



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

Ψηφιακές Εφαρμογές και Καινοτομία

Ανάλυση συναισθήματος στον Τουρισμό

Κοντοζίδης Διονύσιος Παύλος

A.M.: inf.dig2105

Επιβλέπουσα καθηγήτρια:
Ελένη Χριστοπούλου

Κέρκυρα

15 Νοεμβρίου 2023

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση συναισθήματος στον Τουρισμό

Κοντοζίδης Διονύσιος Παύλος

A.M.: inf.dig2105

ΕΠΙΒΛΕΠΟΥΣΑ

ΣΥΝ - ΕΠΙΒΛΕΠΩΝ

Ελένη Χριστοπούλου, ΕΔΙΠ

**Θεόδωρος Ανδρόνικος,
Αναπληρωτής Καθηγητής**

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΥΝ - ΕΠΙΒΛΕΠΩΝ

Δημήτριος Ρίγγας, ΕΔΙΠ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Abstract

Today, the mass of data in the Worldwide Web is massive, a phenomenon that is caused by the rapid expansion of technology around the world, in addition to the numerous content sharing apps that are available on the internet. This, causes new needs to appear. One of these, is the need for the automated analysis of texts from companies that intend to observe what is trending and dissatisfies or satisfies it's clients, so that they can adjust their business accordingly. For that, the Sentiment Analysis technique is used fast and efficiently, a technique that contributes to the discovery of the emotions behind those texts from clients. A field that this method is specifically helpful, is that of the Tourism sector. Especially, for countries like Greece that attract waves of tourists each year.

This paper has to do with the sentiment analysis of multimedia, and specifically the texts that are relevant to the tourism sector of the Corfu prefecture. For that to occur, a dataset is produced from the external data that come from specific popular tourism platforms, that is the Tripadvisor and Booking platforms. Afterwards, the information is modified, so that it can be converted into the proper format for the analysis of machine learning models, and later the results are investigated. The approaches that will be used is those that have too do with a preconstructed dictionary, with words and sentiment polarities, in addition to machine learning models for the user-specific processing, specific to the dataset that the algorithm takes as input.

SUBJECT AREA: Sentiment Analysis, Data Mining

KEYWORDS: big data, natural language processing, tourism, Corfu, Tripadvisor, Booking

Περίληψη

Σήμερα, ο όγκος των δεδομένων στον Παγκόσμιο Ιστό είναι τεράστιος, ένα φαινόμενο που προκύπτει από την ταχύτατη εξάπλωση της τεχνολογίας ανά τον κόσμο, καθώς και από τις πολλαπλές εφαρμογές κοινοποίησης περιεχόμενου που διατίθενται στο διαδίκτυο. Αυτό, προκαλεί νέες ανάγκες να εμφανίζονται. Μία από αυτές είναι και η ανάγκη αυτοματοποιημένης ανάλυσης κειμένων από εταιρείες που σκοπεύουν να παρατηρήσουν τι βρίσκεται στις τάσεις και δυσανασχετεί ή ευχαριστεί τους πελάτες τους, για να προσαρμόσουν την επιχείρησή τους ανάλογα. Για την εφαρμογή αυτού, χρησιμοποιείται γρήγορα και αποδοτικά η τεχνική Sentiment Analysis (Ανάλυση Συναισθήματος), τεχνική που συνεισφέρει στην ανακάλυψη του συναισθήματος πίσω από τα κείμενα πελατών. Ένα πεδίο που αυτή η μέθοδος είναι ιδιαίτερα βοηθητική, είναι αυτό του Τουρισμού. Ειδικά, για χώρες όπως την Ελλάδα που ελκύουν κύματα τουριστών κάθε χρόνο.

Η εργασία έχει να κάνει με την ανάλυση συναισθημάτων σε πολυμέσα, και συγκεκριμένα σε κείμενα που αφορούν τον τομέα του τουρισμού στον νομό Κέρκυρας. Για να πραγματοποιηθεί αυτό, παράγεται ένα σύνολο δεδομένων από τα εξαγόμενα στοιχεία που προέρχονται από συγκεκριμένες δημοφιλείς ταξιδιωτικές πλατφόρμες, δηλαδή τις Tripadvisor και Booking πλατφόρμες. Στη συνέχεια, επεξεργάζονται οι πληροφορίες, έτσι ώστε να μετατραπούν σε σωστή μορφή για την ανάλυση από μοντέλο μηχανικής μάθησης και ύστερα εξετάζονται τα αποτελέσματα. Οι προσεγγίσεις που θα χρησιμοποιηθούν είναι αυτές που αφορούν την χρήση προκατασκευασμένου λεξικού με λέξεις και τις αντίστοιχες τιμές πολικότητας, καθώς και μοντέλα μηχανικής μάθησης για την εξατομικευμένη επεξεργασία συγκεκριμένα στο σύνολο δεδομένων που δέχεται ως είσοδο ο αλγόριθμος.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Ανάλυση Συναισθήματος, Εξόρυξη δεδομένων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: "μεγάλα" δεδομένα, επεξεργασία φυσικής γλώσσας, τουρισμός, Κέρκυρα, Tripadvisor, Booking

Περιεχόμενα

1	Εισαγωγή	10
1.1	Σκοπός & Στόχοι	12
1.2	Δομή εργασίας	12
1.2.1	Μεθοδολογία	12
1.2.2	Οργάνωση	13
2	Βιβλιογραφική ανασκόπηση	13
2.1	Ιστορία	17
3	Ερευνητικός χώρος	18
3.1	Τουρισμός	19
3.1.1	Τουρισμός στην Ελλάδα	19
3.2	Ανάλυση συναισθημάτων	20
3.2.1	Επεξεργασία με την χρήση λεξικού	21
3.2.2	Επεξεργασία με την χρήση αλγορίθμων μηχανικής μάθησης	26
3.3	Web scraping	28
3.4	Τουριστικές πλατφόρμες	31
3.4.1	Σχόλια	31
3.5	Μετρικές αξιολόγησης	33
3.6	Μοντέλα	34
3.6.1	Bag-of-Words	34
3.6.2	TF-IDF	35
3.7	Πρόβλημα ανισορροπίας δεδομένων	37
4	Μεθοδολογική διαδικασία	38
4.1	Συλλογή Δεδομένων	39
4.1.1	Συλλογή λίστας ξενοδοχείων	40
4.1.2	Εξαγωγή στοιχείων	48
4.2	Προεπεξεργασία	63
4.3	Ανάλυση Δεδομένων	64
4.4	Ταξινόμηση Συναισθήματος	75
4.5	Τελική τιμή συναισθήματος	75
5	Αποτελέσματα	76
5.1	Tripadvisor Αξιοθέατα	77
5.2	Tripadvisor Ξενοδοχεία	79
5.3	Booking Ξενοδοχεία	80
6	Συμπεράσματα και μελλοντικές προσθήκες	81
6.1	Μελλοντικές Προσθήκες	81
6.2	Συμπεράσματα	82

List of Figures

1	Μεθοδολογία Ανάλυσης Συναισθήματος	12
2	Γράφημα που αφορά τον τύπο κανονικοποίησης	23
3	Παράδειγμα υπερεπιπέδου	27
4	Διάγραμμα με τα βήματα μεθολογίας	38
5	Πλήθος κειμένων κριτικών (στα αριστερά) και βαθμολογίας κριτικών (στα δεξιά) . .	65
6	Ποσότητα βαθμολογιών ανά τιμή βαθμολογίας	66
7	WordCloud Tripadvisor Αξιοθέατων	67
8	Γράφημα με σύνολα από κάθε χρονιά στο σετ δεδομένων των αξιοθέατων	67
9	Πλήθος θετικών κριτικών (στα αριστερά), αρνητικών κριτικών (κεντρικά) και βαθμολογίας κριτικών (στα δεξιά)	68
10	Πλήθος βαθμολογιών ανά τιμή βαθμολογίας σε κλίμακα 1-10	69
11	WordCloud Booking θετικών κριτικών για ξενοδοχεία	70
12	WordCloud Booking αρνητικών κριτικών για ξενοδοχεία	70
13	Γράφημα με σύνολα από κάθε χρονιά στο σετ δεδομένων των ξενοδοχείων του booking	71
14	Πλήθος κειμένων κριτικών (στα αριστερά) και βαθμολογιών (στα δεξιά)	72
15	Πλήθος βαθμολογιών ανά τιμή βαθμολογίας στα ξενοδοχεία του Tripadvisor	73
16	WordCloud Tripadvisor κριτικών για ξενοδοχεία	73
17	Γράφημα με σύνολα από κάθε χρονιά στο σετ δεδομένων των ξενοδοχείων του tripadvisor	74
18	Αναπαράσταση στοιχείων	76
19	Αποτελέσματα στο σύνολο δεδομένων των αξιοθέατων με διάφορες τεχνικές sampling	77
20	Αποτελέσματα στο σύνολο δεδομένων των ξενοδοχείων από το Tripadvisor με διάφορες τεχνικές sampling	79
21	Αποτελέσματα στο σύνολο δεδομένων των ξενοδοχείων από Booking με διάφορες τεχνικές sampling	80

List of Images

1	Βιβλιοθήκες για την συλλογή ξενοδοχείων από το Booking.com	40
2	Ο ιστότοπος και τα headers που θα χρησιμοποιηθούν ως παράμετροι της βιβλιοθήκης requests	40
3	Το κύριο τμήμα της εφαρμογής με την επανάληψη για την συλλογή των καταλυμάτων	41
4	HTML κώδικας όπου εμφανίζεται το στοιχείο που αναζητείται από την εφαρμογή .	42
5	Διαγραφή τυχόν σημείων στίξης που συγκρούονται με το μορφότυπο csv και αποθήκευση σε αρχείο	42
6	Εισαγωγή διαφορετικού σετ βιβλιοθηκών για την χρήση selenium αντί για requests .	43
7	Εγκατάσταση selenium και εισαγωγή παραμέτρων	43
8	Ιστοσελίδα προς επίσκεψη και άρνηση cookies μέσω εντολών	44
9	Πάτημα του κουμπιού που δείχνει το μπλέ βέλος για την εμφάνιση του κουμπιού άρνησης των cookies	44
10	Πάτημα του κουμπιού που δείχνει το μπλέ βέλος για την άρνηση των cookies	45
11	Κεντρικός βρόγχος για την εύρεση όλων των ξενοδοχείων και των συνδέσμων	45
12	Παράδειγμα πηγαίου κώδικα από ένα ξενοδοχείο κατά την αναζήτηση στο Tripadvisor	46
13	Κατά την εύρεση στοιχείων από το πρόγραμμα, δημιουργείται αρχείο (αν δεν υπάρχει) και έπειτα συμπληρώνεται	46
14	Παράδειγμα πηγαίου κώδικα από ένα αξιοθέατο κατά την αναζήτηση στο Tripadvisor	47
15	Εύρεση στοιχείων στον πηγαίο κώδικα βάσει του xpath	47
16	Αποθήκευση στοιχείων στις λίστες	47
17	Διάβασμα αρχείου ξενοδοχείων Κέρκυρας	48
18	Πάτημα του κουμπιού που αναδεικνύεται με κόκκινο χρώμα για το άνοιγμα του tab με τις κριτικές	49
19	Εμφάνιση του συνδέσμου στην μπάρα διευθύνσεων του φυλλομετρητή μετά το άνοιγμα του tab με τις κριτικές	49
20	Παράδειγμα σχολίου στον ιστοχώρο Booking	50
21	Άνοιγμα tab με κριτικές και απόρριψη cookies	50
22	Καθαρισμός λιστών, ορισμός μεταβλητών start και end και επίσκεψη στην σελίδα με τα ξενοδοχεία	51
23	Εύρεση στοιχείων και έλεγχος για τυχόν σφάλματα	51
24	Μπάρα paginator για την αλλαγή σελίδας	51
25	Αν τα στοιχεία στην σελίδα έχουν πάνω από μία σελίδα εξάγεται ο αριθμός της τελευταίας σελίδας	52
26	Κεντρικός βρόγχος για την άντληση των επιθυμητών στοιχείων	53
27	Γέμισμα των λιστών με τα δεδομένα	53
28	Μετακίνηση στην επόμενη σελίδα, στην περίπτωση όπου υπάρχει	54
29	Αναμονή ανά κάθε ξενοδοχείο και σελίδα κριτικής, καθώς και παρουσίαση μηνύματος κατά το τέλος διαβάσματος οποιοδήποτε ξενοδοχείου	54
30	Δημιουργία ίσου πλήθους λιστών και φόρτωση στο αρχείο δεδομένων	55
31	Βιβλιοθήκες για το πρόγραμμα που σχετίζεται με τα αξιοθέατα	56
32	Ορισμός των επαναλήψεων για την εξαγωγή των στοιχείων	56
33	Ορισμός επόμενου συνδέσμου που θα επισκεφθεί ο φυλλομετρητής	56
34	Παράδειγμα ξενοδοχείου με μέση βαθμολογία και πλήθος κριτικών από το 1 έως το 5	57
35	Παράδειγμα κριτικής και εμφάνιση δεδομένων που αποθηκεύονται σε αρχείο	57
36	Επίσκεψη στη σελίδα και έλεγχος για σφάλμα κατά την μετάδοση των πακέτων . .	58

37	Εξαγωγή βαθμολογίας	58
38	Σημείο στον πηγαίο κώδικα από όπου εξάγεται η βαθμολογία κάθε ατομικής κριτικής	59
39	Εξαγωγή ημερομηνίας υποβολής σχολίου	59
40	Κλάση div που αναζητείται από την εφαρμογή	59
41	Εύρεση και αποθήκευση τίτλων σχολίων και κειμένων σχολίων	60
42	Σημείο στον κώδικα HTML με το κείμενο και τίτλο σχολίου	60
43	Διαγραφή σημείων όπου εμφανίζεται ο χαρακτήρας semicolon και αποθήκευση λοιπών στοιχείων	61
44	Αποθήκευση σε αρχείο ανά δέκα ξενοδοχεία με την διαδικασία μετατροπής των λυστών στο ίδιο μέγεθος	61
45	Εύρεση βαθμολογίας στον πηγαίο κώδικα και διαγραφή άχρηστων χαρακτήρων . . .	62
46	Σημείο στον πηγαίο κώδικα με την τιμή βαθμολογίας κάθε κριτικής	62
47	Έλεγχος σχετικά με το αν υπάρχουν κριτικές στην τωρινή σελίδα και αποθήκευση λοιπών στοιχείων σε λίστες	62

List of Tables

1	Σύγκριση βιβλιογραφίας	16
2	Πίνακας με την λέξη και την κατηγορία, βάσει τον διαχωρισμό του συστήματος . . .	22
3	Παρουσίαση αποτελεσμάτων με τις λέξεις και τα αντίστοιχα συναισθήματα	22
4	Στίλνες αρχείου δεδομένων Tripadvisor	32
5	Στίλνες αρχείου δεδομένων Booking	32
6	Αποτελέσματα από παράδειγμα TF-IDF	37

Πρόλογος

Η μελέτη αυτή προέκυψε από την θέληση μου να εξερευνήσω τις πρόσφατες αλλαγές που υπέστη η χώρα μας την περίοδο του Κορονοϊού, αλλά και τις συμπτώσεις που παρατηρήθηκαν από αυτές τις αλλαγές. Ο τουρισμός επιλέχθηκε ως κύριο θέμα, διότι αποτελεί μία από τις μεγαλύτερες βιομηχανίες συσσώρευσης πλούτου της χώρας μας. Επίσης, έγινε επιλογή της Κέρκυρας ως δείγμα για την εξερεύνηση του θέματος, μιας και αποτελεί ένα από τα πιο δημοφιλείς νησιά με τουρίστες να το επισκέπτονται καθ' όλο τον χρόνο. Για να πραγματοποιηθεί αυτή η διαδικασία, απαιτείται η ανάλυση των κειμένων πελατών από τις δημοφιλέστερες τουριστικές πλατφόρμες. Αυτή η διπλωματική εργασία (capstone project) συντάχθηκε στο Ιόνιο Πανεπιστήμιο και συγκεκρωμένα στο μεταπτυχιακό τμήμα «Ψηφιακές Εφαρμογές και Καινοτομία». Η επίβλεψη της εργασίας πραγματοποιήθηκε από την διδάσκουσα Ελένη Χριστοπούλου, μέλος του Εργαστηριακού Διδακτικού Προσωπικού (ΕΔΙΠ).

1 Εισαγωγή

Η διακίνηση δεδομένων στον Παγκόσμιο σήμερα ξεπερνάει τα μεγέθη των GBs και TBs καθημερινά (Duarte, F., 2023). Αυτό προκύπτει από την ευκολία χρήσης συσκευών σήμερα, με άτομα από όλες τις ηλικίες να έρχονται όλο και πιο πολύ σε επαφή με την τεχνολογία, καθώς και από το γεγονός ότι υπάρχει ένας συνεχώς αυξανόμενος αριθμός από εφαρμογές που επιτρέπουν στον χρήστη να κοινοποιήσει περιεχόμενο. Το κοινοποιημένο περιεχόμενο αυτό μπορεί να αφορά κείμενα, ήχους, βίντεο αλλά και εικόνες, με τα στοιχεία αυτά να διαμοιράζονται ανά τον κόσμο ελεύθερα, χωρίς γεωγραφικούς περιορισμούς που εμποδίζουν στην εξάπλωση των πληροφοριών αυτών. Αυτό, προκαλεί ένα πρόβλημα με τον όγκο των δεδομένων που κυκλοφορεί στο διαδίκτυο να είναι εξωπραγματικά μεγάλος, καθώς και νέες ανάγκες. Οι ανάγκες αυτές αφορούν την ανάγκη από τις επιχειρήσεις να συμμορφώνονται με τις εποχές και να ψηφιοποιούνται, δηλαδή να παρουσιάζουν τα αγαθά ή τις υπηρεσίες που προσφέρουν ψηφιακά με τέτοιο τρόπο που να μπορούν να αυξήσουν τους πελάτες τους δια τον Παγκόσμιο Ιστό, κάτι ιδιαίτερα εφικτό σήμερα με όλο και περισσότερο κόσμο να έχει πρόσβαση στο ίντερνετ και δεδομένου του αυξημένου καταναλωτισμού που παρατηρείται σήμερα. Οι υπηρεσίες αυτές πρέπει να είναι ελκυστικές και να επιτρέπουν τον χρήστη να εξυπηρετεί τις ανάγκες του γρήγορα και διαδραστικά. Διάφορες βιομηχανίες επωφελούνται από αυτό και έχουν ιδιαίτερα μεγάλη ζήτηση διαρκώς ανά τον κόσμο

Μια τέτοια βιομηχανία που είναι ιδιαίτερα δημοφιλής είναι ο τομέας του τουρισμού. Ο τομέας αυτός αποτελεί μια ιδιαίτερα μεγάλη βιομηχανία που αποτελεί σημαντική πηγή εσόδων σε τουριστικές χώρες, μεταξύ των οποίων και της Ελλάδας. Αυτή η βιομηχανία αποτελεί έναν βασικό πυλώνα στήριξης της ελληνικής οικονομίας, με την Ελλάδα να αποτελεί έναν ελκυστικό τουριστικό προορισμό που επιφέρει σημαντικά κέρδη στο κράτος ετησίως. Φυσικά, μία τόσο δημοφιλής βιομηχανία απαιτεί την ανάγκη από τους επισκέπτες, που επισκέπτονται είτε για λόγους αναψυχής είτε για την εργασία τους, να έχουν εύκολα πρόσβαση σε κάποια πράγματα που αφορούν τον τομέα αυτόν. Τέτοια είναι, τα καταλύματα που θα διαμείνουν, τα μέσα μεταφοράς που θα τους μεταφέρουν στον προορισμό τους γρήγορα, τα καλύτερα εστιατόρια που θα τους επιτρέψουν να έρθουν σε επαφή με νέες γεύσεις από μία διαφορετική κουζίνα, αλλά και διάφορα αξιοθέατα που πιθανώς θα επισκεφθούν κατά την επίσκεψή τους. Για αυτό τον λόγο, οι κριτικές είναι απαραίτητες και επηρεάζουν σημαντικά την αντίληψη και επιλογή ενός μελλοντικού τουρίστα.

Ένας επιχειρηματίας πρέπει να είναι σε θέση να παρακολουθεί αυτές τις κριτικές, για να μπορεί να εξακριβώσει αν η εμπειρία προηγούμενων επισκεπτών ήταν ικανοποιητική, που σημαίνει ότι δεν χρειάζεται να αλλάξει κάτι στον τρόπο που προσφέρει το προϊόν του, ή αν η εμπειρία ενός προηγούμενου επισκέπτη ήταν δυσάρεστη, που δείχνει ότι απαιτείται ταχύτατη αλλαγή στον τρόπο που λειτουργεί η επιχείρησή του. Η Ανάλυση Συναισθήματος (Sentiment Analysis) συνεισφέρει στην διαδικασία αυτή. Επιτρέπει στους επιχειρηματίες να παρακολουθήσουν τα συναισθήματα από προηγούμενους πελάτες και να καταλάβουν τις ανάγκες τους, καθώς και να προσαρμόσουν την υπηρεσία που προσφέρουν σε αυτές τις ανάγκες.

Όπως έχει προαναφερθεί, η Ελλάδα αποκτά ετησίως σημαντικό πλούτο από τούτη την βιομηχανία. Αυτό προκύπτει από τα κύματα επισκεπτών που εμφανίζονται κάθε χρόνο, συνήθως κατά την περίοδο του Καλοκαιριού. Οι επισκέπτες επισκέπτονται τα νησιά, για να βιώσουν νέες εμπειρίες και να δουν νέα τοπία. Ένα ιδιαίτερα δημοφιλές νησί αποτελεί και η Κέρκυρα. Σε σχέση με κάποια άλλα νησιά, κύματα τουριστών ρέουν καθ' όλον τον χρόνο, και όχι μόνο τους καλοκαιρινούς μήνες, αν και σε μικρότερη συχνότητα. Οι τουρίστες επισκέπτονται δημοφιλείς πλατφόρμες για να εντοπίσουν τις καλύτερες προσφορές που θα καλύψουν τις ανάγκες τους. Κάποιες από τις δημοφιλέστερες πλατφόρμες είναι το Booking και το Tripadvisor, με κριτικές από χιλιάδες πελάτες από άτομα διαφόρων εθνών.

Η εργασία αυτή περιέχει συναισθηματική ανάλυση κριτικών που αφορά τα αξιοθέατα και ξενοδοχεία από τον νομό Κέρκυρα, από τις παραπάνω πλατφόρμες, για την εξέταση των δεδομένων. Η ανάλυση πραγματοποιείται με δύο βασικές προσεγγίσεις. Η μία προσέγγιση έχει να κάνει με την χρήση έτοιμων λέξεων με προκαθορισμένες τιμές συναισθήματος, δηλαδή την προσέγγιση με λεξικό. Από την άλλη, μία άλλη προσέγγιση που εξετάζεται είναι η χρήση αλγορίθμων υπολογιστή για την ανάλυση των δεδομένων. Σε σύγκριση με την προηγούμενη προσέγγιση, η προσέγγιση με αλγορίθμους επιτρέπει την δημιουργία νέων τιμών συναισθήματος, συγκεκριμένων στα δεδομένα του χρήστη. Η εξέταση της αποδοτικότητας αυτών των προσεγγίσεων πραγματοποιείται με κάποιες μεθόδους. Οι μέθοδοι αυτοί είναι τα Vader, TextBlob και Sentiwordnet για τα έτοιμα λεξικά, και Linear Regression, LinearSVC και Naive Bayes για τους αλγορίθμους.

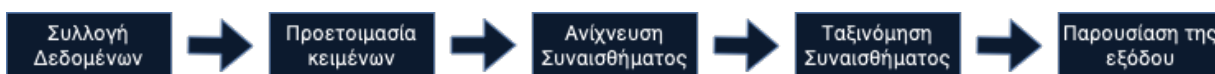
1.1 Σκοπός & Στόχοι

Ο σκοπός είναι η ανάλυση και η ταξινόμηση των κριτικών που αφορούν τον τουρισμό βάσει των βαθμολογιών χρηστών, καθώς και η παροχή σύστασης για μια πιθανή νέα κριτική από τους χρήστες. Οι στόχοι της εργασίας αναφέρονται παρακάτω.

- Να παρατηρηθούν οι τάσεις από τις τουριστικές δραστηριότητες, καθώς και να αναλυθεί η ποσότητα της ψυχαγωγίας των τουριστών, μετά από την εμπειρία τους
- Να αναλυθούν οι προτιμήσεις των καταναλωτών σχετικά με τα αξιοθέατα και τα ξενοδοχεία του νομού Κέρκυρας, από δημοφιλείς ταξιδιωτικές πλατφόρμες

1.2 Δομή εργασίας

1.2.1 Μεθοδολογία



Διάγραμμα 1: Μεθοδολογία Ανάλυσης Συναισθήματος

Το τμήμα της μεθοδολογίας έχει να κάνει με την μετατροπή αδόμητων κειμένων σε πληροφορία. Τα βήματα είναι τα παρακάτω (Barney, N., 2023):

- Συλλογή δεδομένων
- Προετοιμασία κειμένων
- Ανίχνευση Συναισθήματος
- Ταξινόμηση Συναισθήματος
- Παρουσίαση της εξόδου

Το πρώτο βήμα είναι η συλλογή δεδομένων. Πιθανώς, αποτελεί το σημαντικότερο βήμα από όλα τα στάδια της μεθοδολογίας, διότι η ποιότητα των δεδομένων επηρεάζει το τελικό αποτέλεσμα. Η διαδικασία αυτή είτε πραγματοποιείται με την χρήση κάποιου API ή προγράμματος για την εξαγωγή πληροφοριών από ιστοσελίδες, είτε από την εφαρμογή βάσης δεδομένων. Η χρήση API ή προγράμματος έχει να κάνει με την κατασκευή συνόλου δεδομένων από τον ερευνητή, με κείμενα να προέρχονται συνήθως από διάφορα δημόσια φόρουμ, blog και κοινωνικά μέσα. Σε αντίθεση, η άλλη προσέγγιση έχει να κάνει με την χρήση κάποιου προ-κατασκευασμένου συνόλου στοιχείων από κάποια βάση δεδομένων που είναι διαθέσιμη στο διαδίκτυο. Μετά από την συλλογή των πληροφοριών, ξεκινάει το βήμα της προετοιμασίας των δεδομένων.

Στο στάδιο της προετοιμασίας των δεδομένων, επεξεργάζονται οι πληροφορίες στο σύνολο δεδομένων για να μετατραπούν σε μία μορφή που είναι αποδοτική για την ανάλυση συναισθήματος. Για παράδειγμα, πολλαπλές απόψεις και διάφορα συναισθήματα παρουσιάζονται στα κείμενα και κάνουν το αρχείο να είναι τεράστιο και ανοργάνωτο. Σε αυτό βοηθάει η διαδικασία προετοιμασίας των κειμένων. Η προετοιμασία (ή προ-επεξεργασία) αφαιρεί το μη σχετικό περιεχόμενο που δεν θεωρείται αναγκαίο για τους σκοπούς της έρευνας. Τα πιο συνηθισμένα βήματα στο στάδιο αυτό είναι το tokenization, η ληματοποίηση, το stemming, η αφαίρεση stopwords και η μετατροπή

κεφαλαίων γραμμμάτων σε πεζά. Στη συνέχεια, ανιχνεύεται το συναίσθημα.

Η ανίχνευση του συναισθήματος εξερευνά την αντικειμενικότητα κάθε στοιχείου που οδηγεί στην ταξινόμηση του συναισθήματος. Στο επόμενο βήμα κατηγοριοποιείται κάθε αντικειμενική πρόταση, ανάλογα με την πολικότητα συναισθήματος (Αρνητική, Ουδέτερη, Θετική). Τέλος, γίνεται παρουσίαση της εξόδου, όπου η σημαντική πληροφορία που εξορύχτηκε παρουσιάζεται συνήθως με την βοήθεια διαγράμματος ή γραφήματος.

1.2.2 Οργάνωση

Στο πρώτο κεφάλαιο αναλύεται η περίληψη όπου αναλύεται συνοπτικά το κείμενο. Στη συνέχεια, στο δεύτερο κεφάλαιο αναφέρεται η εισαγωγή της μελέτης, με το τρίτο κεφάλαιο να ακολουθεί για την ανάλυση δομής εργασίας. Ακόμα, παρουσιάζεται η βιβλιογραφική ανασκόπηση στο τέταρτο κεφάλαιο και ο ερευνητικός χώρος εργασίας στο πέμπτο κεφάλαιο. Μετέπειτα, αναλύεται η κύρια τεχνική που εφαρμόζεται στην εργασία, της Ανάλυσης Συναισθήματος, στο έκτο κεφάλαιο, ενώ στο έβδομο κεφάλαιο εμφανίζεται η μεθοδολογική διαδικασία. Τέλος, στο όγδοο κεφάλαιο αναφέρονται τα συμπεράσματα και οι μελλοντικές προτάσεις για την βελτίωση.

2 Βιβλιογραφική ανασκόπηση

Στο διαδίκτυο υπάρχουν αρκετές μελέτες που εφαρμόζουν την τεχνική της Ανάλυσης Συναισθημάτων, αλλά είναι λιγιστές αυτές που αφορούν συγκεκριμένα τον τομέα της φιλοξενίας και τουρισμού.

Όμοια με τα δεδομένα που θα αναλυθούν σε αυτή την εργασία είναι η μελέτη (Valdivia, A., Luzón, M. V., & Herrera, F., 2017) που αφορά αποκλειστικά τα δεδομένα που εξάγονται από την δημοφιλή πλατφόρμα τουρισμού Tripadvisor. Στη συγκεκριμένη εργασία μελετώνται οι κριτικές που σχετίζονται με τρία δημοφιλή μνημεία της Ισπανίας: τα Alhambra, Mezquida Cordoba και Sagrada Familia. Θεωρείται ότι οποιαδήποτε κριτική περιέχει μία ή δύο φούσκες είναι αρνητική, όταν περιέχει τρεις είναι ουδέτερη και όταν ξεπερνάει τις τρεις, τότε είναι θετική. Στην συνέχεια, εφαρμόζονται τέσσερις μέθοδοι Ανάλυσης Συναισθημάτων και συγκρίνονται τα αποτελέσματα. Οι μέθοδοι αυτοί είναι τα SentiStrength, Bing, Syuzhet και CoreNLP, που ανήκουν στην προσέγγιση της Ανάλυσης Συναισθημάτων βάσει λεξικού. Τα αποτελέσματα που επιδεικνύονται στη μελέτη αυτή είναι αποθαρρυντικά, με τις τεχνικές αυτές να δείχνουν ότι δεν μπορεί να κριθεί σωστά από το κείμενο το συναίσθημα που αφορά τις κριτικές. Συγκεκριμένα, τα SentiStrength και Syuzhet ταξινομούν σωστά το 57.48% και 45.10% των αρνητικών κριτικών αντίστοιχα, αλλά αποτυγχάνουν στην ταξινόμηση των θετικών σχολίων με περίπου 20% των σχολίων να ταξινομούνται λανθασμένα. Σε αντίθεση, το Bing και CoreNLP ταξινομούν πιο σωστά τις αρνητικές κριτικές, αλλά ταξινομούν λανθασμένα πάνω από 30% των θετικών σχολίων, που ξεπερνάει τα λανθασμένα αποτελέσματα των προηγούμενων μεθόδων.

Μία μελέτη που επίσης αναλύει το θέμα της Ανάλυσης Συναισθημάτων στον τομέα του τουρισμού με την ίδια προσέγγιση είναι αυτή που εξετάζει τρεις διαδικτυακές υπηρεσίες και συγκρίνει τα αποτελέσματα. Οι υπηρεσίες αυτές είναι το AlchemyAPI, το Semantria και το Text2Data (Gao, S., Hao, J., & Fu, Y., 2015). Το σετ δεδομένων που χρησιμοποιήθηκε αντλήθηκε από την σελίδα "<http://arguana.com>" που περιέχει 2100 κείμενα με κριτικές που αφορούν ορισμένες δημοφιλείς τοποθεσίες γραμμένες σε αγγλική γλώσσα. Το σετ δεδομένων χειροκίνητα έχει ταξινομηθεί με τέτοιο τρόπο, έτσι ώστε κάθε κείμενο να ανήκει σε μία από τρεις κατηγορίες, θετική κριτική, αρνητική ή τεκμηριωμένη πληροφορία. Τα αποτελέσματα από την έρευνα είναι αρκετά θετικά, με τα αποτελέσματα από το AlchemyAPI να είναι τα υψηλότερα με precision 88.36%. Οι θετικές κριτικές ταξινομούνται πιο σωστά με 91.14% ανάκληση, ενώ συγκριτικά οι αρνητικές κριτικές ταξινομούνται σωστά με 84.03% ανάκληση. Παρόμοια, αν και με συγκριτικά πιο χαμηλά αποτελέσματα, είναι το Text2Data με ανάκληση ίση με 89.56% για τις θετικές κριτικές, ενώ για τις αρνητικές το ποσοστό αυτό μεταβάλλεται σε 64.07%. Ωστόσο, η έρευνα ανακάλυψε ότι το AlchemyAPI και το Text2Data δεν δουλεύουν σωστά με ουδέτερες κριτικές, καθώς δεν ταξινομούσαν σχεδόν κανένα κείμενο στην ουδέτερη κατηγορία. Επίσης, παρατηρήθηκε ότι η ανάκληση για το Semantria ήταν πολύ χαμηλή, με μόνο το 37.31% από τις αρνητικές κριτικές να ταξινομήθηκε σωστά. Συγκριτικά, το recall από τις θετικές κριτικές ήταν πιο σωστό με ποσοστό 82.23%. Παρόλα αυτά, το ποσοστό αυτό είναι αρκετά πιο χαμηλό από τις άλλες δύο υπηρεσίες και, σε συνδυασμό με το ποσοστό ανάκλησης των αρνητικών κριτικών που είναι εξαιρετικά χαμηλό, δείχνει ότι αποτέλεσε την χειρότερη μέθοδο από αυτές που εξετάστηκαν. Η έρευνα παρατήρησε τρία κύρια προβλήματα, δηλαδή ότι το AlchemyAPI έχει περιορισμούς σχετικά με το μήκος των κειμένων καθώς δεν δέχεται πολύ μικρά κείμενα, το Text2Data είναι πολύ αργό, δηλαδή μη αποδοτικό καθώς είναι χρονοβόρο, και τα τρία εργαλεία υποστηρίζουν περιορισμένο αριθμό γλωσσών. Παρεμπιπτόντως, είναι σημαντικό να αναφερθεί ότι και οι τρεις μέθοδοι που εξετάστηκαν είχαν καλύτερα αποτελέσματα στις θετικές κριτικές ξενοδοχείων, ενώ υστερούσαν στην εξέταση των αρνητικών σχολίων.

Στην προσέγγιση όπου χρησιμοποιείται τεχνητή νοημοσύνη, και συγκεκριμένα αλγόριθμοι μηχανικής μάθησης ανήκει η μελέτη όπου ερευνώνται τα κείμενα τουρισμού από το δημοφιλές κοινωνικό μέσο Twitter, όσον αφορά την περίοδο της πανδημίας COVID, στην Ταϊλάνδη (Leelawat, N. et al., 2022). Τα σχόλια που μελετήθηκαν σχετίζονταν με τις κύριες πόλεις της Ταϊλάνδης, δηλαδή το Μπανγκόκ (Bangkok), το Πουκέτ (Phuket) και το Chiang Mai. Οι αλγόριθμοι μηχανικής μάθησης που εξετάστηκαν είναι ο CART, ο Random Forest και ο Support Vector Machine. Στην έρευνα δημιουργήθηκαν τρία σετ δεδομένων, ένα για την κάθε πόλη, με αυτό του Bangkok να περιέχει 61.985 στοιχεία, αυτό του Phuket να έχει 83.902 δεδομένα και αυτό του Chiang Mai να περιέχει 4.693 στοιχεία. Η έρευνα έδειξε ότι το συναίσθημα ανιχνεύθηκε βέλτιστα στην πόλη Bangkok με πολύ υψηλές τιμές ακρίβειας και F1-score, με επίσης θετικά αποτελέσματα ακρίβειας και F1-score στο Πουκέτ, αν και σχετικά χαμηλότερα συγκριτικά με αυτά του Bangkok. Ωστόσο, τα αποτελέσματα που σχετίζονταν με την πόλη Chiang Mai ήταν αρκετά χαμηλά, εξαιτίας του μικρού πλήθους στοιχείων μέσα στο σετ. Ομοίως στην ανάλυση σχετικά με την πρόθεση, τα αποτελέσματα είχαν την ίδια κατάταξη με αυτά του Bangkok να επιστρέφουν υψηλότερες τιμές ακρίβειας και F-Measure, του Πουκέτ να ακολουθεί και τελευταία να είναι αυτά που σχετίζονται με το Chiang Mai.

Ακόμα, μία ενδιαφέρουσα μελέτη πάνω στο θέμα της ανάλυσης συναισθημάτων στο συγκεκριμένο πεδίο, είναι η έρευνα όπου αναλύονται ισπανικά κείμενα για εστιατόρια και καταλύματα που προέρχονται από την σελίδα του Tripadvisor (García, A., Gaines, S., & Linaza, M. T., 2012). Φυσικά, η εύρεση της σωστής πολικότητας συναισθήματος που αφορά κείμενα στην ισπανική γλώσσα είναι πιο δύσκολη διαδικασία από την αντίστοιχη στην αγγλική γλώσσα. Για αυτό, εφαρμόζεται ένα δικό τους λεξικό που περιέχει 6000 ισπανικές λέξεις με την πολικότητα, ανεξαρτήτου του περιεχομένου των προτάσεων, καθώς και άλλοι επιπλέον κανόνες που λαμβάνουν υπόψη και άλλα στοιχεία σε μία πρόταση. Για παράδειγμα, ένας από τους κανόνες αφορά τα συμφραζόμενα μίας πρότασης, όπου μπορεί να παρατηρηθεί αν η πρόταση είναι θετική από τις επιμέρους λέξεις που χρησιμοποίησε ένας συγγραφέας, αλλά και το αντίθετο. Δίχως να πραγματοποιηθεί κάποια προεπεξεργασία, τα αποτελέσματα ήταν εξαιρετικά, με ακρίβεια στην εύρεση πολικότητας συναισθήματος να φτάνει στα 80% για τα σύνολα δεδομένων ξενοδοχείων και καταλυμάτων και 70.3% σε ένα μικρότερο σύνολο δεδομένων 40 σχολίων με χαρακτηριστικά ταξινομημένα σε κατηγορίες. Η έρευνα εντόπισε μία σχέση μεταξύ της αντικειμενικότητας και του μήκους των κριτικών. Επίσης, βρέθηκε ότι τα αποτελέσματα ήταν χειρότερα όσον αφορά τις αρνητικές κριτικές, σε σύγκριση με τις θετικές κριτικές.

Παρομοίως, αποτελεί η χρήση BERT μοντέλου για την ανάλυση ισπανικών κριτικών που βρίσκονται στον ιστότοπο TripAdvisor (Vásquez, J. et al., 2021). Σε αυτή την μελέτη το μοντέλο τροποποιήθηκε με συγκεκριμένες ρυθμίσεις που επέστρεψαν την μέγιστη απόδοσή. Στη συνέχεια, εφαρμόστηκε το γνωστό TF-IDF και ταξινομήθηκαν τα δεδομένα με το μοντέλο Λογιστικής Παλινδρόμησης. Τα αποτελέσματα εξετάστηκαν με τον τύπο μέσου απολύτου σφάλματος. Η έρευνα αυτή επέστρεψε τα καλύτερα αποτελέσματα από όλες τις ομάδες που συμμετείχαν, αλλά το ποσό μέσου απολύτου σφάλματος παρέμεινε αρκετά υψηλό. Αυτό οφείλεται σε κάποιους περιορισμούς όπως είναι το σετ δεδομένων και την υπολογιστική δύναμη του συστήματος που εκτελέστηκε το μοντέλο.

Άλλη μία έρευνα είναι αυτή που έχει να κάνει με την ανάλυση συναισθημάτων σε big data χρησιμοποιώντας το open-source framework hadoop (Vishnupriya, S. et al., 2015). Το hadoop χρησιμοποιείται για την αποθήκευση και διαχείριση μεγάλου όγκου δεδομένων. Ο σκοπός της εργασίας ήταν να αναλυθούν ταχύτατα big data, δίχως να θυσιάσει η ακρίβεια του μοντέλου. Τα αποτελέσματα τους ήταν ικανοποιητικά με την ακρίβεια να φτάνει το 75% και τονίστηκε ότι το ίδιο μοντέλο μπορεί να εφαρμοστεί και σε μεγαλύτερα σετ δεδομένων.

Μία ενδιαφέρουσα έρευνα που ασχολείται με την ανάλυση συναισθήματος σε δεδομένα από το twitter είναι όμοια με την προηγούμενη έρευνα, καθώς χρησιμοποιεί το hadoop για την αποθήκευση και ανάλυση όγκου δεδομένων (Rodrigues, A. P. et al., 2019). Παρόλα αυτά, εφαρμόζει επίσης και το Map Reduce προγραμματιστικό μοντέλο για την διαχείριση των σφαλμάτων. Εφαρμόζεται το μοντέλο Hybrid Lexicon-Naïve Bayes Classifier (HL-NBC) στα πειράματά τους. Παρατηρήθηκε ότι ο ταξινομητής που εφάρμοσαν, επέστρεψε θετικά αποτελέσματα που ξεπέρασαν τους άλλους ταξινομητές. Ως εξίσου, το προτεινόμενο hadoop framework, σε συνδυασμό με το Map Reduce προγραμματιστικό μοντέλο που προτείνουν αποτελεί μία αποδοτική λύση.

No.	Βιβλιογραφία	Πηγή δεδομένων	Τύπος αντ.	Πλήθος	Valence	Feature Extraction	Data Relationship Analysis	Lexicon based method	Ανάλυση μελέτης
1	Valdivia, A. et al. (2017)	Tripadvisor	Μνημεία	Nan		✓		✓	Καταμέτρηση ικανοποίησης πελατών και σύγκριση SAMs
2	Gao, S. et al. (2015)	Tripadvisor	Ξενοδοχεία	2100				✓	Καταμέτρηση ικανοποίησης πελατών και έλεγχος αποτελεσμάτων
3	Garcia, A. et al. (2012)	Tripadvisor	Ξενοδοχεία και Εστιατόρια	1994	✓			✓	Δημιουργία προσωποποιημένου λεξικού και έλεγχος αποτελέσματος
4	Vásquez, J. et al. (2021)	Tripadvisor	Αξιοθέατα	5197		✓			Χρήση μεθόδους BERT και κατηγοριοποίηση της άποψης σχετικά με έναν τουρίστα
5	Leelawat, N. et al. (2022)	Twitter	Πόλεις	150850		✓			Ανάλυση αγγλικών tweets που αφορούν τις κύριες πόλεις της Ταϊλάνδης
6	Vishnupriya, S. et al. (2015)	Twitter	Λάπτοπ	50					Έυρεση τεχνικής για την αποδοτική εφαρμογή ανάλυσης συναισθημάτων.
7	Rodrigues, A. P. et al. (2019)	Twitter	Nan	5000 - 1000000		✓		✓	Χρήση HL-NBC και συνδυασμό hadoop με Map Produce για την ανάλυση συναισθήματος.

Πίνακας 1: Σύγκριση βιβλιογραφίας

2.1 Ιστορία

Ρωμαϊκή Αυτοκρατορία

Την περίοδο της Ρωμαϊκής Αυτοκρατορίας (από το 27 π.Χ μέχρι το 476 μ.Χ), εφαρμόστηκαν ταξίδια που αφορούσαν στρατιωτικό σκοπό, πολιτικό σκοπό ή το εμπόριο, καθώς και την μεταφορά των μινυμάτων από την κεντρική διοίκηση στις μακρινές υποταγμένες περιοχές τους (Cooking, T., 2021). Επίσης, αυτό το μέσο ήταν αναγκαίο για τους τεχνίτες και αρχιτέκτονες της εποχής που καλούνταν να σχεδιάσουν και να κατασκευάσουν τα μεγαλοπρεπή παλάτια και τις κατακόμβες. Στην Αρχαία Ελλάδα, οι άνθρωποι ταξίδευαν για να παρακολουθήσουν τους Ολυμπιακούς Αγώνες. Φυσικά, και οι συμμετέχοντες αλλά και οι θεατές των παιχνιδιών απαιτούσαν κατάλυμα για την διαμονή τους αλλά και προσφορά υπηρεσιών για την διατροφή τους. Ακόμα, είναι σημαντικό να αναφερθεί ότι εκείνη την περίοδο, οι πλούσιοι Ρωμαίοι πολίτες, ταξίδευαν στα παραθαλάσσια εξοχικά τους σε περιοχές όπως η Ελλάδα ή η Αίγυπτος για τουριστικούς σκοπούς.

Μεσαίωνας

Την περίοδο του Μεσαίωνα (από το 500 μ.Χ μέχρι το 1400 μ.Χ), εφαρμόστηκαν πολλές μεταφορές για θρησκευτικούς σκοπούς (Newton A., 2013). Ήταν συχνό φαινόμενο η διέλευση πιστών στις ιερές πόλεις τους, όπως οι μουσουλμάνοι στην Μέκκα και οι χριστιανοί στην Ιερουσαλήμ και Ρώμη. Επιπροσθέτως, τον 16ο αιώνα υπήρχε μία αύξηση στο εμπόριο, κάτι που οδήγησε στην εμφάνιση νέου τύπου τουριστών με σκοπό την επέκταση της γνώσης και εμπειρίας τους.

17ο αιώνα

Από τον 17ο μέχρι και τον 19ο αιώνα, οι απόγονοι των αριστοκρατών της Βρετανίας ταξίδευαν σε περιοχές της Ευρώπης, όπως η Ιταλία, Γερμανία και Γαλλία, για αρκετά μεγάλα χρονικά διαστήματα (δύο ή τριών ετών), με σκοπό την αύξηση των γνώσεών τους (Europeana, 2020). Αυτό το φαινόμενο ονομάστηκε "Grand Tour", και ήταν αναγκαίο στην δημιουργία μελλοντικών πολιτικών ηγετών και λοιπών διαχειριστών (The National Archives, 2021).

Βιομηχανική Επανάσταση

Την εποχή της Βιομηχανικής Επανάστασης (από το 1750 μέχρι και το 1850) στην Ευρώπη κατασκευάστηκαν τα θεμέλια για τον μαζικό τουρισμό (Popp A., 2009). Την περίοδο εκείνη, οι άνθρωποι άρχισαν να απομακρύνονται από την γεωργία και να επεκτείνονται σε άλλους τομείς, που θυμίζουν πιο πολύ την σημερινή εποχή, όπως την εργασία σε πόλεις ή εργοστάσια και γενικότερα τον αστικό τρόπο ζωής. Με την εγκαθίδρυση αυτού του συστήματος παρατηρήθηκαν ραγδαίες αυξήσεις στον πλούτο και το επίπεδο εκπαίδευσης των πολιτών της μεσαιας τάξης, καθώς και αύξηση του ελεύθερου χρόνου τους. Το φαινόμενο αυτό επιπλέον προκάλεσε την εισαγωγή νέων κλάδων στον χώρο του τουρισμού, κάτι που προέκυψε από την υψηλή ζήτηση για τουριστικές δραστηριότητες στις διακοπές. Επιπλέον, είναι σημαντικό να αναφερθεί ότι εκείνο το διάστημα το ταξίδι για τα οφέλη υγείας έγινε απαραίτητο, κάτι που είχε ως αποτέλεσμα την επίσκεψη σε λουτροπόλεις από πολλούς ευκατάστατους Ευρωπαίους πολίτες, καθώς και σε παραθαλάσσια εξοχικά στην Αγγλία.

19ο αιώνα με 20ο αιώνα

Τον 19ο με 20ο αιώνα υπήρχαν δραστικές αλλαγές στον κοινωνικό και τεχνολογικό τομέα, κάτι που επέφερε τεράστιες αλλαγές στον τομέα του τουρισμού (IEG - EGO, 2010). Αυτές οι ραγδαίες αλλαγές που πραγματοποιήθηκαν προκάλεσαν την δημιουργία μέσω μεταφοράς που ήταν ταχύτατα, ασφαλή αλλά και συγκριτικά φθηνότερα, κάτι που επέτρεψε την συχνότερη μετακίνηση των πολιτών χαμηλότερων κλάσεων, σε σχέση με την προηγούμενη περίοδο που ήταν πιο κοινό φαινόμενο η μετακίνηση μόνο των εύπορων πολιτών. Ακόμα, τον 19ο αιώνα, δημιουργήθηκε ο σιδηρόδρομος, ενώ τον 20ο αιώνα κατασκευάστηκε και το αεροσκάφος για την μεταφορά επιβατών. Ο δεύτερος Παγκόσμιος Πόλεμος (1939 με 1945) προκάλεσε και μεγάλες αλλαγές στα συστήματα επικοινωνίας και στην μετακίνηση δια του αέρα, που διευκόλυνε την μεταφορά συγκριτικά με τα παλαιότερα χρόνια.

Δεκαετία του 1980

Την δεκαετία του 1980 υπήρχε μεγάλη άνθηση στον τουρισμό, με ταξίδια να πραγματοποιούνται για λόγους εργασίας ή για αναψυχή. Οι ταξιδιώτες έψαχναν για διάφορες υπηρεσίες που συμπεριλαμβάνουν ταξίδια για περιπέτεια, ταξίδια για οικοτουρισμό και πολυτελές ταξίδια για αναψυχή (IEG - EGO, 2010). Έτσι, υπήρχε ισχυρή επέκταση στην εργασία που αφορά τα ταξίδια, αλλά και αύξηση στους τουριστικούς προορισμούς. Ακόμα, η πτώση του Τείχους του Βερολίνου το 1989 οδήγησε στην εξαύλωση του κομμουνισμού στην Ευρώπη και μετέτρεψε χώρες όπως την Ρωσία και την Τσεχία σε δημοφιλείς τουριστικούς προορισμούς για αναψυχή αλλά και εργασία.

Από την δεκαετία του 1990 μέχρι σήμερα

Από την δεκαετία του 1990 μέχρι σήμερα, τα συστήματα κρατήσεων υιοθέτησαν πιο ανεπτυγμένες τεχνολογικές πρακτικές. Με αυτό τον τρόπο, έγινε δυνατή η επιλογή από μία μεγάλη γκάμα ταξιδιωτικών υπηρεσιών, όπως είναι η κράτηση ξενοδοχείου, η ενοικίαση αυτοκινήτων, το κλείσιμο εισιτηρίου για κρουαζιέρες και τρένα, καθώς και η αγορά εισιτηρίων για το θέατρο δια μέσω του αναβαθμισμένου συστήματος κρατήσεων. Η εισαγωγή των ηλεκτρονικών εισιτηρίων επέφερε κάποια προνόμια στα αεροδρόμια, καθώς μειώθηκε η γραφειοκρατία και το κόστος των εισιτηρίων. Ταυτόχρονα, οι επιβάτες δεν χρειάστηκαν πλέον να ανησυχούν για τυχόν απώλειες των εισιτηρίων τους ή για την μεταφορά του δελτίου των εισιτηρίων σε φυσική μορφή. Επίσης, τα ηλεκτρονικά εισιτήρια διευκόλυναν την επικοινωνία και διαχείριση των πελατών από τις αεροδρομικές εταιρείες. Οι ταξιδιώτες μπορούσαν πλέον να συνδεθούν στο ίντερνετ και να ερευνήσουν ταχύτατα πληροφορίες για τον ταξιδιωτικό τους προορισμό, καθώς και να κάνουν κράτηση εισιτηρίου και καταλύματος διαμονής εύκολα από τον προσωπικό τους υπολογιστή που διατίθεται στην οικία τους.

3 Ερευνητικός χώρος

Σε αυτή την ενότητα θα αναλυθεί περαιτέρω το θεωρητικό υπόβαθρο της εργασίας. Αυτό μπορεί να εμπεριέχει κάποιες μεθόδους, παραμέτρους και γενικότερα την θεωρία πίσω από ότι εφαρμόστηκε για τους στόχους της εργασίας.

3.1 Τουρισμός

Ο τουρισμός, σε γενικούς όρους, είναι το ταξίδι που υφίσταται για τον σκοπό της ψυχαγωγίας ή για την εργασία ενός ατόμου (Walton, J. K., 2023). Πιο συγκεκριμένα, ο όρος «τουρισμός» αναφέρεται στο κοινωνικό, πολιτιστικό και οικονομικό φαινόμενο που απαιτεί την μετακίνηση των ατόμων σε περιοχές και μέρη που βρίσκονται μακριά από το σύνθηες περιβάλλον τους, για προσωπικούς ή επαγγελματικούς σκοπούς. Αυτοί οι άνθρωποι αποκαλούνται επισκέπτες ή τουρίστες και ο τουρισμός ασχολείται με τις δραστηριότητες που εμπλέκονται σε αυτή την διαδικασία, κάποιες από τις οποίες περιέχουν τα έξοδα που πραγματοποιούν (United World Tourism Organization, 2008). Εν κατακλείδι, η ορολογία αυτή δεν είναι απλά η μεταφορά των ανθρώπων για πολλαπλούς σκοπούς, όπως η εργασία ή η ψυχαγωγία, αλλά το σύνολο όλων των δραστηριοτήτων, υπηρεσιών και εμπλεκόμενων τομών που αποτελούν αυτήν την μοναδική τουριστική εμπειρία.

Είναι σύνθηες να προκαλείτε σύγχυση μεταξύ του νοήματος κάποιων ορολογιών που αναφέρονται σε αυτόν τον τομέα, δηλαδή ανάμεσα στις λέξεις «τουρισμός», «ταξίδι» και «φιλοξενία». Ωστόσο, αυτές οι τρεις ορολογίες εννοιολογικά διαφέρουν αισθητά η μία από την άλλη. Αρχικά, όπως εξηγήθηκε προηγουμένως, ο όρος «τουρισμός» καλύπτει τις δραστηριότητες, τους τομείς και την βιομηχανία που εμπλέκονται σε μια τουριστική εμπειρία. Σχετικά με τον όρο που καλύπτει το σύνολο των διαδικασιών, ο όρος «ταξίδι» είναι πιο συγκεκριμένα, η δραστηριότητα της μετακίνησης από τοποθεσία σε τοποθεσία για οποιοδήποτε σκοπό, και συχνότερα για τον σκοπό της ψυχαγωγίας και της αναψυχής. Σε αντίθεση με τους προηγούμενους όρους, η «φιλοξενία» μπορεί να οριστεί ως την βιομηχανία της βοήθειας ατόμων, έτσι ώστε να νιώσουν πιο ευπρόσδεκτοι και να απολαύσουν τον χρόνο τους κατά την διάρκεια του ταξιδιού αναψυχής τους ή εργασίας τους.

Γενικότερα, ο τουρισμός είναι μία από τις ταχύτερα αναπτυσσόμενες βιομηχανίες στον κόσμο και αποτελεί μία σημαντική μορφή δημιουργίας συναλλάγματος, που προκύπτει από τις πολλαπλές συναλλαγές μέσω διαφορετικών νομισμάτων από όλο τον κόσμο, και απασχόλησης για πολλές χώρες. Η λέξη "tour" στο "tourism" προέρχεται από την λατινικής προελεύσεως λέξη "tornus", που εννοιολογικά αναφέρεται σε ένα εργαλείο για την κατασκευή κύκλου.

3.1.1 Τουρισμός στην Ελλάδα

Οι περιοχές της Ελλάδας και ειδικότερα τα νησιά της χώρας είναι από τους δημοφιλέστερους προορισμούς αναψυχής στην Ευρώπη (Krinis, N., 2022). Μάλιστα, το 2022 η Ελλάδα κέρδισε την έκτη θέση στην λίστα με τις καλύτερες χώρες για να ταξιδέψει κανείς με 89.97 πόντους, λίγο μπροστά από χώρες όπως το Ηνωμένο Βασίλειο και την Δανία. Οι θέσεις αυτές προέκυψαν από τα Conde Nast Traveller Reader's Choice Awards του 2022, με την Πορτογαλία να κερδίζει την πρώτη θέση και την Ιαπωνία να βρίσκεται κοντά στην δεύτερη θέση. Σύμφωνα με το άρθρο της European Travel Commission του 2023 η Ελλάδα είναι η 5η πιο επιθυμητή χώρα στην Ευρωπαϊκή Ένωση όσον αφορά τον προορισμό ταξιδιών των πολιτών, με μοναδικές χώρες της Ευρώπης που είναι πιο αρεστές να είναι η Γαλλία, η Ιταλία, η Ισπανία και η Γερμανία. Σε επιμέρους αγορές, η Ελλάδα βρίσκεται στην 5η θέση στις προτιμήσεις των Γερμανών, στην 4η θέση στις προτιμήσεις των Γάλλων και στην 3η θέση στις προτιμήσεις των Πολωνών.

3.2 Ανάλυση συναισθημάτων

Ανάλυση συναισθημάτων (Sentiment Analysis) αποτελεί η διαδικασία ανάλυσης ψηφιακού περιεχομένου με σκοπό την ταξινόμησή του βάσει του συναισθήματος που εμφανίζεται στο περιεχόμενο αυτό (Alaei, A. et al., 2019). Το συναίσθημα, ή πιο συγκεκριμένα το συναισθηματικό ύφος, μπορεί να είναι θετικό, αρνητικό ή ουδέτερο. Ο τομέας της Ανάλυσης Συναισθημάτων γενικότερα προσπαθεί να εφαρμόσει υπολογιστικούς αλγορίθμους για την αποκωδικοποίηση και ποσοτικοποίηση των συναισθημάτων που χρησιμοποιήθηκε σε διαφόρων τύπων πολυμέσων, όπως είναι τα κείμενα, οι εικόνες και τα βίντεο. Αυτή η τεχνική είναι μία από τις πιο συνηθισμένες τεχνικές ταξινόμησης κειμένου. Στον τομέα του τουρισμού γίνεται χρήση της για να ερευνηθεί το ποσοστό ικανοποίησης των πελατών, όσον αφορά την εμπειρία του ταξιδιού καθώς και την εμπειρία φιλοξενίας που βίωσαν οι επισκέπτες. Ουσιαστικά, η τεχνική αυτή αποτελεί έναν τρόπο ερμάνευσης κειμένου μέσω μηχανικής μάθησης για την εκφορά νοήματος από το κείμενο αυτό. Στον τομέα αυτό υπάρχουν δύο μέθοδοι για την ποσοτικοποίηση των κειμένων, η μέθοδος που αφορά την χρήση λεξικού και η μέθοδος που αφορά την χρήση αλγορίθμων μηχανικής μάθησης.

Η χρήση της τεχνικής αυτής είναι σημαντική και εφαρμόζεται σε διάφορους τομείς όπως σε σελίδες που περιέχουν έναν εξωφρενικά μεγάλο όγκο τυπωμένων στοιχείων, με ένα παράδειγμα να είναι οι κριτικές. Προφανώς, δεν έχει την δυνατότητα ο ιδιοκτήτης κάποιας σελίδας που διαχειρίζεται την σελίδα να αναλύει χειροκίνητα κάθε κριτική που έχει στον ιστότοπο του, ειδικά αν οι κριτικές είναι σε κλίμακα των χιλίων ή των εκατομμυρίων. Για αυτό τον λόγο, γίνεται χρήση αυτής της τεχνικής, διότι είναι σε θέση να επεξεργαστεί αδόμητα κείμενα γρήγορα και με τρόπο έτσι ώστε να εξοικονομεί πόρους, όπως είναι τα χρήματα. Παρομοίως, αν μία επιχείρηση επιθυμεί να αναλύσει τα δεδομένα χρησιμοποιώντας έναν μεγάλο αριθμό ατόμων, η διαδικασία θα ήταν χρονοβόρα με αποτέλεσμα η επιχείρηση να έχει πιθανές απώλειες, εφόσον πρέπει μία επιχείρηση να γνωρίζει τι συμβαίνει σε πραγματικό χρόνο, όσον αφορά την λειτουργία της επιχείρησής της. Για παράδειγμα, πρέπει μια επιχείρηση να έχει την δυνατότητα να γνωρίζει τι γνώμη έχουν οι χρήστες της υπηρεσίας που διαθέτουν, και για τα εμπλεκόμενα τμήματα που αφορούν την υπηρεσία αυτή, και να είναι σε θέση να στοχεύσει σωστά το υπεύθυνο τμήμα, σε θέματα που προέκυψαν ή ενδέχεται να προκύψουν, επιλύοντας το πρόβλημα άμεσα και αποφεύγοντας πιθανές απώλειες. Ακόμα, με την τεχνική αυτή υπάρχει συνέπεια, διότι όσο το σύστημα συνεχίζει να εκπαιδεύεται με δεδομένα, τόσο η ανάλυση εφαρμόζεται στα στοιχεία με μεγαλύτερη επιτυχία. Σε αντίθεση, όταν ένα άτομο αναλύει κάποιο κείμενο πιθανότατα θα υπάρχει κάποια προκατάληψη σχετικά με τις πεποιθήσεις, ιδέες, αλλά και προσωπικές εμπειρίες του, που ενδέχεται να επηρεάσει σημαντικά την τελική απόφασή του σχετικά με το ύφος του κειμένου.

Στην ταξινόμηση με την χρήση Sentiment Analysis μπορεί να εμφανιστούν εύκολες περιπτώσεις, αλλά προφανώς και δύσκολες περιπτώσεις (Pandey, P., 2021). Κάποια εύκολα σχόλια είναι οι προτάσεις:

- The circus was fun!
- I had a great time at the cinema!
- This book was the best!
- That movie sequel was awful!
- Ew! The food they served us at the restaurant was disgusting!

Σε αυτές τις προτάσεις είναι εύκολο να αναγνωριστεί το ύφος του κειμένου, διότι αναφέρεται ξεκάθαρα στο τυπωμένο κείμενο. Όμως, υπάρχουν και περιπτώσεις όπου η επεξεργασία κειμένου αποτελεί μεγαλύτερο πρόβλημα. Κάποια πιθανά δύσκολα σχόλια είναι τα:

- I did not mind the loud sound effects at the start of the movie (Φράση που περιέχει άρνηση με θετικό συναίσθημα)
- This movie was LIT! No cap! (ίσως δυσκολευτεί να αναγνωρίσει μοντέρνους διαδικτυακούς όρους)
- Oh, just perfect. I waited like two months for the album to come out, but now I have to wait another three months. (Σαρκασμός)
- There are times like these, that I hate this company. (Αλλάζει το ύφος με την χρήση του "There are times like these")

3.2.1 Επεξεργασία με την χρήση λεξικού

Η προσέγγιση αυτή εφαρμόζει ένα ήδη κατασκευασμένο λεξικό συναισθημάτων για να βαθμολογήσει ένα έγγραφο, αθροίζοντας τις τιμές συναισθημάτων από κάθε λέξη που περιέχεται στο έγγραφο αυτό. Το λεξικό συναισθημάτων πρέπει να περιέχει κάθε λέξη με την αντίστοιχη τιμή συναισθήματος που αφορά την λέξη αυτή. Επιπλέον, οι αρνητικές μορφές των λέξεων πρέπει να εμφανίζονται σε ξεχωριστές θέσεις μέσα στο λεξικό, και να δίνεται υψηλότερη έμφαση σε αυτές, συγκριτικά με τις θετικές μορφές των λέξεων. Σε αυτή την προσέγγιση, εφαρμόζεται ένα σύνολο κανόνων. Τα βήματα που ακολουθούνται συνήθως για την ανάλυση κειμένων είναι:

1. Εξαγωγή των απαραίτητων δεδομένων από την ιστοσελίδα]
2. Tokenize το κείμενο, δηλαδή διαχωρισμός κάθε ατομικής λέξης από το κείμενο.
3. Διαγραφή των άρθρων και άλλων ειδών stopwords από τα κείμενα. Stopwords αποτελούν αυτές οι λέξεις που θεωρούνται ασήμαντες για την ανάλυση κειμένου. Κάποια παραδείγματα είναι οι λέξεις: για, από, τα, ενώ, αυτοί κ.α.
4. Διαγραφή των σημείων στίξης (σε ορισμένες περιπτώσεις που απαιτείται η διαδικασία αυτή)
5. Εκτέλεση του κειμένου στο λεξικό για την εξαγωγή της τιμής (σκορ) που αφορά το συναισθηματικό ύφος που παρατηρήθηκε στο κείμενο

Παράδειγμα χρήσης

Ας υποθέσουμε ότι πρέπει να εξετασθεί μία πρόταση με αυτή την προσέγγιση.

« Ο Δημήτρης είναι καλός άνθρωπος.»

Το σύστημα αρχικά θα διαχωρίσει τις λέξεις από την πρόταση οπότε θα προκύψει:

Λέξη	Κατηγορία
Ο	stopword
Δημήτρης	
είναι	stopword
καλός	
άνθρωπος	
.	σημείο στίξης

Πίνακας 2: Πίνακας με την λέξη και την κατηγορία, βάσει τον διαχωρισμό του συστήματος

Τα stopwords αφαιρούνται γιατί δεν υπολογίζονται στην διαδικασία της ανάλυσης συναισθημάτων. Επιπλέον, στην προκειμένη περίπτωση, αφαιρείται και το σημείο στίξης λόγω του ότι δεν επηρεάζεται το αποτέλεσμα της ανάλυσης συναισθημάτων από το σημείο στίξης. Στην συνέχεια, θα προσπαθήσει το σύστημα να ορίσει κάποια βαθμολογία για τις λέξεις που απομείνανε με σκοπό την εύρεση του ύφους συναισθήματος στην πρόταση. Με αυτό τον τρόπο θα προκύψουν τα παρακάτω αποτελέσματα:

Λέξη	Συναίσθημα
Δημήτρης	ουδέτερο
καλός	θετικό
άνθρωπος	ουδέτερο

Πίνακας 3: Παρουσίαση αποτελεσμάτων με τις λέξεις και τα αντίστοιχα συναισθήματα

Όπως φαίνεται, από κάθε λέξη που παρουσιάζεται στην πρόταση εξάγεται κάποιο συμπέρασμα με την βοήθεια της προσέγγισης αυτής. Όμως, η λέξη άνθρωπος και η λέξη Δημήτρης, δεν επηρεάζουν το τελικό αποτέλεσμα, καθώς δεν έχουν κάποιο συναίσθημα που να μπορεί να ερμηνεύσει με σιγουριά ο υπολογιστής ότι έχει θετικό ή αρνητικό συναισθηματικό ύφος. Έτσι, αυτά ταξινομούνται ως ουδέτερα και η λέξη «καλός» ταξινομείται σωστά ως θετική. Εν κατακλείδι, η πρόταση κατατάσσεται ότι περιέχει γραπτό λόγο με θετικό συναισθηματικό ύφος, μετά από την ανάλυση συναισθημάτων.

Αυτό, αποτελεί ένα απλό παράδειγμα που δεν συναντάται συχνά στον πραγματικό κόσμο, διότι συνήθως μία πρόταση εμπλουτίζεται με ένα πλήθος από λέξεις, με αποτέλεσμα ο λόγος να γίνεται πιο πολύπλοκος και πιο δυσνόητος από τον υπολογιστή. Για αυτό, δημιουργούνται λεξικά συναισθημάτων με τιμές για λέξεις που λαμβάνουν υπόψη και το πόσο έντονο είναι το συναίσθημα, δηλαδή το πόσο θετικό ή αρνητικό είναι το κείμενο. Για παράδειγμα, ένα κείμενο μπορεί να είναι θετικό, αλλά υπολογίζεται και αν είναι καλό, πολύ καλό ή εξαιρετικά καλό το θετικό συναίσθημα της πρότασης. Επιπροσθέτως, θεωρείται σημαντικός και ο υπολογισμός της υποκειμενικότητας ή της αντικειμενικότητας στο κείμενο, καθώς και το νόημα του κειμένου.

Υπάρχουν διάφορα τέτοια λεξικά που είναι ιδιαίτερα δημοφιλείς, με τα πιο γνωστά να είναι το VADER, το TextBlob και το Sentiwordnet. Στην συγκεκριμένη εργασία θα εφαρμοστούν τα Vader και TextBlob.

Vader

Το Vader αποτελεί ένα εργαλείο που εφαρμόζεται με σκοπό την ανάλυση συναισθημάτων σε κείμενα και επιτρέπει την κατάταξη κειμένων ανάλογα με το αν το ύφος κειμένου θεωρείται θετικό ή αρνητικό, αλλά και το πόσο έντονα είναι τα συναισθήματα που παρουσιάζονται στο κείμενο αυτό (Hutto, C. et al., 2014). Είναι ιδιαίτερα χρήσιμο στην ανάλυση κειμένων από κοινωνικά μέσα, διότι είναι ειδικά κατασκευασμένο για την ανάλυση κειμένων από τέτοιες πηγές. Παρέχεται στα πακέτα της βιβλιοθήκης NLTK και προσφέρει μετρήσεις για κάθε λέξη. Οι τιμές που μπορεί να επιστρέφει σε κάθε λέξη είναι:

- Αρνητικές: Μέχρι τιμή ίση με -4 που σημαίνει ότι το ύφος είναι εξαιρετικά αρνητικό
- Θετικές: Μέχρι τιμή ίση με +4 που σημαίνει ότι το ύφος που παρατηρήθηκε είναι εξαιρετικά θετικό
- Ουδέτερες: Οι τιμές σε αυτή την κατάταξη ορίζονται με τιμή ίση με μηδέν και μπορεί να αφορούν λέξεις που είναι όντως ουδέτερες ή τυχόν στοιχεία που απουσιάζουν από το σετ δεδομένων (NaN).

Ωστόσο, οι τιμές που μπορεί να επιστρέφει σε κάθε πρόταση είναι στην κλίμακα των [-1. +1], με την τιμή -1 να είναι η κατηγορία όπου ταξινομούνται τα στοιχεία με αρνητικό συναίσθημα και την τιμή +1 να ανήκουν φυσικά τα στοιχεία με θετικό συναίσθημα. Σε αυτό το σημείο, μπορεί να αναρωτηθεί κανείς το πώς γίνεται η κλίμακα της τιμής κάθε λέξης να είναι από -4 έως 4, ενώ κάθε πρότασης να είναι από -1 έως 1. Αυτό το φαινόμενο προκύπτει επειδή χρησιμοποιείται κανονικοποίηση (normalization) στον υπολογισμό της τιμής πρότασης. Για να συμβεί αυτό εφαρμόζεται ο τύπος κανονικοποίησης:

$$\frac{x}{\sqrt{x^2+a}}$$

,όπου το a , που χρησιμοποιείται για το normalization, στον τύπο του VADER είναι σταθερά ίση με τιμή 15, ενώ το x υπολογίζεται από το άθροισμα της τιμής συναισθήματος κάθε λέξης που αποτελεί την πρόταση.



Διάγραμμα 2: Γράφημα που αφορά τον τύπο κανονικοποίησης

Στο γράφημα παρατηρείται ο τύπος αυτός, και όπως φαίνεται όσο αυξάνεται η τιμή του x , τόσο πλησιάζει περισσότερο τις τιμές -1 και $+1$. Με τον ίδιο τρόπο, όταν παρουσιάζονται πολλές λέξεις μέσα σε ένα έγγραφο και εφαρμόζεται VADER ανάλυση, τότε οι τιμές κάθε πρότασης θα είναι κοντά στις τιμές -1 και $+1$. Αυτό δείχνει ότι το VADER εφαρμόζεται βέλτιστα σε μικρά έγγραφα όπως tweets και προτάσεις, και όχι σε έγγραφα με μεγάλο πλήθος στοιχείων.

Το πακέτο αυτό μπορεί να εκτελεσθεί άμεσα σε σετ δεδομένων που δεν περιέχουν κλάσεις. Επιπλέον, είναι σε θέση να καταλάβει ότι λέξεις όπως «αγαπώ», «απολαμβάνω», «αρέσω» είναι λέξεις με θετικό συναισθηματικό ύφος, καθώς και το νόημα πίσω από αυτές τις λέξεις, σε τέτοιο βαθμό που μπορεί να κρίνει ότι η φράση «δεν μου αρέσει» περιέχει αρνητικό συναισθηματικό ύφος. Ακόμα, λαμβάνει υπόψη και το αν τα γράμματα είναι κεφαλαία ή αν εμπεριέχεται στην πρόταση κάποιο ειδικό σημείο στίξης, όπως το θαυμαστικό (!). Για παράδειγμα, αν παρατηρηθεί η λέξη «ΑΡΕΣΩ» με κεφαλαία γράμματα, θεωρεί ότι το συναίσθημα πίσω από την λέξη είναι ισχυρότερο, σε σύγκριση με την ίδια λέξη που παρουσιάζεται με μικρά γράμματα. Η παρατήρηση αυτή στα κεφαλαία γράμματα, πιθανώς προκύπτει από την χρήση των κειμένων στα social media, όπου όταν εμφανίζεται κείμενο με κεφαλαία γράμματα, τότε θεωρείται ότι ο συγγραφέας φωνάζει.

Συνεπώς, το Vader δεν λαμβάνει υπόψη μόνο τις λέξεις αλλά και επιπρόσθετα χαρακτηριστικά όπως αν παρουσιάζεται κάποιο σημείο στίξης, αν η λέξη είναι γραμμένη με κεφαλαίους χαρακτήρες κ.α. Τα επιπρόσθετα στοιχεία που αναλύονται και βαθμολογούνται κατάλληλα από το Vader μπορούν να ανήκουν σε μία από πέντε κατηγορίες. Αναλυτικά:

- **Σημεία στίξης:** Ενισχύονται οι τιμές συναισθημάτων ανάλογα με το αν παρουσιάζονται ορισμένα σημεία στίξης, όπως το ερωτηματικό (?) και το θαυμαστικό (!), και ανάλογα το πλήθος των σημείων στίξης που παρουσιάζονται στην πρόταση. Το VADER πρώτα αναλύει την πρόταση και εξάγει κάποιες τιμές συναισθήματος. Αν το συναισθηματικό ύφος είναι θετικό, τότε αθροίζεται στην βαθμολογία μία συγκεκριμένη τιμή στην περίπτωση θαυμαστικού (0.292) και ερωτηματικού (0.18). Στην αντίθετη περίπτωση, δηλαδή εκεί όπου το ύφος που ανιχνεύθηκε αποδείχθηκε ότι είναι αρνητικό, τότε αυτές οι τιμές αφαιρούνται από την βαθμολογία.
- **Κεφαλαία γράμματα:** Αν παρουσιάζεται κάποια λέξη που θεωρείται θετική ή αρνητική και παρουσιάζεται με κεφαλαία γράμματα στην πρόταση, τότε αθροίζεται στην τιμή μία τιμή 0.733 στην περίπτωση που το ύφος είναι θετικό, ενώ στην περίπτωση όπου το ύφος είναι αρνητικό το ποσό αυτό αφαιρείται.
- **Βαθμός (Επίρρηση):** Στην περίπτωση αυτή, μπορεί να υπάρχει κάποια λέξη που τροποποιεί τον βαθμό μίας δεύτερης λέξης, δηλαδή το πόσο έντονο είναι το συναίσθημα γύρω από την δεύτερη λέξη αυτή. Ένα παράδειγμα, αποτελεί η φράση «λίγο αργό» ή «πολύ ωραίο», όπου στην πρώτη περίπτωση μειώνεται το πόσο έντονη είναι η τιμή συναισθήματος στην λέξη «αργό», ενώ στην δεύτερη αυξάνεται η τιμή συναισθήματος στην λέξη «ωραίο». Το πόσο πρέπει να προστεθεί ή να μειωθεί στην τιμή της δεύτερης λέξης εξαρτάται από παράγοντες, όπως την απόσταση μεταξύ της πρώτης λέξης και της δεύτερης λέξης στην πρόταση. Στην περίπτωση όπου η πρώτη λέξη βρίσκεται ακριβώς δίπλα από την δεύτερη λέξη, τότε αθροίζεται ή αφαιρείται μία τιμή ίση με 0.293 στο συναισθηματικό σκορ της πρότασης. Αν υπάρχουν και άλλες λέξεις μεταξύ τότε στην απόσταση μίας λέξης αθροίζεται ή αφαιρείται το 95% του 0.293, στην τρίτη λέξη το 90% κ.ο.κ.
- **Εναντιωματικός σύνδεσμος:** Ελέγχεται αν στην πρόταση εμφανίζεται το «αλλά» στην ελληνική γλώσσα ή το αντίστοιχο "but" στην αγγλική γλώσσα. Οι σύνδεσμοι αυτοί χρησι-

μπορούνται για να ενώσουν δύο αντίθετα συναισθήματα, αλλά το κύριο νόημα συνήθως εξάγεται από το δεύτερο κομμάτι της πρότασης. Για παράδειγμα, σε μία πρόταση «Πέρασα ωραία στον κινηματογράφο, αλλά δεν θα ξαναπήγαινα», το πρώτο κομμάτι δηλαδή το «Πέρασα ωραία στον κινηματογράφο» έχει ύφος θετικό αλλά το δεύτερο κομμάτι «δεν θα ξαναπήγαινα» έχει ύφος αρνητικό και αποτελεί αυτό που συμβάλλει στην απόφαση του κυρίαρχου συναισθήματος της πρότασης, που προφανώς είναι το αρνητικό. Στο VADER η τιμή συναισθήματος στην πρόταση αλλάζει με την τιμή συναισθήματος της κάθε λέξης στην πρόταση που είναι πριν το «αλλά» να μειώνεται κατά 50%, ενώ στο μέρος της πρότασης που είναι μετά το «αλλά» να αυξάνεται η τιμή κατά 50%.

- **Tri-grams:** Στην πρόταση πριν το χαρακτηριστικό που είναι συναισθηματικά φορτωμένο, είτε με θετικό είτε με αρνητικό ύφος, εξετάζονται τα τρία προηγούμενα χαρακτηριστικά { λέξεις, για να ανακαλυφθεί αν υπάρχει αναιρεση του συναισθήματος. Αν παρουσιάζεται στις τρεις προηγούμενες λέξεις νόημα που καταργεί το συναίσθημα στην συναισθηματικά φορτωμένη λέξη, τότε πολλαπλασιάζεται η τιμή συναισθήματος της λέξης αυτής επί μία τιμή ίση με -0.74.

TextBlob

Το TextBlob αποτελεί ένα εργαλείο που επίσης εφαρμόζει ένα σετ κανόνων με την χρήση λεξικού, όπως και το VADER, δηλαδή αποτελεί εργαλείο που ακολουθεί την προσέγγιση της Ανάλυσης Συναισθήματος βάσει λεξικού που αναφέρθηκε προηγουμένως (Bonta, V. et al., 2019). Αποτελεί μία απλή βιβλιοθήκη που επιτρέπει πολύπλοκες αναλύσεις και τροποποιήσεις σε σετ κειμένων. Όπως και με το VADER, επιστρέφεται η τιμή συναισθήματος κάθε πρότασης. Ωστόσο, στο TextBlob επιστρέφεται και η αντικειμενικότητα της πρότασης, που διαφέρει με το VADER όπου επιστρέφεται το πόσο έντονο είναι το συναίσθημα. Η τιμή του συναισθήματος κάθε πρότασης ανήκει στην κλίμακα από -1 έως και +1. Προφανώς, το -1 ορίζεται στις προτάσεις που εμφανίζεται αρνητικό συναισθηματικό ύφος, ενώ το +1 στις προτάσεις όπου παρουσιάζεται θετικό συναισθηματικό ύφος.

Η αντικειμενικότητα μετατρέπει σε αριθμό την ποσότητα της προσωπικής άποψης που παρουσιάζεται σε κάθε πρόταση, σε σχέση με τις τεκμηριωμένες πληροφορίες που γράφονται στην πρόταση αυτή. Όταν ο αριθμός της αντικειμενικότητας στην πρόταση είναι υψηλότερος, τότε αυτό σημαίνει ότι η πρόταση επηρεάζεται περισσότερο από την προσωπική άποψη του συγγραφέα, και όχι από πραγματικά στοιχεία. Για τον υπολογισμό της αντικειμενικότητας συμπεριλαμβάνεται μία επιπλέον παράμετρος, που υπολογίζει το αν μία λέξη τροποποιεί το συναίσθημα μίας άλλης λέξης. Αυτό είναι κάτι αντίστοιχο με την τρίτη κατηγορία που αναλύθηκε στο Vader, με την ονομασία «Βαθμός». Στην Ελληνική γλώσσα αλλά και την Αγγλική τα επιρρήματα είναι τα σχήματα λόγου που τροποποιούν τον βαθμό της ακόλουθης λέξης.

Θα αναφερθούν κάποια παραδείγματα για την καλύτερη επεξήγηση επί του θέματος. Ένα παράδειγμα αποτελεί η πρόταση «Δεν μου άρεσε το ταξίδι, γιατί ήταν πολύ βαρετό». Σε αυτή την πρόταση το TextBlob θα απέδιδε τιμή συναισθήματος στην πρόταση ίση με -1, δηλαδή ότι το συναίσθημα που παρατηρήθηκε ήταν εξαιρετικά αρνητικό, ενώ στην αντικειμενικότητα η τιμή θα αντιστοιχούσε σε +1, διότι είναι περισσότερο προσωπική άποψη παρά τεκμηριωμένη πληροφορία. Ελέγχοντας το νόημα της πρότασης με την κατάταξη από το TextBlob, μπορεί να παρατηρηθεί ότι η πρόταση ταξινομήθηκε σωστά. Όμως, υπάρχει και περίπτωση όπου τα στοιχεία δεν κατατάσσονται με επιτυχία. Ένα παράδειγμα όπου αποτυγχάνει η διαδικασία αυτή αποτελεί η πρόταση

«Ο Γιάννης με βοήθησε στην εργασία, αλλά θα προτιμούσα να την κάνω μόνος μου». Σε αυτή την πρόταση επιστρέφεται ότι η αντικειμενικότητα και η τιμή συναισθήματος ισούνται με 0. Το μηδέν επιστρέφεται, διότι το «αλλά» στην πρόταση και το «βοήθησε» συγκρούονται, με το ένα να αναιρεί το συναισθήμα από το άλλο. Παρόλα αυτά, σε απλές περιπτώσεις και πιο συχνά το TextBlob λειτουργεί σωστά. Ένα ακόμα παράδειγμα αποτελεί η πρόταση «Το φαγητό ήταν τέλειο». Σε αυτό η τιμή συναισθήματος θα είναι ίση με την μέγιστη πιθανή τιμή, δηλαδή 1, ενώ η αντικειμενικότητα θα είναι ίση με 0.75 που σωστά θεωρείται ότι αυτό αποτελεί προσωπική άποψη σε μεγάλο βαθμό, καθώς το γούστο στο φαγητό δεν είναι σταθερά το ίδιο στους ανθρώπους.

3.2.2 Επεξεργασία με την χρήση αλγορίθμων μηχανικής μάθησης

Παρόλο που η χρήση έτοιμων λεξικών/σετ κανόνων είναι αρκετά αποδοτική, οι κανόνες αυτοί μπορεί να μην δουλεύουν πολύ καλά για όλα τα δεδομένα που τίθεται να αναλύσει μια εφαρμογή Ανάλυσης Συναισθημάτων. Προφανώς, δεν γίνεται να έχει χρησιμοποιηθεί κάθε κείμενο για να δημιουργηθεί το λεξικό και επιπλέον η γλώσσα, ειδικά αυτή που εφαρμόζεται σε τυπωμένα κείμενα στο διαδίκτυο, αλλάζει συνεχώς. Έτσι, το άτομο που επιθυμεί να αναλύσει τα συναισθήματα σε κάποιο κείμενο, πρέπει να εντοπίσει μία εναλλακτική μέθοδο που να είναι αποδοτική. Φυσικά, μπορεί να κατασκευάσει το δικό του λεξικό με διάφορες λέξεις για την ανάλυση του συναισθήματος. Όμως, αυτή η διαδικασία σίγουρα θα είναι δύσκολη και χρονοβόρα. Για αυτό τον λόγο, υπάρχει η αυτοματοποιημένη μέθοδος με αλγορίθμους μηχανικής μάθησης που εφαρμόζει αυτή την διαδικασία συγκεκριμένα στα δεδομένα του χρήστη, συσχετίζοντας τις λέξεις που παρέχονται με τις αντίστοιχες ετικέτες συναισθημάτων αλγοριθμικά (Ahmed, H. et al., 2021).

Linear Regression

Η Γραμμική Παλινδρόμηση, η αλλιώς Linear Regression, είναι ένας στατιστικός αλγόριθμος, όπου προβλέπεται μία τιμή Y , από τα παρεχόμενα X χαρακτηριστικά. Ουσιαστικά, χρησιμοποιείται για να προβλέψει μία τιμή μεταβλητής βασισμένη στην τιμή μιας άλλης μεταβλητής (Vijay K., 2023). Η μεταβλητή που τίθεται να προβλέψει ονομάζεται εξαρτημένη μεταβλητή, ενώ η μεταβλητή που χρησιμοποιείται για την πρόβλεψη ονομάζεται ανεξάρτητη μεταβλητή. Με την βοήθεια της μηχανικής μάθησης εξετάζονται τα δεδομένα του συνόλου δεδομένων για να εντοπιστούν σχέσεις ανάμεσα στα στοιχεία. Στη συνέχεια, οι σχέσεις αυτές τοποθετούνται κατά μήκος του άξονα X/Y , με μία ευθεία γραμμή να τρέχει μέσα από αυτές για να προβλέψει περαιτέρω σχέσεις. Ο Linear Regression, υπολογίζει το πως η είσοδος X (λέξεις και φράσεις) συσχετίζεται στην Y έξοδος (πολικότητα συναισθήματος). Με αυτόν τον τρόπο, θα μπορεί να παρατηρηθεί το που οι φράσεις και οι λέξεις, τοποθετούνται πάνω στην κλίμακα της πολικότητας από πολύ θετικό συναισθήμα σε πολύ αρνητικό και ανάμεσα από αυτά.

Δεδομένου ενός συνόλου δεδομένων $\{y_1, x_{i1}, \dots, x_{ip}\}_{i=1}^n$ από n στατιστικές μονάδες, το μοντέλο της γραμμικής παλινδρόμησης θεωρεί ότι η σχέση μεταξύ της εξαρτημένης μεταβλητής y και του διανύσματος με τα χαρακτηριστικά X είναι γραμμική. Η σχέση αυτή διαμορφώνεται μέσα από μία μεταβλητή λάθους (error) ε_i - μια απαραίτητη τυχαία μεταβλητή που προσθέτει θόρυβο (noise) στην γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής του διανύσματος. Οπότε, το μοντέλο θα καταλήξει στην μορφή:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \dots, n$$

όπου το T δείχνει ότι γίνεται μεταφορά στον πίνακα, έτσι ώστε το $x_i^T \beta$ να είναι το εσωτερικό

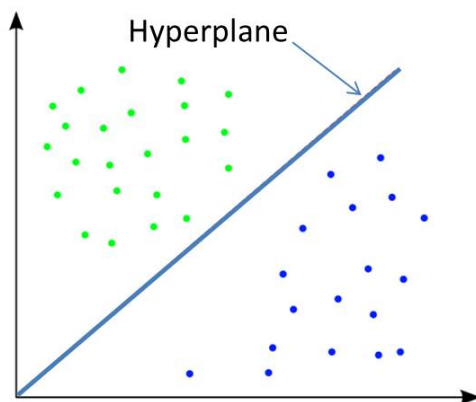
γινόμενο των διανυσμάτων x_i και β . Συχνά, σε μελέτες η εξίσωση της γραμμικής παλινδρόμησης είναι στην μορφή:

$$y = x\beta + \varepsilon$$

που προκύπτει από την πιο πάνω εξίσωση.

LinearSVC

Το LinearSVC είναι παρόμοιο με την Γραμμική Παλινδρόμηση (Linear Regression), με την έννοια ότι και τα δύο προσπαθούν να ανακαλύψουν υπερεπίπεδα που βέλτιστα θα διαχωρίζουν τα στοιχεία ανάμεσα στις κλάσεις (Scikit-learn, n.d.). Στην python γλώσσα προγραμματισμού είναι ενσωματωμένο στα πακέτα SVM, που αφορούν συστήματα υποστήριξης αποφάσεων. Τα υπερεπίπεδα αυτά εμφανίζονται στην παρακάτω εικόνα για την καλύτερη επεξήγηση του θέματος.



Διάγραμμα 3: Παράδειγμα υπερεπιπέδου

Είναι σημαντικό να αναφερθεί ότι σε έναν δισδιάστατο χώρο ένα υπερεπίπεδο είναι μία ευθεία γραμμή που διαχωρίζει το διάστημα σε δύο μισά, ενώ σε έναν τρισδιάστατο χώρο αποτελεί ένα επίπεδο που διαχωρίζει το διάστημα σε δύο μισά. Όπως μπορεί να παρατηρηθεί στην εικόνα, το υπερεπίπεδο είναι η ευθεία γραμμή που βρέθηκε να διαχωρίζει βέλτιστα τα πράσινα στοιχεία από τα μπλέ στοιχεία, εφόσον αφορά έναν δισδιάστατο χώρο με μόνους άξονες τον X και Y.

Naive Bayes

Ο Naive Bayes είναι ένας αλγόριθμος πιθανοτήτων που εφαρμόζεται σε πολλές περιπτώσεις. Είναι απλός και ευνόητος, αλλά παρόλα αυτά είναι αρκετά αποτελεσματικός. Για αυτό τον λόγο χρησιμοποιείται και από υπηρεσίες e-mail για την εύρεση spam μηνυμάτων. Επιπλέον, είναι ιδιαίτερα αποτελεσματικός και σε μεγάλα σύνολα δεδομένων, επειδή είναι γρήγορος και απλός. Αποτελεί επιβλεπόμενο ταξινομητή, με την έννοια ότι ταξινομεί τα αντικείμενα που περιέχουν ήδη κλάσεις (labels). Αυτές οι κλάσεις είναι οι κατηγορίες οι οποίες ανήκει κάποιο στοιχείο. Για παράδειγμα, ένα Opel Corsa στοιχείο μπορεί να ανήκει σε μια κατηγορία "Αυτοκίνητο" η οποία περιέχει ένα ευρύ φάσμα αντικειμένων, σε ένα σύνολο δεδομένων, αλλά μπορεί σε ένα άλλο σύνολο δεδομένων να ανήκει σε κατηγορία "B-segment" που περιγράφει όλα τα αυτοκίνητα που ανήκουν συγκεκριμένα στην κατηγορία αυτών των μικρών αυτοκινήτων.

Το μοντέλο αυτό βασίζεται στο θεώρημα του Bayes που ασχολείται με τις δεσμευμένες πιθανότητες (Gandhi, R., 2018). Δεσμευμένη πιθανότητα αποτελεί η πιθανότητα να υφίσταται ένα γεγονός, δεδομένου ότι ένα άλλο γεγονός συνέβη προηγουμένως. Ο τύπος είναι:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Στον τύπο αυτό, υπολογίζεται η πιθανότητα να συμβεί το ενδεχόμενο A, δεδομένου ότι έχει συμβεί ήδη το ενδεχόμενο B. Άρα, το B είναι η απόδειξη που έχουμε και το A είναι η υπόθεση που κάνουμε. Η υπόθεση στο θεώρημα Bayes είναι ότι τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο, που σημαίνει ότι η ύπαρξη ενός συγκεκριμένου χαρακτηριστικού δεν επηρεάζει τα υπόλοιπα. Για αυτόν τον λόγο ονομάζεται Naive, δηλαδή αφελής.

3.3 Web scraping

Η τεχνική web scraping υλοποιείται σε μελέτες όπου είναι αναγκαία η εξαγωγή δεδομένων από τις ιστοσελίδες στον χώρο του διαδικτύου (Hasson, E., 2019). Πιο συγκεκριμένα, πραγματοποιείται αρχικά η μεταφόρτωση (fetching) του κώδικα και στην συνέχεια υφίσταται επεξεργασία (processing) σε αυτόν, με σκοπό την απόκτηση στοιχείων που θα συμπεριληφθούν σε ένα αρχείο δεδομένων. Έτσι, τα δεδομένα θα είναι διαθέσιμα σε μια τοπική βάση δεδομένων ή κάποιον αποθηκευτικό χώρο, ώστε να μπορεί ανά πάσα στιγμή να πραγματοποιηθεί επεξεργασία των στοιχείων και έπειτα ερμίνευση των αποτελεσμάτων. Η εξαγωγή των δεδομένων συνήθως είναι χρονοβόρα και για αυτό πρώτα γίνεται η εξαγωγή στοιχείων και μετά η ανάλυση που είναι αρκετά ταχύτερη σε σχέση με το πρώτο στάδιο της εξαγωγής. Η εξαγωγή των δεδομένων αυτών είναι χρονοβόρα εξαιτίας κάποιων βασικών παραγόντων, όπως είναι ο όγκος των δεδομένων, η ταχύτητα της σύνδεσης, τον χρόνο ανταπόκρισης από τις σελίδες, το αν επιτρέπει η σελίδα έναν μεγάλο αριθμό συνεχών αιτημάτων, την ταχύτητα των εργαλείων που εφαρμόζουν τεχνική scraping κ.ο.κ.

Δύο βιβλιοθήκες που εφαρμόζονται συχνότερα για την εκτέλεση της τεχνικής αυτής με σκοπό την εξαγωγή χαρακτηριστικών και κειμένων από ιστοχώρους είναι η βιβλιοθήκη selenium και requests της γλώσσας Python. Η γλώσσα Python συχνά εφαρμόζεται γιατί είναι πιο ευνόητη, μιας και είναι πιο κοντά στην ανθρώπινη γλώσσα. Επιπλέον, δεν απαιτεί ιδιαίτερα προσεκτική μεταχείριση σε σχέση με άλλες γλώσσες προγραμματισμού όπως την C++ ή την Java που έχουν μια αυστηρότερη δόμηση με κανόνες που πρέπει να ακολουθούνται ρητά για την σωστή εκτέλεση του προγράμματος. Για παράδειγμα, στην γλώσσα C++ συμπεριλαμβάνεται η `int main()` συνάρτηση και η βιβλιοθήκη `iostream` για την εκτέλεση του προγράμματος. Ακόμα, όλες οι γραμμές κώδικα εφαρμόζονται γύρω από άγκιστρα που εμφανίζονται στην αρχή και στο τέλος των συναρτήσεων αυτών. Αντίθετα από αυτό το σύνθετο φαινόμενο στις γλώσσες προγραμματισμού, η Python υιοθετεί μια πιο απλή τακτική με μόνο κριτήριο για την σωστή λειτουργία να αποτελούν τα κενά σε κάθε γραμμή, καθώς δεν απαιτείται κάποια βιβλιοθήκη για την εκτέλεση εφαρμογής και επιπλέον οι συναρτήσεις δεν ανοίγουν και κλείνουν με χαρακτήρες όπως τα άγκιστρα ().

Για την διευκόλυνση κατά την οργάνωση των αρχείων και εντολών, δημιουργήθηκαν δύο προγράμματα μέσω της γλώσσας προγραμματισμού Python, για κάθε αντικείμενο ιστοσελίδας. Ένα απλό πρόγραμμα για την άντληση των τίτλων από τα ξενοδοχεία και τα αξιοθέατα και ένα δεύτερο πρόγραμμα για την επίσκεψη και εξαγωγή των σημαντικών δεδομένων που απαιτούνται για τον σκοπό της εργασίας. Αυτά τα δύο προγράμματα υφίστανται για κάθε θέμα ιστοσελίδας, όπως είναι τα αξιοθέατα και τα ξενοδοχεία. Για παράδειγμα, αν πρέπει να εφαρμοστεί εξαγωγή

στοιχείων για τα αξιοθέατα και τα ξενοδοχεία από μία σελίδα, τότε ο συνολικός αριθμός προγραμμάτων θα είναι ίσος με τέσσερα, διότι υλοποιούνται δύο εφαρμογές για την συλλογή των ονομάτων και συνδέσμων από τα ξενοδοχεία και αξιοθέατα, καθώς και δύο λογισμικά για την αποθήκευση στοιχείων από κάθε ένα από αυτά τα αξιοθέατα και ξενοδοχεία. Για την υλοποίηση της εφαρμογής αυτής απαιτήθηκε η εφαρμογή τεχνικής web scraping, αφού τα στοιχεία είναι διαθέσιμα στο διαδίκτυο.

Selenium Webdriver

Το Selenium Webdriver είναι ένα εργαλείο που επιτρέπει την αυτοματοποιημένη δοκιμή διάφορων εφαρμογών διαδικτύου και πολλαπλών φυλλομετρητών, και αποτελεί τον πυρήνα του Selenium Framework (Unadkat, J., 2023). Είναι ανοικτού κώδικα, δηλαδή έχει αποθετήριο στο github με όλο τον κώδικα που χρησιμοποιήθηκε για την κατασκευή του εργαλείου και υποστηρίζει ένα πλήθος από γλώσσες προγραμματισμού και φυλλομετρητές. Κάποιες από τις γλώσσες προγραμματισμού που υποστηρίζει είναι η Python, η Java, η Javascript, η C#, η PHP και η Perl. Ακόμα, υποστηρίζει πολλές εφαρμογές για την περιήγηση στο διαδίκτυο με το εργαλείο να είναι σε δυνατότητα να υποστηρίζει τα Google Chrome, Mozilla Firefox, Opera, Safari, Internet Explorer και HtmlUnit. Παρεμπιπτόντως, είναι σημαντικό να αναφερθεί ότι η εφαρμογή μπορεί να εκτελεστεί και σε όλα τα δημοφιλή λειτουργικά συστήματα, δηλαδή στα Microsoft Windows, Linux, MacOS και Solaris.

Παλιότερα, το εργαλείο απαιτούσε την χειροκίνητη κατάσταση ενός driver (π.χ. chromedriver) για την επικοινωνία με τον φυλλομετρητή, αλλά πλέον αυτή η διαδικασία αυτοματοποιήθηκε με την εισαγωγή κάποιων επιπλέον πακέτων κατά την κατασκευή κάποιας εφαρμογής Python. Το εργαλείο χρησιμοποιεί Javascript, οπότε μπορεί να αλληλοεπιδράσει με το περιεχόμενο μίας σελίδας, όπως ένα κουμπί, σε αντίθεση με το requests. Ακόμα, προσφέρει ένα εύρος από οδηγίες που μπορεί να ενεργοποιήσει ο χρήστης όταν στέλνει κάποιο αίτημα. Αυτές μπορεί να αφορούν την εκτέλεση της εφαρμογής στο παρασκήνιο (headless), την εισαγωγή cookies ή user agent, την τροποποίηση του μεγέθους από το παράθυρο φυλλομετρητή, την εκτέλεση αιτήματος σε ανώνυμη λειτουργία (incognito), την απενεργοποίηση των extensions ή popups, την εκτέλεση σε πλήρης οθόνη (maximized) κ.α.

Requests

Η requests βιβλιοθήκη αποτελεί μία από τις πιο εγκαταστημένες βιβλιοθήκες για την γλώσσα προγραμματισμού Python. Το εργαλείο αυτό επιτρέπει τον χρήστη να στείλει εύκολα HTTP αιτήματα σε οποιαδήποτε ιστοσελίδα επιλέγει ο χρήστης, ενώ αφού πραγματοποιηθεί η διαδικασία επιστρέφει πίσω ένα αντικείμενο από την ανταπόκριση, με περιεχόμενο την απάντηση, την κωδικοποίηση ιστοσελίδας, κωδικό κατάστασης κ.α (Chandra, R. et al., 2015). Κάποιοι από τους συχνότερους κωδικούς κατάστασης ισούνται με την τιμή 200, αν το αίτημα ολοκληρώθηκε επιτυχώς, ενώ μπορεί επίσης κάποιος χρήστης να αντιμετωπίσει τον κωδικό 401, αν το αίτημα απαγορεύτηκε από την ιστοσελίδα, 404 αν δεν βρέθηκε η ιστοσελίδα, καθώς και διάφορα άλλα. Όμοια με το εργαλείο webdriver της Selenium, η βιβλιοθήκη αυτή προσφέρει ένα πλήθος από παραμέτρους. Αυτά είναι:

- method: Ορίζεται η μέθοδος του αιτήματος. Η μέθοδος μπορεί να είναι μία από τα GET, OPTIONS, HEAD, POST, PUT, PATCH και DELETE και είναι απαραίτητη στον ορισμό της εντολής.

- url: Φυσικά, αναφέρεται στην ιστοσελίδα που θα εφαρμοστεί το αίτημα και είναι ένα ακόμα απαραίτητο στοιχείο στην συνάρτηση της μεθόδου.
- params: Δέχεται κάποιο λεξικό, κάποια λίστα από πλειάδες ή κάποια bytes για το αίτημα που θα αποστείλει στον ιστότοπο. Η προεπιλογή είναι None, δηλαδή δεν απαιτείται ο ορισμός της για την εκτέλεση του αιτήματος.
- allow_redirects: Ευθύνεται για την ενεργοποίηση ή απενεργοποίηση του redirection. Η προεπιλογή είναι True.
- auth: Ενεργοποιείται κάποια HTTP αυθεντικοποίηση που αφορά τον ιστότοπο. Η προεπιλογή είναι None.
- cert: Έχει να κάνει με το αν υπάρχει κάποιο αρχείο για την ταυτοποίηση ή κάποιο κλειδί. Όμοια με το auth και params.
- cookies: Τα cookies που σχετίζονται με την σελίδα. Επίσης λαμβάνει τιμή None κατά την μη τροποποίηση από τον χρήστη.
- headers: Τα headers κατά την αποστολή του αιτήματος. Για παράδειγμα, από ποιον φυλλομετρητή αποστέλλεται το αίτημα, την κωδικοποίηση σελίδας κτλ. Δεν χρειάζεται η υλοποίηση του πεδίου headers για την αποστολή του αιτήματος.
- proxies: Πρωτόκολλο που αφορά τα proxies του χρήστη, για την σύνδεση με διαφορετικές IP διευθύνσεις στον ιστότοπο. Δεν απαιτείται η συμπλήρωση του πεδίου αυτού για την κανονική λειτουργία του προγράμματος.
- stream: Αν στα δεδομένα της ιστοσελίδας θα εφαρμόζεται απευθείας λήψη ή θα γίνεται ροή των δεδομένων. Ορίζεται ως False, αν δεν επεξεργαστεί από τον χρήστη.
- timeout: Οδηγία που αφορά τον μέγιστο χρόνο αναμονής για την δημιουργία σύνδεσης με τον ιστότοπο και/η λήψη απάντησης από τον ιστότοπο. Από προεπιλογή παραμένει ανοιχτή η σύνδεση καθ'όλη την διάρκεια του προγράμματος ή μέχρι να δοθεί οδηγία για τον τερματισμό του από τον χρήστη.
- verify: Για την επιβεβαίωση της λειτουργίας του πρωτοκόλλου TLS που προσφέρει ασφάλεια κατά την σύνδεση μεταξύ του αποστολέα και του παραλήπτη. Σε κάθε αίτημα αυτή η οδηγία είναι ενεργοποιημένη, κάτι που φυσικά μπορεί να αλλάξει αν ο χρήστης επιθυμεί με την απενεργοποίησή της.

BeautifulSoup

Η βιβλιοθήκη BeautifulSoup βιβλιοθήκη επιτρέπει την άντληση δεδομένων από HTML και XML αρχεία και, με την χρήση κάποιας παραμέτρου, την ανάλυση των αρχείων αυτών γρήγορα και εύκολα (Richardson, L., 2007). Ακόμα, παρέχει την δυνατότητα της περιήγησης, αναζήτησης και τροποποίησης του δέντρου ανάλυσης που επιστρέφεται μετά την ανάλυση των παραπάνω αρχείων. Είναι χρήσιμη για περιπτώσεις όπου απαιτείται η εξαγωγή στοιχείων από το διαδίκτυο, δηλαδή όταν πρέπει να εφαρμοστεί η τεχνική web scraping. Χρησιμοποιείται σε συνδυασμό με κάποιο εργαλείο όπως το Selenium και Requests, για την ανάλυση των κειμένων. Στην περίπτωση που πρέπει να εφαρμοστεί σε συνδυασμό με το requests, ορίζεται ως πρώτη είσοδο το «page.text», που αποτελεί τον πηγαίο κώδικα HTML μετά από την πραγματοποίηση αιτήματος σε κάποιον

ιστότοπο. Στην άλλη περίπτωση, δηλαδή όταν εφαρμόζεται μαζί με το selenium εργαλείο, εισάγεται το "driver.page_source" που επιστρέφει τον κώδικα HTML, όμοια με την requests βιβλιοθήκη.

3.4 Τουριστικές πλατφόρμες

Tripadvisor

Το TripAdvisor περιέχει κριτικές για μια πληθώρα από ξενοδοχεία, αξιοθέατα, πτήσεις και εστιατόρια. Όλα αυτά χρησιμοποιούνται στον τουρισμό με τις πτήσεις να πραγματοποιούνται εμφανώς για την μεταφορά των τουριστών στον προορισμό τους, τα ξενοδοχεία για την διαμονή των πολιτών κατά την αποβίβαση στον προορισμό τους, τα αξιοθέατα ως τις τοποθεσίες που μπορεί να επισκεφθεί ένας τουρίστας κατά τον περίγυρό του στον τόπο τουρισμού και τα εστιατόρια για να ανακαλύψει κάποιος τουρίστας εξαιρετικές προτάσεις για την διατροφή του. Για αυτό τον λόγο, οι κριτικές συμβάλλουν σημαντικά στην επιλογή των ταξιδιωτών σε έναν τουριστικό προορισμό. Για παράδειγμα, ένα ξενοδοχείο με μέση κριτική στα τέσσερα αστέρια αποτελεί έναν ιδανικότερο προορισμό για τους τουρίστες, συγκριτικά με ένα άλλο ξενοδοχείο που μπορεί να έχει τρία αστέρια ως μέση βαθμολογία. Φυσικά, η επιλογή του ταξιδιωτή επηρεάζεται και από όλους παράγοντες, όπως είναι το χρηματικό ποσό που μπορεί και επιθυμεί να δαπανήσει, η τοποθεσία της κατοικίας αλλά και ο αριθμός των κριτικών. Ένα ξενοδοχείο με ένα μικρό πλήθος κριτικών δεν είναι αξιόπιστο, όταν συγκρίνεται με άλλα ξενοδοχεία που περιέχουν χιλιάδες κριτικές στην ψηφιακή πλατφόρμα του ταξιδιωτικού πρακτορείου.

Booking

Το Booking είναι μία ταξιδιωτική πλατφόρμα που παρουσιάζει στοιχεία για ένα πλήθος ξενοδοχείων και κατοικιών και επικοινωνεί απευθείας με τους οικοδεσπότες. Οι κριτικές στην σελίδα αυτή είναι λιγότερες και δεν υφίστανται αυτόματα μετάφραση κατά την συλλογή τους.

3.4.1 Σχόλια

Οι κριτικές στην σελίδα του Tripadvisor μεταφράζονται αυτόματα στην γλώσσα των επισκεπτών ιστοσελίδας για την καλύτερη εξυπηρέτησή των χρηστών. Παρόλα αυτά, η αυτόματη μετάφραση που εφαρμόζεται σε όλες τις κριτικές παρουσιάζει προβλήματα, διότι η αυτόματη μετάφραση μπορεί να μην είναι σωστή, με αποτέλεσμα το κείμενο να είναι δυσνόητο από τους χρήστες. Αυτό το φαινόμενο παρουσιάζεται περισσότερο στην μετάφραση από ξένες γλώσσες στην ελληνική γλώσσα, κάτι που ίσως παρατηρείται γιατί το σύστημα που ευθύνεται για την μετάφραση τροφοδοτείται με λιγότερα κείμενα ως είσοδο (Input) σε σχέση με άλλες πιο δημοφιλείς γλώσσες, όπως η εξαιρετικά δημοφιλής αγγλική γλώσσα. Για αυτό τον λόγο, αλλά ταυτόχρονα λαμβάνοντας υπόψη και τις περισσότερες πηγές που εκτελούνται για το Sentiment Analysis που είναι σε αγγλική γλώσσα, έγινε μεταφόρτωση κειμένων από την αγγλική σελίδα (<https://tripadvisor.com>) αλλά και την ελληνική σελίδα (<https://tripadvisor.com.gr>) της εταιρείας. Το μεγαλύτερο πλήθος σχολίων είναι ήδη στην αγγλική γλώσσα, κάτι που είναι ωφέλιμο μιας και η αυτόματη μετάφραση συνήθως επιστρέφει δυσνόητα κείμενα ιδίως σε πολύπλοκες προτάσεις.

Στην περίπτωση της σελίδας Booking, οι κριτικές που αντλούνται είναι στην γλώσσα την οποία τυπώθηκαν, οπότε υπάρχει ένα μεγάλο πλήθος από σχόλια γραμμένα σε γλώσσες που μπορεί να μην είναι γνώριμες στον επισκέπτη της πλατφόρμας.

Η εφαρμογή που δημιουργήθηκε συλλέγει όλες τις κριτικές ανεξάρτητα από την γλώσσα τους, και στην συνέχεια κατά την επεξεργασία, πριν εκκινηθεί η διαδικασία ανάλυσης συναισθημάτων, κρατάει μόνο τα στοιχεία τα οποία είναι ωφέλιμα στην έρευνα, δηλαδή τα σχόλια στην αγγλική και στην ελληνική γλώσσα.

Οι βασικές στήλες στο αρχείο csv που παράγεται μετά από την τεχνική web scraping συμπεριλαμβάνουν βασικά στοιχεία που θεωρήθηκαν αναγκαία για την σωστή ταξινόμηση των δεδομένων.

Τα στοιχεία από το Tripadvisor είναι ως εξής:

Αριθμός index	Τίτλος κριτικής	Ημερομηνία κριτικής	Κείμενο Κριτικής	Βαθμολογία κριτικής	Τίτλος	Σύνδεσμος
---------------	-----------------	---------------------	------------------	---------------------	--------	-----------

Πίνακας 4: Στήλες αρχείου δεδομένων Tripadvisor

Τα στοιχεία από το booking έχουν μικρές διαφορές από τα στοιχεία του Tripadvisor και είναι ως εξής:

Αριθμός index	Τίτλος κριτικής	Ημερομηνία κριτικής	Χώρα	Θετική κριτική	Αρνητική κριτική	Βαθμολογία κριτικής	Τίτλος	Σύνδεσμος
---------------	-----------------	---------------------	------	----------------	------------------	---------------------	--------	-----------

Πίνακας 5: Στήλες αρχείου δεδομένων Booking

Τα στοιχεία από τα δύο σετ δεδομένων αναλύονται ως εξής:

- Αριθμός index: Αριθμός που κρατάει σειρά στο αρχείο. Προστίθεται 1 για κάθε γραμμή/παράδειγμα στο σετ δεδομένων. Για παράδειγμα, η 1η γραμμή έχει index 0, η 2η γραμμή έχει index 1, κ.ο.κ. Φυσικά, η γραμμή με τις στήλες αγνοείται.
- Τίτλος κριτικής: Ο τίτλος της κριτικής, έτσι όπως παρουσιάζεται στην ιστοσελίδα.
- Ημερομηνία κριτικής: Ημερομηνία υποβολής σχολίου στην ιστοσελίδα.
- Χώρα: Η χώρα του χρήστη, έτσι όπως έχει προσδιοριστεί στον ιστότοπο.
- Θετική κριτική: Τα κείμενα κριτικής που έχει αναφέρει ο χρήστης ως θετικά στην πλατφόρμα του Booking.
- Αρνητική κριτική: Τα κείμενα κριτικής που έχει αναφέρει ο χρήστης ως αρνητικά στην πλατφόρμα του Booking.
- Κείμενο κριτικής: Το κείμενο κριτικής από την σελίδα του Tripadvisor. Δεν έχει ειδικά τμήματα για τα θετικά και τα αρνητικά όπως το Booking. Για αυτό εισάγονται σε ένα πεδίο, συνδυασμένα.
- Βαθμολογία κριτικής: Το σκορ του χρήστη που αφορά το συγκεκριμένο κατάλυμα ή αξιοθέατο. Οι τιμές στο σκορ του Booking είναι από το ένα μέχρι το δέκα (1-10), ενώ στο Tripadvisor είναι από το ένα μέχρι το πέντε (1-5).
- Τίτλος: Ο τίτλος του καταλύματος ή αξιοθέατου από το οποίο έγινε εξαγωγή των στοιχείων
- Σύνδεσμος: Ο σύνδεσμος (URL) το οποίο αφορά το κατάλυμα ή το αξιοθέατο από το οποίο έγινε εξαγωγή των στοιχείων.

Το πλήθος από ξενοδοχεία και αξιοθέατα ήταν αρκετά μεγάλο, οπότε η εξαγωγή στοιχείων από τον πηγαίο κώδικα (html) των ιστοσελίδων και η μετέπειτα αποθήκευση σε αρχεία έγινε σε μικρά κομμάτια για να μην παρουσιαστούν τυχόν λάθη με αποτέλεσμα να χαθεί η προγενέστερη πρόοδος στην διαδικασία αυτή. Για αυτό εισάγεται στον δίσκο αρχείο csv με τα στοιχεία σε κομμάτια ανά δέκα ξενοδοχεία ή αξιοθέατα. Με αυτό τον τρόπο, αν παρουσιαστεί κάποιο σφάλμα κατά την εκτέλεση της διαδικασίας αυτής, τότε μπορεί στιγμιαία να γίνει επανεκκίνηση της διαδικασίας από το τελευταίο στάδιο, που εξοικονομεί χρόνο και πηγές από το σύστημα.

3.5 Μετρικές αξιολόγησης

Οι συνήθεις τύποι από τις μετρικές που μελετώνται μετά την εκτέλεση των δοκιμών είναι (Tigerschiold, T., 2022):

- Ορθότητα (Accuracy): $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
- Ακρίβεια (Precision): $Precision = \frac{TP}{TP+FP}$
- Ανάκληση (Recall): $Recall = \frac{TP}{TP+FN}$
- F1-Score: $F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall}$

Υποθέτοντας ότι έχουμε δύο κλάσεις, 0 και 1, οι παράμετροι TP (True Positive), TN (True Negative), FP (False Positive) και FN (False Negative) υπολογίζονται ως εξής:

TP = Όσα παραδείγματα ταξινομήθηκαν στην κλάση 0 και πραγματικά ανήκουν στην κλάση 0
 TN = Όσα παραδείγματα ταξινομήθηκαν στην κλάση 1 και πραγματικά ανήκουν στην κλάση 1
 FP = Όσα παραδείγματα ταξινομήθηκαν στην κλάση 0, ενώ ανήκουν πραγματικά στην κλάση 1
 FN = Όσα παραδείγματα ταξινομήθηκαν στην κλάση 1, ενώ ανήκουν στην κυριολεξία στην κλάση 0

Φυσικά, στην περίπτωση που υπάρχουν τρεις ή παραπάνω κλάσεις, οι τύποι αλλάζουν. Για να παρουσιαστούν καλύτερα οι αλλαγές, έγινε προσθήκη ενός i σε κάθε παράμετρο, με το i να αναπαριστά την κάθε κλάση. Μετά τις αλλαγές, οι τύποι είναι ως εξής:

TP_i = Όσα παραδείγματα ταξινομήθηκαν στην κλάση i και πραγματικά ανήκουν στην κλάση i
 TN_i = Όσα παραδείγματα δεν ταξινομήθηκαν στην κλάση i και πραγματικά δεν ανήκαν σε αυτή την κλάση
 FP_i = Όσα παραδείγματα ταξινομήθηκαν στην κλάση i , ενώ δεν ανήκαν πραγματικά σε αυτή την κλάση
 FN_i = Όσα παραδείγματα δεν ταξινομήθηκαν στην κλάση i , ενώ ανήκαν πραγματικά σε αυτή την κλάση.

Η Ορθότητα (Accuracy) είναι μετρική που συνοψίζει την απόδοση ενός μοντέλου ταξινόμησης ως το πλήθος των σωστών προβλέψεων δια τον αριθμό των συνολικών προβλέψεων.

Η Ακρίβεια (Precision) είναι η ικανότητα του μοντέλου ταξινόμησης να προσδιορίζει μόνο τις σχετικές περιπτώσεις στο σύνολο δεδομένων, εφόσον ο τύπος είναι ο αριθμός των σωστών θετικών προβλέψεων δια όλες τις θετικές προβλέψεις. Είναι σημαντικό να αναφερθεί ότι η ακρίβεια είναι μέτρο ποιότητας σε ένα μοντέλο ταξινόμησης.

Η Ανάκληση (Recall) είναι η ικανότητα του μοντέλου ταξινόμησης να προσδιορίσει όλες τις σχετικές περιπτώσεις στο σύνολο δεδομένων, εφόσον ο τύπος είναι ο αριθμός των σωστών θετικών προβλέψεων δια όλες τις πραγματικά θετικές προβλέψεις. Σε αντίθεση με την ακρίβεια, η ανάκληση είναι μέτρο ποσότητας σε ένα μοντέλο ταξινόμησης.

Το F1-Score είναι ο αρμονικός μέσος μεταξύ της ακρίβειας και της ανάκλησης. Αποτελεί μετρική μηχανικής μάθησης που υπολογίζει το γενικό accuracy ενός μοντέλου, εφόσον συνδυάζει τα precision και recall. Ουσιαστικά, παρουσιάζει το πόσες φορές το μοντέλο έκανε σωστή πρόβλεψη σε ολόκληρο το σετ δεδομένων.

Το Support είναι ο αριθμός των πραγματικών προβλέψεων στο σύνολο δεδομένων (Kohli, S., 2022).

3.6 Μοντέλα

3.6.1 Bag-of-Words

Το μοντέλο Bag-of-Words είναι μία αναπαράσταση που μετατρέπει το κείμενο σε σταθερού μεγέθους διανύσματα που μετράνε το πόσες φορές εμφανίστηκε η κάθε λέξη (Great Learning Team, 2022). Αποτελεί έναν απλό και ευάλωτο τρόπο να εξάγεις χαρακτηριστικά από έγγραφα. Το μοντέλο αυτό περιλαμβάνει δύο πράγματα:

- Λεξιλόγιο γνωστών λέξεων
- Μέτρηση παρουσίας γνωστών λέξεων

Ονομάζεται bag of words, η τσάντα λέξεων (στα ελληνικά), διότι κάθε πληροφορία που αφορά την δομή ή την σειρά των λέξεων απορρίπτεται. Αυτό προκύπτει επειδή το μοντέλο ασχολείται μόνο με το αν εμφανίζονται γνωστές λέξεις αλλά και το πόσες φορές εμφανίστηκαν αυτές.

Παράδειγμα

Ας υποθέσουμε ότι έχουμε δύο προτάσεις στο έγγραφό μας. Οι προτάσεις είναι ως εξής:

- Πρόταση 1: Έχω ραντεβού με τον γιατρό αύριο.
- Πρόταση 2: Πέρασα ωραία εχθές με τον Γιάννη στον χορό.

Το μοντέλο θα έπαιρνε την κάθε λέξη στην πρόταση και θα έβρισκε τις συχνότητες. Οι μοναδικές λέξεις είναι:

- "έχω"
- "ραντεβού"
- "με"
- "τον"
- "γιατρό"
- "αύριο"

- "πέρασα"
- "ωραία"
- "εχθές"
- "Γιάννη"
- "στον"
- "χορό"

Συνεπώς έχουμε 12 μοναδικές λέξεις από έγγραφο με 14 συνολικές λέξεις (corpus).

Στη συνέχεια, πρέπει να εντοπισθούν οι συχνότητες των παραπάνω λέξεων στην κάθε πρόταση. Για παράδειγμα, για την πρώτη πρόταση θα βρούμε τις συχνότητες ως εξής:

- "έχω" = 1
- "ραντεβού" = 1
- "με" = 1
- "τον" = 1
- "γιατρό" = 1
- "αύριο" = 1
- "πέρασα" = 0
- "ωραία" = 0
- "εχθές" = 0
- "Γιάννη" = 0
- "στον" = 0
- "χορό" = 0

Άρα το διάνυσμα της πρώτης πρότασης, δηλαδή της "Έχω ραντεβού με τον γιατρό αύριο.", είναι [1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0]. Με τον ίδιο τρόπο, το διάνυσμα της δεύτερης πρότασης, δηλαδή της "Πέρασα ωραία εχθές με τον Γιάννη στο χορό", είναι [0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1].

3.6.2 TF-IDF

Το TF-IDF είναι αλγόριθμος που χρησιμοποιεί την συχνότητα των λέξεων για να προσδιορίσει το πόσο σχετικές είναι οι λέξεις σε ένα δοσμένο έγγραφο (Anirudha S., 2021). Σε αντίθεση με το BOW μοντέλο, το TF-IDF παρέχει πληροφορία για τις πιο σημαντικές λέξεις, καθώς και για τις ασήμαντες. Το TF αποτελεί ένα μέτρο του πόσο συχνά μια λέξη παρουσιάζεται σε ένα έγγραφο, ενώ το IDF δείχνει το πόσο σημαντική ήταν εκείνη η φράση. Ο πολλαπλασιασμός αυτών των δύο, δημιουργεί την τιμή του TF-IDF. Οι ορολογίες είναι οι παρακάτω:

- Term Frequency (Συχνότητα όρου): $tf_{t,d}$ = πλήθος του t (όρου) στο έγγραφο / πλήθος λέξεων στο d (έγγραφο)

- Document Frequency (Συχνότητα εγγράφου): df_t = πλήθος των εγγράφων που παρουσιάζεται το t (όρος)
- Inverse Document Frequency (Αντίστροφη συχνότητα εγγράφων): $idf_t = \frac{N}{df_t}$

Κάποια προβλήματα παρουσιάζονται στο IDF, όπως όταν υπάρχει μεγάλο πλήθος εγγράφων (corpus). Για παράδειγμα, αν είχαμε πάνω από εκατομμύρια έγγραφα, οι τιμές του IDF θα ήταν τρομακτικά υψηλές. Για αυτό, χρησιμοποιείται ο λογάριθμος του IDF. Επιπλέον, υπάρχει περίπτωση να έχουμε παρανομαστή μηδέν, αν δεν παρουσιάζεται ο όρος σε ένα έγγραφο. Αυτό επιλύεται με την προσθήκη μονάδας στην τιμή του παρανομαστή. Έτσι ο τύπος μετατρέπεται σε:

$$idf_t = \log\left(\frac{N}{df_t + 1}\right)$$

και ο τελικός τύπος σε:

$$w_{t,d} = tf_{t,d} * idf_t$$

Παράδειγμα

Ας υποθέσουμε ότι έχουμε δύο προτάσεις.

- Πρόταση 1: Πέρασα ωραία με την Μαρία σήμερα.
- Πρόταση 2: Πέρασα άσχημα στην τάξη σήμερα.

Από αυτές τις προτάσεις, παρατηρείτε ότι οι μοναδικές λέξεις είναι:

- "πέρασα"
- "ωραία"
- "με"
- "την"
- "Μαρία"
- "σήμερα"
- "άσχημα"
- "στην"
- "τάξη"

Συνεπώς, έχουμε 9 μοναδικές λέξεις από ένα corpus με 11 συνολικές λέξεις. Ο αναλυτικός πίνακας παρουσιάζεται παρακάτω.

Λέξη	tf_1	tf_2	idf	$tf * idf_1$	$tf * idf_2$
πέρασα	1/6	1/5	$\log 2/2 = 0$	0	0
ωραία	1/6	0	$\log 2/1 = 0.3$	0.05	0
με	1/6	0	$\log 2/1 = 0.3$	0.05	0
την	1/6	0	$\log 2/1 = 0.3$	0.05	0
Μαρία	1/6	0	$\log 2/1 = 0.3$	0.05	0
σήμερα	1/6	1/5	$\log 2/2 = 0$	0	0
άσχημα	0	1/5	$\log 2/1 = 0.3$	0	0.06
στην	0	1/5	$\log 2/1 = 0.3$	0	0.06
τάξη	0	1/5	$\log 2/1 = 0.3$	0	0.06

Πίνακας 6: Αποτελέσματα από παράδειγμα TF-IDF

Οι λέξεις που έχουν TF-IDF μηδέν είναι ασήμαντες, ενώ οι σχέσεις που έχουν κάποια τιμή θεωρούνται πιο σημαντικές. Αυτές είναι τα "ωραία", "με", "την", "Μαρία", "σήμερα", "άσχημα", "στην" και "τάξη". Το υψηλότερο σκορ TF-IDF δείχνει ότι οι λέξεις είναι πιο σχετικές, οπότε οι λέξεις "άσχημα", "στην" και "τάξη" θεωρούνται πιο σχετικές από τις υπόλοιπες, εφόσον έχουν μεγαλύτερη τιμή.

3.7 Πρόβλημα ανισορροπίας δεδομένων

Το πρόβλημα της ανισορροπίας δεδομένων έχει να κάνει με την παρουσία κλάσεων με άνιση κατανομή παραδειγμάτων σε ένα σετ δεδομένων (Mazumder, S., 2023). Βασικά, το κύριο πρόβλημα στην επεξεργασία δεδομένων είναι ότι δεν είναι πάντα εξισορροπημένα τα στοιχεία στις κλάσεις, με τις κλάσεις να μην περιέχουν ίδιο ή περίπου το ίδιο πλήθος στοιχείων ταξινομημένα στις κλάσεις. Φυσικά, πρέπει να υπάρχει αισθητή διαφορά στο πλήθος των κλάσεων για να παρουσιαστεί αυτό το πρόβλημα. Αυτό σημαίνει ότι αν είχαμε δύο κλάσεις με πλήθος στοιχείων 50 στην μία και 45 στην άλλη, δεν παρουσιάζεται αυτό το φαινόμενο. Σε αντίθεση, αν για παράδειγμα είχαμε δύο κλάσεις "Ναι" και "Όχι", με 90% των στοιχείων να ταξινομούνται στην "Ναι" και 10% στην "Όχι", τότε πρέπει να βρεθούν τρόποι τα στοιχεία αυτά να γίνουν περίπου ίσα για να εκπαιδευτεί το σύστημα καλύτερα.

Ένα σύστημα που κατατάσσει τις πληροφορίες μόνο σε κλάση "Ναι", που αφορά το 90% των στοιχείων, σίγουρα θα έχει ακρίβεια 90%, αλλά θα είναι λανθασμένο γιατί δεν μπορεί να ταξινομήσει τίποτα σε άλλη κλάση. Με λίγα λόγια, δεν μπορεί να διαχωρίσει τα δεδομένα και τυφλά ταξινομεί όλα τα στοιχεία στην μία κλάση, που επιστρέφει υψηλά αποτελέσματα εφόσον το πλήθος των στοιχείων στην πολυπληθέστερη κλάση είναι τόσο μεγάλο. Αυτό δείχνει ότι το μοντέλο είναι αποτυχημένο.

Για την διαχείριση αυτού του προβλήματος μπορούν να εφαρμοστούν κάποιες τεχνικές. Μία από τις πιο δημοφιλείς είναι το data sampling. Το data sampling περιέχει μία συλλογή τεχνικών που μεταμορφώνει (transform) το σετ εκμάθησης για να εξισορροπήσει, ή τουλάχιστον να εξισορροπήσει καλύτερα, την κατανομή των κλάσεων. Ανάμεσα σε αυτές είναι το Undersampling, το Oversampling και το SMOTE (Shelke, M. S. et al., 2017).

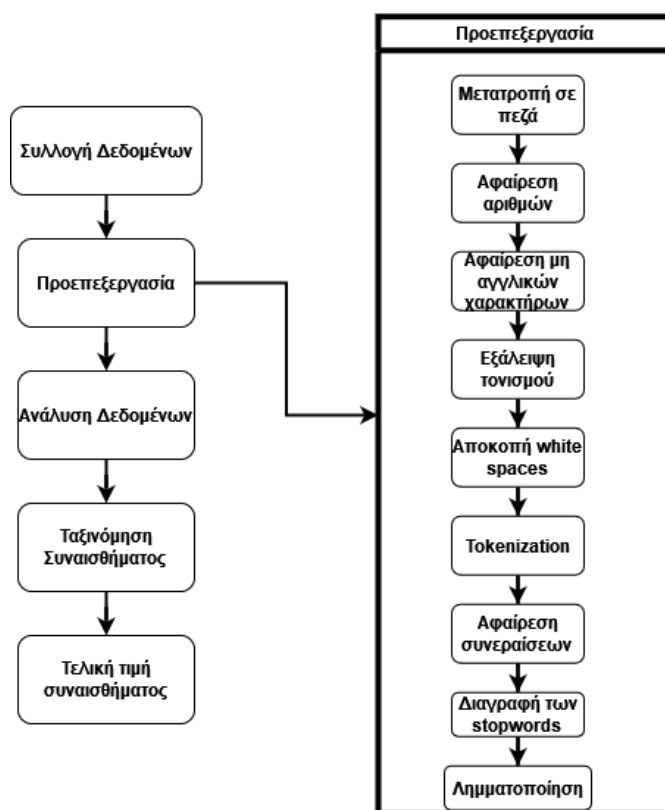
Το Undersampling ισορροπεί τα δεδομένα κρατώντας όλα τα δεδομένα της κλάσης μειονότητας και μειώνοντας τα δεδομένα της κλάσης με το μεγαλύτερο πλήθος, διαγράφοντας παραδείγματα από τις πολυπληθέστερες κλάσεις. Σε αντίθεση, το Oversampling προσθέτει περισσότερα στοιχεία στην κλάση μειονότητας μέχρι αυτή η κλάση να περιέχει περίπου το ίδιο πλήθος στοιχείων με την πιο δημοφιλή κλάση, αντιγράφοντας παραδείγματα από τις πολυπληθέστερες κλάσεις. Το SMOTE είναι μία κλάση Oversampling, που προσθέτει συνθετικά στοιχεία στην κλάση μειονότητας, και όχι ήδη υπαρκτά στοιχεία (αντίγραφα) από την πιο δημοφιλή κλάση.

Σε αυτή την εργασία θα πραγματοποιηθούν δοκιμές με τα εξής:

- Random Undersampling : Τυχαία διαγραφή παραδειγμάτων από τις κλάσεις πλειονότητας, μέχρι το πλήθος που περιέχουν να είναι περίπου ίσο με αυτό της κλάσης μειονότητας.
- Random Oversampling: Προσθήκη παραδειγμάτων στις κλάσεις μειονότητας με την τυχαία επιλογή παραδειγμάτων από την κλάση πλειονότητας.
- SMOTE: Προσθήκη παραδειγμάτων στις κλάσεις μειονότητας με συνθετικά παραδείγματα που προκύπτουν από την κλάση πλειονότητας. Τα συνθετικά παραδείγματα αυτά είναι όμοια με αυτά που υπάρχουν στο σετ δεδομένων, αλλά όχι ακριβώς τα ίδια.

4 Μεθοδολογική διαδικασία

Σε αυτό το παράρτημα πραγματοποιείται αναλυτική αναφορά σε κάθε βήμα της μεθοδολογικής διαδικασίας.



Διάγραμμα 4: Διάγραμμα με τα βήματα μεθοδολογίας

Παραπάνω εμφανίζεται το διάγραμμα με την μεθοδολογία της εργασίας.

4.1 Συλλογή Δεδομένων

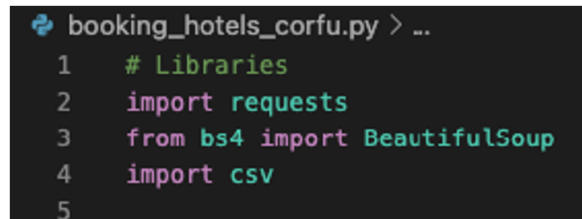
Αρχικά, το πρώτο βήμα που εμπλέκεται σε αυτή την διαδικασία είναι η απόκτηση των δεδομένων. Προφανώς για να υφίσταται προεπεξεργασία σε ένα πλήθος δεδομένων πρέπει πρώτα να έχει γίνει συλλογή κάποιων στοιχείων για να χρησιμοποιηθούν σε ένα πρόβλημα ανάλυσης συναισθημάτων. Τα δεδομένα αυτά που είναι σε σωστή μορφή ονομάζονται σετ δεδομένων και είναι συνήθως σε μορφή csv ή xlsx. Στη συγκεκριμένη περίπτωση, τα αρχεία με τα δεδομένα είναι σε μορφή CSV, με σημείο στίξης για τον διαχωρισμό των στοιχείων να είναι το ελληνικό ερωτηματικό (η semicolon στα αγγλικά). Το πρόγραμμα που εξάγει τα δεδομένα αποθηκεύει τα στοιχεία σε πολλαπλά αρχεία, οπότε πρώτα ενώνονται σε ένα αρχείο. Αυτό εξασφαλίζει ότι ακόμα και αν η διαδικασία αποθήκευσης των δεδομένων σταματήσει απότομα, θα είναι σίγουρο ότι δεν θα έχει χαθεί η πρόοδος.

Το σημείο ενδιαφέροντος που αποτελεί το θέμα εξερεύνησης σε αυτή την διπλωματική αποτελεί ο νομός Κέρκυρας και διάφορες κριτικές που έχουν τυπωθεί για την τοποθεσία αυτή. Το θέμα ορίστηκε ως ο νομός Κέρκυρας, διότι αποτελεί μέρος που επισκέπτονται συχνότερα τουρίστες από τα Επτάνησα. Επιπλέον, το νησί αποτελεί αξιοσημείωτο δείγμα τουρισμού στην χώρα, εφόσον δέχεται τουρίστες καθ' όλην τον χρόνο και περιέχει μεγάλη παρουσία κριτικών. Όπως είναι γνωστό τα Επτάνησα αποτελούν τα επτά νησιά του Ιονίου Πελάγους με αυτά να είναι η Κέρκυρα, η Λευκάδα, η Ζάκυνθος, η Κεφαλονιά, τα Κύθηρα, η Ιθάκη και οι Παξοί. Οι κριτικές που διατίθενται σε δημοφιλή ταξιδιωτικά γραφεία μπορεί να αφορούν αξιοθέατα, μέρη διαμονής ή την μεταφορά του τουρίστα στην περιοχή, με κάποιο μεταφορικό μέσο όπως ένα αεροπλάνο από κάποια Αεροπορική εταιρεία.

Για ένα μεγάλο διάστημα κυκλοφορούν στην αγορά διάφορα εργαλεία για την διευκόλυνση των χρηστών όσον αφορά την ταχύτατη εύρεση και κράτηση μερών διαμονής. Κάποια δημοφιλείς εργαλεία αυτά που γνωρίζει κάθε υπεύθυνος τουρίστας, αποτελούν σελίδες όπως το Booking, το TripAdvisor και το Trivago. Το Booking είναι πλατφόρμα με την οποία συνεργάζεται κάποιος ιδιώτης κατοικίας για να ενοικιάσει την κατοικία που του ανήκει σε ενδιαφερόμενους επισκέπτες που επιθυμούν να ταξιδέψουν γύρω από την τοποθεσία της κατοικίας αυτής. Σε αντίθεση, ιστοσελίδες όπως το TripAdvisor είναι, όπως αναφέρεται και στην ονομασία, εργαλεία που προσφέρουν προτάσεις στους επισκέπτες βάσει κάποιων στοιχείων όπως των τιμών που έχουν σε διάφορα ταξιδιωτικά γραφεία (Φερ' ειπείν το Booking). Με αυτό τον τρόπο, ο φιλοξενούμενος μπορεί να ανακαλύψει την βέλτιστη επιλογή με τα περισσότερα προνόμια στην ελάχιστη τιμή. Παρομοίως, το Trivago επίσης αποτελεί πλατφόρμα όπου κάποιος που επιθυμεί να ταξιδέψει σε κάποιον προορισμό μπορεί να συγκρίνει τιμές ανάμεσα στα ξενοδοχεία.

4.1.1 Συλλογή λίστας ξενοδοχείων

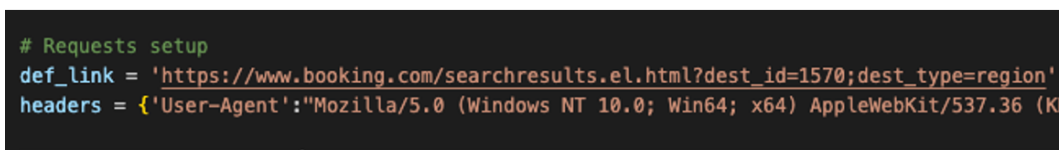
Booking Ξενοδοχεία



```
booking_hotels_corfu.py > ...
1  # Libraries
2  import requests
3  from bs4 import BeautifulSoup
4  import csv
5
```

Εικόνα 1: Βιβλιοθήκες για την συλλογή ξενοδοχείων από το Booking.com

Για την εξαγωγή στοιχείων από την σελίδα με τα ξενοδοχεία στην πλατφόρμα Booking έγινε εισαγωγή της βιβλιοθήκης requests, της βιβλιοθήκης bs4 που περιέχει το πακέτο BeautifulSoup και εισαγωγή βιβλιοθήκης csv. Το requests είναι υπεύθυνο για την επίσκεψη των ιστοσελίδων και επιστροφή των στοιχείων, το BeautifulSoup από το bs4 για την ερμάνευση και εξαγωγή των δεδομένων και το csv για την αποθήκευση των δεδομένων σε αρχείο csv.



```
# Requests setup
def_link = 'https://www.booking.com/searchresults.el.html?dest_id=1570;dest_type=region'
headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (K
```

Εικόνα 2: Ο ιστότοπος και τα headers που θα χρησιμοποιηθούν ως παράμετροι της βιβλιοθήκης requests

Στην συνέχεια, εισάγεται η διεύθυνση της ιστοσελίδας με τα ξενοδοχεία του νομού Κέρκυρας στην μεταβλητή def_link, για να χρησιμοποιηθεί μετέπειτα στην διαδικασία εξαγωγής των απαραίτητων στοιχείων. Επιπλέον, στην μεταβλητή headers ορίζονται τα user-agent, για να παρουσιάζεται στην ιστοσελίδα σωστό αίτημα από browser και όχι αίτημα από αυτοματοποιημένο bot. Γενικότερα, όταν επισκέπτεται κάποιος χρήστης σε ιστοσελίδες με browser, η ιστοσελίδα έχει την δυνατότητα να εντοπίσει κάποιες βασικές πληροφορίες από την συσκευή του χρήστη, εφόσον αλληλοεπιδράει ο χρήστης με το περιεχόμενό της. Στη συγκεκριμένη περίπτωση, τα στοιχεία που στέλνονται κατά την δημιουργία αιτήματος είναι από ποιο λογισμικό μπορεί να προέρχεται το αίτημα αυτό, ένα πολύ βασικό στοιχείο που συμπεριλαμβάνεται πάντα σε αιτήματα ιστοσελίδων μαζί με το λειτουργικό σύστημα το οποίο έχει ο χρήστης. Εξαιρετικά γνωστές εφαρμογές user-agent, είναι φυλλομετρητές όπως το Chrome της Google, το Firefox της Mozilla, το Safari της Apple κ.α. Για τους σκοπούς της εφαρμογής συλλογής δεδομένων στην εργασία, έγινε αποστολή αιτημάτων που εμφανίζουν ότι προέρχεται από το λογισμικό Opera με το λογισμικό συστήματος που εκτελεί το αίτημα να είναι το γνωστό λειτουργικό σύστημα Windows 10. Συγκεκριμένα, τα δεδομένα εισήχθησαν από την συσκευή που χρησιμοποιήθηκε για τους σκοπούς της εργασίας, που έχει εγκατεστημένο το Opera Browser GX, καθώς και το λειτουργικό σύστημα Windows 10.

```
# Loop until last page
while add <= 975:
    page = requests.get(url, headers=headers)
    if page.status_code != 200:
        print('Error with ' + url)
        break
    soup = BeautifulSoup(page.text, 'html.parser')
    listings = soup.find_all('a', {'data-testid': 'title-link'})
    # Saving hotel titles and links per page
    for i in range(len(listings)):
        title = listings[i].find('div').text
        link = listings[i]['href']
        titles.append(title)
        links.append(link)
    add += 25
    url = def_link + '&offset=' + str(add)
```

Εικόνα 3: Το κύριο τμήμα της εφαρμογής με την επανάληψη για την συλλογή των καταλυμάτων

Όπως παρουσιάζεται στην εικόνα, αρχικοποιείται μία μεταβλητή `add` με μηδενική τιμή, καθώς και δύο κενές λίστες `titles` και `links`, όπου θα χρησιμοποιηθούν για την αποθήκευση των τίτλων και των συνδέσμων ξενοδοχείων αντίστοιχα.

Η λογική πίσω από την επιλογή του 975 ως το όριο για τον βρόγχο, είναι το γεγονός ότι στην πλατφόρμα όταν αλλάζει σελίδα ο χρήστης προστίθεται το string "&offset=" με αθροίσεις ανά 25 να προστίθεται στο τέλος κάθε τέτοιου string. Αυτό το 25 προκύπτει επειδή κάθε σελίδα εμφανίζει στο μέγιστο 25 αντικείμενα. Για παράδειγμα, όταν πατηθεί το κουμπί για την μετάβαση στην δεύτερη σελίδα, ο σύνδεσμος στην πάνω μπάρα αναζήτησης του φυλλομετρητή μετατρέπεται στο ίδιο που ήταν προηγουμένως, με την διαφορά ότι έγινε προσθήκη "&offset=25". Παρομοίως στην επόμενη σελίδα θα μετατραπεί σε "&offset=50", στην επόμενη "&offset=75" κ.ο.κ. Επαναλαμβάνοντας την διαδικασία αυτή για την κάθε σελίδα ξενοδοχείου παρατηρείται ότι στην τελευταία σελίδα το string αυτό μετατρέπεται σε "&offset=975", κάτι που σημαίνει ότι δεν υπάρχει άλλη σελίδα για να πραγματοποιηθεί εξαγωγή δεδομένων, οπότε η διαδικασία τερματίζει σε εκείνο το σημείο. Ορίζεται όριο μικρότερο ίσο του αριθμού αυτού, αν και θα μπορούσε να εφαρμοστεί όριο μικρότερο των χιλίων μονάδων με ίδιο αποτέλεσμα.

Μέσα στον βρόγχο, αρχικά πραγματοποιείται το αίτημα στην σελίδα που έχει οριστεί πιο πάνω στο πρόγραμμα και εισχωρείτε στην μεταβλητή `page`. Για την υλοποίηση αυτή, γίνεται εφαρμογή της `get` συνάρτησης της βιβλιοθήκης `requests` με είσοδο τον σύνδεσμο του ιστότοπου σε συνδυασμό με το λεξικό `headers`. Στη συνέχεια, ελέγχεται αν το αίτημα ολοκληρώθηκε επιτυχώς για να συνεχίσει το πρόγραμμα την λειτουργία του. Αν το αίτημα ολοκληρώθηκε επιτυχώς τότε προχωράει η εφαρμογή στις επόμενες εντολές, ενώ σε αντίθετη περίπτωση εμφανίζεται μήνυμα με την σελίδα που προέκυψε το σφάλμα, όπου και τερματίζεται ο βρόγχος με την εντολή "`break`". Το `soup` είναι η μεταβλητή που εισάγεται το αποτέλεσμα του `Beautifulsoup` που αναλύει τον πηγαίο κώδικα από το `page` με την οδηγία "`html.parser`".

```

▼ <a href="https://www.booking.com/hotel/gr/corfu-sky-
loft.el.html?aid=356980&label=gog235jc-
1DCAQoggJCC3JlZ2lvd18xNTcwSAhYA2hciaEBmAEIuAEHyAEN2AED6AEB-
AFCiATBqATDuAlnn0SiBzACAdTCJDFxMzdh7DFmlTT17jctNGlW7S1h7DF0
LTYyMGM4NTg1NWYxN9gCB0ACAQ&ucfs=1&arphpl=1&dest_id=1570&
dest_type=region&group_adults=2&req_adults=2&no_rooms=1&
group_children=0&req_children=0&hpos=2&hapos=2&
sr_order=popularity&srpvid=034869732fa10182&
sreporch=1683557992&from=searchresults#hotelTmpl" class=
"e13098a59f" target="_blank" rel="noopener noreferrer"
data-testid="title-link"> Event
<div data-testid="title" class="fcab3ed991 a23c043802">
Sky Loft Corfu Old Town Apartments</div>

```

Εικόνα 4: HTML κώδικας όπου εμφανίζεται το στοιχείο που αναζητείται από την εφαρμογή

Μετάπειτα, πραγματοποιείται αναζήτηση στο soup για το αντικείμενο `a` στο html με `data-testid` ίσου ισοδυναμεί με `title-link`. Αυτό το αντικείμενο προκύπτει από την παραπάνω εικόνα. Όπως μπορεί να παρατηρηθεί, το αντικείμενο αυτό με την διεύθυνση ενός ξενοδοχείου περιέχει ένα αναγνωριστικό `data-testid` που ισούται με `title-link`. Με την βοήθεια αυτού του αναγνωριστικού γίνεται εφικτή η αναζήτηση των στοιχείων που απαιτούνται για τον σκοπό της εργασίας. Αφού βρεθούν όλα τα αντικείμενα `a` από κάθε ξενοδοχείο στη σελίδα, πραγματοποιείται βρόγχος με τον τίτλο του κάθε ξενοδοχείου και τον σύνδεσμο του. Ο τίτλος βρίσκεται σε ένα αντικείμενο `div` που αποτελεί παιδί του `a`, ενώ ο σύνδεσμος βρίσκεται στο `href` όπως φαίνεται και στην παραπάνω εικόνα. Όταν η διαδικασία ολοκληρωθεί και τα δεδομένα εισαχθούν στις λίστες επιτυχώς, τότε τελειώνει ο βρόγχος, η τιμή της μεταβλητής `add` αυξάνεται κατά 25 μονάδες και ο σύνδεσμος τροποποιείται κατάλληλα με την νέα τιμή της μεταβλητής `add`.

```

# Replace any hidden commas that may mess with the csv format
titles = [item.replace(',', '') for item in titles]

# Saving output to file
with open('hotels_corfu.csv', 'w', newline='', encoding='utf-8') as csvfile:
    writer = csv.writer(csvfile, quotechar = '|', quoting=csv.QUOTE_MINIMAL)
    data = list(zip(titles, links))
    for row in data:
        row = list(row)
        writer.writerow(row)

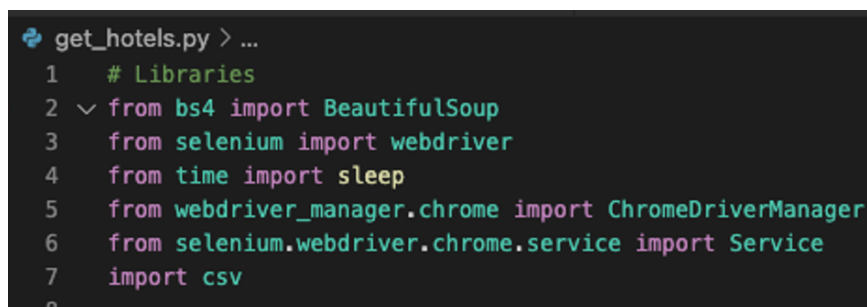
```

Εικόνα 5: Διαγραφή τυχόν σημείων στίξης που συγκρούονται με το μορφότυπο csv και αποθήκευση σε αρχείο

Το πρόγραμμα διαγράφει κάθε κόμμα στίξης που εμφανίζεται στους τίτλους καταλυμάτων, που θα μπορούσε να προκαλέσει σφάλμα στο αρχείο csv, και μετά αποθηκεύονται τα δεδομένα σε αρχείο csv. Για να μην οριστεί από την Python κάποιο `newline` σύμβολο που δεν είναι επιθυμητό, ορίστηκε ως κενό και χρησιμοποιήθηκε η κωδικοποίηση `utf-8`. Μετά έγινε λίστα με τον συνδυασμό των `titles` και `links`. Τελικά, πραγματοποιήθηκε εγγραφή όλων των δεδομένων στο αρχείο για την χρήση στο επόμενο στάδιο, της εξαγωγής στοιχείων.

Tripadvisor Ξενοδοχεία

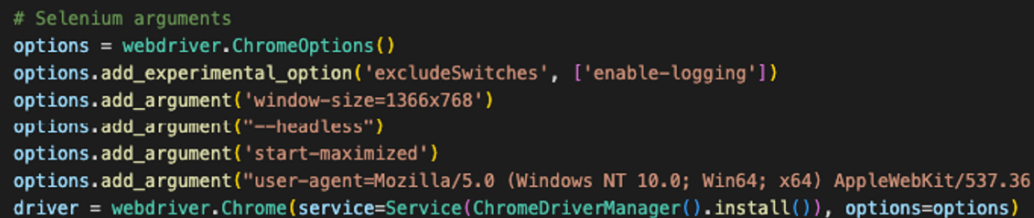
Στο Tripadvisor η διαδικασία γίνεται πιο πολύπλοκη, διότι δεν σου επιτρέπει να συλλέξεις πολλά δεδομένα καταλυμάτων κατά την μία εκκίνηση του προγράμματος. Για αυτό τον λόγο, η διαδικασία συλλογής καταλυμάτων μετατράπηκε στην σταδιακή συλλογή ξενοδοχείων με την δημιουργία δύο αρχείων. Ενός αρχείου για την προσθήκη ξενοδοχείων από το άλλο αρχείο και το άλλο αρχείο για την προσθήκη (append) ξενοδοχείων σε κάθε εκκίνηση. Συνοπτικά, το πρόγραμμα εκτελείται από το 0 έως το 30, από το 30 έως το 60 κ.ο.κ. Κάποιες φορές μπορεί να μην επιστρέφει αποτέλεσμα, οπότε εμφανίζεται ένα μήνυμα λάθους από την εφαρμογή και το πρόγραμμα εκτελείται ξανά στο σημείο που δεν ολοκληρώθηκε επιτυχώς.



```
get_hotels.py > ...
1  # Libraries
2  from bs4 import BeautifulSoup
3  from selenium import webdriver
4  from time import sleep
5  from webdriver_manager.chrome import ChromeDriverManager
6  from selenium.webdriver.chrome.service import Service
7  import csv
8
```

Εικόνα 6: Εισαγωγή διαφορετικού σεν βιβλιοθηκών για την χρήση selenium αντί για requests

Οι βιβλιοθήκες σε αυτό το αρχείο, τροποποιούνται με την κύρια διαφορά, ότι πλέον δεν εφαρμόζεται η requests βιβλιοθήκη για την εξαγωγή του πηγαίου κώδικα, αλλά η βιβλιοθήκη selenium που λειτουργεί σαν αυτοματοποιημένο bot που επισκέπτεται τις σελίδες και μπορεί να αλληλοεπιδράει με το περιεχόμενο χρησιμοποιώντας το Chrome. Παρόλα αυτά, η βιβλιοθήκη bs4 ορίζεται και σε αυτή την εφαρμογή, διότι δεν απαιτήθηκε αλλαγή στον τρόπο ανάλυσης πηγαίου κώδικα. Παρακάτω, προσδιορίζεται το webdriver από το selenium, το οποίο απαιτείται για την λήψη των ρυθμίσεων που θα προσδιοριστούν για την σωστή εκτέλεση της εφαρμογής. Το sleep επίσης χρησιμοποιείται από την βιβλιοθήκη time, διότι το selenium επισκέπτεται τις σελίδες όπως ένας χρήστης, οπότε είναι πιο αργό σχετικά με την απλή βιβλιοθήκη requests. Το ChromeDriverManager από την βιβλιοθήκη webdriver_manager επίσης εφαρμόζεται για την αυτόματη λήψη και εισαγωγή του αρχείου chromedriver που είναι αναγκαίο σε κάθε εκτέλεση με selenium. Το Service χρησιμοποιείται σε συνδυασμό με το ChromeDriverManager και το csv για την σωστή αποθήκευση του αρχείου με μορφότυπο csv.



```
# Selenium arguments
options = webdriver.ChromeOptions()
options.add_experimental_option('excludeSwitches', ['enable-logging'])
options.add_argument('window-size=1366x768')
options.add_argument('--headless')
options.add_argument('start-maximized')
options.add_argument('user-agent=Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36')
driver = webdriver.Chrome(service=Service(ChromeDriverManager().install()), options=options)
```

Εικόνα 7: Εγκατάσταση selenium και εισαγωγή παραμέτρων

Για τον σωστή λειτουργία του Selenium απαιτούνται κάποιες ρυθμίσεις. Στην προκειμένη περίπτωση ορίζεται το αντικείμενο ChromeOptions στην μεταβλητή options. Στο επόμενο στάδιο,

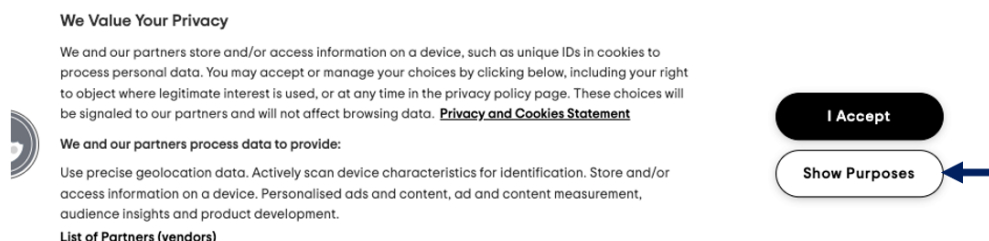
δίνεται εντολή έτσι ώστε να μην πραγματοποιείται κάποια καταγραφή (log) στο τερματικό του συστήματος και πιο κάτω ορίζεται το μέγεθος του παραθύρου φυλλομετρητή, εφόσον θα εκτελέσει αυτόματα την εφαρμογή chrome. Η οδηγία "headless" χρησιμοποιείται για να μην εμφανίζεται το παράθυρο chrome στον χρήστη, αλλά αντιθέτως να εκτελείται σιωπηλά στο παρασκήνιο του συστήματος. Επιπλέον, ορίζεται το "start-maximized" για να ξεκινήσει σε πλήρης οθόνης το chrome. Έπειτα, εφαρμόζεται ένα στοιχειώδες header, όπως ακριβώς και στην βιβλιοθήκη requests που αναλύθηκε προηγουμένως. Τέλος, ορίζεται το driver με που είναι το αποτέλεσμα μετά την εκκίνηση του selenium με τις ρυθμίσεις αυτές.

```
def_url = 'https://www.tripadvisor.com.gr/Hotels-g189458-oa60-Corfu_Ionian_Islands-Hotels.html'
tripadvisor_titles = []
tripadvisor_links = []

driver.get(def_url)
sleep(5)
element = driver.find_element('xpath', '//*[@id="onetrust-pc-btn-handler"]')
element.click()
sleep(2)
element2 = driver.find_element('xpath', '//*[@id="onetrust-pc-sdk"]/div[3]/div[1]/button[1]')
element2.click()
sleep(2)
```

Εικόνα 8: Ιστοσελίδα προς επίσκεψη και άρνηση cookies μέσω εντολών

Όπως αναφέρθηκε και στο προηγούμενο πρόγραμμα, χρησιμοποιείται ένας σύνδεσμος που σε μεταφέρει στην σελίδα με όλα τα ξενοδοχεία της Κέρκυρας, με την διαφορά ότι πλέον πραγματοποιείται αναφορά στην σελίδα του Tripadvisor και όχι στην σελίδα του Booking. Μετέπειτα, αποστέλλεται αίτημα για την μετάβαση στην ιστοσελίδα αυτή. Σε αυτό το στάδιο αναστέλλεται η λειτουργία της εφαρμογής για ένα διάστημα πέντε δευτερολέπτων, έτσι ώστε να έχει αρκετό χρόνο να φορτώσει όλα τα στοιχεία η σελίδα για την σωστή εκτέλεση των επόμενων εντολών. Στη συνέχεια, εμφανίζεται στην σελίδα οθόνη σχετικά με την αποδοχή ή μη των cookies, όπως σε κάθε σελίδα, όπου και πατείται το κουμπί άρνησης.



Εικόνα 9: Πάτημα του κουμπιού που δείχνει το μπλέ βέλος για την εμφάνιση του κουμπιού άρνησης των cookies

Η πρώτη εντολή find_element υλοποιείται για να βρεθεί το αντικείμενο που γράφει "Show Purposes", που εμφανίζεται με ένα μπλέ βέλος στην παραπάνω εικόνα. Από κάτω η εντολή εφαρμόζεται για να πατηθεί το κουμπί αυτό με μία αναμονή δύο δευτερολέπτων να ακολουθεί για να υπάρχει χρόνος για την εμφάνιση της επόμενης οθόνης. Σε εκείνο το σημείο πατιέται το κουμπί για την άρνηση όλων των cookies, αφού αναζητηθεί και βρεθεί με επιτυχία η τοποθεσία του κουμπιού αυτού με παρόμοιο τρόπο όπως και στο προηγούμενο στοιχείο.

About Your Privacy

We process your data to deliver content or advertisements and measure the delivery of such content or advertisements to extract insights about our website. We share this information with our partners on the basis of consent and legitimate interest. You may exercise your right to consent or object to a legitimate interest, based on a specific purpose below or at a partner level in the link under each purpose. These choices will be signaled to our vendors participating in the Transparency and Consent Framework.

[PRIVACY AND COOKIES STATEMENT](#)

Allow All

Manage Consent Preferences



Reject All

Confirm My Choices

Εικόνα 10: Πάτημα του κουμπιού που δείχνει το μπλέ βέλος για την άρνηση των cookies

Στην παρακάτω εικόνα μπορεί να παρατηρηθεί ο βρόγχος που εφαρμόζεται για την άντληση των δεδομένων. Γνωρίζοντας ότι ο ιστότοπος περιέχει 2147 καταλύματα ορίζεται βρόγχος που να ξεκινάει από το 0 μέχρι το 2147 με βήμα 30, κάτι που σημαίνει ότι σε κάθε επανάληψη θα αυξάνεται κατά 30 μονάδες η τιμή του *i*. Στην φάση αυτή, εκτυπώνεται το *i* για να εμφανιστεί μήνυμα στο τερματικό σε κάθε επανάληψη. Το νέο url σελίδας ορίζεται ως το link που προσδιορίστηκε στην μεταβλητή *def_link*, με την τιμή του *oa60* να τροποποιείται με την τιμή του *i*. Εφόσον υπάρχει γνώση ότι η ιστοσελίδα περιέχει 30 στοιχεία σε κάθε σελίδα και το *oa* αλλάζει με αθροίσεις του 30, έτσι αλλάζει το link κατάλληλα. Η διαφορά με το booking είναι ότι εκεί εμφανίζονται 25 στοιχεία σε κάθε σελίδα, ενώ σε αυτό 30 στοιχεία. Από κάτω αποστέλνεται αίτημα για την επίσκεψη στον ιστότοπο και χρησιμοποιείται το BeautifulSoup για την ανάλυση του πηγαίου κώδικα σελίδας, μετά από την παύση ενός δευτερολέπτου για την φόρτωση του περιεχομένου.

```
for i in range(0, 2147, 30):
    print(i)
    new_url = def_url.replace('oa60', 'oa{}'.format(str(i)))
    driver.get(new_url)
    sleep(1)
    soup = BeautifulSoup(driver.page_source, 'html.parser')
    div = soup.find_all('div', {'data-automation' : 'hotel-card-title'})
    for item in div:
        tripadvisor_titles.append(item.find('a').find('div').get_text())
        tripadvisor_links.append(item.find('a')['href'])
```

Εικόνα 11: Κεντρικός βρόγχος για την εύρεση όλων των ξενοδοχείων και των συνδέσμων

Το *div* δέχεται όλα τα *div* στοιχεία που περιέχουν το *data-automation* με τιμή το "hotel-card-title" που είναι ένα αναγνωριστικό μοναδικό σε κάθε ξενοδοχείο. Στην συνέχεια για το μήκος του *div* εισάγονται στην λίστα τα σημεία που υπάρχουν τα αντικείμενα που ψάχνουμε. Στην παρακάτω εικόνα εμφανίζεται ένα παράδειγμα ενός ξενοδοχείου. Όπως φαίνεται περιέχει *div* με *data-automation* hotel-card-title που έχει ως παιδί το *a* στοιχείο με τον σύνδεσμο για την

ιστοσελίδα του συγκεκριμένου καταλύματος. Επιπλέον, το ίδιο το a έχει το δικό του παιδί, με αυτό να είναι το div που περιέχει τον τίτλο του ξενοδοχείου.

```
<div class="jsTLT K" data-automation="hotel-card-title">
  <a target="_blank" href="/Hotel_Review-g775844-d1415171-Reviews-Atlantica_Grand_Mediterraneo_Resort-Ermones_Corfu_Ionian_Islands.html" class="BMQDV _F G- wSSLs SwZTJ F6wzt ukgoS">
    <div class="nBrpc Wd o W">
      "61. "
      <!-- -->
      "Atlantica Grand Mediterraneo Resort"
    </div>
  </a>
</div>
```

Εικόνα 12: Παράδειγμα πηγιαίου κώδικα από ένα ξενοδοχείο κατά την αναζήτηση στο Tripadvisor

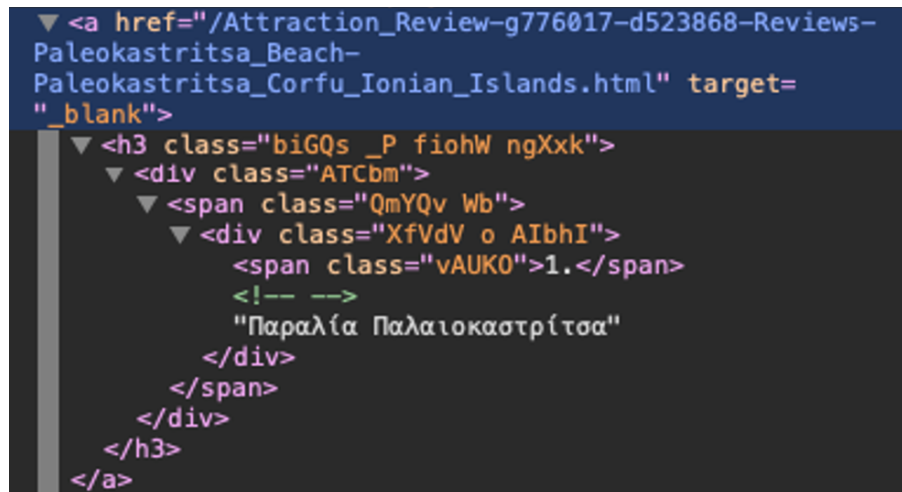
Στο τέλος γίνεται η ίδια διαδικασία όπως και στο προηγούμενο πρόγραμμα για το booking, δηλαδή διαγράφονται τα σημεία στίξης και εισάγονται τα αντικείμενα σε αρχείο csv, με την μόνη διαφορά ότι όλα είναι υπό συνθήκη και ότι εφαρμόζεται πλέον append και όχι write όπως στην προηγούμενη υλοποίηση. Το append (a) είναι για να συμπληρωθεί το αρχείο αν υπάρχει. Αν δεν υπάρχει ήδη στον φάκελο της εφαρμογής, τότε δημιουργείται αυτόματα το αρχείο και μετά συμπληρώνεται.

```
if len(tripadvisor_links) > 0:
    with open('tripadvisor_hotels.csv', 'a', newline='', encoding='utf-8') as csvfile:
        writer = csv.writer(csvfile, quotechar='|', quoting=csv.QUOTE_MINIMAL)
        data = list(zip(tripadvisor_titles, tripadvisor_links))
        for row in data:
            row = list(row)
            writer.writerow(row)
else:
    print('Error in getting hotels!')
```

Εικόνα 13: Κατά την εύρεση στοιχείων από το πρόγραμμα, δημιουργείται αρχείο (αν δεν υπάρχει) και έπειτα συμπληρώνεται

Tripadvisor Αξιοθέατα

Η υλοποίηση της εφαρμογής για την συλλογή δεδομένων είναι σχεδόν ίδια με αυτή των ξενοδοχείων, αφού αφορούν την ίδια σελίδα. Φυσικά, έχουν αλλαγές σχετικά με το ποιες κλάσεις αναζητούνται.



Εικόνα 14: Παράδειγμα πηγαίου κώδικα από ένα αξιοθέατο κατά την αναζήτηση στο Tripadvisor

Σε αυτό το σημείο εμφανίζεται η κλάση που αναζητούμε στον πηγαίο κώδικα. Το `a` αντικείμενο περιέχει το `href` που απαιτείται για την εργασία και το `span` αντικείμενο συμπεριλαμβάνει την ονομασία του αξιοθέατου που πρέπει να εξαχθεί. Σε αντίθεση με τα ξενοδοχεία, στη σελίδα με τα αξιοθέατα δεν υπάρχουν μοναδικά αναγνωριστικά που αφορούν αποκλειστικά τα αξιοθέατα, οπότε έγινε εφαρμογή της συνάρτησης `find_elements()` που εντοπίζεται μέσα στα στοιχεία της βιβλιοθήκης `selenium`. Το `find_elements()` λαμβάνει δύο εισόδους, μία είσοδο που προσδιορίζεται ο τύπος στοιχείου που πρέπει να εντοπιστεί και στην δεύτερη είσοδο το στοιχείο. Στην προκειμένη περίπτωση ορίζεται ως τύπου στοιχείου το `xpath` και ως στοιχείο η επιθυμητή διαδρομή για το συγκεκριμένο αντικείμενο που πρέπει να εντοπιστεί.

```

titles = driver.find_elements('xpath', '//*[@id="lithium-root"]/main/div[1]/div/div[3]/div/div[2]/ \
div[2]/div[2]/div/div[3]/div/div[2]/div/div/section/div/div/ \
div/article/div[2]/header/div/div/a[1]/h3/div/span/div')
links = driver.find_elements('xpath', '//*[@id="lithium-root"]/main/div[1]/div/div[3]/div/div[2]/ \
div[2]/div[2]/div/div[3]/div/div[2]/div/div/section/div/div/ \
div/article/div[2]/header/div/div/a[1]')

```

Εικόνα 15: Εύρεση στοιχείων στον πηγαίο κώδικα βάσει του `xpath`

Όπως παρουσιάζεται στην εικόνα, δίνεται ως στοιχείο η διαδρομή των αντειμένων στον πηγαίο κώδικα με τον χαρακτήρα `backslash` στο τέλος γραμμών για να χωριστεί το `string` της διαδρομής, στην ίδια εντολή, σε πολλές γραμμές σωστά.

```

for item in titles:
    tripadvisor_titles.append(item.text)
for link in links:
    tripadvisor_links.append(link.get_attribute('href'))

```

Εικόνα 16: Αποθήκευση στοιχείων στις λίστες

Στο επόμενο στάδιο εισάγονται τα δεδομένα στις λίστες με την διαφορά ότι, εφόσον δεν γίνεται εφαρμογή του BeautifulSoup, δεν υπάρχει `get_text()` συνάρτηση για την άντληση κειμένων από τα δεδομένα και τα attributes δεν εξάγονται με τετράγωνες αγκύλες αλλά με την συνάρτηση `get_attribute()`. Τέλος, η διαδικασία αποθήκευσης σε αρχείο είναι η ίδια με το προηγούμενο που αναλύθηκε στα ξενοδοχεία.

4.1.2 Εξαγωγή στοιχείων

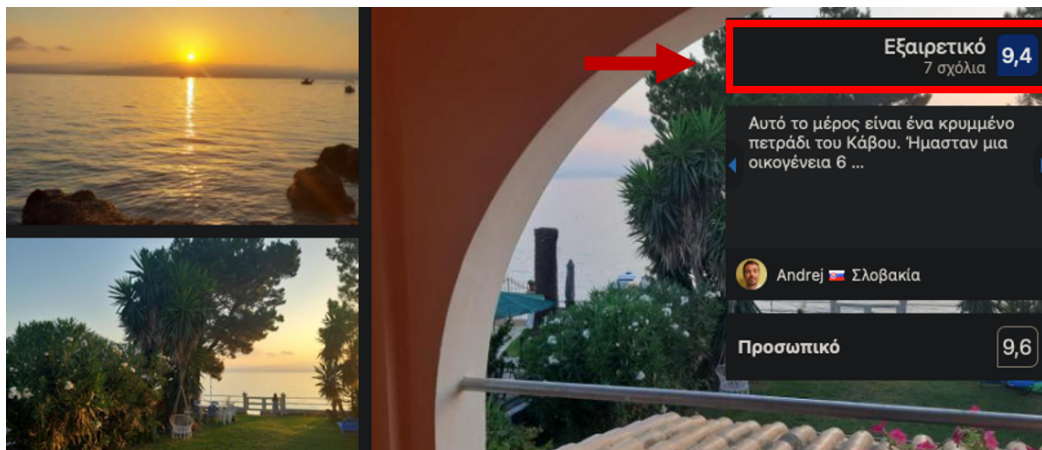
Booking Ξενοδοχεία

Για την υλοποίηση της διαδικασίας αυτής φορτώνονται τα στοιχεία του Selenium χωρίς το BeautifulSoup, ακριβώς όπως το τελευταίο πρόγραμμα που αναλύθηκε στην εργασία. Ακόμα, προστίθεται στις βιβλιοθήκες η βιβλιοθήκη pandas για την δημιουργία και διαχείριση DataFrame. Τα DataFrame αυτά αποτελούν μορφή που είναι πιο εύκολα επεξεργάσιμη και θα χρησιμοποιηθεί για την δημιουργία αρχείου csv με όλα τα δεδομένα. Στην συνέχεια φορτώνονται τα στοιχεία από το csv που δημιουργήθηκε προηγουμένως. Τα στοιχεία εντοπίζονται εντός του φακέλου με την ονομασία «eptanisa» στο αρχείο "hotels_corfu.csv". Αφού διαβαστεί το αρχείο από το πρόγραμμα εισάγεται βρόγχος για κάθε γραμμή του αρχείου. Αν υπάρχει κάποιο όνομα που δεν υπάρχει ήδη μέσα στην λίστα με τους τίτλους ξενοδοχείων, διότι μπορεί να προκύψει κάποιο ξενοδοχείο πάνω από μία φορά, τότε εισάγεται το περιεχόμενο αυτό εντός της λίστας με την ονομασία «hotel_names» και το στοιχείο μετά την στίξη, που είναι ο σύνδεσμος που πρέπει να προσπελαστεί για την αποστολή των αιτημάτων στην σελίδα, καταχωρείται μέσα στην λίστα με την ονομασία "hotel_links".

```
# Loading data from file
with open('./eptanisa/hotels_corfu.csv', 'r', encoding = 'utf-8') as csvfile:
    for row in csvfile:
        if row.split(',')[0] not in hotel_names:
            hotel_names.append(row.split(',')[0])
            hotel_links.append(row.split(',')[1])
```

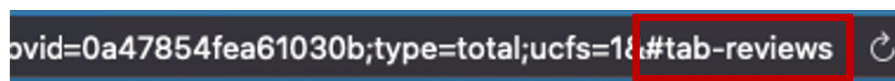
Εικόνα 17: Διάβασμα αρχείου ξενοδοχείων Κέρκυρας

Στην συνέχεια αποστέλλεται αίτημα στην σελίδα με τις κριτικές. Η τοποθεσία που συμπεριλαμβάνει τις κριτικές στον ιστοχώρο εντοπίζεται εντός κάποιων side panel, που εμφανίζεται στην οθόνη του χρήστη αφού πατήσει πάνω σε ένα κουμπί που αναφέρει το πλήθος κριτικών και την μέση βαθμολογία καταλύματος.



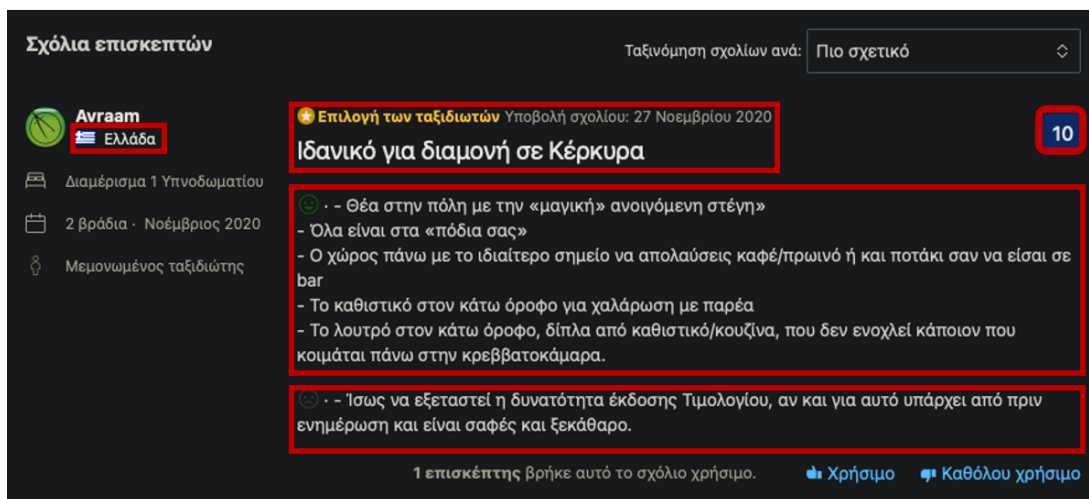
Εικόνα 18: Πάτημα του κουμπιού που αναδεικνύεται με κόκκινο χρώμα για το άνοιγμα του tab με τις κριτικές

Όπως μπορεί να παρατηρηθεί στο παραπάνω παράδειγμα, υπάρχει η μέση βαθμολογία καταλύματος σε αριθμητική τιμή μαζί με το πλήθος σχολίων και την μέση βαθμολογία σε μορφή κειμένου. Αυτό, σημαδεύεται με ένα κόκκινο περίγραμμα και ένα βέλος στο αριστερό μέρος για την διευκόλυνση στην εξήγηση του περιεχομένου στην εικόνα.



Εικόνα 19: Εμφάνιση του συνδέσμου στην μπάρα διευθύνσεων του φυλλομετρητή μετά το άνοιγμα του tab με τις κριτικές

Αφού πατηθεί το κουμπί με τις κριτικές, που σημαδεύτηκε στην εικόνα με τα σχόλια, εμφανίζεται ένα panel με τα σχόλια και ο σύνδεσμος τροποποιείται με την προσθήκη του string "#tab-reviews" στο τέλος του συνδέσμου στην μπάρα διευθύνσεων του φυλλομετρητή. Το panel που προβάλλεται περιέχει κάποια δεδομένα όπως οι ξεχωριστές κριτικές για το δωρεάν WIFI, αν συμπεριλαμβάνεται μέσα στις παροχές που προσφέρει το κατάλυμα που νοίκιασε ο πελάτης. Επιπλέον, περιέχει κριτική για το προσωπικό, την άνεση, τις παροχές, την σχέση ποιότητας/τιμής, την καθαριότητα αλλά και την τοποθεσία που τοποθετείται το κατάλυμα που επέλεξε ο τουρίστας.



Εικόνα 20: Παράδειγμα σχολίου στον ιστοχώρο Booking

Οι ατομικές κριτικές βρίσκονται στην παραπάνω μορφή, και σημαδεύονται τα στοιχεία που εξαγονται κατά την εκτέλεση του προγράμματος με κόκκινο χρώμα. Αυτά είναι:

- Ο τίτλος σχολίου που συμπεριλαμβάνει και το αν το σχόλιο αυτό αποτελεί "Επιλογή των ταξιδιωτών", δηλαδή είναι το top σχόλιο, στο κείμενο, που εντοπίζεται στο πάνω αριστερό κουτί με κόκκινο χρώμα,
- η ημερομηνία υποβολής σχολίου που εντοπίζεται δεξιά από τον τίτλο και περιέχει την ημερομηνία που ο χρήστης υποβάλλει την κριτική του στον ιστότοπο,
- τα θετικά σημεία που εμφανίζονται εντός της σελίδας με ένα χαρούμενο πράσινο emoji, που βρίσκονται ακριβώς κάτω από τον τίτλο στο μεγάλο κουτί της εικόνας,
- τα αρνητικά σημεία που προβάλλονται με μία γκριζα δυσάρεστη φατσούλα στο κάτω μέρος και τοποθετούνται κάτω από τα θετικά σχόλια,
- η χώρα προέλευσης του χρήστη που καταχώρησε την κριτική που βρίσκεται στο αριστερό πάνω μέρος κάτω από το username, και
- την βαθμολογία του σχολίου στο δεξί μέρος, που στο συγκεκριμένο παράδειγμα ορίζεται ως την αριθμητική τιμή δέκα, κάτι που σημαίνει ότι ο χρήστης δεν εντόπισε κανένα πρόβλημα που να ήταν δυσάρεστο για να μειώσει την εμπειρία του.

```
# Access review tabs
driver.get(hotel_links[0] + '#tab-reviews')
# Reject all cookies
driver.find_element('xpath', '//*[@id="onetrust-reject-all-handler"]').click()
```

Εικόνα 21: Άνοιγμα tab με κριτικές και απόρριψη cookies

Οι εντολές για την προσπέλαση των tabs είναι η προσθήκη του string "#tab-reviews" για να εμφανιστεί στο πρόγραμμα το tab που περιέχει τις κριτικές και να επιτευχθεί η επιθυμητή

εξαγωγή δεδομένων. Επίσης, με παρόμοιο τρόπο, όπως και στο Tripadvisor, πραγματοποιείται άρνηση των cookies με το πάτημα του ανάλογου κουμπιού.

```
start = 0
end = 10
for k in range(0, len(hotel_links)):
    titles.clear()
    countries.clear()
    dates.clear()
    pos_reviews.clear()
    neg_reviews.clear()
    reviewscore.clear()
    hotel_name.clear()
    hotel_link.clear()
    for j in range(start, end):
        driver.get(hotel_links[j] + '#tab-reviews')
```

Εικόνα 22: Καθαρισμός λιστών, ορισμός μεταβλητών start και end και επίσκεψη στην σελίδα με τα ξενοδοχεία

Όπως φαίνεται δημιουργούνται δύο μεταβλητές start και end με τιμές μηδέν και δέκα αντίστοιχα και εισάγεται βρόγχος που ξεκινάει από το μηδέν μέχρι την τιμή του συνολικού αριθμού καταλυμάτων. Στην συνέχεια, καθαρίζονται τα δεδομένα από τις λίστες, διότι σε αυτές αποθηκεύονται τα στοιχεία ανά δέκα ξενοδοχεία σε ξεχωριστό αρχείο για την αποφυγή τυχόν σφαλμάτων, και κατασκευάζεται ένας ακόμα βρόγχος από τις τιμές start και end που ορίστηκαν προηγουμένως για την επίσκεψη με σκοπό την εξαγωγή των στοιχείων από τις σελίδες αυτές μέσω του chrome φυλλομετρητή που χρησιμοποιεί το selenium.

```
try:
    page_list = driver.find_elements('xpath', '//*[@id="review_list_pa
except StaleElementReferenceException:
    try:
        page_list = driver.find_elements('xpath', '//*[@id="review_list
    except StaleElementReferenceException:
        page_list = []
        pass
except NoSuchElementException:
    try:
        page_list = driver.find_elements('xpath', '//*[@id="review_list
    except NoSuchElementException:
        page_list = []
        pass
```

Εικόνα 23: Εύρεση στοιχείων και έλεγχος για τυχόν σφάλματα

Πριν την εκκίνηση του επόμενου βρόγχου για την συλλογή των στοιχείων σε κάθε σελίδα των κριτικών, δημιουργείται μία λίστα με όλα τα στοιχεία του paginator της σελίδας, που είναι υπεύθυνο για την εμφάνιση των σελίδων και την μετακίνηση σε αυτές.



Εικόνα 24: Μπάρα paginator για την αλλαγή σελίδας

Αυτός ο paginator τοποθετείται στο κάτω μέρος της οθόνης του χρήστη και παρουσιάζεται στην παραπάνω φωτογραφία. Επειδή υπάρχει περίπτωση να μην ορίζεται ο paginator και ένα κατάλυμα να έχει μόνο μία σελίδα εισάγεται μέσα σε μπλοκ try except, για να μην διακόπτεται η λειτουργία του προγράμματος αν προκύψει κάποιο σφάλμα. Δοκιμάζεται πρώτα η εντολή για την εύρεση όλων των στοιχείων. Αν προκύψει σφάλμα StaleElementException, το οποίο σημαίνει ότι το αντικείμενο που αναζητείται δεν είναι προσβάσιμο, τότε επαναλαμβάνεται η εύρεση στοιχείων. Το στοιχείο μπορεί να μην είναι προσβάσιμο, διότι δεν έχει φορτωθεί πλήρως το περιεχόμενο και για αυτό επαναλαμβάνεται αίτημα για την εύρεση στοιχείων. Σε περίπτωση που δεν καταφέρει να βρει στοιχεία και εμφανιστεί ξανά το ίδιο λάθος, τότε απλά προχωράει στο επόμενο στάδιο με την εντολή pass, για να κάνει exit από τα μπλοκ try except. Ένα ακόμα σφάλμα που μπορεί να παρουσιαστεί είναι το NoSuchElementException, δηλαδή ότι δεν βρέθηκε κανένα τέτοιο στοιχείο στον ιστότοπο. Σε αυτή την περίπτωση, ξαναδοκιμάζει να εντοπίσει τα στοιχεία και αν δεν εμφανίσει αποτέλεσμα γίνεται pass, δηλαδή παράκαμψη των εντολών αυτών και το πρόγραμμα προχωράει στις επόμενες εντολές. Στις περιπτώσεις που γίνεται pass, καθαρίζονται τα περιεχόμενα της λίστας page_list για να μην επαναληφθούν από το προηγούμενο ξενοδοχείο.

```
if len(page_list) > 1:
    last_page = page_list[len(page_list) - 1].text
    last_page = int(last_page)
else:
    last_page = 1
```

Εικόνα 25: Αν τα στοιχεία στην σελίδα έχουν πάνω από μία σελίδα εξάγεται ο αριθμός της τελευταίας σελίδας

Αν η λίστα του paginator έχει πλήθος στοιχείων μεγαλύτερο από ένα τότε εισάγεται στην μεταβλητή last_page ο αριθμός του προτελευταίου στοιχείου με το string του αριθμού τελευταίας σελίδας, διότι το άκρως τελευταίο στοιχείο είναι το κουμπί της επόμενης σελίδας, που έχει string με τον αριθμό από την τιμή της επόμενης σελίδας. Για παράδειγμα, αν ο χρήστης βρίσκεται στην πρώτη σελίδα, ο αριθμός του τελευταίου στοιχείου paginator θα είναι ίσος με δύο, αν ο χρήστης βρίσκεται στην δεύτερη σελίδα τότε ο αριθμός του τελευταίου στοιχείου θα είναι ίσος με τρία κ.ο.κ. Το string που αντλήθηκε σε αυτό το κομμάτι κατόπιν μετατρέπεται σε αριθμό για τον βρόγχο που θα ακολουθήσει. Σε οποιαδήποτε άλλη περίπτωση, δηλαδή στο σημείο που το πλήθος της λίστας είναι ίσο με ένα ή μηδέν, τότε το last_page θα γίνει 1 για να εφαρμοστεί μόνο μία επανάληψη για τα περιεχόμενα της τωρινής σελίδας.

```

for i in range(0, last_page):
    # Dates
    try:
        datesum = driver.find_elements('xpath', '//*[@id="review_list_pag
    except StaleElementReferenceException:
        try:
            datesum = driver.find_elements('xpath', '//*[@id="review_list
        except StaleElementReferenceException:
            continue
    except NoSuchElementException:
        try:
            datesum = driver.find_elements('xpath', '//*[@id="review_list
        except NoSuchElementException:
            continue

```

Εικόνα 26: Κεντρικός βρόγχος για την άντληση των επιθυμητών στοιχείων

Σε αυτό το στάδιο εφαρμόζεται εκκίνηση του επόμενου βρόγχου και αναζητούνται τα αντικείμενα σε αυτή την μορφή που παρατηρήθηκε προηγουμένως στην λίστα με τα στοιχεία του paginator. Αυτή η διαδικασία επαναλαμβάνεται για το κάθε ένα δεδομένο που πρέπει να εξαχθεί.

```

for country in countrysum:
    countries.append(country.text)

for date in datesum:
    dates.append(date.text.replace('Υποβολή σχολίου: ', ''))
    hotel_name.append(hotel_names[j])
    hotel_link.append(hotel_links[j])

```

Εικόνα 27: Γέμισμα των λιστών με τα δεδομένα

Οι χώρες και όλα τα υπόλοιπα αντικείμενα εισάγονται όπως φαίνεται στην εικόνα, δηλαδή εξάγεται μόνο το κείμενο των στοιχείων html που βρέθηκε προηγουμένως, ενώ για τις ημερομηνίες διαγράφεται το κομμάτι του string "Υποβολή σχολίου: " για να μείνει μόνο το χρήσιμο κομμάτι, της ημερομηνίας. Ταυτόχρονα, προστίθεται το όνομα του ξενοδοχείου και ο σύνδεσμος για να γνωρίζουμε σχετικά με το ποια κατοικία αναφερόμαστε σε κάθε κριτική

```

if len(page_list) > 0:
    try:
        element = driver.find_element('xpath', '//*[@id="review_list"]')
        element.location_once_scrolled_into_view
        element.click()
        driver.switch_to.default_content()
    except StaleElementReferenceException:
        try:
            element = driver.find_element('xpath', '//*[@id="review_list"]')
            element.location_once_scrolled_into_view
            element.click()
            driver.switch_to.default_content()
        except StaleElementReferenceException:
            break
    except NoSuchElementException:
        try:
            element = driver.find_element('xpath', '//*[@id="review_list"]')
            element.location_once_scrolled_into_view
            element.click()
            driver.switch_to.default_content()
        except NoSuchElementException:
            break
else:
    break

```

Εικόνα 28: Μετακίνηση στην επόμενη σελίδα, στην περίπτωση όπου υπάρχει

Αν υπάρχει και άλλη σελίδα στο `page_list`, τότε προσπαθεί να βρεθεί το κουμπί της επόμενης σελίδας για να επιτευχθεί μετάβαση στην επόμενη σελίδα και να αντληθεί το περιεχόμενό της. Για αυτό, πρώτα αναζητείται το αντικείμενο με την εντολή `find_element()`. Στην συνέχεια, κινείται το bot από το πάνω μέχρι το κάτω μέρος της οθόνης στον φυλλομετρητή (όπως κινεί και ένας κανονικός χρήστης με την ροδέλα του ποντικιού του) μέχρι να βρεθεί η τοποθεσία του κουμπιού, όπου και πατιέται. Αν η αναζήτηση αυτή είναι μάταια ή το πλήθος των σελίδων είναι ίσο με μηδέν τότε διακόπτεται η διαδικασία και το πρόγραμμα προχωράει στο επόμενο ξενοδοχείο.

```

        break
    sleep(1.5)
    print('finished {}'.format(j))
    sleep(2)

```

Εικόνα 29: Αναμονή ανά κάθε ξενοδοχείο και σελίδα κριτικής, καθώς και παρουσίαση μηνύματος κατά το τέλος διαβάσματος οποιoδήποτε ξενοδοχείου

Στο κάτω μέρος του βρόγχου ορίζεται ένα διάστημα ενάμιση δευτερολέπτου για να φορτωθεί η επόμενη σελίδα σωστά και ύστερα επαναλαμβάνεται η διαδικασία για κάθε σελίδα που περιέχει κριτικές. Αφού τερματίσει ο βρόγχος αυτός, εμφανίζεται ένα μήνυμα που περιέχει τον αριθμό του καταλύματος που έχει πλέον τελειώσει στο τερματικό του χρήστη. Στο επόμενο βήμα, δίνεται εντολή για μία μικρή αναστολή του προγράμματος των δύο δευτερολέπτων και μετά προχωράει στο επόμενο ξενοδοχείο.

```

countries += ['-'] * (max_length - len(countries))
titles += ['-'] * (max_length - len(titles))
pos_reviews += ['-'] * (max_length - len(pos_reviews))
neg_reviews += ['-'] * (max_length - len(neg_reviews))
reviewscore += ['-'] * (max_length - len(reviewscore))
dates += ['-'] * (max_length - len(dates))
hotel_name += ['-'] * (max_length - len(dates))
hotel_link += ['-'] * (max_length - len(dates))
data = {'Title' : titles,
        'Date' : dates,
        'Country': countries,
        'Positive Review' : pos_reviews,
        'Negative Review' : neg_reviews,
        'Review Score' : reviewscore,
        'Hotel Title' : hotel_name,
        'Hotel Link' : hotel_link}

df = pd.DataFrame.from_dict(data)
#df.to_csv('out_forgotten.csv')
df.to_csv('out_corfu_reviews({},{}).csv'.format(str(start), str(end)), sep=';')
start += 10
end += 10

```

Εικόνα 30: Δημιουργία ίσου πλήθους λιστών και φόρτωση στο αρχείο δεδομένων

Μετά την επανάληψη που περιέχει τις τιμές των μεταβλητών start και end εισάγεται η μεταβλητή max_length που περιέχει το μέγιστο από όλες τις λίστες, που αναζητείται επειδή μπορεί να μην συμπεριλαμβάνεται κάποιο στοιχείο σε ορισμένα σχόλια. Παρόλο που οι λίστες μπορεί να μην έχουν το ίδιο πλήθος αντικειμένων, η δημιουργία του αρχείου csv απαιτεί ρητά την ύπαρξη ίδιου μήκους λιστών. Έτσι, αυξάνεται το μήκος των λιστών αυτών με την διαφορά μεταξύ του μέγιστου μήκους λίστας και του πλήθους της τωρινής λίστας. Στο αριστερό μέρος πολλαπλασιάζεται το αποτέλεσμα με μία λίστα που περιέχει ως μοναδικό στοιχείο τον χαρακτήρα "-", για την εισαγωγή αυτού του χαρακτήρα σε κάθε επιπλέον δεδομένο που προστίθεται στην λίστα. Ουσιαστικά αθροίζεται μία λίστα με μία ακόμα λίστα που έχει μήκος την διαφορά που αναφέρθηκε προηγουμένως και μοναδικό περιεχόμενο ένα πλήθος χαρακτήρων "-". Φυσικά, οι λίστες hotel_name hotel_name συμπληρώνονται με την διαφορά μεταξύ του μέγιστου μήκους και του μήκους της λίστας με τις ημερομηνίες, εφόσον συμπληρώνονται στην ίδια επανάληψη με τις ημερομηνίες, με αποτέλεσμα να έχουν το ίδιο πλήθος στοιχείων. Έπειτα, υλοποιείται ένα λεξικό με τους τίτλους στο αριστερό μέρος στις στήλες και στο δεξί μέρος τα στοιχεία που θα βρίσκονται κάτω από τις συγκεκριμένες στήλες. Για παράδειγμα, το "Title" είναι ένας τίτλος στήλης που προφανώς θα περιέχει τα δεδομένα από την λίστα "titles" με όλους τους τίτλους των σχολίων. Στην επόμενη φάση, κατασκευάζεται DataFrame από λεξικό, διότι αυτή η μορφή δεδομένων της βιβλιοθήκης pandas αποτελεί δομή που μπορεί να επεξεργαστεί πιο εύκολα. Στο επόμενο βήμα δημιουργείται αρχείο με την τιμή start και end στον τίτλο, αυξάνεται το start και end κατά δέκα για να προχωρήσει η διαδικασία στα επόμενα δέκα ξενοδοχεία και το πρόγραμμα επαναλαμβάνει τον βρόγχο για τα επόμενα δέκα ξενοδοχεία. Στο αρχείο αυτό που δημιουργείται, ορίζεται διαχωριστής μεταξύ στοιχείων ως τον χαρακτήρα semicolon (;), για να αποφευχθούν τυχαία σφάλματα με περισσότερα από ένα κόμμα στις γραμμές. Φυσικά, τα προβλήματα αυτά μπορεί να προκύψουν, επειδή στα κείμενα εφαρμόζονται πολλά τέτοια σημεία στίξης, όπως τον χαρακτήρα κόμμα (,).

Tripadvisor Αξιοθέατα

```
# Libraries
import requests
from bs4 import BeautifulSoup
import csv
import re
import pandas as pd
from time import sleep
from requests import exceptions
```

Εικόνα 31: Βιβλιοθήκες για το πρόγραμμα που σχετίζεται με τα αξιοθέατα

Οι βιβλιοθήκες είναι σχεδόν ίδιες με αυτές που χρησιμοποιήθηκαν και στο προηγούμενο αντικείμενο. Όμως, σε αυτό το πρόγραμμα δεν χρησιμοποιήθηκε το αργό selenium, αλλά το ταχύτερο requests. Επίσης έγινε προσθήκη της βιβλιοθήκης re για την χρήση regular expressions και του πακέτου exceptions που βρίσκεται μέσα στην βιβλιοθήκη requests, για την αποφυγή κάποιων σφαλμάτων που μπορεί να προκύψουν κατά την λειτουργία της εφαρμογής. Ακόμα, η ανάγνωση του αρχείου έγινε με τον ίδιο τρόπο, όπως στο προηγούμενο πρόγραμμα. Μολονότι υπήρχε βρόγχος και σε αυτή την εφαρμογή, που παρουσιάζει ομοιότητες με τον προηγούμενο που αναλύθηκε, η υλοποίηση της επανάληψης εφαρμόζεται με διαφορετικό τρόπο.

```
while True:
    for i in range(start, end):
        print(i)
        for j in range(0, 99990, 10):
            print(j)
```

Εικόνα 32: Ορισμός των επαναλήψεων για την εξαγωγή των στοιχείων

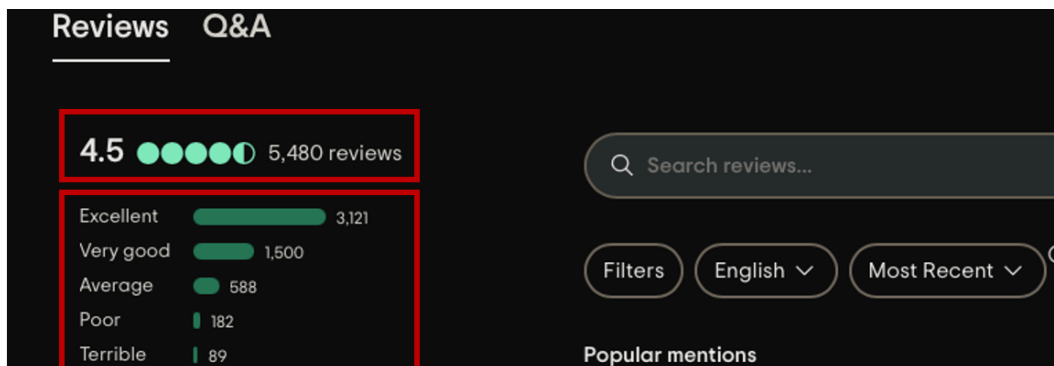
Για την επανάληψη αυτή ορίζεται ατέρμων βρόγχος που τελειώνει όταν δεν έχει απομείνει κάποιο ξενοδοχείο για να προσπελάσει η εφαρμογή. Επιπροσθέτως, το κάτω μέρος παραμένει το ίδιο με την μεταβλητή start και end. Επίσης, ορίζεται μία ακόμα επανάληψη με έναν εξωφρενικό αριθμό που θα διακόπτεται μετά από την ανάγνωση και αποθήκευση κάθε κριτικής σε κάποιο ξενοδοχείο.

```
next_link = tripadvisor_links[i].split('Reviews-')[0] + \
    'or{}-'.format(str(j)) + \
    tripadvisor_links[i].split('Reviews-')[1]
```

Εικόνα 33: Ορισμός επόμενου συνδέσμου που θα επισκεφθεί ο φυλλομετρητής

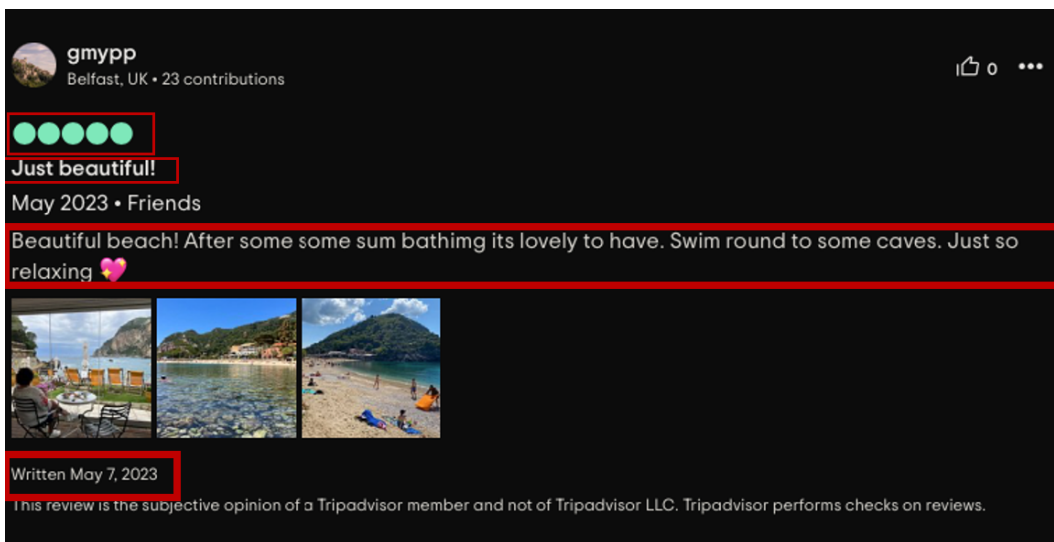
Η κάθε ιστοσελίδα στο TripAdvisor, περιέχει τις κριτικές χωρίς να εμφανίζονται τα δεδομένα αυτά σε διαφορετικό tab όπως στις κριτικές του Booking. Ωστόσο, όταν αλλάξεις σελίδα σε αυτή την ιστοσελίδα συμπεριλαμβάνεται ένα string που περιέχει "or" που το ακολουθεί το πλήθος

των κριτικών σε κάθε σελίδα. Εφόσον η κάθε σελίδα δέχεται μέχρι δέκα κριτικές, ο αριθμός μετά από κάθε string "or" θα είναι το βήμα στην for, που έχει οριστεί πιο πάνω ως την τιμή δέκα.



Εικόνα 34: Παράδειγμα ξενοδοχείου με μέση βαθμολογία και πλήθος κριτικών από το 1 έως το 5

Οι κριτικές σε κάθε κατάλυμα τοποθετούνται στο κάτω μέρος από την ατομική σελίδα κάθε ξενοδοχείου, που συμπεριλαμβάνει τις γενικές πληροφορίες από το ξενοδοχείο αυτό. Όπως παρατηρείται, περιέχει τον μέσο όρο που είναι 4.5 στην προκειμένη περίπτωση και το πλήθος των κριτικών. Από κάτω εμφανίζεται το πλήθος των κριτικών ανάλογα σε ποια κατηγορία ανήκει η κριτική αυτή. Στην συγκεκριμένη περίπτωση οι περισσότερες κριτικές είναι εξαιρετικά θετικές, ενώ σε σύγκριση εμφανίζεται ένα πολύ μικρό πλήθος με εξαιρετικά αρνητικές κριτικές από τους χρήστες. Επίσης, υπάρχει ένα μικρό πλήθος από ασθενές κριτικές και ένα αρκετό, αλλά συγκριτικά σπανιότερο από θετικές κριτικές, πλήθος με μέτριες κριτικές.



Εικόνα 35: Παράδειγμα κριτικής και εμφάνιση δεδομένων που αποθηκεύονται σε αρχείο

Τα δεδομένα που εξάγονται στο συγκεκριμένο πρόγραμμα είναι η τιμή της βαθμολογίας που εμφανίζεται στο πάνω μέρος στις πράσινες φούσκες, ο τίτλος κειμένου του κάθε σχολίου, το κείμενο που θα αναλυθεί για τους στόχους της εργασίας, καθώς και η ημερομηνία υποβολής σχολίου.

```
try:
    page = requests.get(next_link, headers= headers)
except exceptions.ChunkedEncodingError:
    sleep(3)
    page = requests.get(next_link, headers=headers)
```

Εικόνα 36: Επίσκεψη στη σελίδα και έλεγχος για σφάλμα κατά την μετάδοση των πακέτων

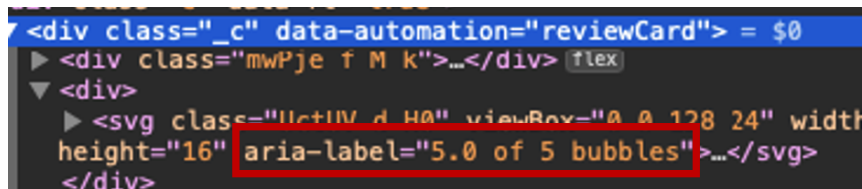
Για την πρόσβαση στα δεδομένα της κάθε ιστοσελίδας, προστίθεται μέσα στον τελευταίο βρόγχο αυτό το μπλόκ εντολών. Αυτές οι εντολές χρειάστηκαν να ενσωματωθούν στο πρόγραμμα, διότι εμφάνιζε σε κάποια διαστήματα ένα σφάλμα "ChunkedEncodingError" που σταματούσε απότομα την λειτουργία της εφαρμογής. Το πρόβλημα αυτό έχει να κάνει με τον απότομο τερματισμό της συνεδρίας (session) μεταξύ της εφαρμογής και της ιστοσελίδας, με αποτέλεσμα την επιστροφή ελλιπών δεδομένων που δεν μπορούν πλέον να ερμηνευτούν από το πρόγραμμα, εφόσον έχουν απώλειες στα στοιχεία τους. Το πρόβλημα με τα ελλιπή δεδομένα συνήθως προκαλείται από την μεριά της ιστοσελίδας και δεν έχει σχέση με την εφαρμογή, αφού η σελίδα σταματάει την ενεργή συνεδρία. Φυσικά, το φαινόμενο αυτό μπορεί να εμφανίζεται και επειδή πραγματοποιούνται πολλαπλά αιτήματα σε κοντινά διαστήματα ή να υπάρχει κάποιο όριο χρόνου στην διάρκεια συνεδρίας. Για την διαχείριση αυτού, εφαρμόστηκε ένα πακέτο από το exceptions που αναλαμβάνει την αντιμετώπιση αυτού του σφάλματος, με μια μικρή αναμονή τριών δευτερολέπτων και επανάληψη του αιτήματος από το σημείο το οποίο σταμάτησε η άντληση των δεδομένων.

```
soup = BeautifulSoup(page.text, 'html.parser')
rating_list = soup.select("div._c div > svg[aria-label]")
for rate in rating_list:
    if 'of 5 bubbles' in rate['aria-label']:
        review_ratings.append(rate['aria-label'].replace('of 5 bubbles', ''))
try:
    reviewdate = soup.find_all('div', class_ = 'TreSq')
except:
    print('error')
    break
```

Εικόνα 37: Εξαγωγή βαθμολογίας

Αφού χρησιμοποιηθεί η βιβλιοθήκη requests για να επιστρέφει τα στοιχεία της σελίδας, γίνεται εφαρμογή του BeautifulSoup για την ανάλυσή τους. Για τις βαθμολογίες που βρίσκονται στην σελίδα, χρησιμοποιείται η εντολή soup.select() που συμπεριλαμβάνεται στα πακέτα του BeautifulSoup. Αυτή η εντολή δέχεται ως είσοδο το css selector από την σελίδα και επιστρέφονται τα στοιχεία. Στην συγκεκριμένη περίπτωση πρέπει να επιστρέφονται τα aria-label από το στοιχείο που ορίζονται οι πράσινες φούσκες στον ιστότοπο. Το attribute aria-label εμπεριέχεται μέσα στο element svg και για αυτό πραγματοποιείται αναζήτηση για αυτό. Όμως, δεν είναι δυνατή η ταυτοποίηση του αντικείμενου δίχως να υπάρχει κάποιο μοναδικό αναγνωριστικό που αφορά αποκλειστικά και μόνο κάθε κριτική στην σελίδα. Δυστυχώς εμφανίζεται και σε άλλα αντικείμενα svg με attribute aria-label οπότε πραγματοποιείται εξέταση στο html για την εύρεση parent elements που έχουν μοναδικά αναγνωριστικά μόνο για τις συγκεκριμένες κριτικές. Μετά από έρευνα παρατηρείται ότι αυτό είναι το div με κλάση "_c", που συναντάται μόνο σε κάθε κριτική της σελίδας. Έτσι, ορίζεται αυτό στην αρχή του css selector, με ένα ακόμα div αντικείμενο, που είναι

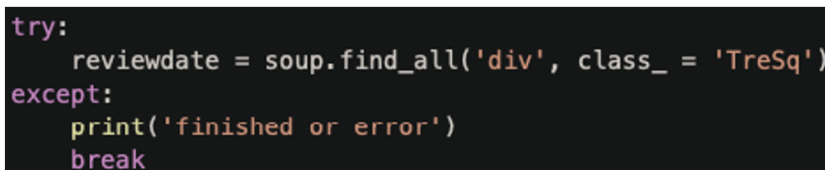
παιδί του πρώτου στοιχείου να ακολουθεί, και τελευταίο να αποτελεί το svg αντικείμενο στον selector που είναι το παιδί του παιδιού του πρώτου αντικειμένου.



```
<div class="_c" data-automation="reviewCard"> = $0
  <div class="mwPje f M k">...</div> flex
  <div>
    <svg class="UctIV d H0" viewBox="0 0 128 24" width=
    height="16" aria-label="5.0 of 5 bubbles">...</svg>
  </div>
```

Εικόνα 38: Σημείο στον πηγαίο κώδικα από όπου εξάγεται η βαθμολογία κάθε ατομικής κριτικής

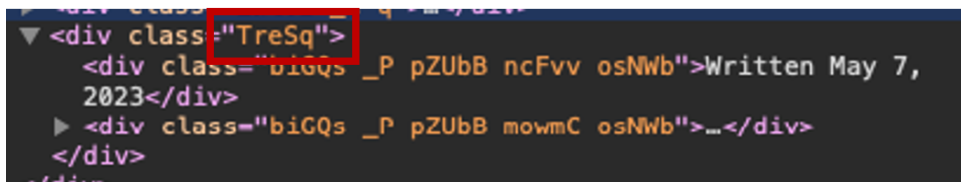
Παραπάνω παρουσιάζεται το html που αφορά τα αντικείμενα αυτά. Στην αρχή υπάρχει το div με κλάση "_c", που περιέχει πολλαπλά div παιδιά, αλλά επειδή εφαρμόζεται αναζήτηση συγκεκριμένα για κάποιο div με παιδί ένα svg element, επιλέγονται μόνο όλα τα svg αντικείμενα αυτά με "aria-label" και αγνοούνται οι υπόλοιπες div.κλάσεις που δεν έχουν ως παιδιά τα svg στοιχεία. Στην συνέχεια, στο πρόγραμμα αναζητείται μόνο το aria-label attribute που έχει την φράση "of 5 bubbles", που σημαίνει ότι αφορά τις πράσινες φούσκες με την βαθμολογική τιμή της κριτικής. Όταν βρεθεί με επιτυχία, τότε αυτό εισάγεται στην λίστα και διαγράφεται το κομμάτι "of 5 bubbles", για να καταχωρηθεί μόνο ο δεκαδικός αριθμός της κριτικής που είναι σε κλίμακα από το 1.0 μέχρι το 5.0.



```
try:
    reviewdate = soup.find_all('div', class_ = 'TreSq')
except:
    print('finished or error')
    break
```

Εικόνα 39: Εξαγωγή ημερομηνίας υποβολής σχολίου

Γνωρίζοντας ότι η ημερομηνία υποβολής υπάρχει σε κάθε σχόλιο, εισάγεται ένα μπλοκ για την εύρεση του στοιχείου αυτού στα σχόλια και σε περίπτωση που δεν βρεθεί τυπώνεται μήνυμα που γράφει ότι είτε τελείωσε η διαδικασία είτε εμφανίστηκε σφάλμα, στην γραμμή εντολών, και το πρόγραμμα προχωράει στην εξαγωγή κριτικών για το επόμενο ξενοδοχείο. Αυτό συμβαίνει γιατί μετά θα απαιτηθεί η εξαγωγή δεδομένων για τα άλλα δεδομένα, κάτι που δεν μπορεί να επιτευχθεί λόγω του ότι τα στοιχεία δεν θα εμφανίζονται στην σελίδα με κριτικές.



```
<div class="TreSq">
  <div class="biGQs _P pZUBB ncFvv osNWb">Written May 7,
  2023</div>
  <div class="biGQs _P pZUBB mowmC osNWb">...</div>
</div>
```

Εικόνα 40: Κλάση div που αναζητείται από την εφαρμογή

Το όνομα της κλάσης για κάθε ημερομηνία υποβολής σχολίου περιέχει ένα στοιχείο div με την παραπάνω κλάση, δηλαδή με το string "TreSq". Για αυτό, δίνεται εντολή για την εύρεση όλων των div που βρίσκεται η κλάση με αυτό το συγκεκριμένο κείμενο.

```
reviewtitle = soup.select('div._c > div.biGQs._P.fiohW.qWPrE.ncFvv.f0tGX > a > span')
reviewtext = soup.select('div._c > div._T.FKffI > div.fIrGe._T.bgMZj > div > span')
```

Εικόνα 41: Εύρεση και αποθήκευση τίτλων σχολίων και κειμένων σχολίων

Το ίδιο div στοιχείο εφαρμόζεται ως αρχή και σε αυτά τα css selectors, μόνο που καταλήγουν σε διαφορετικά στοιχεία, με περιεχόμενο τα επιθυμητά δεδομένα που πρέπει να εξαχθούν για την διπλωματική εργασία.

```
<div class="_c" data-automation="reviewCard">
  <div class="mwPje f M k">...</div> flex
  <div>...</div>
  <div class="biGQs _P fiohW qWPrE ncFvv f0tGX">
    <a target="_blank" tabindex="0" href="/ShowUserReviews-g776017-d523868-r889418498-Paleokastritsa_Beach-Paleokastritsa_Corfu_Ionian_Islands.html" class="BMQDV _FG- wSSLS SwZTJ FGwzt ukgoS"> Event
      <span class="yCeTE">Just beautiful!</span> = $0
    </a>
  </div>
  <div class="RpeCd">May 2023 • Friends</div>
  <div class="_T FKffI">
    <div class="fIrGe _T bgMZj" style="line-break: normal; cursor: auto;"> Event
      <div class="biGQs _P pZUbB KxBGd">
        <span class="yCeTE">
          "Beautiful beach! After some some sum bathing its lovely to have. Swim round to some caves. Just so relaxing 🍷"
        </span>
      </div>
    </div>
  </div>
```

Εικόνα 42: Σημείο στον κώδικα HTML με το κείμενο και τίτλο σχολίου

Ο τίτλος και το κείμενο επισημαίνονται με μπλε χρώμα στην παραπάνω εικόνα. Στην προκειμένη περίπτωση, το περιεχόμενο του κειμένου με τον τίτλο είναι η φράση "Just beautiful!". Αυτό εντοπίζεται μέσα στο span στοιχείο του πηγαίου κώδικα. Όμως, επειδή η κλάση δεν αποτελεί μοναδικό αναγνωριστικό αποκλειστικά για τις κριτικές της σελίδας, αλλά εμφανίζεται και στα κείμενα των κριτικών, εφαρμόζεται πάλι η αρχή ως το div με την κλάση "_c". Στην συνέχεια, εισάγεται η κλάση του div που περιέχει το a στοιχείο, ένα αντικείμενο που αποτελεί γόνο του span στοιχείου, για την εξαγωγή δεδομένων του τίτλου. Η ίδια διαδικασία υλοποιείται και στο κείμενο, με την διαφορά ότι πραγματοποιείται αναφορά στο επόμενο στοιχείο div με παιδί ένα ακόμα div. Ακόμα, με την σειρά του, και το παιδί αυτό έχει ακόμα ένα στοιχείο div ως παιδί που περιέχει τελικά την επιθυμητή span κλάση, με περιεχόμενο το κείμενο, που αναζητείται για τους σκοπούς της εργασίας.

```

for title in reviewtitle:
    reviewtitles.append(title.get_text().replace(';', ''))
for text in reviewtext:
    reviewtexts.append(text.get_text().replace(';', ''))
for date in reviewdate:
    reviewdates.append(date.find('div').get_text().split('Written ')[1])
    attraction_review_titles.append(tripadvisor_titles[i])
    attraction_review_links.append(tripadvisor_links[i])

```

Εικόνα 43: Διαγραφή σημείων όπου εμφανίζεται ο χαρακτήρας semicolon και αποθήκευση λοιπών στοιχείων

Για κάθε τίτλο αντλείται προφανώς το ίδιο το κείμενο, και φροντίζεται να διαγραφούν όλοι οι χαρακτήρες semicolon (;) που θα μπορούσαν να συγκρουστούν με τον διαχωριστή στο αρχείο csv. Η ίδια διαδικασία επαναλαμβάνεται και στο κείμενο που αφορούν οι τίτλοι σχολίων. Για τις ημερομηνίες, αναζητείται η διαδικασία div που αποτελεί παιδί του αντικειμένου που βρέθηκε με επιτυχία στο προηγούμενο στάδιο, και στη συνέχεια αντλείται το κείμενο του div. Φυσικά, κόβεται το κομμάτι με την λέξη "Written" από το κείμενο αυτό, διότι δεν είναι χρήσιμο για την μελέτη που θα πραγματοποιηθεί αργότερα. Ταυτόχρονα, προστίθενται οι τίτλοι και οι σύνδεσμοι από τα καταλύματα σε δύο ξεχωριστές λίστες, όπως και στο προηγούμενο πρόγραμμα.

```

if i == end - 1:
    max_length = max(len(reviewdates), len(reviewtexts), len(reviewtitles))
    reviewtitles += ['-'] * (max_length - len(reviewtitles))
    reviewdates += ['-'] * (max_length - len(reviewdates))
    reviewtexts += ['-'] * (max_length - len(reviewtexts))
    review_ratings += ['-'] * (max_length - len(review_ratings))
    attraction_review_titles += ['-'] * (max_length - len(attraction_review_titles))
    attraction_review_links += ['-'] * (max_length - len(attraction_review_links))
    data = {'Review Title' : reviewtitles, 'Review Date' : reviewdates, 'Review Ratings' : review_ratings, 'Attraction Review Titles' : attraction_review_titles, 'Attraction Review Links' : attraction_review_links}
    df = pd.DataFrame.from_dict(data)
    df.to_csv('sxolia_axiotheata({},{}).csv'.format(str(start), str(end)))
    reviewdates.clear()
    reviewtexts.clear()
    reviewtitles.clear()
    attraction_review_links.clear()
    attraction_review_titles.clear()
    review_ratings.clear()
    start += 10
    end += 10

```

Εικόνα 44: Αποθήκευση σε αρχείο ανά δέκα ξενοδοχεία με την διαδικασία μετατροπής των λιστών στο ίδιο μέγεθος

Το επόμενο στάδιο είναι το ίδιο με αυτό στο προηγούμενο πρόγραμμα, με την διαφορά ότι όλα βρίσκονται υπό συνθήκη που φροντίζει να αποθηκεύει ανά δέκα ξενοδοχεία. Εισάγεται *end-1*, γιατί όταν για παράδειγμα το *end* θα έχει τιμή ίση με δέκα, τότε θα σημαίνει αυτό ότι στο τέλος του ένατου ξενοδοχείου θα αποθηκεύονται οι κριτικές από τα ξενοδοχεία σε αρχείο, μέχρι εκείνο το σημείο, δηλαδή ανά δέκα στοιχεία ακριβώς (0-9). Με την αύξηση του *start* και *end* θα προχωρήσει η διαδικασία στα επόμενα δέκα στοιχεία (10-19), στην ακόλουθη σειρά ξενοδοχείων (20-29) κ.ο.κ.

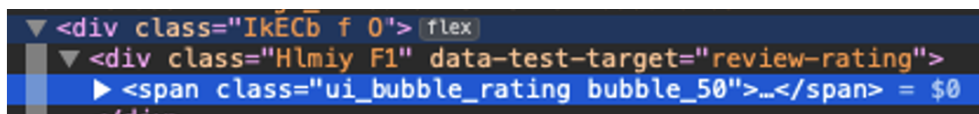
Tripadvisor Ξενοδοχεία

Η δόμηση της εφαρμογής για την εξαγωγή στοιχείων που αφορά τα ξενοδοχεία είναι περίπου η ίδια με αυτή των αξιοθεάτων που αναφέρθηκε προηγουμένως. Ωστόσο, παρουσιάζονται ελάχιστες τροποποιήσεις ή προσθήκες που θα αναλυθούν σε αυτό το κομμάτι.

```
var = soup.select('div.IkECb.f.0 > div > span[class]')
for element in var:
    data = re.sub(r"\D", "", str(element))
    review_rating.append(data)
try:
    reviewdate = soup.find_all('div', class_ = 'IHLTq _Z o')
except:
    print('error')
    break
```

Εικόνα 45: Εύρεση βαθμολογίας στον πηγαίο κώδικα και διαγραφή άχρηστων χαρακτήρων

Στην εφαρμογή υπάρχουν τρεις επαναλήψεις, όπως και στην υλοποίηση για τα αξιοθέατα. Η διαφορά είναι ότι εδώ για την εξαγωγή κριτικής προστίθενται διαφορετικά css selector, όπως αυτό στην λίστα var για την εύρεση κάθε βαθμολογίας κριτικής.



```
<div class="IkECb f 0"> flex
  <div class="Hlmiy F1" data-test-target="review-rating">
    <span class="ui_bubble_rating bubble_50">50</span> = $0
  </div>
```

Εικόνα 46: Σημείο στον πηγαίο κώδικα με την τιμή βαθμολογίας κάθε κριτικής

Στη συνέχεια, εισάγεται επανάληψη για κάθε στοιχείο που βρίσκεται στην λίστα var, και χρησιμοποιείται η βιβλιοθήκη re για την εύρεση regular expressions προτύπων. Στην προκειμένη περίπτωση, εφαρμόζεται η sub από την βιβλιοθήκη re, διότι πρέπει να απομείνει από όλο το κείμενο της κλάσης μόνο ο αριθμός με την τιμή βαθμολογίας, που στην πάνω εικόνα είναι ίσος με 50. Αυτό σημαίνει ότι το υπόλοιπο μέρος, δηλαδή το "ui_bubble_rating bubble_" από το κείμενο πρέπει να διαγραφεί. Έτσι, ορίζεται "\D" στην συνάρτηση για να κρατήσει το πρόγραμμα μόνο τον αριθμό από το κείμενο, αντικαταστήνοντας ότι δεν αποτελεί αριθμό με το άδειο string ". Μετά, αυτός ο αριθμός εισάγεται στην λίστα και η διαδικασία επαναλαμβάνεται για όλο το πλήθος της λίστας var. Πιο κάτω εμφανίζεται το μπλόκ για την εύρεση της ημερομηνίας.

```
if (len(reviewdate) < 5):
    print('finished or error')
    break
for date in reviewdate:
    reviewdates.append(date.find('div', class_ = 'cRVSD').find('span').get_text().split('review ')[1])
    attraction_review_links.append(tripadvisor_links[i])
    attraction_review_titles.append(tripadvisor_titles[i])
for text in reviewtext:
    reviewtexts.append(text.find('span').get_text())
```

Εικόνα 47: Έλεγχος σχετικά με το αν υπάρχουν κριτικές στην τωρινή σελίδα και αποθήκευση λοιπών στοιχείων σε λίστες

Αν τα αποτελέσματα από την εύρεση των ημερομηνιών έχουν μικρότερο πλήθος από πέντε στοιχεία, τότε εμφανίζεται μήνυμα ότι είτε η διαδικασία τερμάτισε επιτυχώς είτε υπήρξε πρόβλημα, και συνεχίζει η επανάληψη για το επόμενο ξενοδοχείο. Στην αρχή, είχε δοκιμαστεί η ίδια συνθήκη με πλήθος ίσο με μηδέν αλλά προέκυψαν κάποιες σελίδες που εμφάνιζαν δύο ή τρία στοιχεία, ενώ δεν περιείχαν στην πραγματικότητα κριτικές, με αποτέλεσμα το πρόγραμμα να συνεχίζει επ' άπειρον. Με την γνώση αυτή, ο αριθμός των στοιχείων τροποποιήθηκε σε πέντε. Έπειτα, για κάθε ατομική ημερομηνία στην λίστα με τις ημερομηνίες, αναζητείται το κείμενο της κλάσης που περιέχει ένα span element και στην συνέχεια χωρίζεται στο σημείο όπου εμφανίζεται το string "review " για να εφαρμοστεί εξαγωγή μόνο της ημερομηνίας και όχι του υπόλοιπου κειμένου. Το υπόλοιπο κείμενο περιέχει το username που δεν απαιτείται πουθενά για τους στόχους της εργασίας. Επιπλέον, για κάθε αριθμό στοιχείων εισάγεται στις λίστες ο τίτλος και ο σύνδεσμος καταλύματος. Ακόμα, για κάθε κριτική βρίσκεται το span στοιχείο και εξάγεται το κείμενο που αφορά την κριτική αυτή, όπου και εισάγεται στην λίστα με τις κριτικές. Η υπόλοιπη διαδικασία είναι ακριβώς η ίδια με το προηγούμενο πρόγραμμα για τα αξιοθέατα, εφόσον αφορά τον ίδιο ιστότοπο.

4.2 Προεπεξεργασία

Το στάδιο προεπεξεργασίας είναι ιδιαίτερα σημαντικό και αφορά την επεξεργασία δεδομένων με σκοπό την μετατροπή τους σε ιδανική μορφή για την ανάλυση από μοντέλα μηχανικής μάθησης. Τα δεδομένα αυτά συνήθως περιέχουν θόρυβο, ελλειπείς τιμές ή μπορεί να είναι σε μορφή που δεν υφίσταται δυνατή η φόρτωση τους σε μοντέλα μηχανικής μάθησης. Για αυτό, υφίσταται κάποια τροποποίηση στα στοιχεία αυτά, για τον καθαρισμό και την προετοιμασία των πληροφοριών με τέτοιο τρόπο, ώστε να γίνει εύκολη η εισαγωγή τους σε μοντέλα μηχανικής μάθησης. Αυτό, ταυτόχρονα κάνει πιο εύκολη την διαδικασία και αυξάνει την αποδοτικότητα και ακρίβεια του μοντέλου μηχανικής μάθησης.

Ο καθαρισμός των δεδομένων είναι ανάμεσα στα βήματα που συμβαίνουν στο στάδιο της προεπεξεργασίας. Ωστόσο, δημοφιλείς λεξικά όπως το VADER προέκυψαν από την εκπαίδευση ωμών κειμένων, από σελίδες όπως κοινωνικά μέσα, κάτι που τα κάνει πιο φιλικά σε καταστάσεις όπου το κείμενο που αντιμετωπίζεται δεν υπέστη κάποια τροποποίηση. Αυτό τα κάνει από τις ειδικές περιπτώσεις που το στάδιο της προεπεξεργασίας μπορεί να βλάψει την απόδοση του μοντέλου, αντί να το επηρεάσει θετικά. Παρόλα αυτά, το στάδιο του καθαρισμού συνήθως βελτιώνει την απόδοση. Τα στάδια του καθαρισμού είναι:

- Μετατροπή γραμμάτων από κεφαλαία σε πεζά
- Αφαίρεση αριθμών
- Αφαίρεση γραμμάτων που δεν ανήκουν στην αγγλική γλώσσα
- Διαγραφή σημείων στίξης (punctuation)
- Αποκοπή λευκών κενών (white spaces)
- Tokenize κειμένου
- Εξάλειψη μη αλφαβητικών tokens
- Διαγραφή stopwords

- Επιδιόρθωση συναιρέσεων (contractions)
- Λημματοποίηση
- Τελική μετατροπή tokenized κειμένου στην πλήρης μορφή

Αρχικά υφίσταται μετατροπή γραμμάτων από κεφαλαία σε πεζά γράμματα, κάτι που θα συνεισφέρει στην μείωση του πλήθους των χαρακτηριστικών. Στη συνέχεια, αφαιρούνται αριθμοί και γράμματα που δεν ανήκουν στην αγγλική γλώσσα, διότι μας ενδιαφέρουν συγκεκριμένα οι αγγλικές λέξεις, εφόσον πραγματοποιείται ανάλυση συναισθήματος και τα κείμενα των κριτικών είναι κυρίως στην αγγλική γλώσσα. Ακόμα, διαγράφονται τα σημεία στίξης, σε ορισμένες δοκιμές όπως στην μηχανική μάθηση. Αυτό συμβαίνει γιατί τα λεξικά συνήθως επωφελούνται από την παρουσία σημείων στίξης. Τα λευκά κενά επίσης σβήνονται, διότι δεν χρησιμεύουν. Γίνεται tokenize κειμένου, που κρατάει ατομικές λέξεις και γίνεται εξάλειψη των tokens που δεν είναι αλφαριθμητικοί, όπως είναι τα σύμβολα. Επιπλέον, διαγράφονται οι άχρηστες λέξεις, που είναι γνωστές ως stop words. Τέτοιες λέξεις περιλαμβάνουν τα άρθρα. Ακολουθεί η επιδιόρθωση των συναιρέσεων, που έχει να κάνει με την παρουσία "κοφτών" λέξεων όπως είναι το "don't" και την επιμήκυνσή του σε "do not". Επόμενη είναι η λημματοποίηση που συνεισφέρει συνήθως σε ταξινόμηση που μας ενδιαφέρει το περιεχόμενο. Η ανάλυση συναισθήματος ασχολείται με το περιεχόμενο, οπότε γίνεται χρήση της διαδικασίας αυτής. Λημματοποίηση είναι η επιλογή του λήμματος (ρίζας) από μία λέξη, δηλαδή η αναγωγή στον πρώτο κλιτικό τύπο. Τα λήμματα προκύπτουν με την χρήση μορφολογικού λεξικού. Για παράδειγμα, η λέξη "running" θα γινόταν "run". Παρομοίως, η λέξη "caring" θα επέστρεφε ως λήμμα το "care" και όχι το "car", διότι διαφέρουν στην σημασία τους. Τέλος, μετατρέπονται από ατομικές λέξεις πίσω σε πλήρη μορφή.

Labels/Κατηγορίες

Τα Labels προκύπτουν στο σύνολο δεδομένων από την τιμή της βαθμολογίας που έχει προσδιορίσει ο χρήστης με την υποβολή της κριτικής του. Η βαθμολογία αυτή είναι από το ένα μέχρι το πέντε στον ιστότοπο του Tripadvisor και από το ένα μέχρι το δέκα στην ιστοσελίδα του Booking. Τα τελικά labels είναι τρία ανάλογα με το συναισθηματικό ύψος που παρουσιάζεται στο κείμενο. Συγκεκριμένα, τα labels μπορεί να είναι "negative", "neutral" και "positive". Φυσικά, το "negative" αντιστοιχεί σε αρνητική κριτική, το "neutral" σε ουδέτερη και το "positive" σε θετική.

Η επιλογή του διαχωρισμού των δεδομένων σε τρεις βασικές κατηγορίες υφίσταται, διότι τα στοιχεία δεν ήταν πολλά και δεν μπορούσε η ταξινόμηση να γίνει σωστά. Τα labels αναλύονται ως εξής:

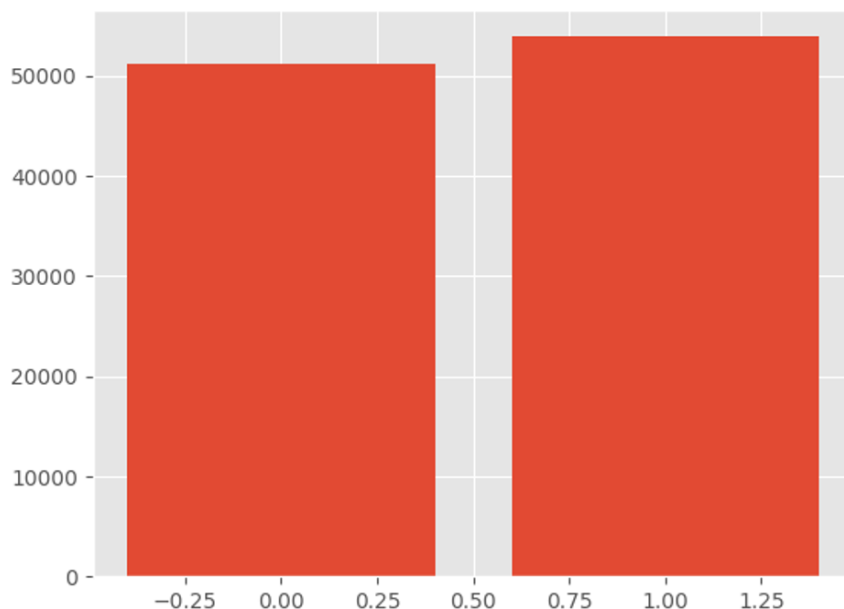
- **Tripadvisor (1-5):** 1-2 αρνητική, 3 ουδέτερη και 4-5 θετική
- **Booking (1-10):** 1-3 αρνητική, 4-7 ουδέτερη και 8-10 θετική.

4.3 Ανάλυση Δεδομένων

Η ανάλυση των δεδομένων έχει να κάνει με την εξέταση των συνόλων δεδομένων για την παρουσίαση σημαντικών ευρημάτων, πριν ξεκινήσει το στάδιο της ταξινόμησης. Αυτό περιλαμβάνει διάφορα διαγράμματα και wordclouds.

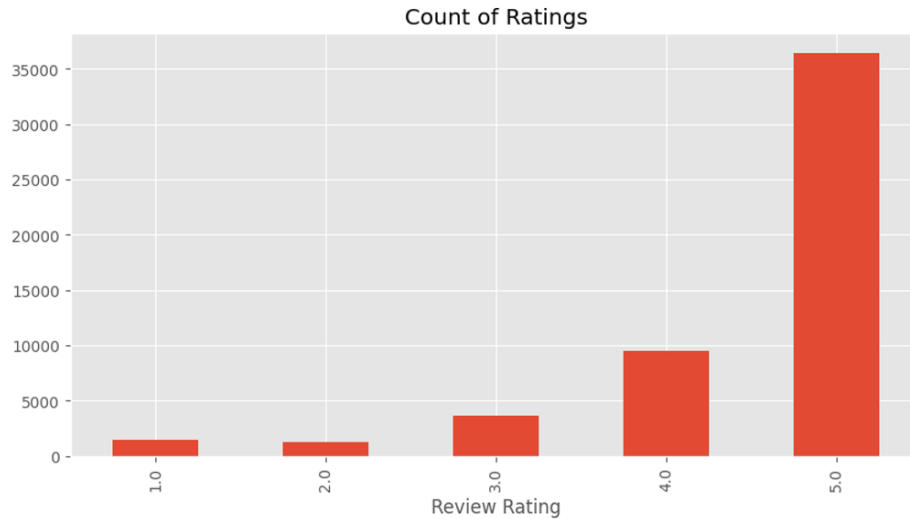
Tripadvisor Αξιοθέατα

Το αρχείο που αντλήθηκε μετά την εξαγωγή των στοιχείων από τα αξιοθέατα του νομού Κέρκυρας, περιέχει 53.879 σειρές. Όμως, ανάμεσα σε αυτά τα στοιχεία παρουσιάζονται και γραμμές με κενά πεδία κειμένου. Αυτό συμβαίνει διότι ένας χρήστης μπορεί να μην έχει υποβάλλει κάποια προαιρετικά πεδία, όπως είναι το κείμενο, όταν υπέβαλλε την κριτική του. Ωστόσο, το αντικείμενο που μελετάται στην εργασία αυτή είναι η ανάλυση συναισθημάτων, κάτι που απαιτεί την παρουσία κειμένων για την ανάλυση και ταξινόμηση των κειμένων, σχετικά με το αν έχουν θετικό, αρνητικό ή ουδέτερο ύφος. Για αυτό, ελέγχεται αρχικά το πλήθος κειμένων και τιμών βαθμολογιών για να ερευνηθεί αν το πλήθος τους είναι το ίδιο, στις κριτικές που αποθηκεύτηκαν στο αρχείο.



Διάγραμμα 5: Πλήθος κειμένων κριτικών (στα αριστερά) και βαθμολογίας κριτικών (στα δεξιά)

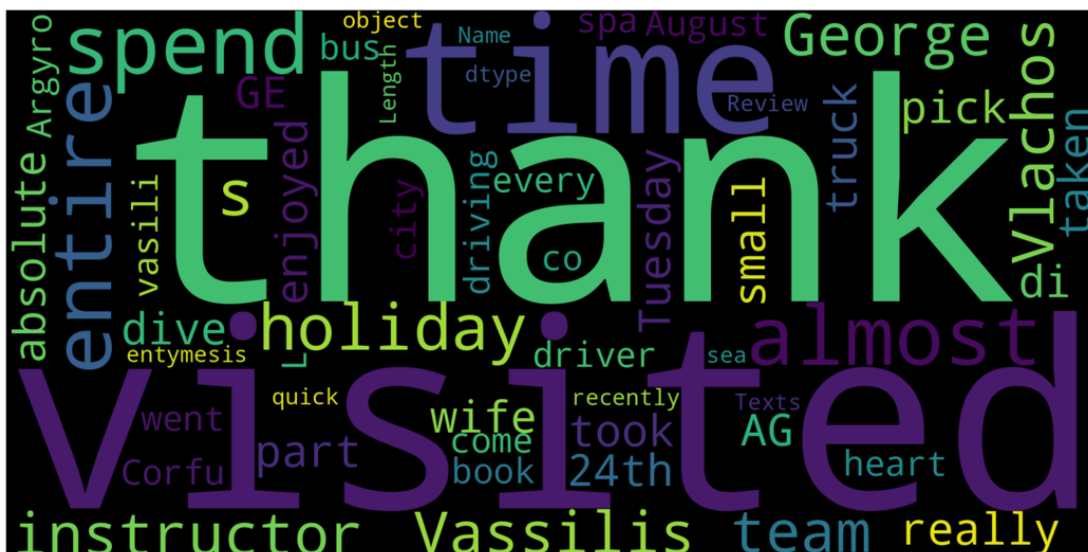
Αριστερά στην εικόνα επιδεικνύεται το πλήθος των κειμένων κριτικών και στα δεξιά παρουσιάζεται το πλήθος των τιμών βαθμολογιών. Είναι φανερό ότι το πλήθος των κειμένων διαφέρει σε έναν μικρό βαθμό από το πλήθος των βαθμολογιών, με τις γραμμές που περιλαμβάνουν κείμενα να είναι λιγότερες. Παρόλα αυτά, δεν απαιτείται κάποια ιδιαίτερη τροποποίηση, διότι κάθε κείμενο έχει την αντίστοιχη τιμή βαθμολογίας του. Έτσι, απλά θα απαιτηθεί η διαγραφή των γραμμών με βαθμολογίες που δεν περιέχουν το αντίστοιχο κείμενο. Με αυτό τον τρόπο όταν εφαρμοστεί, σε επόμενο βήμα, η τεχνική ανάλυσης συναισθημάτων, υπάρχει η τιμή βαθμολογίας που θα εξετάσει σε τι βαθμό ο αλγόριθμος μηχανικής μάθησης ταξινόμισε επιτυχώς τα στοιχεία του αρχείου. Ουσιαστικά, θα εξετάσει πόσο αποτελεσματικός ήταν ο αντίστοιχος αλγόριθμος μηχανικής μάθησης.



Διάγραμμα 6: Ποσότητα βαθμολογιών ανά τιμή βαθμολογίας

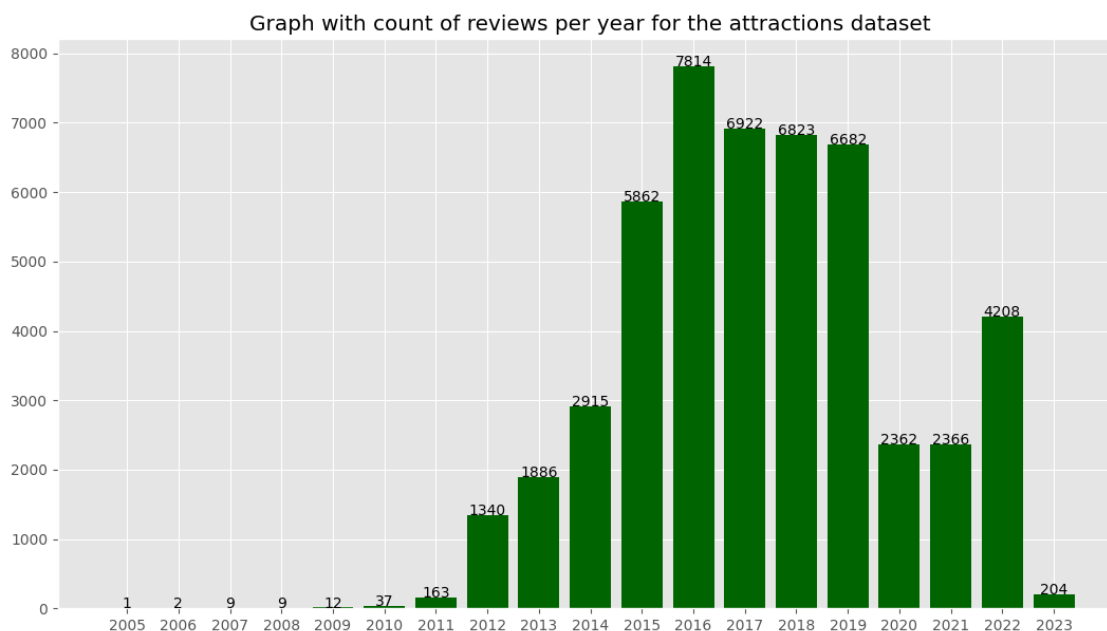
Αφού ερευνήθηκε το πλήθος των τιμών βαθμολογιών και κειμένων, είναι πλέον χρήσιμο να αναλυθούν τα δεδομένα, σε έναν μεγαλύτερο βαθμό. Για παράδειγμα, στις βαθμολογίες μπορεί να ελεγχθεί πόσο είναι το πλήθος τιμών σε κάθε τιμή βαθμολογίας, που κυμαίνεται από το 1 μέχρι το 5, και με αυτό τον τρόπο να ερευνηθεί το πόσο θετικά ή αρνητικά είναι τα δεδομένα που θα υφίσταται αργότερα η τεχνική ανάλυσης συναισθημάτων. Στην παραπάνω εικόνα εμφανίζεται το πλήθος των βαθμολογιών κριτικών σε κάθε τιμή του αρχείου. Στην προκειμένη περίπτωση οι διαθέσιμες τιμές είναι ένα πλήθος πέντε στοιχείων, από τον αριθμό 1 μέχρι τον αριθμό 5. Όπως φαίνεται, υπάρχει ένας εξαιρετικά μεγάλος αριθμός κριτικών που έχει ύψιστες θετικές τιμές, όσον αφορά τα αξιοθέατα του νομού Κέρκυρας, δηλαδή 5 αστέρια. Αυτός ο αριθμός είναι γύρω στις 37.500 τιμές. Επιπλέον, υπάρχει μία ακόμα δημοφιλή τιμή των 4 αστεριών, με ένα πλήθος γύρω στα 9000 στοιχεία. Μετά, εμφανίζεται η τιμή 3 που ξεπερνάει τις πιο χαμηλές τιμές, σε ποσοστό, με ένα πλήθος 3500 στοιχείων. Τέλος, υπάρχει ένα μικρό μεν, αλλά σημαντικό δε, πλήθος στοιχείων με τιμή βαθμολογίας το 1 και το 2, όπου ξεπερνάει το ποσοστό κριτικών με 1 αστέρι τις κριτικές που έχουν τιμή ίση με 2. Το πλήθος κριτικών με 1 αστέρι είναι γύρω στα 2000 στοιχεία, ενώ το πλήθος κριτικών με 2 αστέρια είναι γύρω στα 1800 δεδομένα.

Ένα wordcloud είναι ένα σύννεφο λέξεων που εμφανίζει τις πιο δημοφιλείς λέξεις, με τις πιο συχνές λέξεις να εμφανίζονται με μεγαλύτερη γραμματοσειρά. Παρακάτω θα αναλυθεί το wordcloud για αυτό το σύνολο δεδομένων.



Διάγραμμα 7: WordCloud Tripadvisor Αξιοθέατων

Όπως είναι δυνατόν να παρατηρηθεί, δίχως την επιπλέον προεπεξεργασία, οι δημοφιλέστερες λέξεις είναι τα "thank", "time", "visited" και "entire". Δεν φαίνεται η παρουσία ιδιαίτερων λέξεων που θα βοηθήσουν το σύστημα να βγάλει ένα συμπέρασμα σχετικά με το συναισθηματικό ύφος του κειμένου. Ειδικά με την εμφάνιση ονομάτων όπως είναι τα "George", "Vlachos" και "Vassilis", κάτι που μπορεί να οδηγήσει το σύστημα να βγάλει λανθασμένο συμπέρασμα. Ωστόσο, υπάρχουν κάποιες λέξεις όπως το "thank" και το "enjoyed" που μπορούν να βοηθήσουν στην αναγνώριση θετικού συναισθήματος. Φυσικά, τέτοιες χρήσιμες λέξεις είναι σε πολύ μικρή συχνότητα, με μόνο δύο να εμφανίζονται στο wordcloud.



Διάγραμμα 8: Γράφημα με σύνολα από κάθε χρονιά στο σετ δεδομένων των αξιοθέατων

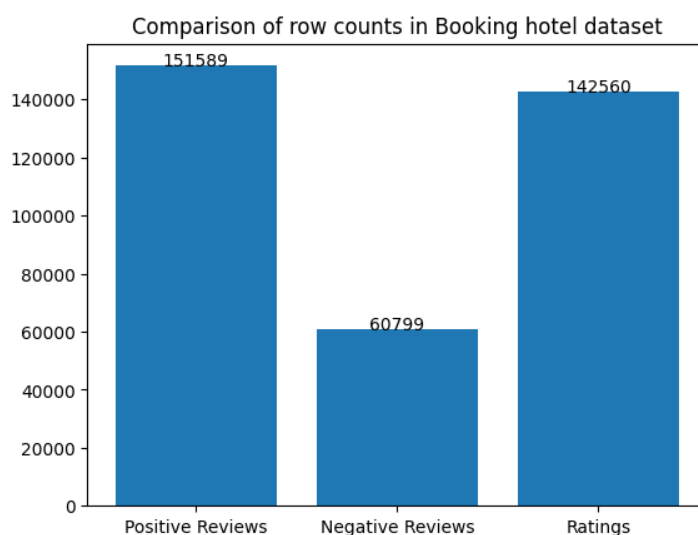
Σύνολικό πλήθος στοιχείων: 49.617 στοιχεία
Τελευταία κριτική: Μάιος 2023

Στο διάγραμμα που αφορά τα αξιοθέατα, τα περισσότερα παραδείγματα προήλθαν από τα δεδομένα που αφορούν την χρονιά του 2016. Συγκεκριμένα, το πλήθος των στοιχείων για αυτή την χρονιά ήταν 7.814. Τα στοιχεία στις επόμενες χρονιές με τα μεγαλύτερα πλήθη ήταν οι ακόλουθες χρονιές, δηλαδή το 2017, 2018 και 2019. Φαίνεται πως δεν υπήρχαν πολλές κριτικές που αφορούν τα αξιοθέατα στις χρονιές 2020 και 2021. Ωστόσο, υπήρχε μια μικρή ανάκαμψη το 2022 με 4.208 κριτικές. Μετά είναι οι χρονιές του 2015, 2014, 2021, 2020, 2013 και 2012, με την σειρά από το μεγαλύτερο πλήθος παραδειγμάτων έως το μικρότερο. Τα υπόλοιπα στοιχεία είναι μικρά σε αριθμό, με το επόμενο να έχει μόνο 204 στοιχεία το 2023. Φυσικά, αυτός ο αριθμός είναι τόσο μικρός επειδή δεν αντλήθηκαν δεδομένα από τους μήνες του Ιουνίου, Ιουλίου και Σεπτεμβρίου, τους μήνες όπου παρουσιάζονται τα μεγαλύτερα κύματα τουριστών. Επίσης, μπορεί να κατανοηθεί ότι το Tripadvisor έχει λίγα δεδομένα για τις χρονιές πριν το 2011, πιθανότητα διότι δεν ήταν τόσο δημοφιλής ο ιστότοπος εκείνο το διάστημα ή επειδή η ιστοσελίδα διαγράφει τα παλιότερα σχόλια.

Κατά την περαιτέρω ανάλυση του γραφήματος, ανιχνεύεται ότι υπήρχε μικρή πτώση του τουρισμού τις χρονιές 2017-2019, με το 2016. Αυτή η πτώση αυξήθηκε σε μεγάλο βαθμό το 2020-2021, προφανώς επειδή προέκυψε ο κορωνοϊός. Ωστόσο, υπήρχε μια σημαντική αύξηση στο πλήθος κριτικών το 2022, με 4.208 κριτικές να αντιστοιχούν στην χρονιά αυτή. Δεν μπορεί να γίνει σύγκριση με την φετινή χρονιά, διότι δεν υπάρχουν αρκετά στοιχεία.

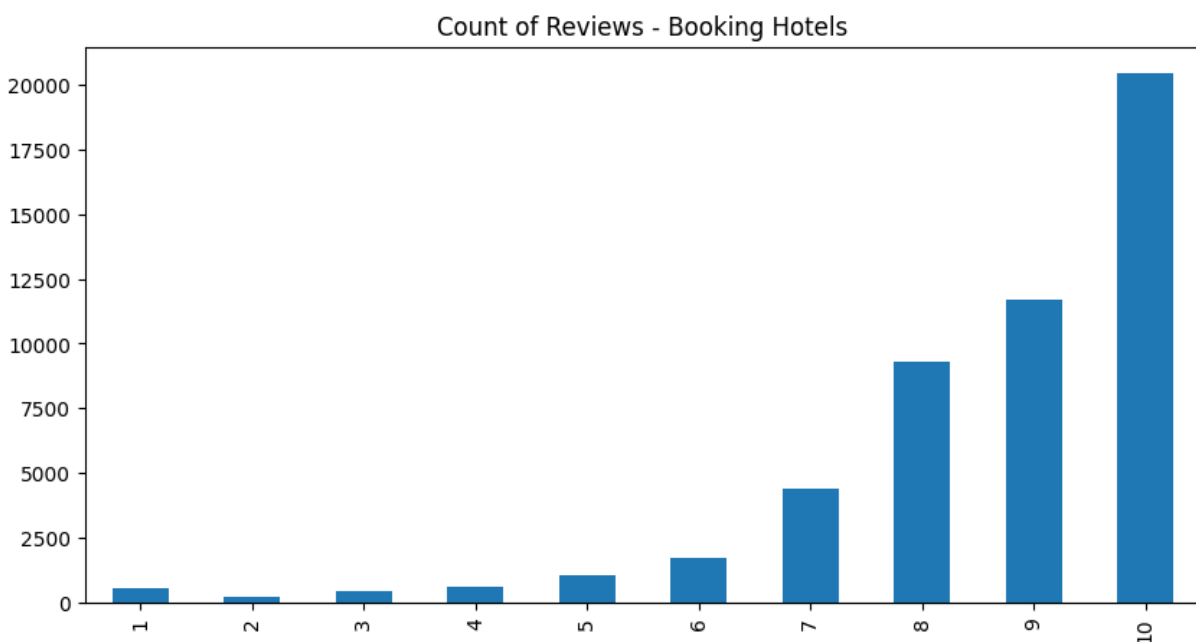
Booking Ξενοδοχεία

Στο Booking τα κείμενα των κριτικών βρίσκονται σε ξεχωριστά τμήματα, ανάλογα με το αν είναι θετικά ή αρνητικά. Συγκεκριμένα, στην ιστοσελίδα εμφανίζονται τα θετικά στοιχεία δίπλα από ένα σύμβολο πρόσθεσης (+), ενώ τα αρνητικά στοιχεία παρουσιάζονται με ένα σύμβολο αφαίρεσης (-).



Διάγραμμα 9: Πλήθος θετικών κριτικών (στα αριστερά), αρνητικών κριτικών (κεντρικά) και βαθμολογίας κριτικών (στα δεξιά)

Όπως μπορεί να παρατηρηθεί, τα περισσότερα στοιχεία αφορούν θετικές κριτικές, αν και υπάρχουν και κάποια επιπλέον αρνητικά στοιχεία. Φυσικά, τα αρνητικά στοιχεία μπορούν να βρεθούν στις κριτικές σε συνδυασμό με τα θετικά στοιχεία, κάτι που σημαίνει ότι δεν μπορεί να βγει συμπέρασμα μόνο από αυτή την εικόνα για τις κριτικές. Παρακάτω, όταν αναλυθούν όλες οι τιμές βαθμολογίας αναλυτικά θα μπορεί να ερευνηθεί το πόσο αρνητικά ή θετικά είναι τα στοιχεία στο σύνολο δεδομένων.

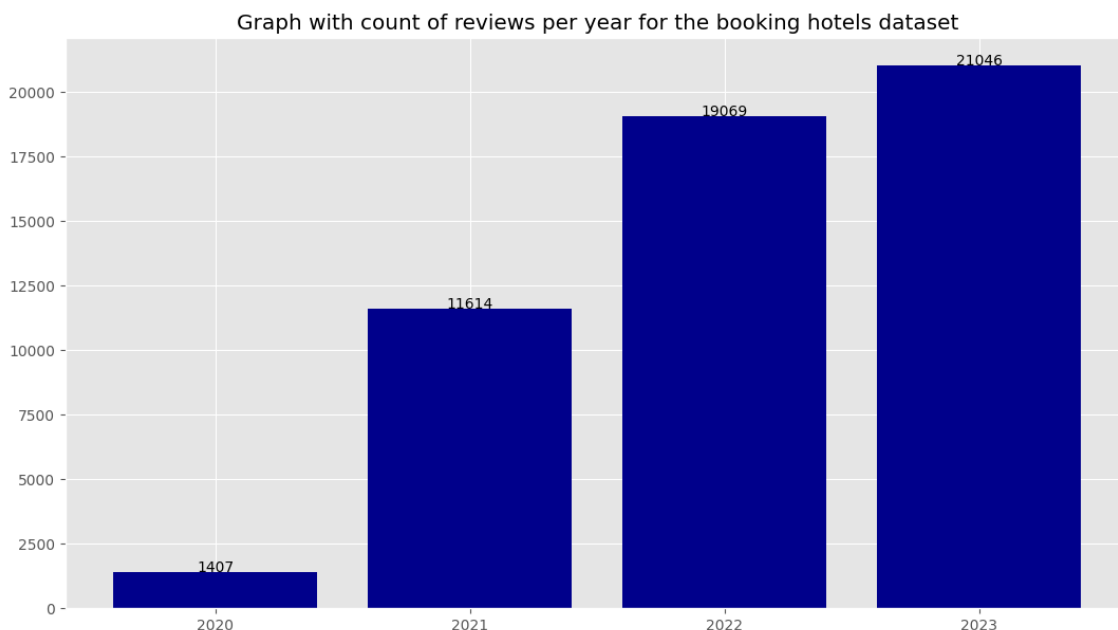


Διάγραμμα 10: Πλήθος βαθμολογιών ανά τιμή βαθμολογίας σε κλίμακα 1-10

Από το πλήθος ανά τιμή βαθμολογίας παρουσιάζεται ότι υπάρχουν γενικότερα υψηλόβαθμες κριτικές στο σύνολο δεδομένων με τα περισσότερα στοιχεία να αφορούν την ύψιστη τιμή βαθμολογίας ίσης με 10. Κάθε μικρότερη τιμή βαθμολογίας (1-9) περιέχει όλο και λιγότερα στοιχεία, με την μόνη τιμή που δεν παρουσιάζεται αυτό το φαινόμενο να είναι η τιμή βαθμολογίας ίση με 2. Παραδόξως, στην τιμή βαθμολογίας ίση με 2 τα στοιχεία είναι τα λιγότερα, με περισσότερα στοιχεία να εμφανίζονται στην προηγούμενη βαθμολογία ίση με 1. Αυτό το φαινόμενο μπορεί να παρατηρηθεί και στα τρία σύνολα δεδομένων, κάτι που δείχνει ότι οι τουρίστες έχουν κάποιους δισταγμούς στην καταχώρηση τιμή βαθμολογίας ίσης με 1. Προφανώς, αυτό έχει να κάνει με το γεγονός ότι η τιμή βαθμολογίας ίση με 1 αποτελεί την χειρότερη βαθμολογία, και λίγα καταλύματα και αξιοθέατα αξίζουν αυτή την βαθμολογία, ακόμα και αν οι τουρίστες έχουν μεγάλες απαιτήσεις.

Υπάρχει μεγάλη διαφορά στο πλήθος ανάμεσα στις τιμές βαθμολογίας. Το πλήθος στην τιμή βαθμολογίας ίσης με 10 είναι περίπου 20000 παραδείγματα. Η αμέσως επόμενη υψηλότερη τιμή, δηλαδή η τιμή ίση με 9, περιλαμβάνει περίπου 11000 παραδείγματα. Ακολουθούν οι τιμές ίσες με 8 και 7, που περιέχουν περίπου 10000 και 5000 παραδείγματα αντίστοιχα. Η τιμή ίση με 6 έχει περίπου 2000 παραδείγματα, με την τιμή 5 να ακολουθεί με περίπου 1000 δείγματα. Τα υπόλοιπα έχουν κάτω από 1000 δείγματα το καθ' ένα. Γενικότερα, ο μεγαλύτερος αριθμός δειγμάτων ανήκει στην κλίμακα 8-10, που έχουμε ορίσει ως την θετική κλάση.

Τα wordcloud με τις δημοφιλέστερες λέξεις μπορούν να παρατηρηθούν παρακάτω.



Διάγραμμα 13: Γράφημα με σύνολα από κάθε χρονιά στο σετ δεδομένων των ξενοδοχείων του booking

Σύνολικό πλήθος στοιχείων: 53.136 στοιχεία

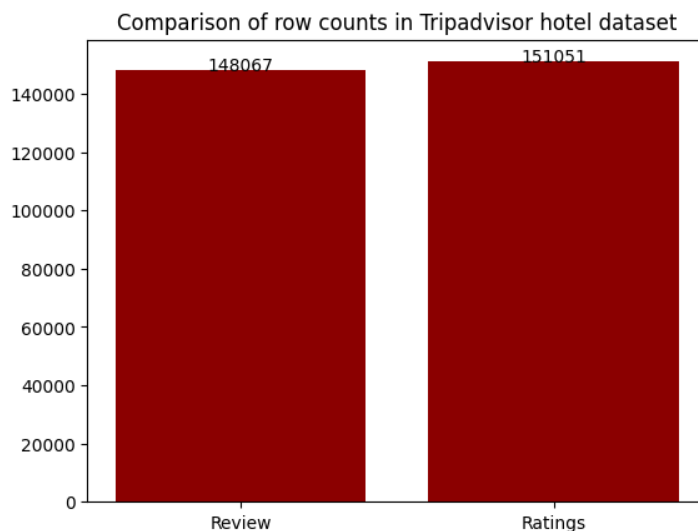
Τελευταία κριτική: Σεπτέμβριος 2023

Παρατηρώντας την εικόνα 13, παρουσιάζεται η υψηλότερα συχνότητα δειγμάτων στην χρονιά του 2023 με 21.046 δείγματα. Η συχνότητα είναι μεγάλη, συγκριτικά με τα αποτελέσματα από το Tripadvisor, διότι τα δεδομένα αντλήθηκαν μετά την περίοδο του καλοκαιριού, όπου οι τουρίστες υποβάλλουν περισσότερες κριτικές. Φυσικά, το προηγούμενο σετ δεδομένων είχε να κάνει με αξιοθέατα, ενώ αυτό με ξενοδοχεία. Στα ξενοδοχεία είναι περισσότερες οι κριτικές από ότι στα αξιοθέατα, όπως μπορεί να παρατηρηθεί και από το μέγεθος του σετ δεδομένων που αφορά τα αξιοθέατα, με εκείνο να είναι το μικρότερο.

Τα στοιχεία στο παραπάνω διάγραμμα πηγαίνουν με ανάποδη σειρά, όσον αφορά την έρευνα για το μεγαλύτερο πλήθος παρατηρήσεων. Με λίγα λόγια, το 2023 προηγείται με το 2022 να είναι η ακόλουθη χρονιά, με τον μεγαλύτερο αριθμό. Στη συνέχεια, είναι το 2021, η επόμενη πολυπληθέστερη χρονιά. Τέλος, είναι η χρονιά του 2020 με το μικρότερο μέγεθος. Φαίνεται καθαρά ότι το 2023 παρουσιάστηκαν περισσότεροι τουρίστες, από ότι στις προηγούμενες χρονιές. Ωστόσο, είναι σημαντικό το γεγονός ότι τα δεδομένα στο σετ δεδομένων προέρχονται μόνο από τις τελευταίες 4 χρονιές, κάτι που οφείλεται στο ότι ο ιστότοπος δεν επιτρέπει την συλλογή κριτικών από ανενεργά ξενοδοχεία ή επειδή διαγράφονται τα παλιά δεδομένα. Παρόλα αυτά, μπορεί να προσδιοριστεί ότι τις προηγούμενες χρονιές, πιο κοντά στην περίοδο του κορονοϊού, υπήρχε μικρότερη συχνότητα τουρισμού, με μία σταδιακή αύξηση στα κύματα τουριστών τα τελευταία χρόνια.

Tripadvisor Ξενοδοχεία

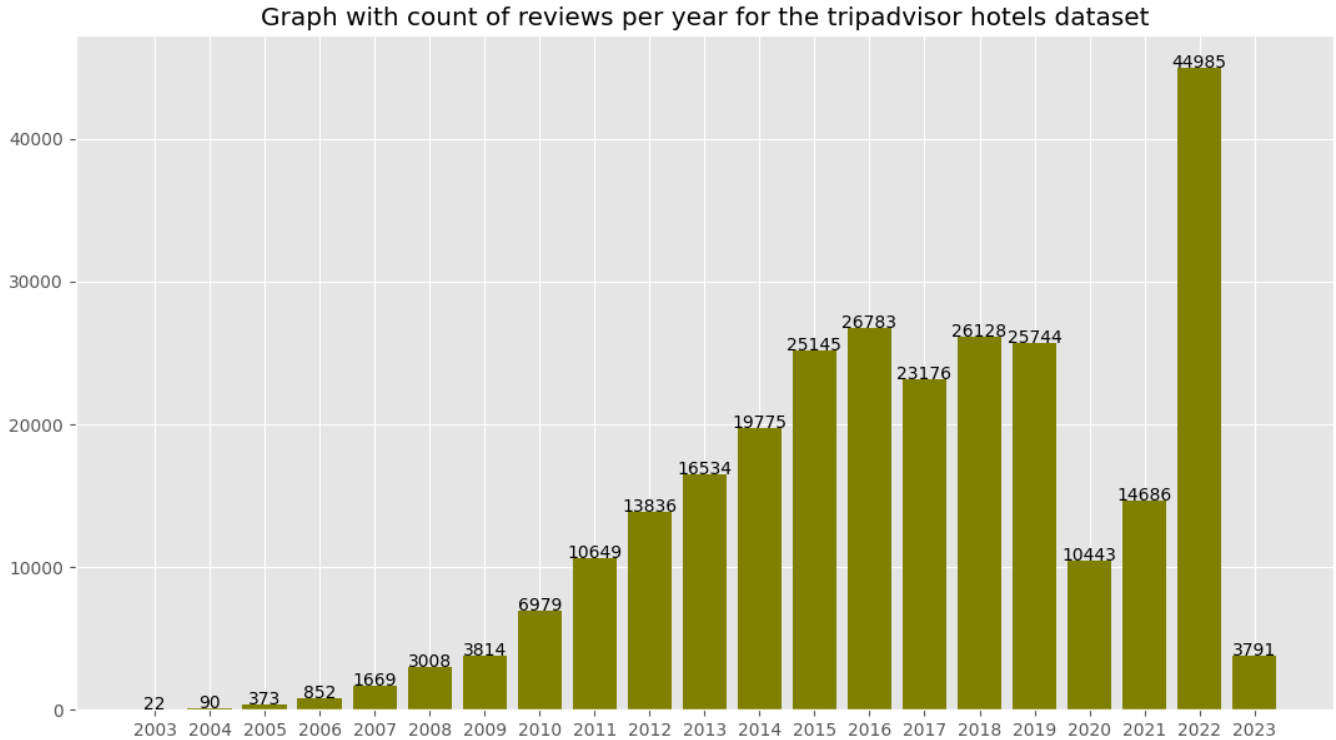
Στα Tripadvisor Ξενοδοχεία υπάρχουν περισσότερα δείγματα από ότι στο Booking. Παρακάτω, παρατηρείται το αποτέλεσμα μετά την σύγκριση της στήλης με τα κείμενα και της στήλης με τις βαθμολογίες, από το σύνολο δεδομένων.



Διάγραμμα 14: Πλήθος κείμενων κριτικών (στα αριστερά) και βαθμολογιών (στα δεξιά)

Όπως φαίνεται δεν υπάρχει μεγάλη διαφορά μεταξύ της στήλης με τα κείμενα από τις κριτικές και της στήλης με τις βαθμολογίες. Εμείς θα κρατήσουμε τις 148067 κριτικές από όλο το σετ, εφόσον τα κείμενα χωρίς την αντίστοιχη τιμή βαθμολογίας δεν μπορούν να χρησιμοποιηθούν για την εκπαίδευση του μοντέλου.

Αν ελεγχθεί αναλυτικά το διάγραμμα με τα πλήθη σε κάθε τιμή βαθμολογίας, μπορεί να ανιχνευτεί ότι ο διαμοιρασμός των στοιχείων ακολουθεί περίπου την ίδια διαδικασία με τα προηγούμενα διαγράμματα. Αυτό σημαίνει ότι τα περισσότερα στοιχεία αντιστοιχούν στην τιμή βαθμολογίας ίσης με 5 (50), ενώ όλο και πιο λίγα στοιχεία έχουν οι τιμές κάτω από αυτήν την τιμή. Ωστόσο, όπως και στα προηγούμενα διαγράμματα, τα στοιχεία από την τιμή βαθμολογίας ίσης με 2 (20) είναι περισσότερα από αυτά στην τιμή βαθμολογίας ίση με 1.



Διάγραμμα 17: Γράφημα με σύνολα από κάθε χρονιά στο σετ δεδομένων των ξενοδοχείων του tripadvisor

Σύνολικό πλήθος στοιχείων: 278.482 στοιχεία
Τελευταία κριτική: Μάιος 2023

Στην εικόνα 17 παρουσιάζεται το γράφημα που αφορά τα ξενοδοχεία του tripadvisor, και συγκεκριμένα τα σύνολα κριτικών ανά χρονιά. Οι κριτικές σε αυτό το γράφημα ξεκινάνε από το 2003. Φαίνεται ότι υπάρχει μία σταδιακή άνοδος στην κατανομή των παραδειγμάτων, με τις εξαιρέσεις να είναι οι χρονιές 2017, 2019, 2020 και 2023. Γενικότερα, η πτώση στα πλήθη είναι μικρή, με τις μόνες σημαντικές σε μέγεθος να παρατηρούνται το 2020 και το 2023. Όμως, το 2020 ήταν εντός της περιόδου του κορονοϊού που φυσικά δεν υπήρχε τόσος τουρισμός. Επίσης, το 2023 δεν συλλέχθηκαν δεδομένα από την περίοδο που έρχεται το μεγαλύτερο κύμα τουρισμού στην χώρα. Παρόλα αυτά, μπορεί να ανιχνευθεί ότι μετά το 2020 υπήρξε μία μικρή αύξηση το 2021 και μία τεράστια το 2022. Προφανώς, το 2022 ήταν η χρονιά που πέρασε το κύμα του κορονοϊού και ο κόσμος πλέον ήταν ελεύθερος να κάνει ταξίδια, είτε για την αναψυχή τους είτε για την εργασία τους. Ωστόσο, το 2022 παρατηρήθηκε τόσο μεγάλη αύξηση στον τουρισμό με 44.985 κριτικές να έχουν υποβληθεί, που ξεπέρασε με διαφορά τις προηγούμενες χρονιές.

4.4 Ταξινόμηση Συναισθήματος

Αρχικά, χωρίζεται το σύνολο δεδομένων σε δύο υποσύνολα, ένα σύνολο εκπαίδευσης (train set) και ένα σύνολο δοκιμής (test set). Η διαδικασία αυτή εφαρμόζεται, διότι το σύστημα πρέπει να φορτωθεί με πολλά στοιχεία για την βέλτιστη απόδοση του μοντέλου και στην συνέχεια πρέπει να δοκιμαστεί το πόσο αποτελεσματικό είναι με στοιχεία που δεν έχει ξαναδεί. Έτσι θα μπορεί να ερευνηθεί το πόσο αποτελεσματικό είναι το μοντέλο αυτό. Στην περίπτωση που δεν πραγματοποιούνταν αυτή η διαδικασία, το μοντέλο φυσικά θα επέστρεφε τέλεια αποτελέσματα αλλά δεν θα ήταν σε θέση να ταξινομήσει με επιτυχία στοιχεία που δεν έχει αντιμετωπίσει στο παρελθόν.

Μετά από τον διαχωρισμό σε δεδομένα εκπαίδευσης και δοκιμής, πραγματοποιείται διαχωρισμός, για άλλη μια φορά, του συνόλου εκπαίδευσης (train set) σε δεδομένα εκπαίδευσης (train set) και δεδομένα επαλήθευσης (validation set). Τα δεδομένα επαλήθευσης είναι σημαντικά και δημιουργούνται για να μπορεί να επαληθευτεί η απόδοση του μοντέλου στο στάδιο της εκπαίδευσης. Με αυτό τον τρόπο, μπορούμε να βρούμε ποιες είναι οι καλύτερες ρυθμίσεις στο μοντέλο, πριν προχωρήσουμε στο τελικό στάδιο της δοκιμής με το σύνολο δεδομένων δοκιμής.

Ο διαχωρισμός των δεδομένων σε train, test και validation σετ έγινε ως εξής:

- **Train set:** 80%
- **Test set:** 20%
- **Validation set:** 20%

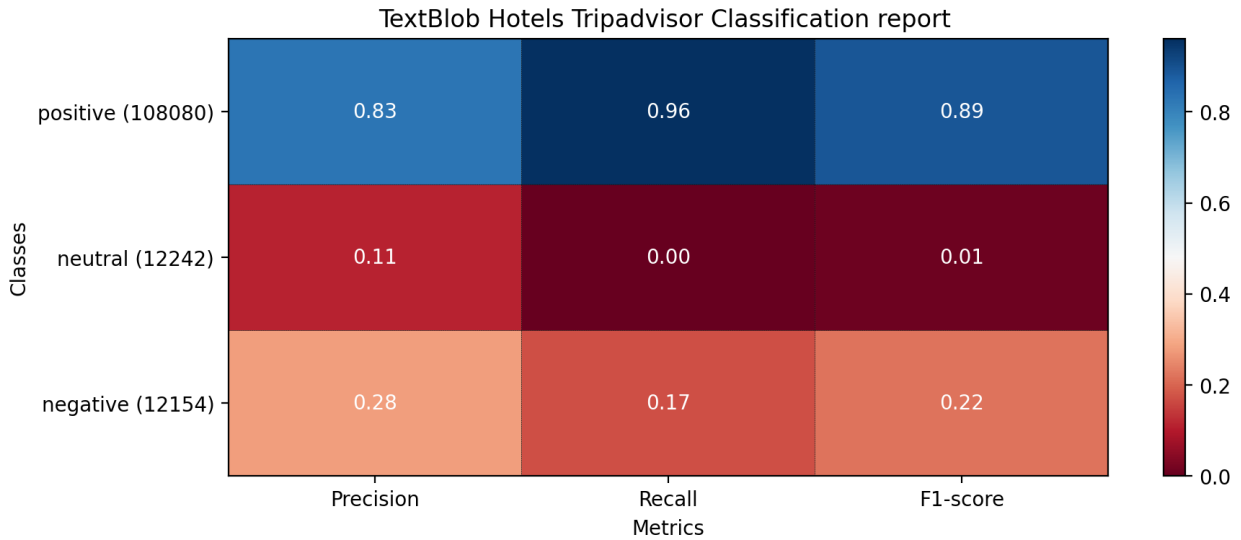
Στη συνέχεια, τα δεδομένα χρησιμοποιούνται είτε με έναν CountVectorizer, για την εφαρμογή Bag-Of-Words μεθόδου, είτε με έναν TfidfVectorizer, για την εφαρμογή της μεθόδου TF-IDF. Τα σετ δεδομένων είναι αρκετά μεγάλα και για αυτό επιλέχθηκε ως ο μέγιστος αριθμός χαρακτηριστικών ένα πλήθος 100.000 χαρακτηριστικών.

Κάποιες από τις ειδικές οδηγίες στους παραπάνω Vectorizer είναι η χρήση μόνο λέξεων στα n-grams, το μέγιστο πλήθος χαρακτηριστικών να είναι ίσο με 100.000, η χρήση κανονικών εκφράσεων για να επιλεχθούν μόνο λέξεις και αριθμοί, καθώς και η επιλογή n-grams, δηλαδή bigrams (2-gram) και trigrams (3-gram). Τα n-grams είναι σειρές λέξεων και χρησιμοποιούνται για να αυξήσουν την απόδοση του ταξινομητή με την προσφορά περισσότερης πληροφορίας. Με την χρήση πάνω από μίας λέξης, γίνεται δυνατή η χρήση ολόκληρων φράσεων που είναι πιο χρήσιμα από ατομικές λέξεις, σε ορισμένες περιπτώσεις. Για παράδειγμα, ας μελετήσουμε την φράση "break a leg". Αν εξετάσεις μόνο το "break" δεν μπορείς να καταλάβεις το νόημα, ενώ αν χρησιμοποιήσεις τις δύο λέξεις που το ακολουθούν και γίνει "break a leg" μπορείς να καταλάβεις ότι είναι μία δημοφιλής φράση που χρησιμοποιείται για να ευχηθείς καλή τύχη σε κάποιον στην αγγλική γλώσσα. Στο countvectorizer χρησιμοποιείται επίσης $df = 0.9$, για να επιλεχθούν μόνο οι σχετικές λέξεις, αντί για το max_features.

4.5 Τελική τιμή συναισθήματος

Η τελική τιμή συναισθήματος που προκύπτει από κάθε δοκιμή εμφανίζεται σε μορφή πίνακα με στοιχεία όπως το accuracy, f1-score, precision, recall και support. Τέλος, παρουσιάζεται το macro average που έχει επιλεχθεί ως το μέτρο σύγκρισης μεταξύ των δοκιμών, διότι θεωρεί ότι όλες οι κλάσεις είναι ισάξιες.

Ένα παράδειγμα από την αναπαράσταση στοιχείων εμφανίζεται παρακάτω:



Διάγραμμα 18: Αναπαράσταση στοιχείων

Η παραπάνω εικόνα αποτελεί ένα παράδειγμα της εξόδου ενός μοντέλου, μετά την δοκιμή. Μπορούν να παρατηρηθούν οι στήλες που αφορούν το Precision, Recall και το F1-Score. Επίσης, το support παρουσιάζεται στις παρενθέσεις, δίπλα από τις κλάσεις. Τέλος, το weighted average στην εικόνα παρατηρείτε στην δεξιά κάθετη στήλη. Φυσικά, στα πειράματα επιλέχθηκε να γίνει χρήση του Macro Average, οπότε αντικαταστάται αυτό το αποτέλεσμα με το Macro Average.

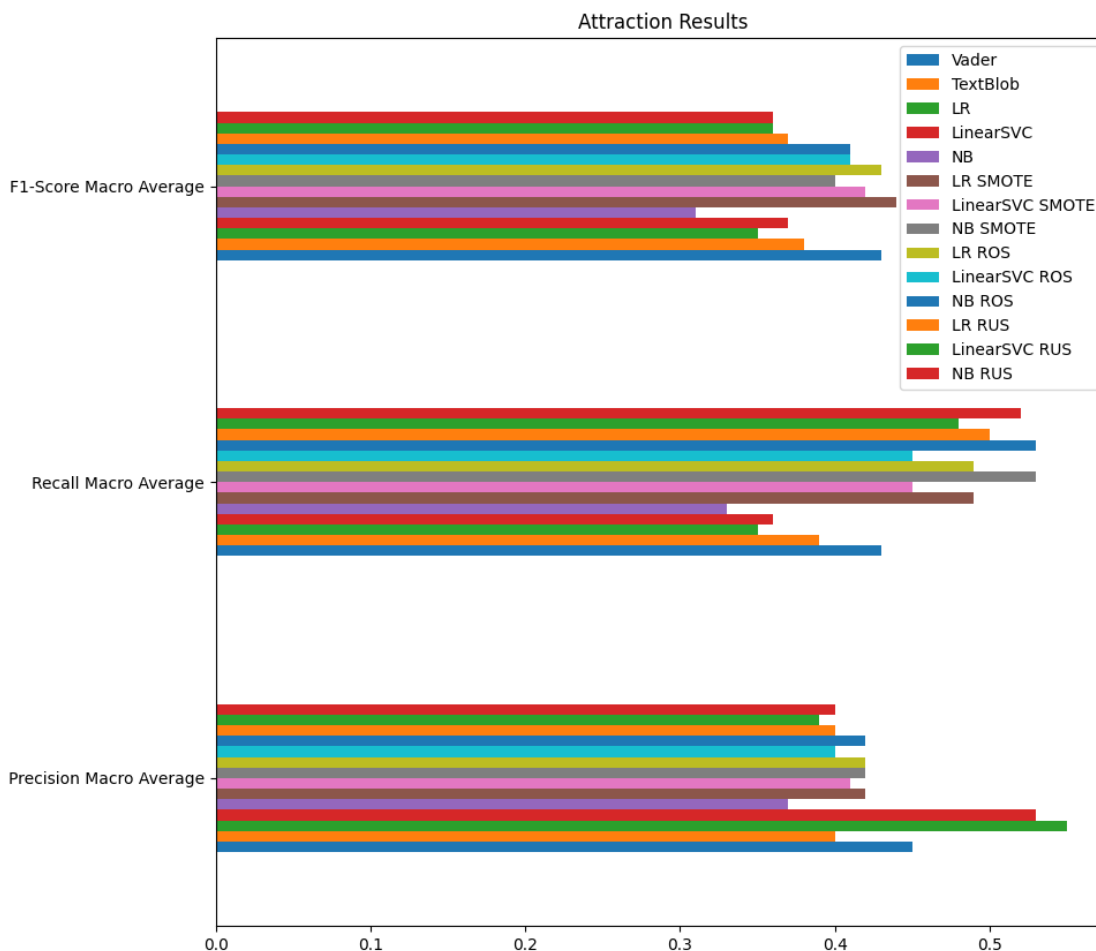
5 Αποτελέσματα

Σε αυτή την ενότητα θα αναλυθούν τα αποτελέσματα από τις έρευνες που διεκπεραιώθηκαν. Τα αποτελέσματα αυτά θα αφορούν τα στοιχεία που προέρχονται από τα ξενοδοχεία και τα αξιοθέατα. Συγκρίθηκαν διάφορα μοντέλα με μέτρο σύγκρισης τις τιμές του Macro Average, μετρικές που μας παρέχουν τον μέσο όρο των πεδίων Precision, Recall και F1-Score για όλα τα σύνολα δεδομένων. Η μετρική αυτή επιλέχθηκε, διότι θεωρεί ότι όλες οι κλάσεις είναι ισάξιες, ανεξαρτήτως από το πλήθος των δειγμάτων που κατατάσσονται σε κάθε κλάση. Ο τύπος του Macro Average είναι ένας πολύ απλός και ευνόητος. Συγκεκριμένα ο τύπος είναι:

$$F1 - Score = \frac{\sum_{i=1}^n}{n}$$

Ο παραπάνω τύπος είναι για τον Macro Average του F1-Score, αλλά λειτουργεί αντίστοιχα και για τα Precision και Recall. Στον αριθμητή ορίζεται το άθροισμά των τιμών (Precision, Recall, F1-Score) των κλάσεων, ενώ στον παρονομαστή εισάγεται το πλήθος των κλάσεων. Για παράδειγμα, αν έχουμε τρεις κλάσεις (0, 1, 2) και επιθυμούμε να ανιχνεύσουμε τον Macro Average του F1-Score, τότε ο τύπος θα αλλάξει και θα γίνει $F1 - Score = \frac{F1_1 + F1_2 + F1_3}{3}$.

5.1 Tripadvisor Αξιοθέατα



Διάγραμμα 19: Αποτελέσματα στο σύνολο δεδομένων των αξιοθεάτων με διάφορες τεχνικές sampling

Στο διάγραμμα εμφανίζονται τα αποτελέσματα για το σύνολο δεδομένων που αφορά τα αξιοθέατα. Όπως μπορεί να παρατηρηθεί, οι μεγαλύτερες τιμές παρατηρήθηκαν στο Precision Macro Average, με μερικές τιμές να αγγίζουν τα 0.55. Επίσης, το Recall Macro Average περιείχε κάποιες τιμές που ξεπερνάνε τα 0.5. Συγκριτικά, οι τιμές στα F1-Score είναι χαμηλές, με τα macro average για το Precision και το Recall να τις ξεπερνάνε. Αυτό προκύπτει επειδή δεν ξεπερνάνε τα 0.5,

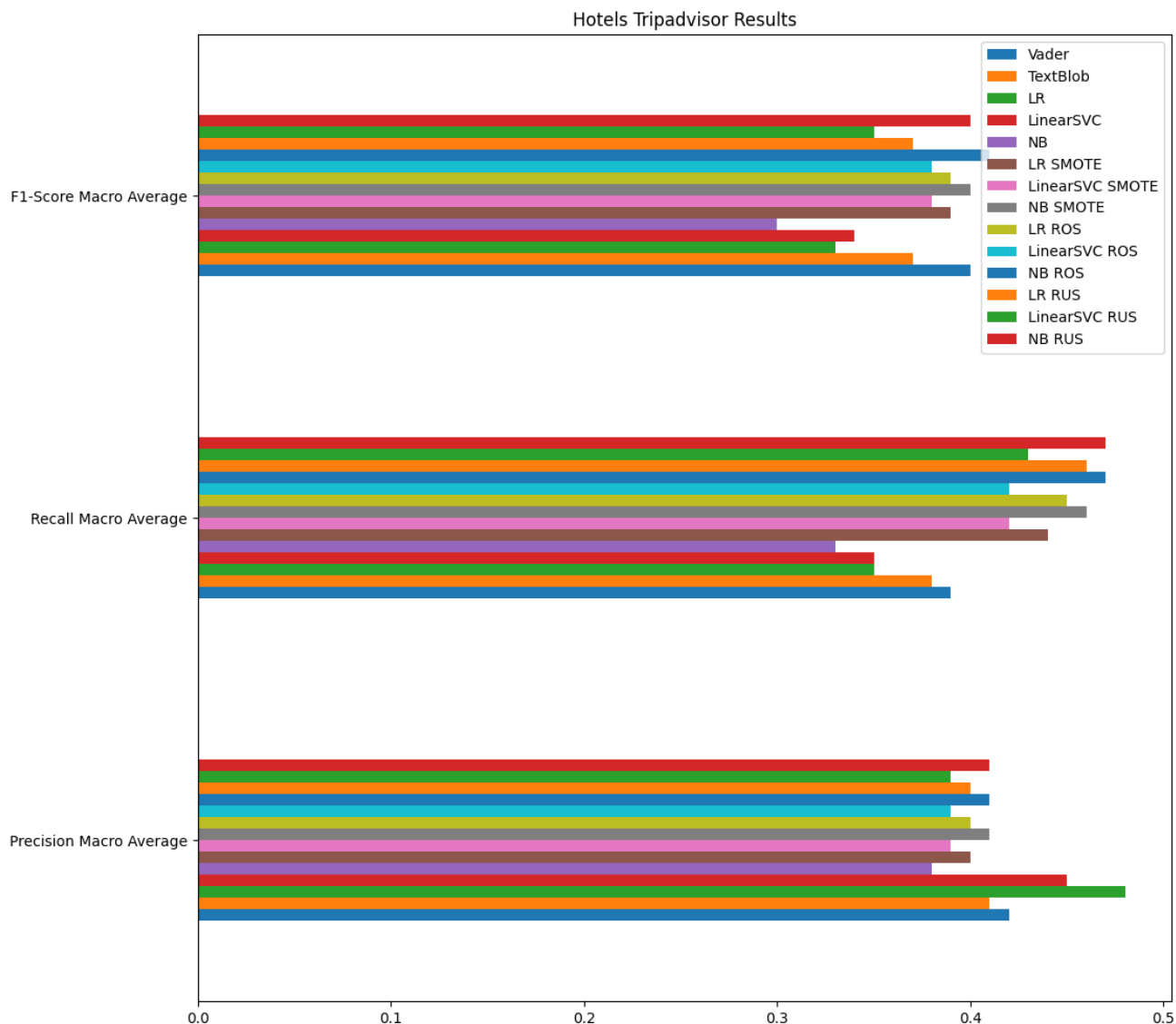
Από το διάγραμμα φαίνεται ότι έχουμε κάποιες με σχετικά υψηλό Precision με χαμηλό Recall, και το αντίθετο. Παρατηρούνται δηλαδή περιπτώσεις όπου οι δοκιμές επέστρεψαν υψηλό Recall με χαμηλό Precision, και το αντίθετο. Όταν συναντάται υψηλό Recall με χαμηλό Precision, αυτό δείχνει ότι το σύστημα επιστρέφει πολλά αποτελέσματα, αλλά τα περισσότερα από τα labels που πρόβλεψε είναι λανθασμένα συγκρίνοντας τα με τα labels από το σετ εκμάθησης. Αντιθέτως, όταν έχουμε υψηλό Precision με χαμηλό Recall, το σύστημα επιστρέφει λίγα αποτελέσματα, αλλά τα περισσότερα από αυτά που πρόβλεψε είναι σωστά συγκρίνοντας τα με τα labels από το σετ εκμάθησης. Εμείς θέλουμε μία ισορροπία μεταξύ των δύο.

Μελετώντας το διάγραμμα, η υψηλότερη τιμή στο Precision επιστρέφεται από τον LinearSVC RUS. Το μοντέλο αυτό εφαρμόζει την μέθοδο Sampling Random Undersampling, που διαγράφει τυχαία παραδείγματα από τις πολυπληθέστερες κλάσεις μέχρι οι κλάσεις να φτάσουν να είναι περίπου το ίδιο μέγεθος με την κλάση μειονότητας. Μετά την εφαρμογή της μεθόδου, εκτελείται η δοκιμή με το μοντέλο LinearSVC. Το αμέσως επόμενο καλύτερο αποτέλεσμα επιστρέφεται από τον NB RUS, που εφαρμόζει την ίδια μέθοδο και μετά δοκιμή με τον Naive Bayes ταξινομητή. Αυτά τα δύο ήταν τα μοναδικά που ξεπέρασαν τα 0.5 στο Precision.

Όσον αφορά τα Recall, το LinearSVC χωρίς την μέθοδο Sampling φαίνεται να επέστρεψε υψηλή τιμή, συγκριτικά με τα υπόλοιπα, με μοναδικά που το ξεπερνάνε να είναι τα NB SMOTE και Vader. Το NB SMOTE αφορά τον ταξινομητή Naive Bayes με εφαρμοσμένη την μέθοδο SMOTE, μία μέθοδος που δημιουργεί συνθετικά παραδείγματα για την εξισορρόπηση. Το Vader αφορά την χρήση λεξικού για την ταξινόμηση συναισθήματος και επέστρεψε τα καλύτερα αποτελέσματα στο Recall.

Τέλος, στο F1-Score, που αποτελεί τον αρμονικό μέσο μεταξύ του Precision και Recall, το LR SMOTE μοντέλο φαίνεται να επιστρέφει την υψηλότερη τιμή. Τα επόμενα καλύτερα μοντέλα είναι τα LR ROS και το NB ROS, με την μέθοδο Random Oversampling να υφίσταται σε LR και NB μοντέλα .

5.2 Tripadvisor Ξενοδοχεία



Διάγραμμα 20: Αποτελέσματα στο σύνολο δεδομένων των ξενοδοχείων από το Tripadvisor με διάφορες τεχνικές sampling

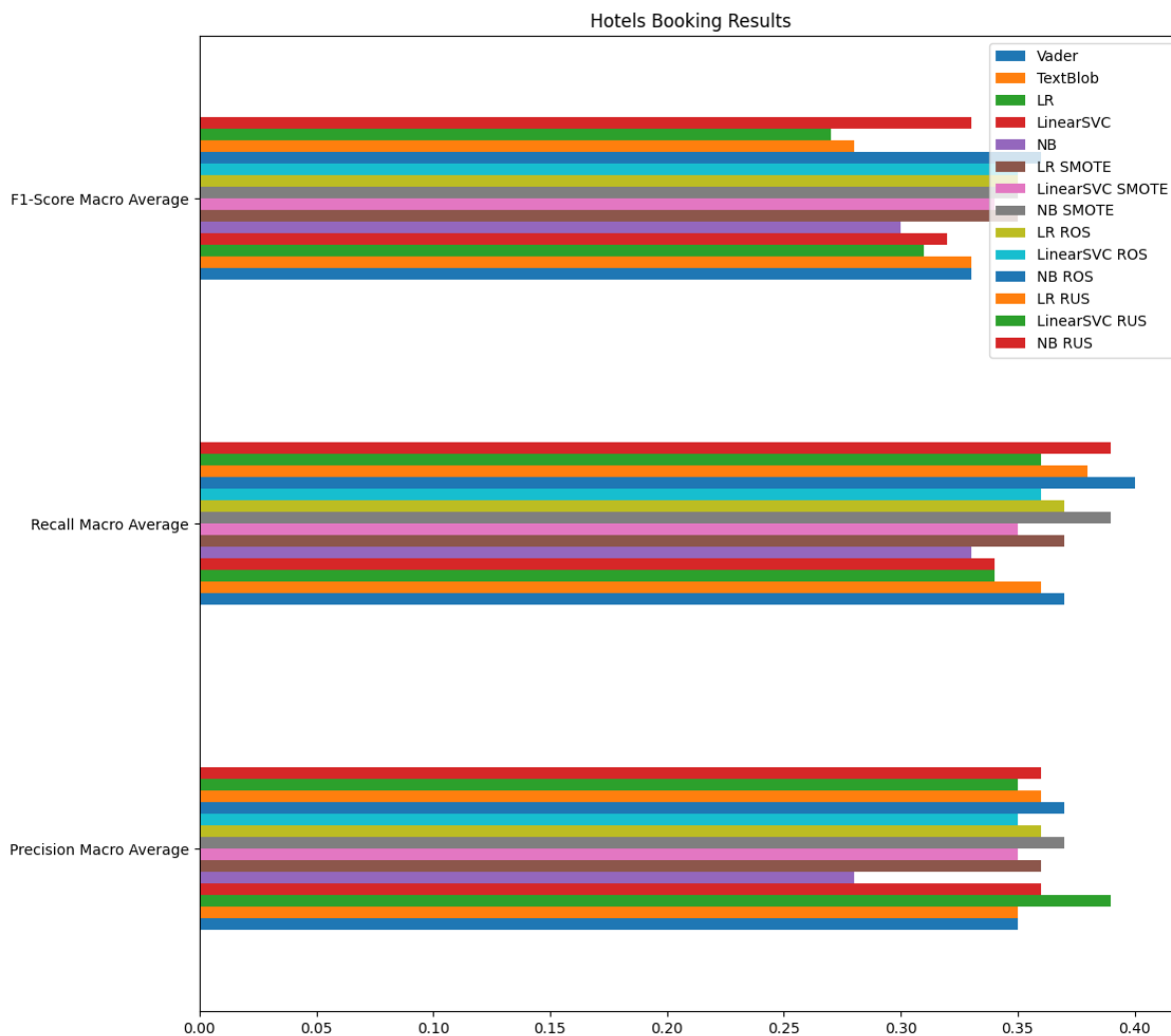
Παρόμοια με το προηγούμενο διάγραμμα, το διάγραμμα που αφορά τα ξενοδοχεία από το Tripadvisor φαίνεται να επέστρεψε την καλύτερη τιμή στο Precision Macro Average. Η τιμή αυτή αφορά το μοντέλο LinearSVC με την εφαρμογή της μεθόδου Sampling RUS (Random Undersampling), που έχει να κάνει με την τυχαία διαγραφή παραδειγμάτων από τις πολυπληθέστερες κλάσεις για την εξισορρόπηση. Το μόνο που πλησίαζε αυτό το μοντέλο σε τιμή είναι το NB RUS, δηλαδή το μοντέλο NB (Naive Bayes) με την εφαρμογή της ίδιας μεθόδου για την εξισορρόπηση. Τα υπόλοιπα κυμαίνονται περίπου στην ίδια τιμή, γύρω στα 0.4, όσον αφορά το Precision. Φυσικά, οι τιμές σε αυτό το διάγραμμα στο Precision είναι μικρότερες από 0.5, ενώ στο προηγούμενο διάγραμμα έφταναν τα 0.55. Αυτό δείχνει ότι τα μοντέλα δεν είχαν την ίδια επιτυχία με το προηγούμενο σετ δεδομένων. Φυσικά, αυτό μπορεί να οφείλεται στο μέγεθος του συνόλου δεδομένου,

που είναι αισθητά μεγαλύτερο από το προηγούμενο σετ.

Όσον αφορά το Recall, το LinearSVC μοντέλο χωρίς εφαρμογή κάποιας μεθόδου Sampling επέστρεψε την υψηλότερη τιμή. Κοντά σε αυτή την τιμή ήταν τα TextBlob και Vader λεξικά. Ακολουθούν τα αποτελέσματα από το μοντέλο NB με εφαρμογή μεθόδου SMOTE και το μοντέλο LR με εφαρμογή μεθόδου ROS.

Στο F1-Score, το VADER λεξικό ξεπέρασε όλα τα υπόλοιπα μοντέλα. Το επόμενο καλύτερο αποτέλεσμα προέρχεται από το LinearSVC χωρίς μέθοδο Sampling. Τέλος, τα NB SMOTE και NB ROS είχαν τις επόμενες υψηλότερες τιμές στο F1-Score. Όλα αυτά φαίνεται πως είχαν καλύτερη ισορροπία ανάμεσα στο Precision και το Recall στις δοκιμές.

5.3 Booking Ξενοδοχεία



Διάγραμμα 21: Αποτελέσματα στο σύνολο δεδομένων των ξενοδοχείων από Booking με διάφορες τεχνικές sampling

Το Booking είχε τις χαμηλότερες τιμές, αν συγκρίνουμε με τα προηγούμενα σύνολα δεδομένων. Στο Macro Average του Recall επιτεύχθηκε η μέγιστη τιμή, με την τιμή να φτάνει περίπου τα 0.4 στο Vader μοντέλο. Το επόμενο καλύτερο αποτέλεσμα, στο Recall, προέκυψε από το LinearSVC. Ακολουθεί το NB SMOTE ως την υψηλότερη τιμή μετά από αυτά, στις δοκιμές με την ανάκληση.

Όσον αφορά το Precision, το LinearSVC RUS φαίνεται ότι είχε την υψηλότερη τιμή, με την τιμή να φτάνει περίπου τα 0.38. Τα επόμενα δύο με την υψηλότερη τιμή είναι τα Vader και NB SMOTE.

Τέλος, στο F1-Score το Vader φαίνεται ότι επέστρεψε την καλύτερη τιμή. Άλλες δοκιμές με σχετικά καλύτερα, σε σχέση με τα υπόλοιπα, είναι τα LR ROS, LinearSVC ROS, NB SMOTE, LinearSVC SMOTE και LR SMOTE.

6 Συμπεράσματα και μελλοντικές προσθήκες

6.1 Μελλοντικές Προσθήκες

- Δοκιμή με περισσότερες μεθόδους επιλογής χαρακτηριστικών: Τα χαρακτηριστικά είναι σημαντικά γιατί αυξάνουν τα αποτελέσματά από την πρόβλεψη των στοιχείων. Φυσικά στην εργασία αυτή έγινε δοκιμή πολλαπλών μεθόδων για την βέλτιστη επιλογή χαρακτηριστικών και προσθήκη των καλύτερων αποτελεσμάτων στην εργασία. Παρόλα αυτά, τα αποτελέσματα μπορούν πιθανότατα να αυξηθούν από διαφορετικούς μεθόδους.
- Δοκιμή με κείμενα από την Ελληνική γλώσσα: Τα κείμενα που χρησιμοποιήθηκαν είναι κυρίως γραμμένα στην Αγγλική γλώσσα. Αυτό συμβαίνει, επειδή τα κείμενα από την Αγγλική γλώσσα είναι πιο δοκιμασμένα και υπάρχουν καλύτερα λεξικά για την επιστροφή στοιχείων, καθώς και καλύτερες επιλογές για την επιλογή χαρακτηριστικών. Μελλοντικά, μπορεί να κατασκευαστεί ένα καλύτερο λεξικό με την συναισθηματική πολικότητα και πιο συμβατό με κείμενα από τουριστικές ιστοσελίδες που θα συνεισφέρει στην δοκιμή με ελληνικά κείμενα. Φυσικά, οι περισσότεροι ξένοι τουρίστες ομιλούν στην Αγγλική γλώσσα και δεν θα υπάρχει μεγάλος αριθμός κειμένων.
- Συλλογή πληροφοριών από περισσότερες ταξιδιωτικές πλατφόρμες: Θα μπορούσε να γίνει σύγκρισή μεταξύ διαφορετικών ταξιδιωτικών ιστότοπων για την συλλογή περισσότερων στοιχείων και σύγκριση μεταξύ των ξεχωριστών μεθόδων κάθε ιστοσελίδας. Αυτό θα επέτρεπε τον πιο ολοκληρωμένο έλεγχο του τουρισμού στην χώρα.
- Εφαρμογή μεθόδων deep learning: Μέθοδοι όπως τα επαναληπτικά νευρωνικά δίκτυα θα επέστρεφαν υψηλότερα αποτελέσματα, καθώς μαθαίνουν. Συγκεκριμένα, θα μπορούσε να χρησιμοποιηθεί το LSTM (Long Short Term Memory) που θυμάται την σειρά των προηγούμενων δεδομένων. Επιλέγει τις χρήσιμες πληροφορίες για να κρατήσει και "ξεχνάει" αυτές που θεωρούνται άχρηστες.

6.2 Συμπεράσματα

Τα καλύτερα αποτελέσματα παρατηρήθηκαν στο σετ δεδομένων που αφορά τα αξιοθέατα του νομού Κέρκυρας, με τα ξενοδοχεία του Tripadvisor να ακολουθούν και τελευταία να είναι τα ξενοδοχεία του Booking. Αυτό προέκυψε από το μεγάλο πλήθος δεδομένων στο σετ με τα ξενοδοχεία από το Tripadvisor και το πλήθος των μη σχετικών λέξεων στο Booking. Η υψηλότερη τιμή που παρατηρήθηκε από όλα τα δεδομένα αφορούσε τα Precision και ήταν περίπου ίση με 0.55, στο σετ δεδομένων από τα αξιοθέατα. Γενικότερα, οι τιμές ήταν σχετικά χαμηλές με την κλάση για τα ουδέτερα στοιχεία συνήθως να αποτυγχάνει ως προς την ταξινόμησή της.

Η εργασία έδειξε ότι ο τουρισμός αυξήθηκε αισθητά μετά την περίοδο του COVID στην χώρα μας, όπως μπορεί να παρατηρηθεί στο γράφημα που αφορά τα ξενοδοχεία στο tripadvisor, με το 2022 να φτάνει τα ύψη, σε σύγκριση με παλαιότερες χρονιές. Το ίδιο σετ δεδομένων δεν συμπεριλάμβανε πληροφορίες για την καλοκαιρινή σεζόν του 2023, αν και οι τιμές αυτές αναμένονται να είναι υψηλές, δεδομένου των τιμών του 2022.

Βιβλιογραφία

- Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment analysis in tripadvisor. *IEEE Intelligent Systems*, 32(4), 72-77.
- Gao, S., Hao, J., & Fu, Y. (2015, June). The application and comparison of web services for sentiment analysis in tourism. In *2015 12th International Conference on Service Systems and Service Management (ICSSSM)* (pp. 1-6). IEEE.
- Keita, Z. (2022, November 21). Social Media Sentiment Analysis In Python With VADER | No Training Required! Medium. <https://towardsdatascience.com/social-media-sentiment-analysis-in-python-with-vader-no-training-required-4bc6a21e87b8>
- García, A., Gaines, S., & Linaza, M. T. (2012). A lexicon based sentiment analysis retrieval system for tourism domain. *Expert Syst Appl Int J*, 39(10), 9166-9180.
- Vásquez, J., Gómez-Adorno, H., & Bel-Enguix, G. (2021, September). Bert-based Approach for Sentiment Analysis of Spanish Reviews from TripAdvisor. In *IberLEF@ SEPLN* (pp. 165-170).
- Duarte, F. (2023b, April 3). Amount of data created daily (2023). Exploding Topics. <https://explodingtopics.com/blog/data-generated-per-day>
- Leelawat, N., Jariyapongpaiboon, S., Promjun, A., Boonyarak, S., Saengtabtim, K., Laosunthara, A., ... & Tang, J. (2022). Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning. *Heliyon*, 8(10). Taylor, P. (2022). Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025. Available on: <https://www.statista.com/statistics/871513/worldwide-data-created>.
- Newton, A. (2013). *Travel and travellers of the Middle Ages*. Routledge.
- Popp, A. (2009). From town to town: how commercial travel connected manufacturers and markets during the industrial revolution. *Journal of Historical Geography*, 35(4), 642-667.
- Travelling for pleasure: a brief history of tourism. (2020, June 16). Europeana. <https://www.europeana.eu/en/blog/travelling-for-pleasure-a-brief-history-of-tourism>
- The National Archives. (2021, August 3). Grand Tour - The National Archives. <https://www.nationalarchives.gov.uk/education/families/holidays-through-history/grand-tour/>
- IEG(<http://www.ieg-mainz.de>). (n.d.). The History of Tourism: Structures on the Path to Modernity | EGO. EGO | Europäische Geschichte Online. <http://ieg-ego.eu/en/threads/europe-on-the-road/the-history-of-tourism>
- Barney, N. (2023, March 10). sentiment analysis (opinion mining). Business Analytics. <https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining>
- Cooking, T. (2021, December 7). Tourism in Ancient Greece and Rome: Sparta, Pompeii & Capri. Medium. <https://medium.com/@travelingcooking/tourism-in-ancient-greece-and-rome-sparta-pompeii-capri-4315e860c374>
- Walton, J. K. (2023, October 28). tourism. Encyclopedia Britannica. <https://www.britannica.com/topic/tourism>

Krinis, N. (2022, December 16). Greece among Europe's top destinations to visit in 2023, says ETC market study. GTP Headlines. <https://news.gtp.gr/2022/12/14/greece-among-europes-top-destinations-to-visit-in-2023-says-etc-market-study/>

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. *Journal of travel research*, 58(2), 175-191.

Pandey, P. (2021, December 11). Sentiment Analysis is difficult, but AI may have an answer. Medium. <https://towardsdatascience.com/sentiment-analysis-is-difficult-but-ai-may-have-an-answer-a8c447110357>

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Bonta, V., Kumares, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(2), 1-6.

Ahmed, H. M., Javed Awan, M., Khan, N. S., Yasin, A., & Faisal Shehzad, H. M. (2021). Sentiment analysis of online food reviews using big data analytics. Hafiz Muhammad Ahmed, Mazhar Javed Awan, Nabeel Sabir Khan, Awais Yasin, Hafiz Muhammad Faisal Shehzad (2021) Sentiment Analysis of Online Food Reviews using Big Data Analytics. *Elementary Education Online*, 20(2), 827-836.

Vijay K. What is Linear Regression?- Spiceworks - Spiceworks. (2023, April 3). Spiceworks. <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/>

Sklearn.svm.LinearSVC. (n.d.-b). Scikit-learn. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

Gandhi, R. (2018, June 26). Naive Bayes classifier - towards data science. Medium. <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>

Hasson, E. (2019, December 29). What is scraping | About price & web scraping tools | Imperva. Learning Center. <https://www.imperva.com/learn/application-security/web-scraping-attack/>

Unadkat, J. (2023, August 23). Selenium Webdriver Tutorial in Java with Examples | BrowserStack. BrowserStack. <https://www.browserstack.com/guide/selenium-webdriver-tutorial>

Chandra, R. V., & Varanasi, B. S. (2015). *Python requests essentials*. Birmingham, UK: Packt Publishing.

Richardson, L. (2007). *Beautiful soup documentation*.

Tigerschiold, T. (2022, November 17). What is Accuracy, Precision, Recall and F1 Score? LabelF Blog. <https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score>

Kohli, S. (2022, May 3). Understanding a Classification Report For Your Machine Learning Model. Medium. <https://medium.com/@kohlshivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397>

Great Learning Team. (2022, October 24). An Introduction to Bag of Words in NLP using Python | What is BoW? Great Learning Blog: Free Resources What Matters to Shape Your Career! <https://www.mygreatlearning.com/blog/bag-of-words/>

Anirudha Simha, Principle Associate Software Engineer, Kai Chatbot Team. (2021, October 6). Understanding TF-IDF for Machine Learning. Capital One. <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/>

Mazumder, S. (2023, September 27). 5 Techniques to Handle Imbalanced Data For a Classification Problem. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/>

Shelke, M. S., Deshmukh, P. R., & Shandilya, V. K. (2017). A review on imbalanced data handling using undersampling and oversampling technique. *Int. J. Recent Trends Eng. Res*, 3(4), 444-449.

Kurian, D. D. M. K., Vishnupriya, S., Ramesh, R., Divya, G., Divya, D., Kurian, M. K., ... & Divya, D. (2015). Big data sentiment analysis using hadoop. *International Journal for Innovative Research in Science and Technology*, 1(11), 92-96.

Rodrigues, A. P., & Chiplunkar, N. N. (2019). A new big data approach for topic classification and sentiment analysis of Twitter data. *Evolutionary Intelligence*, 1-11.