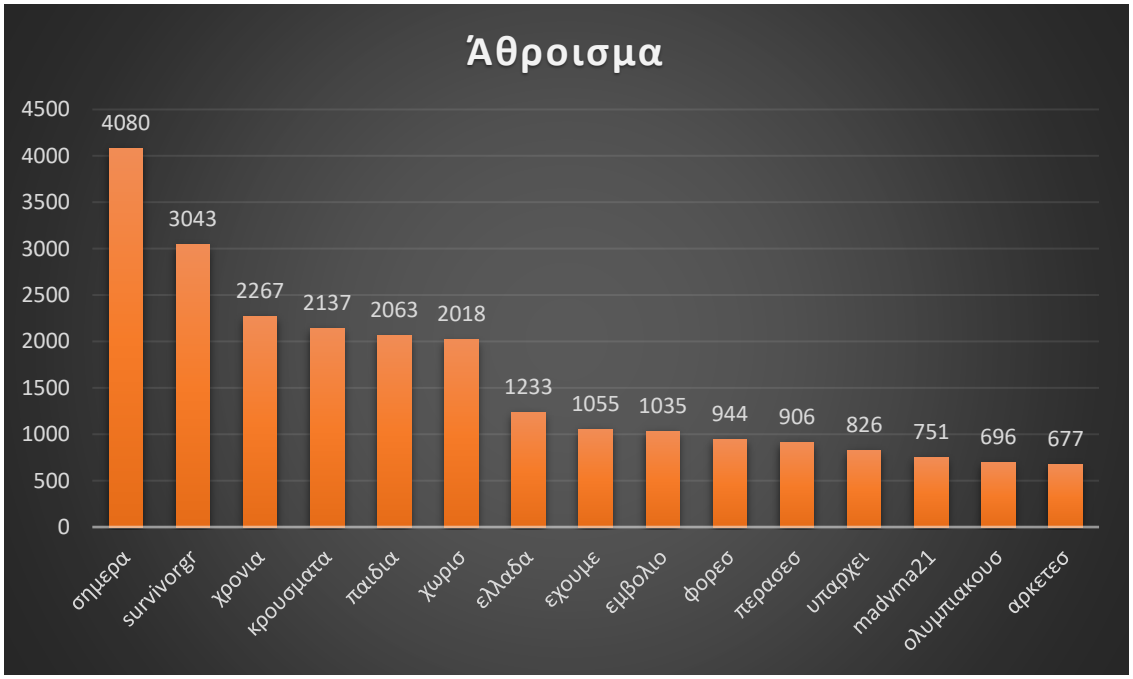
Διάγραμμα 4.15: Μέσος Όρος Ιουνίου



Στον Μέσο Όρο βλέπουμε τις λέξη ¨ερικσεν¨ για την οποία μιλήσαμε.

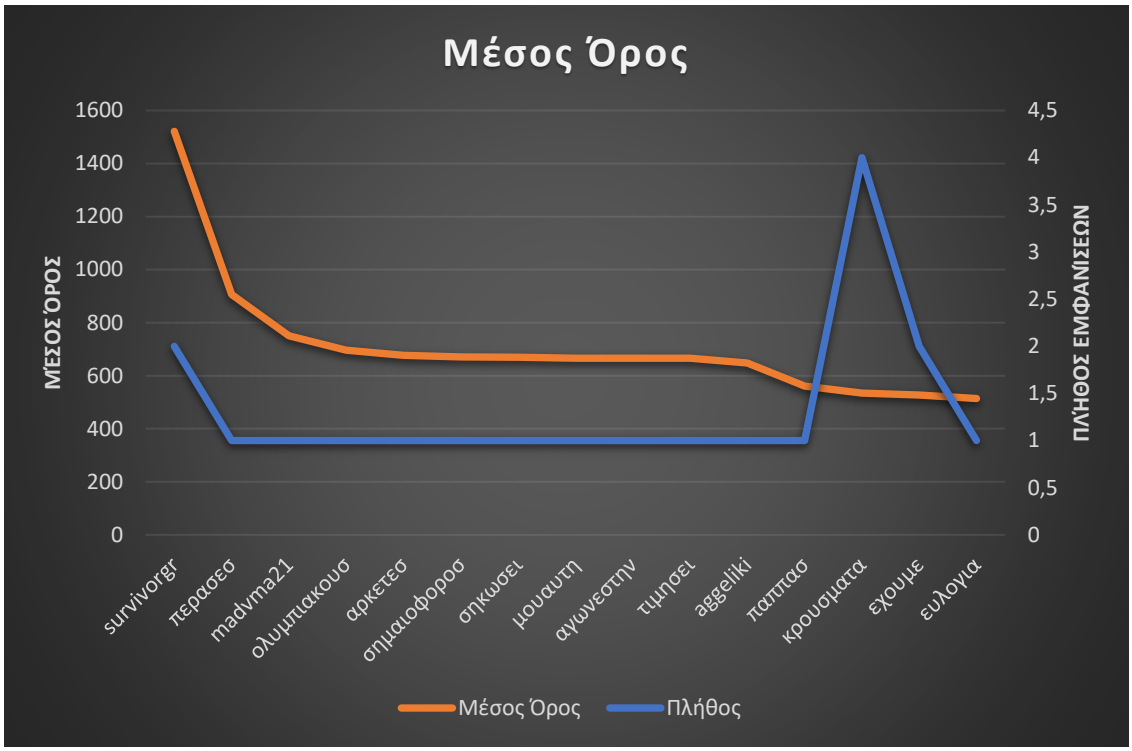
Ιούλιος

Διάγραμμα 4.16: Άθροισμα Ιουλίου



Τον Ιούλιο οι ημέρες εξόρυξης των δεδομένων ήταν λίγες καθώς η τελευταία μέρα συλλογής ήταν η 9η Ιουλίου. Αυτές τις πρώτες ημέρες βλέπουμε την λέξη ¨εμβολιο¨ να εμφανίζεται καθώς η συζήτηση σχετικά με τους εμβολιασμούς εντείνεται.

Διάγραμμα 4.17: Μέσος Όρος Ιουλίου



Στις αρχές Ιουλίου μια λέξη που ξεχωρίζει είναι η λέξη ¨περασεσ¨ καθώς και η λέξη ¨ολυμπιακουσ¨ καθώς πλησιάζουμε στην έναρξη των αγώνων.

4.4 Μελλοντική Έρευνα

Στην εργασία αυτή μελετήσαμε την συχνότητα με την οποία εμφανίζονται κάποιες λέξεις στο Twitter. Το πεδίο είναι ακόμα ανοιχτό για πολλές παραπάνω έρευνες. Κάποιες ιδέες για μελλοντική ενασχόληση είναι οι εξής:

- Εύρεση σύνδεσης μεταξύ των λέξεων παράλληλα με την συχνότητα εμφάνισης.
- Προσθήκη και των emojis ή των url στην ανάλυση.
- Χρήση μηχανικής μάθησης για την ανάλυση συναισθήματος.
- Δημιουργία Ελληνικών βιβλιοθηκών για την καλύτερη ανάλυση στην Ελληνική γλώσσα.
- Διεύρυνση του μοντέλου χρησιμοποιώντας και τονισμένους χαρακτήρες.

- Μεγέθυνση του χρονικού πλαισίου και του ημερήσιου όγκου συλλογής δεδομένων με σκοπό την βαθύτερη ανάλυση.
- Μελέτη διαφορετικών τρόπων ταιριάσματος νημάτων

Βιβλιογραφία

- Peter Jackson and Isabelle Moulinier, Natural Language Processing for Online Applications, 2007, John Benjamins Publishing Company
- Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, 1999, MIT Press, Cambridge
- Wes McKinney, Python for Data Analysis, 2011, O'Reilly
- Αβούρης Νικόλαος, Κουκιάς Μιχαήλ, Παλιουράς Βασίλειος, Σγάρμπας Κυριάκος, Μια εισαγωγή στους υπολογιστές με την γλώσσα Python, 2013, Πανεπιστήμιο Πατρών

Διαδίκτυο:

- https://en.wikipedia.org/wiki/Twitter#2007%E2%80%932010
- https://www.digitalocean.com/community/tutorials/an-introduction-to-oauth-2
- https://en.wikipedia.org/wiki/Natural_language_processing
- https://backlinko.com/twitter-users
- https://www.python.org/
- https://en.wikipedia.org/wiki/Hamming_distance
- https://www.nltk.org/
- https://datareportal.com/reports/digital-2021-greece
- https://en.wikipedia.org/wiki/Data_analysis

Παράρτημα Α

Παρακάτω παρατίθεται ο κώδικας της Python:

Οι γραμμές που ξεκινούν με # είναι σχόλια πάνω στον κώδικά μας.

Συλλογή των Tweets:

```
#καλούμε τις κατάλληλες βιβλιοθήκες
import tweepy
import csv

#κωδικοί για είσοδο στο API
consumer_key = "**************************"
consumer_secret = "**************************"
```

```
access_key = "*****************************"
access_secret = "*************************"

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)

api = tweepy.API(auth, wait_on_rate_limit=True)

#γεωγραφικά και χρονικά κριτήρια επιλογής
search = 'geocode:"37.99231,23.72062,25km" since:2021-7-9 until:2021-
7-10'
#αφαίρεση των retweets
new_search = search + "-filter:retweets"

csvarxeio = open('tweets.csv', 'a')
csvwriter = csv.writer(csvarxeio)

#αναζήτηση με τα παραπάνω κριτήρια και ακόμα κριτήριο γλώσσας
for tweet in tweepy.Cursor(api.search, q=new_search,
result_type='recent', count=100, lang="el", since_id=0).items():
    #αποθήκευση των Tweets με κωδικοποίηση UTF-8
    csvwriter.writerow([tweet.created_at, tweet.text.encode('utf-8'),
tweet.user.screen_name.encode('utf-8')])
```

Επεξεργασία των Tweets:

```
#καλούμε τις κατάλληλες βιβλιοθήκες
import ast
import collections
import csv
import re
import nltk
import pandas as pd
from nltk.corpus import stopwords
import itertools

#καλούμε τις λέξεις διακοπής
stop_words = stopwords.words('greek')

#συνάρτηση που αφαιρεί τα url
def remove_url(txt):
    return " ".join(re.sub("(\w+:\/\/\S+)", "", txt).split())

#συνάρτηση που αποκωδικοποιεί τα κείμενα
def parse_bytes(field):
    result = field
    try:
        result = ast.literal_eval(field)
    finally:
        return result.decode() if isinstance(result, bytes) else
field

#συνάρτηση που καλεί αρχείο και το αποκωδικοποιεί
def my_csv_reader(filename, /, **kwargs):
    with open(filename, 'rt', newline='') as file:
```

```
        for row in csv.reader(file, **kwargs):
            yield [parse_bytes(field) for field in row]

#dictionaty για να αντικαταστήσουμε τα τονισμένα γράμματα με άτονα
xwris_tonous = {'ά': 'α', 'έ': 'ε', 'ί': 'ι', 'ό': 'ο', 'ύ': 'υ',
'ή': 'η', 'ώ': 'ω'}

#συνάρτηση που αφαιρεί όλα τα emojis
def remove_emojis(data):
    emoj = re.compile("["
                      u"\U0001F600-\U0001F64F"  # emoticons
                      u"\U0001F300-\U0001F5FF"  # symbols &
pictographs
                      u"\U0001F680-\U0001F6FF"  # transport & map
symbols
                      u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                      u"\U00002500-\U00002BEF"  # chinese char
                      u"\U00002702-\U000027B0"
                      u"\U00002702-\U000027B0"
                      u"\U000024C2-\U0001F251"
                      u"\U0001f926-\U0001f937"
                      u"\U00010000-\U0010ffff"
                      u"\u2640-\u2642"
                      u"\u2600-\u2B55"
                      u"\u200d"
                      u"\u23cf"
                      u"\u23e9"
                      u"\u231a"
                      u"\ufe0f"  # dingbats
                      u"\u3030"
                      "]+", re.UNICODE)
    return re.sub(emoj, '', data)

#καλούμε το αρχείο csv και το περνάμε από όλες τις συναρτήσεις
c = ' '
decoded = my_csv_reader('tweets.csv', delimiter=',')
for row in decoded:
    if len(row) == 3:
        m = remove_url(row[1])
        m = m.lower()
        m = m.translate(str.maketrans(xwris_tonous))
        m = re.sub('@[^\s]+', '', m)
        m = re.sub(';', '', m)
        m = re.sub('!', '', m)
        m = re.sub('rt', '', m)
        m = re.sub('"', '', m)
        m = re.sub(':', '', m)
        m = re.sub(',', '', m)
        m = re.sub('-', '', m)
        m = re.sub('»', '', m)
        m = re.sub('«', '', m)
        m = re.sub('ς', 'σ', m)
        m = remove_emojis(m)
        #ενώνουμε όλα τα Tweets της ημέρας σε ένα κείμενο
        c += m

#χωρίζουμε σε λίστα λέξεων
wordlist = nltk.word_tokenize(c)
print(len(wordlist))

k = []
```

```
for w in wordlist:
    if w not in stop_words:
        if w not in other_words:
            if len(w) > 4:
                k.append(w)

final = list(itertools.chain(k))

#βρίσκουμε συχνότητα εμφάνισης
counts = collections.Counter(final)

#μετατροπή σε Data Frame
top_20 = pd.DataFrame(counts.most_common(20), columns=['Λέξη',
'21.7.9'])

print(top_20)

#αποθήκευση του Data Frame σε excel
top_20.to_excel(r'C:\Users\admin\Desktop\thesis\Thesis
Project\top_20\21.7.9.xlsx', index=False, header=True)
```

Παράρτημα Β

Greek Stop Words:

αλλα
αν
αντι
απο
αυτα
αυτεσ
αυτη
αυτο
αυτοι
αυτοσ
αυτουσ
αυτων
αἱ
αἳ
αἵ
αὐτόσ
αὐτὸς
αὗ
γάρ
γα
γα^
γε
για
γοῦν
γὰρ
δ'
δέ
δή
δαί
δαίσ
δαὶ
δαὶς
δε
δεν
δι'
διά
διὰ
δὲ
δὴ
δ’
εαν
ειμαι
ειμαστε
ειναι
εισαι
ειστε
εκεινα
εκεινεσ

εκεινη
εκεινο
εκεινοι
εκεινοσ
εκεινουσ
εκεινων
ενω
επ
επι
εἰ
εἰμί
εἰμὶ
εἰς
εἰσ
εἶ
εἶμι
εἶτε
η
θα
ισωσ
κ
καί
καίτοι
καθ
και
κατ
κατά
κατα
κατὰ
καὶ
κι
κἀν
κἂν
μέν
μή
μήτε
μα
με
μεθ
μετ
μετά
μετα
μετὰ
μη
μην
μέν
μὲν
μὴ
μὴν
να
ο
οι
ομωσ
οπωσ
οσο
οτι
οἱ
οἶ
οἷς
οὐ
οὐδ
οὐδέ
οὐδείσ
οὐδεὶς
οὐδὲ
οὐδὲν
οὐκ
οὐχ
οὐχὶ
οὗς
οὔτε
οὕτω
οὕτως
οὕτωσ
οὖν
οὗ
οὗτος
οὗτοσ
παρ
παρά
παρα
παρὰ
περί
περὶ
ποια
ποιεσ
ποιο
ποιοι
ποιοσ
ποιουσ
ποιων
ποτε
που
ποῦ
προ
προσ
πρόσ
πρὸ
πρὸς
πως
πωσ
σε

στη
στην
στο
στον
σόσ
σύ
σύν
σὸς
σὺ
σὺν
τά
τήν
τί
τίς
τίσ
τα
ταῖς
τε
την
τησ
τι
τινα
τις
τισ
το
τοί
τοι
τοιοῦτος
τοιοῦτοσ
τον
τοτε
του
τούσ
τοὺς
τοῖς
τοῦ
των
τό
τόν
τότε
τὰ
τὰς
τὴν
τὸ
τὸν
τῆς
τῆσ
τῇ
τῶν
τῷ
ωσ
ἀλλ'
ἀλλά
ἀλλὰ
ἀλλ’
ἀπ
ἀπό
ἀπὸ
ἀφ
ἄν
ἃ
ἄλλος
ἄλλοσ
ἄν
ἄρα
ἅμα
ἐάν
ἐγώ
ἐγὼ
ἐκ
ἐμόσ
ἐμὸς
ἐν
ἐξ
ἐπί
ἐπεὶ
ἐπὶ
ἐστι
ἐφ
ἐὰν
ἑαυτοῦ
ἔτι
ἡ
ἢ
ἣ
ἦ
ἧ
ἧς
ἵνα
ὁ
ὃ
ὃν
ὃς
ὅ
ὅδε
ὅθεν
ὅπερ
ὅς
ὅσ
ὅστις
ὅστισ
ὅτε

ὅτι
ὑμόσ
ὑπ
ὑπέρ
ὑπό
ὑπὲρ
ὑπὸ
ὡς
ὡσ
ὣς
ὥστε
ὦ
ὧ