ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Ανάλυση Μεγάλου Όγκου Δεδομένων και Κειμένων με στόχο την Τμηματοποίηση της Αγοράς

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ:

2021-2022

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΤΟΥ ΦΟΙΤΗΤΗ

Πασχαλίδη Μάριου

ΑΜ:2015010115

ΕΠΙΒΛΕΠΩΝ :
ΜΑΤΣΑΤΣΙΝΗΣ ΝΙΚΟΛΑΟΣ
Καθηγητής Πολυτεχνείου Κρήτης

Τριμελής εξεταστική επιτροπή:

Καθηγητής ΝΙΚΟΛΑΟΣ ΜΑΤΣΑΤΣΙΝΗΣ (Επιβλέπων)

Καθηγητής ΕΥΑΓΓΕΛΟΣ ΓΡΗΓΟΡΟΥΔΗΣ

Αναπλ. Καθηγητής ΣΤΕΛΙΟΣ ΤΣΑΦΑΡΑΚΗΣ

Περίληψη

Η ανάλυση μεγάλων όγκων δεδομένων από το διαδίκτυο αποτελεί σύγχρονη και πολύ αξιόπιστη μέθοδο εξαγωγής συμπερασμάτων για καταναλωτικές συμπεριφορές, όπως και κατηγοριοποίησης και ομαδοποίησης καταναλωτών και προϊόντων/επιχειρήσεων.

Στην διπλωματική εργασία, θα πραγματοποιηθεί η εξαγωγή δεδομένων από ιστοσελίδες αξιολόγησης τουριστικού ενδιαφέροντος σε Ελληνικές και ξένες ανταγωνιστικές τουριστικές περιοχές. Θα δοθεί έμφαση στην ανάλυση των αξιολογήσεων των χρηστών και των σχολίων τους. Στα δεδομένα που θα συγκεντρωθούν θα εφαρμοστούν μέθοδοι πολυκριτήριας ανάλυσης αποφάσεων, εξόρυξης δεδομένων και ανάλυσης κειμένων με στόχο την ανάλυση της συμπεριφοράς και τη δημιουργία ομάδων χρηστών/καταναλωτών (τμηματοποίηση αγοράς), την ανάλυση της συμπεριφοράς και την πρόβλεψη καταναλωτικών συμπεριφορών.

Τέλος, θα εξαχθούν συμπεράσματα για τα αποτελέσματα και τη χρήση ομάδας μεθόδων Πολυκριτήριας Ανάλυσης και μεθόδων Τεχνητής Νοημοσύνης και θα διατυπωθούν προτάσεις τμηματοποίησης της αγοράς και βελτίωσης των ανταγωνιστικών χαρακτηριστικών των επιμέρους περιοχών/ξενοδοχείων.

Abstract

Big Data Analysis on information from the internet is the contemporary and most reliable way of forming customer behavioral patterns, as well as classification and clustering of customers and businesses.

In this thesis, data from tourist rating websites were collected with web crawling methods, with specific targeting on Greek and competing regions. User ratings and reviews were targeted with more importance. On the whole total of the collected data, we applied Multi Criteria Analysis methods, data mining and text analysis with the aim of analyzing and predicting customer behavior.

Finally, we present our conclusions on the results and use of Multi Criteria Analysis Methods and Artificial Intelligence algorithms, suggest ways of clustering the market and the strategic enhancement on specific criteria based on competition.

Ευχαριστίες

Για την εκπόνηση αυτής της εργασίας θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή μου Νικόλαο Ματσατσίνη για την συνεχή του καθοδήγηση και το ενδιαφέρον του, όχι μόνο για την εκπόνηση αυτής της εργασίας, αλλά και για την εις βάθος κατανόηση της ουσίας του προβλήματος.

Πίνακας Περιεχομένων

Πίνακας Σχημάτων

Πίνακας Πινάκων

Πίνακας Διαγραμμάτων

Κεφάλαιο 1: Εισαγωγή

1.1 Σκοπός Εργασίας

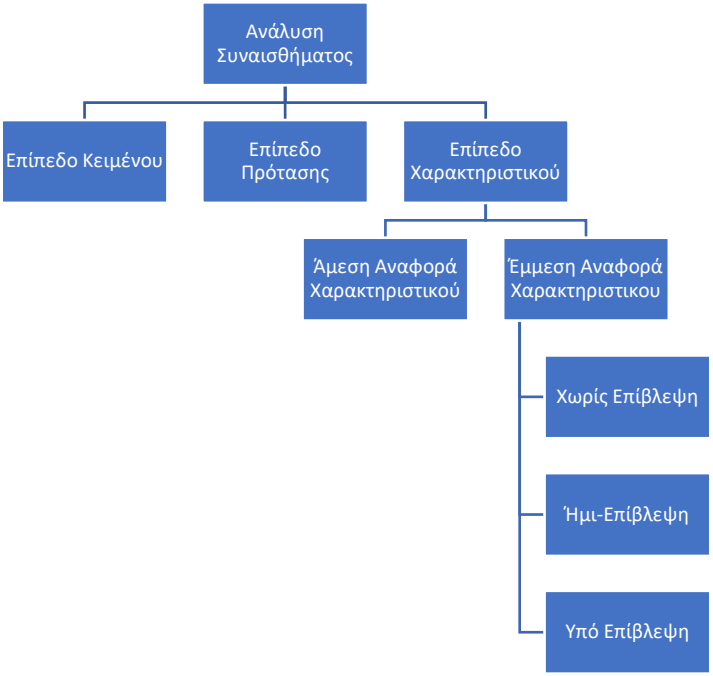
Η βιομηχανία του τουρισμού αποτελεί μία από τις σημαντικότερες πηγές εσόδων στην Ελλάδα. Ο αυξανόμενος ανταγωνισμός στην αγορά και η εμφάνιση καινούριων ειδών προσφοράς, οδηγεί κάθε επιχείρηση να επανεξετάσει και να οργανώσει εκ νέου τις δραστηριότητές της, δίνοντας έμφαση σε στοχευμένα χαρακτηριστικά.

Ο τουρισμός ανήκει στον τομέα των υπηρεσιών, οπότε ο σημαντικότερος παράγοντας αξιολόγησης των διαφορετικών επιχειρήσεων αποτελεί η γνώμη τους. Παλιότερα αυτή συγκεντρωνόταν μέσω ερωτηματολογίων και άλλων χρονοβόρων διαδικασιών, σήμερα όμως, υπάρχουν πολλές εφαρμογές και ιστοσελίδες στις οποίες χρήστες δημοσιεύουν την γνώμη και την αξιολόγησή τους για την εμπειρία τους. Η ανάλυση μεγάλων όγκων δεδομένων αξιοποιεί αυτά τα δεδομένα και αποτελεί σύγχρονη και πολύ αξιόπιστη μέθοδο εξαγωγής συμπερασμάτων για καταναλωτικές συμπεριφορές.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η ανάλυση της συμπεριφοράς των χρηστών/καταναλωτών (τμηματοποίηση αγοράς), από την ανάλυση συναισθήματος και εξαγωγή προτιμήσεων (θεμάτων αναφοράς) από τα σχόλια των χρηστών. Στη συνέχεια, διαμορφώνονται τα αποτελέσματα των προηγούμενων αναλύσεων κατάλληλα για πολυκριτήρια ανάλυση (UTASTAR) σε συνδυασμό με τα δεδομένα των επιχειρήσεων/ξενοδοχείων με σκοπό την τμηματοποίηση της αγοράς των ξενοδοχείων και την ανάλυση του ανταγωνισμού των ελληνικών ξενοδοχείων. Τέλος, θα εξαχθούν συμπεράσματα και θα διατυπωθούν προτάσεις τμηματοποίησης της αγοράς και βελτίωσης των ανταγωνιστικών χαρακτηριστικών των επιμέρους περιοχών/ξενοδοχείων.

1.2 Ανάλυση Κειμένου

Στην περίπτωση των σχολίων σε σελίδες αξιολόγησης, η ανάλυση κειμένου έχει ως σκοπό την Ανάλυση Συναισθήματος (Sentiment Analysis) ή Εξόρυξη Γνώμης (Opinion Mining) δηλαδή την υπολογιστική μέθοδο αναγνώρισης και προσδιορισμό της ανθρώπινης γνώμης ή συναισθήματος, το οποίο αναφέρεται σε διαφορετικά πεδία.
Η Ανάλυση Συναισθήματος όπως εξηγείται και στις εργασίες (Mohammad Tubishat, 2018) και (Medhat, Hassan, & Korashy, 2014) μπορεί να κατηγοριοποιηθεί ανάλογα με το ιεραρχικό επίπεδο που γίνεται η ανάλυση σε: Ανά Κείμενο, Ανά Πρόταση και Ανά Χαρακτηριστικό. Επίσης, μπορούν να χωριστούν ανάλογα από την ανάγκη από επίβλεψη για την λειτουργία του αλγορίθμου από τον αναλυτή σε: Χωρίς Επίβλεψη, Υπό Επίβλεψη και Ημι-Επίβλεψη.
Στην (Mohammad Tubishat, 2018) οι μέθοδοι Ανάλυσης Συναισθήματος που μελετήθηκαν χωρίστηκαν σύμφωνα στο Σχήμα 1:



Σχήμα 1: Κατηγοριοποίηση Αλγορίθμων Ανάλυσης Συναισθήματος

Στην περίπτωση Απόδοσης Συναισθήματος για συγκεκριμένο Χαρακτηριστικό, η αναφορά του χαρακτηριστικού μπορεί να γίνεται στο κείμενο/πρόταση: Έμμεσα ή Άμεσα όπως φαίνεται στο παρακάτω παράδειγμα, το χαρακτηριστικό της Τιμής αναφέρεται έμμεσα (φθηνό) ενώ το χαρακτηριστικό της ποιότητας ήχου άμεσα (ποιότητα ήχου):

Το ηχείο είναι αρκετά φθηνό, αλλά έχει καλή ποιότητα ήχου.

Επειδή στη γλώσσα είναι πολύ συχνό να αναφερθούμε σε Χαρακτηριστικά Έμμεσα οι περισσότεροι αλγόριθμοι που χρησιμοποιούνται είναι ικανοί να αναγνωρίσουν και την έμμεση αναφορά των χαρακτηριστικών.

Οι τρόποι που επιτυγχάνεται η Εξόρυξη Γνώμης συνήθως είναι: Μέσω λεξικών κανόνων, Μέσω Μηχανικής Μάθησης ή Υβριδική χρήση και των δύο μεθόδων. Η δημιουργία λεξικών κανόνων απαιτεί ένα οργανωμένο λεξικό με την κατηγοριοποίηση των λέξεων ως μέρη του λόγου και απόδοση συναισθήματος από πριν σε αυτές, και δημιουργία συντακτικών κανόνων που να κάνουν χρήση του λεξικού. Η μηχανική μάθηση κάνει χρήση μεγάλου όγκου δεδομένων κατάλληλα διαμορφωμένου, στον οποίο εκπαιδεύεται και είναι σε θέση να δώσει τιμές, της μορφής των δεδομένων που εκπαιδεύτηκε, σε καινούρια δεδομένα.

Η πιο συνηθισμένη κατηγορία αλγορίθμων είναι χωρίς επίβλεψη που μπορούν να αναγνωρίσουν χαρακτηριστικά που αναφέρονται έμμεσα. Αυτό συμβαίνει διότι συνήθως αναλύεται μεγάλος όγκος δεδομένων και αυτοί οι αλγόριθμοι είναι οι ταχύτεροι, καθώς και γιατί υπάρχει η ανάγκη της όσο το δυνατόν λιγότερης εμπλοκής του αναλυτή λόγω της αδυναμίας διαχείρισης του μεγάλου όγκου.

Στις περισσότερες αναλύσεις χρησιμοποιούνται κείμενα στα Αγγλικά ή τα Κινέζικα, καθώς είναι πιο εύκολα διαθέσιμα αυτά τα δεδομένα και οι περισσότεροι αλγόριθμοι έχουν αναπτυχθεί πάνω σε αυτές τις γλώσσες.

1.3 Υφιστάμενη Κατάσταση

Στη συνέχεια θα αναλυθούν συνοπτικά μερικές εργασίες στις οποίες επιχειρείται παρόμοιος σκοπός με την παρούσα, δηλαδή ανάλυση συναισθήματος σε σχόλια σε ξενοδοχεία και πολυκριτήρια ανάλυση των αποτελεσμάτων.

Στην εργασία (Priyantina & Sarno, 2019) επιχειρείται δημιουργία μοντέλου εξόρυξης δεδομένων από σχόλια σε ξενοδοχεία με τη χρήση των μεθόδων Latent Dirichlet Allocation, Term Frequency – Inverse Cluster Frequency (LDA & TF-ICF) και Long-Short Term Memory (LSTM) για την απόδοση σημασιολογικής ομοιότητας σε σχέση με τα 5 επιλεγμένα χαρακτηριστικά (Καθαριότητα, Άνεση, Υπηρεσίες, Φαγητό και Τοποθεσία) τα οποία χρησιμοποιεί η σελίδα αξιολόγησης. Αρχικά για κάθε χαρακτηριστικό επιλέχθηκε ένα μικρό λεξικό με λέξεις-κλειδιά των χαρακτηριστικών που να αφορούν τα ξενοδοχεία. Η διαδικασία της μεθοδολογίας ξεκινά με τη συλλογή των δεδομένων (σχόλια) από την ιστοσελίδα Traveloka.com και στη συνέχεια πραγματοποιείται προεπεξεργασία των δεδομένων για την οποία επιλέχθηκαν τα ακόλουθα βήματα: 1) Tokenize (διαχωρισμός των λέξεων στο κείμενο του σχολίου), 2) Normalization (μετατροπή όλων των γραμμάτων σε μικρά), 3) Απομάκρυνση των σημείων στίξης, 4) Stemming (επαναφορά των λέξεων στην αρχική ρίζα, κλίση ή χρόνο τους, 5) Απομάκρυνση των stopwords (λέξεις όπως δηλαδή άρθρα, αντωνυμίες, επιρρήματα, ρήματα που δεν προσφέρουν πληροφορία κ.α.). Τέλος τους εφαρμόζεται 6) Αλγόριθμος ελέγχου της ορθογραφίας των λέξεων.
Έπειτα εφαρμόζεται ο αλγόριθμος LDA στα επεξεργασμένα σχόλια ώστε να εξαχθούν τα κρυμμένα πιο συχνά θέματα αναφοράς που προκύπτουν από τη συλλογή των σχολίων. Στα αποτελέσματα του αλγορίθμου υπολογίζεται η πιθανότητα κάθε σχολίου και κάθε λέξης να ανήκει σε ένα από τα θέματα. Τα θέματα αυτά θα συγκριθούν με τα αρχικά επιλεγμένα χαρακτηριστικά και τα λεξικά που δημιουργήθηκαν για αυτά.
Ο έλεγχος της σημασιολογικής ομοιότητας των λέξεων επιχειρήθηκε σε τρεις παραλλαγές οι οποίες διαφέρουν στο ποσοστό χρήσης του πίνακα συχνοτήτων της μεθόδου TF-ICF. Σε όλες τις παραλλαγές χρησιμοποιείται το μοντέλο που προέκυψε από τον LDA στον οποίο εισάγονται οι λέξεις-κλειδιά από το αρχικό λεξικό των χαρακτηριστικών. Έτσι το μοντέλο του LDA υπολογίζει την σημασιολογική ομοιότητα κάθε λέξης των σχολίων με κάθε χαρακτηριστικό. Στις παραλλαγές επεκτείνεται το αρχικό λεξικό από τον πίνακα του TF-ICF σε ποσοστά 20% και 100% συμπερίληψης του πίνακα. Ο πίνακας του αλγορίθμου TF-ICF υπολογίζεται μέσω της συχνότητας εμφάνισης των λέξεων στις συστάδες που συνιστούν τα χαρακτηριστικά. Έτσι, τελικά υπολογίζεται η σημασιολογική ομοιότητα των λέξεων και σε συνέχεια των σχολίων με κάθε χαρακτηριστικό, και του αποδίδεται αυτό με την μεγαλύτερη τιμή.
Η απόδοση συναισθήματος επιτυγχάνεται με την εκπαίδευση του νευρωνικού δίκτύου LSTM. Για την δημιουργία της εισόδου του νευρωνικού χρησιμοποιείται το έτοιμο λεξικό διάνυσμα του Stanford GloVe, σύμφωνα με το οποίο προβάλλονται οι λέξεις των σχολίων και χρησιμοποιούνται

από το μοντέλο SentiWordNet. Το λεξικό GloVe επεκτείνεται από τον πίνακα του TF-ICF ώστε να περιέχονται όλες οι λέξεις των σχολίων.
Στα συμπεράσματα της εργασίας αναφέρεται η αποτελεσματικότητα των παραλλαγών της μεθοδολογίας, με τον συνδυασμό LDA + TF-ICF 100%, να έχει την καλύτερη απόδοση στην απόδοση χαρακτηριστικού στα σχόλια, και να ταξινομεί με ακρίβεια 85%, ενώ η απόδοση συναισθήματος να επιτυγχάνεται σε ακρίβεια 93%.

Στην εργασία (Nie, Tian, & Wang, 2020) επιχειρείται συγκριτική ανάλυση ξενοδοχείων με βάση τα δεδομένα που προκύπτουν από την επεξεργασία των κειμένων των σχολίων τους. Για την εξαγωγή των δεδομένων εφαρμόζεται σημασιολογική απόδοση και ανάλυση συναισθήματος στα σχόλια, τα οποία δεδομένα χρησιμοποιούνται σε πολυκριτήρια ανάλυση.
Αρχικά γίνεται η επιλογή της ιστοσελίδας TripAdvisor.com ως η δημοφιλέστερη σελίδα τουριστικών αξιολογήσεων, από όπου θα συλλεχθούν τα σχόλια και επιλέχθηκαν 30 ξενοδοχεία στα οποία θα αναφέρονται τα σχόλια. Τα χαρακτηριστικά σύμφωνα με τα οποία επιλέχθηκε να γίνει η σημασιολογική απόδοση στα σχόλια είναι τα χαρακτηριστικά που δίνει την δυνατότητα το TripAdvisor στους χρήστες να βαθμολογήσουν τα ξενοδοχεία, δηλαδή Αξία, Τοποθεσία, Δωμάτιο, Καθαριότητα, Υπηρεσία και Ποιότητα Ύπνου.
Έπειτα εφαρμόστηκε η προεπεξεργασία των κειμένων των σχολίων με 1) Διαχωρισμό των λέξεων, 2) Χαρακτηρισμός τους ως μέρη του λόγου, 3) Απομάκρυνση των stop words και τέλος 4) Απομάκρυνση λέξεων που εμφανίζονται σε πολύ μικρή συχνότητα και προσφέρουν μικρή πληροφορία.
Για την ανάλυση συναισθήματος εκπαιδεύτηκε το μοντέλο word2vec με βάση το λεξικό συναισθήματος HowNet ώστε με είσοδο το προεπεξεργασμένο σχόλιο να αποδίδει τιμή 0-1 που να χαρακτηρίζει το συναίσθημα του σχολίου (αρνητικό-θετικό). Στη συνέχεια διαχωρίζεται η συνεχής τιμή συναισθήματος σε επιλεγμένα διαστήματα για μία επταβάθμια κλίμακα, έτσι ώστε να μετασχηματιστεί η γνώμη του σχολίου στη μορφή εξαγωγής γνώμης που χρησιμοποιούν συνήθως τα ερωτηματολόγια. Τέλος, στην μεθοδολογία προστέθηκαν κανόνες καλύτερου διαχωρισμού του συναισθήματος, όπως για παράδειγμα εάν υπάρχει λέξη που αντιστρέφει το νόημα πριν τη λέξη συναισθήματος, απόδοση του αντίθετου συναισθήματος.
Στην απόδοση χαρακτηριστικού στα σχόλια χρησιμοποιήθηκε και πάλι το μοντέλο word2vec, το οποίο υπολόγιζε την ομοιότητα των λέξεων του σχολίου με τις λέξεις σε ειδικά δημιουργημένο λεξικό του κάθε χαρακτηριστικού. Το αθροιστικό αυτό αποτέλεσμα σημασιολογικής ομοιότητας του σχολίου με, σε συνδυασμό με τα παραπάνω αποτελέσματα απόδοσης συναισθήματος, τελικά αναπαρίσταται ως πιθανοτικό μοντέλο τύπου LDA, στο οποίο σε κάθε σχόλιο και για κάθε κριτήριο εμφανίζεται το ποσοστό ομοιότητας στην επταβάθμια κλίμακα συναισθήματος.
Το τελικό αποτέλεσμα (απόδοση συναισθήματος και σημασιολογική ομοιότητα με τα κριτήρια) αθροίζεται για κάθε ξενοδοχείο και υπολογίζεται για αυτά η βαθμολογία για κάθε χαρακτηριστικό, δημιουργώντας τον πολυκριτήριο πίνακα.
Ο πολυκριτήριος πίνακας σε συνέχεια χρησιμοποιείται για τον υπολογισμό των βαρών για κάθε κριτήριο που εκφράζουν συνολικά τους σχολιαστές.
Τέλος, γίνεται ανάλυση της ακρίβειας των αποτελεσμάτων της μεθοδολογίας με βάση τα βάρη των κριτηρίων που υπολογίστηκαν και την τελική κατάταξη που προέκυψε από την ανάλυση, σε σχέση με άλλες μεθοδολογίες και τα δικά τους αποτελέσματα.

Στην εργασία (Guo, Barnes, & Jia, 2017) επιχειρείται εξόρυξη γνώμης σε σχόλια σε τουριστικά καταλύματα και ανάλυση της συμπεριφοράς των σχολιαστών.
Αρχικά, έγινε η συλλογή 250.000 σχολίων από 25,670 ξενοδοχεία, 16 χωρών για περίοδο 3 μηνών, από τη σελίδα του TripAdvisor και τα στοιχεία των σχολιαστών που εμφανίζονται στην ιστοσελίδα.
Στη συνέχεια, πραγματοποιήθηκε προεπεξεργασία των κειμένων των σχολίων έτσι ώστε να εισαχθούν στον αλγόριθμο LDA για την εξαγωγή των κρυμμένων θεμάτων αναφοράς.
Επίσης, έγινε στατιστική ανάλυση στην συχνότητα εμφάνισης των λέξεων στα σχόλια.
Μετά την εξαγωγή των κρυμμένων θεμάτων, επιχειρήθηκε η ομαδοποίηση και ο χαρακτηρισμός των θεμάτων που προέκυψαν ως αποτέλεσμα του LDA. Σύμφωνα με τον χαρακτηρισμό των θεμάτων έγινε η αξιολόγηση των αποτελεσμάτων, καθώς και η σύγκρισή τους σε σχέση με προηγούμενες μελέτες μέσω του στατιστικού μέτρου Jaccard Coefficient. Η αξιολόγηση των αποτελεσμάτων του LDA έγινε από δύο ερευνητές οι οποίοι έκριναν την ακρίβεια εξαγωγής των θεμάτων σε δείγμα των δεδομένων.
Έπειτα, πραγματοποιήθηκε στατιστική ανάλυση των αποτελεσμάτων και των δεδομένων των σχολιαστώ. Αρχικά, έγινε ανάλυση της συχνότητας εμφάνισης των θεμάτων αυτών στα σχόλια και διαχωρισμός τους ανάλογα με την δυνατότητα ελέγχου από το ξενοδοχείο πάνω σε αυτά. Στη συνέχεια, στατιστικά αναπαράσταση εμφάνισης των θεμάτων σε σχόλια ανάλογα με το φύλλο του σχολιαστή καθώς και την ηλικία του.
Τέλος, έγινε διαλογή των σχολίων με πλήρη βαθμολόγηση στα χαρακτηριστικά του TripAdvisor και πραγματοποιήθηκε Regression Analysis ώστε να υπολογιστεί πόσο επηρεάζει κάθε χαρακτηριστικό τη γνώμη των χρηστών.

Στην εργασία (Kumar & Parimala, 2020), αρχικά, γίνεται η συλλογή σχολίων σε προϊόντα από τη σελίδα της Amazon, στα οποία έγινε προεπεξεργασία μέσω του πακέτου της python Natural Language ToolKit με διαχωρισμό των λέξεων. Χαρακτηρισμό τους ως μέρη του λόγου, απομάκρυνση των stopwords και έλεγχο ορθογραφίας.
Στη συνέχεια στα προεπεξεργασμένα σχόλια έγινε εξαγωγή των πιο συχνών θεμάτων αναφοράς μέσω της μεθόδου Term Frequency, τα οποία συγκρίθηκαν με αυτά που συλλέχθηκαν από ερωτηματολόγια και τη γνώμη ειδικών. Τα χαρακτηριστικά που συμβάδιζαν σε όλες τις μεθόδους κρατήθηκαν και χρησιμοποιήθηκαν ως κριτήρια αξιολόγησης των προϊόντων. Οι λέξεις που προέκυψαν από αυτή τη μέθοδο χρησιμοποιήθηκαν για την απόδοση χαρακτηριστικού στα σχόλια.
Για την ανάλυση συναισθήματος επιλέχθηκε να χρησιμοποιηθούν μόνο οι λέξεις που χαρακτηρίστηκαν ως επίθετα και μέσω το εκπαιδευμένου λεξικό συναισθήματος SentiWordNet.
Οπότε για κάθε σχόλιο, μετά από την παραπάνω διαδικασία εύρεσης αναφοράς στα χαρακτηριστικά, εφαρμόστηκε η απόδοση συναισθήματος σε εύρος τριών λέξεων πριν και μετά την λέξη του χαρακτηριστικού και στη συνέχεια υπολογίστηκε αθροιστικά το συναίσθημα κάθε σχολίου σε κλίμακα 0-1 μετά από κανονικοποίηση.
Τα αποτελέσματα της ανάλυσης συναισθήματος για τα κριτήρια στα σχόλια χρησιμοποιήθηκαν για τον υπολογισμό του πολυκριτήριου πίνακα για τα προϊόντα (κάμερες) στα επιλεγμένα κριτήρια. Για κάθε προϊόν υπολογίστηκαν τα βάρη για κάθε κριτήριο ανάλογα με την συχνότητα αναφοράς του κριτηρίου στα σχόλια για το προϊόν, τα οποία βάρη επίσης χρησιμοποιήθηκαν στη δημιουργία του πολυκριτήριου πίνακα.

Τέλος γίνεται μελέτη στο πώς επηρεάζεται η κατάταξη των εναλλακτικών, από την εφαρμογή της πολυκριτήριας ανάλυσης Weighted Sum Method, εάν μειωθούν τα κριτήρια που προέκυψαν στον πολυκριτήριο πίνακα. Τα αποτελέσματα των παραλλαγών της πολυκριτήριας ανάλυσης συγκρίνονται με την κατάταξη των εναλλακτικών από ειδικούς ώστε να βρεθεί το πιο ακριβές μοντέλο.

Στην εργασία (Sutherland & Sim, 2020) επιχειρείται η εξαγωγή των πιο σημαντικών κριτηρίων για τους τουρίστες από διαδικτυακά σχόλια σε καταλύματα στην Νότια Κορέα, μέσω του αλγόριθμου LDA.
Αρχικά έγινε η συλλογή των σχολίων από διαφορετικές ιστοσελίδες αξιολόγησης, μαζί με πληροφορίες των σχολιαστών και των καταλυμάτων. Τα σχόλια, στη συνέχεια, κατηγοριοποιήθηκαν κατά τοποθεσία (αστική-επαρχιακή) και κατά είδος καταλύματος.
Στη συνέχεια πραγματοποιήθηκε προεπεξεργασία στα σχόλια, με αφαίρεση των stopwords, χωρισμό του κειμένου ανά λέξη, ανά ομάδα λέξεων ή ανά φράση με σκοπό τη διατήρηση όσο το δυνατόν περισσότερου νοήματος και επαναφορά των λέξεων στην αρχική τους ρίζα.
Υπολογίστηκε ο αριθμός των θεμάτων για τον LDA όπου τα θέματα που προκύπτουν να έχουν τον καλύτερο διαχωρισμό μεταξύ τους. Εφαρμόστηκε ο αλγόριθμος για αυτόν τον αριθμό θεμάτων στα προεπεξεργασμένα σχόλια και αξιολογήθηκαν και τους αποδόθηκε σημασία και τίτλος από μία επιτροπή αναλυτών και εργαζομένων στον τομέα του τουρισμού. Η επιτροπή εξέτασε τις πιο σημαντικές λέξεις αλλά και τα σχόλια με το μεγαλύτερο ποσοστό για κάθε θέμα, ώστε τελικά να χαρακτηρίσουν ομόφωνα το κάθε θέμα ως ένα κριτήριο ικανοποίησης των σχολιαστών. Τέλος, ομαδοποιήθηκαν τα θέματα σε γενικότερες κατηγορίες (Γενικά Χαρακτηριστικά Ικανοποίησης και Κίνητρα επιπλέον επίσκεψης, Χαρακτηριστικά σχετικά με την Υπηρεσία, Χαρακτηριστικά σχετικά με την Τοποθεσία, Χαρακτηριστικά σχετικά την κατάσταση του Δωματίου και Χαρακτηριστικά σχετικά με το Περιβάλλον του Ξενοδοχείου (Εγκαταστάσεις-Παροχές και Ατμόσφαιρα)
Τέλος, τα θέματα που προέκυψαν χρησιμοποιήθηκαν για τον στατιστικό έλεγχο της κατανομής της αναφοράς των θεμάτων στα σχόλια ανά κατηγορία που χωρίστηκαν τα ξενοδοχεία (αστικά-επαρχιακά και είδος καταλύματος).

Συνολικά από την υφιστάμενη κατάσταση που παρουσιάστηκε στις παραπάνω εργασίες, συμπεραίνουμε ότι τα κοινά στοιχεία συμπυκνώνονται στις παρακάτω παρατηρήσεις.

Καθώς όλες οι εργασίες στηρίζονται σε σχόλια και αξιολογήσεις από πλατφόρμες στο διαδίκτυο είναι σημαντική η εξαγωγή των πιο συχνών χαρακτηριστικών – θεμάτων αναφοράς που απασχολούν τους σχολιαστές.

Στη συνέχεια, είναι σημαντική, κατά την εφαρμογή της ανάλυσης συναισθήματος στα σχόλια, η δυνατότητα αναγνώριση της αναφοράς στα χαρακτηριστικά και η αναγνώριση του συναισθήματος σχετικά με το συγκεκριμένο θέμα.

Τέλος, χρειάζεται τρόπος μετατροπής και συμπύκνωσης της πληροφορίας από τις προηγούμενες διαδικασίες, με στατιστικές μεθόδους ή μεθόδους πολυκριτήριας ανάλυσης για την δημιουργία χρήσιμης πληροφορίας για τους σχολιαστές.

Κεφάλαιο 2: Θεωρητικό Υπόβαθρο

2.1 Εξόρυξη Χαρακτηριστικών

Στη συνέχεια θα παρουσιαστούν συνοπτικά μέθοδοι Εξόρυξης Χαρακτηριστικών από κείμενα όπως παρουσιάζονται στην εργασία (Chetty, Gautami, & Naganna, 2015).

2.1.1 Επιλογή Χαρακτηριστικών Κειμένου

Το πρώτο βήμα στο πρόβλημα της εύρεση των χαρακτηριστικών από ένα σύνολο κειμένων είναι η επεξεργασία και Επιλογή των Χαρακτηριστικών του κειμένου (Feature Selection). Κύριες μέθοδοι της Επιλογής των Χαρακτηριστικών του κειμένου αποτελούν:

- Συχνότητα Παρουσίας των λέξεων ή συμπτυγμάτων λέξεων (Term Frequency)
- Απόδοση ως μέρη του λόγου στις λέξεις (Parts Of Speech)
- Εύρεση λέξεων ή φράσεων συναισθήματος (Opinion words and phrases)
- Εύρεση λέξεων ανατροπής νοήματος (Negations)
- Δημιουργία Bag of Words, δηλαδή στατιστική απεικόνιση των λέξεων με διατήρηση της σειράς με την οποία εμφανίζονται
- Επαναφορά των λέξεων στην αρχική τους κλίση-ρίζα (Stemming)

2.1.2 Στατιστικές Μέθοδοι Επιλογής Χαρακτηριστικών

Η μέθοδος Point-wise Mutual Information (PMI) αποτυπώνει την σχέση μεταξύ λέξης και κλάσης μέσω της συχνότητας ταυτόχρονης εμφάνισης τους. Η σχέση αυτή υπολογίζεται

$$M_i(w) = \log\left(\frac{F(w)\cdot p_i(w)}{F(w)\cdot P_i}\right) = \log\left(\frac{p_i(w)}{P_i}\right)$$

(1)

όπου η προβλεπόμενη ταυτόχρονη εμφάνιση της κλάσης i με την λέξη w υπολογίζεται από P_i x F(w) ενώ η πραγματική από F(w) x p_i(w).

Η μέθοδος χ^2 υπολογίζει επίσης την σχέση μεταξύ λέξεων και κλάσεων σε ένα κείμενο ή σύνολο κειμένων ως εξής:

$$\chi_i^2 = \frac{n\cdot F(w)^2\cdot\left(p_i(w)-P_i\right)^2}{F(w)\cdot(1-F(w))\cdot P_i\left(1-P_i\right)}$$

(2)

όπου n είναι ο αριθμός των κειμένων, p_i(w) η πιθανότητα της κλάση i για τα κείμενα που περιέχουν την λέξη w, P_i το ποσοστό των κειμένων της κλάσης i και F(w) το ποσοστό των κειμένων που περιέχουν την λέξη w.

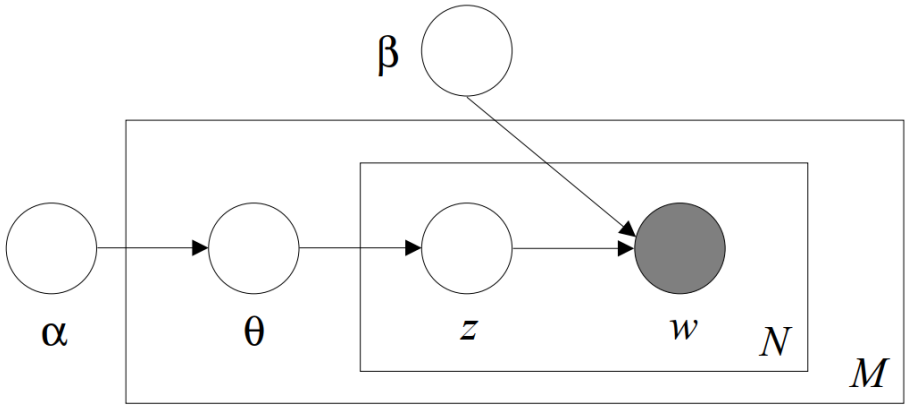
Η διαφορά των μεθόδων PMI και χ^2 είναι ότι τα αποτελέσματα της σχέσης των λέξεων με τα χαρακτηριστικά στην μέθοδο χ^2 είναι κανονικοποιημένα, κάνοντας τη σύγκριση των αποτελεσμάτων των λέξεων του ίδιου χαρακτηριστικού πιο ακριβή. Για αυτό το λόγο η μέθοδος χ^2 χρησιμοποιείται περισσότερο σε αλγορίθμους όπως ο SVM

Ένας άλλος τρόπος επιλογής χαρακτηριστικών είναι η Εύρεση Κρυμμένων Σημασιολογικών Εννοιών (Latent Semantic Indexing). Όπως εξηγείται στην εργασία (Rosario, 2000), h μέθοδος αυτή πραγματοποιείται χωρίς επίβλεψη και επιχειρείται η μείωση των διαστάσεων των δεδομένων και τελικά ο υπολογισμός γραμμικής σχέσης μεταξύ των χαρακτηριστικών. Αυτό επιτυγχάνεται μέσω του αλγορίθμου Principal Component Analysis η οποία μειώνει τον αριθμό των διαστάσεων με βάση τη διατήρηση της μέγιστης ποσότητας πληροφορίας, δηλαδή κατανομή στις κλάσεις.

Τέλος, σημαντικός αλγόριθμος επιλογής χαρακτηριστικών είναι ο Latent Dirichlet Allocation (LDA). Όπως προτάθηκε από τους (Blei, David, Andrew, & Michael, Allocation), ο LDA είναι μοντέλο πιθανοτικής γενίκευσης σε μία συλλογή κειμένων ο οποίος βασίζεται στην ιδέα ότι κάθε κείμενο αναπαριστά τυχαία μίξη κρυμμένων θεμάτων και κάθε θέμα αντιστοιχεί σε μία κατανομή λέξεων. Για τον LDA υποθέτουμε την παρακάτω διαδικασία γενίκευσης για κάθε κείμενο w σε συλλογή κειμένων D:

1. Επιλογή N ~ Poisson(ξ)
2. Επιλογή θ ~ Dir(θ)
3. Για κάθε N λέξεις w_n:
    a. Επιλογή θέματος z_n ~ Πολυωνυμική Κατανομή (θ)
    b. Επιλογή λέξης w_n από $p(w_n \mid z_n, \beta)$, πρόβλεψη πιθανότητας (πολυωνυμική κατανομή) για το θέμα z_n

Στο Σχήμα 2 απεικονίζεται η λειτουργία του μοντέλου του LDA όπου το εξωτερικό πλαίσιο αντιστοιχεί στα κείμενα ενώ το εσωτερικό αντιστοιχεί στην κατανομή των λέξεων στα θέματα για το κείμενο.

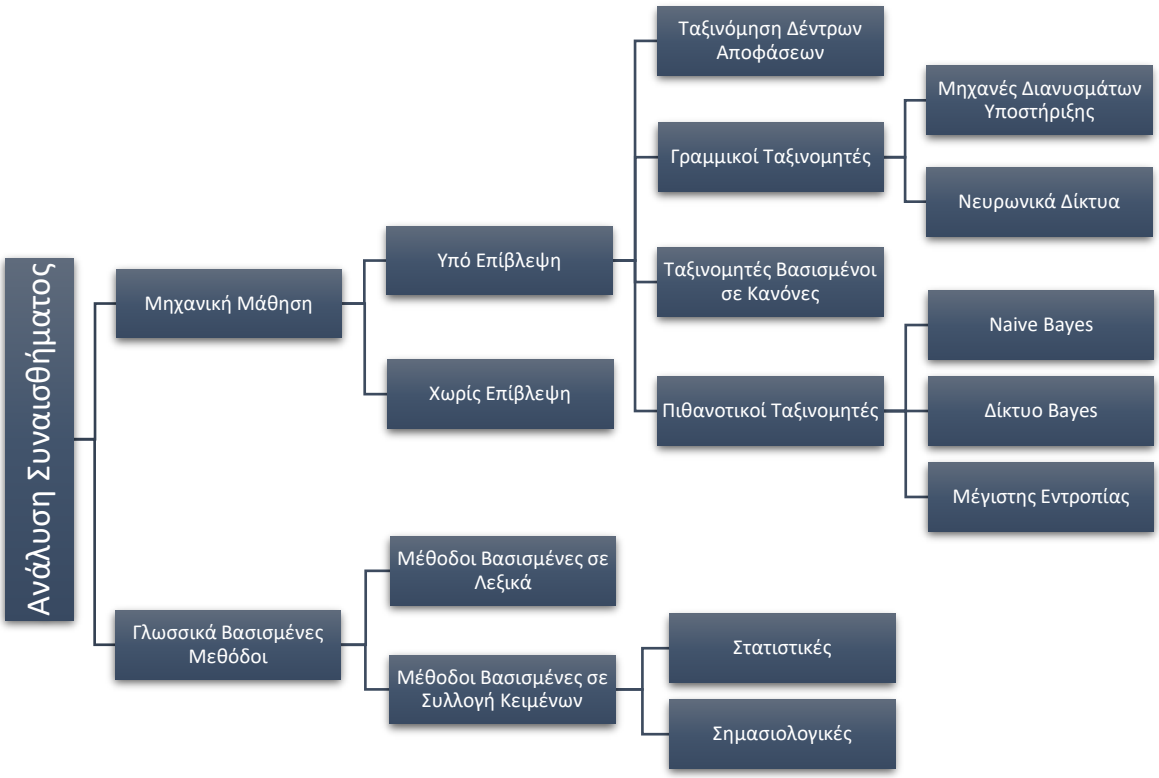


Σχήμα 2: Μοντέλο LDA

Ο αλγόριθμος LDA μπορεί να χρησιμοποιηθεί για την εξαγωγή των κρυμμένων θεμάτων σε συγκεκριμένη συλλογή κειμένων, αλλάί και για την δημιουργία μοντέλου ταξινόμησης κειμένων, όπου εκπαιδεύεται σε μία συλλογή κειμένων και στη συνέχει μπορεί να ταξινομήσει καινούρια κείμενα στα θέματα που έχει παράξει. Επίσης, μπορεί να χρησιμοποιηθεί και για Συνεργατικό φιλτράρισμα με την εκπαίδευση αντίστοιχου μοντέλου.

2.2 Ανάλυση Συναισθήματος

Στη συνέχεια θα παρουσιαστούν τεχνικές Ανάλυσης Συναισθήματος με βάση την εργασία (Medhat, Hassan, & Korashy, 2014). Στο Σχήμα 3 απεικονείζεται ο διαχωρισμός των τεχνικών ταξινόμησης συναισθήματος και στη συνέχεια θα περιγραφουν αναλυτικότερα.



Σχήμα 3: Ταξινόμηση Αλγορίθμων Ανάλυσης Συναισθήματος

2.2.1 Μέθοδος Μηχανικής Μάθησης

Οι μέθοδοι ταξινόμησης συναισθήματος που βασίζονται σε τεχνικές μηχανικής μάθησης απαιτούν μεγάλη ομάδα δεδομένων στα οποία εκπαιδεύονται ώστε να μπορούν να ταξινομήσουν καινούρια αντίστοιχα δεδομένα και χωρίζονται ανάλογα με την ανάγκη από τον αναλυτή για την λειτουργεία

τους. Οι υπό επίβλεψη μέθοδοι χρειάζονται χαρακτηρισμένα δεδομένα σε κλάσεις στα οποία εκπαιδεύονται. Η δημιουργία και εύρεση κατάλληλα χαρακτηρισμένων δεδομένων είναι δύσκολη, οπότε σε αυτήν την περίπτωση χρησιμοποιούνται μέθοδοι χωρίς ή με ήμι- επίβλεψη. Αυτοί οι αλγόριθμοι χρησιμοποιούν συνδυασμό μεθόδων όπως η ανάλυση των κειμένων σε επίπεδο προτάσεων ή λέξεων και υπολογισμός της ομοιότητας τους με βάση ειδικά λεξικά. Σε μία άλλη μέθοδο χρησιμοποιήθηκε ο αλγόριθμος LDA στον οποίο τα διαφορετικά συναισθήματα αντιμετωπίστηκαν σαν θέματα και επιτεύχθηκε η αναγνώριση και ταξινόμησή τους.

- Πιθανοτικοί Ταξινομητές: Είναι μοντέλα μίξης, όπου κάθε κλάση (συναίσθημα) θεωρείται ως μέρος του συνόλου μίξης και υπολογίζεται η πιθανότητά του στο σύνολο.

    - Naive Bayes: Ο αλγόριθμος λειτουργεί μετατρέποντας τα χαρακτηριστικά του κειμένου σε Bag-Of-Words, το οποίο αγνοεί την πληροφορία θέσης των λέξεων και θεωρεί ανεξάρτητα τα χαρακτηριστικά του κειμένου μεταξύ τους. Στη συνέχεια υπολογίζει την πιθανότητα τα χαρακτηριστικά αυτά να ανήκουν σε κάθε κλάση (συναίσθημα) με βάση τα χαρακτηρισμένα δεδομένα εισόδου. Τέλος, χρησιμοποιούνται οι πιθανότητες αυτές για την ταξινόμηση νέων δεδομένων.
    - Δίκτυο Bayes: Η διαφορά του με τον αλγόριθμο Naïve Bayes είναι ότι σε αυτόν τον αλγόριθμο τα χαρακτηριστικά του κειμένου θεωρούνται ως πλήρως εξαρτημένα τα μεταξύ τους. Έτσι δημιουργείται δίκτυο που αποτυπώνει την σχέση των χαρακτηριστικών και τις πιθανότητες των χαρακτηριστικών να ανήκουν στα συναισθήματα. Το αρνητικό αυτού του αλγορίθμου η υψηλή υπολογιστή ακρίβεια που απαιτεί.
    - Μέγιστης Εντροπίας: Ο αλγόριθμος λειτουργεί μετατρέποντας τα χαρακτηρισμένα δεδομένα εισόδου σε διανύσματα και υπολογίζοντας βάρη για αυτά τα διανύσματα. Στη συνέχεια, για την ταξινόμηση νέων δεδομένων, γίνεται αντίστοιχη μετατροπή τους σε διανύσματα και με τα υπολογισμένα βάρη γίνεται η ταξινόμησή τους στα συναισθήματα. Μεγάλο πλεονέκτημα αυτού του αλγορίθμου είναι ότι όταν υπάρχουν συλλογές κειμένων παράλληλα σε διαφορετικές γλώσσες, τα αποτελέσματα του αλγορίθμου για μία γλώσσα λειτουργούν και για τις άλλες.

- Γραμμικοί Ταξινομητές: Λειτουργούν μέσω τον γραμμικό διαχωρισμό του επιπέδου στο οποίο αποτυπώνονται οι συχνότητες των λέξεων στα κείμενα, με σκοπό τον διαχωρισμό των κλάσεων (συναισθημάτων)

    - Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines): Κατά τον αλγόριθμο αυτόν υπολογίζεται το επίπεδο του γραμμικού διαχωρισμού με τον μέγιστο διαχωρισμό των κλάσεων (συναισθημάτων). Παράγει καλά αποτελέσματα όταν τα δεδομένα είναι κείμενα, γιατί παρά την αραιή τους φύση τείνουν να χωρίζονται καλά γραμμικά.

- Νευρωνικά Δίκτυα: Δέχονται σαν είσοδο τις συχνότητες των λέξεων στα κείμενα, εκπαιδεύουν το δίκτυο των νευρώνων υπολογίζοντας τα βάρη κάθε νευρώνα με σκοπό την βελτιστοποίηση της ταξινόμησης των συναισθημάτων (εξόδων) των δεδομένων.

- Δέντρα Αποφάσεων: Λειτουργούν με ιεραρχική κατάταξη αποφάσεων (κανόνων) που χωρίζουν τα δεδομένα εισόδου με σκοπό τα τελικά φύλλα να περιλαμβάνουν έναν ελάχιστο αριθμό δεδομένων και α ταξινομούνται σε μία μόνο κλάση (συναίσθημα). Οι κανόνες αυτοί χρησιμοποιούνται για την ταξινόμηση νέων δεδομένων.

- Ταξινομητές Βασισμένοι σε Κανόνες: Δημιουργούν κανόνες με βάση την ύπαρξη χαρακτηριστικών κειμένου στα δεδομένα, τα οποία να αντιστοιχούν σε κάποια κλάση. Για καθένα από τα χαρακτηριστικά αυτά υπολογίζονται οι δείκτες support δηλαδή ο αριθμός εμφάνισης του χαρακτηριστικού στα δεδομένα και confidence δηλαδή το ποσοστό παρουσίας του χαρακτηριστικού στην κλάση.

2.2.2 Γλωσσικά Βασισμένες Μέθοδοι

- Μέθοδοι Βασισμένες σε Λεξικά: Στις μεθόδους αυτές γίνεται αρχικά η επιλογή λέξεων από τα δεδομένα που να προσδίδουν ισχυρό συναίσθημα και μέσω μοντέλων όπως το WordNet και thesaurus γίνεται εύρεση νέων λέξεων συνωνύμων ή ανωνύμων των αρχικών. Η διαδικασία αυτή συνεχίζεται έως ότου να συμπεριληφθούν όλες οι λέξεις των δεδομένων στο λεξικό, και στη συνέχεια πραγματοποιείται η απόδοση συναισθήματος.

- Μέθοδοι Βασισμένες σε Μεγάλες Συλλογές Κειμένων: Οι μέθοδοι αυτές λύνουν το πρόβλημα αναγνώρισης συναισθήματος που απευθύνεται σε συγκεκριμένο χαρακτηριστικό. Λειτουργούν με την αναγνώριση συντακτικών μοτίβων και λέξεων συναισθήματος στις συλλογές κειμένων. Παράλληλα ελέγχουν τη συνέπεια συναισθήματος από συνδετικές λέξεις οι άλλες που να αντιστρέφουν το νόημα που τις ακολουθεί.

    - Στατιστικές Μέθοδοι: Κατά τις μεθόδους αυτές γίνεται στατιστική ανάλυση των λέξεων σε μεγάλες συλλογές κειμένων που τους έχει αποδοθεί πολικότητα συναισθήματος (αρνητικό-ουδέτερο-θετικό) και με βάση τη στατιστική σχέση των λέξεων με τα κείμενα αλλά και των λέξεων μεταξύ τους γίνεται απόδοση πολικότητας στις λέξεις.
    - Σημασιολογικές Μέθοδοι: Χρησιμοποιούν τις συλλογές κειμένων για να δημιουργήσουν πλαίσιο υπολογισμού σημασιολογικής ομοιότητας μεταξύ των λέξεων με μοντέλα όπως το WordNet. Με αυτόν τον τρόπο γίνεται αντιστοίχιση

συνωνύμων λέξεων σε γνωστές ή άγνωστες λέξεις και με την χρήση λεξικών συναισθήματος πραγματοποιείται η ταξινόμηση συναισθήματος.

2.3 Lexalytics / Semantria

Για την εργασία αυτή, στην ανάλυση συναισθήματος χρησιμοποιήθηκε η πλατφόρμα Semantria της Lexalytics. Θα αναλυθούν συνοπτικά παρακάτω οι μέθοδοι που χρησιμοποιούνται με βάση την παρουσίαση από την σελίδα (Lexalytixs\Semantria, n.d.).

- Πραγματοποιείται Ανάλυση συναισθήματος σε Επίπεδο Κειμένου, σε Επίπεδο Χαρακτηριστικού αλλά και σε Επίπεδο Πρόταση-Φράσεις τις οποίες χωρίζει ακολουθώντας ειδική μεθοδολογία
- Γλωσσικά Βασισμένες Μέθοδοι: Λειτουργούν σε όλα τα επίπεδα (Κειμένου, Πρότασης-Φράσης και Χαρακτηριστικού)
    - Απόδοση στις λέξεις και φράσεις ως μέρος του λόγου (POS tagging) και χρήση αυτών σε ενισχυμένο λεξικό συναισθήματος που λαμβάνει υπόψιν την πληροφορία pos tag και αποδίδει στις λέξεις/φράσεις βαθμό συναισθήματος από -1 έως +1.
    - Αναγνώριση λέξεων που τροποποιούν την ισχύ ή κατεύθυνση του συναισθήματος και απόδοση τιμή πολλαπλασιαστή σε αυτές.
    - Άθροισμα του συνολικού συναισθήματος του κάθε κειμένου.
- Μέθοδοι Μηχανικής Μάθησης: Λειτουργούν μόνο σε Επίπεδο Κειμένου
    - Χρήση μοντέλου που έχει εκπαιδευτεί σε κείμενα που είχαν χαρακτηριστεί ως προς το συναίσθημα, για απόδοση βαθμού συναισθήματος στα νέα κείμενα. Λόγω της αδυναμίας αναγνώρισης χαρακτηριστικών του κειμένου λειτουργεί μόνο στο επίπεδο κειμένου.
- Κατηγοριοποίηση και Εξόρυξη Χαρακτηριστικών: Γίνεται αναγνώριση της αναφοράς στα κείμενα και κατηγοριοποίησή τους τόσο Γενικών, πιο συμπεριληπτικών χαρακτηριστικών, όσο και συγκεκριμένων και πιο ειδικών χαρακτηριστικών.
    - Αναγνώριση μέσω ερωτήσεων (Queries): Αναγνώριση χαρακτηριστικών με εξειδικευμένη αναζήτηση. Προσφέρεται η δυνατότητα επιλογής και τροποποίησης από τον χρήστη.
    - Κατηγορίες (Categories): Αναγνώριση των σημαντικών χαρακτηριστικών σε ένα κείμενο. Εκτός από το ίδιο το χαρακτηριστικό, επιστρέφεται και ο βαθμός σημαντικότητάς του στο σύνολο του κειμένου και έχει τη δυνατότητα γενίκευσης , αλλά λειτουργεί καλά σε μεγάλα κείμενα όπου προσφέρεται περισσότερη πληροφορία. . Προσφέρεται η δυνατότητα επιλογής και τροποποίησης από τον χρήστη.
    - Αυτόματες Κατηγορίες: Γίνεται αναγνώριση της αναφοράς σε χαρακτηριστικά με βάση τις κατηγορίες της Wikipedia. Έχει πολλά είδη και μεγάλο εύρος στα χαρακτηριστικά που μπορεί να αναγνωρίσει αλλά δεν υπάρχει η δυνατότητα επιλογής και τροποποίησης από τον χρήστη.

- Μηχανική Μάθηση: Εκπαιδευμένο μοντέλο που αναγνωρίζει θέματα αναφοράς με βάσει του συνόλου κειμένων στα οποία εκπαιδεύτηκε. Αναγνωρίζει μεγάλο εύρος και με δυνατότητα επιλογής πολύ ειδικευμένων σημείων αναφοράς, αλλά εξαρτάται σε μεγάλο βαθμό από τα κείμενα εκπαίδευσης και δεν προσφέρεται στον χρήστη καμία επιλογή τροποποίησης.
- Αναγνώριση Τίτλων Φορέων: Δυνατότητα Αναγνώρισης Τίτλων Φορέων όπως ονόματα εταιριών, ανθρώπων κ.α.. Προσφέρεται η δυνατότητα επιλογής αι τροποποίησης από τον χρήστη.
- Αναγνώριση Θεμάτων (Themes): Με βάση γλωσσικών κανόνων που βασίζονται σε επιλεγμένα μοτίβα σε μέρη του λόγου (pos tag) γίνεται αναγνώριση λέξεων ή φράσεων κλειδιά. Αυτές οι λέξεις/φράσεις κλειδιά ανήκουν σε προτάσεις του κειμένου και γίνεται έλεγχος της σχέσης τους με όλο το κείμενο όπου υπολογίζεται η σημαντικότητά τους στο νόημα του κειμένου. Επίσης υπολογίζεται το συναίσθημα που αναφέρεται στα συγκεκριμένα θέματα.

2.4 Πολυκριτήρια Ανάλυση

Όπως περιγράφεται στο βιβλίο των στο άρθρο των (Siskos, Grigoroudis, & Matsatsinis, 2016) και στο βιβλίο των (Jacquet-Lagreze & Siskos, 1982), στην θεωρεία αποφάσεων όταν εμπλέκονται πολλά κριτήρια εμφανίζεται το πρόβλημα της τελικής απόφασης και οι μεθοδολογίες που οδηγούν σε αυτήν ανήκουν στην οικογένεια της Πολυκριτήρια Ανάλυση. Οι βασικές κατηγορίες μεθόδων προσέγγισης του προβλήματος της Πολυκριτήριας Ανάλυσης αποτελούν μέθοδοι ή μοντέλα που επιτρέπουν τον συμψηφισμό των διαφορετικών κριτηρίων και οδηγούν στις δράσεις πάνω στις διαφορετικές εναλλακτικές, ή μέθοδοι που οδηγούν στην τελική απόφαση με την ενεργή συμμετοχή του Αποφασίζοντα.

Παρακάτω θα παρουσιαστούν δύο αλγόριθμοι Πολυκριτήριας Ανάλυσης (UTASTAR και TOPSIS) που χρησιμοποιήθηκαν στην εργασία για την κατάταξη των εναλλακτικών των ξενοδοχείων. Και οι δύο αλγόριθμοι ανήκουν στην πρώτη κατηγορία, δηλαδή χρησιμοποιούν μεθόδους συμψηφισμού των διαφορετικών ώστε να καταλήξουν σε συμπεράσματα για τα βάρη των κριτηρίων και την κατάταξη των εναλλακτικών αντίστοιχα.

2.4.1 UTASTAR

Ο αλγόριθμος UTASTAR προτάθηκε από τους (Siskos & Yannacopoulos, 1985) και αποτελεί μία βελτιωμένη έκδοση της μεθόδου UTA.

Η μέθοδος UTA προτάθηκε από τους (Jacquet-Lagreze & Siskos, 1982) με σκοπό την εξαγωγή συναρτήσεων προστιθέμενης αξίας για τα κριτήρια από μία δοσμένη κατάταξη εναλλακτικών με τιμές στα συγκεκριμένα κριτήρια. Το μοντέλο σύνθεσης των κριτηρίων στη μέθοδο UTA είναι το εξής:

$$u(g) = \sum_{i=1}^{n} p_i u_i(g_i)$$

(3)

το οποίο υπόκειται στους περιορισμούς κανονικοποίησης:

$$\begin{cases} \sum_{i=1}^{n} p_i = 1 \\ u_i(g_{i*}) = 0\,, \qquad u_i(g_i^*) = 1, \qquad \forall \mathrm{i} = 1,2,\dots,\mathrm{n} \end{cases}$$

(4)

όπου u_i=1,2,….n είναι αύξουσες συναρτήσεις των g_i που καλούνται συναρτήσεις μερικής αξίας κανονικοποιημένες στο διάστημα [0,1] και p_i τα βάρη των κριτηρίων.

Οι συναρτήσεις μερικής αξίας καθώς και ολικής αξίας έχουν ίδια μονοτονία με τα κριτήρια στα οποία αντιστοιχούν και στην περίπτωση των συναρτήσεων ολικής αξίας υπόκεινται στις παρακάτω ιδιότητες:

$$\begin{cases} u[g(a)] > u[g(a)] \Leftrightarrow a > b\ (\pi\rho o\tau\acute{\iota}\mu\eta\sigma\eta) \\ u[g(a)] = u[g(a)] \Leftrightarrow a \sim b\ (\alpha\delta\iota\alpha\varphi o\rho\acute{\iota}\alpha) \end{cases}$$

(5)

Η μέθοδος UTA λαμβάνει μία μορφή της συνάρτησης προστιθέμενης αξίας χωρίς βάρη ισοδύναμη με την (3), η οποία προϋποθέτει την προτιμησιακή ανεξαρτησία των κριτηρίων για τον αποφασίζοντα και είναι ως εξής:

$$u(g) = \sum_{i=1}^{n} u_i(g_i)$$

(6)

το οποίο υπόκειται στους περιορισμούς κανονικοποίησης:

$$\begin{cases} \sum_{i=1}^{n} u_i(g_i^*) = 1 \\ u_i(g_{i^*}) = 0\,, \qquad \forall \mathrm{i} = 1,2,\dots,\mathrm{n} \end{cases}$$

(7)

Με βάση το μοντέλο προσθετικής αξίας και μέσω των κανόνων σχέσης προτίμησης, η ολική αξία κάθε εναλλακτικής a ∈ A_R εκφράζεται ως εξής:

$$u'[g_i(a)] = \sum_{i=1}^{n} u_i[g_i(a)] + \sigma(\alpha), \qquad \forall\, \alpha \in A_R$$

(8)

όπου σ(α) είναι το ενδεχόμενο σφάλμα της συνάρτησης αξίας $u'[g_i(a)]$ το οποίο εισάγεται με σκοπό την ελαχιστοποίησή του, κατά την λογική του γραμμικού προγραμματισμού, ώστε να ελαχιστοποιηθεί και η διασπορά των σημείων της μονότονης καμπύλης. Για τον υπολογισμό των αντίστοιχων συναρτήσεων μερικής αξίας σε μία κατά τμήματα γραμμική μορφή, όπως προτάθηκε από τους (Jacquet-Lagreze, E., & Siskos, J. (1982). Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. European journal of operational research, 10(2), 151-164, χρησιμοποιείται η γραμμική παρεμβολή. Οπότε, για κάθε κριτήριο το διάστημα [g_{i*}, g_i^*] χωρίζεται σε (a_i - 1) ίσα διαστήματα με τα άκρα g_i^j να δίνονται από τη σχέση:

$$g_i^j = g_{i^*} \frac{j-1}{a_i - 1}(g_i^* - g_{i^*}),\ \forall\, j = 1,2,\dots,a_i$$

(9)

Στη συνέχεια, η μερική αξία μίας εναλλακτικής a υπολογίζεται κατά προσέγγιση με γραμμική παρεμβολή,

$$u_i[g_i(a)] = u_i\left(g_i^j\right) + \ g_{i^*} \frac{g_i(a) - g_i^j}{g_i^{j+1} - g_i^j}\left[u_i\left(g_i^{j+1}\right) - u_i\left(g_i^j\right)\right],\ \ \gamma\iota\alpha\ g_i(a) \in [g_i^j - g_i^{j+1}]$$

(10)

Ο τρόπος με τον οποίο κατατάσσονται οι εναλλακτικές a του συνόλου A_R, είναι σε σειρά ανάλογα με την συνολική βαθμολογία τους σε όλα τα κριτήρια $u'[g_i(a)]$. Ο κανόνας σύγκρισης κάθε ζεύγους εναλλακτικών (a_k, a_{k+1}) θα καταλήγει σε προτίμηση (a_k< ή > a_{k+1}) ή αδιαφορία (a_k ~ a_{k+1}). Αυτό επιτυγχάνεται μέσω του ορισμού ελάχιστου θετικού κατωφλιού δ με βάση το οποίο θα διακρίνεται η αδιαφορία από την προτίμηση. Οι κανόνες αυτοί παρουσιάζονται παρακάτω:

Η διαφορά των ολικών αξιών των εναλλακτικών a_k, a_{k+1}:

$$\Delta(\alpha_k,\ \alpha_{k+1}) = \ u'[g_i(a_k)] - u'[g_i(a_{k+1})]$$

(11)

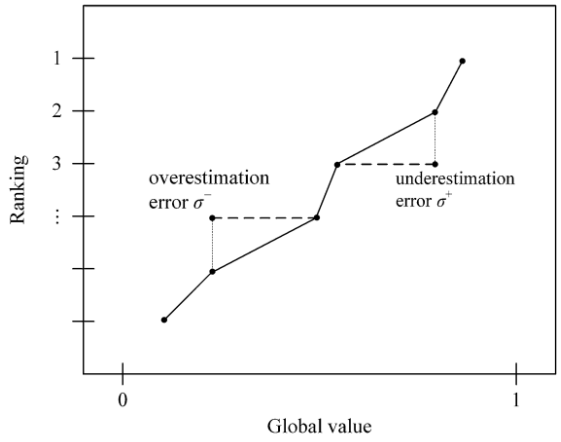
Η σύγκριση των εναλλακτικών a_k, a_{k+1}:

$$\begin{cases} \Delta(\alpha_k,\ \alpha_{k+1}) \geq \delta \quad \Rightarrow \ \alpha_k > \alpha_{k+1}\ (\pi\rho o\tau\acute{\iota}\mu\eta\sigma\eta) \\ \Delta(\alpha_k,\ \alpha_{k+1}) = 0 \Rightarrow \ \alpha_k \sim \alpha_{k+1}\ (\alpha\delta\iota\alpha\phi o\rho\acute{\iota}\alpha) \end{cases}$$

(12)

Η μέθοδος UTASTAR αποτελεί βελτιωμένη έκδοση της αρχικής μεθόδου UTA που αναλύθηκε παραπάνω.

Αυτό που διαφοροποιεί την UTASTAR είναι η αλλαγή του υπολογισμού του σφάλματος που από ένα μοναδικό θετικό σφάλμα σ(α) των εναλλακτικών $\alpha \in A_R$, χρησιμοποιήθηκε ένα διπλό σφάλμα $\sigma^+(\alpha)$ και $\sigma^-(\alpha)$ τα οποία προσφέρουν καλύτερη εκτίμηση της αξίας της εναλλακτικής μειώνοντας την υποτίμηση ή υπερτίμηση. Αυτό επιτυγχάνεται μέσω της δυνατότητας που προσφέρει το διπλό σφάλμα προσέγγισης σημείων δεξιά και αριστερά της καμπύλης, από την πρόσθεση/αφαίρεση μίας ποσότητας αξίας χωρίς να επηρεαστούν οι άλλες αξίες, όπως φαίνεται στο παρακάτω γράφημα:



Σχήμα 4: Απεικόνιση Υπολογισμού Ολικής Χρησιμότητας UTASTAR

Η συνάρτηση υπολογισμού των ολικών αξιών των εναλλακτικών στη μέθοδο UTASTAR είναι:

$$u'[g(a)] = \sum_{i=1}^{n} u_i[g_i(a)] - \sigma^+(\alpha) + \sigma^-(\alpha), \qquad \forall\, \alpha \in A_R$$

(13)

Μία ακόμα διαφορά της UTASTAR αποτελεί η χρήση των θετικών μεταβλητών μονοτονίας των κριτηρίων w_{ij} οι οποίες χρησιμοποιούνται στο μετασχηματισμό μεταβλητών και υπολογίζονται ως εξής:

$$w_{ij} = \ u_i\left(g_i^{j+i}\right) - u_i\left(g_i^{j}\right) \ \geq 0, \qquad \forall\, i = 1,2,..,n\ \kappa\alpha\iota\ j = 1,2,..,a_i - 1$$

(14)

Στη συνέχεια παρουσιάζονται συνοπτικά τα βήματα για την εφαρμογή της UTASTAR:

Βήμα 1: Έκφραση των ολικών αξιών ως σχέση των μερικών αξιών και των μεταβλητών μονοτονίας:

$$\begin{cases} u_i(g_i^1) = 0\,, & \forall\, \mathrm{i} = 1,2,\dots,\mathrm{n} \\ u_i(g_i^j) = \sum_{i=1}^{j-1} w_{it} = 1\,, & \forall\, \mathrm{i} = 1,2,\dots,\mathrm{n}\ \kappa\alpha\iota\ \forall\, j = 1,2,\dots,a_i - 1 \end{cases}$$

(15)

Βήμα 2: Υπολογισμός της διαφοράς αξίας των εναλλακτικών του A_R ανά δύο με την εισαγωγή διπλού σφάλματος:

$$\Delta(\alpha_k,\, \alpha_{k+1}) = u[g(a_k)] - \sigma^+(\alpha) + \sigma^-(\alpha) - u[g(a_{k+1})] + \sigma^+(\alpha) - \sigma^-(\alpha)$$

(16)

Βήμα 3: Λύση του Γραμμικού Προβλήματος, με τον ορισμό του δ ως μικρή θετική ποσότητα:

$$\begin{cases} [min]_Z = \sum_{k=1}^{m} [\sigma^+(\alpha_k) + \sigma^-(\alpha_k)] \\ \upsilon\pi\acute{o}: \\ \Delta(\alpha_k,\, \alpha_{k+1}) \geq \delta\ \alpha\nu\ \alpha_k > \alpha_{k+1},\ \ \forall\, k \\ \Delta(\alpha_k,\, \alpha_{k+1}) = 0\ \alpha\nu\ \alpha_k \sim \alpha_{k+1},\ \ \ \forall\ k \\ \sum_{i=1}^{n} \sum_{j=1}^{a_i - 1} w_{ij} = 1 \\ w_{ij} \geq 0,\ \ \sigma^+(a_k) \geq 0, \sigma^-(a_k) \geq 0,\ \ \forall\ i,j,k \end{cases}$$

(17)

Βήμα 4: Εφαρμογή μεταβελτιστοποίησης, μέσω του ελέγχου κοντινών λύσεων του γραμμικού προβλήματος, υπολογίζοντας το βαρύκεντρο των προσθετικών συναρτήσεων αξίας που μεγιστοποιούν τις παρακάτω αντικειμενικές συναρτήσεις:

$$u_i(g_i^*) = \sum_{j=1}^{a_i - 1} w_{ij}\,, \qquad \forall\, \mathrm{i} = 1,2,\dots,\mathrm{n}$$

(18)

Στο πολύεδρο των περιορισμών του γραμμικού προβλήματος (17) επιβάλλεται επιπλέον περιορισμός:

$$\sum_{k=1}^{m}[\sigma^{+}(\alpha_k)+\sigma^{-}(\alpha_k)] \leq z^{*}+\varepsilon$$

(19)

όπου z^* είναι η βέλτιστη λύση του γραμμικού προβλήματος μετά το Βήμα 3, και ε είναι ένας πολύ μικρός θετικός αριθμός.

Μετά από συγκριτική ανάλυση σε ένα σύνολο πειραματικών δεδομένων οι Siskos & Yannacopoulos απέδειξαν ότι η μέθοδος UTASTAR προσφέρει καλύτερα αποτελέσματα από το μοντέλο της UTA σε ένα πλήθος δεικτών, όπως τον αριθμό των ελάχιστων απαραίτητων επαναλήψεων της simplex, τον δείκτη του Ταφ του Kendall και τη διασπορά του συνόλου των σφαλμάτων των κριτηρίων.

2.4.2 TOPSIS

Ο αλγόριθμος TOPSIS όπως αναλύεται στην εργασία (Lai, Liu, & Hwang, 1994) ανήκει στην κατηγορία των τεχνικών Πολυκριτήριας Ανάλυσης που δεν απαιτούν πρότερη πληροφορία προτίμησης για την κατάταξη των εναλλακτικών. Η βασική ιδέα, σύμφωνα με την οποία πραγματοποιεί την κατάταξη των εναλλακτικών, χρησιμοποιεί ένα σημείο αναφοράς το οποίο υπολογίζεται ως η ιδανική εναλλακτική για τα δεδομένα του προβλήματος και στη συνέχεια πραγματοποιείται η κατάταξη με βάση την απόσταση των εναλλακτικών από την ιδανική λύση. Για τον υπολογισμό της απόστασης υπάρχουν διαφορετικές προσεγγίσεις όπως στον προγραμματισμό στόχου όπου χρησιμοποιείται η μέθοδος σταθμισμένου μέσου με απόλυτες τιμές ενώ στην προσέγγιση ολικών κριτηρίων χρησιμοποιείται η μέθοδος Minslowski L_p κατά την οποία η απόσταση d_p από την ιδανική λύση f^* υπολογίζεται ως εξής:

$$d_p = \left\{\sum_{t=1}^{k}(f_t^{*}-f_t)^{p}\right\}^{1/p}$$

(20)

όπου $p \geq 1$.

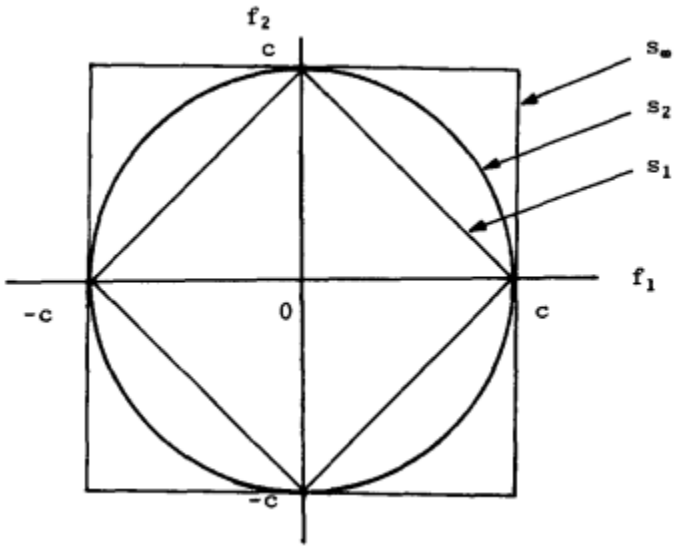
Η επιλογή της τιμής του p είναι πολύ σημαντική γιατί από αυτήν κρίνεται ο τύπος της απόστασης όπου για p=1 αντιστοιχεί η απόσταση d_1 (Manhattan), για p=2 η απόσταση d_2 (Ευκλείδεια) ενώ για $p=\infty$ αντιστοιχεί η απόσταση d_∞ (Tchebycheff). Όμως, επειδή είναι αδύνατο να υπολογιστούν απευθείας οι αποστάσεις αυτές χρησιμοποιείται η κανονικοποιημένη εκδοχή η οποία υπολογίζεται ως εξής:

$$d_p = \left\{\sum_{t=1}^{k}[\frac{f_t^* - f_t}{f_t^*}]^p\right\}^{1/p}$$

(21)

, με p≥1.

Οι αποστάσεις και πάλι εξαρτώνται από την επιλογή του p=1,2 ή ∞ και η αναπαράσταση των καμπύλων των αποστάσεων για τις τιμές του p στο δισδιάστατο χώρο φαίνεται στο παρακάτω γράφημα:



Σχήμα 5: Απεικόνιση υπολογισμού αποστάσεων για την κατάταξη TOPSIS

Κάποιες άλλες παραλλαγές της οικογένειας των μεθόδων της TOPSIS, που χρησιμοποιούνται στην κατάταξη των εναλλακτικών, βασίζονται στη χρήση διαφορετικού σημείου αναφοράς.

Ένας τρόπος (PIS) χρησιμοποιεί την ιδανική λύση ως σημείο αναφοράς και βαθμολογεί τις εναλλακτικές με βάση την μικρότερη σταθμισμένη απόσταση από την ιδανική λύση. Αυτός ο τρόπος δίνει προτεραιότητα στην μεγιστοποίηση του οφέλους. Ενώ η παραλλαγή (NIS) χρησιμοποιεί ως σημείο αναφοράς την χειρότερη δυνατή λύση και βαθμολογεί τις εναλλακτικές με βάση την μεγαλύτερη σταθμισμένη απόσταση από αυτήν. Αυτή η παραλλαγή δίνει προτεραιότητα στην ελαχιστοποίηση του κόστους. Υπάρχει επίσης και η παραλλαγή που συμβιβάζει τους δύο αυτούς τρόπους, καθώς προκύπτουν συχνά διαφορετικά αποτελέσματα. Η συμβιβαστική αυτή εναλλακτική προτάθηκε από τους Hwang και Yoon και υπολογίζει την απόσταση ως εξής:

$$d_p = \left\{\sum_{t=1}^{k}[\frac{f_t^* - f_t(x)}{f_t^* - f_t^-}]^p\right\}^{1/p}$$

(22)

με $p \geq 1$

Όπου η ιδανική λύση (PIS) είναι f^* και η χειρότερη λύση (NIS) είναι f^-.

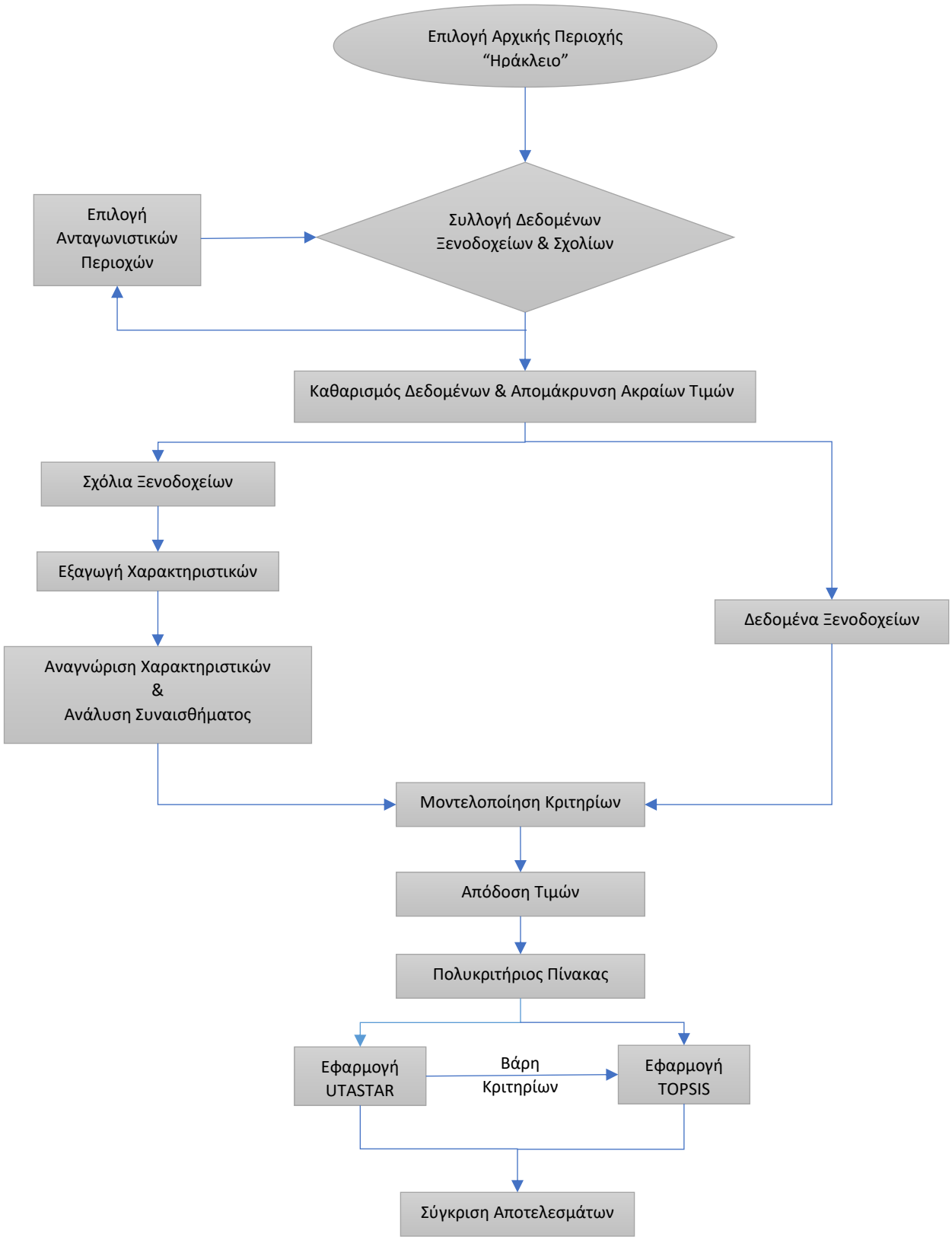
Οι διαφορετικές αυτές παραλλαγές είναι στην διάθεση του αναλυτή για να αξιολογήσει τα αποτελέσματά τους και να επιλέξει την κατάταξη που εκφράζει καλύτερα το πρόβλημα και τον αποφασίζοντα.

Κεφάλαιο 3: Προτεινόμενη Μεθοδολογία

3.1 Εισαγωγή

Στο Σχήμα 7 παρουσιάζεται η αρχιτεκτονική της μεθοδολογίας που ακολουθήθηκε στην εργασία και στο Σχήμα 6 ομαδοποιημένα σε τίτλους. Παρακάτω αναφέρονται συνοπτικά αυτά με την σειρά που πραγματοποιήθηκαν. Αρχικά πραγματοποιείται η συλλογή των δεδομένων, η οποία αποτελείται από δύο βήματα. Το πρώτο βήμα είναι η συλλογή των δεδομένων των Ξενοδοχείων και των σχολίων τους στα αγγλικά από το TripAdvisor για την αρχική περιοχή (στην συγκεκριμένη εργασία επιλέχθηκε το Ηράκλειο) και στο δεύτερο βήμα αναλύονται τα στοιχεία των σχολιαστών και γίνεται εύρεση των πιο συχνών προορισμών πέρα από τον αρχικό στους οποίους σχολιάζουν, και πραγματοποιείται η συλλογή των δεδομένων των Ξενοδοχείων και των Σχολίων τους, των ανταγωνιστικών περιοχών. Στη συνέχεια, πραγματοποιείται καθαρισμός των δεδομένων και η απομάκρυνση των ακραίων τιμών. Ακολουθεί η εφαρμογή αλγορίθμων στα σχόλια για την εξαγωγή των χαρακτηριστικών αρχικά (Αλγόριθμος LDA) και στη συνέχεια η αναγνώριση αναφοράς των χαρακτηριστικών και ανάλυση συναισθήματος στην αναφορά των χαρακτηριστικών (Πλατφόρμα Lexalytics/Semantria). Με αυτόν τον τρόπο εξάγεται η γνώμη των σχολίων πάνω στα επιλεγμένα χαρακτηριστικά. Έπειτα τα δεδομένα ανάλυσης των σχολίων και τα δεδομένα των ξενοδοχείων συνδυάζονται για την Μοντελοποίηση Κριτηρίων. Η επιλογή των κριτηρίων στηρίχθηκε στα χαρακτηριστικά που εξήχθησαν από τα σχόλια, τα οποία αποτελούν τα χαρακτηριστικά των ξενοδοχείων που δίνουν περισσότερη σημασία οι σχολιαστές. Με βάση την μοντελοποίηση κριτηρίων και τα διαθέσιμα δεδομένα έγινε η απόδοση τιμών για την δημιουργία του Πολυκριτήριου Πίνακα ο οποίος υπολογίσθηκε για τα ξενοδοχεία για να είναι κατάλληλος ώστε να εφαρμοστεί η Πολυκριτήρια Ανάλυση UTASTAR από την οποία έγινε εξαγωγή των βαρών των κριτηρίων. Τα βάρη αυτά μαζί με τον Πολυκριτήριο Πίνακα χρησιμοποιήθηκαν στην πολυκριτήρια ανάλυση TOPSIS. Τέλος, έγινε σύγκριση των αποτελεσμάτων των μεθόδων UTASTAR και TOPSIS και ανάλυση της αγοράς των ξενοδοχείων των περιοχών που μελετήθηκαν. Στη συνέχεια του κεφαλαίου αναπτύσσονται αναλυτικά όλα τα βήματα.

Σχήμα 6: Παρουσίαση Μεθοδολογίας ανά κεφάλαιο



Σχήμα 7: Παρουσίαση Αρχιτεκτονικής Μεθοδολογίας

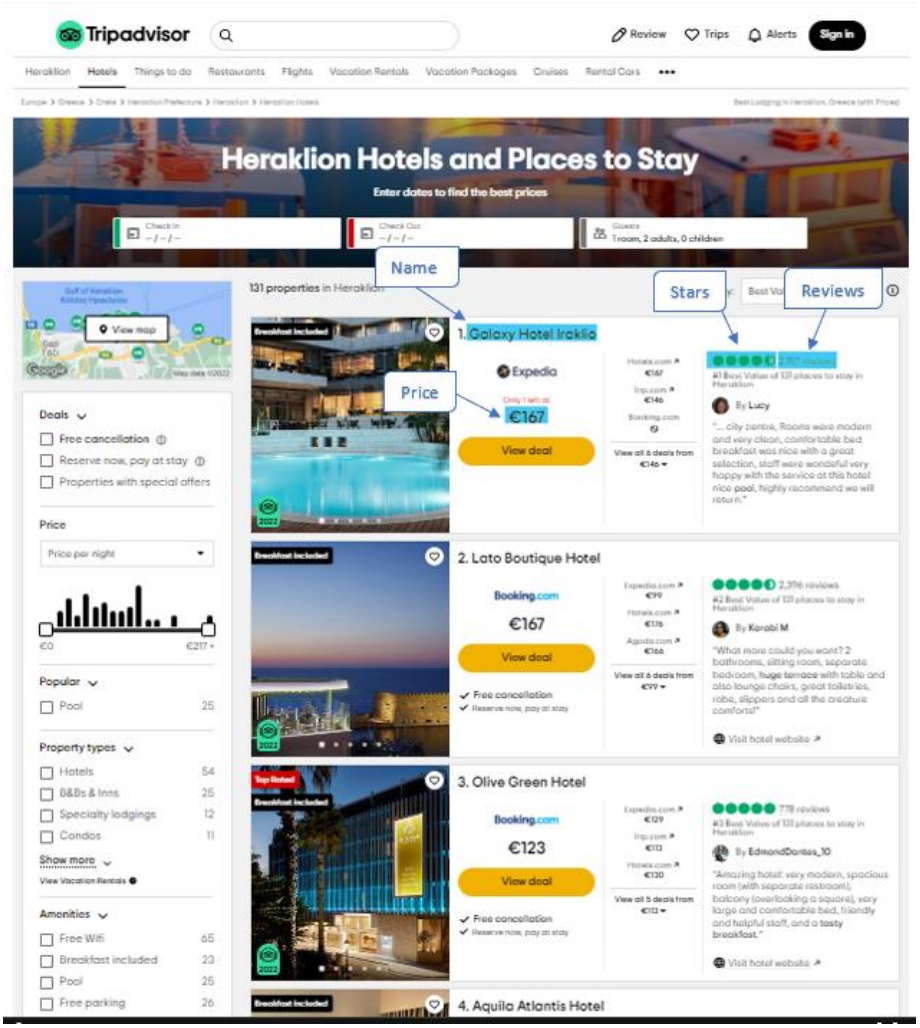
3.2 Συλλογή Δεδομένων

Για τη συλλογή δεδομένων των ξενοδοχείων και των σχολίων τους επιλέχθηκε η ιστοσελίδα TripAdvisor, ως η δημοφιλέστερη σελίδα τουριστικών αξιολογήσεων. Ως βάση τουριστικού προορισμού και είδους ξενοδοχείων επιλέχθηκε το Ηράκλειο, σύμφωνα με το οποίο θα επιλεχθούν ανταγωνιστικές περιοχές και ξενοδοχεία.

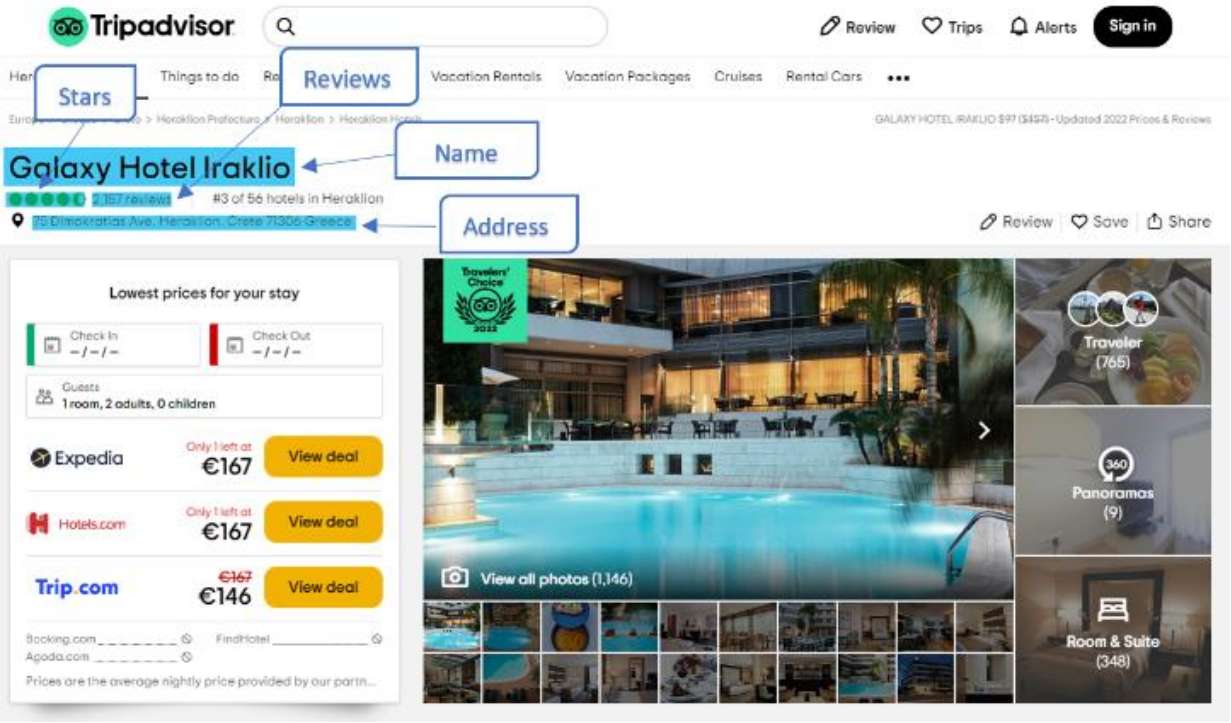
Αρχικά πραγματοποιήθηκε η συλλογή των δεδομένων των ξενοδοχείων και των σχολίων τους με τη μέθοδο Web Crawling μέσω Python από την σελίδα TripAdvisor.com για τον νομό του Ηρακλείου. Επιλέχθηκαν μόνο τα σχόλια στα αγγλικά όπως τα χωρίζει το TripAdvisor για να είναι κατάλληλα για την ανάλυση συναισθήματος μετά, αποτελούσαν εξάλλου την μεγαλύτερη πλειοψηφία.

Η πλοήγηση στην ιστοσελίδα και τα δεδομένα που συλλέχθηκαν από αυτή παρουσιάζονται στο Σχήμα 8 με τις αντίστοιχες εικόνες

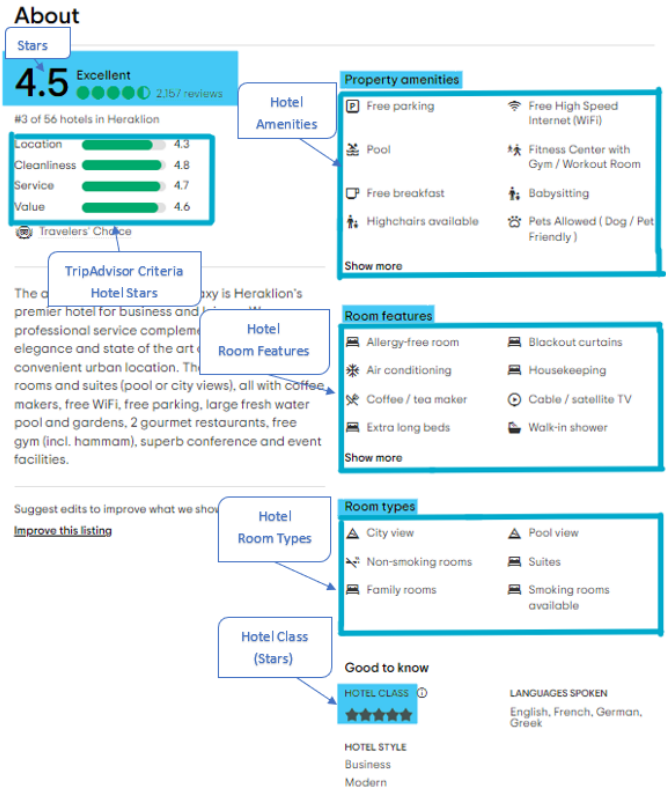
TripAdvisor Πλοήγηση ανά Περιοχή



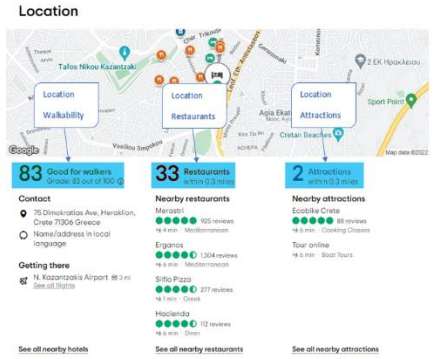
Επιλογή Ξενοδοχείου



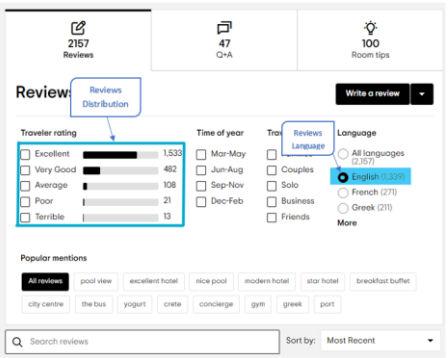
Πληροφορίες Παροχών Ξενοδοχείου



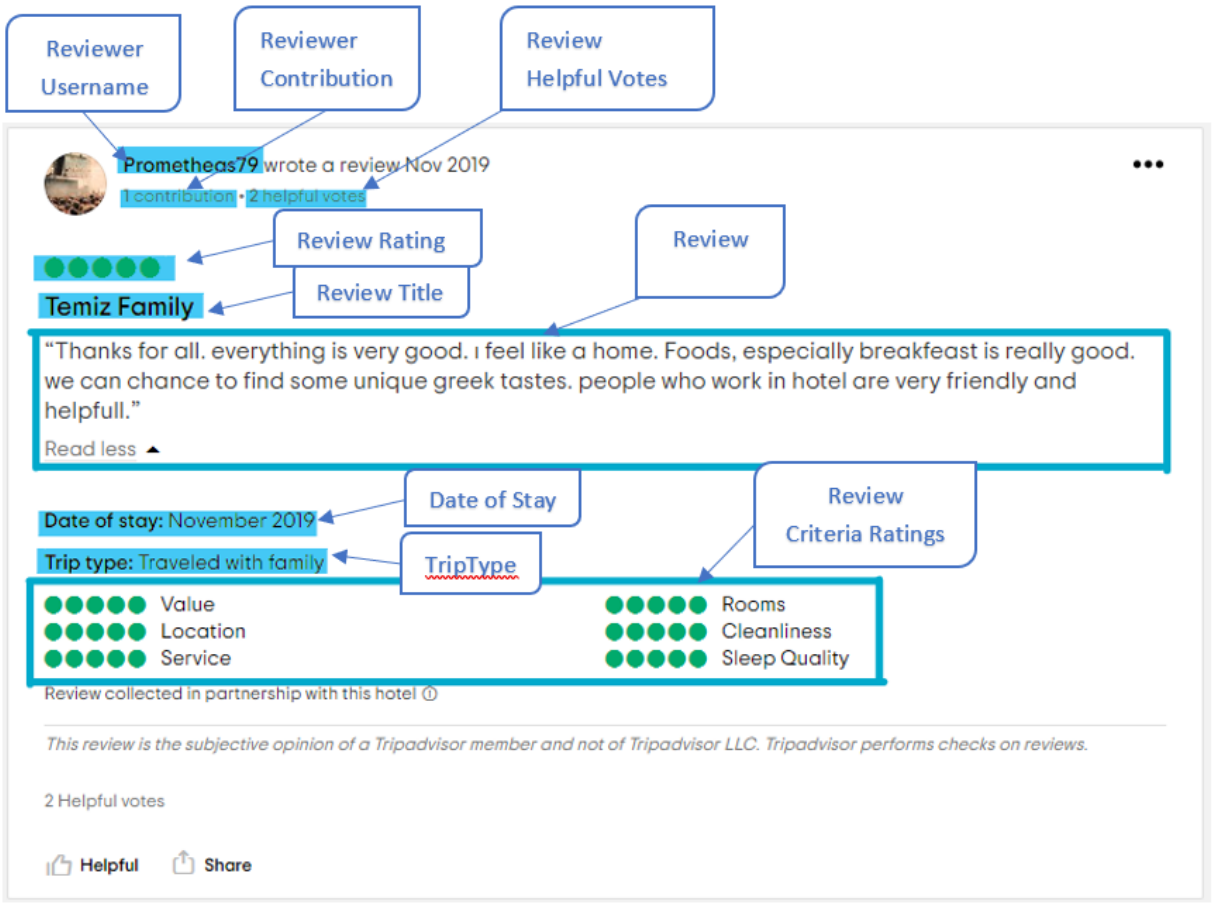
Πληροφορίες Τοποθεσίας Ξενοδοχείου



Επιλογές Σχολίων



Σχόλιο



Σχήμα 8: Εικόνες παρουσίασης πλοήγησης στα ξενοδοχεία στο TripAdvisor & παρουσίαση των δεδομένων που συλλέχθηκαν

Στη συνέχεια, σε δείγμα σχολίων, επιλέχθηκαν οι σχολιαστές και σύμφωνα με τα δεδομένα που φαίνονται στον Πίνακα 1 συλλέχθηκαν οι τοποθεσίες των ξενοδοχείων που σχολιάζουν αυτοί οι σχολιαστές. Τα δεδομένα αυτά καθαρίστηκαν αφαιρώντας τα ξενοδοχεία στα οποία αρχικά είχαν σχολιάσει και τα ξενοδοχεία τα οποία ήταν στην ίδια χώρα με την χώρα του σχολιαστή, όπου την είχε δηλώσει. Παρακάτω φαίνονται οι πιο συχνές χώρες και οι 10 πιο συχνές περιοχές:

Χώρες	Ποσοστό επί του Συνόλου	Περιοχές	Ποσοστό επί του Συνόλου
Ελλάδα	16,0768%	Ηράκλειο	3,1414%
		Ρόδος	1,5790%

		Κέρκυρα	1,4134%
		Χανιά	1,3305%
		Ζάκυνθος	1,0435%
Ισπανία	13,5869%	Μαγιόρκα	1,9378%
		Τενερίφη	1,7170%
		Λανθαρότε	1,5845%
		Φουερτεβεντούρα	0,9717%
Ηνωμένο Βασίλειο	11,8312%	Λονδίνο	1,2753%
Ιταλία	4,726%		
Τουρκία	3,6217%		
Γαλλία	3,1469%		
Πορτογαλία	2,4016		
Ταυλάνδη	1,8771		
Κύπρος	1,7888		

Πίνακας 1: Κατανομή Συλλογής Σχολίων σε Χώρες και Περιοχές

Λόγω γεωγραφίας επιλέχθηκαν οι χώρες Ελλάδα, Ισπανία, Ιταλία, Τουρκία, Γαλλία, Πορτογαλία και Κύπρος, στις οποίες με βάση τα παραπάνω δεδομένα και έρευνα αγοράς, οι περιοχές κάθε χώρας που επιλέχθηκαν είναι οι εξής:

Ελλάδα: Ηράκλειο, Χανιά, Ρόδος, Ζάκυνθος και Κέρκυρα

Ισπανία: Μαγιόρκα, Τενερίφη, Λανθαρότε, Φουέρτεβεντούρα και Γκραν Κανάρια

Ιταλία: Σαρδηνία και Ακτή Αμάλφι

Τουρκία: Αττάλεια και Μούγκλα

Γαλλία: Νίκαια, Μασσαλία, Κορσική, Μπορντώ και Καλαί

Πορτογαλία: Μαδέιρα, Πόρτο και Αλγκάρβε

Κύπρος: όλα τα καταλύματα στη σελίδα

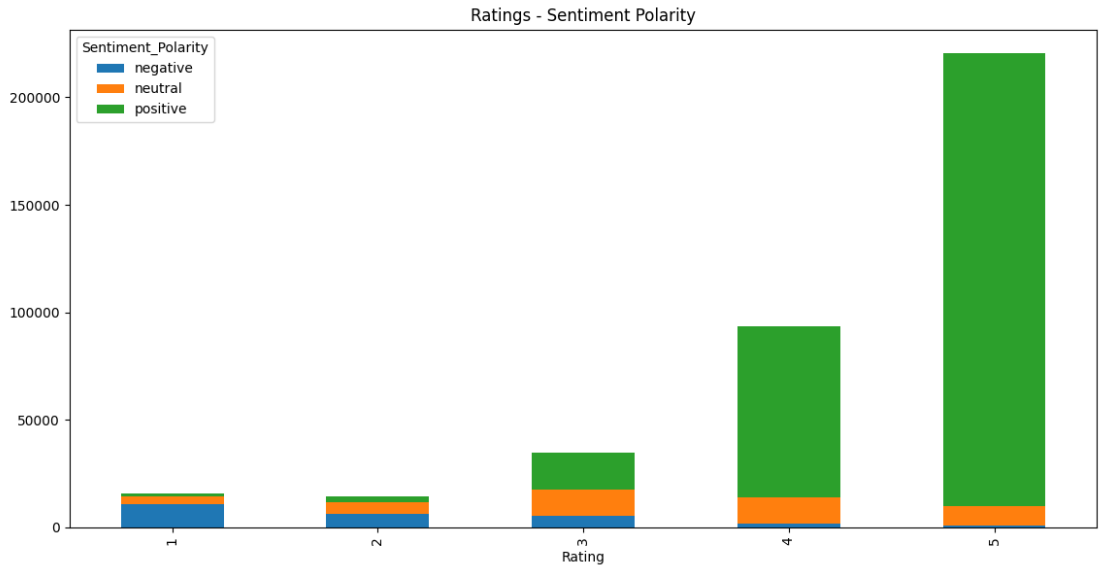
3.3 Ανάλυση Συναισθήματος

Για την Ανάλυση Συναισθήματος στα Σχόλια των ξενοδοχείων χρησιμοποιήθηκε η πλατφόρμα Semantria της Lexalytics, η οποία έδωσε και αποτελέσματα για την εξαγωγή των θεμάτων αναφοράς τα οποία θα αναλυθούν παρακάτω.
Όπως αναλύθηκε και στην ενότητα §2.3, η μέθοδος που χρησιμοποιείται στην πλατφόρμα για την ανάλυση συναισθήματος αποτελεί συνδυασμό Γλωσσικά Βασισμένων Μεθόδων και Μηχανικής Μάθησης, η οποία πραγματοποιεί ανάλυση συναισθήματος σε επίπεδο Κειμένου, Επίπεδο Φράσης Πρότασης και Επίπεδο Χαρακτηριστικού. Στην εργασία αυτή μας ενδιαφέρει η ανάλυση συναισθήματος επιπέδου Κειμένου για την εξαγωγή του συνολικού συναισθήματος του σχολίου, και επιπέδου Χαρακτηριστικού για την εξαγωγή του συναισθήματος στο Χαρκτηριστικό αναφοράς. Τα αποτελέσματα που προκύπτουν από την Ανάλυση Συναισθήματο στα Σχόλια παρουσιάζουν τα παρακάτω στατιστικά:

Αρνητικά: 25.199 Ουδέτερα: 42.598 Θετικά: 310.818

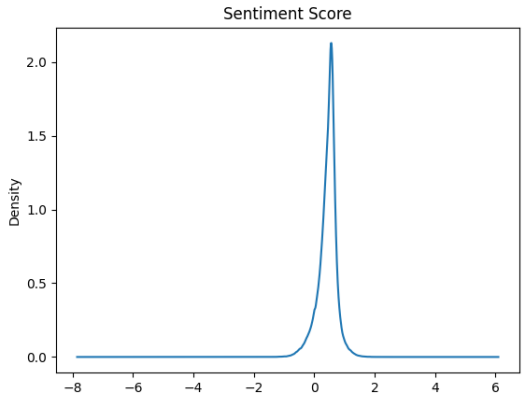
Η κατανομή της πολικότητας των σχολίων είναι λογική και συμβαδίζει με τα αποτελέσματα άλλων ερευνών καθώς και με τις αξιολογήσεις στα σχόλια όπως φαίνεται και στο Διάγραμμα 1. Στο Διάγραμα 1, με μπλε χρώμα φαίνονται τα αρνητικά σχόλια και φαίνεται να υπερτερούν στις αξιολογήσεις με 1 και 2 αστέρια, ενώ με πράσινο σχρώμα πράσινο, τα θετικά σχόλια, υπερτερούν στις αξιολογήσεις με 4 και 5 αστέρια και τέλος στις αξιολογήσεις με 3 αστέρια φαίνεται περίπου ίση κατανομή αρνητικών, θετικών αλλά και ουδέτερων σχολίων, κάτι που είναι λογικό και δείχνει τον διαχωρισμό θετικών και αρνητικών αξιολογήσεων.

Η σχέση των Αξιολογήσεων των σχολίων με την Πολικότητά τους από την Ανάλυση Συναισθήματος:



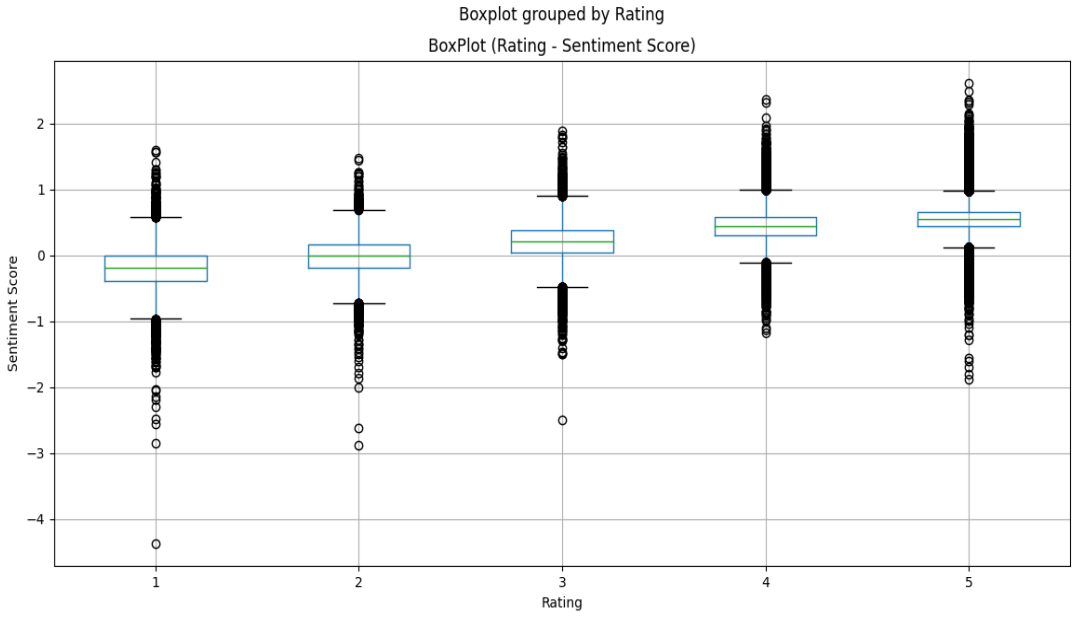
Διάγραμμα 1: Κατανομή Συναισθήματος στα σχόλια ανά Βαθμολογία

Στη συνέχεια ακολουθεί το γράφημα Πυκνότητας του Βαθμού συναισθήματος των σχολίων. Στο Διάγραμμα 2 φαίνεται πως ο βαθμός συναισθήατος για τα σχόλια που μελετήθηκαν είναι μετατοπισμένος προς τα θετικά με μέσο όρο το 0,45. Αυτό συμβαδίζει με την κατανομή των βαθμολογήσεων (κατά πολύ περισσότερες οι υψηλές/θετικές βαθμολογίες από τις χαμηλές/αρνητικές), αλλά και με άλλες έρευνες καθώς συνηθίζεται στα σχόλια σε ξενοδοχείανα υπερτερούν σε μεγάλο βαθμό τα θετικά σχόλια.



Διάγραμμα 2: Κατανομή Πυκνότητας Συναισθήματος στα Σχόλια

Το Διάγραμμα 3 είναι ένα boxplot οργανωμένο ως προς τις αξιολογήσεις και δείχνει την κατανομή του Βαθμού Συναισθηματος (Sentiment Score) των σχολίων. Όπως φαίνεται όσο αυξάνεται η αξιολόγηση αυξάνεται και η βαθμολογία συναισθήματος ενώ οι περισσότερες βαθμολογίες βρίσκονται στο έυρος [-1, 1] αλλά ύπαρχουν και μερικές ακραίες τιμές που το υπερβαίνουν, όμως οι περισσότερες από αυτές δεν ξεπερνούν το διάστημα [-2, 2].



Διάγραμμα 3: Boxplot Κατανομής Συναισθήματος Σχολίων ανά Βαθμολογία

3.4 Εξαγωγή Θεμάτων Αναφοράς

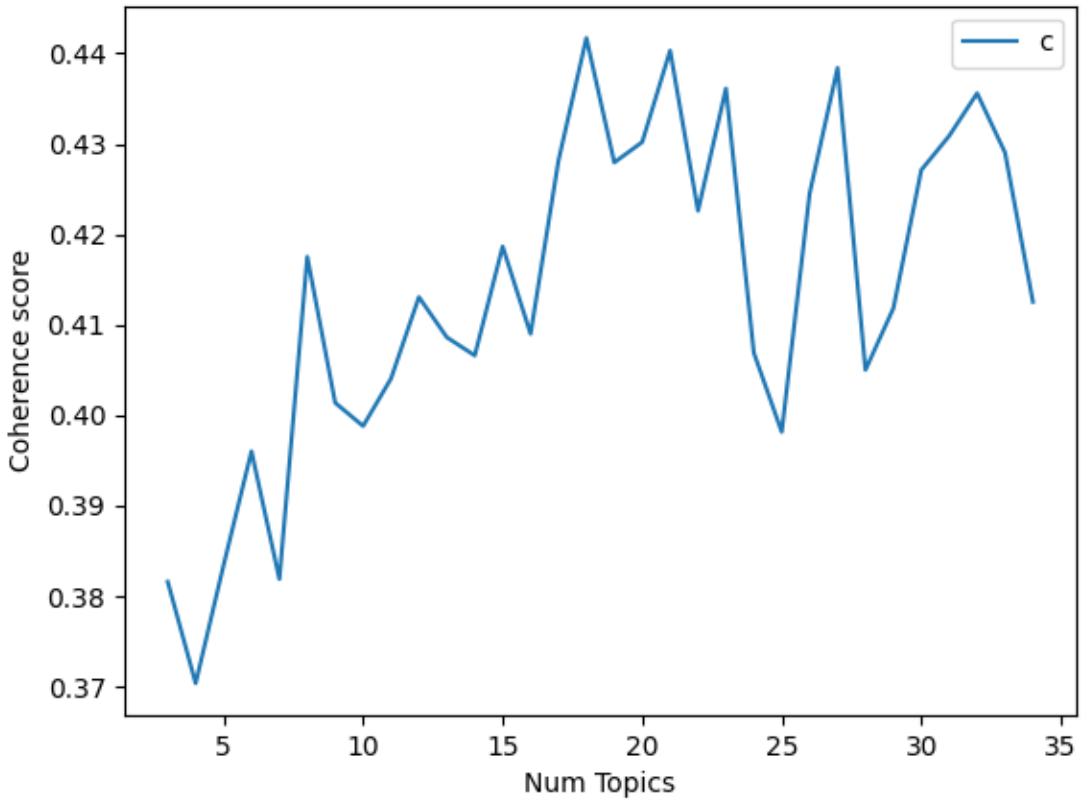
Σε αυτό το στάδιο της μεθοδολογίας θα επειχηρηθεί η εξαγωγή των πιο συχνών και σημαντικών για την ξιολόγηση των ξενοδοχείων Θεμάτων Αναφοράς των Σχολίων. Στη συνέχεια, τα θέματα θα ομαδοποιηθούν σε Χαρακτηριστικά των Ξενοδοχείων τα οποία θα χρησιμοποιηθούν για την Μοντελοποίηση των κριτηρίων. Τα Θέματα Αναφοράς θα εξαχθούν από την εφαρμογή του αλγορίθμου LDA στα Σχόλια και από την επεξεργασία των λέξεων/φράσεων κλειδιά από τα αποτελέσματα της Semantria όπως παρουσιάζεται παρακάτω.

- Αλγόριθμος LDA:

Για την εφαρμογή του αλγορίθμου Latent Dirichlet Allocation στα κείμενα των σχολίων αρχικά πραγματοποιήθηκε η προεπεξεργασία τους με τα παρακάτω βήματα:

1. Tokenization
2. Αφαίρεση StopWords και Σημείων Στίξεως και άλλων χαρακτήρων
3. Ένωση λέξεων με συχνή σειριακή εμφάνιση ανά δύο ή ανά τρεις (Bigram, Trigram)
4. Pos Tagging
5. Stemming

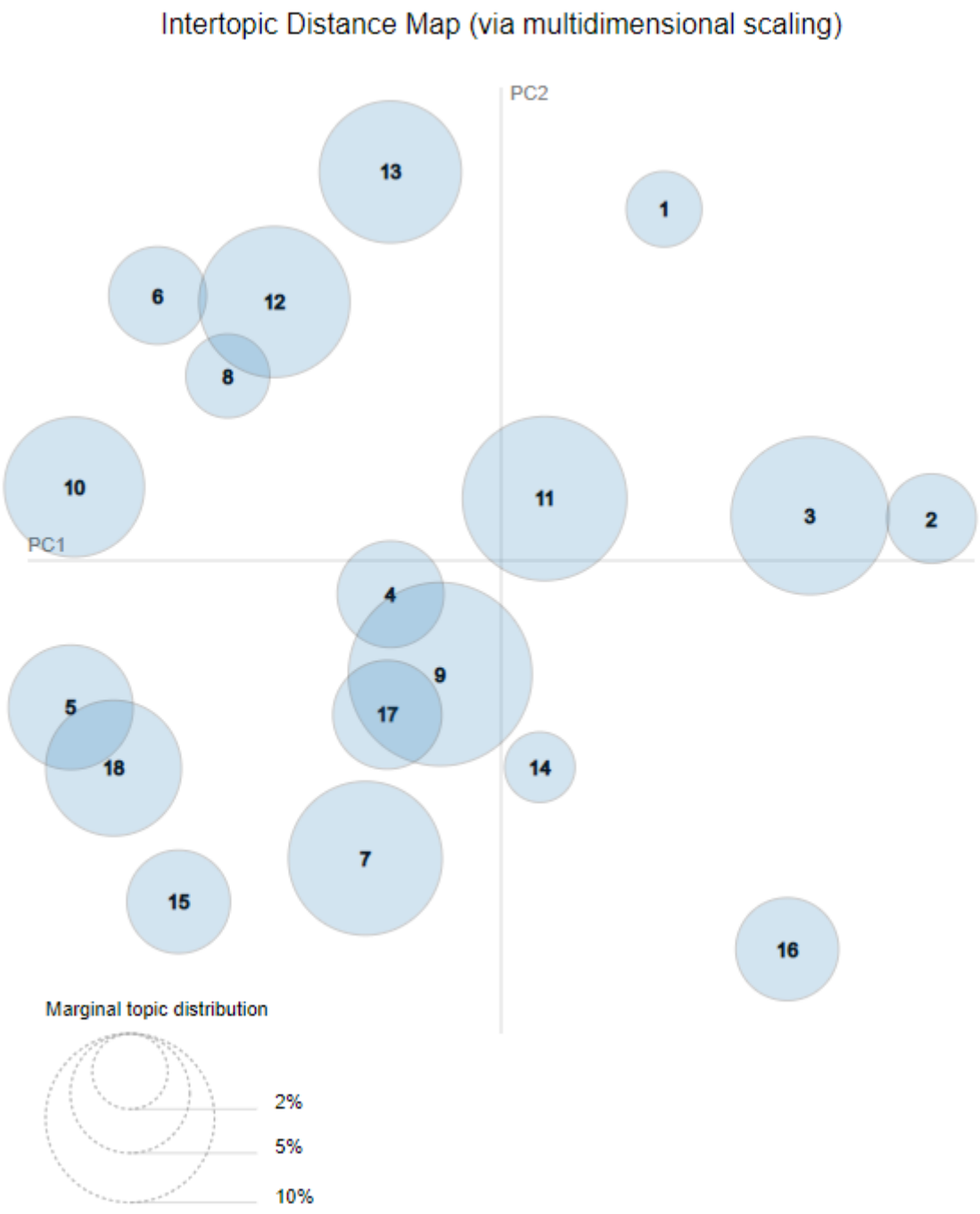
Για την επιλογή του αριθμού των θεμάτων που θα εφαρμοστεί ο αλγόριθμος έγινε ο έλεγχος για 3 έως 35 θέματα και αξιολογήθηκαν τα αποτελέσματα μέσω του μοντέλου CoherenceModel. Το μοντέλο περιγράφεται στο άρθρο (Röder, Both, & Hinneburg, 2015) και χρησιμοποιείται για τον υπολογισμό της «συνοχής» των θεμάτων που παράγουν αλγόριθμοι όπως ο LDA. Με αυτόν τον τρόπο αξιολογείται η «καθαρότητα» της πληροφορίας των θεμάτων. Το μοντέλο βασίζεται στο συνδασμό του μετασχηματισμό συνημιτόνου και άλλων μεθόδων και επιστρέφει τιμές από 0 έως 1 με το 1 να είναι το ιδανικό αποτέλεσμα. Στο άρθρο μελετήθηκαν τα αποτελέσματα Coherence Score σε διαφορετικά σετ δεδομένων
Για την εργασία αυτή εφαρμόστηκε ο αλγόριθμος LDA στα επεξεργασμένα σχόλια για διαφορετικό αριθμό θεμάτων από 3 έως 34 και για τα θέματα που προέκυψαν υπολογίσθηκε το σκορ μέσω του CoherenceModel όπως φαίνεται στο Διάγραμμα 4 όπου το καλύτερο αποτέλεσμα προέκυψε για 18 θέματα με coherence score = 0, 438.



Διάγραμμα 4: Γράφημα Αριθμού Θεμάτων – Coherence Score από την εφαρμογή του LDA

Σύμφωνα με το μοντέλο CoherenceModel για 18 θέματα προκύπτουν τα θέματα με την καλύτερη συνοχή, όμως για τη βελτίωση του διαχωρισμού τους, στη συνέχεια και για αυτόν τον αριθμό θεμάτων, έγινε ξανά η εφαρμογή του αλγορίθμου με παραμέτρους (alpha = symmetric, eta = auto, minimum_propability = 0,1, iterations = 135) οι οποίες θα βοηθούσαν στον καλύτερο διαχωρισμό των θεμάτων μεταξύ τους, και το αποτέλεσμα είχε coherence score = 0,432.

Η αποτύπωση των θεμάτων σε δύο διαστάσεις φαίνεται στο Σχήμα 9.



Σχήμα 9: Απεικόνιση κατανομής λέξεων σε προβολή στο δισδιάστατο χώρο οργανωμένες στα θέματα από την εφαρμογή του LDA

Στη συνέχεια ακολουθεί Πίνακας 2 με 50 λέξεις με την πιο συχνή εμφάνιση για κάθε θέμα και η απόδοσή χαρακτηριστικού (aspect) σε κάθε θέμα με βάση το σύνολο των λέξεων στο θέμα.

Θέματα	Χαρακτηριστικά	Συχνότερες Λέξεις Θεμάτων
Θέμα 1	**Διασκέδαση – Φιλικό προς την Οικογένεια**	kid, child, family, old, young, adult, club, year, entertainment, daughter, wedding, son, age, love, play, activity, baby, pool, resort, german, group, great, travel, park, party, parent, children, fun, show, month, husband, boy, food, team, 41nglish, teenager, lot, cater, entertain, barely, disco, even, guest, day, speak, especially, join, little, heavy, girl
Θέμα 2	**Τιμή – Σχόλια**	star, review, say, read, pay, bad, expect, people, money, think, look, previous, however, many, place, food, write, negative, well, book, price, comment, poor, guest, thing, positive, see, rating, never, much, give, seem, experience, find, complain, know, ever, value, rate, disappointed, feel, standard, want, believe, need, time, holiday, opinion, let, date
Θέμα 3	**Κράτηση – Check In**	day, get, book, arrive, go, tell, ask, reception, night, take, give, check, leave, say, time, even, find, first, morning, make, come, late, hour, early, guest, move, pay, problem, see, next, never, manager, know, back, wait, try, call, however, arrival, want, last, flight, put, pm, due, show, start, way, look, receptionist
Θέμα 4	**Μεταφορά**	bus, car, town, taxi, euro, trip, airport, take, hire, island, local, day, get, boat, road, visit, rhode, stop, cost, way, go, ride, worth, rent, drive, also, hour, minute, corfu, transfer, easy, min, cheap, travel, explore, find, free, walk, park, book, tour, euros, see, good, bike, crete, price, want, beach, well
Θέμα 5	**Δωμάτιο – Θέα**	view, sea, breakfast, balcony, great, lovely, terrace, night, restaurant, helpful, large, overlook, location, comfortable, floor, bed, friendly, good, small, bathroom, top, well, roof, excellent, suite, spacious, beautiful, modern, area, front, enjoy, side, sit, right, pool, clean, upgrade, fantastic, recommend, also, little, book, wonderful, nice, rooftop, definitely, town, bay, huge, look
Θέμα 6	**Ατμόσφαιρα – Θέα**	view, beautiful, place, amazing, recommend, perfect, wonderful, villa, love, stunning, day, enjoy, relax, highly, hill, village, want, sea, night, absolutely, go, trip, back, take, delicious, dinner, mountain, positano, get, definitely, look, step, spend, visit, drive, great, incredible, walk, private, come, service, way, breakfast, location, set, little, 41nglis, gorgeous, experience, island

Θέμα 7	**Δωμάτιο – Αυτονομία**	apartment, pool, clean, area, bed, towel, bar, good, day, well, need, sun, small, plenty, bedroom, week, balcony, large, kitchen, change, walk, reception, lovely, friendly, helpful, return, bathroom, spacious, little, basic, shower, bit, also, restaurant, get, daily, keep, problem, complex, size, fridge, enough, comfortable, sunbed, find, quite, supermarket, time, road, available
Θέμα 8	**Εξυπηρέτηση – Ατμόσφαιρα**	old, breakfast, place, feel, make, house, porto, owner, experience, chania, detail, comfortable, town, guest, home, wonderful, charming, welcome, boutique, local, little, building, attention, design, host, city, wine, well, expectation, modern, beautiful, warm, locate, heart, real, 42nglish42, recommendation, unique, charm, style, visit, serve, beautifully, crete, traditional, find, interesting, character, exceed, hospitality
Θέμα 9	**Παροχές Ξενοδοχείου – Διασκέδαση**	food, pool, good, bar, drink, beach, day, restaurant, go, inclusive, evening, get, entertainment, time, clean, week, lovely, eat, nice, great, water, night, choice, meal, lot, also, holiday, bit, always, main, really, plenty, area, well, little, much, serve, find, bed, friendly, sun, lunch, sea, snack, 42ngli, wine, sunbed, quite, return, never
Θέμα 10	**Κολύμβηση – Παροχές Ξενοδοχείου**	pool, beach, area, lovely, restaurant, resort, relax, great, bar, well, quiet, food, beautiful, friendly, garden, walk, excellent, place, holiday, clean, plenty, family, recommend, swimming, enjoy, good, village, week, look, also, away, visit, want, return, sun, small, sea, swim, helpful, lot, minute, main, couple, private, ground, time, perfect, day, fantastic, short
Θέμα 11	**Διασκέδαση**	go, good, bar, great, get, holiday, clean, night, back, pool, lovely, day, food, people, time, week, place, really, want, come, friendly, make, strip, say, always, beach, nice, drink, also, amazing, walk, basic, friend, well, enough, eat, need, much, away, look, recommend, brilliant, cheap, pay, thing, never, love, location, price, bit
Θέμα 12	**Εξυπηρέτηση - Προσωπικό**	make, thank, great, food, always, good, friendly, amazing, clean, special, service, holiday, restaurant, well, helpful, excellent, recommend, time, smile, welcome, really, work, wonderful, help, also, day, team, fantastic, feel, return, reception, bar, manager, definitely, much, especially, lovely, happy, perfect, visit, back, experience, enjoy, beautiful, mention, ever, pool, hard, week, sure

Θέμα 13	**Ατμόσφαιρα**	year, time, back, return, next, go, family, holiday, lovely, great, visit, make, clean, friendly, week, food, place, fantastic, much, last, love, come, welcome, feel, thank, home, friend, book, well, first, apartment, recommend, good, see, definitely, day, enough, always, run, look, second, wait, wonderful, amazing, trouble, helpful, excellent, already, many, say
Θέμα 14	**-**	nice, really, place, big, d, good, s, also, lot, kind, like, thing, bit, people, river, pretty, apt, re, I, think, quite, enjoy, madeira, breakfast, small, kitchen, clean, little, super, come, see, 43nglish4343e, house, actually, hob, minibar, bridge, surprise, price, compliment, rare, be, ps, garden, person, anyway, around, cozy, 43nglish43, almost
Θέμα 15	**Φαγητό**	breakfast, good, fresh, choice, coffee, fruit, dinner, buffet, meal, clean, evening, tea, food, bread, eat, friendly, selection, excellent, well, menu, lunch, helpful, hot, egg, include, cheese, 43nglish, serve, meat, cook, restaurant, comfortable, dish, option, night, cold, small, half_board, juice, offer, also, fish, available, order, day, salad, cereal, plenty, continental, variety
Θέμα 16	**Δωμάτιο**	bed, shower, bathroom, night, water, door, sleep, floor, work, small, bad, breakfast, tv, pay, wifi, dirty, wall, hot, towel, noise, even, smell, get, window, toilet, open, coffee, old, day, fridge, balcony, bath, clean, thing, euro, also, extra, poor, make, noisy, look, hear, pillow, sheet, cold, hard, change, place, hostel, next
Θέμα 17	**Παροχές Ξενοδοχείου – Χώροι Αθλητισμού – Χώροι Ευεξίας**	good, service, well, quality, high, excellent, spa, standard, price, facility, food, offer, restaurant, star, experience, breakfast, overall, course, gym, quite, location, free, also, golf, need, level, bodrum, area, time, buffet, massage, provide, new, however, dinner, wifi, bit, season, include, expect, many, expensive, pool, option, treatment, class, small, first, improve, point
Θέμα 18	**Τοποθεσία – Περπάτημα – Αξιοθέατα**	walk, good, beach, clean, location, great, minute, nice, restaurant, close, friendly, town, breakfast, helpful, min, away, recommend, city, centre, shop, old, small, street, main, also, place, quiet, night, bus, locate, bar, right, area, short, center, parking, perfect, road, value, comfortable, easy, excellent, well, lot, need, far, free, definitely, price, really

Πίνακας 2: Απόδοση Τίτλων των Θεμάτων του LDA

- Semantria:

Κατά την Ανάλυση Κειμένου ο αλγόριθμος της Semantria επιστρέφει επίσης Θέματα Αναφοράς και Χαρακτηριστικά που αναγνωρίζει από τα κείμενα. Στα δικα μας σχόλια των ξενοδοχείων οι τίτλοι των πιο συχνών Θεμάτων/Χαρακτηριστικών στις κατηγορίες *topics* και *auto_categories* είναι τα εξής:

topics : Hotels, Food, Beverages, Automotive, Sports, Travel, Marriage, Technology, Real Estate, Space, Education, Weather, Law, Environment, Software and Internet, Health, Popular Culture, Business, Art, Fashion, Disasters, Mobile Devices, Banking, Traditional Energy, Advertising

auto_categories : Food, Beverages, Home_And_Lodgings, Desserts, Alcohol, Business, Audio, Condiments_And_Sweeteners, Atmosphere, Water, Personal_Care, Architecture, Cleaning, Glass, Automobiles, Laundry, Energy, Leisure, Space, Family, Finance, Agriculture, Holidays, Broadcasting, Mobile_Phones

Τα πιο χρήσιμα αποτελέσματα στην αναγνώρηση των θεμάτων αναφοράς προήλθαν από την κατηγορία Themes στην ανάλυση κειμένου της Semantria. Όπως είχε αναλυθεί στην ενότητα §2.3, στην κτηγορία Themes γίνεται αναγνώριση στις προτάσεις των λέξεων/φράσεων κλειδιά που να αποδίδουν το κύριο θέμα αναφοράς της πρότασης. Γίνεται η επιλογή δύο ή περισσότερων λέξεων και όχι μίας για την αποφυγή «θορύβου» και την καλύτερη αποτύπωση του νοήματος της πρότασης. Επίσης, υπολογίζεται η ισχύς ή αλλιώς πόσο σημαντικό αποτελεί το συγκεκριμένο θέμα για το σύνολο του κειμένου στον όρο strength_score καθώς και η απόδοση συναισθήματος που αναφέρεται σε αυτό το θέμα αναφοράς.

Επειδή οι φράσεις κλειδιά (themes) αποτελούν μέρος του κειμένου των σχολίων, χρειάστηκε επιπλέον επεξεργασία τους για την εξαγωγή των πιο σημαντικών χαρακτηριστικών με την πιο συχνή αναφορά. Για την ομαδοποίηση των λέξεων στις φράση, έγινε επαναφορά τους στην αρχική ρίζα και εφαρμόστηκε αλγόριθμος αναγνώρισης λέξεων με συχνή ταυτόχρονη εμφάνιση και αυτές οι λέξεις συμπτύχθηκαν ως μία λέξη. Στη συνέχεια, υπολογίσθηκε η συχνότητα αναφοράς των επεργασμέων πλέον λέξεων και έγινε αναπαράστασή τους στο Σχήμα 10 με τη μορφή Γραφήματος Νέφους Λέξεων:

Σχήμα 10: Γράφημα WordCloud πιο συχνών φράσεων κλειδιά της Semantria για τα Σχόλια

Στη συνέχεια πραγματοποιήθηκε ομαδοποίηση των πιο συχνών λέξεων με σκοπό την εμφάνιση εννοιών χρήσιμων για την αξιολόγηση ξενοδοχείων. Τα αποτελέσματα αυτής της διαδικασίας, με τις πιο συχνές λέξεις/φράσεις κάθε χαρακτηριστικού παρουσιάζονται στον Πίνακα 3:

Χαρακτηριστικά	*Λέξεις*	Χαρακτηριστικά	*Λέξεις*
Παροχές Ξενοδοχείου	*garden, facility, building, accommodation, making_facility, amenity, flower, court_yard*	**Τοποθεσία**	*location, island, mountain*
Χώροι κ' Δραστηριότητες Ευεξίας	*spa, massage, turkish_bath, steam, spa_treatment, mud_bath*	**Μεταφορά**	*bus, road, bus_stop, parking, taxi, boat_trip, car, flight, bus_station, transfer, boat, car_park, drive, ride, harbour, shuttle_bus, free_parking, station, airport, port*
Ίντερνετ	*free_wifi, wifi, internet, free_wi, wi_fi, internet_connection*	**Κοντινή Περιοχή**	*area, local, surrounding, right_next, narrow_street, rhodes_town, super_market, local_knowledge, local_produce, pedestrian_street, fishing_village, ghost_town, surrounding_countryside, winding_road*
Αθλητικές Εγκαταστάσεις κ' άλλα Σπορ	*park, golf, gym, water_park, golf_course, tennis_court, water_sport, sport, table_tennis, football, fitness, water_polo, yoga, court, tennis, golf_buggy, golf_cart, crazy_golf, golfing, golfer*	**Βόλτα**	*walking_distance, street, walk, short_walk, minute_walk, main_road, nearby, main_strip, min_walk, walking, easy_access, distance, central, min, center, steep_hill, central_location, hill, residential, promenade*
Χώροι Αναψυχής κ' Διασκέδαση	*bar, drink, terrace, entertainment, fun, cocktail, music, party, club, evening_entertainment, activity, wedding, beer, rooftop, show, entertainment_team, roof, soft_drink, roof_terrace, night_life*	**Αξιοθέατα**	*town, shop, village, city, supermarket, centre, shopping, city_center, surroundings, town_centre, square, city_centre, market, attraction, historic, busy_road, river, pub, corfu_town, shopping_centre*
Φιλικό προς την Οικογένεια	*family, animation_team, team, kid, child, kid_club, baby, age, boy, school, wedding_anniversary, play, daughter, son,*	**Φαγητό**	*food, restaurant, breakfast, buffet, meal, fresh, kitchen, coffee, delicious, dish, lunch, wine,*

	animation, childrens, wife, anniversary, age_group, honeymoon		*snack_bar, glass, fresh_fruit, snack, fish, salad, bottle, tea*
Κολύμβηση	*swimming, swim, swimmer, pool, pool_area, swimming_pool, poolside beach, sun_ beach, sun_bed, water, sea, sun_lounger, sandy_beach, sunbeds, bay, sand, ocean, clear_water, coast, sunbed, sunbathing, seaside, salt_water, coastal, life_guard, sandy, seafront*	**Εστιατόριο**	*dining, evening_meal, dinner, menu, course, waiter, tavernas, chef, taverna, portion, carte, dinning, waitress, resturant, tavern, head_waiter, wine_list*
Δωμάτιο	*floor, spacious, ground_floor, top_floor, condition, th_floor, wall, decor, comfy, key, design, studio, unit, decent_size, rd_floor, storage_space, tidy, decoration*	**Πρωινό/ Μπουφές**	*half_board, self_catering, buffet_style, board, coffee_machine, catering, scrambled_egg, continental_breakfast, boiled_egg, continental, sandwich, fried_egg, wide_selection, milk, omelette, bacon, sausage, cereal*
Παροχές Δωματίου	*room, balcony, air_conditioning, air_con, fridge, patio, furniture, channel, equipment, english_channel, wardrobe, aircon, air_conditioner, sofa, wardrobe_space, mini_fridge, minibar*	**Υπηρεσίες**	*service, guest, hospitality, customer_service, serving, order, customer, security, security_guard, guest_relation, returning_guest, travel_agency, fellow_guest, travel_company, travel_kettle, agent, travel_cot, male_receptionist, assistant_manager, willing*
Θέα	*view, sea_view, window, scenery, seaview, panoramic, seaviews, oceanview*	**Προσωπικό**	*staff, friendly, helpful, friendly_staff, much_trouble, manager, family_run, owner, host, extremely_helpful, special_thanks, professional, rude, polite, management, extra_mile, always_happy, rep, advice, attitude*

Διαμέρισμα/ Αυτονομία	*apartment, private, house, villa, bungalow, washing_machine, studio_apartment, hostel, oven, appartment, ironing_board*	**Κράτηση/ Check In**	*reception, reception_staff, front_desk, book, booking, receptionist, early_morning, lobby, check, desk, information, recommendation, tip, reservation, travel_agent, booked, arrival, late_check, tour_operator, late_arrival*
Είδη Δωματίων	*suite, junior_suite, en_suite, ensuite, honeymoon_suite, executive_suite*	**Καθαριότητα**	*clean, spotlessly_clean, cleaning, dirty, spotless, maid_service, cleaning_lady, immaculately_clean, cleaner, housekeeping, maid, immaculate, laundry, cleanliness, washing, chamber_maid, spotlessly, uncleaned*
Τουαλέτα/ Μπάνιο	*bathroom, towel, shower, hot_water, fresh_towel, bath, toilet, shower_head, shower_curtain, hair_dryer, hot_tub, toilet_roll, jacuzzi, toilet_paper, water_pressure, toiletry, shower_gel, shower_cubicle, powerful_shower, jacuzzi_bath*	**Ατμόσφαιρα**	*time, quiet, experience, relaxing, atmosphere, noise, noisy, peaceful, enjoyable, happy, special_mention, weather, welcome, welcoming, safe, smell, loud, environment, loud_music, delightful*
Ύπνος	*bed, bedroom, comfortable_bed, sheet, double_bed, bedroom_apartment, comfy_bed, sleep, bed_linen, mattress, sofa_bed, bedding, linen, pillow, king_size, sleeping, light_sleeper, bedroomed_apartment, mattress_topper, sleepless_night*	**Αξία**	*value, luxury, reasonable, included*
		Τιμή	*free, price, inclusive, cheap, reasonable_price, including, expensive, money, rate, complimentary, deal, budget, cost, worth, package, euro,*

			charge, extra_cost, extra_charge, luxurious
		Σχόλια	*star, review, bad_review, previous_review, comment, star_rating, negative_review, trip_advisor, negative_comment, rating, previous_reviewer, reading_review, read_review, reviewer, reviewing, reviewes*

Πίνακας 3: Οργάνωση των φράσεων κλειδιά της Semantria σε Θέματα

3.5 Μοντελοποίηση Κριτηρίων

Στην Μεθοδολογία Μοντελοποίησης Κριτηρίων για την εργασία η Προβληματική είναι **Προβληματική Τύπου γ**, αφού έχει σκοπό την Αξιολόγη των Ξενοδοχείων και τις στοιχειωδεις επιπτώσεις αποτελούν τα ξενοδοχεία από τις διαφορετικές χώρες που συλλέχθηκαν και έμειναν μετά τον καθαρισμό των δεδομένων.

Οι στοιχειώδεις επιπτώσεις των εναλλακτικών για το συγκεκριμένο πρόβλημα προκύπτουν από:

i. Τα διαθέσιμα δεδομένα των Ξενοδοχείων από το TripAdvisror
ii. Τα Θέματα/Χαρακτηριστικά που προέκυψαν από την ανάλυση των Σχολίων:
    a. Από τον LDA
    b. Από τον χαρακτηρισμό και την ομαδοποίηση των λέξεων κλειδιά της Semantria

Οπότε οι στοιχειώδεις επιπτώσεις τελικά είναι οι εξής:

Υποδομές/ Παροχές Ξενοδοχείου	Δωμάτιο	Τοποθεσία	Υπηρεσίες	Φαγητό	Αξία
Κολύμβηση	Παροχές Δωματίου	Μεταφορά	Προσωπικό	Εστιατόριο Ξενοδοχείου	Τιμή
Αθλητικές Εγκαταστάσεις κ' άλλα Σπορ	Τουαλέτα/ Μπάνιο	Κοντινή Περιοχή	Κράτηση/ Check In	Πρωινό/ Μπουφές	Γνώμη από Σχόλια

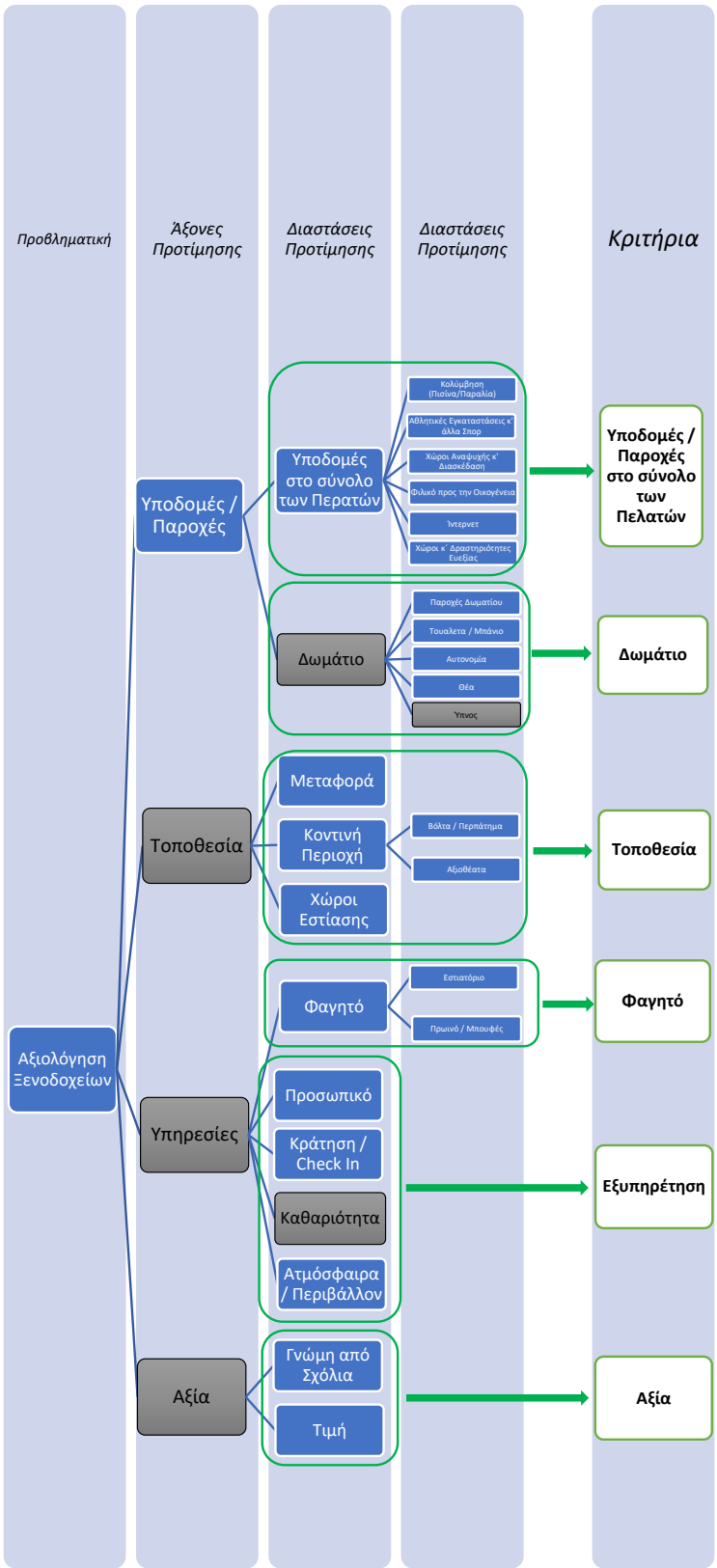
Χώροι Αναψυχής κ' Διασκέδαση	Αυτονομία Δωματίου/ Διαμέρισμα	Βόλτα/Περπάτημα	Καθαριότητα		
Φιλικό προς την Οικογένεια	Θέα	Αξιοθέατα	Ατμόσφαιρα/ Περιβάλλον		
Ίντερνετ	Ύπνος	Χώροι Εστίασης	Υπηρεσίες		
Υποδομές/Παροχές Ξενοδοχείου	Δωμάτιο	Τοποθεσία			

Πίνακας 4: Πίνακας Στοιχειωδών Επιπτώσεων από την Μοντελοποίηση Κριτηρίων από τα αποτελέσματα του LDA και της Semantria

Οι Άξονες Προτίμησης που προέκυψαν από την κατηγοριοποίηση των στοιχειοδών επιπτώσεων είναι οι εξής: **Παροχές / Υποδομές Ξενοδοχείου**, **Τοποθεσία**, **Υπηρεσίες** και **Αξία**

Στο Σχήμα 11 γίνεται αναπαράσταση της Μοντελοποίησης Κριτηρίων, όπου και φαίνονται οι στοιχειώδεις επιτπώσεις κάθε άξονα προτίμησης. Στην τελευταία στήλη φαίνονται τα τελικά κριτήρια και ο τρόπος διαμόρφωσής τους.

Στο σχλημα τα κουτιά με γκρι χρώμα περιέχουν τα χαρακτηριστικά στα επιτρέπεται η αξιολόγηση των ξενοδοχείο στα σχόλια από το TripAdvisor.



Σχήμα 11: Γράφημα Μοντελοποίησης Κριτηρίων

Στη συνέχεια, ο Πίνακας 5 περιέχει τους κανόνες απόδοσης τιμών (διάταξης) στις υποκατηγορίες των κριτηρίων που δημιουργήθηκαν στο γράφημα.

Κριτήρια	Διάσταση		0	1	2	3	4
Υποδομές / Παροχές στο σύνολο των Πελατών	***Κολύμβηση***		Τίποτα	Πισίνα ή Παραλία	Πισίνα και Παραλία	-	-
	Ίντερνετ		Καθόλου ή Επί Πληρωμή	Δωρεάν WiFi	-	-	-
	Χώροι κ' Δραστηριότητες Ευεξίας		Τίποτα	Σπα ή Μασάζ	Σπα και Μασάζ	-	-
	Αθλητικές Εγκαταστάσεις κ' άλλα Σπορ		Τίποτα	Γυμναστήριο	Εξειδικευμένο Άθλημα-Δραστηριότητα	-	-
	Χώροι Αναψυχής κ' Διασκέδασης		Τίποτα	Ένας Χώρος	Δύο ή Περισσότεροι Χώροι	-	-
	Φιλικό προς την Οικογένεια		Τίποτα	Μία Παροχή	Δύο Παροχές και Babysitting	-	-
Δωμάτιο	***Αξιολόγηση TripAdvisor Δωματίου***		Τιίποτα ή 1	1,5 ή 2	2,5 ή 3	3,5 ή 4	4,5 ή 5
	Αξιολόγηση TripAdvisor Ύπνου		Τιίποτα ή 1	1,5 ή 2	2,5 ή 3	3,5 ή 4	4,5 ή 5
	Παροχές κ' Είδη Δωματίου		Μηδαμινές	Βασικές	Στοιχειώδεις	Επιπλέον	Πολυτελείς
Τοποθεσία	***Αξιολόγηση TripAdvisor***		Τιίποτα ή 1	1,5 ή 2	2,5 ή 3	3,5 ή 4	4,5 ή 5
	Βόλτα / Περπάτημα		Τίποτα	Προτείνεται η χρήση Αυτοκινήτου	Κάπως Περπατήσιμη	Καλή Για Περπάτημα	Τέλεια για Περπάτημα
	Εστιατόρια	***Εύρος έως 0,3 μιλίων***	[0, 10]	(10 , 25]	(10 , 80]	(80 , 160]	>160

		Εύρος από 0,3 έως 0,75 μιλίων	[0, 25]	(25, 80]	(80, 160]	>160	-
		Εύρος από 0,75 έως 5 μιλίων	[0,80]	(80, 160]	>160	-	-
	Αξιοθέατα	***Εύρος έως 0,3 μιλίων***	[0, 1]	(1 , 10]	(10 , 25]	(25 , 80]	>80
		Εύρος από 0,3 έως 0,75 μιλίων	[0, 10]	(10, 25]	(25, 80]	>80	-
		Εύρος από 0,75 έως 5 μιλίων	[0,25]	(25, 80]	>80	-	-
	Μεταφορά		Τίποτα	Δωρεάν Πάρκινγκ ή/και Μεταφορά από Αεροδρόμιο	Δωρεάν Πάρκινγκ ή/και Μεταφορά από Αεροδρόμιο και Ενοικιαζόμενα	-	-
Φαγητό	***Φαγητό Εντός του Ξενοδοχείου (Εστιατόριο/Πρωινό /Μπουφές)***		Τίποτα	Διαθέσιμο Πρωινό / Μπουφές	Δωρεάν Πρωινό	Εστιατόριο	Πρωινό Και Εστιατόριο Και Επιπλέον Υπηρεσία
Εξυπηρέτηση	***Αξιολόγηση Υπηρεσιών TripAdvisor***		Τιίποτα ή 1	1,5 ή 2	2,5 ή 3	3,5 ή 4	4,5 ή 5
	Αξιολόγηση Καθαριότητας TripAdvisor		Τιίποτα ή 1	1,5 ή 2	2,5 ή 3	3,5 ή 4	4,5 ή 5
	Κράτηση / Check In		Τίποτα	Φύλαξη Αποσκευών	Θυρωρός / Φύλαξη	-	-

			ή/και 24ωρη Εξυπηρέτηση			
	Καθαριότητα	Τίποτα	Δυνατότητα Πλύσης Ρούχων	-	-	-
Αξία	***Αξιολόγηση TripAdvisor***	Τιίποτα ή 1	1,5 ή 2	2,5 ή 3	3,5 ή 4	4,5 ή 5
	Τιμή	>122	(82 , 122]	(82 , 60]	(42 , 60]	≤42
	Γνώμη Από Σχόλια (από Συνάρτηση)	Πολύ Κακή Γνώμη	Κακή Γνώμη	Μέτρια	Καλή Γνώμη	Πολύ Καλή Γνώμη

Πίνακας 5: Πίνακας Απόδοσης Τιμών στα Κριτήρια και τις Διαστάσεις τους

* Ειδική Συνάρτηση «Γνώμη Από Σχόλια»: Σε αυτή την συνάρτηση επιχειρείται να υπολογιστεί η γνώμη των χρηστών για το ξενοδοχείο από τα άλλα σχόλια στο TripAdvisor. Για τα 5 σχόλια στο ξενοδοχείο με το μεγαλύτερο αριθμό σε "Helpful Votes" δηλαδή τα σχόλια που οι χρήστες ψήφησαν σαν πιο χρήσιμα, χρησιμοποιείται ο αριθμός των Helpful Votes και η Βαθμολογία Συναισθήματος από την ανάλυση συναισθήματος. Για κάθε ξενοδοχείο γίνεται άθροισμα της Βαθμολογίας Συναισθήματος των Σχολίων επί ένα βάρος που υπολογίζεται για κάθε σχόλιο ως ο αριθμός των Helpful Votes προς το συνολικό άθροισμα των Helpful Votes για τα 5 σχόλια.

Η Βαθμολογία κάθε Ξενοδοχείου στα κριτήρια υπολογίζεται με το άθροισμα των κανονικοποιημένων Διαστάσεων (δηλαδή Βαθμολογία Διάστασης Ξενοδοχείου / Μέγιστη Βαθμολογία Διάστασης) και διαίρεση στη συνέχεια με τον αριθμό των διαστάσεων του κριτηρίου. Οπότε η βαθμολογία σε κάθε κριτήριο θα είναι μεταξύ 0 και 1

3.6 Πολυκριτήριος Πίνακας

Όπως παρουσιάστηκε παραπάνω, σύμφωνα με τους κανόνες του Πίνακα 5, δημιουργήθηκε ο πολυκριτήριος πίνακας με τα κριτήρια Υποδομές/Παροχές στο σύνολο των πελατών (*Facilities*), Δωμάτιο (*Room*), Τοποθεσία (*Location*), Φαγητό (*Food*), Εξυπηρέτηση (*Service*) και Αξία (*Value*), εναλλακτικές το σύνολο των ξενοδοχείων που μελετήθηκαν και για την κατάταξή τους (*Ranking*) χρησμοποιήθηκε η βαθμολογία τους σε αστέρια από το TripAdvisor, με 5 αστέρια την μέγιστη τιμή και τα ξενοδοχεία που είχαν αυτή την αξιολόγηση πήραν κατάταξη 1, με 4.5 αστέρια πήραν κατάταξη 2 κ.ο.κ.

Οι τιμή των ξενοδοχείων σε κάθε κριτήριο υπολογίστηκε με βάση τις βαθμολογίες του στις διαστάσεις του σε αυτό το κριτήριο και η διαδικασία που ακολουθήθηκε είναι η εξής:

i. Κανονικοποίηση της βαθμολογίας του ξενοδοχείου σε κάθε διάσταση του κριτηρίου στο [0, 1] με βάση την ελάχιστη και την μέγιστη δυνατή βαθμολογία της κάθε διάστασης.
ii. Υπολογισμός του μέσου όρου των κανονικοποιημένων βαθμολογιών των διαστημάτων του κριτηρίου.

Κατά αυτόν τον τρόπο ο πολυκριτήριος πίνακας περιέχει γραμμές με τα ξενοδοχεία και στήλες με τις βαθμολογίες των ξενοδοχείων αυτών στα κριτήρια. Οι βαθμολογίες αυτές θα λαμβάνουν τιμές από 0 έως 1, με 0 την χειρότερη και 1 την καλύτερη τιμή, αφού είναι αποτέλεσμα μέσου όρου τιμών στο διάστημα [0,1].

Κεφάλαιο 4: Αποτελέσματα

4.1 Εφαρμογή UTASTAR

Ακολουθώντας τη μεθοδολογία που παρουσιάστηκε, στον πολυκριτήριο πίνακα που δημιουργήθηκε στο κεφάλαιο §3.6, εφαρμόστηκε ο αλγόριθμος UTASTAR με σκοπό την εξαγωγή των βαρών προτιμήσεων των χρηστών στα κριτήρια του πίνακα.

Οι παράμετροι που επιλέχθηκαν για την εφαρμογή της UTASTAR είναι ίδιοι για όλα τα κριτήρια και είναι:

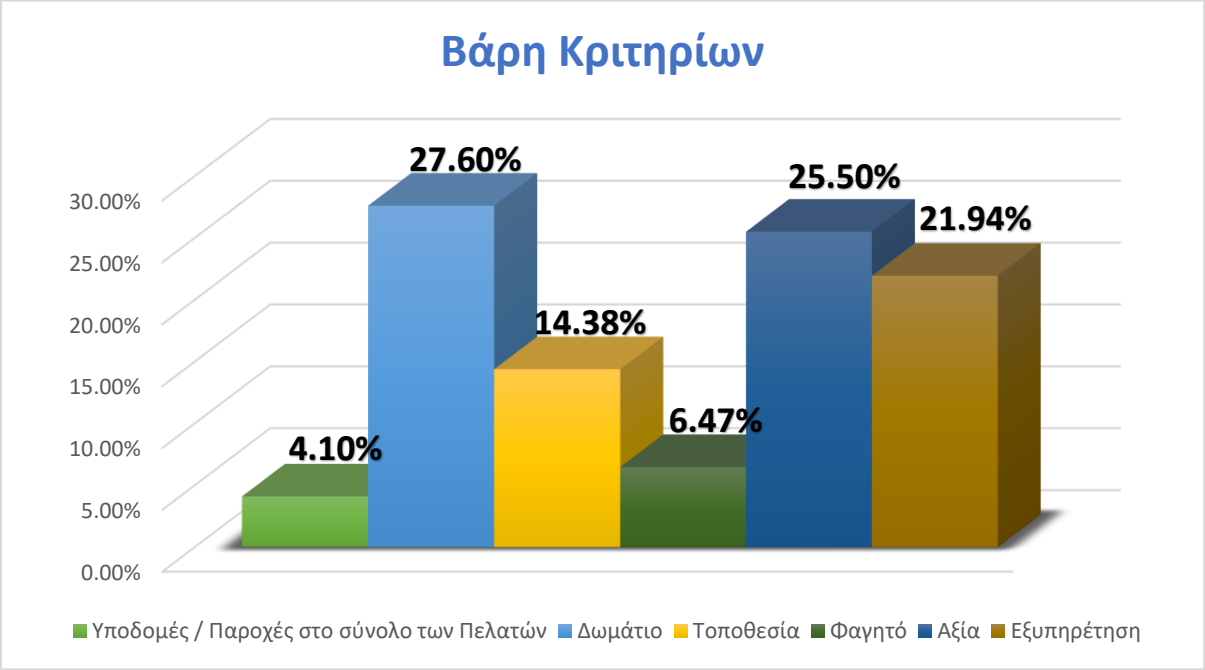
Μονοτονία	Είδος	Χειρότερη Τιμή	Καλύτερη Τιμή	Διαστήματα (a)
Αύξουσα	Συνεχή	0	1	10

Πίνακας 6: Πίνακας παραμέτρων UTASTAR

Για πιο αντιπροσωπευτικά αποτελέσματα ο αλγόριθμος εφαρμόστηκε 10 φορές σε τυχαίο δείγμα μεγέθους 10% του συγκεντρωτικού Πολυκριτήριου Πίνακα, και στη συνέχεια υπολογίστηκε ο μέσος όρος των βαρών της κάθε επανάληψης. Τα αποτελέσματα της της διαδικασίας παρουσιάζονται στον Πίνακα 7:

Επαναλήψεις \ Κριτήρια	Υποδομές / Παροχές στο σύνολο των Πελατών	Δωμάτιο	Τοποθεσία	Φαγητό	Εξυπηρέτηση	Αξία	ΤΑΦ του Kendall
Επανάληψη 1	0	0.25	0.178	0.113	0.224	0.234	**0,4128**
Επανάληψη 2	0.015	0.487	0.138	0.035	0.203	0.123	**0,481**
Επανάληψη 3	0.105	0.263	0.13	0.143	0.215	0.144	**0,4512**
Επανάληψη 4	0	0.225	0.081	0.081	0.413	0.201	**0,3609**
Επανάληψη 5	0	0.189	0.291	0	0.487	0.033	**0,4205**
Επανάληψη 6	0.021	0.331	0.134	0.05	0.006	0.457	**0,4861**
Επανάληψη 7	0.003	0.251	0.132	0.12	0.15	0.344	**0,3558**
Επανάληψη 8	0.125	0.188	0.082	0.032	0.272	0.301	**0,4152**
Επανάληψη 9	0.001	0.349	0.123	0.073	0.147	0.307	**0,4208**
Επανάληψη 10	0.141	0.228	0.149	0	0.432	0.05	**0,3119**
Μέσος Όρος	**0,041**	**0,276**	**0,1438**	**0,0647**	**0,255**	**0,2194**	**0,4116**

Πίνακας 7: Πίνακας Αποτελεσμάτων UTASTAR



Διάγραμμα 5: Βάρη Κριτηρίων

4.2 Εφαρμογή TOPSIS

Αρχικά με βάση τα βάρη των κριτηρίων που προέκυψαν από την εφαρμογή της UTASTAR στο σύνολο των εξεταζόμενων ξενοδοχείων, έγινε η εφαρμογή της TOPSIS στα δεδομένα και αναλύθηκαν τα αποτελέσματά της δίνοντας βάση στην ομοιότητα με την Καλύτερη και Χειρότερη λύση της TOPSIS των ξενοδοχείων και τη σχέση της με την συνολική αξιολόγηση του TripAdvisor των ξενοδοχείων. Η αξιολόγηση αυτή χρησιμοποιήθηκε στην UTASTAR ως κατάταξη των ξενοδοχείων και βάση αυτής υπολογίστηκαν τα βάρη των κριτηρίων. Οπότε στο βήμα αυτό αναζητείται και θα δηλώνει την ορθότητα των αποτελεσμάτων η συσχέτιση των αποτελεσμάτων της TOPSIS με την αξιολόγηση του TripAdvisor.

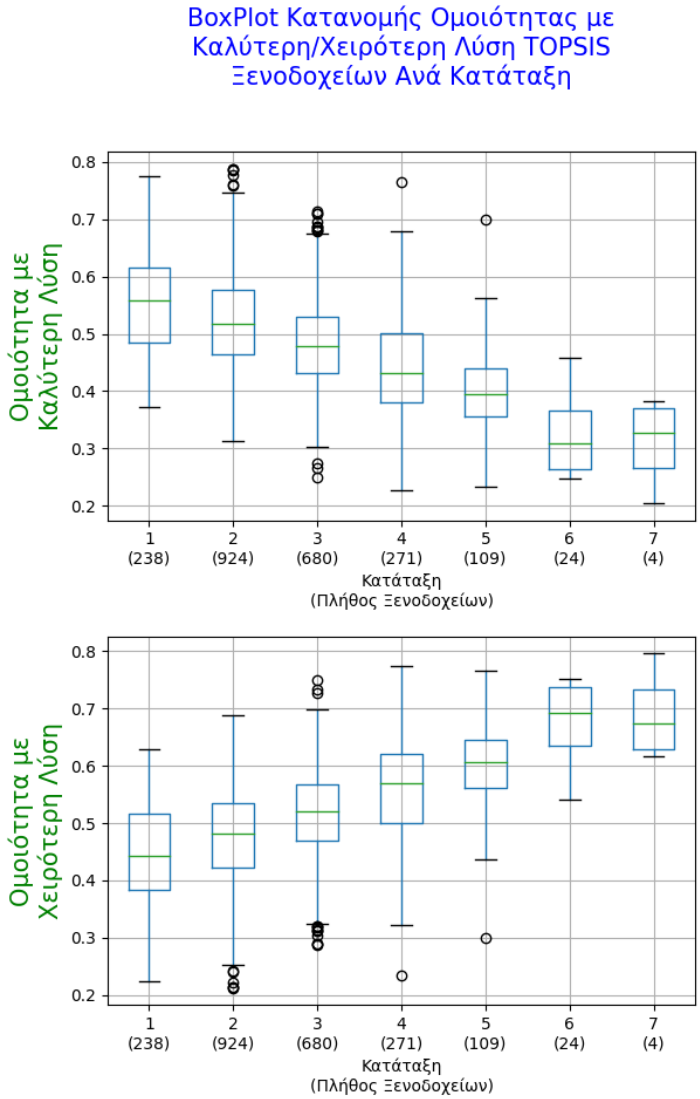
Για τον υπολογισμό της ομοιότητας/απόστασης από την ιδανική λύση χρησιμοποιήθηκε η ευκλείδεια απόσταση (p=2) για την επίτευξη πιο ακριβών αποτελεσμάτων.

4.3 Σύγκριση Αποτελεσμάτων

Σε αυτό το μέρος της εργασίας πραγματοποιήθηκε η σύγκριση των αποτελεσμάτων εφαρμογής των πολυκριτήριων μεθόδων όπως περιγράφηκε στα κεφάλαια §4.1 και §4.2. Αρχικά, για τη δημιουργία του Διαγράμματος 6 χωρίστηκαν τα δεδομένα ανά Κατάταξη (εισαγωγή στην εφαρμογή της UTASTAR), και αναπαρίστανται στον οριζόντιο άξονα. Στις ομάδες αυτές των δεδομένων εφαρμόστηκε η TOPSIS, με εισαγωγή των βαρών από το αποτέλεσμα εφαρμογής της UTASTAR. Από την εφαρμογή της TOPSIS χρησιμοποιήθηκε το στοιχείο ομοιότητας με την Καλύτερη/Χειρότερη Λύση TOPSIS το οποίο κυμαίνεται από [0, 1] όπου 0 σημαίνει καθόλου ομοιότητα με την λύση που μελετάται και 1 ταύτιση με αυτήν. Τα αποτελέσματα αυτά αποτυπώθηκαν για κάθε ομάδα δεδομένων με τη μορφή boxplot το οποίο αναπαριστά τη διακύμανση του μεγαλύτερου μέρους των δεδομένων (εντός του κουτιού), ενώ εκτός αυτού είναι οι ακραίες τιμές, και με την οριζόντια γραμμή η μέση τιμή τους.

Στο διάγραμμα αυτό είναι επιθυμητή η συνεχής μείωση της μέσης τιμής και της διακύμανσης της ομοιότητας με την Καλύτερη Λύση TOPSIS, ανά χειρότερη κατάταξη. Η μορφή του διαγράμματος αποτελεί δείκτη ποιότητας των δεδομένων και ορθότητας των αναλύσεων, οπότε όσο περισσότερο προσομοιάζουν την ιδανική κατάσταση τόσο πιο σωστές είναι οι αναλύσεις.

Στο δεύτερο διάγραμμα του Διαγράμματος 6, στον κάθετο άξονα χρησιμοποιείται το αποτέλεσμα της TOPSIS της ομοιότητας με την Χειρότερη Λύση TOPSIS όπου και θα ισχύουν τα αντίστροφα από ότι περιγράφηκε παραπάνω. Στην επεξήγηση του διαγράμματος και στις υπόλοιπες αναλύσεις επιλέχθηκε μόνο το αποτέλεσμα της ομοιότητας με Καλύτερη λύση TOPSIS για την αποφυγή πλεονασμού καθώς η ομοιότητα με τη Χειρότερη λύση είναι η ακριβώς αντίστροφη λύση και δεν προσφέρει κάποια επιπλέον πληροφορία.

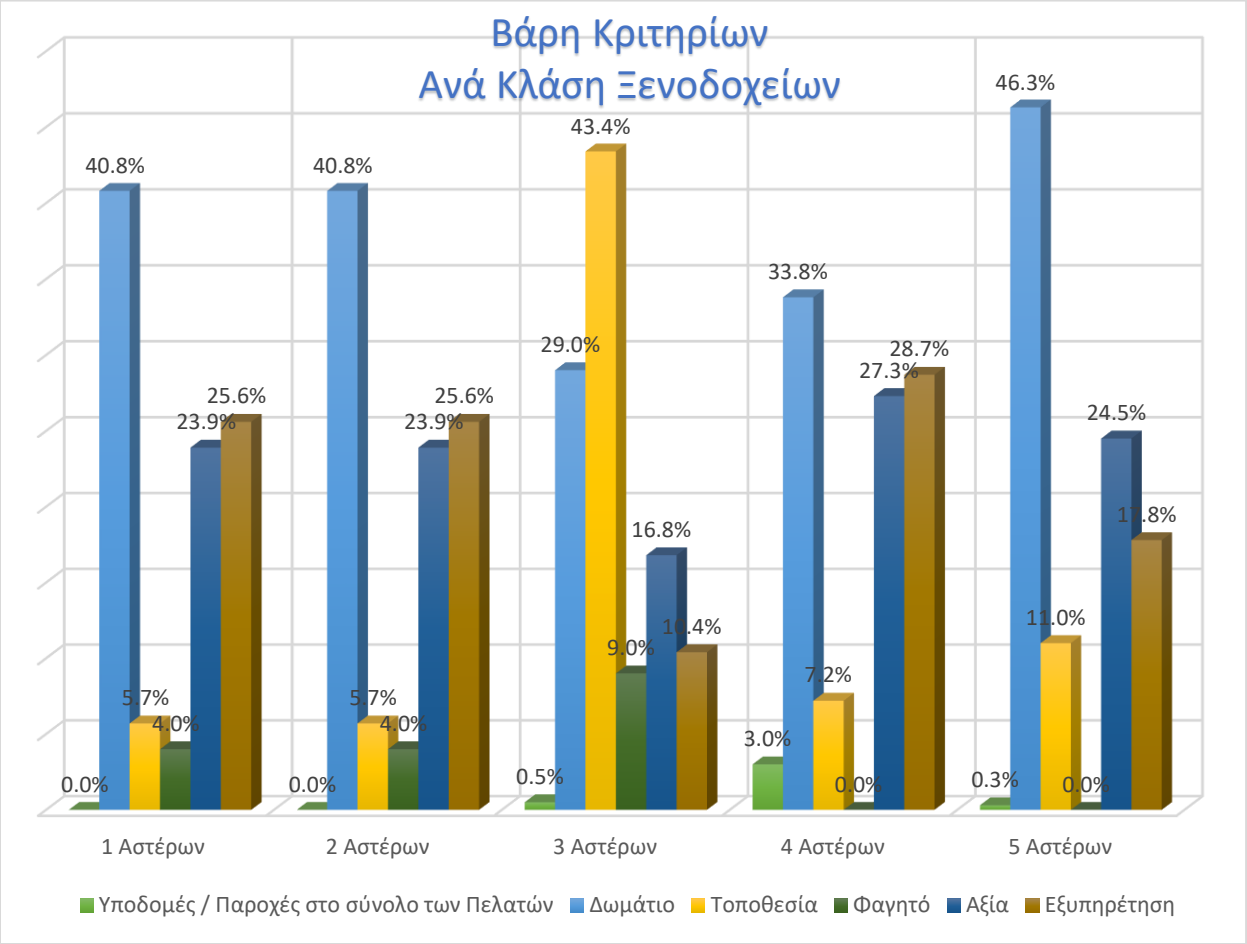


Διάγραμμα 6: BoxPlot Κατανομής Ομοιότητας με Καλύτερη/Χειρότερη Λύση TOPSIS Ξενοδοχείων ανά Κατάταξη

Από το Διάγραμμα 6 φαίνεται πράγματι η μείωση της διακύμανσης και της μέσης τιμής ανά κατάταξη. Από το διάγραμμα όμως δεν φαίνεται κάποιος ευκρινής διαχωρισμός των ομάδων των δεδομένων (κατάταξη) καθώς η μείωση των τιμών γίνεται με συνεχή τρόπο, οπότε δεν μπορούμε να εξάγουμε άλλα συμπεράσματα για τα δεδομένα.

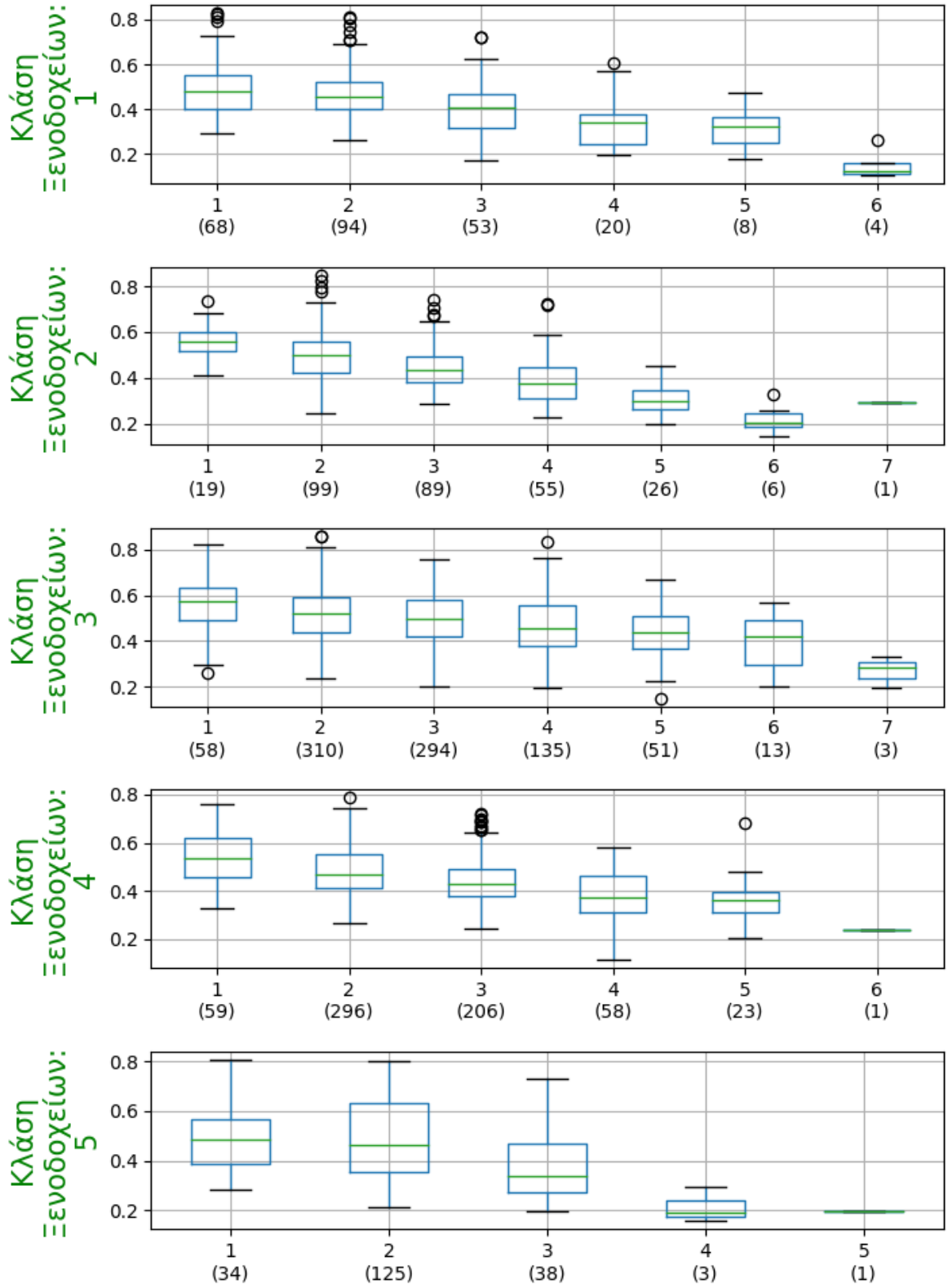
Για αυτό το λόγο στη συνέχεια επιχειρήθηκε η επανάληψη της παραπάνω διαδικασίας με επιπλέον ομαδοποίηση και διαχωρισμό των δεδομένων.

Σε επόμενο βήμα πραγματοποιήθηκε ξανά η παραπάνω διαδικασία, αφού τα ξενοδοχεία Ομαδοποιήθηκαν ανά Κλάση (Αστέρια).



Διάγραμμα 7: Βάρη Κριτηρίων ανά Κλάση

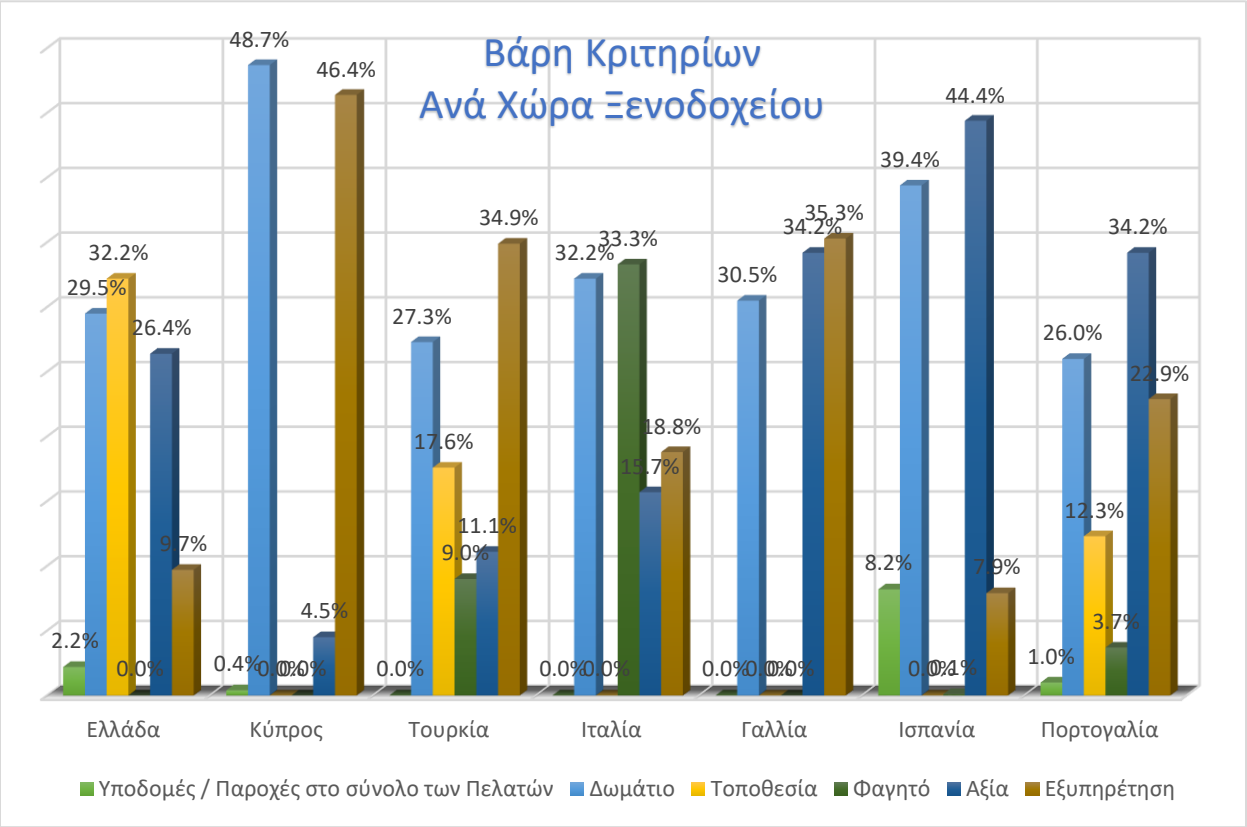
Στην ομαδοποίηση των ξενοδοχείων ανά Κλάση Αστέρων, Διάγραμμα 7, τα βάρη των κριτηρίων δεν παρουσιάζουν κάποια ιδιαίτερη διαφοροποίηση στις κλάσεις με εξαίρεση την κλάση των 3 Αστέρων όπου παρουσιάζεται μεγάλη αύξηση του βάρους του κριτηρίου της Τοποθεσίας, όπου και γίνεται το βάρος με το ισχυρότερο βάρος. Η ομάδα αυτή αποτελεί και σημαντικό ποσοστό του συνόλου των δεδομένων, κάτι που σημαίνει ότι δεν είναι αποτέλεσμα ακραίων τιμών, και μπορούν να εξαχθούν συμπεράσματα για τα ξενοδοχεία 3 Αστέρων. Δηλαδή, συμπεραίνεται ότι για τα ξενοδοχεία μεσαίας κλάσης η τοποθεσία παίζει τον πιο σημαντικό παράγοντα σε αντίθεση με τα υπόλοιπα ξενοδοχεία.



Διάγραμμα 8: BoxPlot Κατανομής Κατάταξης Ξενοδοχείων από την TOPSIS σε σχέση με την κατάταξή τους από το TripAdvisor Ανά Κλάση Ξενοδοχείου

Από το Διάγραμμα 8, για την ομάδα των ξενοδοχείων 3 Αστέρων, που παρατηρήθηκε διαφοροποίηση στα βάρη των κριτηρίων σε σχέση με τις υπόλοιπες ομάδες, παρατηρούνται υψηλότερες τιμές ομοιότητας με την καλύτερη Λύση TOPSIS στο σύνολο των ομάδων κατάταξης, σε σχέση με τις υπόλοιπες ομάδες Αστέρων. Αυτό σημαίνει, λόγω της σημαντικής αύξησης του βάρους του κριτηρίου της τοποθεσίας στην ομάδα των ξενοδοχείων 3 Αστέρων, ότι για τα ξενοδοχεία αυτά υπάρχει μικρότερη διακύμανση των χαρακτηριστικών τους (καλύτερο ξενοδοχείο σε σχέση με το χειρότερο) στα κριτήρια που αξιολογούνται ως σημαντικότερα (Τοποθεσία, Αξία, Δωμάτιο).

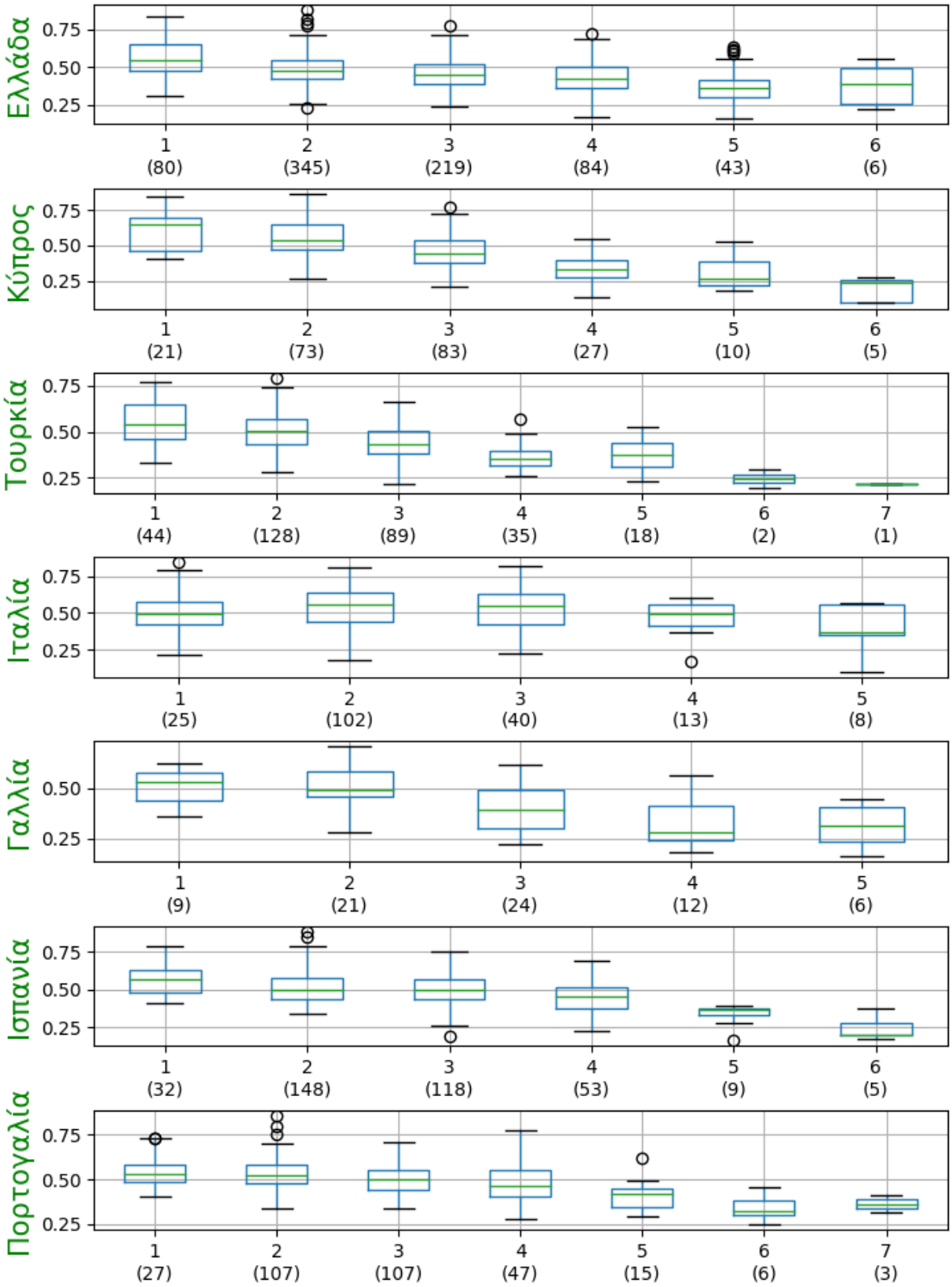
Σε επόμενο βήμα πραγματοποιήθηκε ξανά η παραπάνω διαδικασία, αφού τα ξενοδοχεία Ομαδοποιήθηκαν ανά Χώρα των Ξενοδοχείων.



Διάγραμμα 9: Βάρη Κριτηρίων Ανά Χώρα Ξενοδοχείων

Από το Διάγραμμα 9, η διαφορετική σύσταση των βαρών των κριτηρίων ανά χώρα δηλώνει ότι κάθε χώρα προσελκύει διαφορετικό τύπο τουριστών οι οποίοι έχουν διαφορετικές προτιμήσεις και αξιολογούν διαφορετικά τη σημαντικότητα των κριτηρίων.

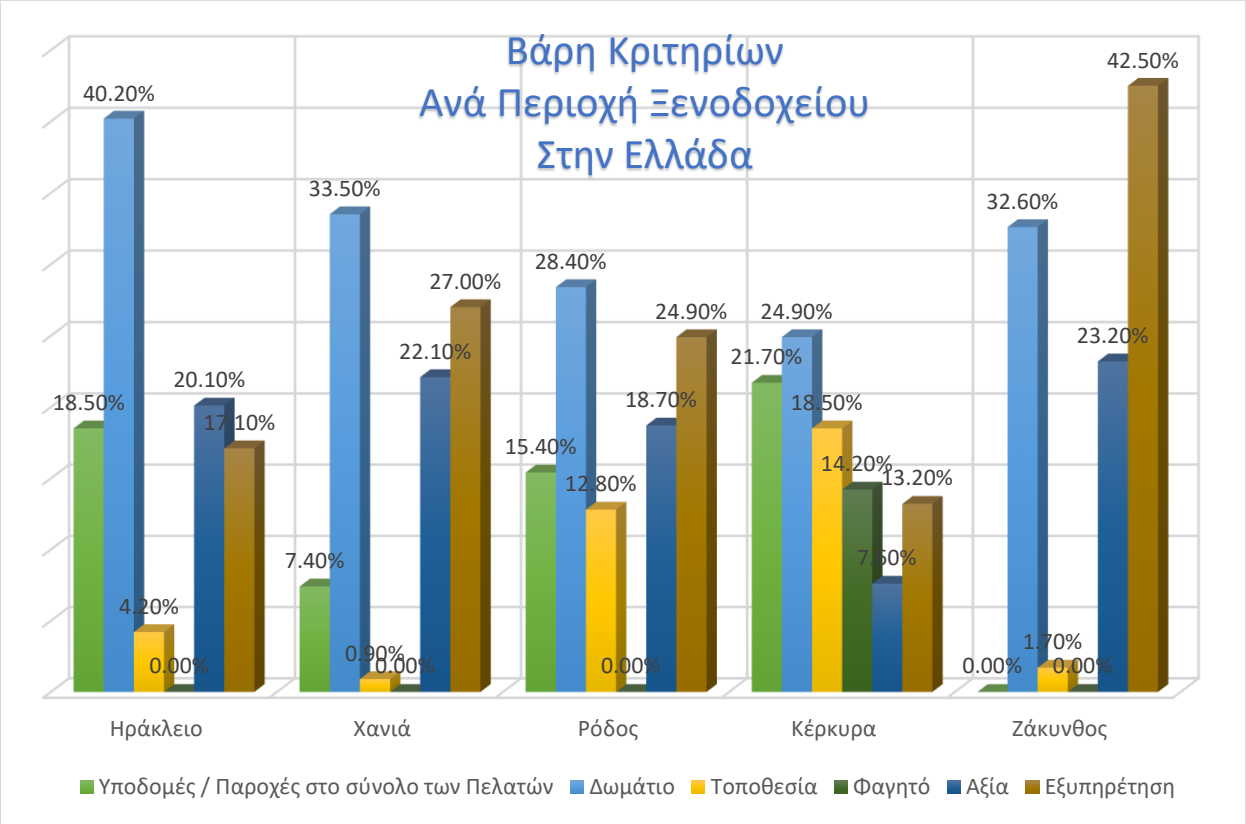
Για τις χώρες εκτός της Ελλάδας θα πρέπει να ληφθεί υπόψιν ότι έγινε συλλογή μικρότερου αριθμού δεδομένων από λιγότερες περιοχές οπότε ίσως να μην είναι αντιπροσωπευτικά τα αποτελέσματα για το σύνολο της χώρα, αλλά για τις περιοχές αυτής.



Διάγραμμα 10: BoxPlot Κατανομής Κατάταξης Ξενοδοχείων από την TOPSIS σε σχέση με την κατάταξή τους από το TripAdvisor Ανά Χώρα Ξενοδοχείου

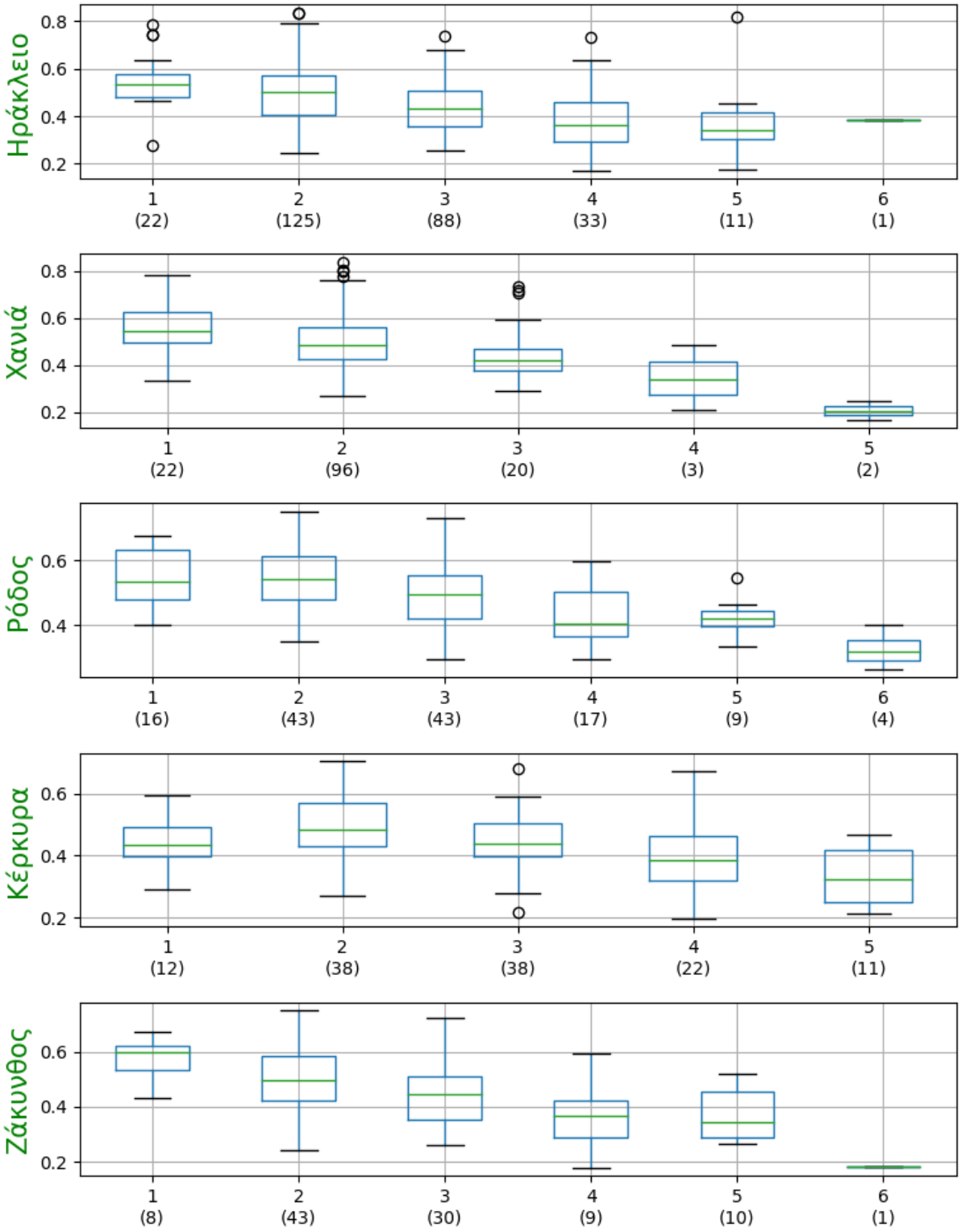
Από το Διάγραμμα 10, για την Ιταλία εμφανίζεται μη ομαλή μείωση των τιμών στις Κατατάξεις των ξενοδοχείων και μικρή διακύμανσή τους. Η Ιταλία, από το Διάγραμμα 9 των βαρών, είναι η μόνη χώρα με υψηλό βάρος στο κριτήριο του φαγητού, οπότε τα προβληματικά της αποτελέσματα από την εφαρμογή της TOPSIS μπορούν να εξηγηθούν από την ελλιπή πληροφορία των δεδομένων για το κριτήριο του φαγητού. Για την Πορτογαλία, παρουσιάζεται μικρή διακύμανση στα αποτελέσματα της TOPSIS ανάμεσα στις κατατάξεις και σχετικά μικρές τιμές. Αυτό μπορεί να εξηγηθεί από μεγάλη διαφοροποίηση των καλύτερων ξενοδοχείων, τα οποία αποτελούν ακραίες τιμές, σε σχέση με τα υπόλοιπα για τη σύσταση των βαρών που αντιστοιχούν στην χώρα αυτή.

Σε επόμενο βήμα πραγματοποιήθηκε ξανά η παραπάνω διαδικασία, αφού τα ξενοδοχεία Ομαδοποιήθηκαν ανά Περιοχή της Ελλάδας.



Διάγραμμα 11: Βάρη Κριτηρίων Ανά Περιοχή της Ελλάδας

Από το Διάγραμμα 11, οι περιοχές του Ηρακλείου, των Χανίων, και της Ρόδου παρουσιάζουν μικρή διαφοροποίηση στα βάρη των κριτηρίων, το οποίο σημαίνει ότι προσελκύουν παρόμοιο τύπο τουριστών. Η Ζάκυνθος εμφανίζει μηδενικό βάρος στο κριτήριο των Υποδομών και η Κέρκυρα είναι η μοναδική περιοχή με υψηλή τιμή στο κριτήριο του Φαγητού.



Διάγραμμα 12: BoxPlot Κατανομής Κατάταξης Ξενοδοχείων από την TOPSIS σε σχέση με την κατάταξή τους από το TripAdvisor Ανά Περιοχή της Ελλάδας

Από το Διάγραμμα 12, για την περιοχή της Κέρκυρας εμφανίζονται παρόμοια αποτελέσματα με αυτά της Ιταλίας από το Διάγραμμα 10, όπου και στις δύο περιπτώσεις εμφανίζεται υψηλό βάρος στο κριτήριο του Φαγητού, και επιβεβαιώνεται ότι για το κριτήριο του φαγητού τα δεδομένα είναι ελλιπή το οποίο έχει εκφράζεται από τη μη συσχέτισης της κατάταξης των χρηστών τω ξενοδοχείων με την κατάταξη της TOPSIS.

4.4 Ανάλυση Αγοράς

Στη συνέχεια παρουσιάζονται πίνακες στους οποίους ομαδοποιούνται τα ξενοδοχεία ανά Αστέρια/Κλάση και ανά Χώρα και Περιοχή της Ελλάδας αντίστοιχα και εμφανίζονται η μέση κατάταξη τους (ποσοστό ομοιότητας με την καλύτερη λύση) από την εφαρμογή της TOPSIS σε όλα τα ξενοδοχεία της ίδιας Κλάσης. Η μέση βαθμολογία τους στο Κριτήριο με το μεγαλύτερο βάρος που προέκυψε από την UTASTAR, και τις επιδόσεις τους στις διαστάσεις αυτού του Κριτηρίου. Αυτός ο πίνακας δίνει την δυνατότητα διερεύνησης περιθωρίων βελτίωσης της κατάταξης ξενοδοχείων μέσω στοχευμένου σχεδιασμού στα Κριτήρια και Διαστάσεις με συγκριτικά πλεονεκτήματα. Οι παρακάτω πίνακες σχεδιάστηκαν με από τη σκοπιά της Ελλάδας και του Ηρακλείου πιο συγκεκριμένα ακολουθώντας την αντίστροφη διαδικασία που ακολουθήθηκε στη συλλογή δεδομένων.

Στον παρακάτω πίνακα τα ξενοδοχεία ομαδοποιήθηκαν ανά Αστέρια/Κλάση και στη συνέχεια ανά Χώρα. Με βάση τα αποτελέσματα του πίνακα (Βάρη Κριτηρίων Ανά Κλάση Ξενοδοχείου) επιλέχθηκε το κριτήριο με το μεγαλύτερο βάρος και παρουσιάστηκαν οι μέσες επιδόσεις των Χωρών στο Κριτήριο και τις Διαστάσεις του.

Κλάση (Αστέρια) Ξενοδοχείου	Χώρα	Μέση Κατάταξη TOPSIS	Βαθμολογία Κριτηρίου	Βαθμολογία Διαστάσεων Κριτηρίου			
5			Δωμάτιο (Βάρος: 0,463)	Παροχές κ' Είδη Δωματίου	Αξιολόγηση TripAdvisor Ύπνου	Αξιολόγηση TripAdvisor Δωματίου	Μέσο Συναίσθημα Σχολίων Δωματίου
	Ιταλία	56.3%	0.529	1.75	4.25	4.5	1.408
	Ελλάδα	51.8%	0.486	1.255	4.17	4.128	1.779
	Πορτογαλία	48.2%	0.453	0.619	4.286	4.286	1.945
	Ισπανία	44.0%	0.439	1.083	4	3.833	1.572
	Τουρκία	44.7%	0.438	0.714	4.214	4	1.883
	Γαλλία	40.4%	0.403	0	5	4	2.286
	Κύπρος	38.6%	0.4	0.667	4.111	4	1.305
4			Δωμάτιο (Βάρος: 0,338)	Παροχές κ' Είδη Δωματίου	Αξιολόγηση TripAdvisor Ύπνου	Αξιολόγηση TripAdvisor Δωματίου	Μέσο Συναίσθημα Σχολίων Δωματίου
	Ιταλία	53.6%	0.535	1.667	4.4	4.4	1.731
	Κύπρος	50.3%	0.492	1.321	4.179	3.964	1.972

	Ελλάδα	51.5%	0.477	1.129	4.235	4.094	1.855		
	Ισπανία	47.6%	0.46	1.021	4.208	3.979	1.845		
	Πορτογαλία	46.3%	0.44	0.759	4.138	4.103	1.733		
	Τουρκία	46.5%	0.429	0.8	4.267	4	1.603		
	Γαλλία	39.2%	0.396	0.6	3.9	3.8	1.558		
3			Τοποθεσία (Βάρος: 0,434)	Μεταφορά	Βόλτα/ Περπάτημα	Αξιοθέατα	Εστιατόρια	Αξιολόγηση TripAdvisor Τοποθεσίας	Μέσο Συναίσθημα Σχολίων Δωματίου
	Πορτογαλία	48.8%	0.578	1.353	2.686	1.725	2.118	4.275	1.707
	Τουρκία	48.2%	0.559	1.288	2.932	1.153	1.712	4.508	1.997
	Ισπανία	49.1%	0.553	1.538	2.354	1.215	1.831	4.4	1.763
	Ελλάδα	46.8%	0.529	1.601	2.072	1.254	1.486	4.308	1.711
	Ιταλία	49.1%	0.528	1.345	2.207	1.483	1.414	4.483	1.802
	Κύπρος	44.7%	0.511	1.409	2.318	1.068	1.386	4.227	1.804
	Γαλλία	42.8%	0.478	1.214	1.643	1.143	1.429	4.5	1.714
2			Δωμάτιο (Βάρος: 0,408)	Παροχές κ' Είδη Δωματίου	Αξιολόγηση TripAdvisor Ύπνου	Αξιολόγηση TripAdvisor Δωματίου	Μέσο Συναίσθημα Σχολίων Δωματίου		
	Ιταλία	57.9%	0.56	1.769	4.385	4.308	2.111		
	Ισπανία	47.6%	0.486	1.389	4.167	4.056	1.681		
	Τουρκία	45.9%	0.469	0.971	4.229	4.114	1.919		
	Ελλάδα	47.9%	0.465	0.982	4.127	4.091	1.861		
	Πορτογαλία	44.8%	0.463	0.733	4.133	4.333	1.898		
	Κύπρος	42.4%	0.462	1.188	4.062	3.875	1.762		
	Γαλλία	43.9%	0.427	0.857	4.143	4	1.501		
1			Δωμάτιο (Βάρος: 0,408)	Παροχές κ' Είδη Δωματίου	Αξιολόγηση TripAdvisor Ύπνου	Αξιολόγηση TripAdvisor Δωματίου	Μέσο Συναίσθημα Σχολίων Δωματίου		
	Ιταλία	52.8%	0.521	1.472	4.346	4.228	1.954		
	Ελλάδα	49.8%	0.479	1.19	4.23	4.153	1.725		
	Ισπανία	47.5%	0.47	1.104	4.189	4.135	1.715		
	Τουρκία	47.4%	0.465	1.05	4.196	4.173	1.662		
	Κύπρος	46.3%	0.46	1.016	4.205	4.082	1.742		
	Πορτογαλία	46.4%	0.451	0.801	4.199	4.184	1.767		
	Γαλλία	45.1%	0.441	0.9	4.125	3.975	1.699		

Πίνακας 8: Παρουσίαση Κατάταξης με βάση τα Αποτελέσματα της Ανάλυσης των Χωρών για το Κριτήριο με το Μεγαλύτερο Βάρος, Ανά Κλάση

Στη συνέχεια, μελετήθηκαν τα ξενοδοχεία της Ελλάδας ανά Περιοχή. Στους παρακάτω πίνακες τα ξενοδοχεία ομαδοποιήθηκαν ανά Αστέρια/Κλάση και ανά Περιοχή της Ελλάδας. Σε κάθε πίνακα εμφανίζονται οι μέσες επιδόσεις των ξενοδοχείων της περιοχής για τα Κριτήρια **Τοποθεσία**, **Δωμάτιο** και **Αξία** αντίστοιχα και τις Διαστάσεις τους, που είναι τα κριτήρια με τα μεγαλύτερα βάρη για τα ξενοδοχεία της Ελλάδας.

Τοποθεσία (Βάρος: 0,322)									
Κλάση (Αστέρια) Ξενοδοχείου	Περιοχή	Μέση Κατάταξη TOPSIS	Τοποθεσία	Μεταφορά	Βόλτα/ Περπάτημα	Αξιοθέατα	Εστιατόρια	Αξιολόγηση TripAdvisor Τοποθεσίας	Μέσο Συναίσθημα Σχολίων Δωματίου
5	*Ρόδος*	56.7%	0.687	1.778	3.111	2.444	2.444	4.389	1.982
	Ηράκλειο	41.9%	0.521	1.444	2.222	1.167	1.389	4.361	1.86
	Χανιά	46.8%	0.511	1.818	1.909	0.727	1.091	4.455	1.864
	Ζάκυνθος	41.8%	0.477	1.25	2	1.25	1.25	4.25	1.472
	Κέρκυρα	35.1%	0.446	1.4	1.8	0.6	1.2	4	1.536
4	*Ρόδος*	48.5%	0.593	1.545	2.545	1.818	2.182	4.227	1.7
	Ηράκλειο	44.7%	0.523	1.425	2.2	1.325	1.575	4.275	1.641
	Χανιά	41.7%	0.496	1.4	2.1	1.2	1.3	4.2	1.603
	Ζάκυνθος	44.4%	0.471	1.769	1.231	0.692	0.923	4.346	1.998
	Κέρκυρα	39.9%	0.464	1.636	1.636	0.727	0.909	4.091	1.822
3	*Ρόδος*	46.9%	0.557	1.75	2.042	1.542	1.708	4.25	1.642
	Χανιά	47.3%	0.546	1.583	2.042	1.458	1.667	4.5	1.637
	Ηράκλειο	41.9%	0.525	1.702	2.064	1.128	1.404	4.17	1.773
	Ζάκυνθος	41.3%	0.513	1.235	2.412	1.294	1.471	4.441	1.551
	Κέρκυρα	39.3%	0.503	1.538	1.923	1	1.269	4.346	1.834
2	Ηράκλειο	51.8%	0.532	1.476	2.095	1.333	1.524	4.238	2.038
	Χανιά	53.1%	0.497	1.667	1.583	0.917	1.167	4.5	1.861
	Ρόδος	49.0%	0.479	1.667	1.5	1	1	4.083	1.949
	Κέρκυρα	44.7%	0.459	1.4	1.7	0.9	1.1	4.25	1.538
	Ζάκυνθος	38.2%	0.425	1.167	1.667	1	0.833	3.917	1.712
1	Χανιά	50.9%	0.552	1.554	2.072	1.542	1.687	4.53	1.672
	Ρόδος	48.9%	0.534	1.61	2.061	1.39	1.573	4.128	1.755
	Ζάκυνθος	47.3%	0.523	1.438	2.094	1.375	1.406	4.43	1.75
	Ηράκλειο	47.2%	0.519	1.455	2.136	1.201	1.474	4.328	1.762
	Κέρκυρα	43.8%	0.485	1.551	1.71	0.928	1.145	4.29	1.842

Πίνακας 9: Παρουσίαση Κατάταξης με βάση τα Αποτελέσματα της Ανάλυσης των Περιοχών της Ελλάδας για το Κριτήριο της Τοποθεσίας, Ανά Κλάση

Δωμάτιο (Βάρος: 0, 295)							
Κλάση (Αστέρια) Ξενοδοχείου	Περιοχή	Μέση Κατάταξη TOPSIS	Δωμάτιο (Βάρος: 0,408)	Παροχές κ' Είδη Δωματίου	Αξιολόγηση TripAdvisor Ύπνου	Αξιολόγηση TripAdvisor Δωματίου	Μέσο Συναίσθημα Σχολίων Δωματίου
5	*Χανιά*	46.8%	0.555	1.818	4.364	4.545	1.697
	Ρόδος	56.7%	0.492	1	4.111	4.333	2.016
	Ζάκυνθος	41.8%	0.485	2.25	3.75	3.25	1.353
	Κέρκυρα	35.1%	0.453	1.4	4	3.6	1.628
	Ηράκλειο	41.9%	0.449	0.778	4.222	4.111	1.849
	Χανιά	44.4%	0.523	1.308	4.538	4.308	2.131

4	*Κέρκυρα*	39.9%	0.477	1.364	4.091	3.909	1.727
	Ηράκλειο	44.7%	0.471	1.15	4.2	4.075	1.742
	Ρόδος	48.5%	0.463	0.727	4.273	4.182	2.079
	Ζάκυνθος	41.7%	0.46	1	4.1	4	1.844
3	*Χανιά*	47.3%	0.53	1.792	4.417	4.292	1.584
	Ρόδος	46.9%	0.527	1.792	4.208	4.292	1.543
	Ηράκλειο	41.9%	0.473	1.319	4.043	3.957	1.664
	Ζάκυνθος	41.3%	0.467	1.118	4.059	4.059	1.746
	Κέρκυρα	39.3%	0.465	1.038	4.077	4	1.883
2	*Χανιά*	53.1%	0.524	1.333	4.5	4.5	1.894
	Ρόδος	49.0%	0.5	1.167	4.167	4.167	2.097
	Ηράκλειο	51.8%	0.46	0.952	4	3.952	1.963
	Κέρκυρα	44.7%	0.451	0.9	4.2	4.2	1.608
	Ζάκυνθος	38.2%	0.357	0.333	3.667	3.5	1.62
1	*Χανιά*	50.9%	0.513	1.398	4.53	4.434	1.698
	Ρόδος	48.9%	0.493	1.415	4.11	4.012	1.803
	Ηράκλειο	47.2%	0.47	1.091	4.208	4.175	1.689
	Ζάκυνθος	47.3%	0.469	1.109	4.156	4.125	1.703
	Κέρκυρα	43.8%	0.449	0.971	4.13	3.957	1.762

Πίνακας 10: Παρουσίαση Κατάταξης με βάση τα Αποτελέσματα της Ανάλυσης των Περιοχών της Ελλάδας για το Κριτήριο του Δωματίου, Ανά Κλάση

Αξία (Βάρος: 0, 264)							
Κλάση (Αστέρια) Ξενοδοχείου	Περιοχή	Μέση Κατάταξη TOPSIS	Αξία	Βαθμολογία Τιμής	Πρώτη Εντύπωση από Σχόλια	Αξιολόγηση TripAdvisor Αξίας	Μέσο Συναίσθημα Σχολίων Δωματίου
5	Ζάκυνθος	41.8%	0.643	3.75	3	4.125	1.158
	Ηράκλειο	41.9%	0.629	3.333	3.111	4.25	1.05
	Ρόδος	56.7%	0.617	2.778	3.222	4.278	1.188
	Κέρκυρα	46.8%	0.61	2.545	3.364	4.364	1.028
	Χανιά	35.1%	0.601	3.2	3.2	3.8	0.966
4	Ηράκλειο	44.7%	0.666	3.925	3.3	4.225	0.983
	Ζάκυνθος	41.7%	0.666	3.9	3.4	4.15	0.955
	Κέρκυρα	39.9%	0.653	3.545	3.273	4.182	1.164
	Χανιά	44.4%	0.652	2.923	3.154	4.5	1.572
	Ρόδος	48.5%	0.632	3.182	3.273	4.091	1.193
3	Ζάκυνθος	41.3%	0.652	3.706	3.118	4.235	1.135
	Χανιά	47.3%	0.648	3.125	3.333	4.396	1.202
	Ηράκλειο	41.9%	0.634	3.404	3.298	4.074	1.01
	Κέρκυρα	39.3%	0.624	3.038	3.269	4.096	1.188
	Ρόδος	46.9%	0.615	3.25	3.042	4.188	0.994
2	Χανιά	53.1%	0.693	3.917	3.5	4.458	1.043
	Ηράκλειο	51.8%	0.647	3.762	3.048	4.071	1.222
	Ρόδος	49.0%	0.639	3	3.667	4	1.125
	Ζάκυνθος	38.2%	0.635	3.833	3	3.917	1.137

	Κέρκυρα	44.7%	0.622	3.4	3	4.1	1.121
1	*Ηράκλειο*	47.2%	0.653	3.539	3.279	4.195	1.148
	Χανιά	50.9%	0.648	3.253	3.277	4.44	1.096
	Κέρκυρα	43.8%	0.645	3.681	3.174	4.152	1.029
	Ζάκυνθος	47.3%	0.638	3.312	3.297	4.195	1.065
	Ρόδος	48.9%	0.637	3.537	3.146	4.061	1.138

Πίνακας 11: Παρουσίαση Κατάταξης με βάση τα Αποτελέσματα της Ανάλυσης των Περιοχών της Ελλάδας για το Κριτήριο της Αξίας, Ανά Κλάση

4.4 Ανάλυση Τουριστών/Σχολιαστών

Αυτό το τμήμα της εργασίας είχε σκοπό την εξαγωγή συμπερασμάτων συγκεκριμένα για τους **Σχολιαστές**, τα χαρακτηριστικά και τις προτιμήσεις τους, για το σύνολο των δεδομένων που συλλέχθηκαν.

Για κάθε ξενοδοχείο, επιλέχθηκαν τα σχόλια με ψήφους *Helpful Votes* παραπάνω από τον μέσο όρο των σχολίων του. Στη συνέχεια από την ανάλυση εξαγωγής θεμάτων αναφοράς της Semantria που είχε πραγματοποιηθεί στα σχόλια, και με επιπλέον επεξεργασία επιλογής μόνο των ουσιαστικών φαίνεται παρακάτω το WordCloud με τις λέξεις (ουσιαστικά) – θέματα αναφοράς σε αυτά τα σχόλια που ψηφίστηκαν ως τα πιο βοηθητικά από τους χρήστες:

Σχήμα 12: Γράφημα WordCloud για τα Σχόλια με σημαντικό αριθμό με Helpful Votes

Σε αντιπαραβολή με το WordCloud στο σύνολο των σχολίων του Σχήματος 10, το γράφημα αυτό δείχνει πιο καθαρά τις προτιμήσεις των Σχολιαστών, και αν ληφθεί υπόψιν το γεγονός ότι αυτά τα σχόλια ψηφίστηκαν ώς χρήσιμα από τους Σχολιαστές, σημαίνει ότι δείχνουν και με μεγαλύτερη ακρίβεια τι θεωρούν πιο σημαντικό. Οπότε, προτείνεται σε επόμενη εργασία να δωθεί μεγαλύτερη βαρύτητα σε σχόλια με μεγάλο αριθμό *Helpful Votes.*

Τη συνέχεια, πραγματοποιήθηκε ανάλυση των σχολιαστών που είχαν σχολιάσει σε δύο ή περισσότερα διαφορετικά ξενοδοχεία των δεδομένων που αναλύθηκαν, όπου η μία από αυτές ήταν το Ηράκλειο. Οι 10 πιο συχνές περιοχές τις οποίες επισκέπτονται (σχολιάζουν στα ξενοδοχεία τους) είναι τα παρακάτω:

Κατάταξη	Περιοχές
1	*Ρόδος*
2	*Κύπρος*
3	*Μούγκλα*
4	*Μαγιόρκα*
5	*Κέρκυρα*
6	*Χανιά*
7	*Αττάλεια*
8	*Αμάλφι Ακτή*
9	*Πόρτο*
10	*Μαδέιρα*

Πίνακας 12: Κατάταξη Περιοχών που σχολιάζουν περισσότερο Τουρίστες που επισκέπτονται τα ξενοδοχεία του Ηρακλείου

Στη συνέχεια παρουσιάζεται πίνακας για κάθε Χώρα και στη συνέχεια Περιοχή της Ελλάδας, οι 5 πιο δημοφιλείς Χώρες καταγωγής των Σχολιαστών των Ξενοδοχείων τους. Να σημειωθεί ότι το σύνολο των σχολίων που συλλέχθηκαν είναι στα Αγγλικά, το οποίο σημαίνει ότι η κατανομή των χωρών των σχολιαστών θα είναι ενισχυμένη για τις Αγγλόφωνες Χώρες και λιγότερο για τις χώρες που δεν χρησιμοποιούνται πολύ τα Αγγλικά.

Χώρα Ξενοδοχείων	Χώρα Σχολιαστών
Ελλάδα	Ηνωμένο Βασίλειο
	ΗΠΑ
	Ελλάδα
	Ισραήλ
	Αυστραλία
Ισπανία	Ηνωμένο Βασίλειο
	ΗΠΑ
	Ιρλανδία
	Ισπανία
	Γερμανία
Ιταλία	Ηνωμένο Βασίλειο
	ΗΠΑ
	Αυστραλία
	Καναδάς
	Ιταλία
Τουρκία	Ηνωμένο Βασίλειο
	ΗΠΑ
	Τουρκία
	Αυστραλία
	Ιρλανδία

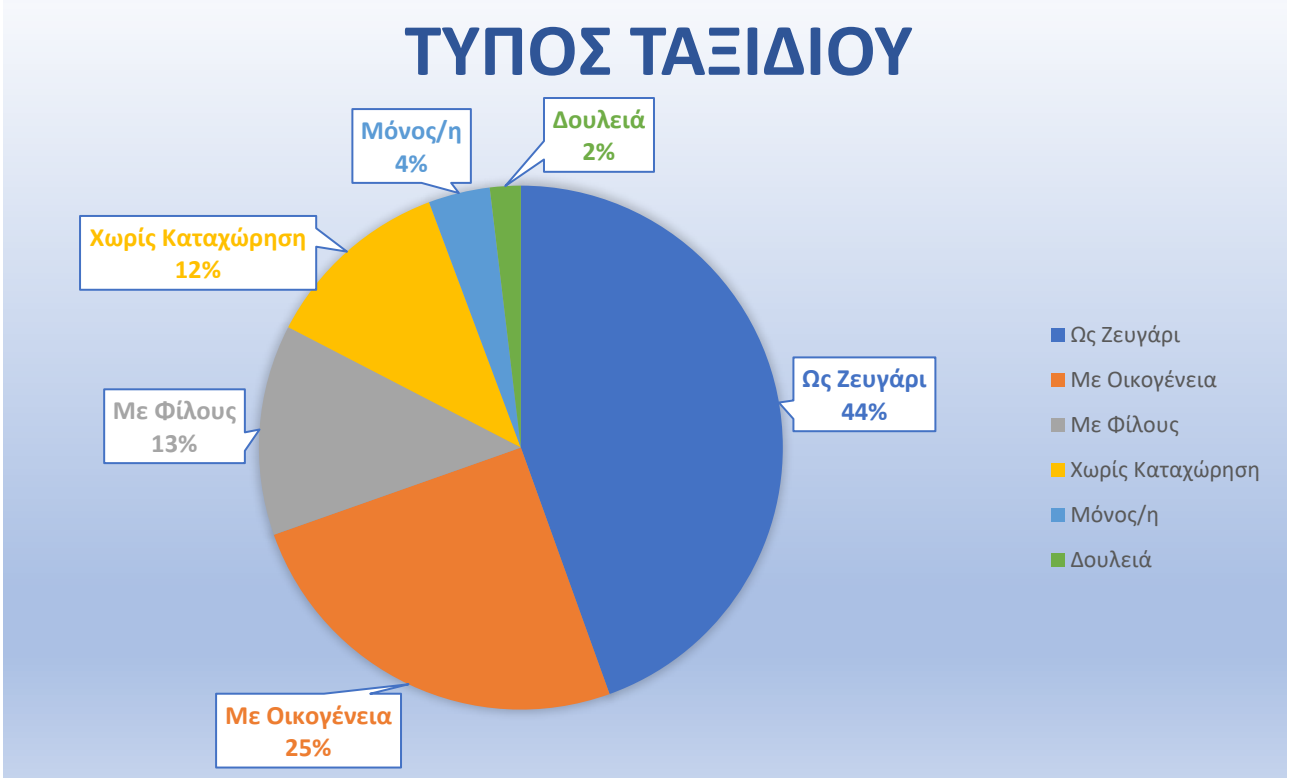
Χώρα Ξενοδοχείων	Χώρα Σχολιαστών
Κύπρος	Ηνωμένο Βασίλειο
	Κύπρος
	ΗΠΑ
	Ισραήλ
	Ελλάδα
Γαλλία	Ηνωμένο Βασίλειο
	ΗΠΑ
	Αυστραλία
	Γαλλία
	Καναδάς
Πορτογαλία	Ηνωμένο Βασίλειο
	ΗΠΑ
	Ιρλανδία
	Πορτογαλία
	Καναδάς

Πίνακας 13: Πίνακας με τις πιο δημολείς Χώρες Καταγωγής των Σχολιαστών Ανά Χώρα των Ξενοδοχείων

Περιοχή Ξενοδοχείων	Χώρα Σχολιαστών	Περιοχή Ξενοδοχείων	Χώρα Σχολιαστών
Ηράκλειο	Ηνωμένο Βασίλειο		
	ΗΠΑ		
	Ελλάδα		
	Ισραήλ		
	Ιρλανδία		
Χανιά	Ηνωμένο Βασίλειο	Κέρκυρα	Ηνωμένο Βασίλειο
	ΗΠΑ		ΗΠΑ
	Ελλάδα		Ελλάδα
	Νορβηγία		Ιρλανδία
	Καναδάς		Αυστραλία
Ρόδος	Ηνωμένο Βασίλειο	Ζάκυνθος	Ηνωμένο Βασίλειο
	ΗΠΑ		ΗΠΑ
	Ελλάδα		Ελλάδα
	Ισραήλ		Ιρλανδία
	Ιρλανδία		Αυστραλία

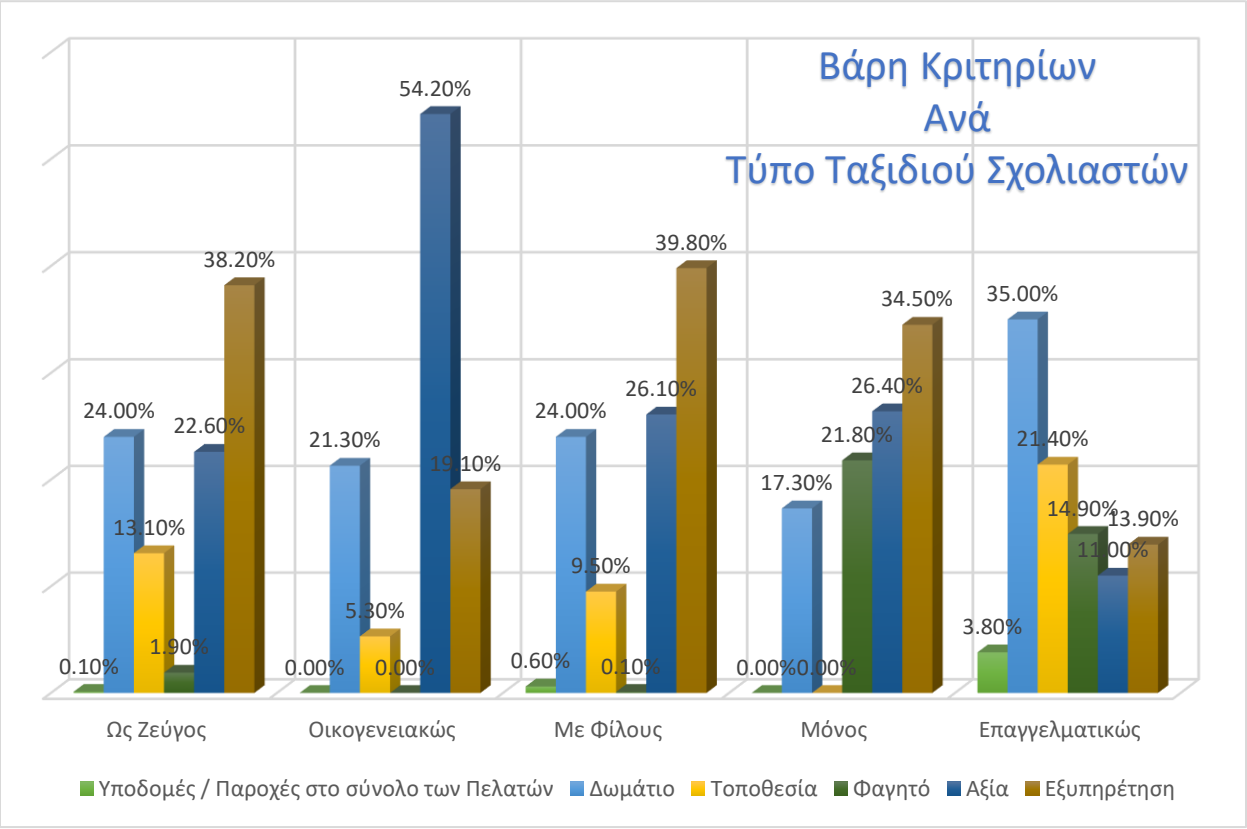
Πίνακας 14: Πίνακας 13: Πίνακας με τις πιο δημολείς Χώρες Καταγωγής των Σχολιαστών για τα Ξενοδοχεία Ανά Περιοχή της Ελλάδας

Ακολουθεί η ανάλυση των σχολιαστών που πραγματοποιήθηκε με βάση τον Τύπο Ταξιδιού. Οι κατηγορίες που χωρίζει το TripAdvisor και η κατανομή τους στα δεδομένα είναι όπως φαίνονται στο γράφημα.



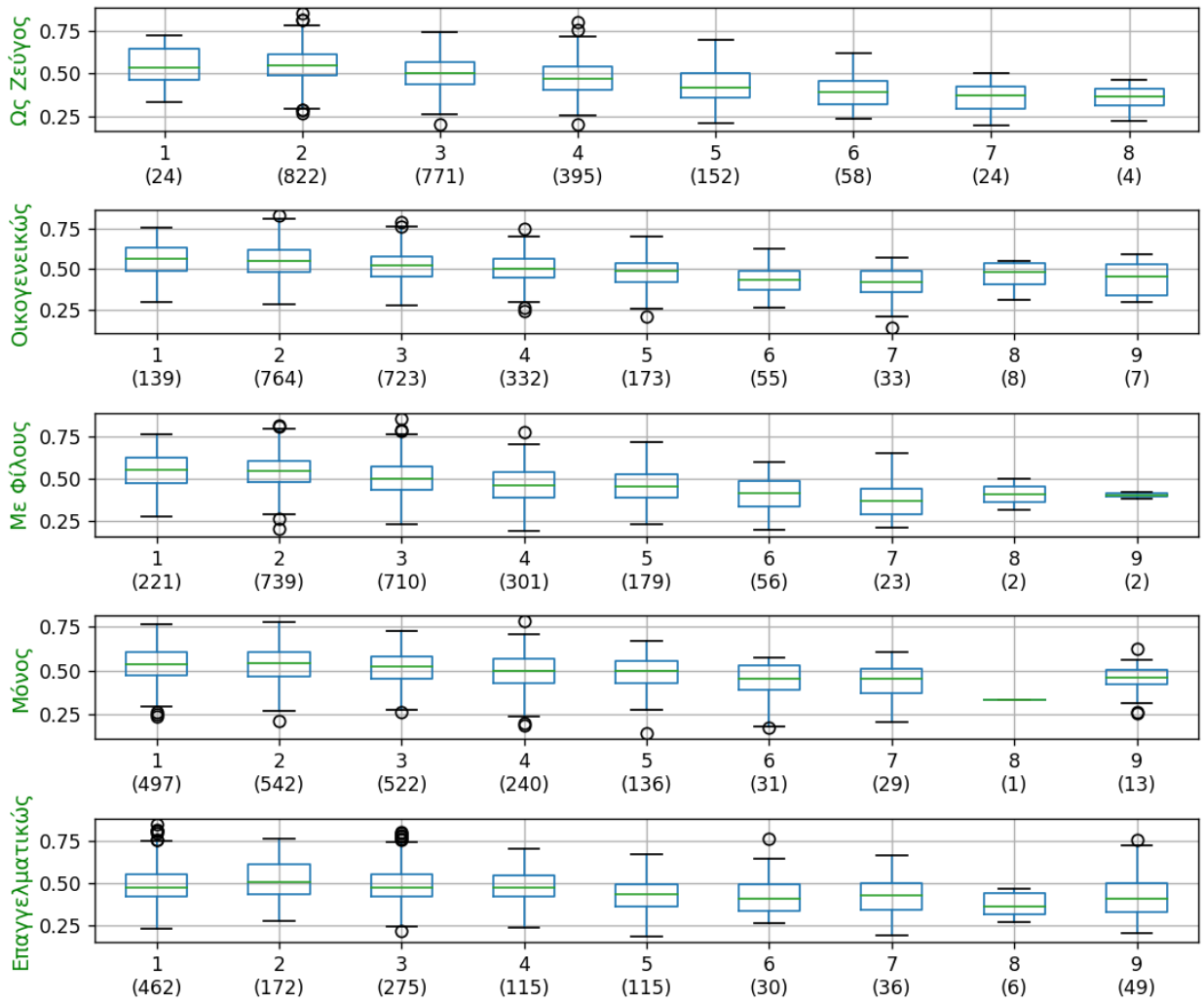
Σχήμα 13: Κατανομή Σχολίων ανά Τύπο Ταξιδιού

Στη συνέχεια χωρίστηκε το σύνολο των δεδομένων των σχολίων σε ομάδες ανά Τύπο Ταξιδιού, με σκοπό της εξαγωγή συμπερασμάτων για τους σχολιαστές κάθε ομάδας. Με βάση τη μέση αξιολόγηση κάθε ομάδας υπολογίστηκε η κατάταξη των ξενοδοχείων, η οποία με τη σειρά της χρησιμοποιήθηκε στο πολυκριτήριο πίνακα. Για τον υπολογισμό των βαρών των κριτηρίων για κάθε ομάδα σχολιαστών, χρησιμοποιήθηκε η κατάταξη των ξενοδοχείων τους ως εισαγωγή στην UTASTAR και έγινε εφαρμογή του αλγορίθμου. Τα αποτελέσματα της διαδικασίας αυτής παρουσιάζονται παρακάτω.



Διάγραμμα 13: Βάρη Κριτηρίων με βάση τον Τύπο Ταξιδιού των σχολιαστών

Από το Διάγραμμα 13, για τους τύπου ταξιδιού που αποτελούν την πλειονότητα των σχολίων (Ζεύγος, Οικογενειακώς και με φίλους) δεν διαφαίνεται κάποια εξαιρετική διαφορά στην κατανομή των βαρών, μεταξύ τους και με τα συνολικά βάρη, με εξαίρεση τα δύο πιο σημαντικά κριτήρια αξία και εξυπηρέτηση, με την Αξία να είναι πιο σημαντικό για την οικογένεια και την Εξυπηρέτηση για τους άλλους δύο. Για τις κατηγορίες Μόνος και Επαγγελματικώς είναι μικρό το δείγμα που υπάρχει για τα δεδομένα οπότε οι διαφοροποιήσεις που παρουσιάζονται μάλλον οφείλονται εκεί και δεν είναι αντιπροσωπευτικές του τύπου ταξιδιού.



Διάγραμμα 14: BoxPlot Κατανομής Κατάταξης Ξενοδοχείων από την TOPSIS σε σχέση με την κατάταξή τους από το TripAdvisor Ανά Τύπο Ταξιδιού

Στη συνέχεια τα βάρη από την εφαρμογή της UTASTAR χρησιμοποιήθηκαν ως είσοδο στην TOPSIS μαζί με τον πολυκριτήριο πίνακα, με σκοπό τη σύγκριση των αποτελεσμάτων του αλγορίθμου, ομοιότητας των ξενοδοχείων της καλύτερης λύσης TOPSIS, με την κατάταξη τους ανά Τύπο Ταξιδιού. Τα αποτελέσματα της ανάλυσης αυτής φαίνονται στο Διάγραμμα 14 σε BoxPlot κατανομής και για τις ομάδες (Ζεύγος, Οικογενειακώς και Με Φίλους) έχουν την φθίνουσα πορεία ανά χειρότερη κατάταξη που είναι το επιθυμητό αποτέλεσμα και είναι σημάδι ότι τα βάρη αυτά είναι όντως αντιπροσωπευτικά των ομάδων. Για τις κατηγορίες Μόνος και Επαγγελματικώς το δείγμα δεν είναι αρκετό για να εξαχθούν συμπεράσματα, όπως εξηγήθηκε προηγουμένως, για αυτό το λόγο και τα γραφήματά τους δεν έχουν τις επιθυμητές ιδιότητες.

Κεφάλαιο 5: Συμπεράσματα

Η εργασία αυτή είχε ως σκοπό την ανάλυση της αγοράς των ξενοδοχείων του Ηρακλείου και των ανταγωνιστικών περιοχών του και την ανάλυση των προτιμήσεων και χαρακτηριστικών των σχολιαστών που επισκέφτηκαν τις περιοχές αυτές. Το σύνολο των δεδομένων που μελετήθηκαν προήλθε από την συλλογή "web crawling" από την ιστοσελίδα *Tripadvisor* . Για την πραγματοποίηση των παραπάνω αναλύσεων χρησιμοποιήθηκαν τεχνικές ανάλυσης κειμένου (εξόρυξης θεμάτων αναφοράς και ανάλυσης συναισθήματος) μέσω βιβλιοθηκών της *Python* και αλγόριθμοι πολυκριτήριας ανάλυσης.

Τα αποτελέσματα που προέκυψαν αμφίβολης ποιότητας καθώς από την εφαρμογή της UTASTAR επιτεύχθηκαν χαμηλές τιμές του δείκτη t Kendall, αλλά η επαλήθευση των αποτελεσμάτων από την TOPSIS δείχνει ενθαρρυντικά σημάδια.

Για την βελτίωση των αναλύσεων σε επόμενες αναλύσεις θα ήταν χρήσιμο να δοθεί περισσότερη έμφαση στην επιλογή των ανταγωνιστικών περιοχών και του είδους των ξενοδοχείων που θα επιλεχθούν ώστε να είναι πιο συγγενή τα δεδομένα. Κάτι τέτοιο θα έχει ως αποτέλεσμα μικρότερη διαφοροποίηση μεταξύ των δεδομένων και αυξημένο δείκτη εμπιστοσύνης των αποτελεσμάτων.

Σχετικά με τις μελλοντικές προεκτάσεις που μπορούν να γίνουν στην μεθοδολογία της εργασίας, είναι η χρήση πιο εξελιγμένων μεθόδων *deep learning* οι οποίες χρησιμοποιούν μηχανική μάθηση και εκπαιδεύονται από τα ίδια τα δεδομένα, οι οποίες θα είναι πιο αποτελεσματικές σε δεδομένα μεγάλου όγκου και θα βοηθήσουν στην εξαγωγή πηγαίων χαρακτηριστικών των δεδομένων.

Βιβλιογραφία

1. Mohammad Tubishat, & Norisma Idris, & Mohammad A.M. Abushariah, (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges, Information Processing & Management, 54(4), 545-563, (Mohammad Tubishat, 2018)
2. Priyantina, Reza & Sarno, Riyanarto. (2019). Sentiment Analysis of Hotel Reviews Using Latent Dirichlet Allocation, Semantic Similarity and LSTM, International Journal of Intelligent Engineering and Systems, 12, 142-155 (Priyantina & Sarno, 2019)
3. Ru-xin Nie, & Zhang-peng Tian, & Jian-qiang Wang, & Kwai Sang Chin, (2020). Hotel selection driven by online textual reviews: Applying a semantic partitioned sentiment dictionary and evidence theory, International Journal of Hospitality Management, 88 (Nie, Tian, & Wang, 2020)
4. Yue Guo, & Stuart J. Barnes, & Qiong Jia, (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation, Tourism Management, 59, 467-483 (Guo, Barnes, & Jia, 2017)
5. Sutherland I, & Sim Y, & Lee SK, Byun J, & Kiatkawsin K. (2020). Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation. Sustainability. 12(5), :1821 (Sutherland & Sim, 2020)
6. Chetty, & Gautami Tripathi, & Naganna S., (2015), Feature Selection and Classification Approach for Sentiment Analysis, Machine Learning and Applications: An International Journal (MLAIJ), 2, 1-16, (Chetty, Gautami, & Naganna, 2015)
7. Rosario Barbara, (2000), Latent Semantic Indexing: An overview, INFOSYS, 240 (1-16), (Rosario, 2000)
8. Blei, & David M., & Andrew Y. Ng., & Michael I. Jordan, (2003), Latent Dirichlet Allocation, Journal of machine Learning research, 3 (993-1022), (Blei, David, Andrew, & Michael, Allocation)
9. Walaa Medhat, & Ahmed Hassan, & Hoda Korashy, (2014). Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal, 5(4), 1093-1113 (Medhat, Hassan, & Korashy, 2014)
10. Lexalytics, API: Semantria, (url: https://semantria-docs.lexalytics.com/reference/semantria-api), (Lexalytixs\Semantria, n.d.)
11. Jacquet-Lagreze,& Siskos Yannis, (1982), Assessing a set of additive utility functions for multicriteria decision making The UTA method, European Journal of Operational Research, 10(2), 151-164, (Jacquet-Lagreze & Siskos, 1982)
12. Siskos Y., & E. Grigoroudis , & N.F. Matsatsinis (2016), UTA methods, in: S. Greco, M. Ehrgott, J. Figueira (eds.), Multiple Criteria Decision Analysis, - State of the Art – Surveys (2nd Edition), International Series in Operations Research and Management Science, vol. I, pp. 315-362, Springer., (Siskos, Grigoroudis, & Matsatsinis, 2016)
13. Siskos, Y., & Yannacopoulos, D. (1985). UTASTAR: An ordinal regression method for building additive value functions. Investigaçao Operacional, 5(1), 39-53, (Siskos & Yannacopoulos, 1985)

14. Lai Young-Jou, & Liu Ting-Yun, & Huang Ching-Lai, (1994), Topsis for MODM, European Journal of Operational Research, 76(486-500), (Lai, Liu, & Hwang, 1994)
15. Röder M., & Both A., & Hinneburg A., (2015), Exploring the Space of Topic Coherence Measures, WSDM, 15(399*408), (Röder, Both, & Hinneburg, 2015)
16. Gaurav Kumar, & N. Parimala, (2020). A weighted sum method MCDM approach for recommending product using sentiment analysis, International Journal of Business Information Systems, 35(2), 185-203 (Kumar & Parimala, 2020)
17. Michael Röder, & Andreas Both, & Alexander Hinneburg, Exploring the Space of Topic Coherence Measures