ΕΘΝΙΚΟΝ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟΝ ΠΑΝΕΠΙΣΤΗΜΙΟΝ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ & ΠΟΛΙΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΚΟΙΝΩΝΙΑΣ & ΜΕΣΩΝ ΜΑΖΙΚΗΣ ΕΝΗΜΕΡΩΣΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ «ΕΠΙΚΟΙΝΩΝΙΑ ΚΑΙ ΤΑ ΜΕΣΑ ΜΑΖΙΚΗΣ ΕΝΗΜΕΡΩΣΗΣ»

ΚΑΤΕΥΘΥΝΣΗ: ΨΗΦΙΑΚΑ ΜΕΣΑ & ΠΕΡΙΒΑΛΛΟΝΤΑ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ

Θέμα διατριβής Ανάλυση Συναισθήματος και Εξόρυξη Κειμένου σε αξιολογήσεις στο TripAdvisor

ΠΑΠΑΔΟΠΟΥΛΟΥ ΕΥΑΓΓΕΛΙΑ

(Α.Μ.: 9983201636053)

Διπλωματική εργασία που κατατίθεται ως μέρος των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών στην Επικοινωνία και τα Μέσα Μαζικής Ενημέρωσης

Επιβλέπων Καθηγητής:

Μουρλάς Κωνσταντίνος

ΑΘΗΝΑ, ΑΥΓΟΥΣΤΟΣ 2018

Τριμελής Εξεταστική Επιτροπή

Υπογραφή	**Υπογραφή**	**Υπογραφή**
Ονοματεπώνυμο	**Ονοματεπώνυμο**	**Ονοματεπώνυμο**

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή μου κον. Μουρλά Κωνσταντίνο για την εμπιστοσύνη και την δυνατότητα που μου δόθηκε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα αλλά και για την πολύπλευρη υποστήριξη που μου παρείχε, καθ' όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας. Έπειτα, ευχαριστώ ιδιαίτερα την διδάκτωρ Σωτηράκου Κατερίνα για τις συμβουλές και την καθοδήγησή της, που κρίθηκαν πολύτιμες για την ολοκλήρωση αυτής της διπλωματικής. Τέλος, δεν μπορώ να μην ευχαριστήσω την οικογένεια μου και τους φίλους μου για τη συμπαράσταση που επέδειξαν όλο αυτό το διάστημα και κυρίως την φίλη μου Κωσταντίνα Δρίτσα γιατί χωρίς την βοήθεια της, η παρούσα διπλωματική δεν θα είχε επιτευχθεί.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ

ABSTRACT

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΓΛΩΣΣΑΡΙ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

1. ΕΙΣΑΓΩΓΗ 14

2.ΚΡΙΤΙΚΗ ΕΠΙΣΚΟΠΗΣΗ ΕΡΕΥΝΗΤΙΚΩΝ ΠΡΟΣΕΓΓΙΣΕΩΝ……..... 19

3. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΟΡΙΣΜΟΙ 21

3.1.ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ (SENTIMENT ANALYSIS)…..... 23

3.2.1.ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ..... 26

3.2. ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ (TEXT MINING) 30

3.2.1 ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΣΘΗΜΑΤΟΣ…..... 32

3.2.2. ΠΕΔΙΑ ΕΞΟΡΥΞΗΣ ΚΕΙΜΕΝΟΥ……..... 33

3.3 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ: TRIPADVISOR ……36

4.**ΕΡΓΑΛΕΙΑ ΕΡΕΥΝΑΣ..... 39**

4.1. PYTHON……..... 39

4.2. SELENIUM WEB DRIVER 39

4.3. NRC WORD EMOTION ASSOCIATION LEXICON..... 39

4.4. PANDAS LIBRARY..... 40

5. ΜΕΘΟΔΟΛΟΓΙΑ…… 41

5.1. ΑΝΑΚΤΗΣΗ ΔΕΔΟΜΕΝΩΝ..... 43

5.2.ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ..... 47

5.3. ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ..... 50

6. ΕΥΡΗΜΑΤΑ…...…….. 55

6.1 ΑΚΡΟΠΟΛΗ…….….55

6.2 ΠΛΑΚΑ……..……...…… 64

6.3 ΠΑΡΘΕΝΩΝΑΣ……...……...……... 73

6.4 ΑΡΧΑΙΑ ΑΓΟΡΑ ΑΘΗΝΩΝ……...……...……..…80

6.5 ΜΟΥΣΕΙΟ ΑΚΡΟΠΟΛΗΣ..……..87

7. ΣΥΜΠΕΡΑΣΜΑΤΑ……..... 95

7.1 ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ……. 97

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΠΑΡΑΡΤΗΜΑΤΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1: Απεικόνιση Google Trends σχετικά με τον όρο Sentiment analysis

Εικόνα 2: Τεχνικές Ανάλυσης Συναισθήματος

Εικόνα 3: Πεδία Εξόρυξης Κειμένου

Εικόνα 4: Παράδειγμα Pandas DataFrame

Εικόνα 5: HTML TripAdvisor

Εικόνα 6: Αρχική μορφή κώδικα για την ανάκτηση των δεδομένων από το TripAdvisor

Εικόνα 7: Τελική μορφή κώδικα για την ανάκτηση των δεδομένων από το TripAdvisor

Εικόνα 8: Παράδειγμα χρήσης του μοντέλου TF-IDF

Εικόνα 9: Διάγραμμα διάχυσης πολικότητας για την Ακρόπολη

Εικόνα 10: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

Εικόνα 11: Περιοχές με τον μεγαλύτερο αριθμό εμφάνισης θετικών σχολίων

Εικόνα 12: Περιοχές με τον μεγαλύτερο αριθμό εμφάνισης αρνητικών σχολίων

Εικόνα 13: Κατανομή θετικών σχολίων ανά χρονιά

Εικόνα 14: Κατανομή αρνητικών σχολίων ανά χρονιά

Εικόνα 15: Μέσος όρος θετικών αξιολογήσεων ανά μήνα

Εικόνα 16: Μέσος όρος αρνητικών αξιολογήσεων ανά μήνα

Εικόνα 17: Ομαδοποίηση λέξεων με τη μέθοδο K-means

Εικόνα 18: Διάγραμμα διάχυσης πολικότητας για την Πλάκα

Εικόνα 19: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

Εικόνα 20: Περιοχές με τον μεγαλύτερο αριθμό εμφάνισης θετικών σχολίων

Εικόνα 21: Περιοχές με τον μεγαλύτερο αριθμό εμφάνισης αρνητικών σχολίων

Εικόνα 22: Κατανομή θετικών σχολίων ανά χρονιά

Εικόνα 23: Κατανομή αρνητικών σχολίων ανά χρονιά

Εικόνα 24: Μέσος όρος θετικών αξιολογήσεων ανά μήνα

Εικόνα 25: Μέσος όρος αρνητικών αξιολογήσεων ανά μήνα

Εικόνα 26: Ομαδοποίηση λέξεων με τη μέθοδο K-means

Εικόνα 27: Διάγραμμα διάχυσης πολικότητας για την Παρθενώνα

Εικόνα 28: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

Εικόνα 29: Κατανομή θετικών σχολίων ανά χρονιά

Εικόνα 30: Κατανομή αρνητικών σχολίων ανά χρονιά

Εικόνα 31: Μέσος όρος θετικών αξιολογήσεων ανά μήνα

Εικόνα 32: Μέσος όρος αρνητικών αξιολογήσεων ανά μήνα

Εικόνα 33: Μέσος Όρος Συναισθημάτων για τον Παρθενώνα

Εικόνα 34: Ομαδοποίηση λέξεων με τη μέθοδο K-means

Εικόνα 35: Διάγραμμα διάχυσης πολικότητας για το Μουσείο της Ακρόπολης

Εικόνα 36: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

Εικόνα 37: Κατανομή θετικών σχολίων ανά χρονιά

Εικόνα 38: Κατανομή αρνητικών σχολίων ανά χρονιά

Εικόνα 39: Μέσος όρος θετικών αξιολογήσεων ανά μήνα

Εικόνα 40: Μέσος όρος αρνητικών αξιολογήσεων ανά μήνα

Εικόνα 41: Μέσος Όρος Συναισθημάτων για το Μουσείο της Ακρόπολης

Εικόνα 42: Ομαδοποίηση λέξεων με τη μέθοδο K-means

Εικόνα 43: Διάγραμμα διάχυσης πολικότητας για την Αρχαία Αγορά

Εικόνα 44: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

Εικόνα 45: Κατανομή θετικών σχολίων ανά χρονιά

Εικόνα 46: Κατανομή αρνητικών σχολίων ανά χρονιά

Εικόνα 47: Μέσος όρος θετικών αξιολογήσεων ανά μήνα

Εικόνα 48: Μέσος όρος αρνητικών αξιολογήσεων ανά μήνα

Εικόνα 49: Μέσος Όρος Συναισθημάτων για το Μουσείο της Ακρόπολης

Εικόνα 50: Ομαδοποίηση λέξεων με τη μέθοδο K-means

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1: Αριθμός αξιολογήσεων που συλλέχθηκαν

Πίνακας 2: Διάστημα συλλογής δεδομένων

Πίνακας 3: Δεδομένα Ανάλυσης Συναισθήματος Ακρόπολης

Πίνακας 4: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

Πίνακας 5: Μέσος Όρος Συναισθημάτων για Ακρόπολη

Πίνακας 6: Οι 10 πιο συχνές λέξεις για την Ακρόπολη

Πίνακας 7: Δεδομένα Ανάλυσης Συναισθήματος για την Πλάκα

Πίνακας 8: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

Πίνακας 9: Μέσος Όρος Συναισθημάτων για την Πλάκα

Πίνακας 10: Οι 10 πιο συχνές λέξεις για την Πλάκα

Πίνακας 11: Δεδομένα Ανάλυσης Συναισθήματος για τον Παρθενώνα

Πίνακας 12: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

Πίνακας 13: Οι 10 πιο συχνές λέξεις για τον Παρθενώνα

Πίνακας 14: Δεδομένα Ανάλυσης Συναισθήματος για το Μουσείο της Ακρόπολης

Πίνακας 15: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

Πίνακας 16: Οι 10 πιο συχνές λέξεις για το Μουσείο της Ακρόπολης

Πίνακας 17: Δεδομένα Ανάλυσης Συναισθήματος Αρχαίας Αγοράς Αθηνών

Πίνακας 18: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

Πίνακας 19: Οι 10 πιο συχνές λέξεις για την Αρχαία Αγορά Αθηνών

ΓΛΩΣΣΑΡΙ

BOW = Bag of Words

Classifier = Ταξινομητής

Cluster = Ομάδα

Feature = Χαρακτηριστικό

Information Retrieval = Ανάκτηση Δεδομένων

K-means Clustering = Ομαδοποίηση λέξεων με την μέθοδο k-means

Missing values = μεταβλητές που λείπουν

Natural Language Processing = Επεξεργασία Φυσικής Γλώσσας

Reviews = Αξιολογήσεις

Sentiment Analysis = Ανάλυση Συναισθήματος

Stemming = Αφαίρεση κατάληξης λέξεων

Stop words = Συχνά χρησιμοποιούμενες λέξεις όπως προθέσεις, σύνδεσμοι κλπ.

Text Mining = Εξόρυξη γνώμης

TF-IDF = Term Frequency – Inverse Document Frequency

Tokenization = Διανυσματοποίηση

Unigram = Λεκτική Μονάδα

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

CSV = Comma separated value

DM = Data mining

IR = Information Retrieval

NLP = Natural Language Processing

SA = Sentiment analysis

TF-IDF = Term Frequency – Inverse Document Frequency

TM = Text Mining

ΠΡΟΛΟΓΟΣ

Το αντικείμενο της παρούσας διπλωματικής εργασίας είναι η ανάλυση συναισθήματος σε δεδομένα της κοινωνικής πλατφόρμας TripAdvisor με την μέθοδο μη επιβλεπόμενης μάθησης βασισμένη σε λεξικό. Η ανάλυση συναισθήματος, που ονομάζεται επίσης εξόρυξη γνώμης (opinion mining), είναι ο τομέας της επιστήμης που αναλύει τις απόψεις, τα αισθήματα, τις εκτιμήσεις, τις στάσεις, τα συναισθήματα προς οντότητες και τα χαρακτηριστικά γνωρίσματά που εκφράζονται σε γραπτό κείμενο. Τα τελευταία χρόνια έχει παρουσιαστεί ραγδαία ανάπτυξη ως πεδίο έρευνας, εξαιτίας της μεγάλης επιρροής των κοινωνικών δικτύων και των πλατφόρμων κοινωνικής δικτύωσης στην καθημερινότητά μας, και λόγω του ότι παρέχει ένα αυτοματοποιημένο τρόπο για την ανάλυση της γραπτής πληροφορίας που αφθονεί στις διαδικτυακές πηγές.

Η αναγνώριση του συναισθήματος γίνεται σε δύο κατηγορίες, θετικό και αρνητικό συναίσθημα και για τις ανάγκες της εργασίας χρησιμοποιείται ένα σύνολο από περίπου 50.047 αξιολογήσεων από τους χρήστες του TripAdvisor σχετικά με πέντε βασικά σημεία τουρισμού της Αθήνας.

Το πρώτο μέρος της ανάλυσης αναφέρεται στην ανάκτηση των δεδομένων μέσω της χρήσης του scraping και αφού γίνει η προ-επεξεργασία των δεδομένων για όλους τους ειδικούς όρους που απαντώνται σε αυτά και μπορεί να δυσκολεύουν την ανάλυση τους, παράλληλα, εξετάζονται διάφοροι τρόποι εξαγωγής πληροφορίας από τα δεδομένα κειμένου και συγκεκριμένα η κλασσική μέθοδος Bag-of-Words με τις παραλλαγές term frequency και tf-idf (term frequency - inverse document frequency) και οι διανυσματικές αναπαραστάσεις λέξεων που καλούνται word vectors με την χρήση της γλώσσας python, της βιβλιοθήκης της pandas και του NRC-Emotion Lexicon. Οι παραπάνω ιδέες αξιολογούνται όλες στο σύνολο δεδομένων.

Η εργασία καταλήγει στο συμπέρασμα πως οι τεχνικές ανάλυσης συναισθήματος μέσω της χρήσης λεξικού ανταποκρίνονται πολύ καλά στο πρόβλημα παρέχοντας γρήγορες υλοποιήσεις και αξιόπιστες επιδόσεις.

Ακολουθώντας τις διαφορετικές μεθοδολογίες: (α) τη συχνότητα χρήσης μεμονωμένων λέξεων το σχήμα στάθμισης TF-IDF, (β) την εμφάνιση επιλεγμένων λέξεων πολικότητας με τη χρήση λεξικού καθώς και (γ) την τεχνική ομαδοποίησης

K-means Clustering, παράγονται σημαντικές πληροφορίες σχετικά με την άποψη των χρηστών για τα πέντε αυτά τουριστικά σημεία της Αθήνας.

Τα αποτελέσματα της μελέτης αυτής αποδεικνύουν ότι οι αξιολογήσεις των χρηστών, οι οποίες είναι γραμμένες στην αγγλική γλώσσα, είναι δυνατό να ταξινομηθούν, με βάση την συναισθηματική τους πολικότητα, με έναν πολύ αποτελεσματικό τρόπο, παράγοντας επαρκή συμπεράσματα.

Η έρευνα αυτή μπορεί να βοηθήσει στην ελαχιστοποίηση του χρόνου που απαιτείται στην αναζήτηση εξόρυξη πληροφορίας σχετικά με τα πέντε τουριστικά σημεία μελέτης καθώς επίσης να βοηθήσει στην ανάπτυξη του τουρισμού στην περιοχή της Αθήνας και της Ελλάδας γενικότερα.

Λέξεις Κλειδιά: Ανάλυση συναισθήματος, Εξόρυξη κειμένου, TripAdvisor, K-means Clustering

ABSTRACT

The subject of this thesis is the analysis of sentiment in data on the social platform of TripAdvisor with the non-supervised lexical based learning method. Sentiment analysis, also called opinion mining, is the field of science that analyzes opinions, feelings, assessments, attitudes, feelings to entities, and features that are expressed in written text. Recent years have shown rapid growth as a research field, due to the great influence of social networks and social media platforms in our everyday lives, and because it provides an automated way to analyze the written information that abounds in online sources.

Recognition of sentiment occurs in two categories, positive and negative emotions, and for the needs of work, a total of about 49,748 evaluations from the users of TripAdvisor on five key tourist spots in Athens are used.

The first part of the analysis refers to the data recovery through the use of scraping and after pre-processing the data for all the specific terms encountered in them and can make their analysis difficult, while at the same time various ways of extracting information from the data and in particular the classical Bag-of-Words method with the term frequency and tf-idf (term frequency - inverse document frequency) variants and the vector representations of words called word vectors using the python language, the pandas library and the use of NRC-Emotion Lexicon. The above ideas are all evaluated in the dataset.

The work concludes that speech-based emotional analysis techniques respond very well to the problem by providing quick implementations and reliable performance.

Following different methods: (a) the frequency of use of individual words using the TF-IDF method, (b) the appearance of selected polarity of words using a dictionary, and (c) technical K-means Clustering, producing significant information on the users' view of these five tourist spots in Athens.

The results of our study show that we can classify reviews written in English based on their emotional polarity in a very effective way. This research can help minimize the time needed to search for relevant information as well as help in the development of tourism in Athens and Greece in general.

1. ΕΙΣΑΓΩΓΗ

Η ταχεία ανάπτυξη των εφαρμογών του Διαδικτύου στον τουρισμό έχει οδηγήσει σε τεράστιο αριθμό προσωπικών αξιολογήσεων (reviews) για πληροφορίες σχετικά με τα ταξίδια στον Παγκόσμιο Ιστό (Web).

Αυτές οι αξιολογήσεις μπορούν να εμφανιστούν σε διαφορετικές μορφές όπως οι κοινωνικές πλατφόρμες δικτύωσης, blogs, Wiki ή φόρουμ και τα μέσα κοινωνικής δικτύωσης (social media). Το πιο σημαντικό, σε αυτές τις αξιολογήσεις είναι ότι οι πληροφορίες αποτελούν πολύτιμους αρωγούς τόσο για τους ταξιδιώτες όσο και για τους επαγγελματίες για τις διάφορες διαδικασίες κατανόησης και προγραμματισμού.

Ο σχεδιασμός ταξιδιού και οι κρατήσεις ξενοδοχείων μέσω ιστοσελίδων έχει αποκτήσει μεγάλη εμπορική χρήση. Ο διαμοιρασμός στον ιστό έχει γίνει ένα σημαντικό εργαλείο για να εκφράσει τις σκέψεις των πελατών σχετικά με ένα συγκεκριμένο προϊόν ή υπηρεσία. Σύμφωνα με μελέτη του Ευρωβαρομέτρου το 53% των Ευρωπαίων έκλεισαν τις διακοπές τους μέσω του διαδικτύου (2011)[1]. Όπως αναφέρουν οι Elango kai Narayanan (2014) τα τελευταία χρόνια σημειώθηκε ραγδαία ανάπτυξη στο διαδίκτυο, ομάδων συζήτησης (forum) και ιστότοποι αξιολογήσεων (π.χ.www.tripadvisor.com) όπου κρίσιμο χαρακτηριστικό της αξιολόγησης ενός πελάτη είναι το δικό τους συναίσθημα ή η γενική γνώμη - για παράδειγμα αν η αξιολόγηση περιέχει λέξεις όπως "τέλειο", «καλό», «τρομερό» είναι πιθανότατα ένα θετικό σχόλιο. Αν οι κριτικές περιέχουν λέξεις όπως «κακό», «φτωχό», «απαίσιο», είναι πιθανώς αρνητική κριτική.

Ένα εγγενές πρόβλημα των συντριπτικών πληροφοριών στο Διαδίκτυο, ωστόσο, είναι η υπερφόρτωση πληροφοριών, καθώς οι χρήστες απλά δεν μπορούν να διαβάσουν όλες τις διαθέσιμες πληροφορίες. Διαφορετικά από τα ενημερωτικά έγγραφα, τα περισσότερα στοιχεία που μπορεί να απομονώσει κάποιος σε μια πλατφόρμα κοινωνικών μέσων είναι μικρός. Επομένως, πολλά μηνύματα δεν είναι ολοκληρωμένα και ουσιαστικά σημασιολογικά. Η δημοσίευση μηνυμάτων στις πλατφόρμες κοινωνικών μέσων είναι συνήθως θορυβώδης. Πολλά από αυτά περιέχουν ορολογίες και φράσεις που δεν έχουν συνάφεια ή αποτελούν σύντομα μηνύματα, με αποτέλεσμα να θεωρούνται ως θόρυβος, που προκαλούν τη δυσκολία ανίχνευσης και ανίχνευσης υπό-γεγονότων.

[1] http://europa.eu/rapid/press-release_MEMO-12-154_el.htm

Πολλά πεδία της επιστήμης των υπολογιστών έχουν ασχοληθεί τα τελευταία χρόνια, κυρίως λόγω της εξάπλωσης των κοινωνικών δικτύων και των πλατφόρμων κοινωνικής δικτύωσης, με την επεξεργασία φυσικής γλώσσας (natural language processing) και την ανάλυση του τεράστιου όγκου πληροφορίας κειμένου που υπάρχει ελεύθερη στο διαδίκτυο. Όλα αυτά τα μηνύματα, τα σχόλια, οι αξιολογήσεις/κριτικές μπορούν να δώσουν πληροφορίες και να λύσουν προβλήματα για πολλά πεδία, είτε σε ακαδημαϊκό επίπεδο είτε σε επιχειρησιακό. Τέτοιες προσεγγίσεις αποτελούν η ανάλυση συναισθήματος (sentiment analysis) και η εξόρυξη γνώμης (text mining), που αποτελούν τον κορμό αυτής της εργασίας.

Το μεγαλύτερο μέρος της έρευνας στην ανάλυση του συναισθήματος επικεντρώθηκε σε αγγλικά κείμενα που βρίσκονται διαθέσιμα στο διαδίκτυο και ειδικά σε κριτικές ξενοδοχείων, ταινιών και προϊόντων. Η βιβλιογραφία σχετικά με ανάλυση του συναισθήματος σε μνημεία και περιοχές πολιτισμικής σημασίας είναι όμως περιορισμένη.

Με την παρατήρηση αυτή και με δεδομένο ότι ο τουρισμός είναι μια πολύ δημοφιλής βιομηχανία στην Ελλάδα, αποφασίσαμε να εξετάσουμε τις κριτικές και αξιολογήσεις των χρηστών από την ταξιδιωτική πλατφόρμα TripAdvisor στην αγγλική γλώσσα αναφορικά με πέντε πολιτισμικά σημεία στην Αθήνα και συγκεκριμένα την Ακρόπολη, την Πλάκα, τον Παρθενώνα και το Μουσείο της Ακρόπολης και την Αρχαία αγορά των Αθηνών.

Στόχος της εργασίας είναι η ανάλυση συναισθήματος, δηλαδή η εύρεση της πολικότητας (θετικής ή αρνητικής) των αξιολογήσεων από τους χρήστες της πλατφόρμας TripAdvisor αναφορικά με τα σημεία μελέτης. Τι είδους πρόσημο (θετικό ή αρνητικό) έχει στο σύνολο της η άποψη των χρηστών και πως εντοπίζεται γεωγραφικά αλλά και χρονικά. Επιπλέον, θέλουμε να διερευνηθεί ποια είναι τα κοινά χαρακτηριστικά που εντοπίζονται, είτε αρνητικά είτε θετικά, σε αυτές τις αξιολογήσεις για τα πέντε τουριστικά αξιοθέατα που μελετούνται.

Για την εκπόνηση της παρούσας εργασίας πραγματοποιήθηκε η μέθοδος ανάλυσης συναισθήματος (Sentiment Analysis) με στόχο να εξαχθεί η πολικότητα του συναισθήματος του συνόλου των αξιολογήσεων και η τεχνική ομαδοποίησης K-Means (K-means Clustering) για την εξόρυξη κειμένου (Text Analysis), δηλαδή την συγκέντρωση των δεδομένων κειμένου σε ομάδες με νοηματική συνάφεια με στόχο

να εξαχθούν συμπεράσματα γύρω από τις θεματικές που εμφανίζονται ότι αναλύουν οι χρήστες μέσα στις αξιολογήσεις τους (reviews).

Πιο αναλυτικά στο πρώτο κεφάλαιο δίνεται μια μικρή εισαγωγή για την παρούσα μελέτη, στο δεύτερο κεφάλαιο μελετάται μέρος της βιβλιογραφικής ανασκόπησης παρόμοιων εργασιών σχετικά με την ανάλυση συναισθήματος και την εξόρυξη γνώμης, στις οποίες στηρίχθηκε η συγκεκριμένη εργασία. Στο τρίτο κεφάλαιο αναλύεται όλο το θεωρητικό υπόβαθρο που στηρίχθηκε το παρόν έργο, δηλαδή οι έννοιες της ανάλυσης συναισθήματος και της εξόρυξης γνώμης καθώς και οι τεχνικές τους, τα πεδία που εφαρμόζονται και περιορισμοί που συναντιούνται στην μελέτη και χρήση τους. Στο τέταρτο κεφάλαιο επεξηγούνται τα εργαλεία που χρησιμοποιήθηκαν για την πραγμάτωση του έργου, ενώ στο κεφάλαιο πέντε αναλύεται λεπτομερώς η μεθοδολογία της έρευνας αυτής, από την απόκτηση των δεδομένων μέχρι την εξαγωγή συμπερασμάτων. Στο κεφάλαιο έξι παρουσιάζονται αναλυτικά όλα τα ευρήματα για το κάθε τουριστικό σημείο ξεχωριστά, ενώ τέλος στο έβδομο κεφάλαιο αναφέρονται τα συμπεράσματα που εξήχθην από την συγκεκριμένη εργασία και μελλοντικές επεκτάσεις που μπορούν να εφαρμοστούν.

2. ΚΡΙΤΙΚΗ ΕΠΙΣΚΟΠΗΣΗ ΕΡΕΥΝΗΤΙΚΩΝ ΠΡΟΣΕΓΓΙΣΕΩΝ

Η ταξινόμηση ενός κειμενικού αρχείου σύμφωνα με το γενικό συναίσθημα του περιεχομένου του είναι ίσως το πλέον διαδεδομένο πεδίο μελέτης στην ακαδημαϊκή κοινότητα σήμερα. Αρκετές μελέτες ανάλυσης συναισθήματος έχουν ορίσει το TripAdvisor ως πηγή δεδομένων και πεδίο μελέτης στην διεθνή βιβλιογραφία.

Για παράδειγμα, οι ερευνητές Guzman and Maale (2014) έχουν συλλέξει αξιολογήσεις σχετικά με την εφαρμογή TripAdvisor στο Google Play Store για την εξαγωγή χαρακτηριστικών της εφαρμογής για να βοηθήσει τους προγραμματιστές στην βελτίωση της. Ο κύριος στόχος της προσέγγισής τους είναι να εντοπίσουν αυτόματα, τα χαρακτηριστικά της εφαρμογής που αναφέρονται στις αξιολογήσεις των χρηστών, καθώς και τα συναισθήματα και τις απόψεις που σχετίζονται με αυτά τα χαρακτηριστικά, μέσω της επεξεργασίας φυσικής γλώσσας με τη χρήση ενός λεξιλογικού εργαλείου εξαγωγής συναισθήματος, το SentiStrength (Thelwall, Buckley, Paltoglou, Cai, and Kappas, 2010) και τεχνικών εξόρυξης δεδομένων με τη χρήση του αλγορίθμου Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003).

Ο Gräbner (2012) ανέλυσε αξιολογήσεις ξενοδοχείων των χρηστών του TripAdvisor για την ταξινόμηση ”καλών” και ”κακών” απόψεων των πελατών τους, δημιουργώντας μια αξιόπιστη μέθοδο ταξινόμησης των αξιολογήσεων των χρηστών σε ένα ήδη υπάρχον κυρίαρχο σώμα κειμένου χρησιμοποιώντας μια μέθοδος ανάλυσης συναισθήματος βασισμένη σε λεξικό. Επιπλέον, οι Palakvangsa, Ayudhya και άλλοι (2011) προτείνουν ένα σύστημα σύνοψης των σχολίων από τα ταξιδιωτικά κοινωνικά δίκτυα, όπως το TripAdvisor. Ομοίως, οι ερευνητές Marrese-Taylor, Velásquez και Marquez (2014) ανέπτυξαν ένα εργαλείο για την ανάλυση των απόψεων των τουριστών για εστιατόρια καθώς και ξενοδοχεία από μια περιοχή στη Χιλή.

Στη παρούσα διπλωματική συλλέξαμε ένα μεγάλο σώμα δεδομένων από αξιολογήσεις χρηστών από την ταξιδιωτική πλατφόρμα TripAdvisor σχετικά με πέντε βασικά τουριστικά αξιοθέατα της Αθήνας: την Ακρόπολη, τον Παρθενώνα, το Μουσείο της Ακρόπολης, την Πλάκα και την Αρχαία Αγορά των Αθηνών.

Στη συνέχεια προχωρήσαμε σε ανάλυση συναισθήματος με σκοπό να βρούμε την πολικότητα των αξιολογήσεων αυτών, αν έχουν θετικό ή αρνητικό πρόσημο με τη

χρήση του λεξικού συναισθήματος NRC Word-Emotion Association Lexicon (EmoLex). Τέλος, πραγματοποιήθηκε ομαδοποίηση των δεδομένων (K-means clustering) για κάθε σημείο μελέτης με σκοπό να δημιουργηθούν ομάδες δεδομένων με νοηματική συνάφεια, που να δείχνουν κεντρικά σημεία αναφοράς των χρηστών σχετικά με προβλήματα ή πλεονεκτήματα που επαναλαμβάνονται μέσα στο σύνολο των αξιολογήσεων τους.

3. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΟΡΙΣΜΟΙ

3.1 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ (SENTIMENT ANALYSIS-SA)

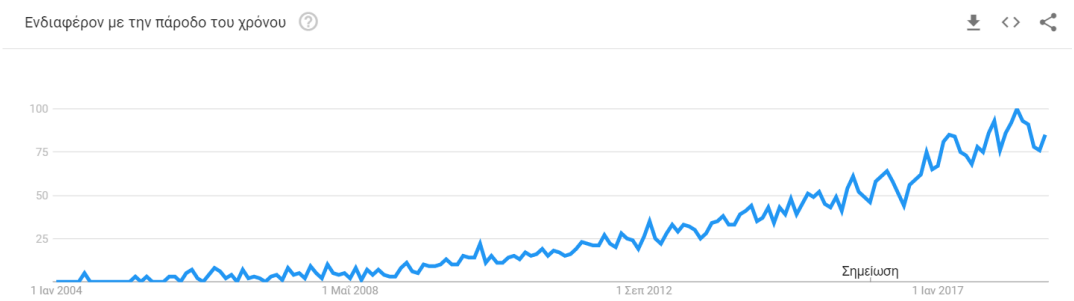
Η ανάλυση συναισθήματος, που ονομάζεται επίσης εξόρυξη γνώμης (opinion mining), είναι ο τομέας της επιστήμης που αναλύει τις απόψεις, τα αισθήματα, τις εκτιμήσεις, τις στάσεις, τα συναισθήματα προς οντότητες και τα χαρακτηριστικά γνωρίσματά που εκφράζονται σε γραπτό κείμενο (Liu, 2014).

Η ανάλυση συναισθήματος είναι μια σειρά από μεθόδους, τεχνικές και εργαλεία ανίχνευσης και εξαγωγής υποκειμενικών πληροφοριών, όπως της γνώμης και των απόψεων, από τη φυσική γλώσσα .

Ο όρος ανάλυση συναισθήματος πιθανότατα πρωτοεμφανίστηκε από τους Nasukawa and Yi (2003), και τον όρο εξόρυξη γνώμης τον συναντάμε για πρώτη φορά σε κείμενο των Dave, Lawrence, and Pennock (2003).

Παρόλα αυτά η έρευνα σχετικά με την ανάλυση συναισθήματος και την εξόρυξη γνώμης εντοπίζεται λίγο νωρίτερα σε μελέτες των Wiebe, (2000), Das and Chen, (2001), Tong, (2001), Morinaga, Yamanishi, Tateishi, και Fukushima, (2002), Turney, (2002).

Τα τελευταία χρόνια εμφανίζεται αυξημένο ενδιαφέρον από τους χρήστες του διαδικτύου για το πεδίο της ανάλυσης συναισθήματος και εξόρυξη γνώμης, όπως διαφαίνεται και από το παρακάτω σχήμα (Google Trends, 2018)[2].



Εικόνα 1: Απεικόνιση Google Trends σχετικά με τον όρο Sentiment analysis

[2] https://trends.google.com/trends/explore?date=all&q=sentiment%20analysis

Σύμφωνα με το λεξικό της Οξφόρδης, η ανάλυση συναισθημάτων αφορά την πολικότητα των απόψεων, που εκφράζονται σε γραπτό κείμενο, δηλαδή αν κάποιος έχει θετική, ουδέτερη ή αρνητική γνώμη αναφορικά με κάποιο θέμα (Schrauwen S., 2010).

Το αντικείμενο της ανάλυσης του συναισθήματος είναι συνήθως ένα προϊόν ή μια υπηρεσία η οποία έχει δημοσιοποιηθεί στο διαδίκτυο. Αυτό μπορεί να εξηγήσει γιατί συχνά η ανάλυση του συναισθήματος (sentiment analysis) και η εξόρυξη γνώμης (opinion mining) χρησιμοποιούνται ως συνώνυμα, αν και είναι πιο ακριβές να βλέπουμε τα συναισθήματα ως συναισθηματικά φορτισμένες απόψεις.

Καθώς η ανάλυση του συναισθήματος μπορεί να προσεγγιστεί ως διαδικασία ταξινόμησης, μπορούμε να σημειώσουμε τρία κύρια επίπεδα ταξινόμησης:

● **Ανάλυση συναισθήματος σε επίπεδο εγγράφου (document-level sentiment analysis):**

Θεωρεί ότι το σύνολο του εγγράφου είναι η βασική μονάδα πληροφοριών η οποία μιλά για ένα θέμα και έχει ως στόχο να την χαρακτηρίσει ότι έχει θετικό ή αρνητικό συναίσθημα προς το θέμα.

● **Ανάλυση συναισθήματος σε επίπεδο πρότασης (Sentence Level Sentiment Analysis):**

Θεωρεί κάθε πρόταση ως βασική μονάδα πληροφορίας. Αρχικά, προσδιορίζει την κάθε πρόταση ως υποκειμενική ή αντικειμενική. Στη συνέχεια, αν η πρόταση χαρακτηριστεί με υποκειμενική χροιά, στοχεύει να την χαρακτηρίσει ως προς την πολικότητα της ως θετική ή αρνητική.

● **Ανάλυση συναισθήματος σε επίπεδο άποψης (Aspect Level Sentiment Analysis):**

Το επίπεδο αυτό σκοπεύει να ταξινομήσει το συναίσθημα σε σχέση με συγκεκριμένες απόψεις των κειμενικών οντοτήτων. Αυτό προϋποθέτει τον προσδιορισμό των οντοτήτων και των προσωπικών τους απόψεων. Παράδειγμα διαφορετικών απόψεων για διαφορετικές πτυχές της ίδιας οντότητας θα μπορούσε να μοιάζει με αυτό " Η εργασία που σχετίζεται με τη διατριβή δεν εξετάζεται διεξοδικά, αλλά τα αποτελέσματα των διατριβών είναι πολύ ενδιαφέροντα », (Medhat, Korashy, 2014).

Είναι σημαντικό να σημειωθεί εδώ ότι δεν υπάρχει καμία ουσιαστική διαφορά μεταξύ της ανάλυσης συναισθήματος σε επίπεδο εγγράφου και της ταξινόμησης σε επίπεδο προτάσεων, καθώς οι προτάσεις μπορούν να θεωρηθούν ως σύντομα έγγραφα (Liu, 2012).

3.1.1 ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Οι τεχνικές ανάλυσης του συναισθήματος μπορούν να κατηγοριοποιηθούν ευρέως σε δύο κύριες κατηγορίες.

● **Προσεγγίσεις μηχανικής μάθησης (Machine Learning Approaches)**

Η πρώτη κατηγορία προσεγγίσεων χρησιμοποιεί μια αναπαράσταση χαρακτηριστικών του κειμένου (textual feature representation) σε συνδυασμό με αλγόριθμους μηχανικής μάθησης (machine learning algorithms) για την εξαγωγή της σχέσης μεταξύ των χαρακτηριστικών, τμημάτων του κειμένου και τις απόψεις που εκφράζονται μέσα στο κείμενο, είτε με προσεγγίσεις μηχανικής μάθησης υπό επίβλεψη ανάλυσης συναισθήματος (supervised learning) είτε με προσεγγίσεις μηχανικής μάθησης χωρίς επίβλεψη ανάλυσης συναισθήματος (unsupervised learning), (Pang, Lee και Vaithyanathan, 2014).

Με τον όρο μηχανική μάθηση υπό επίβλεψη (supervised machine learning sentiment analysis), νοείται ότι ο ταξινομητής λαμβάνει ένα σύνολο από αυτόματα κατηγοριοποιημένα παραδείγματα ως δεδομένα εκπαίδευσης και στη συνέχεια αυτό του δίνει την δυνατότητα να κάνει προβλέψεις για όλα τα αθέατα σημεία (Mohri, Rostamizadeh και Talwalkar, 2012). Με τον όρο μηχανική μάθηση χωρίς επίβλεψη (unsupervised machine learning sentiment analysis), ορίζεται η ταξινόμηση εκείνη που λαμβάνει μη κατηγοριοποιημένα δεδομένα και κάνει αξιολογήσεις ή προβλέψεις για όλα τα αόρατα σημεία του, χωρίς να έχει προηγηθεί κάποια εκπαίδευση σε δεδομένα (Mohri, Rostamizadeh και Talwalkar, 2012). Οι κατηγοριοποιήσεις μπορούν να εκχωρηθούν σε παραδείγματα εκπαίδευσης χειροκίνητα μέσω ανθρώπινης αξιολόγησης του κειμένου ή μπορεί να προκαθοριστεί με βαθμολογίες, όπως παραδείγματος χάριν, γίνεται με τον αριθμό των αστεριών σε μια κριτική (Ghiassi, Skinner και Zimbra, 2013). Η μηχανική μάθηση με μερική επίβλεψη ανάλυση συναισθήματος και η μηχανική μάθηση χωρίς επίβλεψη προτείνονται σαν

τεχνικές όταν δεν υπάρχει ένα αρχικό σύνολο από κατηγοριοποιημένα δεδομένα για την εκπαίδευση του αλγόριθμου ταξινόμησης (Xianghua, Guo, Yanyan και Zhiqian, 2013).

Επιπλέον, οι υβριδικές προσεγγίσεις, που συνδυάζουν την μηχανική μάθηση με επίβλεψη και χωρίς επίβλεψη König and E. Bril, 2006) ή ακόμη και οι τεχνικές μηχανικής μάθησης μερικής επίβλεψης μπορούν να χρησιμοποιηθούν για την ταξινόμηση των συναισθημάτων (Kim και Lee, 2014).

Όσον αφορά την επιλογή χαρακτηριστικών για την ανίχνευση του συναισθήματος, η επεξεργασία της Φυσικής Γλώσσαw (Natural Language Processing) διαδραματίζει σημαντικό ρόλο, καθώς παρέχει χαρακτηριστικά, όπως η συχνότητα των όρων, μέρος των πληροφοριών ομιλίας και συντακτικές εξαρτήσεις (Serrano-Guerrero, Olivas, Romero, Herrera-Viedma, 2015).

Πολλές εμφανείς προσεγγίσεις στην ανάλυση συναισθήματος χρησιμοποιούν τεχνικές μάθησης για να αναπτύξουν ταξινομητές (Pang, Lee & Vaithyanathan, 2002, Gamon, 2004). Μεταξύ των διαφόρων τεχνικών μηχανικής μάθησης , οι ταξινομητές Suppot Vector Machines (SVMs) και Naïve Bayes εφαρμόζονται συνήθως (Ghiassi, Skinner και Zimbra, 2013, Pang, Lee και Vaithyanathan, 2002). Η χρήση της βαθιάς μάθησης, τα τεχνητά νευρωνικά δίκτυα s (such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks) και τα μοντέλα μέγιστης εντροπίας έχουν επίσης επιδείξει επιτυχία στις εφαρμογές ανάλυσης συναισθήματος, κερδίζοντας ολοένα και περισσότερη δημοτικότητα (Ghiassi, Skinner & Zimbra, 2013). Ο ταξινομητής Support Vector Machines μπορεί επίσης να συνδυαστεί με μεθόδους νευρωνικού δικτύου ή με χαρακτηριστικά πυκνής ενσωμάτωσης λέξεων ((Ghiassi, Skinner & Zimbra, 2013).

- **Προσεγγίσεις που βασίζονται σε λεξικό (Lexicon-Based Approaches)**

Η προσέγγιση με λεξικό υπολογίζει το συναίσθημα ενός ήδη υπάρχοντος κειμένου από την πολικότητα των λέξεων ή φράσεων σε αυτό το κείμενο, δηλαδή αν έχουν θετικό, αρνητικό ή ουδέτερο πρόσημο (Turney, 2002). Για αυτή τη μέθοδο απαιτείται η χρήση ενός λεξικού (dictionary) λέξεων με ήδη καταχωρημένη την πολικότητα τους. Μερικά παραδείγματα υφιστάμενων λεξικών

είναι: Opinion Lexicon (Hu and Liu, 2004), SentiWordNet (Esuli and Sebastiani, 2006), AFINN Lexicon (Nielsen, 2011), NRC-Hashtag (Mohammad, Kiritchenko and Zhu 2013).

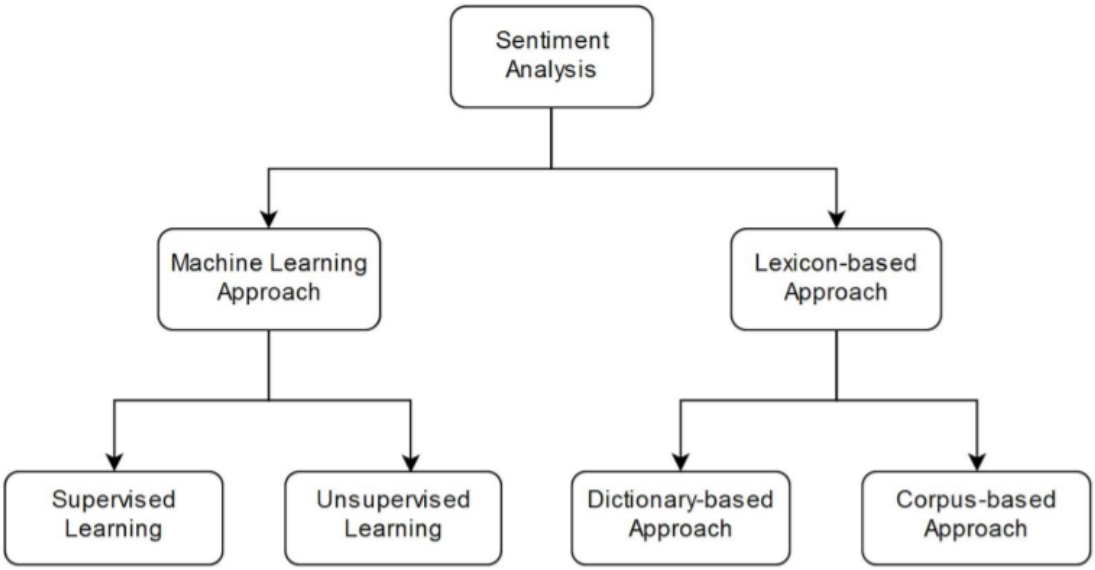
Σε αυτή την προσέγγιση, μπορεί κανείς να βρει δύο υποκατηγορίες. Η βασισμένη σε λεξικό ανάλυση συναισθήματος (dictionary based analysis), όπου η ανάλυση βασίζεται στη χρήση ενός αρχικού συνόλου όρων που συνήθως συλλέγονται και ταξινομούνται χειροκίνητα και η βασισμένη σε σώμα κειμένων προσέγγιση (corpus-based approach), όπου η ανάλυση βασίζεται σε λεξικά που σχετίζονται με το συγκεκριμένο τομέα. Αυτά τα τα λεξικά δημιουργούνται από ένα σύνολο όρων που μεγαλώνει μέσα από την αναζήτηση συναφών λέξεων με στατιστικές ή σημασιολογικές τεχνικές (Serrano, Olivas, Romero, Herrera, 2015).

Η βαθμολογία του συναισθήματος του κειμένου μπορεί να υπολογιστεί ως ο μέσος όρος των πολικοτήτων που μεταφέρει κάθε μια από τις λέξεις στο κείμενο. Η μεθοδολογία για τον υπολογισμό του συναισθήματος μπορεί να περιγραφεί με το ακόλουθα εξής βήματα:

- Προ-επεξεργασία. Το κείμενο υποβάλλεται σε βήματα προ-επεξεργασίας: Προσδιορισμός των μερών του λόγου (Part Of Speech Tagging-POS) , απομάκρυνση καταλήξεων, απομάκρυνση συνδετικών λέξεων (stop-words removal), χειρισμός άρνησης, αναγνώριση λεκτικών μονάδων (tokenisations) σε n-grams. Το αποτέλεσμα της προ-επεξεργασίας είναι ένα σύνολο λεκτικών μονάδων (tokens) ή ένα σύνολο λέξεων (bags of words).
- Έλεγχος κάθε λεκτικής μονάδας (token) για την πολικότητα του στο λεξικό. Κάθε λέξη από το σύνολο των λέξεων (bags of words) συγκρίνεται με το λεξικό. Αν η λέξη βρίσκεται στο λεξικό, τότε η πολικότητα της λέξης αυτής προστίθεται στο άθροισμα του συναισθήματος του κειμένου. Αν είναι η λέξη που δεν βρίσκεται στο λεξικό, η πολικότητα του θεωρείται ίση με μηδέν.
- Υπολογισμός του αθροίσματος του συναισθήματος του κειμένου. Αφού γίνει η κατανομή των βαθμολογιών της πολικότητας όλων των λέξεων που περιέχονται στο κείμενο, η τελική βαθμολογία συναισθήματος του κειμένου υπολογίζεται διαιρώντας το άθροισμα των αποτελεσμάτων των λέξεων που εκφράζουν συναίσθημα με τον αριθμό του συνόλου αυτών των λέξεων.

Μια μεθοδολογία ανάλυσης συναισθημάτων που βασίζεται σε λεξικό για την αγγλική γλώσσα χρησιμοποιήθηκε για την παρούσα εργασία, με σκοπό να αναλύσει αξιολογήσεις χρηστών του TripAdvisor αναφορικά με πέντε βασικά μνημεία και τουριστικά σημεία ενδιαφέροντος της Αθήνας.

Το λεξικό που χρησιμοποιήθηκε για την πραγμάτωση της ανάλυσης είναι το NRC-Emotion Association Lexicon (Emolex), το οποίο πέρα από τον προσδιορισμό της πολικότητας των λέξεων σε θετική και αρνητική, ταξινομεί τις λεκτικές μονάδες ανάλογα με 6 βασικά συναισθήματα (φόβος, θυμός, λύπη, χαρά έκπληξη και προσδοκία).



Εικόνα 2: Τεχνικές Ανάλυσης Συναισθήματος

3.1.2 ΠΕΡΙΟΡΙΣΜΟΙ ΣΤΗΝ ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Οι μέθοδοι ανάλυσης συναισθήματος (Sentiment Analysis Methods-SAM), οι οποίες είναι εκπαιδευμένες για την ανίχνευση της πολικότητας συναισθημάτων, (Cambria και Hussain 2015) μπορούν να ανιχνεύσουν αυτόματα τα συναισθήματα από έγγραφα, προτάσεις ή λέξεις. Μια μεγάλη ποικιλία SAM απευθύνεται στις διάφορες κατηγορίες κειμένων (blogs, κριτικές, tweets κ.ο.κ.). Ωστόσο, η ανάλυση των συναισθημάτων δεν είναι μια τέλεια επιστήμη, ειδικά όταν εφαρμόζεται στα αδόμητα κείμενα που κυριαρχούν στο διαδίκτυο (Khan, Bashir, και Qamar, 2014). Η ανθρώπινη γλώσσα είναι περίπλοκη, έτσι διδάσκοντας μια μηχανή να ανιχνεύει

διαφορετικές γραμματικές αποχρώσεις, πολιτισμικές παραλλαγές, ορολογία και ορθογραφικά λάθη στα μηνύματα στο δίκτυο είναι μια δύσκολη διαδικασία και είναι ακόμη πιο δύσκολο να κατανοήσει αυτόματα πώς το περιεχόμενο μπορεί να επηρεάσει τον τόνο του μηνύματος. Επειδή οι άνθρωποι μπορούν να εφαρμόσουν μια συμφραζόμενη κατανόηση, μπορούν να ερμηνεύσουν διαισθητικά την σκοπιμότητα κάθε γραφής. Οι υπολογιστές, ωστόσο, δυσκολεύονται να κατανοήσουν το πλαίσιο μέσα στο οποίο εκφράζεται μια φράση και να εντοπίσουν για παράδειγμα αν ένα άτομο είναι σαρκαστικό ή όχι.

Μερικά από τις ανοικτές προκλήσεις που αντιμετωπίζει το πεδίο της εξόρυξης κειμένων και της ανάλυσης συναισθήματος συνοψίζονται ως εξής:

• Άρνηση (Negation)

Η άρνηση είναι πολύ σημαντική για την δομή του κειμένου διότι η χρήση αρνήσεως μπορεί να αλλάξει την πολικότητα του κειμένου (Wiegand, Balahur, Roth, Klakow, Montoyo, 2010). Οι όροι άρνησης επηρεάζουν την πολικότητα των συμφραζόμενων λέξεων, αλλά η παρουσία μιας αρνητικής λέξης σε μια πρόταση δεν σημαίνει αυτομάτως ότι όλες οι λέξεις που εκφράζουν κάποιο συναίσθημα είναι ανεστραμμένες. Η άρνηση δεν εκφράζεται μόνο μέσω των κοινών αρνητικών λέξεων, όπως για παράδειγμα είναι στην αγγλική οι λέξεις ‘not’, ‘never’, ‘no’όχι,’, αλλά και από άλλες λεξικολογικές μονάδες.

• Αναφόρηση (Anaphora)

Ένα άλλο σημαντικό εμπόδιο είναι ο χειρισμός της ανάλυσης της αναφόρησης. Αναφόρηση σημαίνει η αναφορά στο ίδιο πράγμα αλλά με διαφορετικές φράσεις ή λέξεις (Ragini και Rubesh , 2016) Αυτό αποτελεί πρόβλημα κυρίως κατά την ομαδοποίηση των οντοτήτων κατά την ανάλυση συναισθήματος βασισμένη στην άποψη (aspect-based sentiment analysis). Η αναφόρηση αποτελεί πρόκληση για την επεξεργασία της φυσικής γλώσσας από τους υπολογιστές , αφού η αναγνώριση της αναφοράς μπορεί να είναι δύσκολη. Για παράδειγμα, ‘I love Acropolis. Its spirit can turn you back thousands of years’. Η λέξη ‘its’ αναφέρεται στην λέξη ‘Acropolis’, άρα δύο διαφορετικές λέξεις εκφράζουν ακριβώς την ίδια έννοια, άρα οι δύο αυτές πτυχές πρέπει να συνδυαστούν προκειμένου να παραχθούν ακριβή αποτελέσματα.

• Διάταξη των λέξεων μέσα στην πρόταση (Word arrangement in a sentence)

Η διάταξη λέξεων σε μια πρόταση διαδραματίζει ζωτικό ρόλο διότι προσδιορίζει τον υποκειμενικό χαρακτήρα του κειμένου. Η σειρά των λέξεων είναι σημαντική για να αποφασιστεί η πολικότητα ενός κειμένου. Σε ένα συγκεκριμένο κομμάτι κειμένου, εάν αλλάξει η σειρά των λέξεων, η πολικότητα του κειμένου επηρεάζεται.

• Έμμεσο συναίσθημα και Σαρκασμός (Implicit Sentiment and Sarcasm):

Ακόμη και χωρίς την παρουσία οποιασδήποτε συναισθηματικά φορτισμένης λέξης, οι προτάσεις μπορούν να κρύβουν κάποιο ένα έμμεσο συναίσθημα. Για παράδειγμα, "Πώς μπορείτε να μου το κάνετε αυτό;" Σε αυτή την πρόταση, καμία από τις λέξεις δεν εκφράζει αρνητική διάθεση, αλλά το νόημα της πρότασης είναι αρνητικό. Έτσι η σημασιολογία αποτελεί πολύ σημαντικό παράγοντα στη σημασιολογική ανάλυση των αρχείων κειμένου.

• Ανίχνευση ανεπιθύμητων μηνυμάτων (Spam Detection)

Οποιοσδήποτε από οποιαδήποτε τοποθεσία μπορεί να εκφράζει τις απόψεις τους στα κοινωνικά μέσα και τις πλατφόρμες δικτύωσης χωρίς να αποκαλύπτει την πραγματική του ταυτότητα. Με αυτόν τον τρόπο, γίνονται πολλά ψεύτικα σχόλια και αξιολογήσεις για την προώθηση των πωλήσεων κάποιας υπηρεσίας ή προϊόντος. Μια τέτοια δραστηριότητα ονομάζεται αποστολή ανεπιθύμητων μηνυμάτων (spamming). Εκτός από τα άτομα, υπάρχουν και εμπορικές εταιρίες που είναι σε αυτή την επιχείρηση με στόχο να εξαπλώνουν ψεύτικες πληροφορίες (Ragini και Rubesh , 2016). Ο προσδιορισμός αυτών των αξιολογήσεων πρόκειται για ένα δύσκολο έργο για να μπορέσει να εξαχθεί το ακριβές συναίσθημα.

• Σύνδεσμοι (Conjunctions)

Η παρουσία συνδέσμων σε μια πρόταση μεταβάλλει ολόκληρη την έννοια της φράσης. Για παράδειγμα: ‘It was an amazing experience, but the service was poor’. Αυτή η πρόταση χωρίζεται σε δύο μέρη. Όταν αναλύουμε το πρώτο μέρος, παίρνουμε ένα θετικό συναίσθημα. Αλλά η παρουσία του συνδέσμου ‘but’ αναστρέφει ολόκληρη την έννοια της φράσης. Συνεπώς, οι σύνδεσμοι θα πρέπει να λαμβάνονται υπόψη για το συναίσθημα ανάλυση.

• Η επίλυση συν-αναφοράς (Co-reference resolution)

Η επίλυση της συν-αναφοράς είναι μία από τις μεγαλύτερες έρευνες προκλήσεις. Αυτό ισχύει περισσότερο σε μέρη όπου χρησιμοποιούνται συγκριτικά κείμενα. Η συν-αναφορά μεταξύ των προτάσεων πρέπει να επιλυθεί αποτελεσματικά προκειμένου να υπάρξει καλύτερη ανάλυση (Ravi και Ravi ,2015).

•Προσαρμογή Τομέα (Domain Adaptation)

Η προσαρμογή τομέα είναι μια άλλη σημαντική πτυχή της ανάλυσης συναισθήματος. Το μεγαλύτερο μέρος των διαθέσιμων λεξικών συναισθήματος είναι γενικού σκοπού. Παρόλο που είναι γενικού σκοπού, είναι σημαντικό να μελετήσουμε τους τρόπους προσαρμογής τους σε κάθε συγκεκριμένο τομέα. Από την άποψη αυτή, υπάρχουν τρία βασικά ζητήματα. Πρώτον, μπορεί η ίδια οντότητα να έχει διαφορετική πολικότητα σε διαφορετικά πεδία. Το δεύτερο θέμα είναι η ανάθεση ενός δείκτη δύναμης για κάθε μία λέξη που εκφράζει συναίσθημα. Τρίτον, είναι η διαφορά στο λεξιλόγιο σε κάθε διαφορετικό τομέα που κάνουν την ανάλυση συναισθήματος μια εφαρμογή που εξαρτάται από τον εκάστοτε τομέα που μελετάει(Ragini και Rubesh , 2016).

- **Emoticons**

Η χρήση emoticons, δηλαδή αυτών των κίτρινων συνήθως φατσών που υπάρχουν πλέον στο διαδίκτυο, αποτελούν σημαντικές ενδείξεις για την έκφραση των συναισθημάτων των ατόμων που τις χρησιμοποιούν, όμως παράλληλα μπορεί να αποτελέσουν εμπόδιο στην επεξεργασία του κειμένου από τις μηχανές. Τα τελευταία χρόνια έχουν δημιουργηθεί λεξικά που συμπεριλαμβάνουν αυτού του είδους τα σύμβολα, με στόχο την αποτελεσματικότερη ανάλυση και επεξεργασία του κειμένου, που τα πλαισιώνουν.

- **Συντομογραφίες (abbreviations)**

Ακόμη μια περίπτωση που δυσχεράνει την απόλυτη κατανόηση και ανάλυση των δεδομένων κειμένου που υπάρχουν στο κοινωνικό ιστό είναι η χρήση συντομογραφιών. Οι συντομογραφίες αποτελούν σύμπτυξη των λέξεων που συμβολίζουν και μπορεί να περιέχουν σημαντική πληροφορία, όπως την έκφραση συναισθήματος. Γι’ αυτό είναι σημαντικό να συμπεριλαμβάνονται στις αναλύσεις κειμένου, αν και είναι αρκετά δύσκολο να επιτευχθεί.

- **Αργώ (slang)**

Η χρήση πρόχειρης και καθημερινής γλώσσας από τους χρήστες κυρίως των μέσων κοινωνικής δικτύωσης αποτελεί συχνό φαινόμενο. Αυτός ο καθημερινός λόγος που συναντιέται κυρίως σε πλατφόρμες όπως το Facebook και το Instagram, μπορεί να υπολανθάνει σημαντικές πληροφορίες για αυτό που θέλει να εκφράσει ο χρήστης. Τέτοιες καταστάσεις ίσως είναι δύσκολο να εντοπιστούν από τους αλγόριθμους μηχανικής μάθησης, αλλά η χρήση λεξικών ίσως προσαρμόζεται καλύτερα αν έχουν καταχωρηθεί στο κορμό των δεδομένων τους άρα μπορούν να εντοπιστούν γρηγορότερα από τις μηχανές ανάλυσης συναισθήματος και εξόρυξης γνώμης.

3.2 ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ (TEXT MINING)

Η εξόρυξη κειμένου (text mining), επίσης γνωστή ως εξόρυξη δεδομένων κειμένου (text data mining) (Hearst, 1997) ή η ανακάλυψη γνώσης από κειμενικές βάσεις δεδομένων (knowledge discovery from textual databases) (Feldman και Dagman, 1995) αναφέρεται γενικά στη διαδικασία εξαγωγής ενδιαφερουσών και μη επουσιωδών μοτίβων ή γνώσης από μη δομημένα έγγραφα κειμένου. Μπορεί να θεωρηθεί ως επέκταση της εξόρυξης δεδομένων (data mining) ή της ανακάλυψης γνώσης από (δομημένες) βάσεις δεδομένων (Kao και Poteet, 2007). Η εξόρυξη κειμένου είναι παρόμοια με την εξόρυξη δεδομένων, με την διαφορά ότι τα εργαλεία εξόρυξης δεδομένων (Navathe, Shamkant, και Ramez, 2000) σχεδιάζονται για να χειρίζονται δομημένα δεδομένα από βάσεις δεδομένων, ενώ η εξόρυξη κειμένου μπορεί επίσης να λειτουργήσει με μη δομημένη ή ημι-δομημένα σύνολα δεδομένων όπως μηνύματα ηλεκτρονικού ταχυδρομείου (emails), έγγραφα κειμένου και αρχεία HTML. Ως αποτέλεσμα, η εξόρυξη κειμένου είναι πολύ καλύτερη λύση.

Η εξόρυξη κειμένου είναι συνήθως η διαδικασία της δόμησης ενός κειμένου (συνήθως η ανάλυση του, μαζί με την προσθήκη κάποιων γλωσσικών χαρακτηριστικών και την απομάκρυνση κάποιων άλλων μετά από την εισαγωγή του σε μια βάση δεδομένων), αποδίδοντας μοτίβα μέσα στο δομημένα δεδομένα και την τελική αξιολόγηση και ερμηνεία του αποτελέσματος.

Ο όρος "εξόρυξη κειμένου" χρησιμοποιείται συνήθως για να δηλώσει οποιαδήποτε σύστημα που αναλύει μεγάλες ποσότητες κειμένου φυσικής γλώσσας και ανιχνεύει

λεξικά ή γλωσσικά πρότυπα χρήσης σε μια προσπάθεια για να εξαχθεί περισσότερο χρήσιμες (αν και μάλλον σωστές) πληροφορίες (Sebastiani, 2002).

Η εξόρυξη κειμένου είναι επίσης ένα πολύ σύνθετο έργο, καθώς περιλαμβάνει την επεξεργασία δεδομένων κειμένου που είναι εγγενώς μη δομημένο και ασαφές. Η εξόρυξη κειμένου είναι ένα διεπιστημονικό πεδίο, το οποίο περιλαμβάνει την ανάκτηση πληροφοριών (information retrieval), ανάλυση κειμένου (text analysis), εξαγωγή πληροφοριών (information extraction), ομαδοποίηση (clustering), κατηγοριοποίηση (categorization), οπτικοποίηση (visualization), τεχνολογία βάσης δεδομένων (database technology), τη μηχανική μάθηση (machine learning) και την εξόρυξη δεδομένων (data mining).

Το πρόβλημα που εισάγεται από την εξόρυξη κειμένου είναι προφανές: φυσικό η γλώσσα αναπτύχθηκε για να επικοινωνούν οι άνθρωποι μεταξύ τους και να καταγράφουν πληροφορίες ενώ οι υπολογιστές απέχουν κατά πολύ από την κατανόηση της φυσικής γλώσσας. Οι άνθρωποι έχουν τη δυνατότητα να διακρίνουν και να εφαρμόζουν τα γλωσσικά μοτίβα στο κείμενο και οι άνθρωποι μπορούν εύκολα να ξεπεράσουν εμπόδια που οι υπολογιστές δεν μπορούν εύκολα να χειριστούν όπως η χρήση της αργκό, ορθογραφικές παραλλαγές και συμφραζόμενα. Ωστόσο, αν και οι γλωσσικές μας δυνατότητες μας επιτρέπουν την κατανόηση των αδόμητων κειμένων, μας λείπει η δυνατότητα των υπολογιστών για επεξεργασία μεγάλων όγκων κειμένου ή σε υψηλές ταχύτητες.

Ένα έγγραφο κειμένου περιέχει χαρακτήρες που μαζί σχηματίζουν λέξεις, οι οποίες μπορούν να συνδυαστούν περαιτέρω για τη δημιουργία φράσεων. Αυτές είναι όλες οι συντακτικές ιδιότητες που αντιπροσωπεύουν από κοινού ήδη καθορισμένες κατηγορίες, έννοιες, αισθήσεις ή έννοιες (Scholtes, 2009). Η εξόρυξη κειμένου πρέπει να αναγνωρίζει, να εξάγει και να χρησιμοποιεί τις πληροφορίες. Αντί να ψάχνει λέξεις, μπορεί να αναζητήσει σημασιολογικά μοτίβα, και αυτό σημαίνει ότι ψάχνει σε υψηλότερο επίπεδο.

3.2.1 ΔΙΑΔΙΚΑΣΙΑ ΕΞΟΡΥΞΗΣ ΚΕΙΜΕΝΟΥ

Η εξόρυξη κειμένου περιλαμβάνει μια σειρά δραστηριοτήτων που πρέπει να εκτελεστούν με την σειρά για την αποτελεσματική εξόρυξη της πληροφορίας. Οι δραστηριότητες αυτές είναι:

- **Προ-επεξεργασία κειμένου (Text Pre-processing):**

Περιλαμβάνει μια σειρά βημάτων:

Α) «Καθαρισμός» κειμένου (Text Cleanup):

«Καθαρισμός» κειμένου σημαίνει απομάκρυνση οποιωνδήποτε περιττών ή ανεπιθύμητων πληροφοριών, όπως η κατάργηση διαφημίσεων από ιστοσελίδες, κανονικοποίηση του κειμένου που μετατρέπεται από δυαδικές μορφές (Kumar, Bhatia, 2013), απομάκρυνση περιττών λέξεων και σημείων στίξεων.

Β) Αναγνώριση λεκτικών μονάδων (Tokenisation):

Η αναγνώριση λεκτικών μονάδων επιτυγχάνεται απλά με το διαχωρισμό του κειμένου σε διαστήματα και στα σημεία στίξης που δεν ανήκουν στις συντομογραφίες που προσδιορίστηκαν στο προηγούμενο στάδιο (Kumar, Bhatia, 2013).

Γ) Κατηγοριοποίηση σε μέρη του λόγου (Part of Speech Tagging-POS):

Η κατηγοριοποίηση σε μέρη του λόγου (POS) σημαίνει την ταξινόμηση σε κατηγορίες των λέξεων για κάθε λεκτική μονάδα. Η κατηγοριοποίηση έχει να αντιμετωπίσει άγνωστες λέξεις και αμφισβητήσιμες αντιστοιχίσεις ετικετών λέξεων. Βασικές προσεγγίσεις (Scholtes, 1995). λειτουργούν σε α) λεξικά που περιέχουν λέξεις μαζί με τις σχετικές κατηγοριοποιήσεις των μερών του λόγου τους καθώς και μορφολογικά και συντακτικά χαρακτηριστικά τους και σε β) κανόνες με ευαίσθητο περιεχόμενο για την επιλογή των κατάλληλων κατηγοριοποιήσεων κατά την εφαρμογή (Kumar, Bhatia, 2013).

- **Μετασχηματισμός κειμένου (δημιουργία χαρακτηριστικών) - Text Transformation (Attribute Generation):**

Ένα έγγραφο κειμένου (text document) αντιπροσωπεύεται από τις λέξεις (χαρακτηριστικά) που περιέχει και τις περιπτώσεις τους (Kumar, Bhatia, 2013).. Δύο

είναι οι κύριες προσεγγίσεις για την αναπαράσταση εγγράφων α) Σύνολο από λέξεις (Bag of words) και β) διάνυσμα διαστήματος (Vector Space).

- **Επιλογή Χαρακτηριστικού (Attribute/Feature Selection):**

Η επιλογή χαρακτηριστικού, γνωστή και ως επιλογή μεταβλητής, είναι η διαδικασία επιλογής ενός υποσυνόλου σημαντικών χαρακτηριστικών για χρήση στο δημιουργία του μοντέλου. Η κύρια υπόθεση όταν χρησιμοποιείτε ένα η τεχνική επιλογής χαρακτηριστικού είναι ότι τα δεδομένα περιέχουν πολλά πλεονάζοντα ή άσχετα χαρακτηριστικά. Τα πλεονάζοντα χαρακτηριστικά είναι αυτά που δεν παρέχουν επιπλέον πληροφορίες. Τα άσχετα χαρακτηριστικά παρέχουν ασήμαντες ή άσχετες πληροφορίες σε οποιοδήποτε κείμενο (Kumar, Bhatia, 2013).. Η τεχνική επιλογής χαρακτηριστικού είναι ένα υποσύνολο του γενικότερου πεδίου της εξαγωγή χαρακτηριστικού (feature extraction).

- **Εξόρυξη δεδομένων (data mining):**

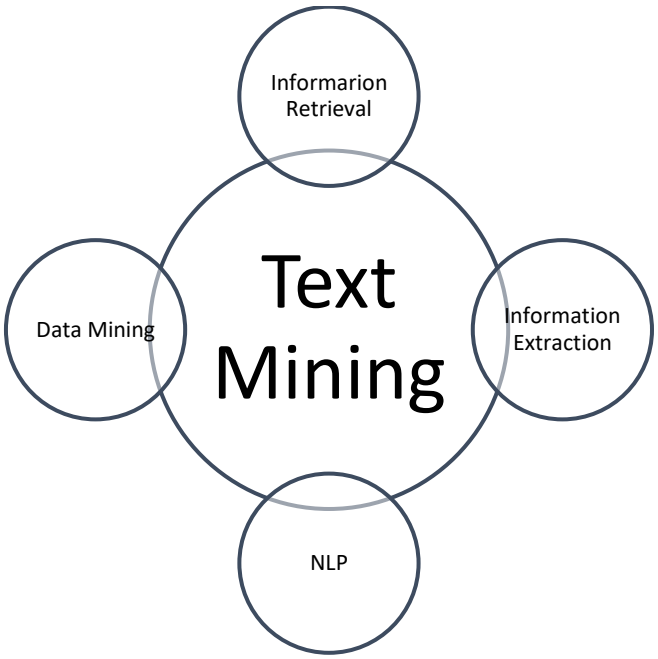
Σε αυτό το σημείο η διαδικασία εξόρυξης κειμένου συγχωνεύεται με την παραδοσιακή διαδικασία εξόρυξης δεδομένων. Τεχνικές κλασική εξόρυξης δεδομένων χρησιμοποιούνται στη δομημένη βάση δεδομένων που προέκυψε από τα προηγούμενα στάδια.

- **Αξιολόγηση:**

Τέλος γίνεται αξιολογήσει του αποτελέσματος που προέκυψε από τα προηγούμενα στάδια. Μετά από την αξιολόγηση, το αποτέλεσμα μπορεί να απορριφθεί ή το αποτέλεσμα που παράγεται μπορεί να χρησιμοποιηθεί ως δεδομένο για το επόμενο σετ ακολουθίας.

3.2.2 ΠΕΔΙΑ ΕΞΟΡΥΞΗΣ ΚΕΙΜΕΝΟΥ

Η ανάλυση κειμένου όπως αναφέρθηκε περιλαμβάνει τα πεδία της ανάκτησης πληροφοριών, τεχνικές εξόρυξης δεδομένων συμπεριλαμβανομένης της σύνδεσης και της ανάλυσης συνδέσεων, της οπτικοποίησης και των προγνωστικών αναλύσεων (Ghosh, Roy, και Bandyopadhyay, 2012). Ο στόχος είναι, ουσιαστικά, να μετατρέψετε το κείμενο (μη δομημένα δεδομένα) σε δεδομένα (δομημένη μορφή) για ανάλυση, μέσω της χρήσης μεθόδων επεξεργασίας της φυσικής γλώσσας (NLP).



Εικόνα 3: Πεδία Εξόρυξης Κειμένου

- **Ανάκτηση πληροφοριών (Information Retrieval-IR)**

Η ανάκτηση πληροφοριών θεωρείται ως επέκταση στο έγγραφο ανάκτησης (document retrieval) όπου γίνεται επεξεργασία των εγγράφων που επιστρέφονται να συμπυκνώσουν ή να εξαγάγουν τις συγκεκριμένες πληροφορίες που ζητήθηκαν από τον χρήστη. Έτσι, μετά την ανάκτηση εγγράφου θα μπορούσε να ακολουθήσει το στάδιο της σύνοψης του κειμένου που εστιάζει στο ερώτημα που θέτει ο χρήστης ή ένα στάδιο εξαγωγής πληροφοριών χρησιμοποιώντας τεχνικές. Τα συστήματα IR βοηθούν στον περιορισμό του συνόλου των εγγράφων που σχετίζονται με ένα συγκεκριμένο πρόβλημα. Καθώς η εξόρυξη κειμένου συνεπάγεται την εφαρμογή πολύ πολύπλοκων αλγορίθμων σε μεγάλες συλλογές εγγράφων, η ανάκτηση της πληροφορίας μπορεί να επιταχύνει την ανάλυση σημαντικά (Ghosh, Roy, και Bandyopadhyay, 2012) μειώνοντας τον αριθμό των εγγράφων για ανάλυση.

- **Εξόρυξη δεδομένων (Data Mining-DM)**

Η εξόρυξη δεδομένων (data mining) μπορεί να περιγραφεί με απλά λόγια ως η αναζήτηση μοτίβων σε δεδομένα. Μπορεί να χαρακτηριστεί πληρέστερα ως η εξαγωγή κρυφών, προηγουμένως άγνωστων και χρήσιμων πληροφοριών (Ghosh,

Roy, και Bandyopadhyay, 2012) από το δεδομένα. Τα εργαλεία εξόρυξης δεδομένων μπορούν να προβλέψουν συμπεριφορές και μελλοντικές τάσεις, επιτρέποντας στις επιχειρήσεις να παράγουν θετικές γνώσεις βασισμένες αποφάσεις. Τα εργαλεία εξόρυξης δεδομένων μπορούν να απαντήσουν στις ερωτήσεις που παραδοσιακά ήταν πολύ χρονοβόρες μέχρι να επιλυθούν. Ψάχνουν μέσα στις βάσεις δεδομένων για κρυφά και άγνωστα μοτίβα, βρίσκοντας κρίσιμες πληροφορίες που οι ειδικοί ενδέχεται να μην μπορούσαν να εντοπίσουν γιατί βρίσκονταν έξω από τις προσδοκίες τους. Ο γενικός στόχος της διαδικασίας εξόρυξης δεδομένων είναι η εξαγωγή πληροφοριών από ένα σύνολο δεδομένων και η μετατροπή του σε μια κατανοητή δομή για περαιτέρω χρήση.

- **Επεξεργασία φυσικής γλώσσας (Natural Language Processing-NLP)**

Η επεξεργασία φυσικής γλώσσας είναι ένα από τα παλαιότερα και πιο δύσκολα προβλήματα στο πεδίο της τεχνητής νοημοσύνης (artificial intelligent-AI). Είναι η μελέτη της ανθρώπινης γλώσσας έτσι ώστε οι υπολογιστές να κατανοούν τις φυσικές γλώσσες όπως κάνουν οι άνθρωποι [3]. Η έρευνα της επεξεργασίας της φυσικής γλώσσας επιδιώκει να απαντήσει το αόριστο ερώτημα για το πώς οι άνθρωποι μπορούν να κατανοήσουν την έννοια μιας φράσης ή ενός εγγράφου. Ποιές είναι οι ενδείξεις που χρησιμοποιούμε για να καταλάβουμε ποιος έκανε τι σε ποιον ή όταν συμβαίνει κάτι, ή ποιο είναι γεγονός και ποια είναι υπόθεση ή πρόβλεψη; Ενώ οι λέξεις –τα ουσιαστικά, τα ρήματα, τα επιρρήματα και τα επίθετα - είναι τα δομικά στοιχεία της έννοιας, είναι ο συσχετισμός μεταξύ τους, εντός της δομή μιας φράσης σε ένα έγγραφο, και μέσα στο πλαίσιο από αυτό που γνωρίζουμε ήδη για τον κόσμο γύρω μας, που παρέχει την πραγματική έννοια ενός κειμένου. Ο ρόλος της επεξεργασίας της φυσικής γλώσσας στην εξόρυξη κειμένου είναι να παραδώσει το σύστημα στο φάση εξαγωγής πληροφοριών ως είσοδο.

Η επεξεργασία φυσικής γλώσσας (Natural Language Processing-NLP), είναι η προσπάθεια εξαγωγής μιας πληρέστερης αναπαράστασης της σημασίας από ελεύθερο κείμενο. Αυτό μπορεί να ερμηνευθεί κατά προσέγγιση ως η κατανόηση του ποιος έκανε τι σε ποιον, πότε, πού, πώς και γιατί. Η NLP συνήθως κάνει χρήση γλωσσικών εννοιών, όπως τα μέρη του λόγου και η γραμματική δομή. Πρέπει να ασχοληθεί με την επανάληψη και τις αμφισημίες. Για να γίνει αυτό, χρησιμοποιεί διάφορες

[3] http://www.scism.lsbu.ac.uk/inmandw/ir/jaberwocky.htm

αναπαραστάσεις γνώσης, όπως ένα λεξικό λέξεων με τις έννοιές τους και τις γραμματικές τους ιδιότητες και ένα σύνολο κανόνων γραμματικής και συχνά άλλους πόρους, όπως την οντολογία των οντοτήτων και των δράσεων, ή ένα λεξικό συνωνύμων ή συντομογραφιών (Kao και Poteet, 2007).

- **Εξόρυξη πληροφοριών (Information Extraction-IE)**

Η εξαγωγή πληροφοριών είναι το έργο αυτόματης εξαγωγής δομημένης πληροφορίας από μη δομημένα ή / και ημιδομημένα αναγνώσιμα έγγραφα από μηχανές. Στις περισσότερες περιπτώσεις αυτή η δραστηριότητα περιλαμβάνει την επεξεργασία κειμένων της ανθρώπινης γλώσσας από μέσα επεξεργασίας φυσικής γλώσσας (NLP). Οι πιο πρόσφατες δραστηριότητες στην επεξεργασία αρχείων πολυμέσων όπως ο αυτόματος σχολιασμός και η εξόρυξη πληροφορίας από εικόνες / ήχο / βίντεο θα μπορούσε να θεωρηθεί ως εξαγωγή πληροφορίας και το καλύτερο πρακτικό και ζωντανό παράδειγμα της εξόρυξης πληροφορίας είναι η μηχανή αναζήτησης της Google (Google Search Engine). Η τελευταία περιλαμβάνει τον προσδιορισμό της γενικής μορφής των πληροφοριών που μας ενδιαφέρει σαν ένα ή περισσότερα υποδείγματα, τα οποία χρησιμοποιούνται για την καθοδήγηση της διαδικασίας εξαγωγής. Τα συστήματα εξόρυξης πληροφοριών εξαρτώνται σε μεγάλο βαθμό από τα δεδομένα που παράγονται από τα συστήματα NLP.

3.3 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ: TRIPADVISOR

Η κοινωνική ταξιδιωτική δικτύωση στο διαδίκτυο έχει αλλάξει τον τρόπο με τον οποίο οι τουρίστες σχεδιάζουν τα ταξίδια τους. Αυτοί οι ιστότοποι επιτρέπουν στους χρήστες να αλληλεπιδρούν και να παρέχουν σχόλια και αξιολογήσεις σε ξενοδοχεία, εστιατόρια ή σε τοπικά τουριστικά σημεία και αξιοθέατα. Μερικά παραδείγματα αυτών των ιστοσελίδων είναι το TravBuddy.com, Travellerspoint, το Yelp και το TripAdvisor.com. Το τελευταίο είναι ίσως η μεγαλύτερη ταξιδιωτική πλατφόρμα στον διαδίκτυο. Δημιουργήθηκε το 2000 και σύμφωνα με την ετήσια χρηματοοικονομική αναφορά του TripAdvisor καλύπτει μέχρι το 2016, 1.035.000 ξενοδοχεία και καταλύματα, 815.000 ενοικιαζόμενα εξοχικά σπίτια, 4.1 εκατομμύρια εστιατόρια και 690.000 χιλιάδες αξιοθέατα και εμπειρίες σε όλο τον κόσμο (TripAdvisor: Ετήσια Οικονομική Έκθεση για το δεύτερο εξάμηνο του 2016, 2016).

Το TripAdvisor είναι ένας ιστότοπος με κεντρική ιδέα ότι οι ταξιδιώτες βασίζονται σε κριτικές άλλων ταξιδιωτών για να προγραμματίσουν τα ταξίδια τους, ή τουλάχιστον μπορεί να συμβάλλουν σε μεγάλο βαθμό στις αποφάσεις τους. Μέχρι το 2016 το TripAdvisor είχε φθάσει τα 385 εκατομμύρια σχόλια και ταξιδιωτικές κριτικές αναφορικά με 6.6 εκατομμύρια καταλύματα, εστιατόρια και αξιοθέατα από τα 5 εκατομμύρια εγγεγραμμένα μέλη του. Το TripAdvisor λειτουργεί σε 48 αγορές παγκοσμίως και υπολογίζει 350 εκατομμύρια μοναδικούς επισκέπτες ανά μήνα κατά μέσο όρο[4].

Οι Web 2.0 ιστότοποι, όπως το Facebook βασίζουν την κοινωνική δικτύωση τους σε κάποιου είδους προηγούμενης σχέσης (Sheldon 2014), ή επάνω σε ισχυρή συνάφεια όσον αφορά τα ενδιαφέροντα ή τις ασχολίες των χρηστών. Σε αντίθεση με αυτά, στη ταξιδιωτική δικτύωση οι χρήστες των ιστότοπων δεν έχουν προϋπάρχουσες ενώσεις. Οι μόνες συνδέσεις τους είναι τα φόρουμ συζητήσεων και οι κριτικές και αξιολογήσεις που δημιουργούνται από άλλους χρήστες (Sheldon, 2014).

Ως απόδειξη της αυξανόμενης σημασίας της αναζήτησης στο διαδίκτυο και της επιρροής στις επιλογές των ταξιδιωτών, μια έρευνα στην βιομηχανία ξενοδοχείων και εστιατόριων στην Ευρώπη το 2007 κατέληξε στο συμπέρασμα ότι το 80% των βρετανών καταναλωτών πραγματοποιεί έρευνα στο διαδίκτυο πριν από οποιαδήποτε κράτηση ξενοδοχείου και οι μισοί από αυτούς, απέφυγαν να κάνουν κράτηση συγκεκριμένου ξενοδοχείου ως συνέπεια της αρνητικής αξιολόγησης σε ιστοσελίδες παρόμοιες με το TripAdvisor (O'Connor, 2018).

Το TripAdvisor, όπως αναφέρθηκε, είναι μια ιστοσελίδα όπου οι περισσότερες πληροφορίες δημοσιεύονται αυτόνομα από τους χρήστες του. Δημοσιεύουν κριτικές, σχόλια και αξιολογήσεις σε προορισμούς, ξενοδοχεία, αξιοθέατα ή οποιοδήποτε άλλο αντικείμενο ή υπηρεσία, που σχετίζεται με τον τουρισμό. Επιπλέον, παρέχεται η δυνατότητα προσθήκης πολυμεσικών στοιχείων, όπως για παράδειγμα φωτογραφίες και βίντεο, ή ταξιδιωτικούς χάρτες προηγουμένων ταξιδιών ή συμμετοχής σε φόρουμ συζήτησης. Οι δημιουργοί του φόρουμ πρέπει να είναι εγγεγραμμένοι χρήστες, πράγμα που τους δίνει τη δυνατότητα να δημιουργήσουν ένα προφίλ που να περιέχει βασικά στοιχεία (ψευδώνυμο, κατοικία κ.λπ.).

[4] http://files.shareholder.com/downloads/AMDA-MMXS5/2604284471x0x903032/5C63829B-B3B4-49D9-8C04-E3C23B6D6A76/TripAdvisor_Reports_Second_Quarter_2016_Financial_Results.pdf

Σύμφωνα με τα βραβεία Travels Choice 2018, η Ακρόπολη κατατάχθηκε 16η στα 25 κορυφαία σημεία ενδιαφέροντος σε όλο το κόσμο και το μουσείο της Ακρόπολης 8ο ανάμεσα στα 25 κορυφαία μνημεία του κόσμου για το έτος 2018 [5].

[5] https://www.tripadvisor.com.gr/TravelersChoice

4. ΕΡΓΑΛΕΙΑ ΕΡΕΥΝΑΣ

4.1 PYTHON

Η Python είναι μια δυναμική γλώσσα προγραμματισμού που για πρώτη φορά ανακοινώθηκε το 1991. Αρχικά σχεδιάστηκε ως μια εύκολη στην εκμάθηση εκπαιδευτική γλώσσα, ενώ τα τελευταία χρόνια η Python κέρδισε δημοτικότητα ως μια παραγωγική πλατφόρμα γενικού σκοπού.

Ωστόσο, η Python μπορεί να χρησιμοποιηθεί για στατιστική επεξεργασία, κυρίως μέσω του αναδυόμενου πλαισίου pandas (pandas framework) - ανοιχτού κώδικα διαχείρισης δεδομένων και ανάλυση βιβλιοθήκη, η οποία παρέχει υψηλής απόδοσης, εύκολο στη χρήση δομών δεδομένων και εργαλεία ανάλυσης δεδομένων. Η βιβλιοθήκη Pandas της Python έχει δημιουργηθεί με στόχο να γίνει το πιο ισχυρό και ευέλικτο εργαλείο ανάλυσης δεδομένων ανοιχτού κώδικα διαθέσιμο σε οποιαδήποτε γλώσσα [6].

4.2 SELENIUM WEB DRIVER

Το Selenium είναι ένα πλαίσιο για την αυτοματοποιημένη δοκιμή εφαρμογών ιστού. Χρησιμοποιώντας το Selenium, μπορεί βασικά να αυτοματοποιηθεί κάθε εργασία στο πρόγραμμα περιήγησής (browser) σαν να έπρεπε να εκτελέσει το έργο ένα πραγματικό άτομο. Η διεπαφή που χρησιμοποιείται για την αποστολή εντολών στα διάφορα προγράμματα περιήγησης ονομάζεται Selenium WebDriver. Οι υλοποιήσεις αυτής της διασύνδεσης είναι διαθέσιμες για κάθε σημαντικό πρόγραμμα περιήγησης, συμπεριλαμβανομένων των προγραμμάτων Mozilla Firefox, Google Chrome και Internet Explorer[7].

4.3 NRC WORD-EMOTION ASSOCIATION LEXICON (EMOLEX)

Το NRC Emotion Lexicon είναι ένας κατάλογος αγγλικών λέξεων και των συσχετισμών τους με οκτώ βασικά αισθήματα (θυμός, φόβος, προσδοκία, εμπιστοσύνη, έκπληξη, θλίψη, χαρά και αηδία) και δύο συναισθήματα (αρνητικά και θετικά). Περιέχει 14,182 λεκτικές μονάδες/λέξεις (unigrams) Για την ανάλυση, χρησιμοποιούνται μόνο "θετικές" και "αρνητικές" αξιολογήσεις με μέγιστη βαθμονόμηση το 1 Όσο πιο κοντά στο 1 βρίσκεται μια λέξη η φράση, τόσο

[6] http://pandas.pydata.org/
[7] https://www.smashingmagazine.com/2018/04/feature-testing-selenium-webdriver/

περισσότερο εκφράζει το εκάστοτε συναίσθημα ή την πολικότητα. Οι σημειώσεις δημιουργήθηκαν χειροκίνητα μέσω της διαδικασία του πληθοπορισμού (crowdsourcing).

4.4 PANDAS LIBRARY

Το εγχειρίδιο της βιβλιοθήκης Pandas ορίζει ένα DataFrame ως μια δισδιάστατη, μεταβλητού μεγέθους μεταβλητή, με δυνητικά ετερογενή δομή δεδομένων σε μορφή πίνακα με άξονες (σειρές και στήλες). Με απλά λόγια, σκεφτείτε ένα DataFrame πίνακα δεδομένων, δηλαδή ένα ενιαίο σύνολο μορφοποιημένων δισδιάστατων δεδομένων, με τα ακόλουθα χαρακτηριστικά:

- Μπορούν να υπάρχουν πολλές σειρές και στήλες στα δεδομένα.
- Κάθε σειρά αντιπροσωπεύει ένα δείγμα δεδομένων,
- Κάθε στήλη περιέχει μια διαφορετική μεταβλητή που περιγράφει τα δείγματα (σειρές).

Τα δεδομένα σε κάθε στήλη είναι συνήθως ο ίδιος τύπος δεδομένων - π.χ. αριθμούς, χορδές, ημερομηνίες. Συνήθως, σε αντίθεση με ένα σύνολο δεδομένων excel, τα dataframes αποφεύγουν να έχουν ελλείπουσες τιμές και δεν υπάρχουν κενά και κενές τιμές μεταξύ γραμμών ή στηλών[8].

```
from bs4 import BeautifulSoup
import requests
import time
import re
from datetime import datetime, timedelta
import pandas as pd
from dirtyclean import clean
acropolis_df = pd.read_csv("acropolis_en.csv")
acropolis_df
```

	date	id	name	origin	review_text	star_no	title
0	2018-06-25	review_590722269	Brendan D	NaN	The Acropolis is for so many reasons a classic...	5	A 'must see in your lifetime' - historic site!
1	2018-06-25	review_590706610	MarcellaAquino	London, United Kingdom	Just for the incredible history of the place i...	4	Expensive, but worthwhile
2	2018-06-25	review_590660610	Tracey C	NaN	The Acropolis is amazing but you need to book ...	4	Amazing
3	2018-06-25	review_590629399	jetsetters000	Melbourne, Australia	We came early morning 9.30am before the heat h...	5	Awe inspiring
4	2018-06-25	review_590626769	carlia s	Red Deer, Canada	Visited the Acropolis with a large group tour....	5	Worth it! Take a tour guide though.
5	2018-06-25	review_590578881	Gary B	Boca Raton, Florida	After a long day of touring, we approsched the...	2	Acropolis refreshment stand is a rip off
6	2018-06-25	review_590567502	Emanuela R	London, United Kingdom	I visited Athens in June and went to the Acrop...	3	Symbol of Democracy, unfortunately no informat...

Εικόνα 4: Παράδειγμα Pandas DataFrame

[8] https://www.shanelynn.ie/using-pandas-dataframe-creating-editing-viewing-data-in-python/

5. ΜΕΘΟΔΟΛΟΓΙΑ

Στην παρούσα εργασία χρησιμοποιήθηκαν τεχνικές επεξεργασίας φυσικής γλώσσας για να εξαγάγουμε την πολικότητα του συναισθήματος των χρηστών σχετικά με πέντε καίρια μνημεία και αξιοθέατα της Αθήνας (Ακρόπολη, Παρθενώνας, Πλάκα, Μουσείο Ακρόπολης και Αρχαία Αγορά Αθηνών) και να τους δώσουμε ένα γενικό σκορ για όλες τις αξιολογήσεις. Τέλος, χρησιμοποιήθηκε η τεχνική ομαδοποίησης K-means (clustering) θέματος με σκοπό να ομαδοποιήσουμε ζευγάρια λέξεων που εμφανίζονται με μεγαλύτερη συχνότητα για κάθε σημείο μελέτης για να εξάγουμε συμπεράσματα σχετικά με αδυναμίες και πλεονεκτήματα που παρουσιάζονται από τους χρήστες μέσα από τις αξιολογήσεις τους για τα σημεία αυτά.

Στόχος της μελέτης μας είναι η ταξινόμηση των αξιολογήσεων των χρηστών με βάση το συναίσθημα που εκφράζεται σε αυτές και εν συνεχεία η εξόρυξη χαρακτηριστικών από το σώμα των δεδομένων μας. Προτιμήσαμε να εφαρμόσουμε μία προσέγγιση βασισμένη σε λεξικό (lexicon-based approach) αναφορικά με την ανάλυση συναισθήματος, ενώ για την εξόρυξη γνώμης χρησιμοποιήσαμε την μέθοδο k-means clustering. Ως εκ τούτου, ξεκινήσαμε συλλέγοντας το σύνολο δεδομένων μας. Επιλέγοντας μεμονωμένες λέξεις (unigrams) ως χαρακτηριστικά, ακολουθήσαμε συνοπτικά τις παρακάτω μεθόδους.

Στην πρώτη μέθοδο, ο αλγόριθμος ταξινόμησης υπολογίζει τη συχνότητα των μεμονωμένων λέξεων εφαρμόζοντας το TF-IDF μοντέλο (TF-IDF model bag-of-words), ενώ στη δεύτερη μέθοδο ο αλγόριθμος μετρά την εμφάνιση επιλεγμένων μεμονωμένων λέξεων, συγκεκριμένα των λέξεων που υπάρχουν στο λεξικό που χρησιμοποιήθηκε στην παρούσα εργασία (NRC-Emolex Lexicon), που εκφράζουν θετικά ή αρνητικό συναίσθημα αφού συνδυάσαμε την συχνότητα εμφάνισης των λέξεων με την συχνότητα εμφάνισης των συναισθηματικά σηματοδοτημένων λέξεων που υπάρχουν στο λεξικό Τέλος, εφαρμόσαμε την τεχνική ομαδοποίησης K-means (K-means clustering) για να ομαδοποιήσουμε σε 20 μεγάλες ομάδες με νοηματική συνάφεια τα μη δομημένα δεδομένα μας με στόχο να εκμαιεύσουμε βασικά σημεία αναφοράς των χρηστών για τα τουριστικά μέρη που μελετάμε.

Για την απόκτηση των δεδομένων μας, δηλαδή των αξιολογήσεων (reviews) από την πλατφόρμα, TripAdvisor χρειάστηκε να χρησιμοποιήσουμε την μέθοδο αυτόματης συλλογής δεδομένων (scraping) από τις ιστοσελίδες του TripAdvisor στα πέντε

γνωστά μνημεία και αξιοθέατα της Αθήνας: Ακρόπολη, Παρθενώνας, Πλάκα, Μουσείο Ακρόπολης και Αρχαία Αγορά Αθηνών, μέσω του Selenium Web Driver, της βιβλιοθήκης Beautiful Soup που προσφέρεται από την Python και της δημιουργίας κώδικα για την συλλογή συγκεκριμένων δεδομένων από την κάθε ιστοσελίδα του TripAdvisor.

Στην συνέχεια, αφού αποκτήθηκαν τα δεδομένα και προ-επεξεργάστηκαν σε μορφή τέτοια όπου να μπορούν να επεξεργαστούν, εφαρμόσαμε την τεχνική της ανάλυσης συναισθήματος στις αξιολογήσεις των χρηστών από το TripAdvisor για τα πέντε βασικά μνημεία και αξιοθέατα της Αθήνας με μια αρκετά απλή και αποτελεσματική μέθοδο. Οι βαθμολογίες του συναισθήματος του συνόλου των αξιολογήσεων μας είναι αποτέλεσμα των άμεσων υπολογισμών που προέρχονται από τις λέξεις που απαρτίζουν το σύνολο των αξιολογήσεων για κάθε πεδίο μελέτης, με τη χρήση του λεξικού συναισθήματος NRC Emotion Lexicon, που αναφέρθηκε σε προηγούμενη ενότητα (Lexicon-based classification). Με αυτό τον τρόπο καταφέραμε να εξάγουμε την συνολική πολικότητα όλων των αξιολογήσεων που συλλέχθηκαν και να βγάλουμε συμπεράσματα για την άποψη των χρηστών για κάθε μνημείο-αξιοθέατο ξεχωριστά.

Το μέγεθος των δεδομένων, δηλαδή το σύνολο των αξιολογήσεων (reviews) των χρηστών που χρησιμοποιήθηκαν, είναι 15.659 αξιολογήσεις για την Ακρόπολη, 6.691 για την Πλάκα, 7.306 για τον Παρθενώνα,18.030 για το Μουσείο της Ακρόπολης και 2.062 για την Αρχαία Αγορά των Αθηνών, όπως απεικονίζεται στον παρακάτω πίνακα.

Σημείο ενδιαφέροντος	**Σύνολο αξιολογήσεων**
Ακρόπολη	15.659
Πλάκα	6.691
Παρθενώνας	7.396
Μουσείο Ακρόπολης	18.219
Αρχαία αγορά Αθηνών	2.082
Σύνολο	50.047

Πίνακας 1: Αριθμός αξιολογήσεων που συλλέχθηκαν

Στην συνέχεια ακολουθήσαμε την K-means μέθοδο ομαδοποίησης των δεδομένων (clustering) για κάθε πεδίο μελέτης ξεχωριστά με στόχο να δημιουργήσουμε 20 ομάδες δεδομένων που να έχουν συνάφεια μεταξύ τους με στόχο να εντοπίσουμε τυχόν αδυναμίες και πλεονεκτήματα των μνημείων αυτών, που αναφέρονται συχνά από τους χρήστες.

5.1 ΑΝΑΚΤΗΣΗ ΔΕΔΟΜΕΝΩΝ (DATA RETRIEVAL)

Ο τύπος εγγράφων που περιλαμβάνει το σύνολο των δεδομένων (είτε πρόκειται για συγκριτικά δοκίμια, κριτικές, σχόλια, tweets κ.λπ.) μπορεί να έχει σοβαρές επιπτώσεις στην ανάλυση της πολικότητας του συναισθήματος τους από τη στιγμή που ορίζει το μέγεθος του εγγράφου και τη γλώσσα που έχει χρησιμοποιηθεί (αν είναι επίσημη ή ανεπίσημη μορφή γλώσσας, οι πολιτιστικές αναφορές που αναφέρονται, είτε υπάρχει σαρκαστική χροιά, είτε πρόκειται για μία μόνο γλώσσα είτε για μείγμα πολλών γλώσσες κ.λπ.).

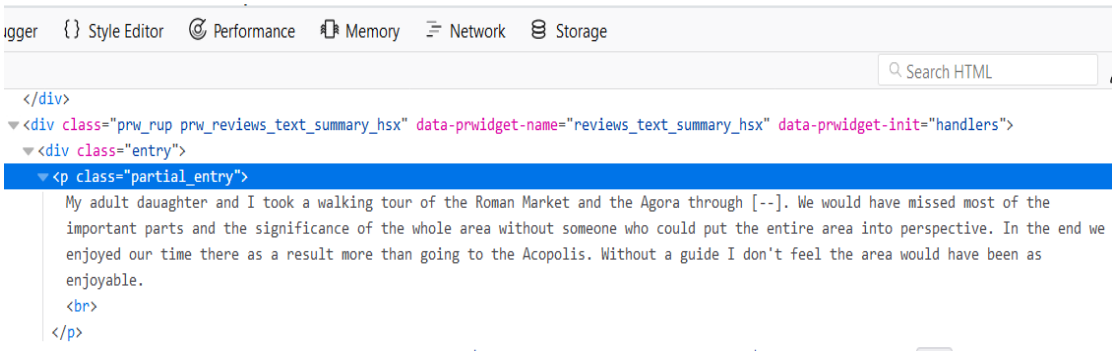
Η παρούσα διπλωματική βασίστηκε σε αξιολογήσεις χρηστών από ιστότοπους κοινωνικής δικτύωσης λόγω της δημοτικότητας αυτών και λόγω της δυνατότητας πρόσβασης σε πολλές χρήσιμες πληροφορίες που οι χρήστες δημοσιεύουν σε τέτοιου είδους ιστοσελίδες. Πιο συγκεκριμένα τα δεδομένα μας συλλέχθηκαν από την πλατφόρμα TripAdvisor, όπου ορίστηκε το πεδίο μελέτης μας (case study).

Χρησιμοποιώντας ένα λογισμικό αυτόματης ανάκτησης δεδομένων (Selenium), συγκεντρώθηκαν συνολικά 50.047 μη κατηγοριοποιημένες αξιολογήσεις/κριτικές σχετικά με τα πέντε σημεία πολιτισμικού και τουριστικού ενδιαφέροντος για την Αθήνα. Αυτές οι αξιολογήσεις περιλαμβάνουν γνώμες γραμμένες στην Αγγλική γλώσσα, από χρήστες κάθε εθνικότητας.

Πιο συγκεκριμένα για την ανάκτηση των δεδομένων μας, όπως αναφέρθηκε χρειάστηκε να χρησιμοποιήσουμε το λογισμικό Selenium, που ουσιαστικά είναι ένα πλαίσιο εργασίας για την αυτοματοποιημένη χρήση εφαρμογών ιστού. Χρησιμοποιώντας το Selenium, μπορεί να αυτοματοποιηθεί κάθε εργασία στο πρόγραμμα περιήγησής (browser) σαν να έπρεπε να εκτελέσει το έργο ένα πραγματικό άτομο. Στη δική μας περίπτωση, το Selenium χρησιμοποιήθηκε για να καλέσουμε το πρόγραμμα περιήγηση μας, που ορίστηκε, και να εκτελέσει τις εντολές

του κώδικα που γράφτηκε για να συλλέξει τα συγκεκριμένα δεδομένα που του ζητήθηκαν με την παράλληλη χρήση της βιβλιοθήκης Beautiful Soup της Python.

Στην συνέχεια ελέγχτηκε η γλώσσα σήμανσης του υπερκειμένου (HyperText Markup Language - HTML) για να κατανοηθεί η γλώσσα γραφής της ιστοσελίδας του TripAdvisor και να αναγνωριστεί ποια ακριβώς σημεία της χρειάζεται να ζητήσουμε από τον περιηγητή για να συλλέξουμε τα απαιτούμενα δεδομένα, όπως διαφαίνεται στην παρακάτω εικόνα.



Εικόνα 5: HTML TripAdvisor

Τα σημεία που συλλέχθηκαν για την εκπόνηση της συγκεκριμένης έρευνας είναι τα εξής: όνομα/ψευδώνυμο χρήστη, χώρα προέλευσης, ημερομηνία δημοσίευσης της αξιολόγησης, ο τίτλος της αξιολόγησης ,το περιεχόμενο της αξιολόγησης και τέλος η βαθμολόγηση του χρήστη για το συγκεκριμένο μνημείο ή τουριστικό σημείο (rating).

Στον κώδικα που αναπτύχθηκε, με τη χρήση της γλώσσας προγραμματισμού Python, ζητήθηκε η αυτόματη ανάκτηση των προαναφερθεισών σημείων για κάθε σελίδα του κάθε τουριστικού σημείου, δημιουργώντας έτσι μια λούπα μέσα στην λίστα με τα ζητηθέντα σημεία, μέχρι να ανακτήσουμε όλα τα δεδομένα μας. Η εικόνα 3 δείχνει τον αρχικό κώδικα που χρησιμοποιήθηκε για την ανάκτηση των δεδομένων μας.

Επιπλέον, στην πρώτη δοκιμή της ανάκτησης των δεδομένων μας, παρουσιάστηκαν προβλήματα που έπρεπε να επιλυθούν για να συλλέξουμε όσο πιο ολοκληρωμένη πληροφορία. Για παράδειγμα, πολλοί χρήστες δεν έδιναν την χώρα προέλευσης τους, όποτε έπρεπε να συμπληρώσουμε τα κενά ως μεταβλητές που λείπουν (missing values) με την ένδειξη ‘null’. Σε αυτό το σημείο είναι σημαντικό να σημειωθεί ότι η πλατφόρμα TripAdvisor δεν θεωρεί ως υποχρεωτικό πεδίο την καταχώρηση της χώρας ή της περιοχής του χρήστη. Επιπρόσθετα για να μπορέσουμε να ομαδοποιήσουμε τα αποτελέσματα μας ανά χώρα χρειάστηκε να διαχωρίσουμε την

πληροφορία της περιοχής προέλευσης, δηλαδή την πόλη και την χώρα, σε όσους χρήστες έδιναν την συγκεκριμένη πληροφορία.

Ένα καίριο πρόβλημα που κληθήκαμε να αντιμετωπίσουμε είναι η μη εμφάνιση όλου του περιεχομένου των αξιολογήσεων, πράγμα που αυτομάτως σήμαινε την ελλιπή συλλογή πληροφορίας άρα και έλλειψη έγκυρου αποτελέσματος. Η λύση δόθηκε με την χρήση του λογισμικού Selenium και της εντολής που δώσαμε στο περιηγητή να 'κλικάρει' την ένδειξη 'more',για να ανοίγει ολόκληρο το περιεχόμενο της εκάστοτε αξιολόγησης χρήστη. Τέλος, μετατρέψαμε την μορφή των αστεριών (rating), από δεκάδες στη κλίμακα 1 έως 5 για την ευκολότερη επεξεργασία των δεδομένων.

```
for review in reviews_num:
    reviews_dict= {}
    review_counter+=1
    print('\nProcessing review No. '+str(review_counter))
    review_id = review.get("id",{"class":"reviewSelector"})  #den 3erw t exei ginei me lir
    reviews_dict['id']= review_id
        #print('ID: '+review_id)
    name=review.find("span", {"class":"expand_inline scrname"}).text
    reviews_dict['name']= name
        #print('User name: '+name)
    title= review.find("span", {"class":"noQuotes"}).text
    reviews_dict['title']= title
        #print('Title: '+title)
    review_text= review.find("p", {"class":"partial_entry"}).text
    review_text = re.sub(r"\n", " ", review_text)
    reviews_dict['review_text']= review_text
        #print('Review content: '+review_text)
    try:
        origin = review.find("span", {"class":"expand_inline userLocation"}).text
    except AttributeError:
        origin = 'null'
    reviews_dict['origin']= origin
        #print('User origin: '+origin)
    date = review.find("span", {"class":"ratingDate relativeDate"}).text
        #print('Date: '+date) #an leei review yesterday prepei na metatrepeis thn plhrofor
    date = dateconverter(date)
        #print('Converted Date: '+date)
    reviews_dict['date']= date
    star_rating=review.find("span", {"class":"ui_bubble_rating"})
        #print('Star rating: '+str(star_rating))
    star_no = ratingconverter(star_rating)
    reviews_dict['star_no']= star_no
        #print('Stars: '+star_no)
```

Εικόνα 6: Αρχική μορφή κώδικα για την ανάκτηση των δεδομένων από το TripAdvisor

Στη συνέχεια, η γλώσσα σήμανσης της ιστοσελίδας της πλατφόρμας TripAdvisor άλλαξε και καθώς τα δεδομένα μας δεν συλλέχθηκαν όλα το ίδιο διάστημα, χρειάστηκε να τροποποιήσουμε τον κώδικα μας για να συμπεριλάβουμε τα νέα στοιχεία. Για παράδειγμα, η μορφή της ημερομηνίας άλλαξε και έτσι δεν χρειάστηκε να χρησιμοποιήσουμε πια μέσα στο κορμό του κώδικα την μέθοδο (def) για απαλοιφή λέξεων και μετατροπή των μηνών σε μορφή αριθμών. Η διαδικασία αυτή πλέον επιτυγχάνεται με την μορφή εντολών κατά την επεξεργασία των δεδομένων στο επόμενο στάδιο ανάλυσης που ακολουθεί.

```
    reviews_dict= {}
    review_counter+=1
    print('\nProcessing review No. '+str(review_counter))
    review_id = review.get("id",{"class":"reviewSelector"})  #den 3erw t exei ginei me li
    reviews_dict['id']= review_id
        #print('ID: '+review_id)
    nametag=review.find("span",{"class":"expand_inline scrname"}) #!!!!!!!
    if nametag is None: #!!!!!!!
        name = review.find("div",{"class":"info_text"}).text #!!!!!!!
    else: #!!!!!!!
        name = nametag.text #!!!!!!!
    reviews_dict['name']= name
        #print('User name: '+name)
    title= review.find("span", {"class":"noQuotes"}).text
    reviews_dict['title']= title
        #print('Title: '+title)
    review_text= review.find("p", {"class":"partial_entry"}).text
    review_text = re.sub(r"\n", " ", review_text)
    reviews_dict['review_text']= review_text
        #print('Review content: '+review_text)
    try:
        origin = review.find("span", {"class":"expand_inline userLocation"}).text
    except AttributeError:
        origin = 'null'
    reviews_dict['origin']= origin
        #print('User origin: '+origin)
    datetag = review.find("span", {"class":"ratingDate relativeDate"}) #!!!!!!!
    if datetag is None: #!!!!!!!
        date = review.find("span", {"class":"ratingDate"})['title'] #!!!!!!!
    else: #!!!!!!!
        date = datetag['title'] #!!!!!!!
        #print('Date: '+date) #an leei review yesterday prepei na metatrepeis thn plhrofo
     date = dateconverter(date) #!!!!!!! removed now
        #print('Converted Date: '+date)
    reviews_dict['date']= date
    star_rating=review.find("span", {"class":"ui_bubble_rating"})
        #print('Star rating: '+str(star_rating))
    star_no = ratingconverter(star_rating)
    reviews_dict['star_no']= star_no
        #print('Stars: '+star_no)
```

Εικόνα 7: Τελική μορφή κώδικα για την ανάκτηση των δεδομένων από το TripAdvisor

Αφού τελειοποιήσαμε τον κώδικα μας και συλλέξαμε τα δεδομένα στην επιθυμητή για επεξεργασία μορφή τους, δημιουργήσαμε μιας μορφής λεξικού (dictionary) για την περαιτέρω επεξεργασία των δεδομένων μας και εισαγάγαμε την βιβλιοθήκη Pandas της Python, που θα αναλυθεί παρακάτω, για να αποθηκεύσουμε τα δεδομένα μας σε μορφή αρχείου csv (comma separated value).

Οι αξιολογήσεις που συλλέχθηκαν, όπως δείχνει ο πίνακας 2 είναι στο σύνολο τους 50.047 για όλα τα τουριστικά σημεία μελέτης και αναφέρονται στην πρώτη δημοσίευση αξιολόγησης μέχρι τον Ιούλιο του 2018. Επειδή εξετάζονται διαφορετικά το καθένα, οι διαφοροποιήσεις που παρατηρούνται στις χρονολογίες δεν αποτελεί εμπόδιο στην μελέτη τους.

Σημείο ενδιαφέροντος	Σύνολο αξιολογήσεων	Ημερομηνία πρώτης αξιολόγησης	Ημερομηνία τελευταίας αξιολόγησης
Ακρόπολη	15.659	7-06-2005	26-06-2018
Πλάκα	6.691	10-04-2012	29-07-2018
Παρθενώνας	7.396	19-1-2006	29-07-2018
Μουσείο Ακρόπολης	18.219	21-06-2009	29-07-2018
Αρχαία Αγορά Αθηνών	2.082	30-09-2005	29-07-2018

Πίνακας 2: Χρονικά διαστήματα συλλογής δεδομένων

5.2 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Ένα μέρος της υλοποίησης αυτού του έργου ήταν αφιερωμένο στην προ-επεξεργασία των δεδομένων. Παρακάτω, παρέχουμε μια λεπτομερή περιγραφή των προκλήσεων που αντιμετωπίσαμε και των βημάτων που ακολουθήσαμε για την αντιμετώπιση τους.

Κατά την προετοιμασία των δεδομένων, τα δεδομένα κειμένου προ-επεξεργάζονται και το μη δομημένο κείμενο μετατρέπεται σε δομημένη μορφή. Μία αξιολόγηση μπορεί να περιέχει πολλές απόψεις για τα δεδομένα, οι οποίες εκφράζονται με διαφορετικούς τρόπους από τους διάφορους χρήστες, οπότε είναι απαραίτητη η όσο το δυνατόν αναλυτικότερη προετοιμασία τους.

Αρχικά, εφόσον ανακτήθηκαν τα δεδομένα και αποθηκεύτηκαν σε μορφή αρχείου csv, όπως ήδη αναφέρθηκε παραπάνω, πραγματοποιήθηκε ένας χειροκίνητος έλεγχος τους για να δούμε τυχόν αδυναμίες και ελλείψεις στην μορφή τους. Για παράδειγμα εντοπίστηκαν στην στήλη με την χώρα προέλευσης (origin) του χρήστη ότι λείπουν δεδομένα, πράγμα που ορίσαμε ως missing values, δηλαδή μεταβλητές που λείπουν και τις αντικαταστήσαμε με την ένδειξη ‘null’. Παρόλα αυτά οι αξιολογήσεις που

περιείχαν missing values χρησιμοποιήθηκαν στην μελέτη μας γιατί ενώ λείπει ένα στοιχείο, μπορούν να δώσουν πληροφορία σχετικά με τη γενική πολικότητα των τουριστικών μνημείων που αναλύουμε μέσω των σχετικών αξιολογήσεων/κριτικών των χρηστών.

Στη συνέχεια, όλα οι χαρακτήρες στο σύνολο των δεδομένων μετατρέπονται σε πεζά γράμματα. Το κείμενο χωρίζεται σε μια ακολουθία από σύμβολα (tokens), που το καθένα αποτελείται από μία μόνο λέξη. Στη συνέχεια όλα τα σύμβολα που ισοδυναμούν με stop words της αγγλικής γλώσσας, των ήδη ενσωματωμένων στην λίστα των stop words του συστήματος λέξεων αναφορικά με τα αγγλικά, καταργούνται από το κείμενο. Οι εξαιρούμενες αυτές λέξεις είναι λέξεις που προκαλούν σύγχυση και αυξάνουν το επίπεδο σφάλματος της ταξινόμησης στα νέα δεδομένα. Τέλος, πραγματοποιείται αποκοπή των καταλήξεων (stemming) με τον αλγόριθμο αποκοπής καταλήξεων Porter εφαρμόζοντας μια επαναληπτική, βασισμένη σε κανόνες αντικατάσταση των επιθημάτων των λέξεων με στόχο να μειώσει το μήκος των λέξεων στο ελάχιστο επιτρεπτό μήκος (Porter,1980). Τα σύμβολα (tokens) χρησιμοποιούνται για τη δημιουργία διανυσμάτων λέξεων (word vectors) αναπαριστώντας αριθμητικά κάθε παράδειγμα και η βαθμολογία TF-IDF κάθε διαθέσιμης λέξης υπολογίζεται .

Τα αποτελέσματα από την προ-επεξεργασία είναι υπό μορφή ενός όρου μήτρας εγγράφου (term document matrix), όπου κάθε σύμβολο αποτελεί τώρα ένα χαρακτηριστικό σε μια κάθετη στήλη και η καθένα αξιολόγηση χρήστη (review) είναι ένα παράδειγμα σε μια οριζόντια σειρά. Οι τιμές στα κελιά είναι η υπολογισμένη TF-IDF βαθμολογία για κάθε λέξη στη διαδικασία δημιουργίας του διανύσματος των λέξεων. Τα παραγόμενα χαρακτηριστικά των λέξεων καθώς και οι βαθμολογίες TF-IDF χρησιμοποιούνται από τον ταξινομητή για να προχωρήσουμε περαιτέρω στην εξόρυξη του κειμένου και της ανάλυσης συναισθήματος.

- **Απομάκρυνση συχνών όρων (Stop words removal)**

Καταργούμε τις λέξεις για την εξάλειψη όρων που είναι πολύ συνηθισμένοι στην αγγλική γλώσσα (π.χ. ‘and’, ‘this’ και ‘is’). Χρησιμοποιούμε την τυπική λίστα των Stop Words που παρέχονται από την βιβλιοθήκη της python.

- **Αποκοπή καταλήξεων (stemming)**

Το stemming συνήθως αναφέρεται στην ευριστική διαδικασία κατά την οποία οι καταλήξεις των λέξεων απομακρύνονται με στόχο να γίνει αυτό σωστά σε όσο το δυνατόν μεγαλύτερο μέρος των λέξεων που απαρτίζουν το κείμενο και συχνά περιλαμβάνει την απομάκρυνση των παράγωγων προσφυμάτων.

Στην διαδικασία της αποκοπής καταλήξεων, γίνεται η μετατροπή των μορφολογικών μορφών μιας λέξης στο θέμα της, υποθέτοντας ότι το καθένα συνδέεται σημασιολογικά. Το θέμα δεν χρειάζεται να είναι μια υπάρχουσα λέξη στο λεξικό, αλλά όλες οι παραλλαγές του θα πρέπει να αντιστοιχούν σε αυτή την μορφή μετά την ολοκλήρωση της αποκοπής της κατάληξης.

Στην υπολογιστική γλωσσολογία, το θέμα αποτελεί το τμήμα εκείνο της λέξης που δεν αλλάζει ποτέ ακόμη και όταν μορφολογικά κλίνεται ενώ το λήμμα είναι η βασική μορφή του ρήματος. Η απομάκρυνση των καταλήξεων είναι συνήθως πιο εύκολο να εφαρμοστεί και να λειτουργήσει ταχύτερά σε μεγάλα αδόμητα δεδομένα κειμένου και η μειωμένη ακρίβεια που ενδέχεται να παρουσιάζει μερικές φορές μπορεί να μην έχει σημασία σε ορισμένες εφαρμογές.

- **TF-IDF (Term Frequency-Inverse Document Frequency)**

TF-IDF είναι μια τεχνική εξόρυξης κειμένου που χρησιμοποιείται για την κατηγοριοποίηση εγγράφων (Martineau και Finin, 2009). Τα αρχικά TF-IDF αντιπροσωπεύουν τις έννοιες term frequency, δηλαδή τον όρο συχνότητας και inverse term frequency, δηλαδή την αντίστροφη συχνότητα όρου και είναι μια μέθοδος που δίνει έμφαση στις λέξεις που εμφανίζονται συχνά σε ένα συγκεκριμένο κείμενο, ενώ στο την ίδια στιγμή αποκλείει τις λέξεις που δεν εμφανίζονται συχνά σε πολλά κείμενα (Ramos, 2003).

Το αποτέλεσμα από την εφαρμογή μόνο του TF είναι ότι οι συχνές λέξεις έχουν υψηλότερη βαθμολογία TF και οι σπάνιες λέξεις - χαμηλότερη βαθμολογία TF. TF-IDF λαμβάνει υπόψη όχι μόνο τη σημασία μιας λέξης σε ένα δεδομένο κείμενο αλλά και της σημασίας του σε ολόκληρο το σώμα του κειμένου. Αυτή η τεχνική μειώνει το βάρος των συχνών λέξεων και αυξάνει αυτό των σπάνιων λέξεων σε ένα σώμα. Η βαθμολογία TF-IDF της μιας λέξης αυξάνεται όταν ο αριθμός των φορών εμφάνισης της σε ένα έγγραφο (TF) αυξάνεται. Εάν ο αριθμός των εγγράφων που περιέχει μια

λέξη είναι αυξημένος (η λέξη εμφανίζεται πιο συχνά στο σώμα), η βαθμολογία TF-IDF της λέξης μειώνεται, διαφορετικά αυξάνεται.

```
%%time
from sklearn.feature_extraction.text import TfidfVectorizer
from textblob import TextBlob
#diwxame to textblob giati argei
def textblob_tokenizer(str_input):
    blob = TextBlob(str_input.lower())
    tokens = blob.words
    words = [token.stem() for token in tokens]
    return words

# Vectorize and save into a new dataframe
vec = TfidfVectorizer(stop_words='english',
                      use_idf=True)

# Fit from the 'text' column of our dataframe
matrix = vec.fit_transform(acropolis_df['review_text'])

# Then turn it into a new dataframe
results = pd.DataFrame(matrix.toarray(), columns=vec.get_feature_names())
```

```
CPU times: user 907 ms, sys: 1.75 s, total: 2.66 s
Wall time: 2.72 s
```

Εικόνα 8: Παράδειγμα χρήσης του μοντέλου TF-IDF

5.3 ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Μετά το στάδιο του scraping, δηλαδή της ανάκτησης των δεδομένων προς μελέτη και της προ-επεξεργασίας των δεδομένων αυτών , η εργασία ακλουθεί τα εξής βήματα: α) πραγματοποίηση ανάλυση συναισθήματος με χρήση λεξικού για τον υπολογισμό της συνολικής πολικότητας (θετική ή αρνητική) των σχολίων που αφήνουν οι χρήστες για κάθε τουριστικό σημείο μελέτης προς εξέταση, ανάλυση επιμέρους θεματικών σχετικά με την χώρα προέλευσης των χρηστών, την εποχικότητα των σχολίων/αξιολογήσεων, την χρονολογική τοποθέτηση και αυξομείωση των αξιολογήσεων αυτών σε όλο το διάστημα που εξετάζουμε, ανάλυση επιμέρους συναισθημάτων που εκφράζονται μέσα στις αξιολογήσεις των χρηστών, σύγκριση της βαθμονόμησης που δίνεται από τους χρήστες στην πλατφόρμα TripAdvisor (rating) σε σχέση με την πολικότητα που εξήχθη από την ανάλυση συναισθήματος β) ομαδοποίηση λέξεων (K-means clustering) με νοηματική συνάφεια με την μέθοδο K-means.

- **Ανάλυση συναισθήματος με χρήση λεξικού (Lexicon-based sentiment analysis)**

Όπως ήδη έχει αναφερθεί η ανάλυση με βάση το λεξικό λειτουργεί με τη σύγκριση των λέξεων σε ένα κείμενο ή ένα σώμα κειμένων σε σχέση με προκαθορισμένα

λεξικά λέξεων. Αυτά τα λεξικά βασίζονται σε θετικές και αρνητικές λέξεις. Κάθε λεξικό περιέχει μια λίστα λέξεων που ο χρήστης πιστεύει ότι αντιστοιχεί σε μια συγκεκριμένη ποιότητα, όπως είναι η θετικότητα (positivity) ή η αίσθηση λύπης (sadness).

Η παρούσα διπλωματική επικεντρώνεται στην αξιολόγηση της θετικότητας (positivity) και αρνητικότητας (negativity) των αξιολογήσεων των χρηστών από το TripAdvisor αναφορικά με την Ακρόπολη, την Πλάκα, τον Παρθενώνα, το Μουσείο της Ακρόπολης και την Αρχαία Αγορά των Αθηνών. Το πρώτο βήμα είναι η εισαγωγή του λεξικού μας και η διανυσματοποίηση των λέξεων που το απαρτίζουν σε λίστες θετικών / αρνητικών λέξεων και εν συνεχεία η εισαγωγή του αρχείο csv με τα δεδομένα προς ανάλυση.

Στη μέθοδο αυτή που βασίζεται σε λεξικό, οι λεκτικές μονάδες (unigrams) που βρίσκονται στο λεξικό περιλαμβάνουν μια βαθμολογία πολικότητας από 0 έως 1 και η συνολική βαθμολογία πολικότητας του κειμένου προς ανάλυση υπολογίζεται στη συνέχεια ως άθροισμα της πολικότητας των λεκτικών μονάδων που εντοπίστηκαν στο κείμενο προς ανάλυση, στη συγκεκριμένη περίπτωση δηλαδή οι αξιολογήσεις/κριτικές των χρηστών από το TripAdvisor.

Κατά συνέπεια, στη δική μας προσέγγιση, εισάγουμε το λεξικό συναισθημάτων (NRC-Emotion Lexicon) για να εντοπίσουμε το συναίσθημα στα δεδομένα από τις αξιολογήσεις, όπου κάθε λέξη στο λεξικό έχει βάρος που αντιπροσωπεύει την πολικότητα 0 έως 1 για κάθε θετική και αρνητική λέξη αντίστοιχα. Εάν η λέξη δεν υπάρχει στο λεξικό, η πολικότητα της είναι μηδέν.

Στη συνέχεια, γίνεται χρήσης της TF-IDF τεχνικής δηλαδή, η τεχνική αυτή δίνει έμφαση στις λέξεις που εμφανίζονται συχνά στο συγκεκριμένο σώμα κειμένου που έχουμε εισάγει, ενώ την ίδια στιγμή αποκλείει τις λέξεις που δεν εμφανίζονται συχνά. Τέλος, μετά τον συνδυασμό των δύο παραπάνω τεχνικών, εξάγεται η πολικότητας κάθε λέξης από το αγγλικό λεξικό και συγκεντρώνονται τα συνολικά μεγέθη και δημιουργείται η πολικότητα ολόκληρου του εισαγόμενου κειμένου.

Αφού λοιπόν εντοπίσαμε την συνολική πολικότητα των αξιολογήσεων για κάθε τουριστικό σημείο ξεχωριστά, στη συνέχεια προχωρήσαμε σε επιμέρους αναλύσεις μέσω των δυνατοτήτων των εντολών της βιβλιοθήκης pandas της python.

Αναφορικά με την χώρα προέλευσης των χρηστών, αν και αντιμετωπίσαμε δυσκολίες καθώς το TripAdvisor δεν θεωρεί την καταχώρηση της χώρας προέλευσής του χρήστη ως υποχρεωτικό πεδίο, ούτε οριοθετεί τις χώρες και τις περιοχές, σε πολλές αξιολογήσεις υπήρχαν ελλείψεις ή σε κάποιες περιπτώσεις η ίδια περιοχή εμφανιζόταν με διαφορετικά τρόπο, όπως παραδείγματος χάριν η τοποθεσία 'Athens' με την τοποθεσία 'Athens, Greece, που αναφέρεται στον ίδιο τόπο με διαφορετικό τρόπο, συγκεντρώθηκαν οι τοποθεσίες με την μεγαλύτερη συχνότητα εμφάνισης αξιολογήσεων.

Αναλύθηκαν επίσης η εποχικότητα και η χρονολογική διάχυση των αξιολογήσεων ανά μήνα και ανά χρονιά, δίνοντας μας αποτελέσματα για τις περιόδους με την εμφάνιση μεγαλύτερης συχνότητας σχολίων για τις τοποθεσίες που εξετάστηκαν ξεχωριστά. Στην συνέχεια εξετάστηκαν τα επιμέρους συναισθήματα που δίνονται από το λεξικό που χρησιμοποιήσαμε ενώ τέλος συγκρίναμε την γενική βαθμονόμηση (rating) που δίνουν οι ίδιοι οι χρήστες για κάθε τουριστικό σημείο που επισκέφτηκαν, που κυμαίνεται στην κλίμακα από 1 (χαμηλότερη βαθμολογία) μέχρι 5 (υψηλότερη βαθμολογία), σε σχέση με το συνολικό βαθμό πολικότητας που εντοπίστηκε από την ανάλυση του συναισθήματος.

- **Ομαδοποίηση (Clustering)**

Η μέθοδος ομαδοποίησης (clustering) μπορεί να χρησιμοποιηθεί για την εύρεση ομάδων εγγράφων με παρόμοιο περιεχόμενο. Το αποτέλεσμα της ομαδοποίησης είναι συνήθως μια διαίρεση, ονομάζεται επίσης ομαδοποίηση P, ένα σετ των ομάδων P. Κάθε ομάδα αποτελείται από μια σειρά εγγράφων d. Αντικείμενα – στη δική μας περίπτωση έγγραφα κειμένων – μιας ομάδας θα πρέπει να είναι είτε παρόμοια είτε ανόμοια με τα έγγραφα άλλων ομάδων. Συνήθως η ποιότητα των ομάδων θεωρείται καλύτερη εάν το περιεχόμενο των έγγραφων μέσα σε κάθε ομάδα (cluster) είναι παρόμοια και μεταξύ των διαφορετικών ομάδων περισσότερο ανόμοια. Οι μέθοδοι ομαδοποίησης ομαδοποιούν τα έγγραφα μόνο λαμβάνοντας υπόψη τη διάχυση τους μέσα στο χώρο του εγγράφου (Hotho, Nurnberger, Fraunhofer, 2005) .

Οι αλγόριθμοι ομαδοποίησης υπολογίζουν τις ομάδες (clusters) με βάση τα χαρακτηριστικά των δεδομένων και τον βαθμό ομοιότητας και ανομοιότητας τους. Ωστόσο, η ιδέα του τι πρέπει να έχει ένα ιδανικό αποτέλεσμα ομαδοποίησης διαφέρει ανάμεσα στις ποικίλες εφαρμογές και μπορεί να διαφέρει ακόμη και μεταξύ των

χρηστών. Κάποιος μπορεί να ασκήσει επιρροή στα αποτελέσματα ενός αλγόριθμου ομαδοποίησης χρησιμοποιώντας μόνο υποσύνολα των χαρακτηριστικών ή προσαρμόζοντας τα χρησιμοποιούμενα μέτρα ομοιότητας και επομένως να ελέγχει την διαδικασία της ομαδοποίησης. Σε ποιο βαθμό το αποτέλεσμα του αλγόριθμου της ομαδοποίησης συμπίπτει με τις ιδέες του χρήστη μπορεί να αξιολογηθεί με μέτρα αξιολόγησης.

- **Ομαδοποίηση με τη μέθοδο K-means (K-means Clustering)**

Το k-means είναι ένας από τους πιο συχνά χρησιμοποιούμενους αλγόριθμους ομαδοποίησης στην πράξη στην τομέα της εξόρυξης δεδομένων και των στατιστικών (Duda και Pattern, 1973). Η διαδικασία που αρχικά προέρχεται από την στατιστική είναι απλή στην εφαρμογή και μπορεί επίσης να εφαρμοστεί σε μεγάλα σύνολα δεδομένων. Αποδείχθηκε ότι ειδικά στον τομέα της ομαδοποίησης κειμένων με την τεχνική K-means επιτυγχάνονται αρκετά έγκυρα αποτελέσματα (Hartigan, 1989).

K-means clustering είναι ένας τύπος μάθησης χωρίς επίβλεψη, ο οποίος χρησιμοποιείται όταν υπάρχουν μη κατηγοριοποιημένα δεδομένα, δηλαδή δεδομένα χωρίς καθορισμένες κατηγορίες ή ομάδες. Ο στόχος αυτού του αλγορίθμου είναι η εύρεση ομάδων στα δεδομένα, που συνδέεται με τον αριθμό των ομάδων που αντιπροσωπεύει η μεταβλητή K. Ο αλγόριθμος λειτουργεί επαναληπτικά για να εκχωρήσει κάθε σημείο δεδομένων σε μία από τις ομάδες K με βάση τα χαρακτηριστικά που παρέχονται. Τα σημεία δεδομένων συγκεντρώνονται σε ομάδες βάσει της ομοιότητας χαρακτηριστικών. Τα αποτελέσματα του αλγόριθμου ομαδοποίησης K-means είναι:

- Τα κέντρα βάρους (centroids) των ομάδων K, τα οποία μπορούν να χρησιμοποιηθούν για την επισήμανση νέων δεδομένων
- Ετικέτες για τα δεδομένα εκπαίδευσης (κάθε σημείο δεδομένων αντιστοιχεί σε ένα ενιαίο σύμπλεγμα)

Αντί να ορίζονται οι ομάδες πριν εξεταστούν τα δεδομένα, αυτή η μέθοδος ομαδοποίησης επιτρέπει να βρεθούν και να αναλυθούν οι ομάδες που έχουν σχηματιστεί με οργανικό τρόπο.

Κάθε κέντρο βάρους (centroid) μίας ομάδας είναι μια συλλογή από τιμές χαρακτηριστικών που ορίζουν τις ομάδες που προκύπτουν. Η εξέταση του βάρους

των χαρακτηριστικών γνωρισμάτων μπορεί να χρησιμοποιηθεί για την ποιοτική ερμηνεία του είδους της ομάδας που αντιπροσωπεύει κάθε ομάδα (cluster).

Προχωρώντας από μια αρχική εκδοχή στην οποία όλα τα έγγραφα διανέμονται σε ένα δεδομένο αριθμό ομάδων (clusters) γίνεται προσπάθεια για βελτίωση της εκδοχής αυτής με συγκεκριμένη αλλαγή της κατανομής των εγγράφων στις ομάδες. Εν τω μεταξύ, υπάρχει ένα σύνολο παραλλαγών σχετικά με την μέθοδο αυτή, ενώ οι βασικές αρχές πηγαίνουν πίσω στον Forgy 1965 ή στον MacQueen 1967 . Στη βιβλιογραφία η κβαντοποίηση K-means (ector quantization k-means) είναι επίσης γνωστή με το όνομα Lloyd Max Algorithm (Gersho και Gray, 1992).

Στην παρούσα εργασία η ομαδοποίηση K-means ουσιαστικά αποτελείται από τα παρακάτω βήματα του αλγορίθμου που τον αποτελεί, όπου ο αριθμός των ομάδων K πρέπει να δοθεί. Στο επόμενο βήμα, εκχωρούνται τα δεδομένου κειμένου στο πλησιέστερο από τα K-centroids (που ονομάζεται επίσης πρωτότυπο της ομαδοποίησης). Στη συνέχεια , ο αλγόριθμος υπολογίζει ένα νέο κέντρο με βάση τις νέες κατανομές. Τα δύο προηγούμενα βήματα επαναλαμβάνονται μέσα σε μια λούπα έως ότου τα κεντροειδή του συμπλέγματος δεν αλλάξουν πλέον. Ο αλγόριθμος αυτός, ουσιαστικά αντιστοιχεί σε μια απλή διαδικασία αναρρίχησης, η οποία τυπικά σταματάει να λειτουργεί όταν βρεθεί σε ένα βέλτιστο σημείο, που δεν επιδέχεται αλλαγές στο σύνολο των ομάδων. Το να βρεθεί το συνολικό βέλτιστο είναι ένα λυμένο πρόβλημα στην επεξεργασία φυσικής γλώσσας (natural language processing).

Λόγω του γεγονότος ότι τα δεδομένα που εξετάζονται στην συγκεκριμένη εργασία είναι ένα μεγάλο σύνολο μη κατηγοριοποιημένων δεδομένων κειμένου με μεγάλο εύρος απόψεων καθώς και σύνολο λέξεων, ορίσαμε την ένδειξη K = 20, έπειτα από αρκετές δοκιμές μέχρι να μην παρουσιάζονται πια μεγάλες διαφορές ανάμεσα στις ομάδες (clusters) που δημιουργήθηκαν. Με αυτό τον τρόπο προσπαθήσαμε να δημιουργήσουμε, 20 μεγάλες ομάδες με νοηματική συνάφεια, που να αποδίδουν κάποιο αποτέλεσμα, δίνοντας μας πληροφορίες σχετικά με κοινά προβλήματα ή πλεονεκτήματα που αναφέρουν οι χρήστες μετά την επίσκεψη τους στα πέντε βασικά σημεία μελέτης.

6. ΕΥΡΗΜΑΤΑ ΕΡΕΥΝΑΣ

Παρακάτω αναλύονται λεπτομερώς τα ευρήματα που εντοπίστηκαν από την ανάλυση των δεδομένων που συλλέχθηκαν για κάθε τουριστικό μνημείο ξεχωριστά. Συνοπτικά, σχετικά με την ανάλυση συναισθήματος των αξιολογήσεων και για τα πέντε σημεία μελέτης, η πολικότητα τους κρίθηκε ως θετική, όπως ήταν το αναμενόμενο αποτέλεσμα, βασιζόμενη στην συνολική βαθμονόμηση που δίνεται από την πλατφόρμα TripAdvisor. Αυτό που αξίζει να σημειωθεί είναι ότι ενώ ο γενικός μέσος όρος βαθμολογίας των χρηστών (rating) μέσα στην πλατφόρμα κυμαίνεται από 4 μέχρι 4.8 στα 5 αστέρια, ποσοστό που αγγίζει δηλαδή το 80% με 96%, η θετική συναισθηματική πολικότητα που εντοπίστηκε κυμαίνεται στο 25% στην κλίμακα από 0 έως 1.

Στη συνέχεια αναλύθηκαν επιμέρους θεματικές σχετικά με την προέλευση των χρηστών, δηλαδή από ποιες περιοχές έγιναν οι περισσότερες κριτικές και κατά συνέπεια, από ποια περιοχή δέχεται η Αθήνα τους περισσότερους επισκέπτες της, σχετικά με την συχνότητα εμφάνισης αξιολογήσεων ανά μήνα και ανά χρονιά, ενώ τέλος συγκρίναμε τον μέσο όρο της θετικής και αρνητικής πολικότητας κάθε σημείου μελέτης με την βαθμολογία των ίδιων των χρηστών από το TripAdvisor (star rating system).

Αναφορικά με την εξόρυξη κειμένου πραγματοποιήθηκε η μέθοδος K-means clustering για κάθε τουριστικό σημείο προς ανάλυση ξεχωριστά, αφού χωρίστηκαν τα δεδομένα μας σε 20 υπό-ομάδες με νοηματική συνάφεια, δίνοντας μας συμπεράσματα για την θεματολογία που συχνά αναφέρεται από τους ίδιους τους χρήστες μέσα στο σώμα των αξιολογήσεων τους.

Παρακάτω παρουσιάζονται τα αποτελέσματα για κάθε τουριστικό σημείο της Αθήνας ξεχωριστά:

6.1 ΑΚΡΟΠΟΛΗ

- **Ανάλυση Συναισθήματος**

Η μέτρηση της πολικότητας του σώματος κειμένου των αξιολογήσεων των χρηστών (review_text column, όπως ορίστηκε στο data frame) για το τουριστικό μνημείο της Ακρόπολης κρίθηκε θετική, με μέσο όρο στο σύνολο των αξιολογήσεων (15.659)

0.25/1 σε σύγκριση με τον μέσο όρο της αρνητικότητας που παρουσιάζουν οι αξιολογήσεις αυτές, όπως βλέπεται και στον παρακάτω πίνακα. Η κλίμακα μέτρησης κυμαίνεται από το 0 (καθόλου θετική ή αρνητική πολικότητα) μέχρι το 1 (απόλυτη θετικότητα ή αρνητικότητα). Αντίθετα, ό μέσος όρος της αρνητικότητα που παρουσιάζουν οι αξιολογήσεις κυμαίνεται στο 0.07. Σε μορφή ποσοστού οι αξιολογήσεις των χρηστών που κρίθηκαν θετικές σε σύγκριση με το σύνολο των αρνητικών αγγίζουν το 79,1%.

Επιπλέον, συγκρίναμε την θετική πολικότητα που εντοπίστηκε από την ανάλυση, με το μέσο όρο της βαθμολογίας (rating) που δίνουν οι ίδιοι οι χρήστες για το κάθε μνημείο. Ο μέσος όρος της βαθμολογίας για την ακρόπολη είναι 4.6 στα 5 αστέρια, ποσοστό που αγγίζει το 92 %. Το ποσοστό αυτό δεν έχει μεγάλη απόκλιση από το 80,4 % της θετικής πολικότητας που εντοπίστηκε, παρόλα αυτά υπάρχει μια διαφορά της τάξης του 12 %, πράγμα που δείχνει ότι οι χρήστες έχουν την τάση να βαθμολογούν ευκολότερα σε σχέση με αυτά που περιγράφουν στις αξιολογήσεις τους.

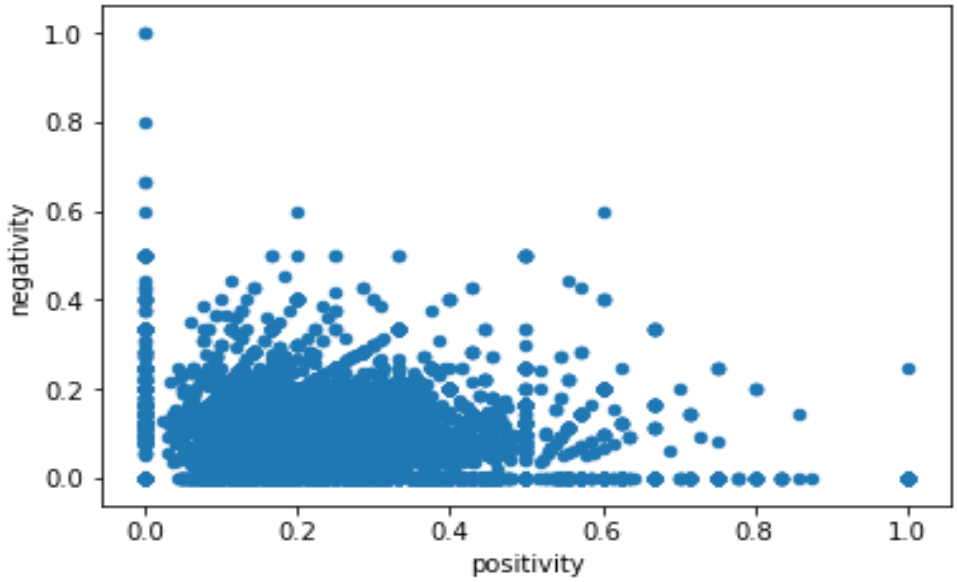
Positivity	0.25/1
Negativity	0.07/1
No of positive reviews	12.392
No of negative reviews	3.266
Percentage of positive reviews	79,1 %
Percentage of negative reviews	20.8 %

Πίνακας 3: Δεδομένα Ανάλυσης Συναισθήματος Ακρόπολης

count 15658.000000 mean 0.254911 std 0.150887 min 0.000000 25% 0.159091 50% 0.250000 75% 0.333333 max 1.000000 Name: positivity, dtype: float64	count 15658.000000 mean 0.073429 std 0.083846 min 0.000000 25% 0.000000 50% 0.058824 75% 0.125000 max 1.000000 Name: negativity, dtype: float64

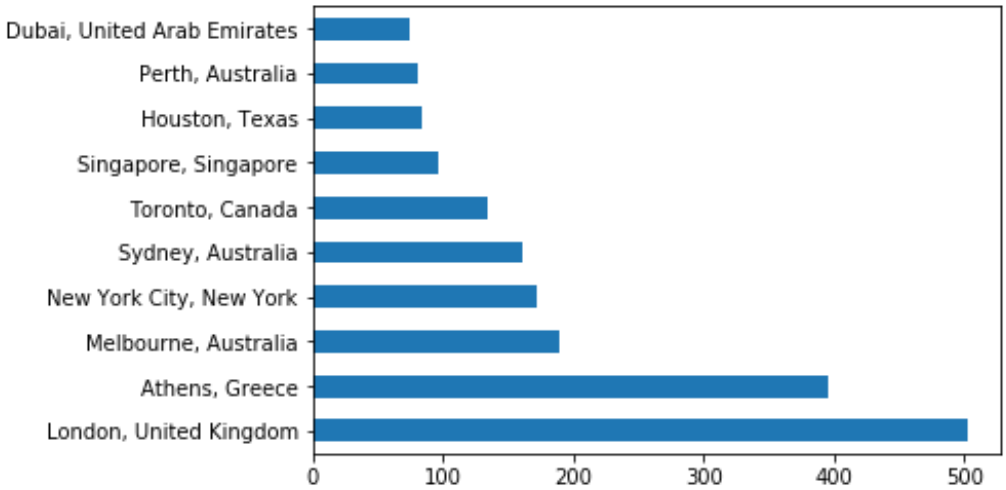
Πίνακας 4: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

Στο παρακάτω διάγραμμα μπορεί να διακριθεί η διάχυση των πολικοτήτων για τα δεδομένα που συλλέχτηκαν για το μνημείο της Ακρόπολης. Είναι φανερό ότι τα δεδομένα παρουσιάζουν μεγαλύτερη συγκέντρωση προς την θετική πολικότητα.



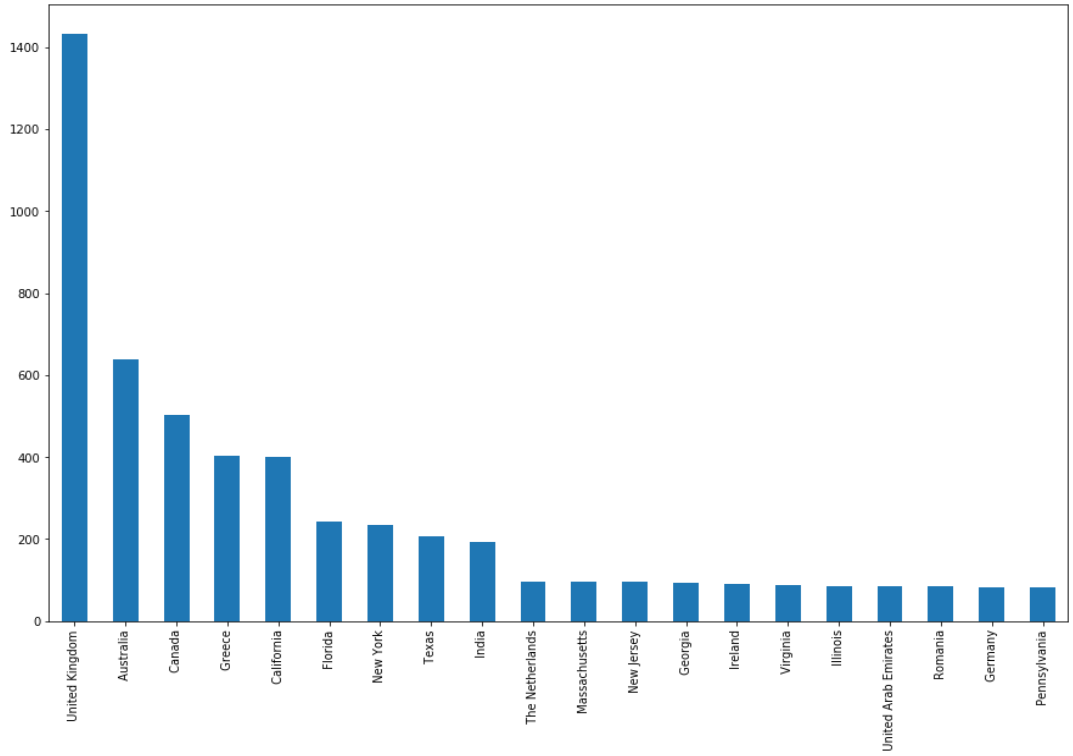
Εικόνα 9: Διάγραμμα διάχυσης πολικότητας για την Ακρόπολη

Αναφορικά με την προέλευση καταγωγής των χρηστών και την συχνότητα εμφάνισης σχολίων από συγκεκριμένες χώρες εντοπίστηκε ότι χρήστες από το Ηνωμένο Βασίλειο και συγκεκριμένα το Λονδίνο εμφανίζονται συχνότερα μέσα στο σύνολο δεδομένων μας, πράγμα που υποδηλώνει ότι οι Άγγλοι υπήκοοι επιλέγουν την Ελλάδα σε μεγαλύτερη συχνότητα από τις υπόλοιπες εθνικότητες, όπως διακρίνεται και από το παρακάτω γράφημα. Στην συνέχεια ακολουθούν σχόλια από χρήστες με τόπο προέλευσης την Αθήνα, την Μελβούρνη, την Νέα Υόρκη , το Σύδνεϋ, το Τορόντο, την Σιγκαπούρη, το Χιούστον, το Περθ της Αυστραλίας και τα Ενωμένα Αραβικά Εμιράτα.

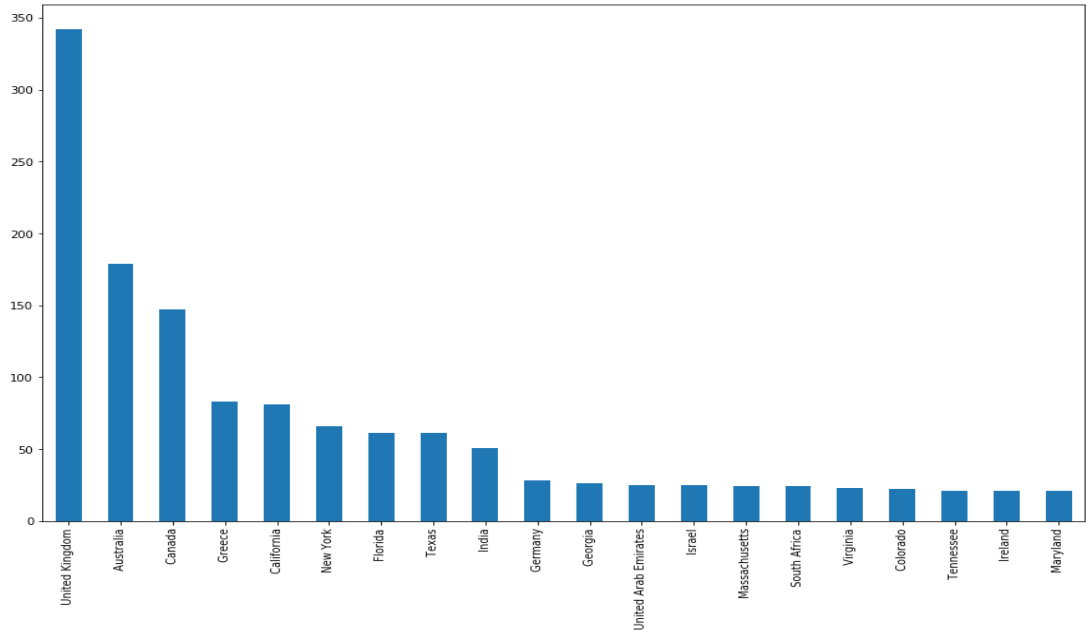


Εικόνα 10: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

Τα πιο θετικά σχόλια εντοπίστηκε ότι προέρχονται από την Αγγλία, την Αυστραλία, τον τον Καναδά, την Ελλάδα και την Καλιφόρνια, ενώ και αντιστοίχως τις πιο αρνητικές αξιολογήσεις παρατηρήθηκε ότι λήφθηκαν από τις ίδιες περιοχές με την ίδια ακριβώς σειρά, όπως διαφαίνεται στα παρακάτω γραφήματα.



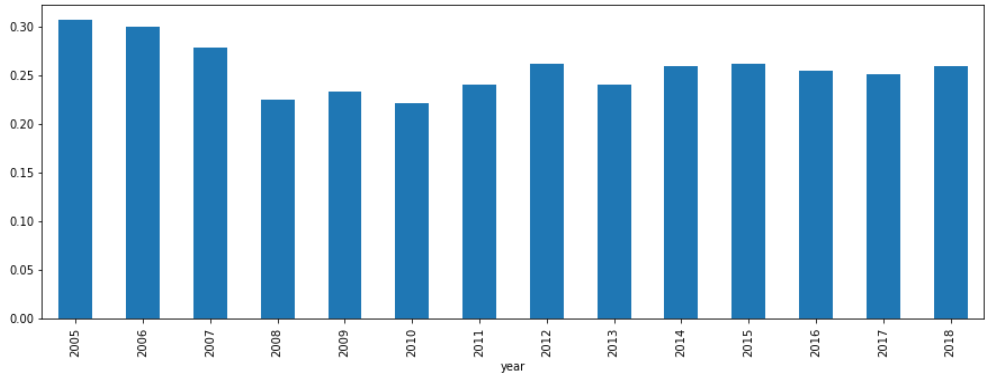
Εικόνα 11: Περιοχές με τον μεγαλύτερο αριθμό εμφάνισης θετικών σχολίων



Εικόνα 12: Περιοχές με τον μεγαλύτερο αριθμό εμφάνισης αρνητικών σχολίων

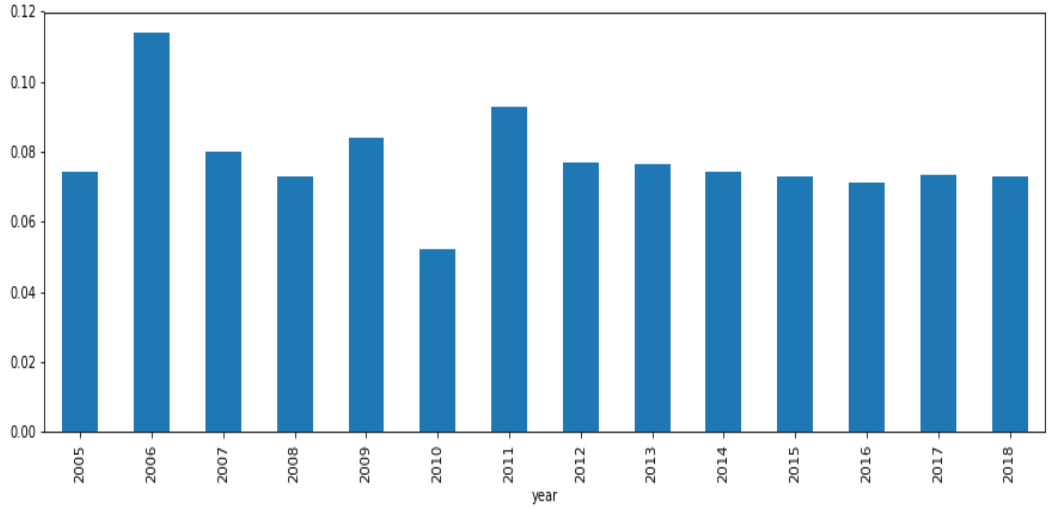
Στη συνέχεια αναλύθηκαν οι αξιολογήσεις που κρίθηκαν με θετική πολικότητα ανά χρονολογία, με αποτέλεσμα να εντοπιστεί ότι οι χρονιές από το 2005 μέχρι το 2007 εμφανίζουν τα περισσότερα θετικά σχόλια όπως διαφαίνεται στο παρακάτω διάγραμμα, και ιδιαίτερα η χρονιά 2005, αποτελεί την χρονιά με την μεγαλύτερη

εμφάνιση θετικών σχολίων, που αποτελεί ένα λογικό αποτέλεσμα καθώς είναι η χρονιά μετά τους Ολυμπιακούς Αγώνες 2004 που πραγματοποιήθηκαν στην Ελλάδα.



Εικόνα 13: Κατανομή θετικών σχολίων ανά χρονιά

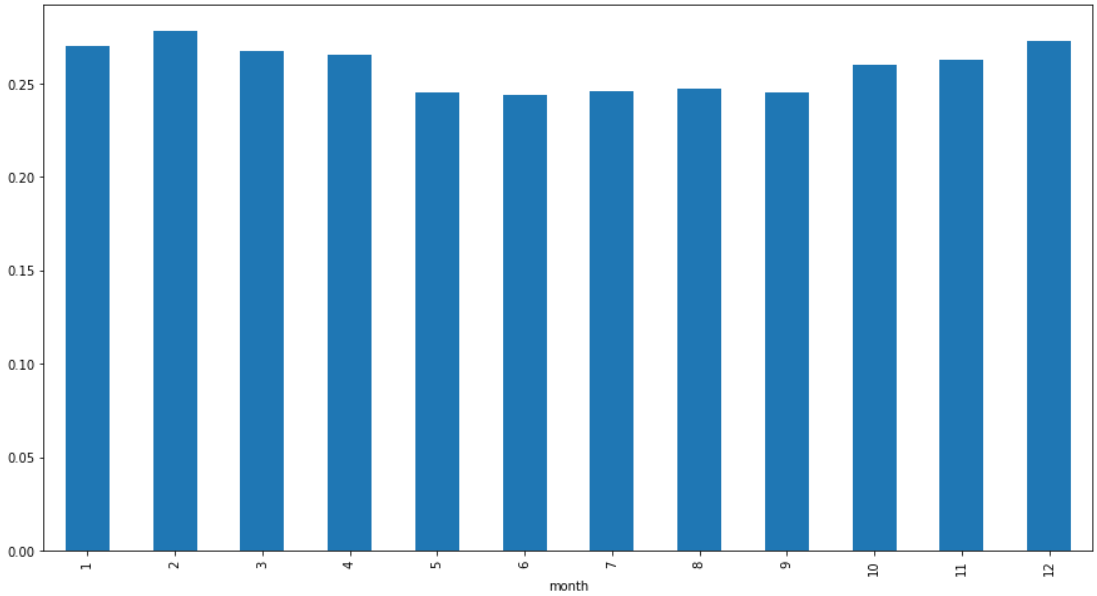
Αναφορικά με τις χρονιές που παρατηρήθηκε ο μεγαλύτερος αριθμός αρνητικών σχολίων αυτές είναι το 2006, το 2011 και το 2009, όπως διαφαίνεται στο γράφημα που ακολουθεί.



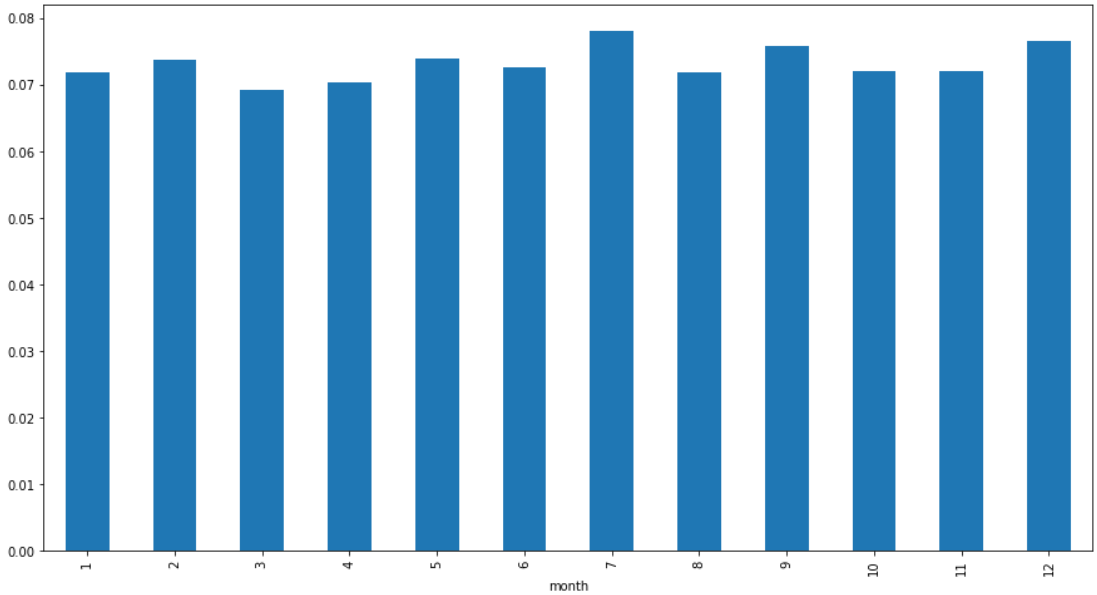
Εικόνα 14: Κατανομή αρνητικών σχολίων ανά χρονιά

Σχετικά με τον ανά μήνα αριθμό εμφάνισης αξιολογήσεων από τους χρήστες της πλατφόρμας TripAdvisor, εντοπίστηκε ότι οι μήνες κατά τους οποίους παρατηρείται μεγαλύτερη συχνότητα σχολίων είναι ο Φεβρουάριος, ο Δεκέμβριος, ο Ιανουάριος και ο Νοέμβριος. Το αποτέλεσμα αλλάζει όσον αφορά τις αρνητικές αξιολογήσεις, όπως μπορεί να διαφανεί στην εικόνα 15 και 16 αντίστοιχα, με τους μήνες Ιούλιο,

Δεκέμβριο και Σεπτέμβριο να εμφανίζουν τον μεγαλύτερο αριθμό αρνητικών σχολίων.



Εικόνα 15: Μέσος όρος θετικών αξιολογήσεων ανά μήνα



Εικόνα 16: Μέσος όρος αρνητικών αξιολογήσεων ανά μήνα

Επιπλέον το συναισθηματικό λεξικό που έχει εισαχθεί για την πραγμάτωση της ανάλυσης συναισθήματος πέρα από την πολικότητα, θετική η αρνητική, που μπορεί να εκφράσει, δίνει την δυνατότητα εντοπισμού και οχτώ επιπλέον συναισθημάτων (εμπιστοσύνη, έκπληξη, λύπη, χαρά, αηδία, προσδοκία, φόβος και θυμό).

Αναφορικά με την Ακρόπολη, τα συναισθήματα που εμφανίζουν μεγαλύτερη ένταση μέσα στο σώμα του κειμένου είναι αυτό της εμπιστοσύνης και της προσδοκίας, με το αίσθημα της εμπιστοσύνης να έχει μεγαλύτερο μέσο όρο σε σχέση με το αίσθημα της προσδοκίας (Πίνακας 5) , πράγμα που είναι αρκετά λογικό, αν αναλογιστεί κανείς ότι η συνολική πολικότητα των αξιολογήσεων είναι θετική.

Συναισθήματα	**Μέσος Όρος**
Εμπιστοσύνη (trust)	0.131
Έκπληξη (surprise)	0.049
Λύπη (sadness)	0.027
Χαρά (joy)	0.079
Αηδία(disgust)	0.013
Προσδοκία (anticipation)	0.111
Φόβος (fear)	0.028
Θυμός (anger)	0.027

Πίνακας 5: Μέσος Όρος Συναισθημάτων για Ακρόπολη

Πριν προχωρήσουμε την ομαδοποίηση των αξιολογήσεων για σκοπό της εύρεσης νοηματικής συνάφειας ανάμεσα στα σχόλια των χρηστών, πραγματοποιήθηκε μια απλή εύρεση των λέξεων που παρουσιάζουν μεγαλύτερη συχνότητα στα δεδομένα κειμένου που έχουμε εισάγει. Στο Πίνακα 6 εικονίζονται οι δέκα πιο συχνά εμφανιζόμενες λέξεις σχετικά με την Ακρόπολη.

Visit	468.531706
History	369.675746
Place	350.829992
Site	251.613893
Worth	251.376998
Time	240.601160
View	240.452688
Top	232.393511
Walk	208.306438
Early	189.995448

Πίνακας 6 : Οι 10 πιο συχνές λέξεις για την Ακρόπολη

- **Ομαδοποίηση (K-means Clustering)**

Τέλος πραγματοποιήθηκε η μέθοδος ομαδοποίησης κειμένου K-means όπως έχει ήδη αναφερθεί. Για να ορίσουμε την μεταβλητή K κατά την επεξεργασία των δεδομένων έγιναν αρκετές δοκιμές, μέχρι να βρεθεί η τιμή του K που δεν εμφανίζει άλλες ομάδες με διαφορετική νοηματική συνάφεια, από αυτές που δημιουργήθηκαν. Η τιμή που ορίστηκε είναι το K να ισούται με το είκοσι (K=20), καταλήγοντας στα ευρήματα της Εικόνας 15.

```
In [83]: print("Top terms per cluster:")
         order_centroids = km.cluster_centers_.argsort()[:, ::-1]
         terms = vec.get_feature_names()
         for i in range(number_of_clusters):
             top_ten_words = [terms[ind] for ind in order_centroids[i, :2]]
             print("Cluster {}: {}".format(i, ' '.join(top_ten_words)))

         Top terms per cluster:
         Cluster 0: acropolis athens
         Cluster 1: city view
         Cluster 2: climb worth
         Cluster 3: shoes wear
         Cluster 4: water bring
         Cluster 5: visit acropolis
         Cluster 6: history amazing
         Cluster 7: ticket sites
         Cluster 8: ancient athens
         Cluster 9: parthenon temple
         Cluster 10: early crowds
         Cluster 11: tour guide
         Cluster 12: loved acropolis
         Cluster 13: line tickets
         Cluster 14: experience amazing
         Cluster 15: years ago
         Cluster 16: beautiful athens
         Cluster 17: site athens
         Cluster 18: construction going
         Cluster 19: place visit
```

Εικόνα 17: Ομαδοποίηση λέξεων με την μέθοδο K-means

Συμπεραίνουμε μέσα από τις ομάδες που δημιουργήθηκαν ότι οι χρήστες είναι θετικά διακείμενοι για την Ακρόπολη, είναι εντυπωσιασμένοι με την θέα, την ιστορία καθώς και ότι αξίζει να βιώσει κάποιος αυτή την εμπειρία, δηλαδή την επίσκεψη στην Ακρόπολη.

Από την άλλη μεριά βέβαια εντοπίζονται και κάποια αρνητικά χαρακτηριστικά όπως είναι η ουρά αναμονής για τα εισιτήρια, ότι υπάρχει πάρα πολύ κόσμος και ότι η Ακρόπολη βρίσκεται ακόμη υπό κατασκευή.

6.2 ΠΛΑΚΑ

- **Ανάλυση Συναισθήματος**

Η μέτρηση της πολικότητας του σώματος κειμένου των αξιολογήσεων των χρηστών για το τουριστικό μνημείο της Πλάκας κρίθηκε επίσης θετική, με μέσο όρο στο σύνολο των αξιολογήσεων (6.691) 0.29/1 σε σύγκριση με τον μέσο όρο της αρνητικότητας που παρουσιάζουν οι αξιολογήσεις αυτές, όπως βλέπεται και στον παρακάτω πίνακα. Η κλίμακα μέτρησης κυμαίνεται από το 0 (καθόλου θετική ή

αρνητική πολικότητα) μέχρι το 1 (απόλυτη θετική ή αρνητική). Αντίθετα, ό μέσος όρος των αρνητικών αξιολογήσεων που παρουσιάζουν τα δεδομένα κυμαίνεται στο 0.06. Σε μορφή ποσοστού οι αξιολογήσεις των χρηστών που κρίθηκαν θετικές σε σύγκριση με το σύνολο των αρνητικών αγγίζουν το 79,7 %.

Επιπλέον, συγκρίναμε την θετική πολικότητα που εντοπίστηκε από την ανάλυση, με το μέσο όρο της βαθμολογίας (rating) που δίνουν οι ίδιοι οι χρήστες για το κάθε μνημείο. Ο μέσος όρος της βαθμολογίας για την ακρόπολη είναι 4.5 στα 5 αστέρια, ποσοστό που αγγίζει το 90 %. Το ποσοστό αυτό δεν έχει μεγάλη απόκλιση από το 79,6 % της θετικής πολικότητας που εντοπίστηκε, παρόλα αυτά υπάρχει μια διαφορά της τάξης του 10,4 %, πράγμα που δείχνει ότι οι χρήστες έχουν την τάση να βαθμολογούν με μεγαλύτερη επιείκεια σε σχέση με αυτά που περιγράφουν στις αξιολογήσεις τους.

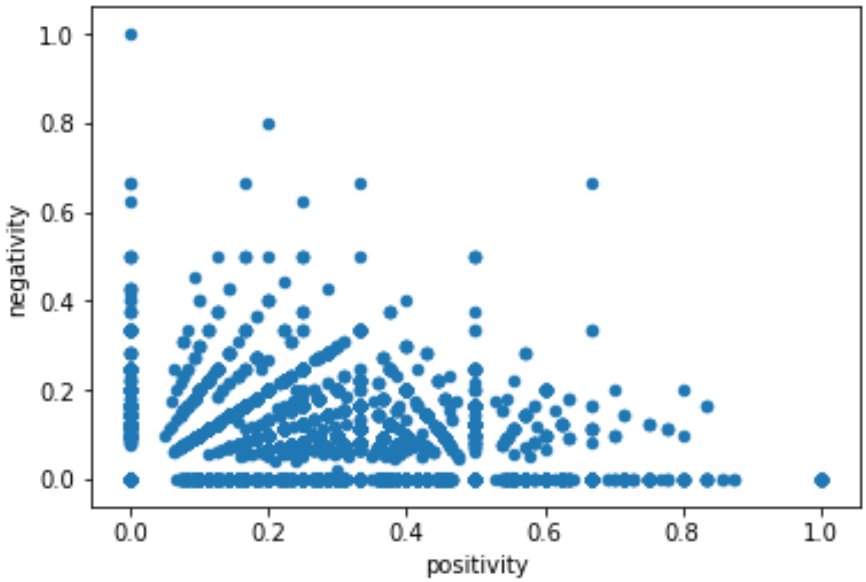
Positivity	0.29/1
Negativity	0.06/1
No of positive reviews	5.335
No of negative reviews	1.335
Percentage of positive reviews	79,7 %
Percentage of negative reviews	20,2 %

Πίνακας 7: Δεδομένα Ανάλυσης Συναισθήματος για την Πλάκα

count 6690.000000 mean 0.287353 std 0.168896 min 0.000000 25% 0.166667 50% 0.272727 75% 0.400000 max 1.000000 Name: positivity, dtype: float64	count 6690.000000 mean 0.065175 std 0.091280 min 0.000000 25% 0.000000 50% 0.000000 75% 0.111111 max 1.000000 Name: negativity, dtype: float64

Πίνακας 8: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

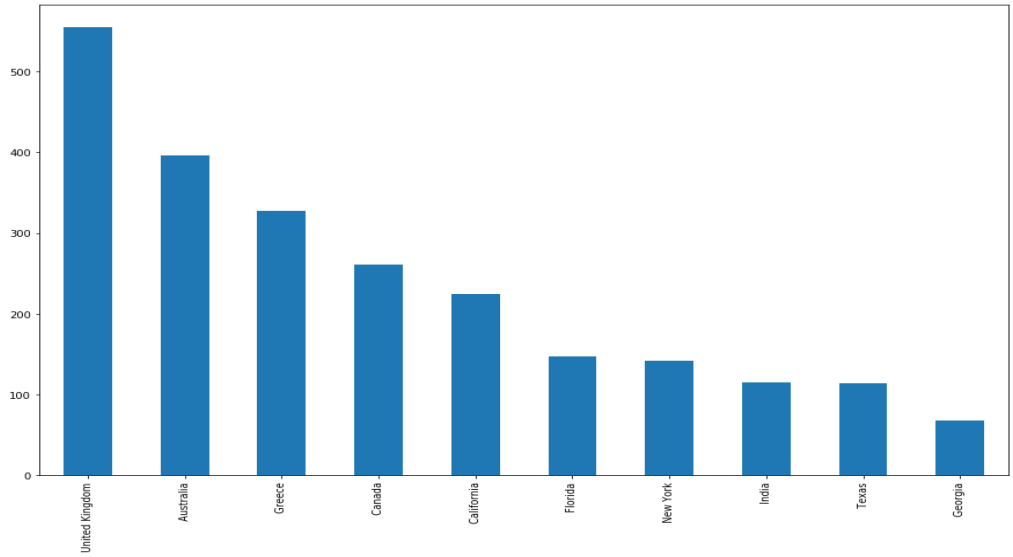
Στο παρακάτω διάγραμμα μπορεί να διακριθεί η διάχυση της πολικότητας για τα δεδομένα που συλλέχτηκαν για το μνημείο της Πλάκας. Είναι φανερό ότι τα δεδομένα παρουσιάζουν μεγαλύτερη συγκέντρωση προς την θετική πολικότητα.



Εικόνα 18: Διάγραμμα διάχυσης πολικότητας για την Πλάκα

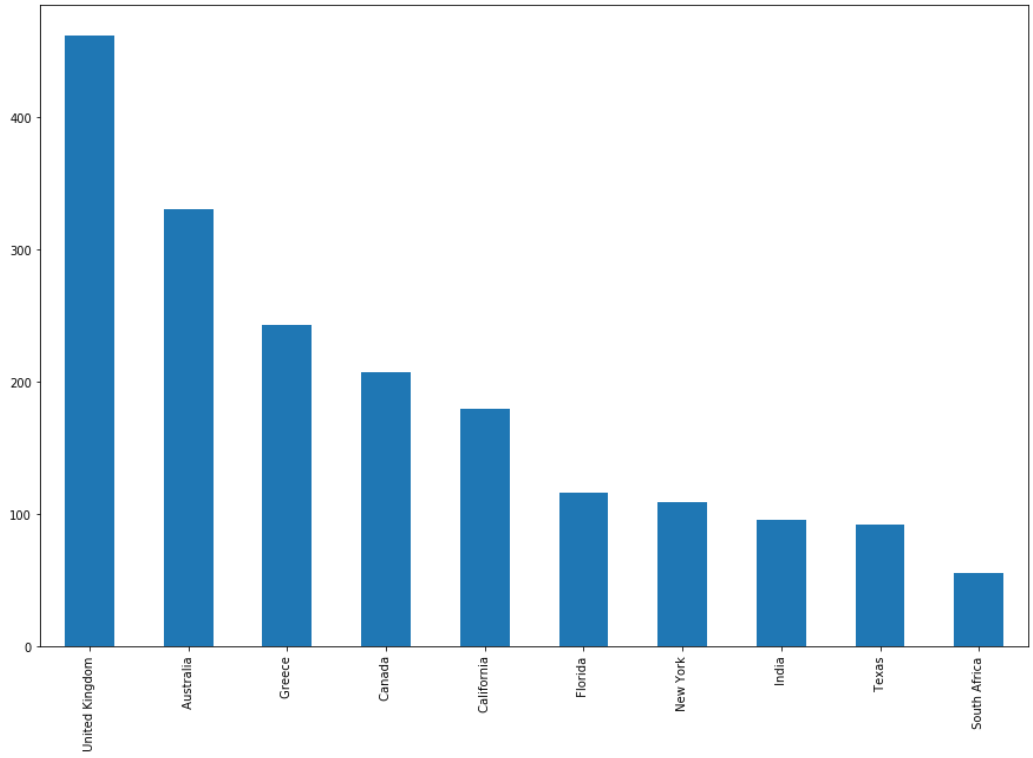
Αναφορικά με την προέλευση καταγωγής των χρηστών και την συχνότητα εμφάνισης σχολίων από συγκεκριμένες χώρες εντοπίστηκε ότι χρήστες από το Ηνωμένο Βασίλειο εμφανίζονται συχνότερα μέσα στο σύνολο δεδομένων μας, πράγμα που

υποδηλώνει ότι οι Άγγλοι υπήκοοι επιλέγουν την Ελλάδα σε μεγαλύτερη συχνότητα από τις υπόλοιπες εθνικότητες, όπως διακρίνεται και από το παρακάτω γράφημα. Στην συνέχεια ακολουθούν σχόλια από χρήστες με τόπο προέλευσης την Αυστραλία, την Ελλάδα, τον Καναδά , την Καλιφόρνια, την Φλόριντα, τη Νέα Υόρκη, την Ινδία, το Τέξας και τη Γεωργία.

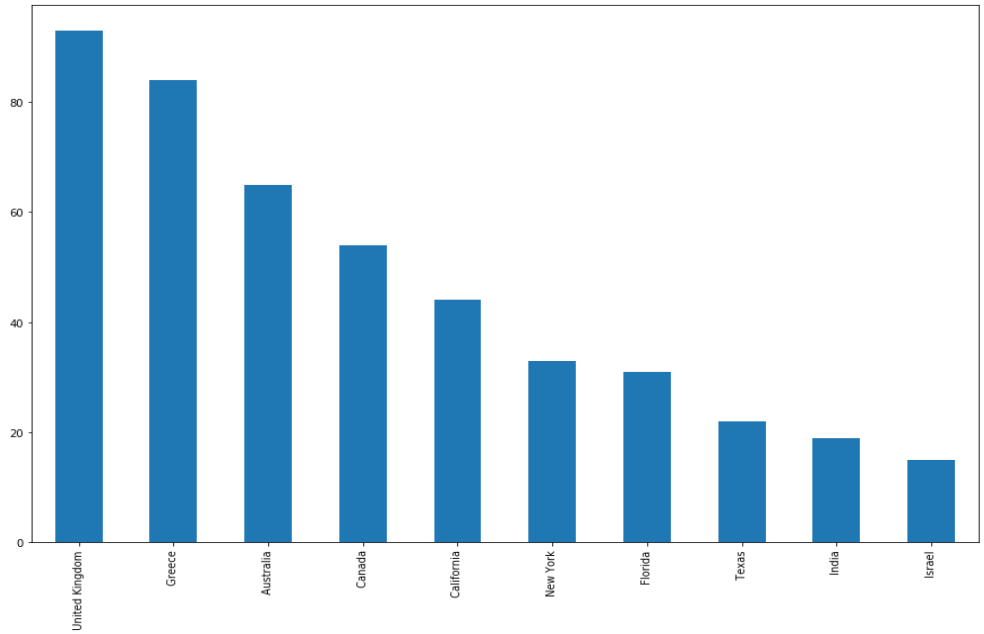


Εικόνα 19: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

Τα πιο θετικά σχόλια εντοπίστηκαν επίσης ότι προέρχονται από τις προαναφερθείσες χώρες καθώς και οι πιο αρνητικές αξιολογήσεις παρατηρήθηκε ότι λήφθηκαν από τις ίδιες χώρες με μια εξαίρεση της χώρας του Ισραήλ, όπως διαφαίνεται στα παρακάτω γραφήματα.



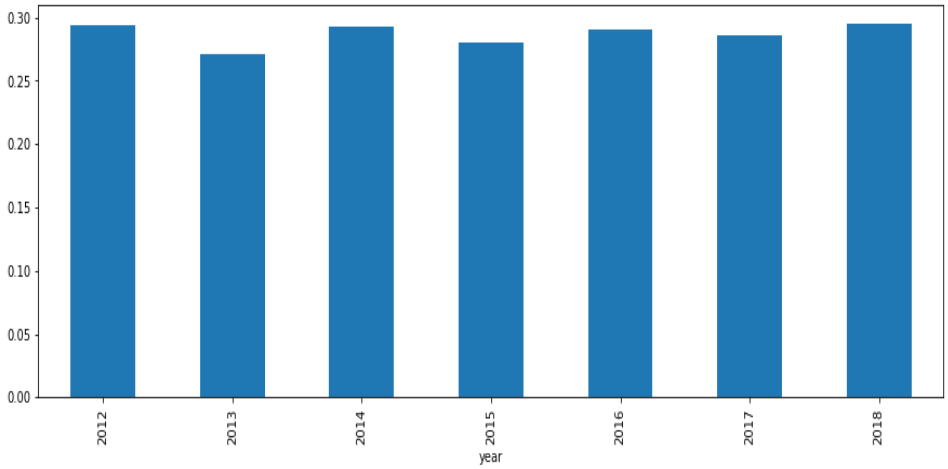
Εικόνα 20: Περιοχές με τον μεγαλύτερο αριθμό εμφάνισης θετικών σχολίων



Εικόνα 21: Περιοχές με τον μεγαλύτερο αριθμό εμφάνισης αρνητικών σχολίων

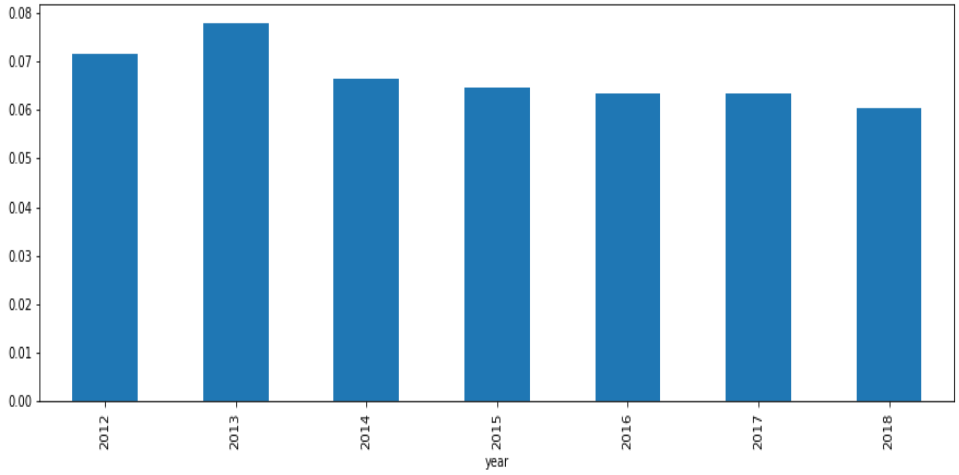
Στη συνέχεια αναλύθηκαν οι αξιολογήσεις που κρίθηκαν με θετική πολικότητα ανά χρονολογία, με αποτέλεσμα να εντοπιστεί ότι οι χρονιές 2018, 2014 και 2016 εμφανίζουν τα περισσότερα θετικά σχόλια όπως διαφαίνεται στο παρακάτω

διάγραμμα, και ιδιαίτερα η χρονιά 2018, αποτελεί την χρονιά με την μεγαλύτερη εμφάνιση σχολίων.



Εικόνα 22: Κατανομή θετικών σχολίων ανά χρονιά

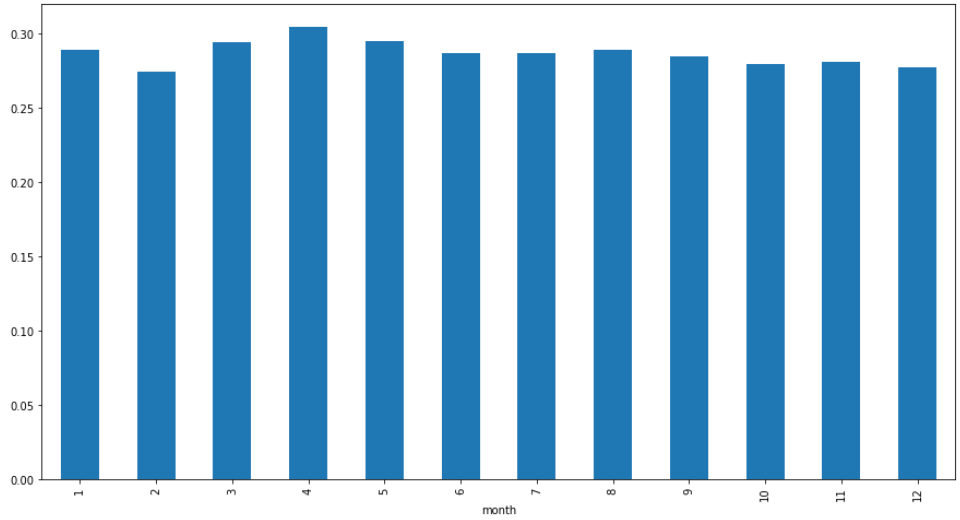
Όσον αφορά τις αξιολογήσεις με αρνητικό πρόσημα ανά χρονιά παρατηρήθηκε ότι το 2013 συγκέντρωσε τα υψηλότερα επίπεδα αρνητικών αξιολογήσεων, με το 2012 και το 2014 να ακολουθούν.



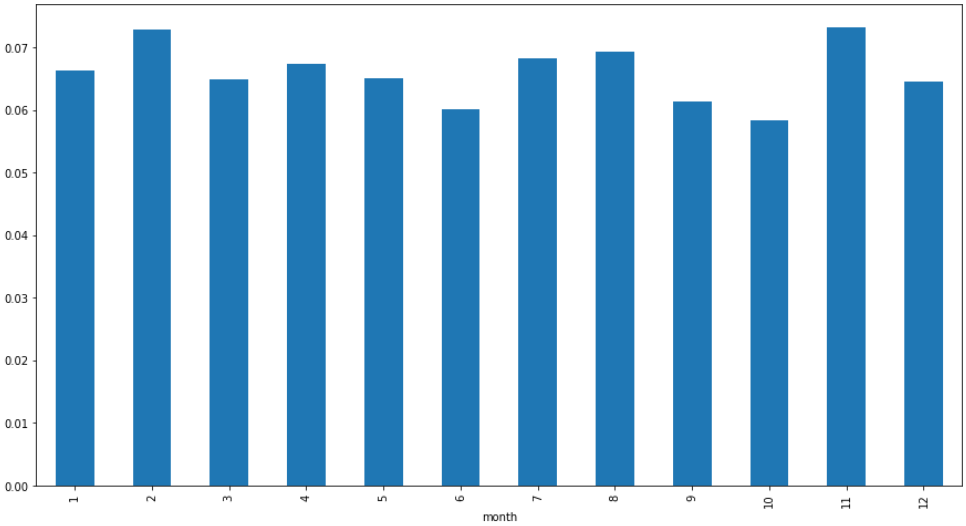
Εικόνα 23: Κατανομή αρνητικών σχολίων ανά χρονιά

Σχετικά με την ανά μήνα συχνότητα εμφάνισης θετικών αξιολογήσεων από τους χρήστες της πλατφόρμας TripAdvisor, εντοπίστηκε ότι οι μήνες Απρίλιος, Μάρτιος και Μάιος περιλαμβάνουν τα πιο θετικά σχόλια κατά μέσο όρο. Το αποτέλεσμα αυτό

διαμορφώνεται διαφορετικά για τις αρνητικές αξιολογήσεις με τους μήνες Νοέμβριο, Φεβρουάριο και Αύγουστο να κατέχουν τις πρώτες θέσεις όπως μπορεί να διαφανεί στην εικόνα 24 και 25 αντίστοιχα.



Εικόνα 24: Μέσος όρος θετικών αξιολογήσεων ανά μήνα



Εικόνα 25: Μέσος όρος αρνητικών αξιολογήσεων ανά μήνα

Επιπλέον το συναισθηματικό λεξικό που έχει εισαχθεί για την πραγμάτωση της ανάλυσης συναισθήματος πέρα από την πολικότητα, θετική η αρνητική, που μπορεί να εκφράσει, δίνει την δυνατότητα εντοπισμού και οχτώ επιπλέον συναισθημάτων, όπως έχει αναφερθεί. Τα συναισθήματα που εντοπίζονται είναι τα εξής: εμπιστοσύνη, έκπληξη, χαρά, αηδία, προσδοκία, φόβος και θυμό και διαμορφώνονται συμφώνα με τον παρακάτω πίνακα.

Αναφορικά με την Πλάκα, τα συναισθήματα που εμφανίζουν μεγαλύτερη ένταση μέσα στο σώμα του κειμένου είναι αυτό της χαράς, της εμπιστοσύνης και της προσμονής σε σύγκριση με τα αρνητικά συναισθήματα που προσφέρει το συναισθηματικό λεξικό που χρησιμοποιήθηκε (Πίνακας 7) , πράγμα που είναι αρκετά λογικό, αν αναλογιστεί κανείς ότι η συνολική πολικότητα των αξιολογήσεων είναι θετική, άρα και τα παραγόμενα συναισθήματα θα έχουν θετικό πρόσημο στο σύνολο τους.

Συναισθήματα	**Μέσος Όρος**
Εμπιστοσύνη (trust)	0.150
Έκπληξη (surprise)	0.073
Λύπη (sadness)	0.032
Χαρά (joy)	0.163
Αηδία (disgust)	0.011
Προσδοκία (anticipation)	0.131
Φόβος (fear)	0.020
Θυμός (anger)	0.015

Πίνακας 9: Μέσος Όρος Συναισθημάτων για την Πλάκα

Πριν προχωρήσουμε την ομαδοποίηση των αξιολογήσεων για σκοπό της εύρεσης νοηματικής συνάφειας ανάμεσα στα σχόλια των χρηστών, πραγματοποιήθηκε μια απλή εύρεση των λέξεων που παρουσιάζουν μεγαλύτερη συχνότητα στα δεδομένα κειμένου που έχουμε εισάγει. Στον παρακάτω πίνακα εικονίζονται οι δέκα πιο συχνά εμφανιζόμενες λέξεις σχετικά με την Πλάκα.

area	250.402731
place	213.387419
walk	142.211640
shopping	125.801691
lots	123.390850
good	118.052090
visit	101.826174
food	96.884306
people	95.646387
tourist	84.988762

Πίνακας 10: Οι 10 πιο συχνές λέξεις για την Πλάκα

- **Ομαδοποίηση (K-means Clustering)**

Τέλος πραγματοποιήθηκε η μέθοδος ομαδοποίησης κειμένου K-means όπως έχει ήδη αναφερθεί. Για να ορίσουμε την μεταβλητή K κατά την επεξεργασία των δεδομένων έγιναν αρκετές δοκιμές, μέχρι να βρεθεί η τιμή του K που δεν εμφανίζει άλλες ομάδες με διαφορετική νοηματική συνάφεια, από αυτές που δημιουργήθηκαν. Η τιμή που ορίστηκε είναι το K να ισούται με το είκοσι (K=20), καταλήγοντας στα ευρήματα της Εικόνας 26.

```
In [104]: print("Top terms per cluster:")
          order_centroids = km.cluster_centers_.argsort()[:, ::-1]
          terms = vec.get_feature_names()
          for i in range(number_of_clusters):
              top_ten_words = [terms[ind] for ind in order_centroids[i, :2]]
              print("Cluster {}: {}".format(i, ' '.join(top_ten_words)))

Top terms per cluster:
Cluster 0: old streets
Cluster 1: eat places
Cluster 2: people watching
Cluster 3: nice walk
Cluster 4: visit athens
Cluster 5: greek food
Cluster 6: spend hours
Cluster 7: city athens
Cluster 8: bars restaurants
Cluster 9: good restaurants
Cluster 10: night day
Cluster 11: small streets
Cluster 12: prices reasonable
Cluster 13: beautiful walk
Cluster 14: street shops
Cluster 15: plaka time
Cluster 16: acropolis close
Cluster 17: lots shops
Cluster 18: shops shop
Cluster 19: great athens
```

Εικόνα 26: Ομαδοποίηση λέξεων με την μέθοδος K-means

Συμπεραίνουμε μέσα από την ομάδες που δημιουργήθηκαν ότι οι χρήστες είναι θετικά διακείμενοι για την περιοχή της Πλάκας, δίνοντας έμφαση στα πολλά μικρά μαγαζιά που υπάρχουν είτε για φαγητό είτε για να κάνουν τις αγορές τους.

Επιπλέον, γίνεται αναφορά στα δρομάκια της Πλάκας, ως μιας όμορφης επιλογής για να κάνει κάποιος την βόλτα του και να περάσει την ώρα του από το πρωί έως και πιο αργά την νύχτα. Τέλος, αυτό που αξίζει να σημειωθεί είναι ότι οι χρήστες αναφέρουν ότι οι τιμές κυμαίνονται σε λογικά πλαίσια αναφορικά με την περιοχή της Πλάκας.

6.3 ΠΑΡΘΕΝΩΝΑΣ

- **Ανάλυση Συναισθήματος**

Η μέτρηση της πολικότητας του σώματος κειμένου των αξιολογήσεων των χρηστών για το τουριστικό μνημείο του Παρθενώνα κρίθηκε επίσης θετική, με μέσο όρο στο σύνολο των αξιολογήσεων (7.396) 0.27/1 σε σύγκριση με τον μέσο όρο της αρνητικότητας που παρουσιάζουν οι αξιολογήσεις αυτές, όπως βλέπεται και στον παρακάτω πίνακα. Η κλίμακα μέτρησης κυμαίνεται από το 0 (καθόλου θετική ή αρνητική πολικότητα) μέχρι το 1 (απόλυτη θετική ή αρνητική). Αντίθετα, ό μέσος όρος των αρνητικών αξιολογήσεων που παρουσιάζουν τα δεδομένα κυμαίνεται στο

0.06. Σε μορφή ποσοστού οι αξιολογήσεις των χρηστών που κρίθηκαν θετικές σε σύγκριση με το σύνολο των αρνητικών αγγίζουν το 77,4 %.

Επιπλέον, συγκρίναμε την θετική πολικότητα που εντοπίστηκε από την ανάλυση, με το μέσο όρο της βαθμολογίας (rating) που δίνουν οι ίδιοι οι χρήστες για το κάθε μνημείο. Ο μέσος όρος της βαθμολογίας για την ακρόπολη είναι 4.7 στα 5 αστέρια, ποσοστό που αγγίζει το 94 %. Το ποσοστό αυτό δεν έχει μεγάλη απόκλιση από το 77,4 % της θετικής πολικότητας που εντοπίστηκε, παρόλα αυτά υπάρχει μια διαφορά της τάξης του 16,6 %, πράγμα που δείχνει ότι οι χρήστες έχουν την τάση να βαθμολογούν με μεγαλύτερη επιείκεια σε σχέση με αυτά που περιγράφουν στις αξιολογήσεις τους.

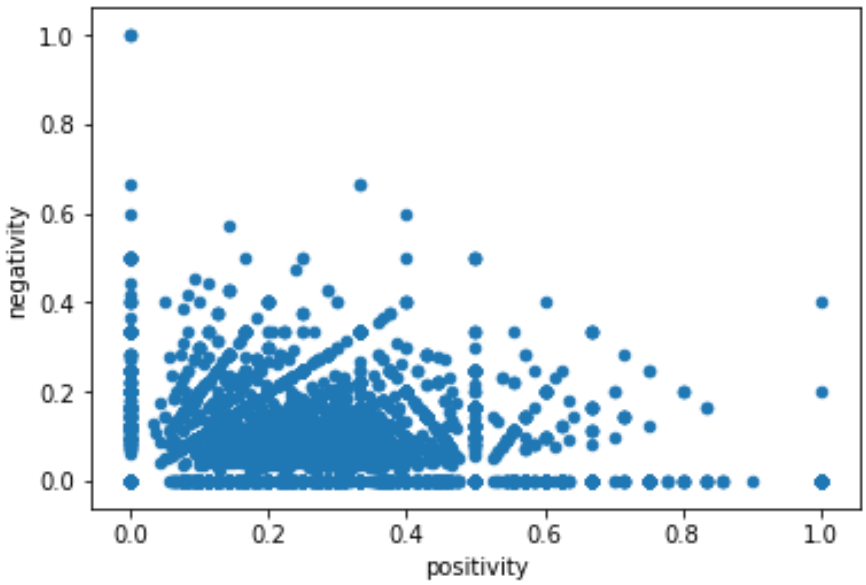
Positivity	0.27/1
Negativity	0.07/1
No of positive reviews	5.731
No of negative reviews	1.665
Percentage of positive reviews	77,4 %
Percentage of negative reviews	20,5 %

Πίνακας 11: Δεδομένα Ανάλυσης Συναισθήματος Παρθενώνα

count 7396.000000	count 7396.000000
mean 0.279332	mean 0.072934
std 0.174052	std 0.095738
min 0.000000	min 0.000000
25% 0.166667	25% 0.000000
50% 0.250000	50% 0.000000
75% 0.375000	75% 0.125000
max 1.000000	max 1.000000
Name: positivity, dtype: float64	Name: negativity, dtype: float64

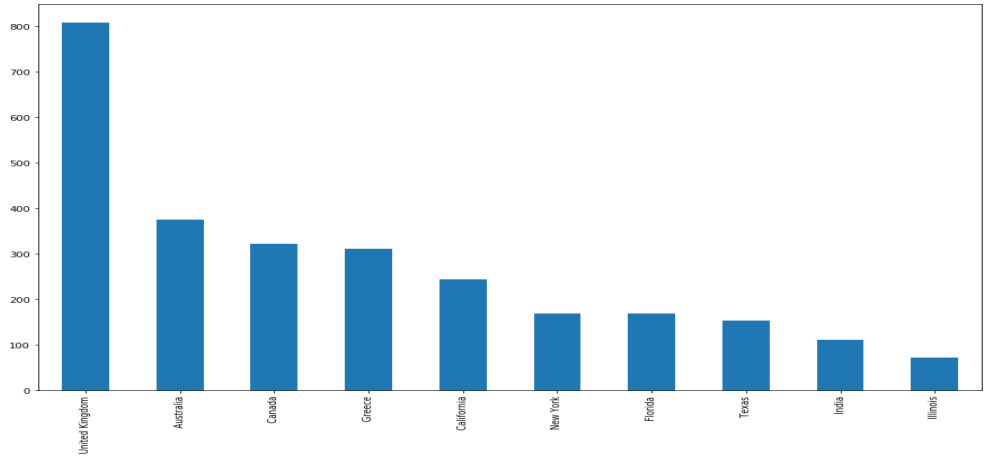
Πίνακας 12: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

Στο παρακάτω διάγραμμα μπορεί να διακριθεί η διάχυση της πολικότητας για τα δεδομένα που συλλέχτηκαν για το μνημείο του Παρθενώνα. Είναι φανερό ότι τα δεδομένα παρουσιάζουν μεγαλύτερη συγκέντρωση προς την θετική πολικότητα.



Εικόνα 27: Διάγραμμα διάχυσης πολικότητας για τον Παρθενώνα

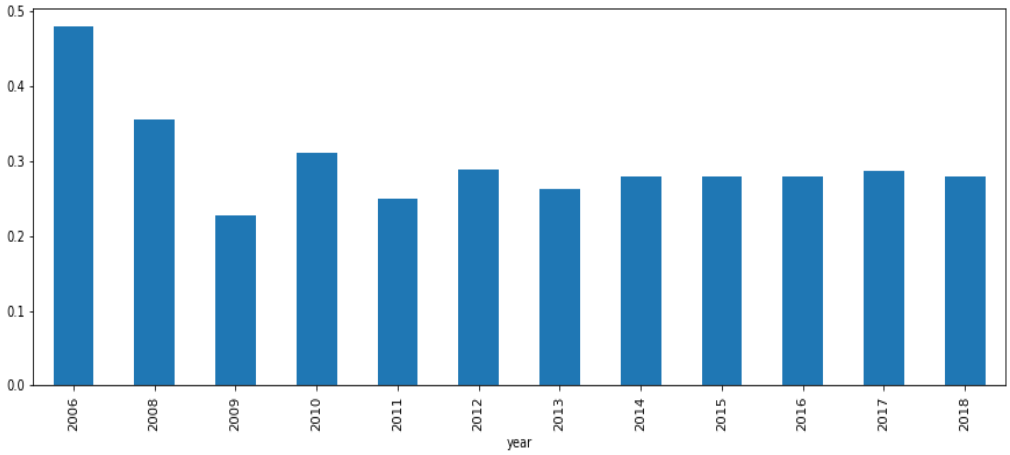
Αναφορικά με την προέλευση καταγωγής των χρηστών και την συχνότητα εμφάνισης σχολίων από συγκεκριμένες χώρες εντοπίστηκε ότι χρήστες από το Ηνωμένο Βασίλειο εμφανίζονται συχνότερα μέσα στο σύνολο δεδομένων μας, πράγμα που υποδηλώνει ότι οι Άγγλοι υπήκοοι επιλέγουν την Ελλάδα σε μεγαλύτερη συχνότητα από τις υπόλοιπες εθνικότητες, όπως διακρίνεται και από το παρακάτω γράφημα. Στην συνέχεια ακολουθούν σχόλια από χρήστες με τόπο προέλευσης την Αυστραλία, τον Καναδά , την Ελλάδα, την Καλιφόρνια, τη Νέα Υόρκη, την Φλόριντα, το Τέξας, την Ινδία, και το Ιλινόις.



Εικόνα 28: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

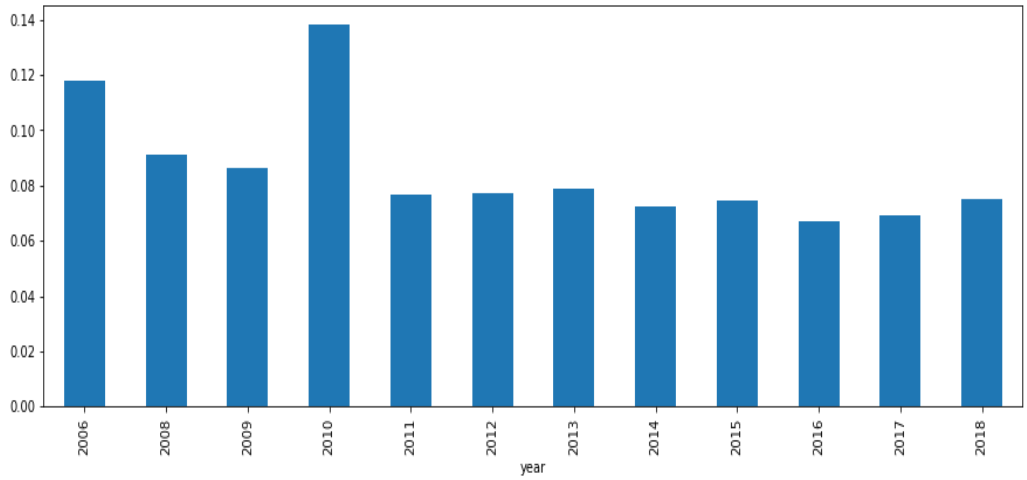
Τα πιο θετικά σχόλια εντοπίστηκαν επίσης ότι προέρχονται από τις προαναφερθείσες χώρες καθώς και οι πιο αρνητικές αξιολογήσεις παρατηρήθηκε ότι λήφθηκαν από τις ίδιες χώρες με μικρές διαφοροποιήσεις κυρίως Πολιτειών της Αμερικής στην τελευταία θέση.

Στη συνέχεια αναλύθηκαν οι αξιολογήσεις που κρίθηκαν με θετική πολικότητα ανά χρονολογία, με αποτέλεσμα να εντοπιστεί ότι οι χρονιές 2006, 2007 και 2010 εμφανίζουν τα περισσότερα θετικά σχόλια όπως διαφαίνεται στο παρακάτω διάγραμμα, και ιδιαίτερα η χρονιά 2006, αποτελεί την χρονιά με την μεγαλύτερη εμφάνιση σχολίων κατά μέσο όρο.



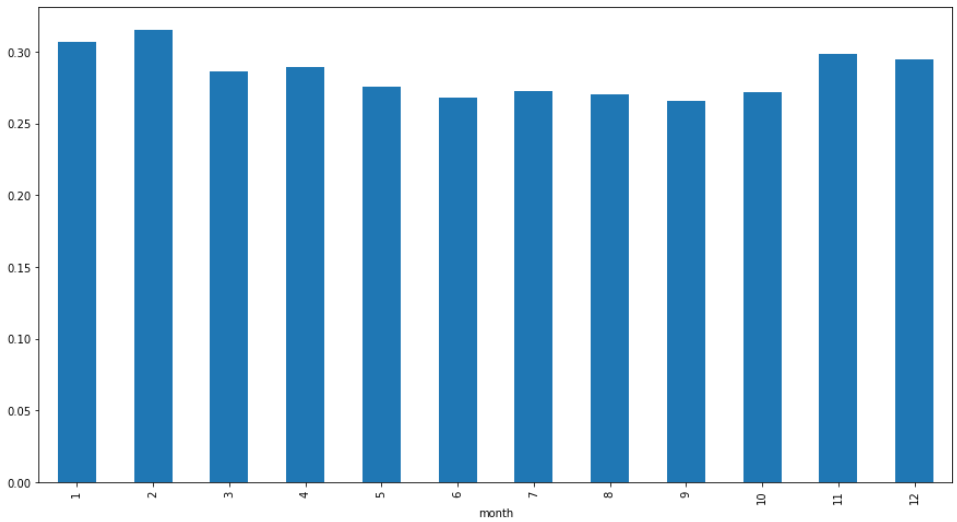
Εικόνα 29: Κατανομή θετικών σχολίων ανά χρονιά

Όσον αφορά τις αξιολογήσεις με αρνητικό πρόσημα ανά χρονιά παρατηρήθηκε ότι το 2010 συγκέντρωσε τα υψηλότερα επίπεδα αρνητικών αξιολογήσεων, με το 2006 και το 2008 να ακολουθούν.

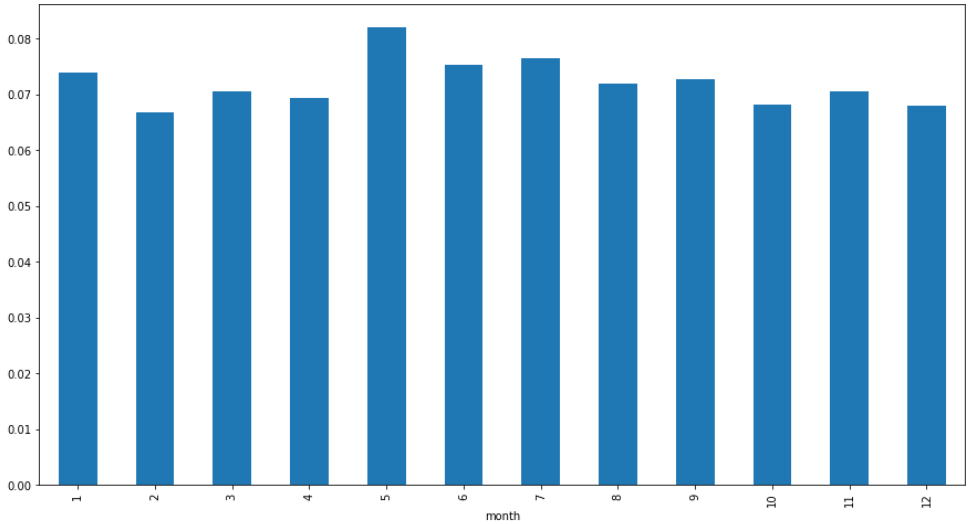


Εικόνα 30: Κατανομή αρνητικών σχολίων ανά χρονιά

Σχετικά με την ανά μήνα συχνότητα εμφάνισης θετικών αξιολογήσεων από τους χρήστες της πλατφόρμας TripAdvisor, εντοπίστηκε ότι οι μήνες Φεβρουάριος, Ιανουάριος και Νοέμβριος περιλαμβάνουν τα πιο θετικά σχόλια κατά μέσο όρο. Το αποτέλεσμα αυτό διαμορφώνεται διαφορετικά για τις αρνητικές αξιολογήσεις με τους μήνες Μάιο, Ιούλιο και Ιούνιο να κατέχουν τις πρώτες θέσεις όπως μπορεί να διαφανεί στα παρακάτω γραφήματα αντίστοιχα. Το πόρισμα αυτό είναι αρκετά λογικό αν αναλογιστεί κανείς την αυξημένη τουριστική κίνηση κατά τους καλοκαιρινούς μήνες.



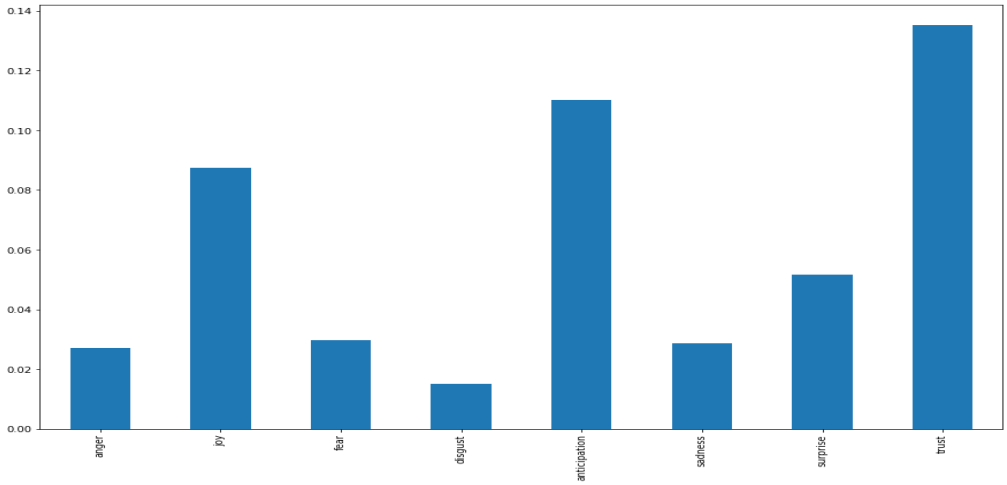
Εικόνα 31: Μέσος όρος θετικών σχολίων ανά μήνα



Εικόνα 32: Μέσος όρος αρνητικών σχολίων ανά μήνα

Επιπλέον το συναισθηματικό λεξικό που έχει εισαχθεί για την πραγμάτωση της ανάλυσης συναισθήματος πέρα από την πολικότητα, θετική η αρνητική, που μπορεί να εκφράσει, δίνει την δυνατότητα εντοπισμού και οχτώ επιπλέον συναισθημάτων (εμπιστοσύνη, έκπληξη, χαρά, αηδία, προσδοκία, φόβος και θυμό).

Αναφορικά με τον Παρθενώνα, τα συναισθήματα που εμφανίζουν μεγαλύτερη ένταση μέσα στο σώμα του κειμένου είναι αυτό της εμπιστοσύνης, της προσμονής και της χαράς, σε σύγκριση με τα αρνητικά συναισθήματα που προσφέρει το συναισθηματικό λεξικό που χρησιμοποιήθηκε (Εικόνα 33) , πράγμα που είναι αρκετά λογικό, αν αναλογιστεί κανείς ότι η συνολική πολικότητα των αξιολογήσεων είναι θετική, άρα και τα παραγόμενα συναισθήματα θα έχουν θετικό πρόσημο στο σύνολο τους.



Εικόνα 33: Μέσος Όρος Συναισθημάτων για τον Παρθενώνα

Πριν προχωρήσουμε την ομαδοποίηση των αξιολογήσεων για σκοπό της εύρεσης νοηματικής συνάφειας ανάμεσα στα σχόλια των χρηστών, πραγματοποιήθηκε μια απλή εύρεση των λέξεων που παρουσιάζουν μεγαλύτερη συχνότητα στα δεδομένα κειμένου που έχουμε εισάγει. Στον παρακάτω πίνακα εικονίζονται οι δέκα πιο συχνά εμφανιζόμενες λέξεις σχετικά με την Ακρόπολη.

Visit	220.927958
History	188.403543
Place	134.870729
Worth	121.025729
Site	107.077235
Time	102.463906
Top	92.254988
View	91.837913
Ancient	89.600100
Walk	87.459728

Πίνακας 13 : Οι 10 πιο συχνές λέξεις για τον Παρθενώνα

- **Ομαδοποίηση (K-means Clustering)**

Τέλος πραγματοποιήθηκε η μέθοδος ομαδοποίησης κειμένου K-means όπως έχει ήδη αναφερθεί. Για να ορίσουμε την μεταβλητή K κατά την επεξεργασία των δεδομένων έγιναν αρκετές δοκιμές, μέχρι να βρεθεί η τιμή του K που δεν εμφανίζει άλλες ομάδες με διαφορετική νοηματική συνάφεια, από αυτές που δημιουργήθηκαν. Η τιμή που ορίστηκε είναι το K να ισούται με το είκοσι (K=20), καταλήγοντας στα ευρήματα της Εικόνας 34.

```
In [64]: print("Top terms per cluster:")
         order_centroids = km.cluster_centers_.argsort()[:, ::-1]
         terms = vec.get_feature_names()
         for i in range(number_of_clusters):
             top_ten_words = [terms[ind] for ind in order_centroids[i, :2]]
             print("Cluster {}: {}".format(i, ' '.join(top_ten_words)))

Top terms per cluster:
Cluster 0: slippery marble
Cluster 1: temple athena
Cluster 2: museum acropolis
Cluster 3: place visit
Cluster 4: tour guide
Cluster 5: list bucket
Cluster 6: ancient monument
Cluster 7: history amazing
Cluster 8: building parthenon
Cluster 9: climb worth
Cluster 10: visit athens
Cluster 11: walk athens
Cluster 12: structure amazing
Cluster 13: experience parthenon
Cluster 14: parthenon athens
Cluster 15: shoes wear
Cluster 16: ticket sites
Cluster 17: ago years
Cluster 18: early crowds
Cluster 19: really parthenon
```

Εικόνα 34: Ομαδοποίηση λέξεων με την μέθοδο K-means

Συμπεραίνουμε μέσα από την ομάδες που δημιουργήθηκαν ότι οι χρήστες είναι θετικά διακείμενοι για το τουριστικό μνημείο του Παρθενώνα, τονίζοντας ότι είναι ένα μνημείο που αξίζει να επισκεφτείς με εντυπωσιακή δόμηση και ιστορικό πλούτο.

Στον αντίποδα, μπορεί να διαπιστωθεί ότι γίνεται λόγος για την δυσκολία ανάβασης στο σημείο αυτό (‘slippery marble, shoes wear’), ενώ ένα ακόμη αρνητικό σημείο που εντοπίζεται ξανά είναι η ουρά για την έκδοση εισιτηρίων, γεγονός που παρατηρήθηκε και στα αποτελέσματα αναφορικά με την Ακρόπολη.

6.4 ΜΟΥΣΕΙΟ ΑΚΡΟΠΟΛΗΣ

- **Ανάλυση Συναισθήματος**

Η μέτρηση της πολικότητας του σώματος κειμένου των αξιολογήσεων των χρηστών για το τουριστικό μνημείο της Πλάκας κρίθηκε επίσης θετική, με μέσο όρο στο σύνολο των αξιολογήσεων (18.219) 0.29/1 σε σύγκριση με τον μέσο όρο της αρνητικότητας που παρουσιάζουν οι αξιολογήσεις αυτές, όπως βλέπεται και στον παρακάτω πίνακα. Η κλίμακα μέτρησης κυμαίνεται από το 0 (καθόλου θετική ή αρνητική πολικότητα) μέχρι το 1 (απόλυτη θετικότητα ή αρνητικότητα). Αντίθετα, ό μέσος όρος της αρνητικότητα που παρουσιάζουν οι αξιολογήσεις κυμαίνεται στο

0.06. Σε μορφή ποσοστού οι αξιολογήσεις των χρηστών που κρίθηκαν θετικές σε σύγκριση με το σύνολο των αρνητικών αγγίζουν το 85%.

Επιπλέον, συγκρίναμε την θετική πολικότητα που εντοπίστηκε από την ανάλυση, με το μέσο όρο της βαθμολογίας (rating) που δίνουν οι ίδιοι οι χρήστες για το κάθε μνημείο. Ο μέσος όρος της βαθμολογίας για το Μουσείο της Ακρόπολης είναι 4.7 στα 5 αστέρια, ποσοστό που αγγίζει το 94 %. Το ποσοστό αυτό δεν έχει μεγάλη απόκλιση από το 85 % της θετικής πολικότητας που εντοπίστηκε, παρόλα αυτά υπάρχει μια διαφορά της τάξης του 9 %, πράγμα που δείχνει ότι οι χρήστες έχουν την τάση να βαθμολογούν ευκολότερα σε σχέση με αυτά που περιγράφουν στις αξιολογήσεις τους.

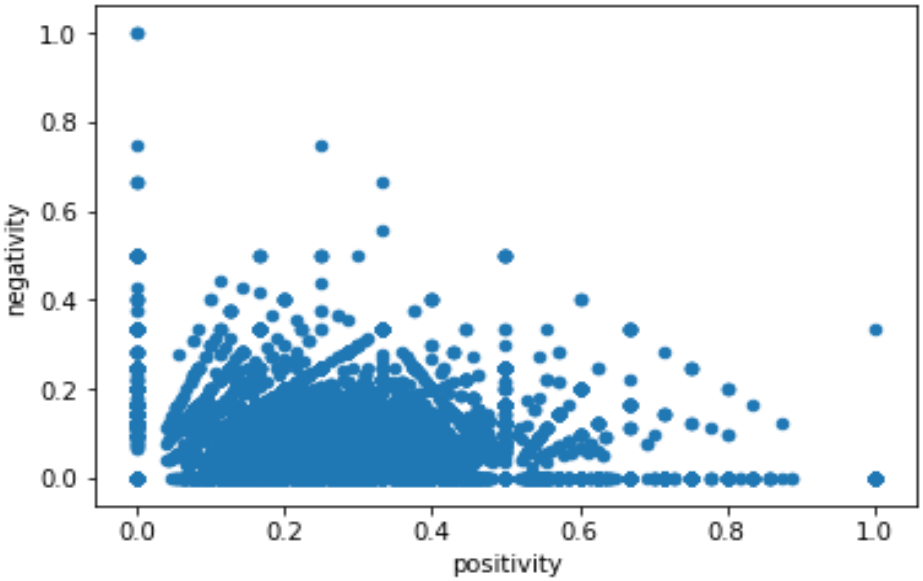
Positivity	0.29/1
Negativity	0.06/1
No of positive reviews	15.502
No of negative reviews	2.717
Percentage of positive reviews	85 %
Percentage of negative reviews	14.9 %

Πίνακας 14: Δεδομένα Ανάλυσης Συναισθήματος Μουσείου Ακρόπολης

count 18219.000000	count 18219.000000
mean 0.293255	mean 0.056848
std 0.161147	std 0.079889
min 0.000000	min 0.000000
25% 0.187500	25% 0.000000
50% 0.285714	50% 0.000000
75% 0.384615	75% 0.100000
max 1.000000	max 1.000000
Name: positivity, dtype: float64	Name: negativity, dtype: float64

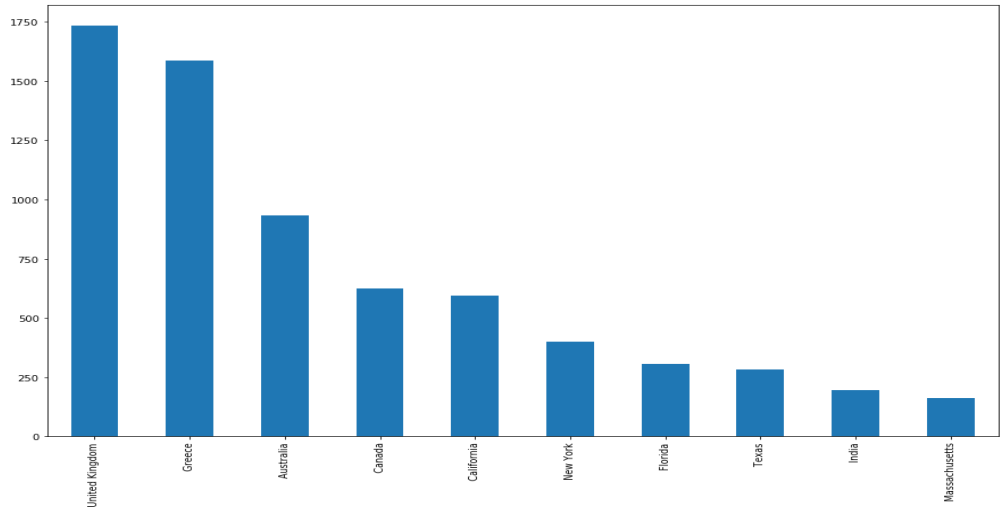
Πίνακας 15: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

Στο παρακάτω διάγραμμα μπορεί να διακριθεί η διάχυση των πολικοτήτων για τα δεδομένα που συλλέχτηκαν για το μουσείο της Ακρόπολης. Είναι φανερό ότι τα δεδομένα παρουσιάζουν μεγαλύτερη συγκέντρωση προς την θετική πολικότητα.



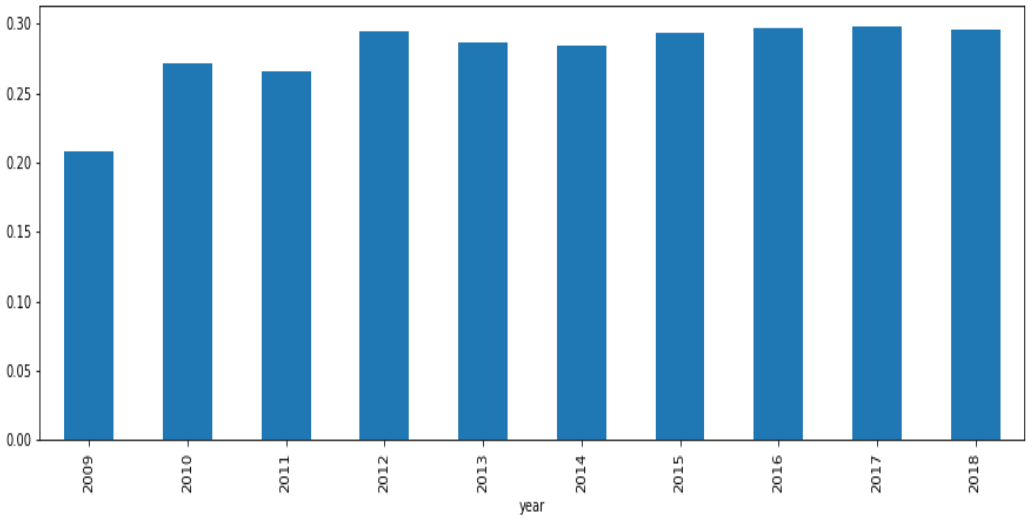
Εικόνα 35: Διάγραμμα διάχυσης πολικότητας για το Μουσείο της Ακρόπολης

Αναφορικά με την προέλευση καταγωγής των χρηστών και την συχνότητα εμφάνισης σχολίων από συγκεκριμένες χώρες εντοπίστηκε ότι χρήστες από το Ηνωμένο Βασίλειο εμφανίζονται συχνότερα μέσα στο σύνολο δεδομένων μας, πράγμα που υποδηλώνει ότι οι Άγγλοι υπήκοοι επιλέγουν την Ελλάδα σε μεγαλύτερη συχνότητα από τις υπόλοιπες εθνικότητες, όπως διακρίνεται και από το παρακάτω γράφημα. Στην συνέχεια ακολουθούν σχόλια από χρήστες με τόπο προέλευσης την Ελλάδα, την Αυστραλία, τον Καναδά, την Καλιφόρνια, τη Νέα Υόρκη, τη Φλόριντα, το Τέξας, την Ινδία και την Μασαχουσέτη.



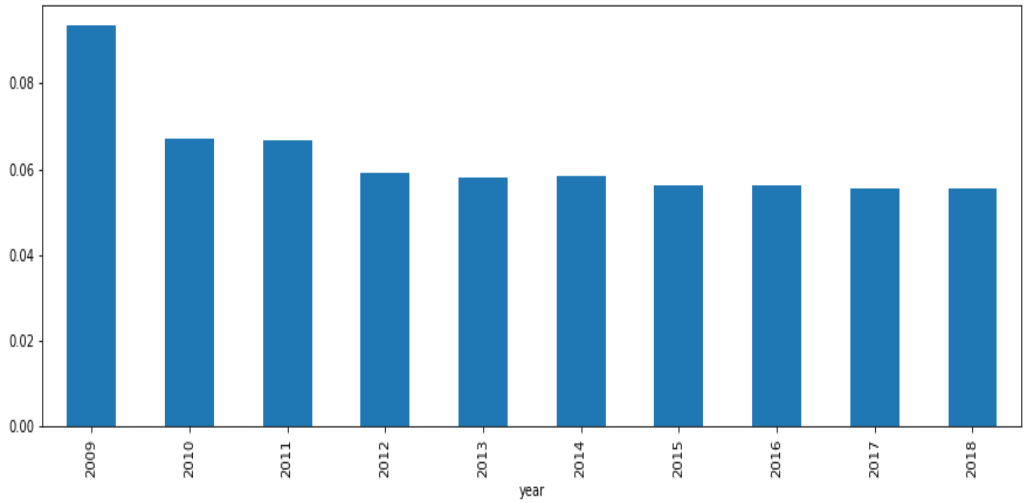
Εικόνα 36: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

Στη συνέχεια αναλύθηκαν οι αξιολογήσεις που κρίθηκαν με θετική πολικότητα ανά χρονολογία, με αποτέλεσμα να εντοπιστεί ότι οι χρονιές 2017, 2016 και 2018 εμφανίζουν τα περισσότερα θετικά σχόλια όπως διαφαίνεται στο παρακάτω διάγραμμα.



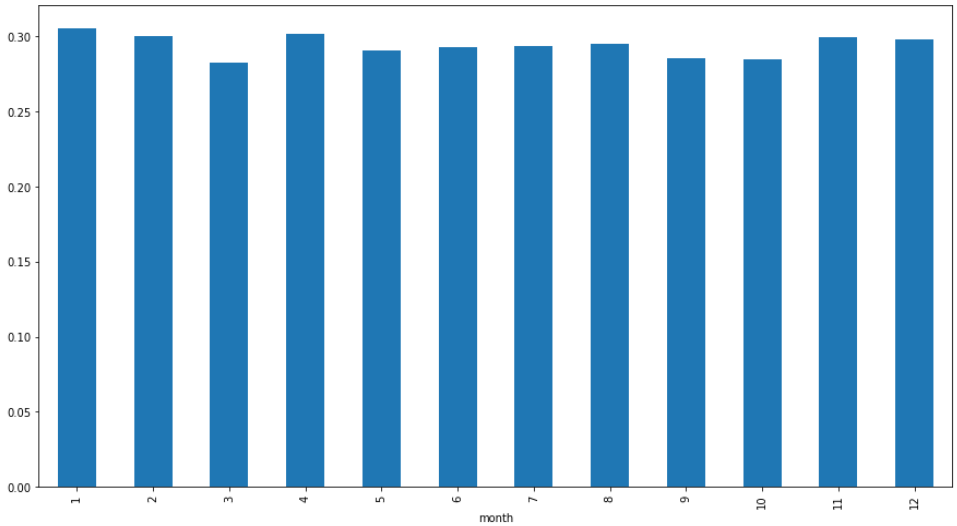
Εικόνα 37: Κατανομή θετικών σχολίων ανά χρονιά

Αναφορικά με τις χρονολογίες που εντοπίστηκαν οι περισσότερες αρνητικές αξιολογήσεις, αυτές είναι το 2009, το 2011 και το 2010, όπως διακρίνεται στην εικόνα παρακάτω.

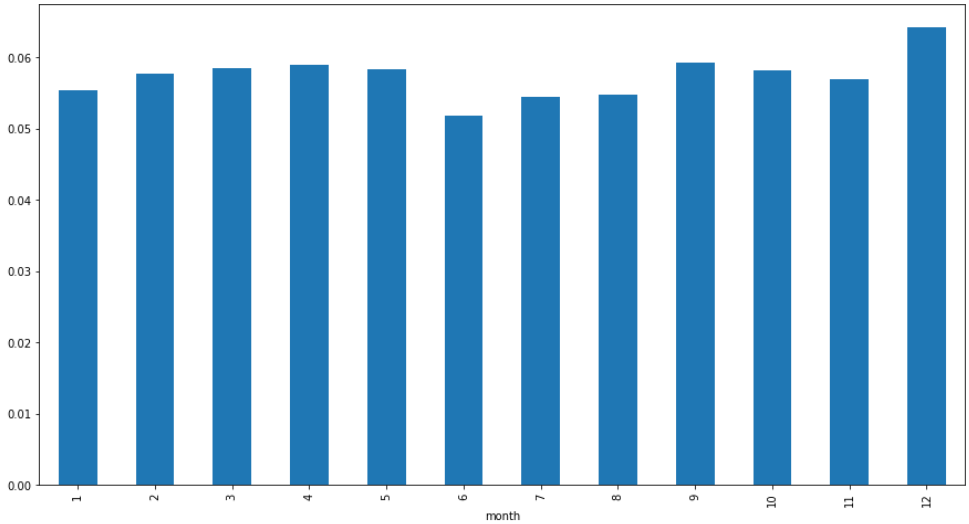


Εικόνα 38: Κατανομή αρνητικών σχολίων ανά χρονιά

Σχετικά με την ανά μήνα συχνότητα εμφάνισης αξιολογήσεων από τους χρήστες της πλατφόρμας TripAdvisor, εντοπίστηκε ότι οι μήνες κατά τους οποίους παρατηρείται μεγαλύτερη συχνότητα θετικών σχολίων είναι ο Ιανουάριος, ο Απρίλιος, ο Ιούλιος, και ο Νοέμβριος. Όσον αφορά τις αρνητικές αξιολογήσεις, ο Δεκέμβριος, ο Σεπτέμβριος και ο Απρίλιος παρουσιάζουν αυξημένα επίπεδα, όπως μπορεί να διαφανεί στα γραφήματα που ακολουθούν. Τα παραπάνω ευρήματα, όπως διακρίνεται δεν αναφέρονται σε καλοκαιρινούς μήνες, γεγονός καθόλα παράξενο γιατί το μουσείο της Ακρόπολης, αποτελεί πόλο πολιτιστικής και τουριστικής έλξης καθ' όλη την διάρκεια του χρόνου ακόμη και για τους ίδιους τους Έλληνες.



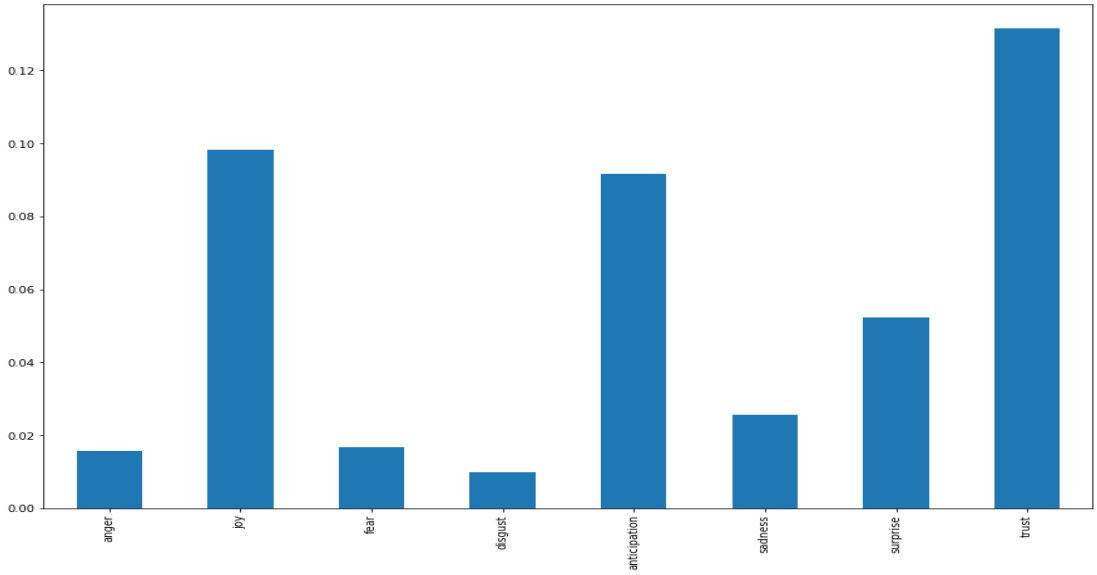
Εικόνα 39: Μέσος όρος θετικών σχολίων ανά μήνα



Εικόνα 40: Μέσος όρος αρνητικών σχολίων ανά μήνα

Επιπλέον το συναισθηματικό λεξικό που έχει εισαχθεί για την πραγμάτωση της ανάλυσης συναισθήματος πέρα από την πολικότητα, θετική η αρνητική, που μπορεί να εκφράσει, δίνει την δυνατότητα εντοπισμού και οχτώ επιπλέον συναισθημάτων (εμπιστοσύνη, έκπληξη, χαρά, αηδία, προσδοκία, φόβος και θυμό).

Αναφορικά με το μουσείο της Ακρόπολη, τα συναισθήματα που εμφανίζουν μεγαλύτερη ένταση μέσα στο σώμα του κειμένου είναι αυτό της εμπιστοσύνης , της προσμονής και της χαράς σε σύγκριση με τα αρνητικά συναισθήματα, όπως ο θυμός και η λύπη (Εικόνα 41) , πράγμα που είναι αρκετά λογικό, αν αναλογιστεί κανείς ότι η συνολική πολικότητα των αξιολογήσεων είναι θετική.



Εικόνα 41: Μέσος Όρος Συναισθημάτων για το Μουσείο της Ακρόπολης

Πριν προχωρήσουμε την ομαδοποίηση των αξιολογήσεων για σκοπό της εύρεσης νοηματικής συνάφειας ανάμεσα στα σχόλια των χρηστών, πραγματοποιήθηκε μια απλή εύρεση των λέξεων που παρουσιάζουν μεγαλύτερη συχνότητα στα δεδομένα κειμένου που έχουμε εισάγει. Στο Πίνακα 14 εικονίζονται οι δέκα πιο συχνά εμφανιζόμενες λέξεις σχετικά με το Μουσείο της Ακρόπολης.

Museum	1786.321267
Visit	577.026929
History	420.244996
Floor	271.175259
Building	246.042451
Place	240.636159
View	238.488180
Good	236.587992
Ancient	236.560558
Time	212.867777

Πίνακας 16: Οι 10 πιο συχνές λέξεις για το Μουσείο της Ακρόπολης

- **Ομαδοποίηση (K-means Clustering)**

Τέλος πραγματοποιήθηκε η μέθοδος ομαδοποίησης κειμένου K-means όπως έχει ήδη αναφερθεί. Για να ορίσουμε την μεταβλητή K κατά την επεξεργασία των δεδομένων έγιναν αρκετές δοκιμές, μέχρι να βρεθεί η τιμή του K που δεν εμφανίζει άλλες ομάδες με διαφορετική νοηματική συνάφεια, από αυτές που δημιουργήθηκαν. Η τιμή που ορίστηκε είναι το K να ισούται με το είκοσι (K=20), καταλήγοντας στα ευρήματα της Εικόνας 42.

```
In [70]: print("Top terms per cluster:")
         order_centroids = km.cluster_centers_.argsort()[:, ::-1]
         terms = vec.get_feature_names()
         for i in range(number_of_clusters):
             top_ten_words = [terms[ind] for ind in order_centroids[i, :2]]
             print("Cluster {}: {}".format(i, ' '.join(top_ten_words)))

         Top terms per cluster:
         Cluster 0: enjoyed thoroughly
         Cluster 1: allowed pictures
         Cluster 2: greece ancient
         Cluster 3: easy museum
         Cluster 4: athens museum
         Cluster 5: modern museum
         Cluster 6: hours spent
         Cluster 7: marbles elgin
         Cluster 8: just amazing
         Cluster 9: floor parthenon
         Cluster 10: museum great
         Cluster 11: acropolis museum
         Cluster 12: loved nice
         Cluster 13: fee entry
         Cluster 14: years old
         Cluster 15: day spent
         Cluster 16: really museum
         Cluster 17: best museums
         Cluster 18: greek history
         Cluster 19: glass floors
```

Εικόνα 42: Ομαδοποίηση λέξεων με την μέθοδο K-means

Συμπεραίνουμε μέσα από τις ομάδες με νοηματική συνάφεια που δημιουργήθηκαν ότι οι χρήστες είναι θετικά διακείμενοι για το μουσείο της Ακρόπολης, τονίζοντας ότι είναι ένα μοντέρνο με γυάλινο πάτωμα και αρχαία πολιτισμικά μνημεία και ότι αξίζει να βιώσει κάποιος αυτή την εμπειρία. Από την άλλη μεριά βέβαια εντοπίζονται και κάποια αρνητικά χαρακτηριστικά όπως είναι η πληρωμή του εισιτηρίου.

6.5 ΑΡΧΑΙΑ ΑΓΟΡΑ ΑΘΗΝΩΝ

- **Ανάλυση Συναισθήματος**

Η μέτρηση της πολικότητας του σώματος κειμένου των αξιολογήσεων των χρηστών για το τουριστικό μνημείο της Αρχαίας Αγοράς κρίθηκε επίσης θετική, με μέσο όρο στο σύνολο των αξιολογήσεων (2.082) 0.25/1 σε σύγκριση με τον μέσο όρο της αρνητικότητας που παρουσιάζουν οι αξιολογήσεις αυτές, όπως βλέπεται και στον παρακάτω πίνακα. Η κλίμακα μέτρησης κυμαίνεται από το 0 (καθόλου θετική ή αρνητική πολικότητα) μέχρι το 1 (απόλυτη θετικότητα ή αρνητικότητα). Αντίθετα, ό μέσος όρος της αρνητικότητα που παρουσιάζουν οι αξιολογήσεις κυμαίνεται στο 0.07. Σε μορφή ποσοστού οι αξιολογήσεις των χρηστών που κρίθηκαν θετικές σε σύγκριση με το σύνολο των αρνητικών αγγίζουν το 72,6%.

Επιπλέον, συγκρίναμε την θετική πολικότητα που εντοπίστηκε από την ανάλυση, με το μέσο όρο της βαθμολογίας (rating) που δίνουν οι ίδιοι οι χρήστες για το κάθε μνημείο. Ο μέσος όρος της βαθμολογίας για την ακρόπολη είναι 4.5 στα 5 αστέρια, ποσοστό που αγγίζει το 90 %. Το ποσοστό αυτό δεν έχει μεγάλη απόκλιση από το 72,6 % της θετικής πολικότητας που εντοπίστηκε, παρόλα αυτά υπάρχει μια διαφορά της τάξης του 17,4 %, πράγμα που δείχνει ότι οι χρήστες έχουν την τάση να βαθμολογούν ευκολότερα σε σχέση με αυτά που περιγράφουν στις αξιολογήσεις τους.

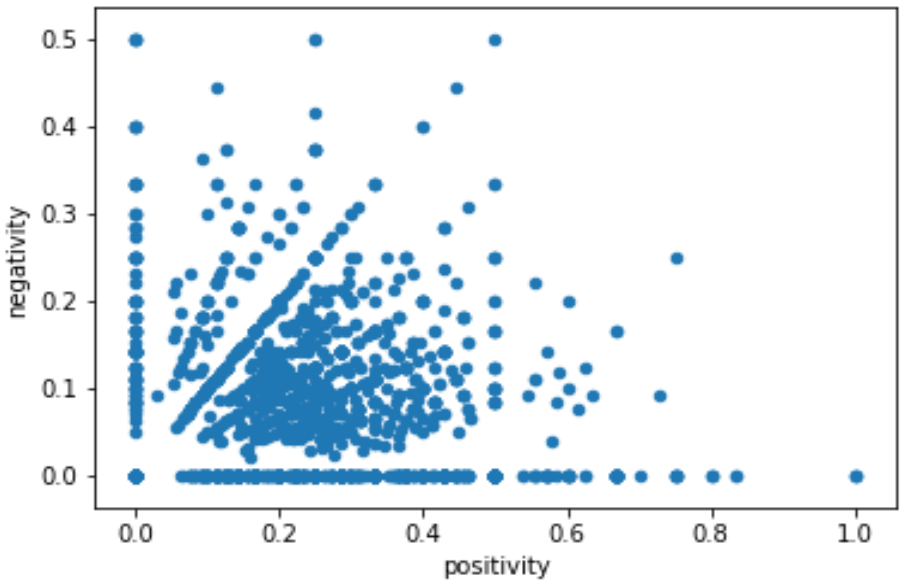
Positivity	0.25/1
Negativity	0.08/1
No of positive reviews	1.570
No of negative reviews	512
Percentage of positive reviews	75,4 %
Percentage of negative reviews	24,5 %

Πίνακας 17: Δεδομένα Ανάλυσης Συναισθήματος Αρχαίας Αγοράς

count 2082.000000 mean 0.248100 std 0.150479 min 0.000000 25% 0.147331 50% 0.238095 75% 0.333333 max 1.000000 Name: positivity, dtype: float64	count 2082.000000 mean 0.082711 std 0.088817 min 0.000000 25% 0.000000 50% 0.076923 75% 0.129480 max 0.500000 Name: negativity, dtype: float64

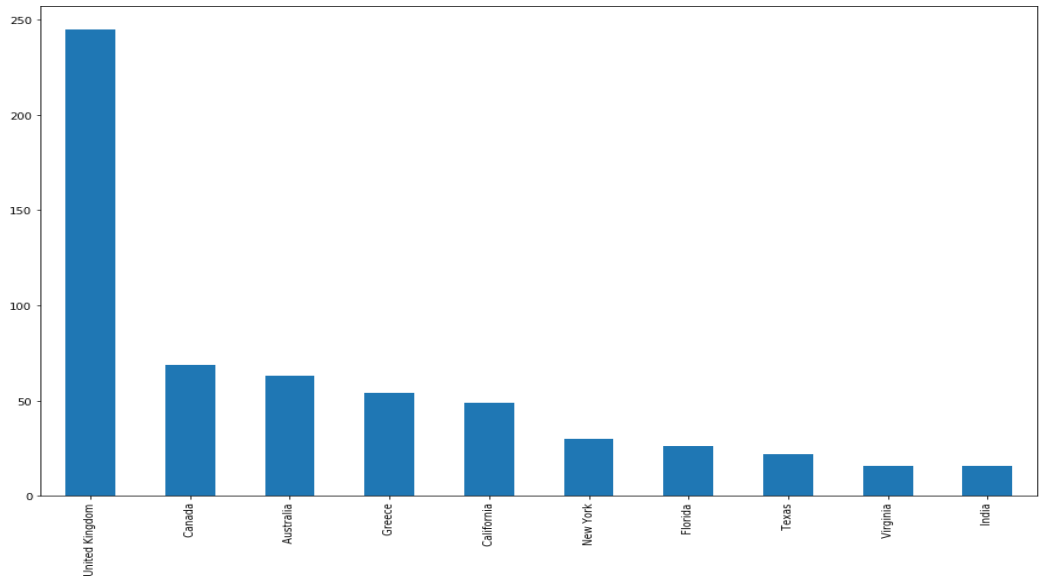
Πίνακας 18: Στατιστικά στοιχεία επεξεργασίας ανάλυσης συναισθήματος

Στο παρακάτω διάγραμμα μπορεί να διακριθεί η διάχυση των πολικοτήτων για τα δεδομένα που συλλέχτηκαν για την Αρχαία Αγορά των Αθηνών. Είναι φανερό ότι τα δεδομένα παρουσιάζουν μεγαλύτερη συγκέντρωση προς την θετική πολικότητα.



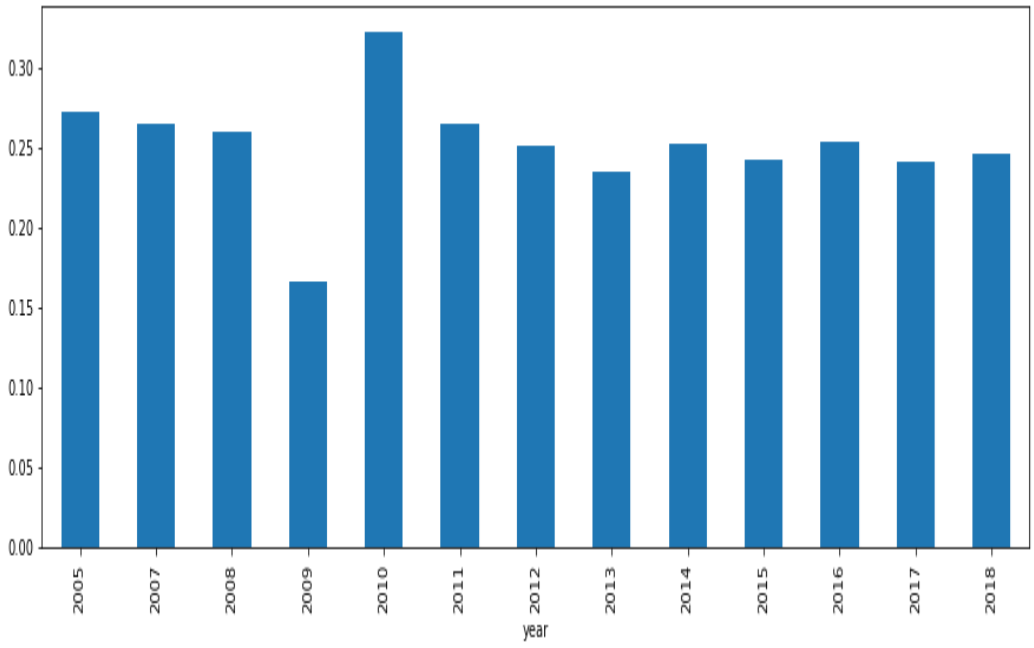
Εικόνα 43: Διάγραμμα διάχυσης πολικότητας για την Αρχαία Αγορά

Αναφορικά με την προέλευση καταγωγής των χρηστών και την συχνότητα εμφάνισης σχολίων από συγκεκριμένες χώρες εντοπίστηκε ότι χρήστες από το Ηνωμένο Βασίλειο εμφανίζονται συχνότερα μέσα στο σύνολο δεδομένων μας, πράγμα που υποδηλώνει ότι οι Άγγλοι υπήκοοι επιλέγουν την Ελλάδα σε μεγαλύτερη συχνότητα από τις υπόλοιπες εθνικότητες, όπως διακρίνεται και από το παρακάτω γράφημα. Στην συνέχεια ακολουθούν σχόλια από χρήστες με τόπο προέλευσης τον Καναδά, την Αυστραλία, την Ελλάδα, την Καλιφόρνια και την Νέα Υόρκη.

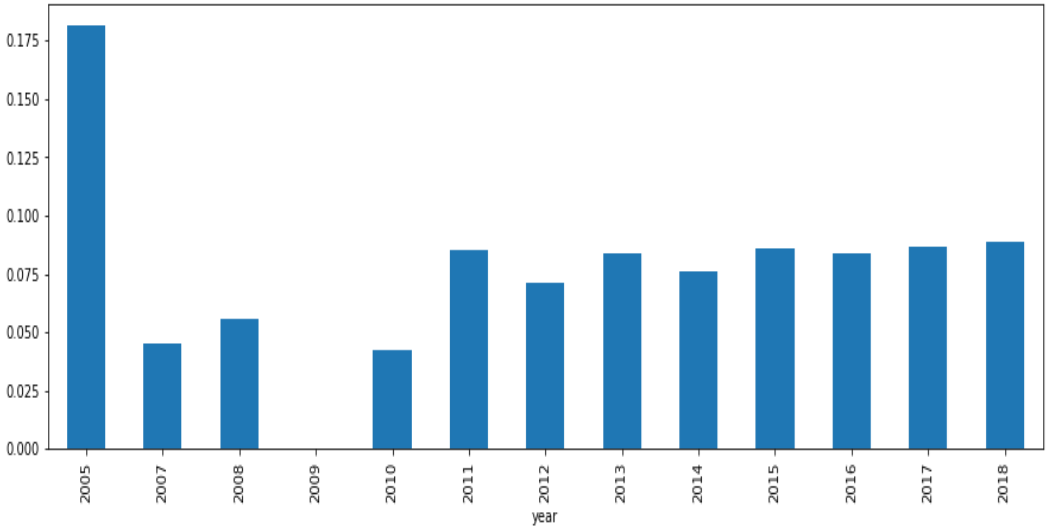


Εικόνα 44: Οι 10 χώρες με την μεγαλύτερη συχνότητα εμφάνισης σχολίων

Στη συνέχεια αναλύθηκαν οι αξιολογήσεις που κρίθηκαν με θετική πολικότητα ανά χρονολογία, με αποτέλεσμα να εντοπιστεί ότι οι χρονιές 2010, 2005 και 2007 εμφανίζουν τα περισσότερα θετικά σχόλια όπως διαφαίνεται στο παρακάτω διάγραμμα, και ιδιαίτερα η χρονιά 2010, αποτελεί την χρονιά με την μεγαλύτερη εμφάνιση θετικών σχολίων ενώ στον αντίποδα αναφορικά με τα έτη που εμφανίζονται τα περισσότερο αρνητικά σχόλια είναι το 2005, το 2018 και το 2015 αντίστοιχα, σύμφωνα με τις εικόνες που ακολουθούν.



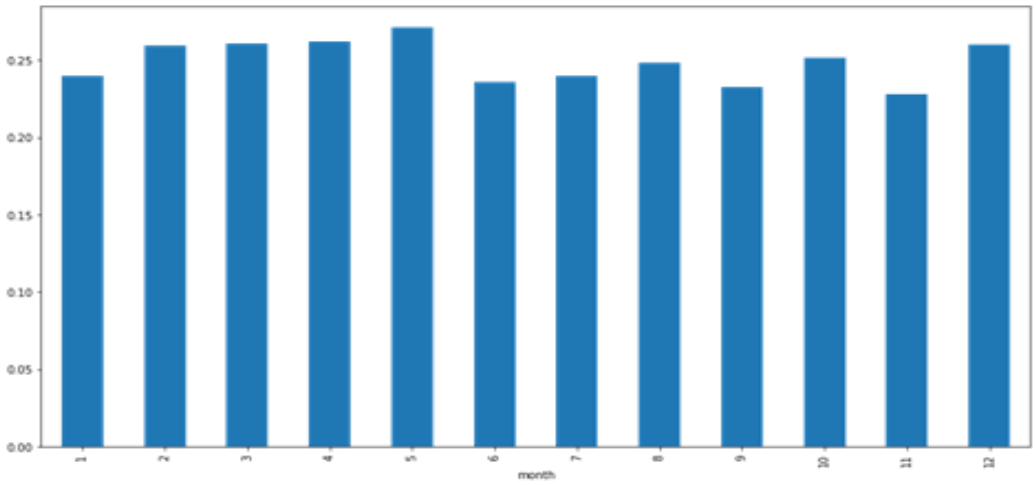
Εικόνα 45: Κατανομή θετικών σχολίων ανά χρονιά



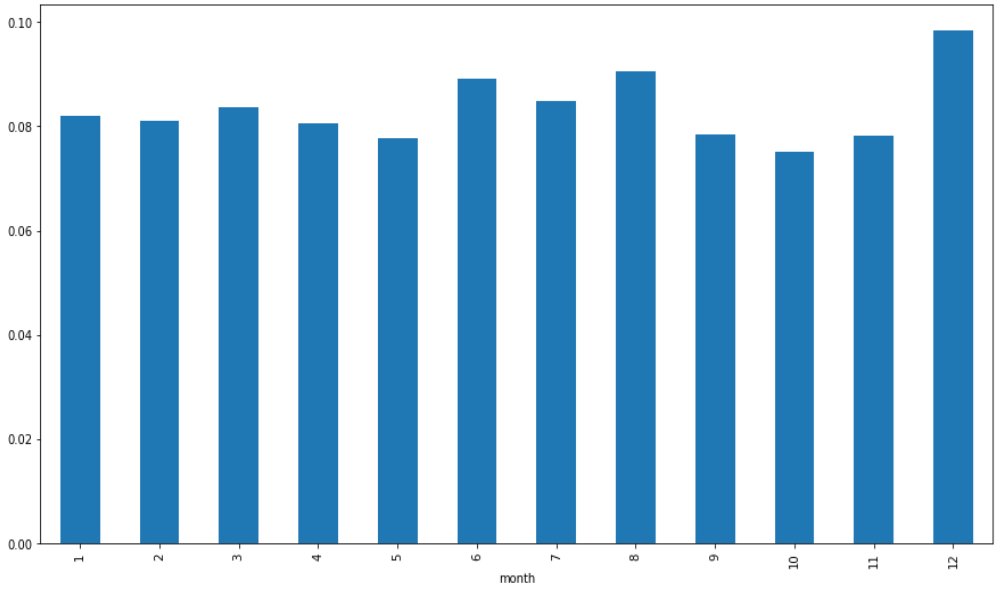
Εικόνα 46: Κατανομή αρνητικών σχολίων ανά χρονιά

Σχετικά με την ανά μήνα συχνότητα εμφάνισης αξιολογήσεων από τους χρήστες της πλατφόρμας TripAdvisor, εντοπίστηκε ότι οι μήνες κατά τους οποίους παρατηρείται μεγαλύτερος αριθμός εμφάνισης θετικών σχολίων είναι ο Μάιος, ο Απρίλιος, ο Δεκέμβριος και ο Μάρτιος. Αυτό το αποτέλεσμα διαμορφώνεται διαφορετικά όσον αφορά τις αρνητικές αξιολογήσεις, όπως μπορεί να διαφανεί στην εικόνα 49 και 50

αντίστοιχα. Το πόρισμα αυτό είναι αρκετά λογικό καθώς η Ελλάδα δέχεται περισσότερους τουρίστες κατά τους καλοκαιρινούς μήνες.



Εικόνα 47: Μέσος όρος εμφάνισης θετικών σχολίων ανά μήνα

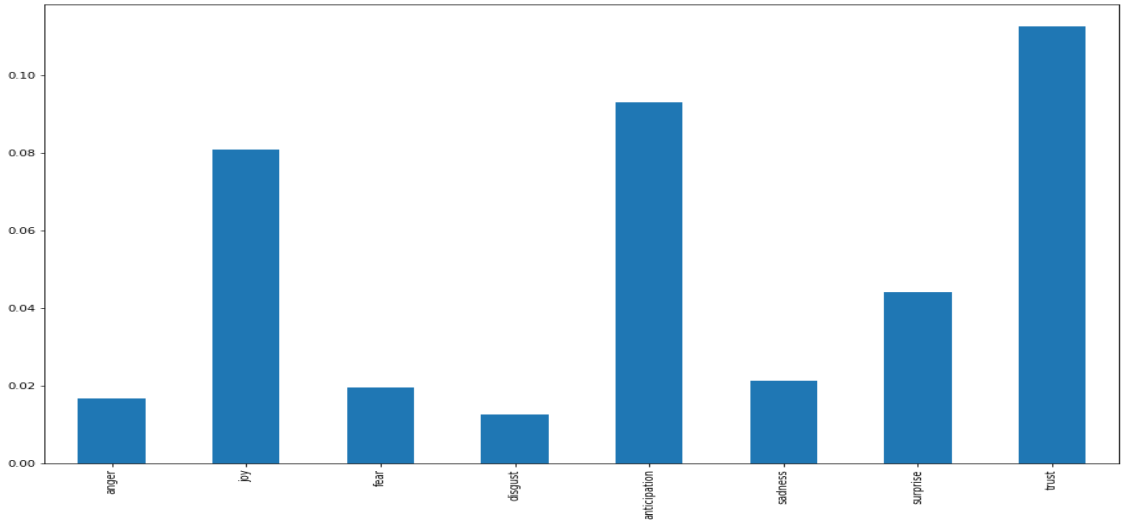


Εικόνα 48: Μέσος όρος εμφάνισης αρνητικών σχολίων ανά μήνα

Επιπλέον το συναισθηματικό λεξικό που έχει εισαχθεί για την πραγμάτωση της ανάλυσης συναισθήματος πέρα από την πολικότητα, θετική η αρνητική, που μπορεί να εκφράσει, δίνει την δυνατότητα εντοπισμού και οχτώ επιπλέον συναισθημάτων (εμπιστοσύνη, έκπληξη, χαρά, αηδία, προσδοκία, φόβος και θυμό).

Αναφορικά με την Αρχαία Αγορά των Αθηνών, τα συναισθήματα που εμφανίζουν μεγαλύτερη ένταση μέσα στο σώμα του κειμένου είναι αυτό της εμπιστοσύνης , της

προσμονής και της χαράς σε σύγκριση με τα αρνητικά συναισθήματα, όπως ο θυμός και η λύπη (Εικόνα 41) , πράγμα που είναι αρκετά λογικό, αν αναλογιστεί κανείς ότι η συνολική πολικότητα των αξιολογήσεων είναι θετική.



Εικόνα 49: Μέσος Όρος Συναισθημάτων για Αρχαία Αγορά

Πριν προχωρήσουμε την ομαδοποίηση των αξιολογήσεων για σκοπό της εύρεσης νοηματικής συνάφειας ανάμεσα στα σχόλια των χρηστών, πραγματοποιήθηκε μια απλή εύρεση των λέξεων που παρουσιάζουν μεγαλύτερη συχνότητα στα δεδομένα κειμένου που έχουμε εισάγει. Στο Πίνακα 18 εικονίζονται οι δέκα πιο συχνά εμφανιζόμενες λέξεις σχετικά με την Αρχαία Αγορά των Αθηνών.

ancient	77.105338
place	55.429737
site	51.441026
visit	50.983678
museum	48.937613
temple	46.612871
ruins	43.678394
walk	41.448425
time	36.778669
area	35.877927

Πίνακας 19 : Οι 10 πιο συχνές λέξεις για την Αρχαία Αγορά

- **Ομαδοποίηση (K-means Clustering)**

Τέλος πραγματοποιήθηκε η μέθοδος ομαδοποίησης κειμένου K-means όπως έχει ήδη αναφερθεί. Για να ορίσουμε την μεταβλητή K κατά την επεξεργασία των δεδομένων έγιναν αρκετές δοκιμές, μέχρι να βρεθεί η τιμή του K που δεν εμφανίζει άλλες ομάδες με διαφορετική νοηματική συνάφεια, από αυτές που δημιουργήθηκαν. Η τιμή που ορίστηκε είναι το K να ισούται με το είκοσι (K=20), καταλήγοντας στα ευρήματα της Εικόνας 50.

```
In [89]: print("Top terms per cluster:")
         order_centroids = km.cluster_centers_.argsort()[:, ::-1]
         terms = vec.get_feature_names()
         for i in range(number_of_clusters):
             top_ten_words = [terms[ind] for ind in order_centroids[i, :2]]
             print("Cluster {}: {}".format(i, ' '.join(top_ten_words)))

         Top terms per cluster:
         Cluster 0: place history
         Cluster 1: way acropolis
         Cluster 2: right hour
         Cluster 3: site market
         Cluster 4: temple hephaestus
         Cluster 5: amazing site
         Cluster 6: parthenon day
         Cluster 7: enjoyed agora
         Cluster 8: years ago
         Cluster 9: lot time
         Cluster 10: great acropolis
         Cluster 11: acropolis entrance
         Cluster 12: early feel
         Cluster 13: athens agora
         Cluster 14: tour audio
         Cluster 15: ancient heart
         Cluster 16: sure know
         Cluster 17: looked imagination
         Cluster 18: museum temple
         Cluster 19: life ancient
```

Εικόνα 50: Ομαδοποίηση λέξεων με την μέθοδο K-means

Συμπεραίνουμε μέσα από την ομάδες που δημιουργήθηκαν ότι οι χρήστες είναι θετικά διακείμενοι για την Αρχαία Αγορά των Αθηνών, είναι εντυπωσιασμένοι με την ιστορικότητα της περιοχής, και ότι αξίζει να βιώσει κάποιος αυτή την εμπειρία. Από την άλλη μεριά βέβαια εντοπίζονται και κάποια ίσως αρνητικά σημεία όπως είναι η σωστή επιλογή ώρας επίσκεψης της Αρχαίας Αγοράς, ώστε να είναι ανοιχτή για το κοινό, καθώς υπάρχουν καθορισμένες ώρες και μέρες που βρίσκεται ανοιχτή.

7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στη παρούσα εργασία έγινε μια προσπάθεια αποτύπωσης της συναισθηματικής πολικότητας των αξιολογήσεων των χρηστών από την κοινωνική πλατφόρμα TripAdvisor, η οποία ορίστηκε το πεδίο μελέτης μας, με τη χρήση συναισθηματικού λεξικού σχετικά με πέντε καίρια σημεία τουρισμού στην Αθήνα. Παράλληλα παρουσιάστηκαν ευρήματα σχετικά με τον τόπο προέλευσης των χρηστών, την εποχικότητα και τις χρονολογίες εμφάνισης των περισσότερων αξιολογήσεων ενώ τέλος έγινε μια προσπάθεια ταξινόμησης του σώματος κειμένου σε ομάδες νοηματικής συνάφειας για την εξαγωγή συμπερασμάτων σχετικά με τις απόψεις που εκφράζουν οι χρήστες μέσα στις αξιολογήσεις τους.

Μέσω της παραπάνω διαδικασίας και έρευνας συμπεράναμε ότι το συναίσθημα που προκύπτει από την ανάλυση της πολικότητας είναι θετικό για όλα τα σημεία έρευνας, όπως αποδεικνύεται και από την προσωπική βαθμολογία που δίνουν οι ίδιοι οι χρήστες στο σύστημα βαθμολόγησης που υπάρχει διαθέσιμο από την ίδια την πλατφόρμα. Αυτό που αξίζει να σημειωθεί εδώ είναι το γεγονός ότι οι χρήστες τείνουν να ψηφίζουν πολύ πιο θετικά σε σύγκριση με αυτό που πραγματικά εκφράζουν. Άρα η άποψη που εκφράζεται από τους χρήστες για τα πέντε αυτά τουριστικά σημεία που μελετήθηκαν είναι στο σύνολο της θετική και συνάδει με την γενική βαθμολόγηση της ίδιας της πλατφόρμας.

Ταυτόχρονα, αναλύθηκαν και συγκεκριμένα συναισθήματα που εντοπίζονται στο σώμα των δεδομένων μας, τα οποία προσφέρονται από το λεξικό που χρησιμοποιήθηκε για την εκπόνηση της έρευνας. Στην πλειονότητα τους τα συναισθήματα με την μεγαλύτερη εμφάνιση κατά μέσο όρο είναι εκείνα της εμπιστοσύνης, της προσδοκίας και της χαράς για όλα τα σημεία μελέτης, γεγονός που δικαιολογείται λόγω της συνολικής θετικής πολικότητας, που παράχθηκε από την έρευνα μας.

Παράλληλα, αναφορικά με την γεωγραφική προέλευση των χρηστών που δημοσίευσαν τις αξιολογήσεις τους στην πλατφόρμα TripAdvisor στα συγκεκριμένα σημεία μελέτης, παρατηρήθηκε ότι οι Βρετανοί, οι Αυστραλοί, οι Καναδοί και οι Αμερικανοί τουρίστες επιλέγουν την Ελλάδα ως προορισμό για τα ταξίδια τους, γεγονός που αφορά και τις θετικά διακείμενες αξιολογήσεις αλλά και τις αρνητικές. Το αποτέλεσμα αυτό δεν μπορεί φυσικά να θεωρηθεί καθολικό ή ρηξικέλευθο, καθώς

πρέπει να ληφθεί υπόψη η δημοτικότητα της πλατφόρμας σε αυτές τις χώρες αλλά και το γεγονός ότι είναι ευρέως γνωστό, ότι κάθε χρόνο επισκέπτονται την χώρα μας, μεγάλο ποσοστό Βρετανών. Παρόλα αυτά τα ευρήματα μας μπορούν επαρκώς να επικοινωνήσουν την κατάσταση της τουριστικής προέλευσης στην χώρα μας και να αποτελέσουν κίνητρο για περαιτέρω μελέτη και έρευνα της προώθησης της Ελλάδας στις συγκεκριμένες περιοχές.

Παρατηρώντας επίσης, την κατανομή των σχολίων ανά έτη, είτε για τις θετικές αξιολογήσεις, είτε για τις αρνητικές, διαπιστώνεται ότι οι χρονιές 2007, 2006 και 2010 έχουν το μεγαλύτερο αριθμό σχολίων με θετικό πρόσημο κατά μέσο όρο, ενώ αντίστοιχα οι αξιολογήσεις που παρουσίασαν το μεγαλύτερο αριθμό αρνητικών αξιολογήσεων είναι το 2006 και το 2011 κατά μέσο όρο. Τα παραπάνω αποτελέσματα μπορούν σε ένα μεταγενέστερο στάδιο έρευνας να αποτελέσουν σημείο μελέτης σχετικά με την εύρεση των αιτιών που ευθύνονται για την παραπάνω κατανομή.

Στην παρούσα έρευνα τέθηκε και το ερώτημα σχετικά με την εποχικότητα εμφάνισης των σχολίων, θεωρώντας ότι οι καλοκαιρινοί μήνες και οι μήνες που διέπουν το καλοκαίρι θα αποτελούν τους μήνες που θα υπάρχουν οι περισσότερες αξιολογήσεις από χρήστες της πλατφόρμας. Μολαταύτα, διαπιστώθηκε ότι αναφορικά με τους μήνες που παρουσιάζουν τα πιο θετικά σχόλια κατά μέσο όρο είναι οι μήνες Φεβρουάριος, Απρίλιος και Μάιος ενώ για τα αρνητικά σχόλια σημειώθηκε ότι οι μήνες Ιούλιος, Δεκέμβριος και Αύγουστος παρουσιάζουν την μεγαλύτερη συχνότητα κατά μέσο όρο για όλα τα σημεία μελέτης. Το τελευταίο εύρημα επιβεβαιώνει την υπόθεση μας ότι οι καλοκαιρινοί μήνες, όπου η τουριστική κίνηση είναι αυξημένη και κατά συνέπεια θα υπάρχουν πιο πολλά σχόλια, παρόλα αυτά εντοπίστηκαν μόνο όσον αφορά τις αρνητικές αξιολογήσεις.

Εν συνεχεία, αναφορικά με την εξόρυξη πληροφορίας από τα δεδομένα μας, η ομαδοποίηση σε clusters μας έδωσε την δυνατότητα να εντοπίσουμε σημεία που αναφέρουν οι ίδιοι οι χρήστες ως θεματικές μέσα στα σχόλια τους. Αυτό που διαφαίνεται ως επί το πλείστον είναι η θετική εντύπωση που λαμβάνουν οι επισκέπτες και συγκεκριμένα αναφέρονται σημεία, όπως η ιστορικότητα των σημείων αυτών, η θέα που προσφέρουν, και η σημασία που έχουν ενώ στα αρνητικά εντοπίζονται οι ουρές για τα εισιτήρια, η πολυκοσμία και η δυσκολία πρόσβασης σε αυτά. Τα παραπάνω ευρήματα μπορούμε να πούμε ότι δείχνουν κάποια πλεονεκτήματα και

μειονεκτήματα των τουριστικών αξιοθέατων, δίνοντας μας το κίνητρο για περαιτέρω βελτίωση αδυναμιών που παρουσιάζει ο τουρισμός στην Αθήνα και διατήρηση των στοιχείων που αποτελούν καίρια σημεία για αυτόν.

Εν κατακλείδι, η παρούσα εργασία, μέσω των μεθόδων που χρησιμοποιήθηκαν για την ανάλυση συναισθήματος και της εξόρυξης γνώμης, αποτελεί μια προσέγγιση για την παρουσίαση του θέματος που ανέλυσε, χωρίς να παράγει καθολικά πορίσματα αλλά θέλοντας να αποτελέσει την βάση για περαιτέρω μελλοντική μελέτη και έρευνα για το πεδίο του τουρισμού στην Ελλάδα γενικότερα.

7.1 ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ

Η παρούσα εργασία παρουσιάζει μια σύντομη έρευνα σχετικά με την ανάλυση συναισθήματος με τη χρήση λεξικού στο πεδίου του τουρισμού. Οι περισσότερες από τις έρευνες επικεντρώνονται στην αγγλική γλώσσα και μπορούν να επεκταθούν σε άλλες γλώσσες για να κατανοήσουν την περιφερειακή τάση. Στην ταξινόμηση των συναισθημάτων, η μηχανική μάθηση και η προσέγγιση με βάση το λεξικό είναι οι πιο ευρέως χρησιμοποιούμενες προσεγγίσεις, αλλά η υβριδική προσέγγιση πρέπει να διερευνηθεί περαιτέρω για καλύτερα αποτελέσματα. Δεδομένου ότι η ανάλυση των συναισθημάτων είναι πρόβλημα που εξαρτάται από το συγκεκριμένο τομέα, υπάρχουν μόνο λίγες περιοχές στις οποίες έχουν πραγματοποιηθεί εργασίες και υπάρχουν πολλά πεδία που πρέπει ακόμη να διερευνηθούν.

Μπορούμε επομένως, να επικεντρωθούμε στην μελέτη του συνδυασμού της μεθόδου μηχανικής μάθησης και στη μέθοδο που βασίζεται σε λεξικό ή ακόμη και να δημιουργήσουμε ένα λεξικό σχετικό με το πεδίο του τουρισμού για εγκυρότερα αποτελέσματα προκειμένου να βελτιωθεί η ακρίβεια του συναισθήματος την ταξινόμηση και την προσαρμοστική ικανότητα σε ποικίλα πεδία και περιοχές σε διαφορετικές γλώσσες.

Τέλος, πρέπει να βρεθεί λύση στους περιορισμούς που συναντώνται στο πεδίο της ανάλυσης συναισθήματος και της εξόρυξης γνώμης, που έχουν αναφερθεί παραπάνω, όπως για παράδειγμα ο εντοπισμός της ειρωνείας και του υπολανθάνοντος συναισθήματος, η χρήση ατημέλητου και πρόχειρου λόγου, η συχνή χρήση των emoticons και επιμέρους ακόμη λεκτικά και συντακτικά φαινόμενα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Bisio, F., Meda, C., Gastaldo, P., Zunino, R., Cambria, E., Bisio, F., … Cambria, E. (2017). Concept-Level Sentiment Analysis with SenticNet. *Socio-Affective Computing*. https://doi.org/10.1007/978-3-319-55394-8_9

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of the 12th international conference on World Wide Web (WWW '03)*. https://doi.org/10.1145/775152.775226

Das, S. R., & Chen, M. Y. (2001). *Yahoo! For Amazon: Sentiment Parsing from Small Talk on the Web*. *SSRN*. https://doi.org/10.2139/ssrn.276189

Duda, R. O., & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. *Leonardo*. https://doi.org/10.2307/1573081

Elmasri, R., & Navathe, S. B. (2010). The Relational Data Model and Relational Database Constraints. In *Fundamentals of Database Systems*.

Esuli, A., Sebastiani, F., & Moruzzi, V. G. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the 5th Conference on Language Resources and Evaluation*. https://doi.org/10.1.1.61.7217

Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). In *The First International Conference on Knowledge Discovery and Data Mining (KDD)*. https://doi.org/10.1.1.47.7462

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*.

Fuchs, M., Ricci, F., Cantoni, L., Ratti, C., & Blat, J. (2012). *Information and communication technologies in tourism 2012 : proceedings of the International Conference in Helsingborg, Sweden, January 25-27, 2012*. *MIT web domain*. https://doi.org/10.1007/978-3-7091-1142-0_3

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING)*. https://doi.org/10.3115/1220355.1220476

Gersho A. and Gray. R.M. (1992).Vector quantization and signal compression. *Kluwer Academic Publishers*.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2013.05.057

Gräbnera, D., & Zankerb, M. (2012). Classification of customer reviews based on sentiment analysis. *... Technologies in Tourism ...*. https://doi.org/10.1007/978-3-7091-1142-0_40

Gray, R. M. (1984). Vector quantization. *IEEE ASSP Magazine*. https://doi.org/10.1109/MASSP.1984.1162229

Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship to information access. *Presentation notes for UW/MS workshop on data mining*

Hartigan, J. A. (1989). *Clustering Algorithms*. *John Wiley & Sons, Inc*. https://doi.org/10.1002/0471725382.scard

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *Ldv Forum*. https://doi.org/10.1111/j.1365-2621.1978.tb09773.x

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*. https://doi.org/10.1145/1014052.1014073

Hu, M. and Liu, B. (2004). Opinion lexicon. Extracted from /www.cs.uic.edu Accessed: 2014-03- 20.

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*. *Natural Language Processing and Text Mining*. https://doi.org/10.1007/978-1-84628-754-1

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.4272

Khan, F. H., Qamar, U., & Bashir, S. (2016). Multi-Objective Model Selection (MOMS)-based Semi-Supervised Framework for Sentiment Analysis. *Cognitive Computation*. https://doi.org/10.1007/s12559-016-9386-8

Kim, K., & Lee, J. (2014). Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2013.07.022

Knauss, E., Damian, D., Poo-Caamano, G., & Cleland-Huang, J. (2012). Detecting and classifying patterns of requirements clarifications. In *2012 20th IEEE International Requirements Engineering Conference, RE 2012 - Proceedings*. https://doi.org/10.1109/RE.2012.6345811

König, A. C., & Brill, E. (2006). Reducing the human overhead in text categorization. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*. https://doi.org/10.1145/1150402.1150474

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. https://doi.org/10.1017/CBO9781139084789

Liu, B. (2012). *Sentiment Analysis and Opinion Mining "Synthesis lectures on human language technologies ". AAAI-2011 Tutorial*. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, J., Gusrialdi, A., Hirche, S., Blood, E. a., Ree, J. D. La, Member, S., … Zhang, Q. (2012). *Foundations of Machine Learning*. *Lecture Notes in Computer*

Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-642-34106-9_15

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. https://doi.org/citeulike-article-id:6083430

Martineau, J., & Finin, T. (2009). Delta TFIDF : An Improved Feature Space for Sentiment Analysis. *AAAI*.

Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2014.05.045

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*. https://doi.org/10.1016/j.asej.2014.04.011

Miguéns, J., Baggio, R., & Costa, C. (2008). NEIN Social media and Tourism Destinations: TripAdvisor Case Study. *Advances in Tourism Research*. https://doi.org/10.1088/1751-8113/44/8/085201

Nasukawa, T., & Jeonghee, Y. (2003). Sentiment Analysis : Capturing Favorability Using Natural Language Processing Definition of Sentiment Expressions. In *The Second International Conference on Knowledge Capture*. https://doi.org/10.1145/945645.945658

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *CEUR Workshop Proceedings*. https://doi.org/10.1016/j.knosys.2015.06.015

O'Reilly, T. (2007). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Design*. https://doi.org/10.2139/ssrn.1008839

Palakvangsa, S., Ayudhya, N.,et al..(2011) “Nebular: A Sentiment Classification System for the Tourism Business,” Proceedins. 8th Int’l Joint Conference. Computer Science and Software Engineering

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP ’02*. https://doi.org/10.3115/1118693.1118704

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*. https://doi.org/10.1108/eb046814

Ramos, J., Eden, J., & Edu, R. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of The First Instructional Conferences on Machine Learning*. https://doi.org/10.1.1.121.1424

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2015.06.015

Rexiline Ragini, J., & Rubesh Anand, P. M. (2016). Sentiment analysis: A comprehensive overview and the state of art research challenges. *Indian Journal of Science and Technology*. https://doi.org/10.17485/ijst/2016/v9i47/108465

Sayantani G., Sudipta R., and. Bandyopadhyay S., K (2013).. A tutorial review on Text Mining Lokesh Kumar et al, Journal of Global Research in Computer Science. *In International Journal of Advanced Research in Computer and Communication Engineering.*

Scholtes, J.,C, (2009).Text-Mining: The next step in search technology. *DESI-III Workshop Barcelona.*

Sebastiani, F. (2002). Machine {Learning} in {Automated} {Text} {Categorization}. *ACM Comput. Surv*. https://doi.org/10.1145/505282.505283

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., Herrera-Viedma, E., Montejo-Ráez, A., Díaz-Galiano, M. C., … Watson, R. T. (2016). Gap between Business

Schools and Presumed. *Communications of the Association for Information Systems*. https://doi.org/10.5465/AMLE.2007.25223459

Scholtes, J.,C., Voutilainen.A., (1995). A syntax-based part of speech analyser. *In Proc. of the Seventh Conference of the European Chapter of the Association for Computation al Linguistics*.

Thelwall, M., Paltoglou, G., Cai, D., & Kappas, A. (2011). Erratum: Sentiment in short strength detection informal text (Journal of the American Society for Information Science and Technology (2010) 61:12 (2544-2558)). *Journal of the American Society for Information Science and Technology*. https://doi.org/10.1002/asi.21475

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. https://doi.org/10.3115/1073083.1073153

Taylor, J., & Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*. https://doi.org/10.1177/1747016117738559

Wiegand, M., Balahur, A., Roth, B., Montoyo, A., & Klakow, D. (2010). A Survey on the Role of Negation in Sentiment Analysis. In *Workshop on Negation and Speculation in Natural Language Processing*. https://doi.org/10.1016/j.athoracsur.2007.06.076

Witten, I. H. (2003). Text mining. *International Journal of Computational Biology and Drug Design*. https://doi.org/10.1504/IJCBDD.2011.041412

Xiang, Z., Du, Q., Ma, Y., Fan, W., Qazi, A., Bux, K., … Bucur, C. (2015). Sentiment Analysis for Hotel Reviews. *Security Informatics*. https://doi.org/10.1016/j.procs.2013.09.094

Xianghua, F., Guo, L., Yanyan, G., & Zhiqiang, W. (2013). Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet

lexicon. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2012.08.003

ΠΑΡΑΡΤΗΜΑ Α

- **Links**

http://sete.gr/_fileuploads/entries/Online%20library/GR/131204_The%20Impact%20of%20Online%20Content%20on%20European%20Tourism_greek.pdf

Python Data Analysis Library — pandas: Python Data Analysis Library." [Online]. Available: http://pandas.pydata.org/. [Accessed: 30-May-2016].

https://www.smashingmagazine.com/2018/04/feature-testing-selenium-webdriver/

TripAdvisor Reports Second Quarter 2016 Financial Results http://files.shareholder.com/downloads/AMDA MMXS5/2604284471x0x903032/5C63829B-B3B4-49D9-8C04 E3C23B6D6A76/TripAdvisor_Reports_Second_Quarter_2016_Financial_Results.pdf

ΠΑΡΑΡΤΗΜΑ Β

- **Scraping (Παράδειγμα Ακρόπολης)**

```
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By
from selenium.webdriver.common.action_chains import ActionChains
from selenium import webdriver
from selenium.common.exceptions import NoSuchElementException, TimeoutExc
ption
import time
from bs4 import BeautifulSoup
import requests
from datetime import datetime, timedelta
import re

def dateconverter(date):
    if 'Reviewed today' in date:
        date = datetime.strftime(datetime.now(), '%Y-%m-%d')
    elif 'Reviewed yesterday' in date:
        date = datetime.strftime(datetime.now() - timedelta(1), '%Y-%m-%d')
    elif '2 days ago' in date:
        date = datetime.strftime(datetime.now() - timedelta(2), '%Y-%m-%d')
    elif '3 days ago' in date:
        date = datetime.strftime(datetime.now() - timedelta(3), '%Y-%m-%d')
    elif '4 days ago' in date:
        date = datetime.strftime(datetime.now() - timedelta(4), '%Y-%m-%d')
    elif '5 days ago' in date:
        date = datetime.strftime(datetime.now() - timedelta(5), '%Y-%m-%d')
    elif '6 days ago' in date:
        date = datetime.strftime(datetime.now() - timedelta(6), '%Y-%m-%d')
    elif '1 week ago' in date:
        date = datetime.strftime(datetime.now() - timedelta(7), '%Y-%m-%d')
    elif '2 weeks ago' in date:
```

```
        date = datetime.strftime(datetime.now() - timedelta(14), '%Y-%m-%d')
    elif '3 weeks ago' in date:
        date = datetime.strftime(datetime.now() - timedelta(21), '%Y-%m-%d')
    elif '4 weeks ago' in date:
        date = datetime.strftime(datetime.now() - timedelta(28), '%Y-%m-%d')
    elif '5 weeks ago' in date:
        date = datetime.strftime(datetime.now() - timedelta(32), '%Y-%m-%d')
    else:
        date = date
    return date

def ratingconverter(star_rating):
    if 'bubble_50' in str(star_rating):
        star_no = '5'
    elif 'bubble_40' in str(star_rating):
        star_no = '4'
    elif 'bubble_30' in str(star_rating):
        star_no = '3'
    elif 'bubble_20' in str(star_rating):
        star_no = '2'
    elif 'bubble_10' in str(star_rating):
        star_no = '1'
    else:
        star_no = '0'
    return star_no

counter=0
page_counter=0
review_counter = 0
url_part1 = "https://www.tripadvisor.com/Attraction_Review-g189400-d198706-Revi
ews-or"
url_part2 = "-Acropolis-Athens_Attica.html"

driver = webdriver.Firefox()
```

```
reviews= []
for counter in range(0,15660,10):

    page_counter+=1
    print('\nProcessing page No. '+str(page_counter)+'\n')
    page_URL = url_part1+str(counter)+url_part2
    print('Link: '+page_URL+'\n')

    driver.get(page_URL)

    if 'taLnk ulBlueLinks' in driver.page_source:
        more = driver.find_element_by_xpath('//span[@class="taLnk ulBlueLinks"]')

        try:
            more.click()
            print('clicked more')
            time.sleep(5)
            html = driver.page_source

        except (NoSuchElementException, TimeoutException):
            print("Exception thrown")
            pass;
    else:
        response = requests.get(page_URL)
        html = response.text #.content -> response as bytes, .text-> response as unicode/
string

    soup = BeautifulSoup(html, "lxml")

    reviews_num = soup.find_all("div", {"class":"reviewSelector"})

    print('Found '+str(len(reviews_num))+' reviews in the page.\n')
    print(type(reviews_num))
```

```
for review in reviews_num:
    reviews_dict= {}
    review_counter+=1
    print('\nProcessing review No. '+str(review_counter))
    review_id = review.get("id",{"class":"reviewSelector"})  #den 3erw t exei ginei me links kai emganizei ta data ths sel 15
    reviews_dict['id']= review_id
        #print('ID: '+review_id)
    name=review.find("span", {"class":"expand_inline scrname"}).text
    reviews_dict['name']= name
        #print('User name: '+name)
    title= review.find("span", {"class":"noQuotes"}).text
    reviews_dict['title']= title
        #print('Title: '+title)
    review_text= review.find("p", {"class":"partial_entry"}).text
    review_text = re.sub(r"\n", " ", review_text)
    reviews_dict['review_text']= review_text
        #print('Review content: '+review_text)
    try:
        origin = review.find("span", {"class":"expand_inline userLocation"}).text
    except AttributeError:
        origin = 'null'
    reviews_dict['origin']= origin
        #print('User origin: '+origin)
    date = review.find("span", {"class":"ratingDate relativeDate"}).text
        #print('Date: '+date) #an leei review yesterday prepei na metatrepeis thn plhr oforia sto x8esino date analoga me th mera pou kaneis to scraping
    date = dateconverter(date)
        #print('Converted Date: '+date)
    reviews_dict['date']= date
    star_rating=review.find("span", {"class":"ui_bubble_rating"})
        #print('Star rating: '+str(star_rating))
```

```
        star_no = ratingconverter(star_rating)
        reviews_dict['star_no']= star_no
            #print('Stars: '+star_no)




        reviews.append(reviews_dict)


print(reviews)
import pandas as pd
df = pd.DataFrame(reviews)
df.to_csv("acropolis_en.csv", index=False)
```