ΕΘΝΙΚΟ
ΜΕΤΣΟΒΙΟ
ΠΟΛΥΤΕΧΝΕΙΟ

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

**ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ “ATHENS MBA”**

Τίτλος μεταπτυχιακής διατριβής

# TOURISM RECOMMENDATION SYSTEM BASED ON USER GENERATED CONTENT

Πανούτσος-Αρχοντής Παναγιώτης

Επιβλέπων:Καρδαράς Δημήτριος, Ε. Καθηγητής

Αθήνα, Φεβρουάριος 2019

## ΔΗΛΩΣΗ ΕΚΠΟΝΗΣΗΣ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

«Δηλώνω υπεύθυνα ότι η συγκεκριμένη μεταπτυχιακή εργασία για τη λήψη του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Διοίκηση Επιχειρήσεων, έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει εγκριθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό. Η εργασία αυτή έχοντας εκπονηθεί από εμένα, αντιπροσωπεύει τις προσωπικές μου απόψεις επί του θέματος. Οι πηγές στις οποίες ανέτρεξα για την εκπόνηση της συγκεκριμένης μεταπτυχιακής αναφέρονται στο σύνολό τους, δίνοντας πλήρεις αναφορές στους συγγραφείς, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο».

Ονοματεπώνυμο Υπογραφή

Πανούτσος-Αρχοντής Παναγιώτης

# Contents

# Abstract

Recommendation Systems (RSs) are software tools and techniques, which provide suggestions for items to users. RS assist users in finding their way through huge databases and catalogues, by filtering and suggesting relevant items taking into account user's preferences (i.e., tastes, interests, or priorities). The explosive growth of available data online not only transformed customers into sophisticated users, who search online for unbiased information, but also created an information overload problem. The aim of this thesis is to utilize user generated content in order to provide successful recommendations to users for tourism services and especially hotels. User generated content is a source with rich customer information, which enable us to capture and understand users' interests and needs. With a web scraping tool users' reviews were extracted form TripAdvisor. These reviews are the base of the analysis and though several approaches a tourism recommendation system is built. The first step of the analysis is users' interests modelling through keyword extraction from the reviews. Users' interests are categorized and form subsets. Collaborative filtering approach is applied and the system is able to generate suggestions based on users' interests. In addition, sentiment analysis is performed to evaluate the polarity of the reviews and classify reviews accordingly. Then content based approach is applied in order to provide recommendations based on the similarity of contents' attributes. Finally, Fuzzy set Qualitative Comparative Analysis is utilized to identify if there is a casual combination between users' interests and provided ratings to the hotels. In this thesis a tourism recommendation system, which can provide personalized suggestions to the user, was designed and implemented successfully.

# Περίληψη

Οι χρήστες κατά την περιήγησή τους στο διαδίκτυο έρχονται αντιμέτωποι με ένα τεράστιο όγκο δεδομένων που ενίοτε τους δυσκολεύει στη εύρεση και επιλογή των αντικειμένων, προϊόντων και υπηρεσιών , που ταιριάζουν περισσότερο στις ανάγκες και τα ενδιαφέροντά τους. Τα συστήματα συστάσεων καλούνται να επιλύσουν το συγκεκριμένο πρόβλημα και λειτουργώντας επιβοηθητικά παρέχουν προτάσεις στους χρήστες, ανάλογα με τα ενδιαφέροντά τους. Σκοπός της εργασίας είναι η αξιοποίηση περιεχομένου που έχει δημοσιευθεί στο διαδίκτυο από χρήστες, προκειμένου να δημιουργηθεί ένα σύστημα συστάσεων που θα παρέχει προτάσεις για τον κλάδο του τουρισμού και κυρίως για ξενοδοχεία. Για την οικοδόμηση του συστήματος συστάσεων έγινε εξόρυξη κριτικών χρηστών από το TripAdvisor. Το πρώτο βήμα της ανάλυσης είναι η εξαγωγή λέξεων-κλειδιών, που περιγράφουν τα ενδιαφέροντα των χρηστών, από τις κριτικές . Οι λέξεις-κλειδιά κατηγοριοποιούνται και σχηματίζουν υποομάδες ενδιαφερόντων. Μέσω συστήματος συστάσεων που βασίζεται στη συνεργασία μπορούμε να παρέχουμε προτάσεις στον χρήστη βάσει των ενδιαφερόντων του. Παράλληλα διεξάγεται συναισθηματική ανάλυση των κριτικών για να κατανοήσουμε την πολικότητα τους, δηλαδή εάν είναι θετικές ή αρνητικές, χωρίς την ύπαρξη βαθμολογίας από τους χρήστες. Μέσω συστήματος συστάσεων βάσει περιεχομένου μπορούμε να προτείνουμε στον χρήστη ξενοδοχεία με παρόμοια χαρακτηριστικά με εκείνο που έχει ήδη αξιολογήσει. Τέλος μέσω ποιοτικής και ποσοτικής ανάλυσης εξετάζεται η αιτιώδης σχέση μεταξύ διαφορετικών ομάδων ενδιαφερόντων των χρηστών και της βαθμολογίας των ξενοδοχείων. Σε αυτή την εργασία σχεδιάστηκε και υλοποιήθηκε επιτυχημένα ένα σύστημα συστάσεων που παρέχει προσωποποιημένες προτάσεις για ξενοδοχεία στους χρήστες.

# Chapter 1

## Introduction

Since the internet is acknowledged as a powerful tool in business processes many researchers have tried to reveal its impact on consumer behavior. Consumer behavior is how people make decisions about what they buy, want, need, or act in regards to a product, service, or company. Consumer behavior is affected by personal, psychological and social factors. Customers decide and act differently, based on their perceptions and attitudes. Personal interests, tastes and opinions vary significantly, while age, gender, culture, background and personal interactions can influence the decision process.

The evolution of the internet has fundamentally changed the way customers perceive and purchase products and services. The growing use of Web 2.0 platforms, like social media and blogs has enabled users not only to access information, but also to contribute and share their opinions. Users have become sophisticated customers, who search online for unbiased information, that will guide them to decide. The explosive growth of available data and Internet users have created an information overload problem. In the past vendors knew their customers personally and could make recommendations to them based on a personal knowledge of past purchases. This type of personal relationship meant that customers would receive great customer service, while vendors were able to reap the benefit of brand loyalty since they understood their customer’s needs, preferences, and even their budget. This fact initiated the development of recommendation systems.

In daily decisions, individuals usually rely on recommendations provided by others. For example, people usually rely on their friends’ opinion when selecting a movie to watch or read a review written by a movie critic. Nowadays the growth of information can overwhelm internet users and lead them to poor decisions. In 2018 there were 4 billion internet users, a number which was increased more than 42% since 2014, 5,2 billion google searches and 22 billion text sent on a daily basis. All this information is not always useful for the user, as the choice paradox occurs. While a large amount of choice is commonly associated with welfare and freedom, too much choice causes the feeling of less happiness and less satisfaction. The available choices should be personalized and become suitable for the needs of each user. As the demand for personalized services in several business sectors increases, recommender systems are emerging and applied in many different domains.

# Chapter 2

## Recommendation systems overview

### 2.1 Introduction

Recommender Systems (RSs) are software tools and techniques, which provide suggestions for items to users. The aim of Recommender System (RS) is to assist users in finding their way through huge databases and catalogues, by filtering and suggesting relevant items taking into account user's preferences (i.e., tastes, interests, or priorities). A RS normally focuses on a specific type of item (e.g., movies, news) and accordingly its design, its graphical user interface, and the core recommendation technique used to generate the recommendations are all customized to provide useful and effective suggestions for that specific type of item.

In order to understand the possible roles a RS can play there must be a discrimination between the role played by the RS on behalf of the service provider from that of the user of the RS. For example, a travel recommendation system is typically introduced by a travel intermediary (TripAdvisor) to sell hotel rooms, while the user access the system to find a suitable room.

A service provider may use RS for several reasons:

- Increase the number of items sold: Sell an additional set of items compared to those usually sold without any kind of recommendation. This goal is achieved because the recommended items are likely to suit the user's needs. Generally, the primary goal for introducing a RS is to increase the conversion rate, i.e., the number of users that accept the recommendation and consume an item, compared to the number of simple visitors that just browse through the information.
- Sell more diverse items: Assist user select items that might be hard to find without a precise recommendation.
- Increase the user satisfaction: A well-designed RS can improve the experience of the user. Effective recommendations and a usable interface will increase the user's evaluation of the system and increase system's usage.
- Increase user fidelity: The longer a loyal user visits the system, the more the recommender output can be effectively customized to match the user's preferences.
- Better understanding of the user's needs

The service provider can re-use the knowledge of user’s preferences for a number of other goals such as improving the management of the item’s stock or production.

Users may also need a RS, if it will support their goals. Herlocker (2004) define eleven popular tasks that a RS can assist in implementing:

- Find some good items: Recommend to a user some items as a ranked list.
- Find all good items: Recommend all the items that can satisfy some user needs.
- Annotation in context: Given an existing context, emphasize on some of them depending on the user’s long-term preferences
- Recommend a sequence: Recommend a sequence of items that is pleasing as a whole. For example, a compilation of musical tracks
- Recommend a bundle: Suggest a group of items that fits well together.
- Just browsing: The user browses without any intention of purchasing an item and the task of the recommender is to help him browse the items that are more likely to fall within the scope of the user’s interests.
- Find credible recommender: Some users do not trust recommender systems thus they play with them to see how good they are in making recommendations.
- Improve the profile: The user provides information to the recommender system about what he likes and dislikes, in order to provide more provide personalized recommendations.
- Express self: For some users is important to contribute with their ratings and express their opinions and beliefs.
- Help others: Some users are happy to contribute with information, because they believe that the community benefits from their contribution.
- Influence others: Users whose main goal is to influence other users into purchasing particular products.

### 2.2 Recommendation techniques

Recommendation systems can be classified on several bases and the categorization is mainly based on the following criteria:

- Data Mining techniques
- Approaches used
- Domain

Recommender systems are generally classified into collaborative filtering (CF), content-based filtering (CB), demographic, community based and hybrid.

Content-based (CB): This technique analyzes attributes of items and generate predictions. Features from the content of items previously evaluated by the user are extracted and are the base of CB filtering. The system learns to recommend items related to positively rated items. The similarity of items is calculated based on the features associated with the compared items. For example, if a user has positively rated a book which belongs to the mystery fiction genre, then the system can learn to recommend other books from this genre. Through the user's previous evaluation, the system can understand the underlying model and provide meaningful recommendations with statistical analysis or machine learning techniques. CB filtering does not need the profiles of other users and can adjust the recommendations quickly, when the user profile changes. The similarity is measured with vector space models, like TF/IDF and probabilistic methods, like naıve Bayes classifier and decision trees.

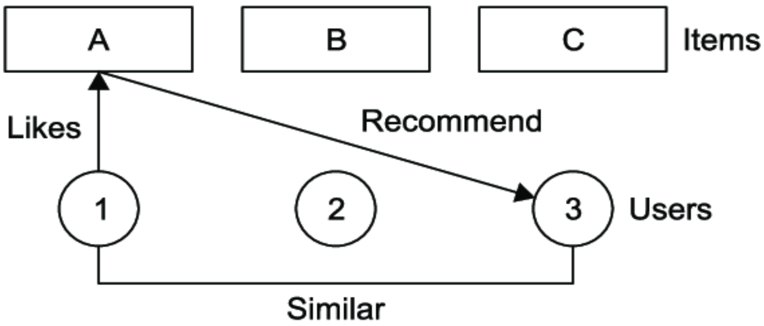
The recommendation process has three steps

- Content analysis: Usually the information has no structure and must be transformed in a structured form, in order to be useful. Data from various information sources are extracted and analyzed with feature extraction techniques. The structured data are the input to the next steps.
- Profile learning: The construction of the user's profile is a generalization of his preferences in the past.
- Filtering: Based on user's profile the system computes similarity of items and makes relevant recommendations

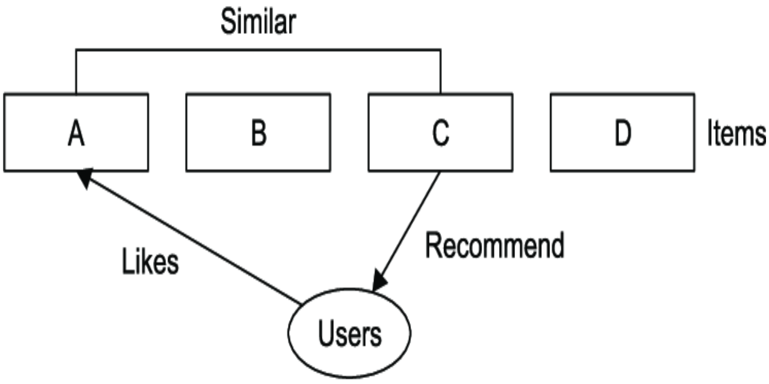
There are two techniques to record user's feedback, implicit and explicit. The implicit way does not need the involvement of the user, as his actions are monitored and evaluated by the system. The explicit technique builds the model based on his likes/dislikes, ratings and text comments.

Collaborative filtering (CF): The idea behind CF is that similar users share similar taste and that similar items are liked by a user. CF recommends a product to the target user based on products already rated by similar users. Unlike CB, collaborative filtering can provide recommendations for items which the target user has never consumed but other users with similar taste have rated positively. There are mainly two approaches to compare the similarity, the neighborhood approach and the latent factor models items (model based).

Neighborhood CF uses ratings already given to items by the users, to predict ratings for new items. Neighborhood CF has two strategies, user based and item based recommendations. User based systems measure the interest of the target user for an item i, comparing the rating to item i from similar users. Similarity between two users is calculated by finding an item that they have both interacted with and by analyzing their behavior with the item.




Item based CF recommends an item i to the target user, by taking into account the ratings of the target user for items similar to the item i.




Similarity Computation

Similarity computation is essential to the identification of the similar users (neighbors) and for the importance which must be given to them. In order to measure the similarity a popular option is Pearson Correlation.

User based

$$\mathrm{PC}(u,v) = \frac{\sum\limits_{i \in \mathcal{I}_{uv}} (r_{ui} - \overline{r}_u)(r_{vi} - \overline{r}_v)}{\sqrt{\sum\limits_{i \in \mathcal{I}_{uv}} (r_{ui} - \overline{r}_u)^2 \sum\limits_{i \in \mathcal{I}_{uv}} (r_{vi} - \overline{r}_v)^2}}$$

Item based

$$\mathrm{PC}(i,j) = \frac{\sum\limits_{u \in \mathcal{U}_{ij}} (r_{ui} - \overline{r}_i)(r_{uj} - \overline{r}_j)}{\sqrt{\sum\limits_{u \in \mathcal{U}_{ij}} (r_{ui} - \overline{r}_i)^2 \sum\limits_{u \in \mathcal{U}_{ij}} (r_{uj} - \overline{r}_j)^2}}$$

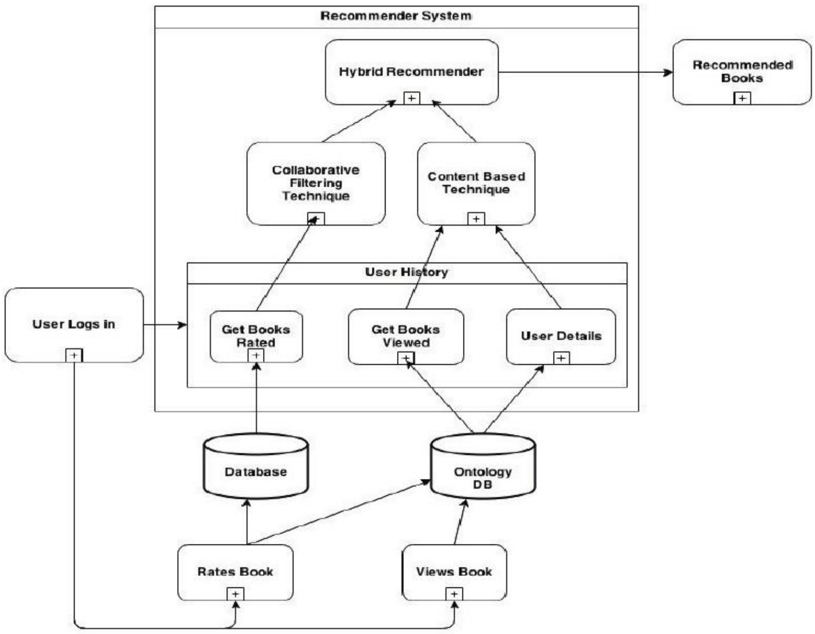
Pearson correlation has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. Neighborhood collaborative filtering is one of the first recommendation techniques proposed and remains popular, due to its simple implementation. Recommendations can be justified to the user with a list of his neighbors and neighborhood CF can be used efficiently to large systems. Comparing user based and item based approach, item based CF is more efficient when the number of users is larger than the number of items. The advantage of item based approach is that item similarity is more stable, as it is not based on user preferences which may change often.

Demographic**:** The core idea of this technique is that different demographic niches have distinctive interests. Recommendations can be based on the language of the user, the country, even his age. For example, a news recommendation system will provide different recommendations to a middle-age male user and to a young woman.

Community based: This technique is based on the interests of the users' friends. It is assumed that the user and his friends share common preferences towards items and the recommendations are generated from the ratings of the community. As social media are a daily routine for the majority of the users, community based RS will gain popularity.

Hybrid: These systems are a combination of above mentioned techniques. A hybrid system overcomes the disadvantages and exploits the advantages of each system. For example, CF suffers from the cold start problem, i.e. CF cannot provide recommendations to new users with limited ratings. CB can provide the necessary information, as the features of the most items are usually available. Generally, hybrid

RS outperform the other techniques, when they are implemented individually. The figure below is an example of a hybrid book recommendation system.



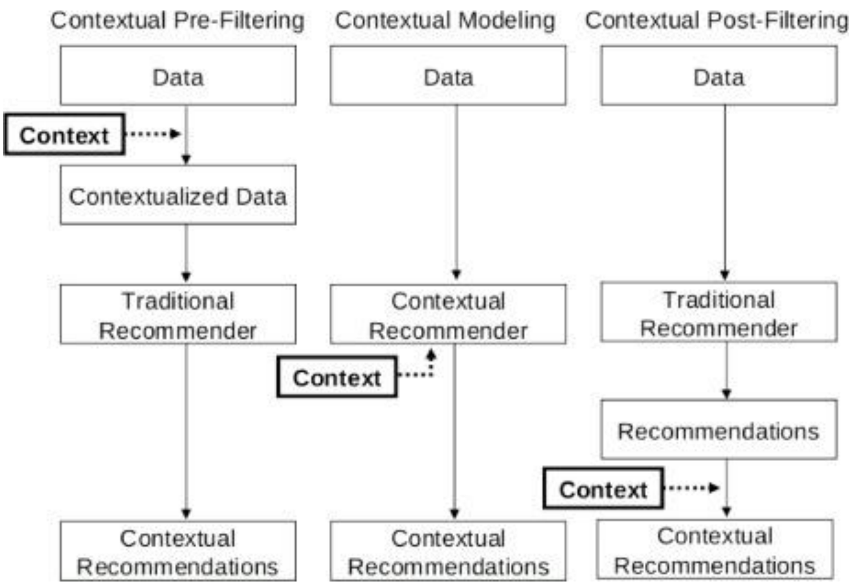

Context- aware recommendation systems

Context is the situation within which something exists or happens, and that can help explain it. In order to improve users' experience, RS must not provide only relevant recommendations but also take into account the context, the circumstances when the recommendations are provided. The users profile can change, as his preferences can differ even in the same day. A user may like to read political news in the morning going to work and sports news in the afternoon, when he returns home. A news RS must not ignore the context, providing only proposals for political news . Consumer behavior is dynamic and a user can have several profiles, based on the context. For example, a user likes romantic movies when he is alone and adventure movies when he is with his friends. Moreover, a user may have different taste when he buys something for himself and when he buys a gift . Users consumer behavior can also vary on the basis of the purchasing situation, i.e. if he purchases items alone from the site of a brand or with the recommendations of a digital assistant in Amazon. Without considering context, RS are trying to predict the rating of the target user for new items, not yet rated. With the context the rating function (R) is:
R: user × item × context → rating

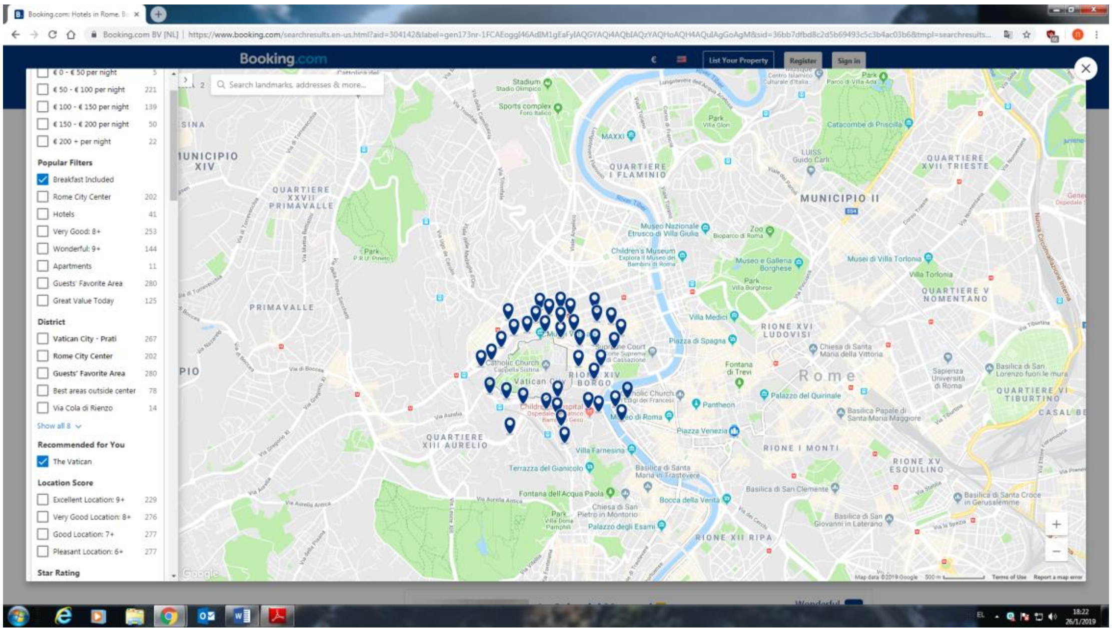
Consider an application for book recommendations to users and the relations with users' consumer behavior. Assume that the user is not willing to order the book online.

- User: the person to whom books are recommended
- Books: the books which can be recommended to the user
- Bookstore: bookstores having available the books, near to the users' location
- Time: when the bookstores are open

The recommendations for books to the user vary and depend on several factors. Recommendations may differ if the user wants to buy a book on Monday afternoon, when the majority of bookstores are closed or he wants to purchase a rare book, available only in few bookstores on Saturday morning.




A typical example of a touristic recommendation system that takes into account context is Booking. When a user searches for a hotel in a destination Booking asks several questions before providing any recommendation. User needs to enter the dates of his trip and the preferred destination. Then the system provides many options in order to customize the recommendations. Filters are available to the user to shorten the list of the proposals. Popular filters are location score, rating of the hotel, property type, landmarks near the hotel, companion, bed preferences and the budget. For instance, a user is travelling alone in Rome and his interested mainly to book a hotel in the Vatican area and secondarily in breakfast. Booking will recommend the hotels shown in the following image and the user can view them in the map.




## 2.3 Advantages and limitations of recommendation techniques

Each type of recommendation technique has strengths and weaknesses. In this section the main advantages and limitations of each technique are examined.

<u>Content based RS:</u>

Advantages:

- User Independence: Content based RS use only ratings already given by the target user to build his profile, while collaborative filtering RS need ratings from other users with similar interests to provide recommendations.
- New item: CB technique can recommend new items, even if there are no ratings for the item and recommendation accuracy is not affected. Instead CF techniques need a significant number of users to rate the item, before providing any recommendation.
- Transparency: As users need explanations to trust a recommendation, CB can specify items description that caused the recommendation.

Limitations:

- Limited content analysis: Content-based recommendations depend on the available features explicitly associated with the items. These features should be in a form that can be automatically parsed by a computer, or manually extracted, which, depending on the domain, can be unfeasible or very difficult to maintain. CB recommendation systems cannot provide suitable suggestions if the analyzed content does not contain enough information to discriminate items the user likes from items the user does not like.

- New user: A user has to show some preference (ratings) for a sufficient number of items before the system can build a reliable content-based user profile. The system cannot recommend items to a new user with no or few ratings.
- Overspecialization. Content-based technique has no method for finding something unexpected(novel). The system recommends similar items with items already rated and the user is restricted by his profile.

<u>Collaborative filtering RS:</u>

Advantages: Collaborative filtering RS have some important advantages over CB. When the content of items is not known, items can be proposed based on other users' ratings.CF can provide novel recommendations, as items with different content are proposed and the quality of an item as an indicator is measured by peers of users.

Limitations:

- Rating data sparsity: The number of observed user-item interactions (e.g. ratings) is generally very small compared to the number of all user-item pairs. This fact may cause CF algorithms to produce unreliable recommendations, since they have been inferred from insufficient data.
- Grey sheep: Since collaborative recommendations rely on the tastes of similar people to suggest new items, when a user has very specific or unusual preferences, it will be more difficult for the system to find good neighbors and recommend interesting items.
- New item: Until a new item has been rated by a significant number of users, a recommender system may not be able to recommend it. Therefore, popular items tend to have advantage in this kind of systems.
- New user: Like in the content-based approaches, until a user has not provided with enough ratings, the system is unable to recommend her interesting, unknown items.

| Problem | Description | CB | CF |
|---|---|---|---|
| Limited content analysis | Items to be recommended must have available data related to their features. The data are often unavailable or incomplete. | YES | NO |
| Overspecialization | Recommenders are trained with the content features of the items. All the recommended items are similar to those already rated. | YES | NO |
| New user | A user has to rate enough items in order to show his preferences. When a new user enters into the system he has no ratings . | YES | YES |
| New item | Items have to be rated by a substantial number of users for being recommended. Recently incorporated items have insufficient ratings. | NO | YES |
| Grey sheep | A user has to be similar to others in the community to receive recommendations. Users whose tastes are unusual may not receive useful suggestions. | NO | YES |
| Rating data sparsity | Ratings are used to train user and item models. The number of available ratings is usually small. | NO | YES |

### 2.4 Recommendation systems evaluation

In general, a recommender system needs to complete the performance evaluation of three stages: offline analysis, user study, and online experiment. Offline analysis does not require user interaction, as it uses datasets to calculate the corresponding evaluation metrics, such as the prediction accuracy and coverage. Offline analysis is the easiest to implement and costs the least among the three types of methods. User study requires testers to use the recommender system, perform a series of tasks, and then answer a set of questions about their experiences on the system, and finally the results of evaluation will be given through statistical analysis. Online experiment executes a large-scale experiment on a deployed recommender system. It evaluates the recommender system by the real tasks executed from real users. The evaluation results of the online experiment are the closest to the real situations when the recommender system runs online.

Offline evaluation

The basic method of offline analysis on recommender systems divides the dataset into the training dataset and testing dataset, and then constructs recommendation models on the training dataset and tests its performance on the testing dataset. These datasets can be used to simulate the interactions between users and the recommender systems. The main targets of offline analysis are to compare the performance of the recommendation algorithms in some metrics, to filter inappropriate algorithms and to remain some candidate algorithms.

User study

User study is an important method for evaluating recommender systems. This method tests the interaction between users and the recommender systems, and can obtain the influence of the recommender systems on the users. User study can also be used in collecting qualitative data, and these data are of great importance in explaining the quantitative results. In order to run the test, some candidates should be recruited to do user study, and be required to do some tasks using the recommender systems. When testers execute the tasks, their behavior is observed and recorded and the situations of their tasks collected, such as which tasks are completed, and how much time is consumed on the tasks and the accuracy of the tasks' results.

Online experiment

Online experiment is to execute a large-scale testing on a recommender system which is already deployed. Online experiment can be used to evaluate or compare different recommender systems by the real tasks carried out by real users. Online experiment can achieve the most real testing results among the three evaluation methods. The advantages of online experiment are that, the entire performance of the recommender systems can be evaluated, such as long-term business profit and users' retention, rather than some single metrics. Therefore, online experiment can be used to understand the impact of the evaluation metrics (such as the accuracy in prediction, diversity in recommendation) on the overall performance of the system.

**Performance evaluation metrics**

Prediction accuracy

The metric of prediction accuracy is essentially about the error of prediction. This is a common metric in various machine learning algorithms evaluation, such as regression or classification. This metric is mainly used to measure the ability to predict users' behaviors. Prediction accuracy is the most important metric in the offline analysis of recommender systems. When calculating prediction accuracy, a set of offline dataset is needed that contains users' scores, such as users' ratings for a product or movie. The dataset is divided into training set and testing set. A users' rating prediction model is trained and then the prediction of users' rating is computed on the testing set. The error is the deviation between the predicted rating and the real rating. There are three metrics to measure the prediction accuracy: Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE), and the formulas are as follows:

Mean Absolute Error: $MAE = \frac{1}{|Q|}\sum_{(u,i)\in Q}|r_{ui} - \hat{r}_{ui}|$

Mean Square Error: $MSE = \frac{1}{|Q|}\sum_{(u,i)\in Q}(r_{ui} - \hat{r}_{ui})^2$

Root Mean Square Error: $RMSE = \sqrt{\frac{1}{|Q|}\sum_{(u,i)\in Q}(r_{ui} - \hat{r}_{ui})^2}$

where $Q$ is the test set, $rui$ represents the user's true ratings, $\hat{r}ui$ represents the prediction ratings of the recommender system. MAE is the simplest, but it does not take into account the direction of the error (positive error or negative error). MSE has a larger penalty on large errors and the squared error does not have an intuitive meaning. Therefore, RMSE is more widely used in computing the prediction accuracy of the recommender system.

The possible results of a recommendation to user can be the following:

| | Recommended | Not recommended |
|---|---|---|
| Used | True-Positive (tp) | False-Negative (fn) |
| Not used | False-Positive (fp) | True-Negative (tn) |

We can use the precision, recall and F-Measure to evaluate the performance of recommender system. The formulas are:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Usually there is a trade off between precision and recall. Which metric is more important depends on the target of the system and the domain.

Coverage: The percentage of items recommended to total items. The term can be also extended to users, as the percentage of users to whom the system provides recommendations to total users.
Diversity: Even if the initial target of a RS is to recommend items based on similarity, it is not always useful. When a user has already bought a new mobile phone the system should stop recommending phones and start suggesting other items, like phone cases. Diversity of recommendations should be taken into consideration when a RS is designed, without reduction in accuracy.
Trust: Trust level refers to the level the user believes that the system provides recommendations appropriate for him. If he likes the suggestions he will continue using the system and trust the recommendations. One way to measure trust is with surveys, asking users whether they like the recommendations, or not. RS should explain to the users why and how specific items are recommended, in order to gain the trust of the user.
Novelty: Novel recommendations are suggestions for items that the user is unfamiliar with. Users can not inform the system for all the items they know. One simple solution is not to recommend items already consumed by the user. For example, a music RS should not only recommend users' favorite tracks, but also new artists.
Serendipity: Serendipity measures the ability of the system to surprise the user with recommendations, by finding something unexpected. Random recommendations can increase serendipity, but on the same time trust is reduced.
Real-time**:** The ability of a RS to provide real time recommendations to users, i.e. to suggest new arrivals. Real time contains two parts. The first is the ability of the system to recommend newly added items to the user. The second is the ability of the system to evaluate users' situation/behavior and make successful recommendations accordingly.
Robust: Recommendation systems are based on users' profiles and the interactions between the users and the system. Users may interact with the system not only to get useful recommendations, but also to manipulate the results. For example, a restaurants' owner may create many user profiles to improve the evaluation of his restaurant, which is considered as an attack to the system. RS build attacking models to identify the attacks and reduce their impact. If the users' rating behavior is not close to real user distribution patterns the system will detect the attack and limit its impact.
Scalability: Recommendation systems aim to assist users in finding their way through huge databases and catalogues. One of their targets is to provide quick results to the users, which affects the properties of the algorithm. Algorithms computational

complexity is measured in terms of time and space requirements. Scalability is tested with growing data sets. The designer of the RS should take into account the possibility the accuracy of a recommendation algorithm to be reduced with the growth of the data sets and assess algorithms potential performance. The response time, the needed time to provide recommendations online, is also calculated and evaluated.

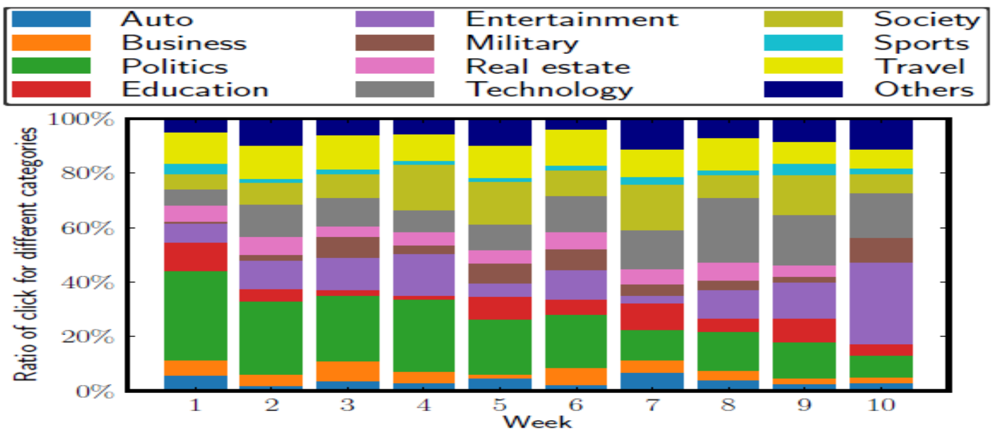
## 2.5 Applications of recommendation systems

Recommendation systems have changed the way users search and purchase products in several domains. The use of RS in commercial applications enhances customer experience, cross-sell/up-sell opportunities and customer loyalty. Businesses can through personalization not only increase sales and improve retention, but also form the consumer behavior by influencing usage patents to the costumers. The most popular areas where RS are applied: news, music, movies, tourism, e-commerce

Movies**:** The best- known example of movie RS is Netflix. The company announced in 2017 a contest with $1 million prize for an algorithm that would increase RS accuracy by 10%. A developer team was awarded in 2009, but the algorithm was never used due to its complexity. Netflix Help Center explains to the costumers how the RS works. The system estimates the likelihood a costumer to watch a movie based on a number of factors, like users' interactions with the system (viewing history, ratings), other members with similar taste and preferences and information about titles (genre, actors, release year). In addition, the system takes into account the time of the day the user watch movies, the device user is watching on and then time user spends on watching. The system does not use demographic information. When a new costumer creates an account, the RS asks the user to select few movie titles he likes. The more recently viewed titles outweigh the initial preferences. Finally, the system personalizes the ranking of each title, based on users' choices and presents the most strongly recommended first. Netflix recommendation system is a hybrid RS which use collaborative filtering and content based approaches. More than 80% of what a typical user watch in Netflix comes from recommendations.
MovieLens is another example of movie RS. It is run by a research lab, which develops tools for data exploration and recommendation and its database contains 26.000.000 ratings for 45.000 movies.

Music: The aim of a music RS is to personalize audio playlists and propose new tracks based on the musical taste of the target user. Spotify is probably the most widely used music RS, with 87 million subscribers. Spotify uses a hybrid recommendation model, combining collaborative filtering, Natural Language Processing and audio models. Audio models analyze the audio tracks and examine time signature, key, mode, tempo, and loudness of each song. Then with item based CF, Spotify recommend tracks to users and solve the problem of popularity bias, from which the most music RS suffer.

News: This category of RS is the most challenging, due to the dynamic nature of news and user preferences. News become outdated very fast and user interest on news may differ, depending on the time. News RS use mainly implicit feedback from the costumers, analyzing the Click Trough Rate (the ratio of users who click on a specific recommendation to the number of total users who view the recommendation). The system is trained to recommend to the user similar news with the articles he has already read. The problem is that users' preferences for news change quickly and with this method, recommendations may be unsuccessful. The following figure represents the clicked categories by a user in ten weeks. Users' interests depend on personal, social and psychological factors. For example, user may enjoy reading sports news in a particular week, because a major sport event is taking place in his town that week or auto news if he is in the process of buying a new car.




Google News, Googles' news aggregator, is a typical news RS, which presents a continuous, customizable flow of articles organized from thousands of publishers and magazines. User can select between personalized news and the most popular articles. The algorithm reviews content automatically, looking for indicators of quality, assessing a story's placement based on the number of user clicks it is attracting, the popular consensus on the trustworthiness of its publisher, the relevance of the story to the reader's current geographical location and the freshness (i.e. publication date and time) of the story in question. Google News is therefore more likely to rank Greek news sites highly when the story concerns a fire in Athens than reports on the same incident from much admired publishers from further afield like The New York Times or Washington Post.

Tourism: Tourism is an activity with complex decision making processes. The tourist has to select destination, restaurants, hotel and take into account several constraints. A tourism RS can make recommendations based on users' interests, Points of Interests, attractions or propose a trip plan. In many cases tourism RS take into account the context, like tourist's current location, weather and the opening hours of the main attractions.

E-commerce: An RS in e-commerce has multiple proposes, mainly to increase the number of products sold. Amazon is the most well-known example of RS implementation. It is estimated that 35% of the Amazon's revenue is generated by the recommendation engine. Amazon currently uses item-to-item collaborative filtering, which scales to massive data sets and produces high-quality recommendations in real time. This type of filtering matches each of the user's purchased and rated items to similar items, then combines those similar items into a recommendation list for the user. Their recommendation algorithm is an effective way of creating a personalized shopping experience for each customer which helps Amazon increase average order value and the amount of revenue generated from each customer.

# Chapter 3

## Methodology

The main objective of the thesis presented here is to utilize user generated content in order to provide successful recommendations to users for tourism services and especially hotels. User generated content is a source of customer information probably more valuable than other types of content, as 86% of the users read online reviews from other users for businesses, products and services. With a web scraping tool, Scrapy, 10276 reviews are extracted from TripAdvisor. This source was selected because TripAdvisor stands out most prominently in terms of usage and content among various travel-related sites that support UGC.

The data include reviews from 10276 individual users for 4153 hotels in Athens, Thessaloniki, Mykonos, Crete and Rome. The hotels' class is 4 and 5 stars and they are rated in a 5-point scale. The first step to build a user interest model is to extract keywords from the reviews, describing their interests. With the free text analysis tool online-utility.org the most frequent words are counted and then grouped in categories. Nine groups are formed that indicate different users' interests. The categories are location, food, service, cleanliness, view, beach/pool, amenities, facilities and bed. Keywords are searched in every review and this way we can understand users' preferences. At this point we know users' profiles and we will exploit this information to provide recommendations.

User based collaborative filtering is applied, in order to recommend to the target user hotels positively rated by other users with common preferences and interests. In this user-based recommendation approach the similarity weight computation will be the guide, which users-neighbors to select and what importance give to them. Pearson correlation is used as a measure of similarity and users with strong linear correlation with the target user will be the base of the recommendations.

Sentiment analysis of the reviews is conducted to determine users' attitude towards hotel. This can be useful in cases where users' ratings are not available. The applications MonkeyLearn and LEXALYTICS are used to identify the polarity of the reviews. In addition, a customized lexicon based approach is applied. The idea behind the approach is to find the most frequent words, describing the feelings of the customer towards the hotels. 75 positive and 35 negative words form a small lexicon, which is the base of the sentiment analysis. In order to analyze and categorize the reviews several rules are tested. Rules are based on the hypothesis that if we subtract the sum of negative words from the sum of positive words and the result is above a threshold we can characterize the review positively or negatively. Then the results of the applications and the custom approach are evaluated and a new hybrid recommendation system is built.

Then a content based approach is applied. The aim of this approach is to recommend to the target user hotels with similar features with the one he has already stayed and rated positively. The attributes of every hotel are found through Booking and TripAdvisor. The similarity of the features is computed with Pearson correlation and sets of similar hotels are formed. Combining the results of the two techniques we can provide recommendations based both in user interests and hotels features.

Another approach to model user interests is by utilizing the Fuzzy set Qualitative Comparative Analysis (FsQCA). QCA examines the similarities and differences between a set of cases to identify conditions that lead to an outcome. The examined outcome is users' ratings. With FsQCA we will determine which sets of users' interests lead to higher rating. Five terms will be tested, location, food, service, cleanliness and view. These terms were selected because they have the highest frequency in the reviews.

Finally, alternative applications of the user generated data beyond recommendation systems will be examined. The data can be used not only by travel intermediates, like TripAdvisor, but also by individual businesses. Competitive analysis can be conducted as the data are a rich source how customers evaluate the hotel and its competition. Moreover, the use of the data from a hotel to provide a unique customer experience will be analyzed.

# Chapter 4

## Case study

### 4.1 Introduction

This chapter focuses on application of RS in tourism. The list of possibilities offered by search engines about destinations, restaurants, hotels, museums or events may be particularly useful, but at the same time overwhelming. RS can assist and provide meaningful suggestions to the users based on their preferences. Travel recommendation systems aim to match the possible alternatives to the user needs, through the analysis of his feedback. Surveys have shown that travel preferences, travel intention and destination choice behavior depend on personality factors, travel experiences, word-of-mouth(WOM) and e-WOM. In the past, when a costumer planned a vacation, he used to address travel agents for recommendations. Within recent years, costumers have become sophisticated users, who search online for unbiased information.

### 4.2 User generated content (UGC)

User-generated content (UGC) is any form of content, such as images, videos, text and audio, that have been posted by users online. There are many types of user-generated content:

- Internet forums, where people talk about different topics.
- Blogs where users can post their opinion about many topics
- Product reviews on a supplier website or in social media
- Wikis such as Wikipedia allow users, sometimes including anonymous users, to edit the content.
- Social networking sites like Facebook, Twitter, Instagram, where users interact with other people chatting, writing messages, or posting images or links.
- Media hosting sites such as YouTube allow users to post content.

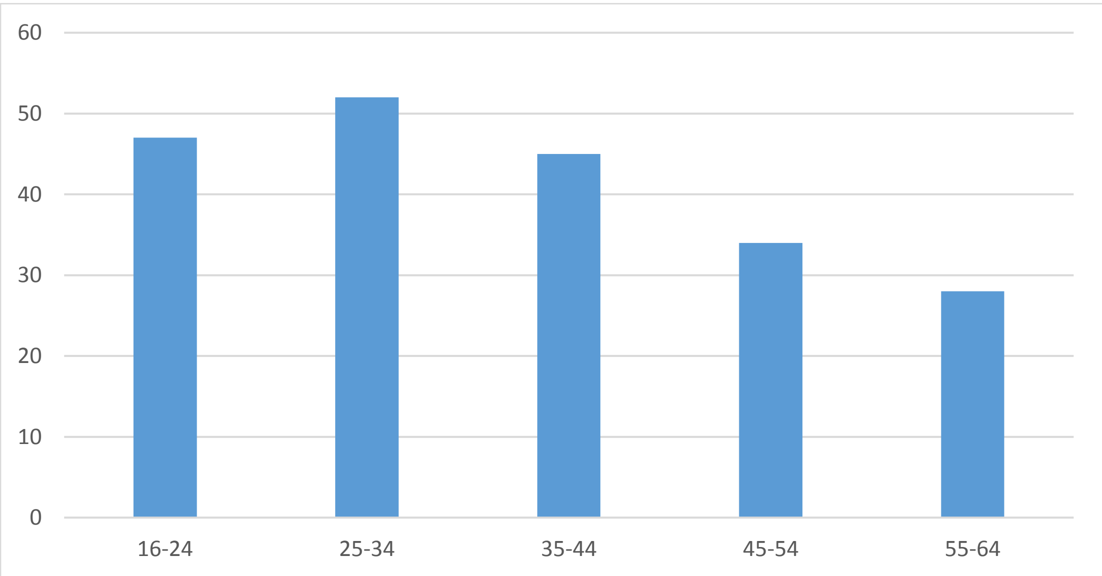
Theories behind the motivation for contributing user generated content range from altruistic, to social and to materialistic. Social incentives allow the user to feel good as an active member of a community and can include relationship between users, such as Facebook's friends, or Twitter's followers. Users also share the experiences that they have while using a particular product/service. This improves the customer experience as they can make informed decisions in buying a product, which makes them smart buyers. Other common social incentives are status, badges or levels within the site, something a user earns when they reach a certain level of participation which may or may not come with additional privileges. Social incentives cost the host site very little and can catalyze vital growth. However, their very nature requires a sizable existing community before it can function. Users reviews analysis can provide new tools to understand user's needs and create new communication channels with them. The significance of user generated content is clearly shown on the following figures.


Do you trust online customer reviews as much as personal recommendation?

| | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|
| Yes, always | - | 8% | 18% | 19% |
| Yes, I believe that the reviews are authentic | 22% | 31% | 27% | 25% |
| Yes, for some types of business , no for others | 34% | 22% | 19% | 20% |
| Yes, if there are multiple costumer review to read | 26% | 19% | 20% | 20% |
| No, I am often skeptical about online reviews | - | - | 12% | 13% |
| No, I don't trust review at all | 17% | 20% | 4% | 3% |

(source:statista.com, USA, 10/2017

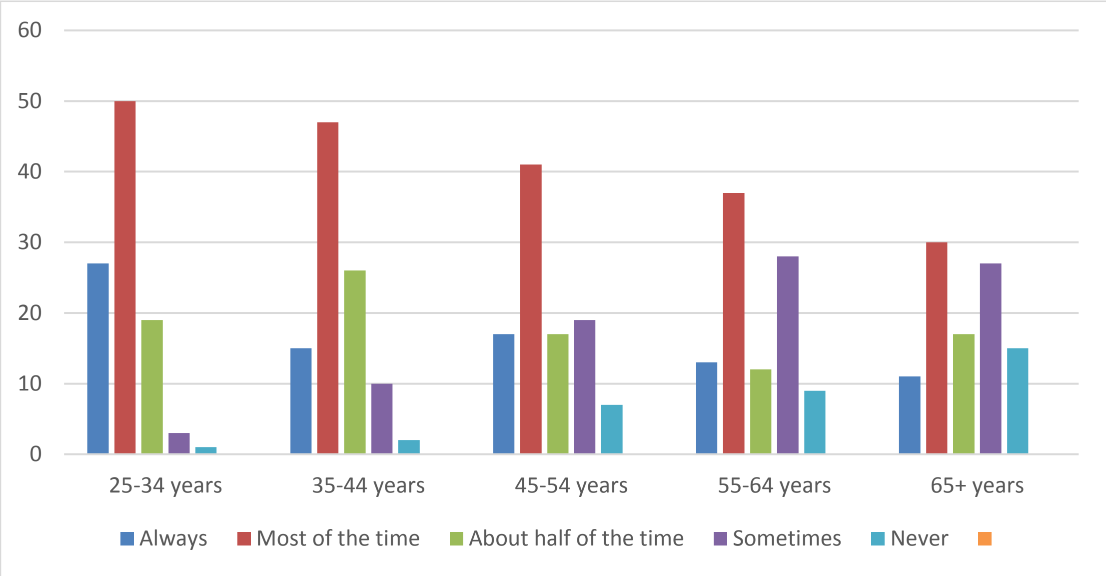
Percentage of global internet users who post reviews online, by age group.




(source:statista.com, worldwide, Q3 2017)

In 2018 55,1% of the global population accessed the internet and over the last two years alone 90 percent of the data in the world was generated. With 86% of the costumers reading reviews from other costumers for businesses, reviews are today a 'power shift' tool, enabling consumers to pull information, rather than having businesses (retailers, service providers) push information to them. In 2017 84% of the costumers trust reviews as personal recommendation at some point, with 45% of them strongly believing in consumer reviews.

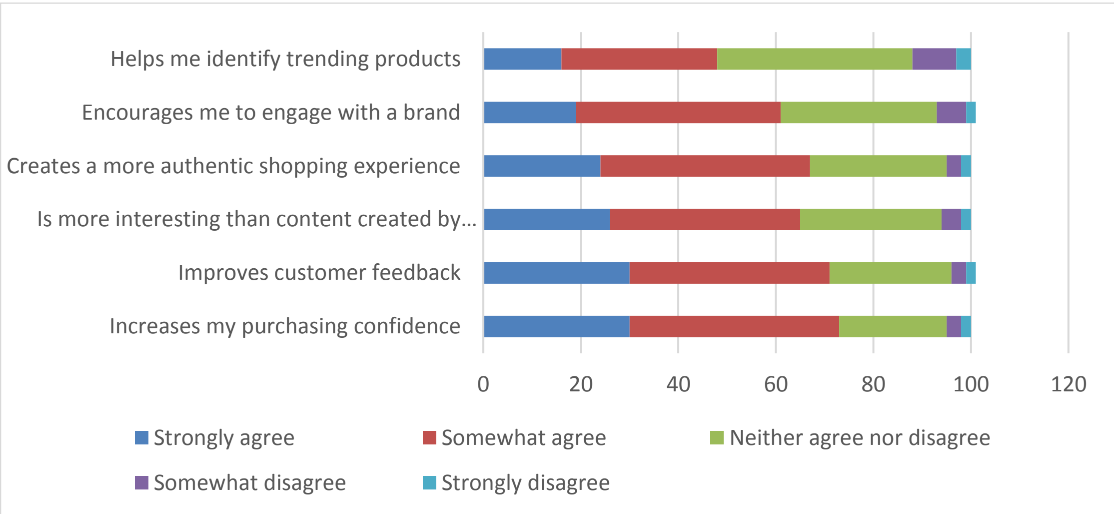
Online review usage frequency prior to new product purchase.




(source:statista.com, USA, 2017)

With 96% of the costumers of the age group 25-34 years using online reviews before purchase a new product more than half of the time and even users of the group 65+ years using for the same reason online reviews 68% more than a half of the time, it is obvious that user generated content plays a very crucial role to the consumer behavior and a very helpful way to understand user's interests.

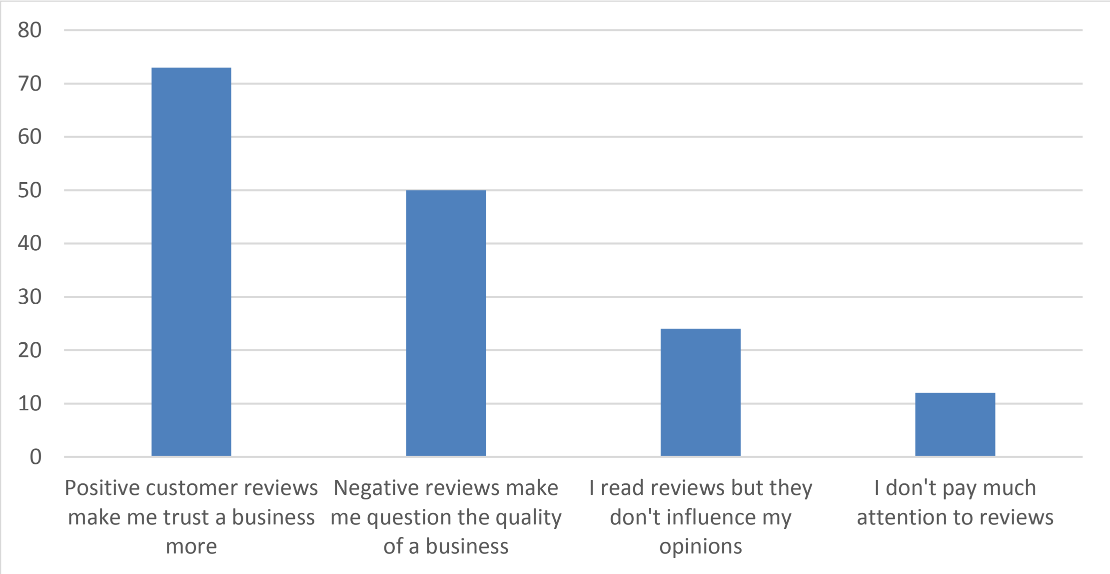
User generated content impact on online shoppers



(source:statista.com, USA, 2017)

In 2021, over 2.14 billion people worldwide are expected to buy goods and services online, up from 1.66 billion global digital buyers in 2016.UGC increases costumers purchasing confidence by 73% and improves customer feedback by 71%. A very interesting fact is that users believe that UGC is more interesting than the content produced by the brand and their getting more engaged with the brand, despite the billions spent on advertisement and market research.

How do online customer reviews affect your opinion of a local business?




With 73% of the costumers saying that positive reviews make them trust a business more and 50% that negative reviews make them question the quality of a business it is undeniable that UGC has transformed the way users judge, decide and finally experience everyday practices at personal and organizational level.

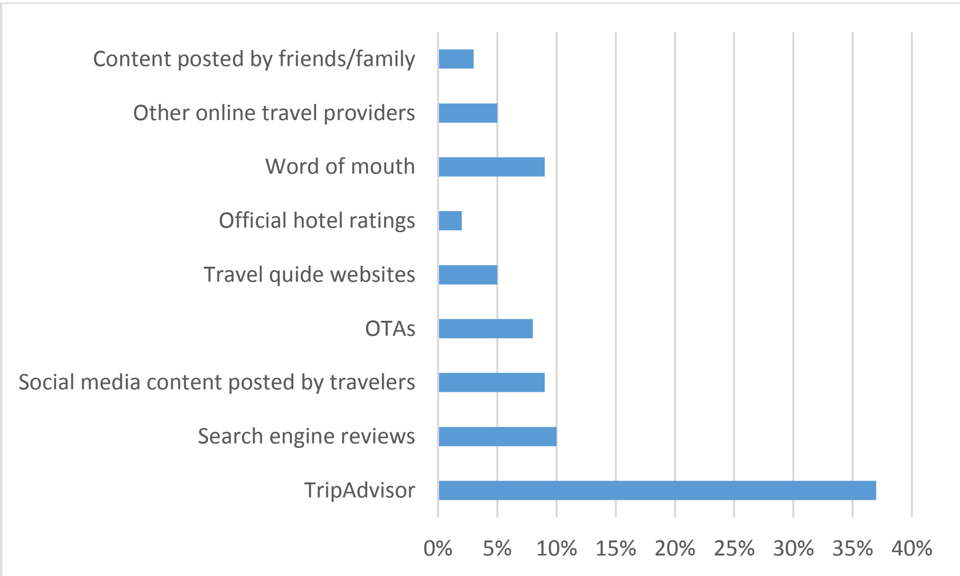
### 4.3 Case study

The aim of this study is to build a user interest model through user generated content analysis and provide successful recommendations to the users. In order to understand consumer behavior through the analysis of UGC and build a useful user interest model, there will be a statistical analysis of costumer reviews with several approaches.

The source of the reviews is TripAdvisor. This source was selected because TripAdvisor stands out most prominently in terms of usage and content among various travel-related sites that support UGC.


With 661 million traveler reviews and 456 million monthly unique visitors, TripAdvisor is one of the most reliable sources. The first question is what drives travelers to start thinking about a trip. Browsing on TripAdvisor inspired 10% of the costumers to visit a destination, while 15% of the costumers are prompted by a personal recommendation. In addition, 40% of the costumers are open to visiting a number of places, when they search for a trip.

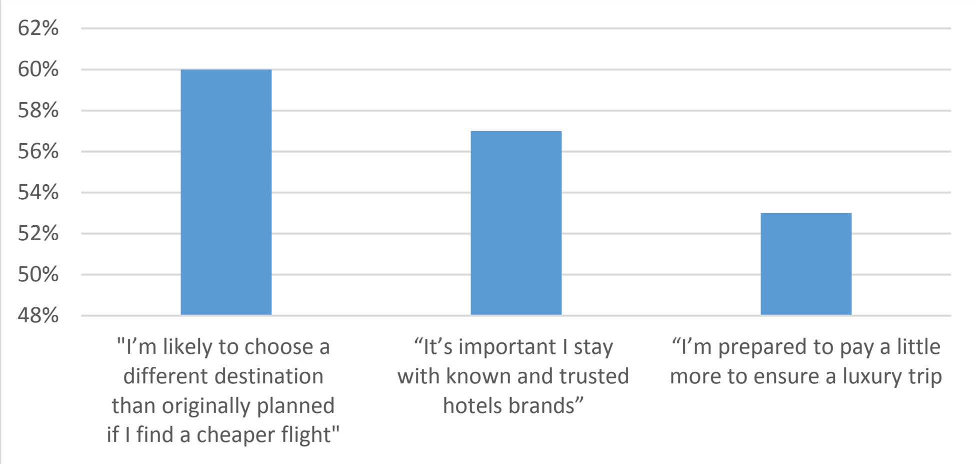
Where did you look for inspiration when considering which destination to visit?




(source: TripBarometer 2017/18, global report)

The second question is which is the path to booking a trip.29% of the costumers arrange transportation to the destination fist, 27% compare carefully all options to find the best option overall and 22% book accommodation first.

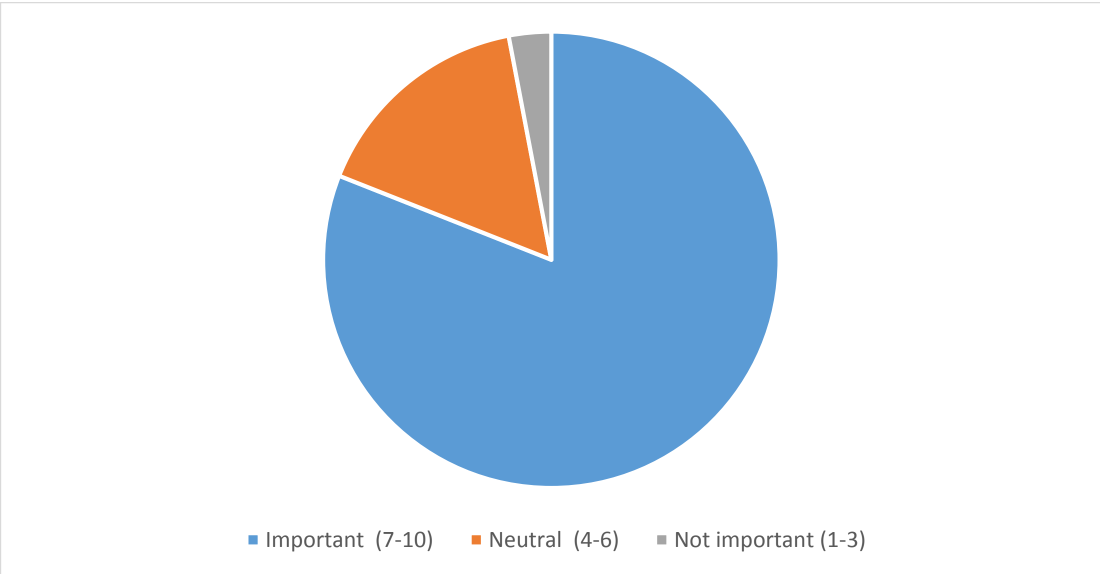
Statements that people have made about their choices when booking a trip.




(source: TripBarometer 2017/18, global report)

From the above we can deduce that TripAdvisor is the largest online travel provider and by far the most influential. In addition, a large proportion of the customer (22%) book accommodation first and it is important for them to stay in trusted hotel brands.

How important are user reviews to you when determining which hotel to stay at?




(source: TripBarometer 2017/18, global report)

The aim of the paper is more specifically to analyze costumer reviews from TripAdvisor, related with accommodation in 5 regions.10276 reviews were extracted from the site and the original format of the data was the following.

| Review's Title | Reviewer's Location | Full Review | Rating | Hotel's Name | Hotel's Location | Hotel's Class |
|---|---|---|---|---|---|---|
| Great location, comfortable.N | Messery, France | Nice. Brilliant location c | 5 of 5 bubbles | The Zillers Boutique Hotel | Athens | 4 Stars |
| Great service and comfort | Cincinnati, Ohio | The upscale hotel Daios | 4 of 5 bubbles | Daios Luxury Living | Thessaloniki | 5 Stars |
| Perfect location for gourmet v | | Nice hotel with friendly : | 4 of 5 bubbles | The Bristol Hotel | Thessaloniki | 5 Stars |
| Best breakfast & service | New York City, New York | I love this hotel, stayed | 5 of 5 bubbles | Archipelagos | Mykonos | 5 Stars |
| Best Hotel in mykonos | Stockholm, Sweden | Good Hospitality & Frie | 5 of 5 bubbles | Kirini - My Mykonos Retreat | Mykonos | 5 Stars |
| Fairly nice hotel, not much ar | Melville, New York | If you want a hotel walk | 3 of 5 bubbles | Apanema Resort | Mykonos | 4 Stars |
| Not worth it!! | California | We stayed at San Anto | 2 of 5 bubbles | San Antonio Summerland Hotel | Mykonos | 4 Stars |
| Spoiled our Wedding Anniversary | | We stayed in a Panora | 3 of 5 bubbles | Aphrodite Beach Hotel | Mykonos | 4 Stars |
| Fantastic experience!!! | | It was unbelievable exp | 5 of 5 bubbles | Tharroe of Mykonos Hotel | Mykonos | 5 Stars |
| Amazing team | London | We've just spent a weel | 5 of 5 bubbles | Tharroe of Mykonos Hotel | Mykonos | 5 Stars |
| Three days wasn't enough | Auburn, Alabama | Wow.......what can we : | 5 of 5 bubbles | Petinos Hotel | Mykonos | 4 Stars |

From the original data it was known the review's title, reviewer's username, reviewer's location, the review, rating of the hotel, hotel's location and hotel's class. The users model will be constructed by analyzing the reviews.

Reviewer's location: more than 35 different countries, mainly in Europe

Rating of the hotel: 5-point scale

Number of users: 10276

Number of hotels: 4153

Hotel's location: Athens, Thessaloniki, Mykonos, Crete, Rome

Hotel's class: 4 and 5 stars

### 4.4 Building users' interests model

The first step of the analysis is keyword extraction from the reviews. Keyword extraction is a process that collects a set of terms, which is an overview of the document. Keyword identifies the core information of the review and this approach can assist to match relevant information from other reviews and then build the model, based on similarity of the user's interests. Keywords can be compounded by one or more words and they can be used to index data to be searched and finally generate tag clouds. The difference with extraction compared to classification is that in classification the result is an associated tag that is usually not present within the text, and therefore has to be predicted or deduced from the text contents. There a lot of different extraction models in order to extract different types of data.

Custom extractors are useful to train a machine learning model to extract pieces of data from a series of texts. The data can be whatever the user define: email addresses, names, products. There are several applications which can help user build his own extraction model. For the purpose of this paper MonkeyLearn is utilized. The user imports the text data directly to the application and specifies the data he will use in order to train the model. With a term frequency analysis tool, like online-utility.org, we can find the most frequent words of the text. Then we can categorize the terms into organized groups and understand the particular interests of each reviewer. For example, the first review:

“Nice. Brilliant location opposite the cathedral. Bed and linen ideal for a good night’s sleep. Good combination of design in neo-classical building. Quiet. The roof terrace is currently very trendy for an early evening drink. Great view. The 8 hours before sunrise cocktail is, incidentally, fun and delicious. Breakfast has a good choice and is good quality. Staff professional and friendly. We will definitely want to revisit.”

From this review we can extract the following words, that describe reviewer’s interests.

Location, bed, sleep, terrace, drink, breakfast, staff

The following table shows words frequencies in the reviews.

| Word | Total number |
|---|---|
| Staff | 6233 |
| Breakfast | 5103 |
| Food | 2877 |
| Pool | 2703 |
| Clean | 4171 |
| Beach | 2067 |
| Restaurant | 3190 |
| Service | 2468 |
| Area | 2156 |
| Location | 3176 |
| Bar | 2418 |
| Walk | 3273 |
| Reception | 1710 |
| Sea | 1514 |
| Bathroom | 1556 |
| Bus | 1861 |
| Located | 1018 |
| Dinner | 917 |
| View | 3005 |
| Balcony | 843 |

The most frequent words are categorized in groups. Each one of the 9 groups describes a different interest of the user. The groups are:

1) Location: location, area, located, walking, walk, metro, car, airport, bus
2) Food: breakfast, dinner, menu, food, restaurant, bar, drinks
3) Service: service, staff, reception
4) Cleanliness: clean, cleanliness, dirty
5) View: view, balcony, window
6) Beach & Pool: beach, pool
7) Amenities: spa, gym
8) Facilities: tv, wifi, Wi-Fi, bathroom, parking, elevator, lift, air condition, kitchen, facilities
9) Bed: bed, sleep, mattress, pillow

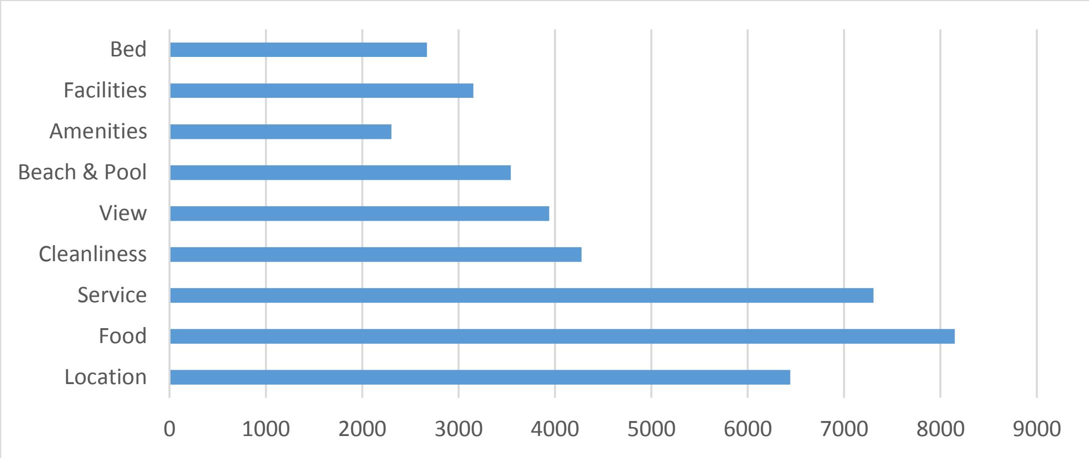
The next step is to search if the keywords are found in each review.

| Reviewer's Username | Full Review | tv | WiFi | wi-fi | bathroom | parking | elevator | lift | air conditi | kitchen | facilities | Facilities |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| themisb | Nice. Brilliant l | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | 1 |
| hlk613 | The upscale ho | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | 1 |
| Somebodyaround | Nice hotel with | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | TRUE | FALSE | TRUE | 1 |
| caronaf1 | I love this hote | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | 0 |
| Zaid A | Good Hospitalit | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | 0 |
| bcmlawer | If you want a h | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | 0 |
| caebayer | We stayed at S | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | 1 |
| Theresa K | We stayed in a | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | 1 |
| Ahmed N | It was unbeliev | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | 0 |

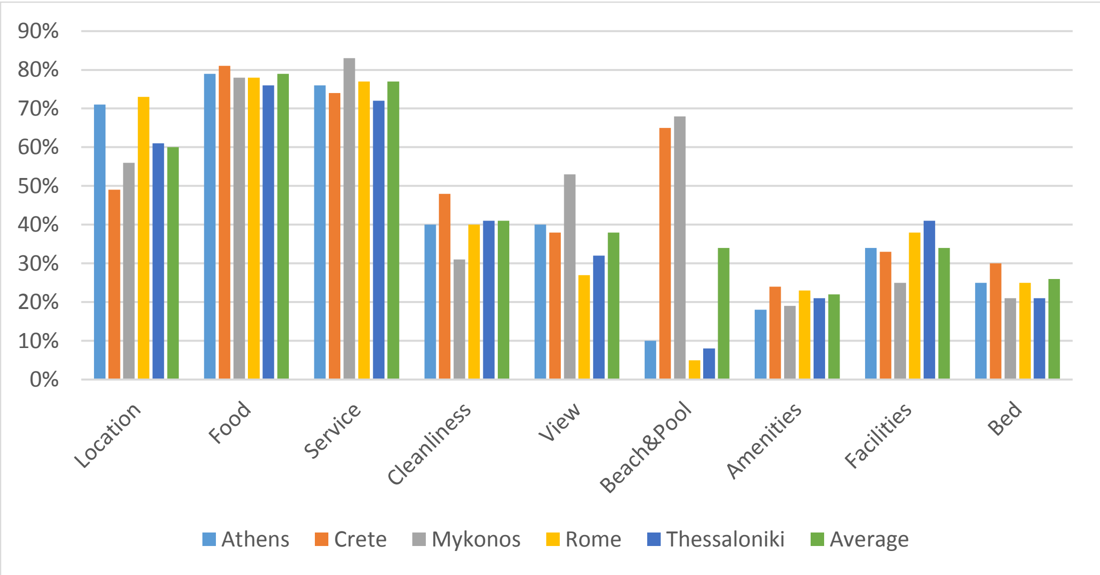
With this approach we know that for example the user $x_1$ has commended the bathroom and the user $x_2$ is interested about the parking, air condition and generally facilities. The specific interests of each user are known and can be grouped together.

| Reviewer's Username | Full Review | Location | Food | Service | Cleanliness | View | Beach&Pool | Amenities | Facilities | Bed |
|---|---|---|---|---|---|---|---|---|---|---|
| themisb | Nice. Brilliar | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| hlk613 | The upscale | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| Somebodyaround | Nice hotel w | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| caronaf1 | I love this ho | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| Zaid A | Good Hospit | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| bcmlawer | If you want a | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| caebayer | We stayed a | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| Theresa K | We stayed in | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| Ahmed N | It was unbel | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| SachaLondon | We've just s | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |

For example, the user $x_y$ whose review was analyzed again before, is interested about location, food, service, view, facilities. The 10276 reviews are categorized in the groups. Below is the number of reviews each group has.

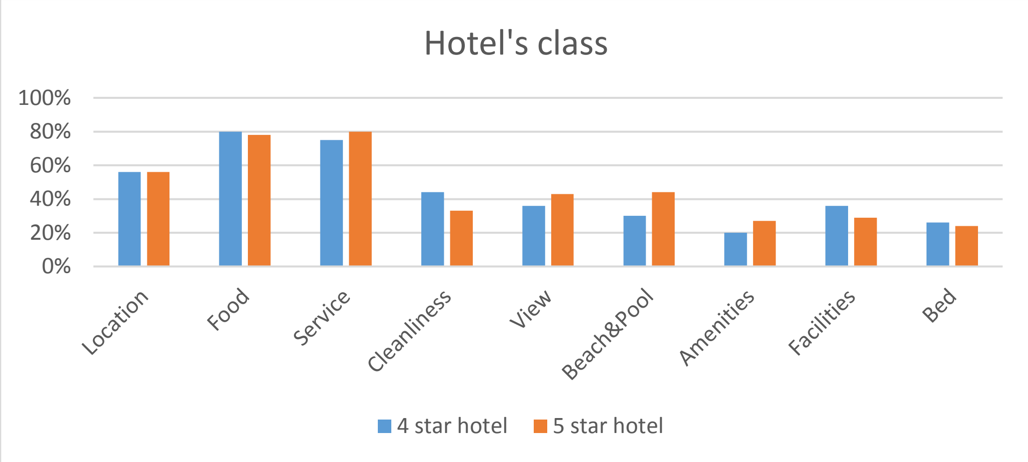



Through the analysis of the groups significant differences have emerged.




The figure shows differences of the users interests based on destination. In the groups food, service, bed, facilities and amenities there are no significant differences. There are great variations within groups location and beach pool, which can be explained by the nature of the trip (beach holidays). As expected, for city-break destinations like Rome and Athens location is more important than Crete and Mykonos which are best described as beach holidays. For the same reason 66% of the reviewers who had visited Crete and Mykonos were interested about Beach& Pool, in contrast with only 6% of the costumers who have visited Rome.

Another way to analyze the reviews is based on the hotel's class. The figure below, which is a comparison between 4-class and 5-class hotel shows that again there is no difference in the groups food, service, bed, amenities. Interesting are the deviations in cleanliness and beach&pool.




### 4.5 Correlation-based similarity

In order to build the model, the similarity of the users must be measured. In this user-based recommendation approach the similarity weight computation will be the guide, which users-neighbors to select and what importance to give them. One popular measure is the Pearson correlation coefficient, a measure of the linear correlation between two variables. Pearson correlation when applied in a sample is commonly represented by $r_{xy}$ .If user 1=x and user 2=y

$$
\begin{aligned}
r_{XY} &= \frac{\text{Degree to which } X \text{ and } Y \text{ vary together}}{\text{Degree to which } X \text{ and } Y \text{ vary separately}} \\
&= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \\
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2},
\end{aligned}
$$

Pearson correlation coefficient has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. Applying the correlation between the reviewer $x_y$ and all the other reviewers, 53 reviewers appear to have exactly the same interests with the examined user(r=1). That means that reviewer's $x_y$ ratings must be considered in order to predict the ratings of the 55 reviewers and make the proper recommendations, as there is a perfect positive linear relationship. In addition, there are 2262 reviewers with correlation coefficient between +0,5 and +0,7, what consists a moderate positive relationship and 358 reviewers with correlation between +0,7 and 1, a strong positive relationship. Reviewer $x_y$ has no linear relationship with 158 reviewers (r=0) and negative relationship with 2581 reviewers(r<0). It is important to highlight that the relation was discovered without taking into consideration destination, the purpose of the trip or the ratings of the hotel. The same technique can be followed for all the reviewers and this way a model on the basis of similarity measurement can be constructed. At this point reviewers have been categorized according to their interests and now ratings should be examined, to understand if the hotels rated by other users with positive linear relationship (r>+0,7) can be successful recommendations.

## 4.6 Rating normalization

When a user rates an item, like a hotel, subjective factors appear and it is not always clear if the rating is positive, negative or neutral. User might be reluctant to give high rating to a hotel he likes or a low in a hotel he dislikes.

Mean-centering

The aim of mean-centering is to assist understand if a rating is positive, negative or neutral by comparing it with the mean rating. If $r_{ui}$ is the rating given from the user to item i , the mean- centered is

$$h(r_{ui}) = r_{ui} - \overline{r}_u.$$

where $\bar{r}_u$ is the mean rating given by the user u.

The 412 reviewers who have strong and perfect linear relationship (r>+0,7) with the target user will form a group and provide recommendations to the user. Their mean rating is 3,93 and only the hotels which gather ratings over the mean will be considered as successful recommendations. If the group is limited only to the 55 reviewers with perfect relationship with the user, then the mean rating will be 3,57.

| Review | Full Revi | Rating | Hotel's | Locatio | Food | Service | Cleanli | View | Beach8 | Amenit | Faciliti | Bed | correl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| themisb | Nice. Brillia | 5 of 5 bubb | The Zillers | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | TRUE |
| holidayfar | My husband | 3 of 5 bubb | SENTIDO F | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0,790569 |
| etrev28 | I am finding | 5 of 5 bubb | Mykonos I | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0,755929 |
| Arjay R | You will be | 5 of 5 bubb | Kouros Ho | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0,790569 |
| Ellen F | This was on | 5 of 5 bubb | Hermes M | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| maggie19 | Spent 3 nigh | 4 of 5 bubb | Areos Hot | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0,790569 |
| | Stayed here | 3 of 5 bubb | Manoulas | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0,755929 |
| hannahpra | Currently st | 5 of 5 bubb | Berg Luxu | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0,755929 |

In case the group of reviewers with strong relationship is chosen, more hotels with higher ratings (over 3,93) will be taken into consideration. The problem is that there are also 9 hotels with mean ratings between 3,57 and 3,93 which could be possible proposals and are excluded from the list of recommendations.

Now we can recommend to the target user hotels which match his interests based on his profile in 5 destinations, with rating over 3,93. For example, in Thessaloniki there are 10 hotels that fit to his interests and needs, with mean rating over 3,93. Of course if we had more reviews from the target user about hotel his profile would be more completed and precise and the recommendations more accurate and useful.

Another aspect is what would change if the target user went for a trip in an island, like Mykonos. As we show before 68% of the reviewers who visited Mykonos are interested in Beach&Pool. The target user, who has commented his trip in Athens was not interested in this group, as expected, because only 10% of Athens' visitors have commended about Beach&Pool. If we want to recommend a hotel in Mykonos we may need to add this group in his interests. This addition forms a new group of users, with different interests and relationships between them. Now there are 214 reviewers with strong relationship with the target user (r>+0,7) and 38 with perfect relationship (r=1). The mean rating of the new users' group is 4,05 and there are 49 hotels in Mykonos that satisfy all the conditions and thus can be recommended.

### 4.7 Sentiment analysis

Oxford Dictionary definition

"The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral."

Sentiment Analysis is a field within Natural Language Processing (NLP) which builds systems that try to identify and extract opinions within text. Humans communicate with words, a form of unstructured data. Unfortunately, computers can not work with unstructured data, as there are no standardized techniques to process them. Humans can understand what an online review of a product really means, the emotions of the writer and his attitude towards the product. Natural Language Processing (NLP), a sub-field of Artificial Intelligence, is focused on enabling computers to understand and process human languages, to get computers closer to a human-level understanding of language.

Since information over the WWW is continuously increasing, billions of texts expressing opinions are available online. These opinions are of great value, if there are transformed in structured data. Public opinions can be extremely useful in commercial applications like marketing analysis, product reviews and customer service.

Any text information can be categorized in two types, facts and opinions. A fact is a statement that can be proven true or false. An opinion is an expression of a person's feelings that cannot be proven. Opinions can be based on facts or emotions and are usually subjective. If a customer wants to buy a product, it would be very useful to read reviews from other customers, but occasionally that could be misleading. Therefore, it is important to be aware of the author's purpose, feelings and choice of language.

Sentiment analysis has two sub problems to solve:

- Subjectivity classification: is the text subjective or objective
- Polarity classification: is the opinion negative, positive or neutral

In this case study we will emphasize only in polarity classification. This type of classification would be very helpful, if ratings were not available. In a review the user talks about the hotel, its features and his experience during the trip. For example, below is a randomly selected review.

"Spent a great time in this lovely hotel strategically located nearby all transport connections. Breakfast was excellent and the view from top floor awesome. Room was quiet and perfectly clean. Most of all we appreciated the very friendly staff, always at your disposal for any questions and requirement. We highly recommend this hotel."

A positive opinion is expressed about the location of the hotel, breakfast, view, cleanliness and the service.

Opinions are distinguished in explicit and implicit. An example of explicit positive review is the following:" Relaxing hotel with nice rooms nice view, lovely pool, near beach kalafatis and agia anna restaurant beach bar cafe. Best hotel ever! "

An implicit opinion on a subject is an opinion implied in an objective sentence. The following review is an implicit negative review:" This is the first time of my life when being on vacation (not mentioning staying in a five start hotel) when I was counting days to go home.".

### Fine grained sentiment analysis

The level of polarity of a review is not limited only to positive, negative and neutral. If precision is required, we should consider a different scale. A five level Likert scale is usually used and the categories are formed accordingly. Feelings are associated with polarity and reviews are categorized on the basis of the reviewer's' feelings. For example, anger is categorized negatively and happiness positively.

Emotion detection is not always an easy process. Emotion detection systems use as source lexicons, like SentiWordNet and SenticNet. A usual problem is that words can have multiple meanings. For example, the word kill usually suggests anger ("the service is killing me"), but it can also be used to describe happiness.

It is estimated that 80% of the available data are in unstructured form and it is time-consuming to analyze them. Through sentiment analysis any user can efficiently analyze text data, like emails and reviews, and find critical information real-time. Moreover, as human's opinion is usually subjective, the user can enhance data consistency by using a centralized sentiment analysis system.

### Rule based approach

Rule based approaches define specific rules that identify the polarity of the opinion.

Rules can be formed with a variety of inputs based on NLP techniques, like tokenization, stemming and POS tagging. Another source to form the rules are lexicons. A very simple rule to find the polarity of a review can be the following.

1) Make one list with positive words, like nice, amazing, beautiful and a second list with negative words like, awful, bad, disaster.

2) Count the number of the words in the text with negative polarity and the words with positive polarity.

3) If the positive words are 30% more frequent than the negative, define the review as positive. If the negative words are 30% more frequent than the positive, define the review as negative. In any other case, define the review as neutral.

For the purposes of this study, two applications will be used to analyze the polarity of the reviews, MonkeyLearn and LEXALYTICS. As we can not analyze all reviews due to the cost, 1000 reviews will be taken as a random sample. Monkeylearn divides the reviews in positive and negative, while LEXALYTICS categorize them as positive, negative and neutral. In addition, both applications measure the confidence of the answer. Below is a table with the results, when the data were entered into the two applications.

| Full Review | Rating | Classification | Confidence | Classificatio | Confidence |
|---|---|---|---|---|---|
| Nice. Brilliant location opposite the | 5 of 5 bubbles | Positive | 0.999 | positive | 0,535 |
| The upscale hotel Daios has much | 4 of 5 bubbles | Positive | 0.901 | positive | 0.277 |
| Nice hotel with friendly staff and free | 4 of 5 bubbles | Positive | 0.992 | neutral | 0,204 |
| I love this hotel, stayed here last ye | 5 of 5 bubbles | Positive | 0.999 | positive | 0,556 |
| Good Hospitality & Friendly Recepc | 5 of 5 bubbles | Positive | 0.978 | positive | 0,603 |
| If you want a hotel walking distance | 3 of 5 bubbles | Positive | 0.873 | positive | 0,465 |
| We stayed at San Antonio Summer | 2 of 5 bubbles | Negative | 0.922 | neutral | 0,033 |
| We stayed in a Panoramic Double I | 3 of 5 bubbles | Negative | 0.999 | neutral | 0,051 |
| It was unbelievable experience!! Ver | 5 of 5 bubbles | Positive | 0.976 | positive | 0,985 |
| We've just spent a week here and c | 5 of 5 bubbles | Positive | 0.997 | positive | 0,332 |
| Wow.......what can we say to give y | 5 of 5 bubbles | Positive | 0.997 | positive | 0,617 |
| A nice hotel with lovely interior and | 4 of 5 bubbles | Negative | 0.639 | positive | 0,315 |
| Mykonos should be so proud for ha | 5 of 5 bubbles | Positive | 1 | positive | 0,548 |

Comparing the results from the first application (MonkeyLearn) with the users' ratings, several conclusions are drawn. The hypothesis is that the hotels with ratings 4 and 5 bubbles should be recommended and the hotels with 1,2 and 3 bubbles should not be recommended, as the average rating is 3, 95. The first system is very effective (94%) at recommending hotels, which based on the hypothesis should be recommended but only 70% effective at excluding hotels, which should not be recommended. The second system (LEXALYTICS) is more sophisticated but the results are very comparable and for that reason it does not add any value to this analysis. In order to demonstrate how a semantic analysis system operates, a customized rule based approach is applied.

The idea behind this approach is to find the most frequent words of the reviews, that describe the feelings of the reviewers towards the hotels. With the free term frequency analysis tool online-utility.org the most frequent words in the reviews are counted. Then a considerably small lexicon is built, with 75 positive words and 35 negative words. The table below shows the most frequent words of every category.

| Positive words | Negative words |
|---|---|
| Good | Small |
| Great | Problem |
| Nice | Noise |
| Lovely | Busy |
| Friendly | Wait |
| Well | Poor |
| Helpful | Noise |
| Amazing | Disappointed |
| Excellent | Unfortunately |
| Best | Fault |

Subsequently the appearance of every word in the reviews is counted and added.

This way we have the sum of the positive words and the sum of the negative words from every review.

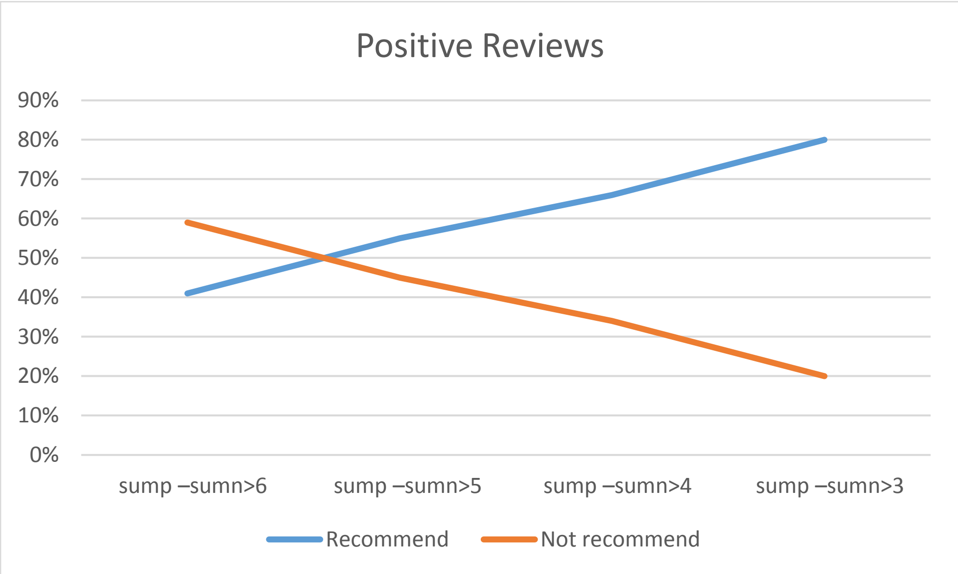
| Full Review | good | great | nice | clean | lovely | friendl | well | helpful | amazin |
|---|---|---|---|---|---|---|---|---|---|
| Nice. Brilliant location opp | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| The upscale hotel Daios ha | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| Nice hotel with friendly sta | FALSE | FALSE | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| I love this hotel, stayed he | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE |
| Good Hospitality & Friendl | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | TRUE | TRUE |
| If you want a hotel walking | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| We stayed at San Antonio S | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| We stayed in a Panoramic I | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| It was unbelievable experi | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE | FALSE |
| We've just spent a week he | FALSE | TRUE | TRUE | FALSE | TRUE | FALSE | FALSE | TRUE | TRUE |
| Wow.......what can we say | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE |

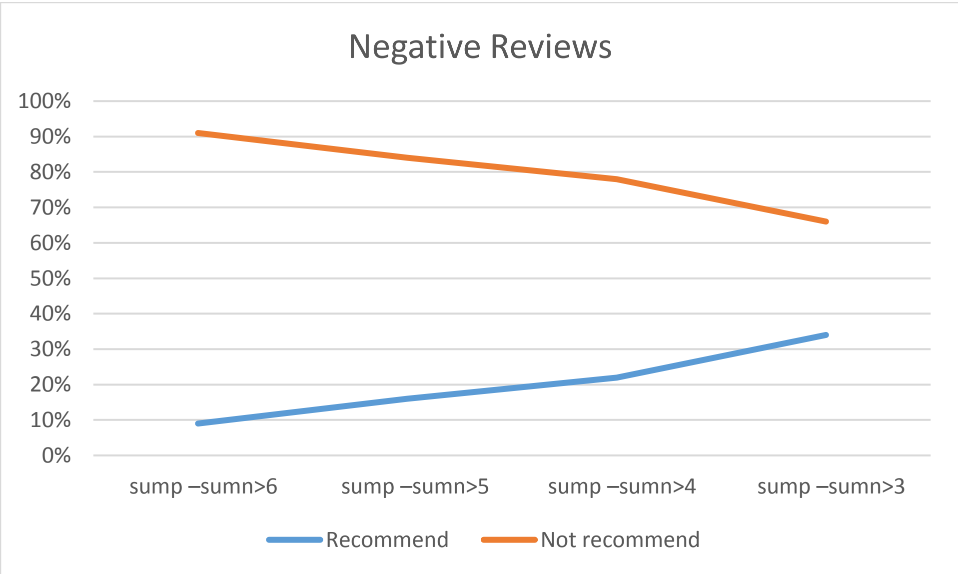
The reviewers are favorably disposed towards the hotels, as 81% of them has given positive review (4 or 5 bubbles). A training set of 1000 reviews will be used, to determine which rule should be applied. From 1000 reviews, based on the lexicon, 11 do not have any positive word and 431 do not have any negative word. From the 431 reviews with no negative words, only 41 reviewers have rated the hotels with 1,2 or 3 bubbles, i.e. negatively. The idea is to subtract the sum of negative words from the sum of positive words and if the result is above a threshold, to characterize the review positively or negatively.

The first threshold is $sum_p - sum_n > 6$. The system polarizes positively 332 reviews (41%) from 802 reviews, which should be recommended and 180 reviews negatively from 198 reviews, which should not be recommended. The accuracy not to recommend reviews, which should not be recommended, when the results are compared with the ratings is very high (91%) but with this threshold 59% of the reviews which express positive attitude towards the hotels are not shown.

Four thresholds are tested, in order the most precise rule to be formed.

- $sum_p - sum_n > 6$
- $sum_p - sum_n > 5$
- $sum_p - sum_n > 4$
- $sum_p - sum_n > 3$



Examining the charts, the first observation is that in both cases the rates are conversely. When the percentage of recommend increases, the percentage of not recommend decreases and conversely, regardless if it is a negative or a positive review. For positive reviews the best threshold is $sum_p - sum_n > 3$, as it returns 80% of the reviews. For negative reviews the best threshold is $sum_p - sum_n > 6$, because it returns only 9% of the reviews. As expected, there is a trade off between precision and recall. When the threshold is reduced, recall increases and precision decreases.

### 4.8 Evaluating the models

In order to evaluate the recommendation systems and measure the success of predictions, precision and recall will be used. Recommendation is viewed as information retrieval task.

| | Recommended | Rejected |
|---|---|---|
| MonkeyLearn | 891 | 109 |
| $sum_p - sum_n > 3$ | 708 | 292 |
| $sum_p - sum_n > 4$ | 572 | 428 |
| $sum_p - sum_n > 5$ | 472 | 528 |
| $sum_p - sum_n > 6$ | 347 | 653 |

Precision: the fraction of relevant items retrieved out of all items retrieved. It is the proportion of recommended reviews that are actually positive.

Recall: the fraction of relevant items retrieves out of all relevant items. It is the proportion of all positive reviews recommended.

$$\text{Precision} = \frac{tp}{tp + fp}$$
$$\text{Recall} = \frac{tp}{tp + fn}$$

The F-Score or F-measure is a measure of a statistic test's accuracy. It considers both precision and recall measures of the test to compute the score. F-Score is a weighted average of the precision and recall, where the best F1 score has its value at 1 and worst score at the value 0.

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

| | Precision | Recall | F1-score |
|---|---|---|---|
| MonkeyLearn | 0,84 | 0,94 | 0,88 |
| $sum_p$-$sum_n$ >3 | 0,90 | 0,80 | 0,84 |
| $sum_p$ –$sum_n$ >4 | 0,92 | 0,66 | 0,77 |
| $sum_p$ –$sum_n$ >5 | 0,93 | 0,55 | 0,69 |
| $sum_p$ –$sum_n$ >6 | 0,94 | 0,41 | 0,57 |

MonkeyLearn has the highest F-Score (0,88), followed by the model with the rule $sum_p$ –$sum_n$>3 (0,84). The aim of this sentiment analysis system is to provide successful recommendations, with high accuracy. The costumers should trust the recommendations, as if they were personal recommendations. As the credibility of the system is the primary target, high precision is the main objective. The cost of False Positive is high, because a costumer will probably stop using the system, if he has a bad experience in a recommended hotel due to a false positive result.

The final sentiment analysis model is a synthesis of MonkeyLearn with the model $sum_p$-$sum_n$ >3. The combination of the two models can give to the costumer the choice to select between more results or higher accuracy.

The customized rule based sentiment analysis system, which was built for the purposes of this study, is more efficient and accurate than expected. It captures the evaluative factors, especially the positive and classifies the reviews effectively. The corpus consists of 10276 TripAdvisor reviews (1.393.058 words) for hotels in 5 different locations. Whether the review is positive or negative is determined through the rating provided by the reviewer. The 110 most frequent words, that express the feelings of the authors towards the hotels are divided in two categories, positive and negative. Five rules are tested to find out which rule provides the best results. The model can be upgraded with the use of weight for every word. For example, the words good and excellent describe both a positive experience, but excellent is more positive. When the words have the same value, the positive weight of the word excellent is underestimated. With the use of weight in every word, the F-Score of the model will probably increase. The model can be enhanced with the addition of words in the lexicon, as a typical lexicon contains thousands of words. In this customized lexicon adjectives are mainly used to determine the polarity of the reviews. The reviewers' interests are categorized in 9 groups and the polarity of the whole review is defined. The main problem is that we do not know the sentiment expressed towards a particular group. For example, a customer may have written a positive review for a hotel, but commented negatively the location of the hotel. With this model we cannot classify the negative comment towards the location. A possible solution to this problem is to form another rule. A search rule can be created that measure the nearness of keywords (location, food, restaurant) to known positive and negative adjectives. The rule will look like:

(location) near (excellent, good, great)

(food) near (bad, horrible, awful)

Then the system will count how many times the keyword location appears near every adjective. With this rule we can understand the authors' attitude towards every group and create more successful recommendations.

### 4.9 Content based recommendations

Content based RS use the content of the items in the database to predict its relevancy with the user profile. User profile reflects users long term interests and it is exploited by the RS, to generate recommendations. Every item in the database has some attributes, that describe it. For example, in a book RS the attributes used are author, publisher, genre, year. The system uses the most similar items to a user's already-rated items to generate recommendations. Similarity is measured with Pearson correlation. The features of the item already rated are compared with the features of other available items and the most similar items are proposed. The aim of this approach is to recommend to the target user hotels with similar features with the one he has already stay and rate positively. A set of 200 hotels will be used as an example how item based RS are applied. The features of the hotels are found through TripAdvisor and Booking. The compared features are: Breakfast included, Location, kitchen facilities, Air-conditioning, Airport shuttle, Parking, Front desk 24/7, Restaurant, Double Bed. The hotel takes 1 as a value in the category, if the feature is available and 0 if it is not. The selected features are indicative and can be changed in accordance with the scope of the analysis. The target user has rated a 4-star hotel in Athens. So the available hotels will be filtered accordingly. Based on the features of hotel x1 the most similar 4-star hotels in Athens are the following.

| Hotel's Name | Hotel's Location | Hotel's Class | Correlation |
|---|---|---|---|
| X1 | Athens | 4 Stars | |
| X51 | Athens | 4 Stars | 0,661437828 |
| X134 | Athens | 4 Stars | 0,661437828 |
| X135 | Athens | 4 Stars | 0,661437828 |
| X140 | Athens | 4 Stars | 1 |
| X195 | Athens | 4 Stars | 1 |
| X199 | Athens | 4 Stars | 0,661437828 |

Based on the similarity 7 hotels can be recommended to the target user, with positive correlation over 0,6. A collaborative filtering approach can be added to the system for more successful and novel recommendations. Leveraging the results of the two approaches, the new system takes into account both users preferences and items features. Applying a user-item approach the most successful recommendation for the target user is hotel 195, because the correlation of the features of the hotels is 1 and the correlation of the users' profiles is 0,86.

### 4.10 Modelling users interests

User generated content, like the costumer reviews from TripAdvisor, can be a very useful source to understand and classify users' interests. Hotels ratings is the outcome of users' experience during their stay. The aim of this chapter is to identify if there is a casual combination between users' interests and provided ratings, by utilizing the Fuzzy set Qualitative Comparative Analysis (FsQCA). Qualitative Comparative Analysis (QCA) is a method that bridges quantitative and qualitative analysis. QCA examines the similarities and differences between a set of cases to identify conditions that lead to an outcome. Set is a group of items. For example, the set of cities in Greece with more than 10.000 citizens. Sets can be subsets of larger sets. FsQCA examines what combination of casual sets is a subset of the outcome. As we are not interested for conditions that are simple presence/absence but for partial membership is the sets, fuzzy sets theory is the most appropriate base for this analysis. FsQCa differs from regression, as regression examines the effect of a single variable ceteris paribus, while FsQCa examines what conditions lead to particular outcome. Fuzzy sets analysis uses the truth table. The truth table is all the possible combination of sets, with one row for each combination. The truth table is used to find which combinations lead to the outcome. If there are k sets, the table will have $2^k$ rows. In this case study the sets are tourism service terms and the outcome is the rating of the hotel by the costumers. Location, service, food, cleanliness and view are the most popular terms and they will be used to identify which combinations best reflect costumer's ratings. The 10276 customers will be divided in 5 groups, based on their ratings. The steps of methodology are shown below:

1) Select the reviews published from every group of users
2) Select the terms which will be the casual combinations
3) Calculate Term Frequency
4) Produce the truth table with all the possible casual combinations
5) Calculate membership degrees for each combination
6) Calculate consistency and coverage with the formulas,

$$Consistency(X \prec Y) = \frac{\sum \min(X,Y)}{\sum X}$$

$$Coverage = \frac{\sum \min(X,Y)}{\sum Y}$$

   Where (X) is the membership deggree of each combination and (Y) the membership degree of the outcome.

7) Select the best combinations with consistency over 0,75

| | | Term Frequency | | | | |
|---|---|---|---|---|---|---|
| Rating | Group of Users | Location | Food | Service | Cleanliness | View |
| 1,00 | 1 | 0,60 | 0,79 | 0,77 | 0,39 | 0,39 |
| 0,75 | 2 | 0,67 | 0,82 | 0,69 | 0,46 | 0,39 |
| 0,5 | 3 | 0,67 | 0,82 | 0,61 | 0,43 | 0,37 |
| 0,25 | 4 | 0,60 | 0,75 | 0,54 | 0,39 | 0,36 |
| 0 | 5 | 0,46 | 0,55 | 0,51 | 0,35 | 0,29 |

The truth table is produced with $2^5$ =32 combinations. Below is a part of the truth table, with the fist 20 combinations.

| Causal Permutation | Location | Food | Service | Cleanliness | View |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 | 1 | 0 |
| 4 | 0 | 0 | 0 | 1 | 1 |
| 5 | 0 | 0 | 1 | 0 | 0 |
| 6 | 0 | 0 | 1 | 0 | 1 |
| 7 | 0 | 0 | 1 | 1 | 0 |
| 8 | 0 | 0 | 1 | 1 | 1 |
| 9 | 0 | 1 | 0 | 0 | 0 |
| 10 | 0 | 1 | 0 | 0 | 1 |
| 11 | 0 | 1 | 0 | 1 | 0 |
| 12 | 0 | 1 | 0 | 1 | 1 |
| 13 | 0 | 1 | 1 | 0 | 0 |
| 14 | 0 | 1 | 1 | 0 | 1 |
| 15 | 0 | 1 | 1 | 1 | 0 |
| 16 | 0 | 1 | 1 | 1 | 1 |
| 17 | 1 | 0 | 0 | 0 | 0 |
| 18 | 1 | 0 | 0 | 0 | 1 |
| 19 | 1 | 0 | 0 | 1 | 0 |
| 20 | 1 | 0 | 0 | 1 | 1 |

The cells in the table take value 1 if true and value 0 if false. The permutation 5 is read (Location=false, Food=false, Service=true, Cleanliness=false, View=false). Then the membership degrees for all the possible combinations for every group of costumers is calculated.

The fuzzy union, is defined as $\mu_{(A\cup B)} = \max(\mu_A, \mu_B)$

The fuzzy intersection is defined as $\mu_{(A\cap B)} = \min(\mu_A, \mu_B)$

The fuzzy complement is calculated as $\mu_{\neg A} 1-\mu_A$

The membership degree of combination no5 for costumer 1 is: $\mu_{C5}$ = $\mu$ (*Location=false* $\cap$ *Food=false* $\cap$ *Service=true* $\cap$ *Cleanliness=false* $\cap$ *View=false)* = $\mu$ (not (*Location), not (Food), Location ,not ( Cleanliness), not (Restaurant)).*

The $\mu$(location=false) = $\mu$((1- $\mu$(Location)) = (1-0,6)=0,4. Similar calculations are performed for all terms thus, $\mu_{C3}$=min(0,4; 0,21; 0,77; 0,61)=0,21. After all membership degrees are calculated the consistency and coverage degrees are computed.

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Sum of Group Combination | Consistency | Coverage |
|---|---|---|---|---|---|---|---|---|
| 1 | 0,21 | 0,18 | 0,18 | 0,25 | 0,45 | 1,27 | 0,645669291 | 0,328 |
| 2 | 0,21 | 0,18 | 0,18 | 0,25 | 0,29 | 1,11 | 0,738738739 | 0,328 |
| 3 | 0,21 | 0,18 | 0,18 | 0,25 | 0,35 | 1,17 | 0,700854701 | 0,328 |
| 4 | 0,21 | 0,18 | 0,18 | 0,25 | 0,29 | 1,11 | 0,738738739 | 0,328 |
| **5** | **0,21** | **0,18** | **0,18** | **0,25** | **0,45** | **1,27** | **0,645669291** | **0,328** |
| 6 | 0,21 | 0,18 | 0,18 | 0,25 | 0,29 | 1,11 | 0,738738739 | 0,328 |
| 7 | 0,21 | 0,18 | 0,18 | 0,25 | 0,35 | 1,17 | 0,700854701 | 0,328 |
| 8 | 0,21 | 0,18 | 0,18 | 0,25 | 0,29 | 1,11 | 0,738738739 | 0,328 |
| 9 | 0,23 | 0,31 | 0,33 | 0,4 | 0,49 | 1,76 | 0,636363636 | 0,448 |
| 10 | 0,23 | 0,31 | 0,33 | 0,36 | 0,29 | 1,52 | 0,736842105 | 0,448 |
| 11 | 0,23 | 0,31 | 0,33 | 0,39 | 0,35 | 1,61 | 0,695652174 | 0,448 |
| 12 | 0,23 | 0,31 | 0,33 | 0,36 | 0,29 | 1,52 | 0,736842105 | 0,448 |
| 13 | 0,4 | 0,33 | 0,33 | 0,4 | 0,51 | 1,97 | 0,664974619 | 0,524 |
| 14 | 0,39 | 0,33 | 0,33 | 0,36 | 0,29 | 1,7 | 0,764705882 | 0,52 |
| 15 | 0,39 | 0,33 | 0,33 | 0,39 | 0,35 | 1,79 | 0,726256983 | 0,52 |
| 16 | 0,39 | 0,33 | 0,33 | 0,36 | 0,29 | 1,7 | 0,764705882 | 0,52 |
| 17 | 0,21 | 0,18 | 0,18 | 0,25 | 0,45 | 1,27 | 0,645669291 | 0,328 |
| 18 | 0,21 | 0,18 | 0,18 | 0,25 | 0,29 | 1,11 | 0,738738739 | 0,328 |
| 19 | 0,21 | 0,18 | 0,18 | 0,25 | 0,35 | 1,17 | 0,700854701 | 0,328 |
| 20 | 0,21 | 0,18 | 0,18 | 0,25 | 0,29 | 1,11 | 0,738738739 | 0,328 |
| 21 | 0,21 | 0,18 | 0,18 | 0,25 | 0,45 | 1,27 | 0,645669291 | 0,328 |
| 22 | 0,21 | 0,18 | 0,18 | 0,25 | 0,29 | 1,11 | 0,738738739 | 0,328 |
| 23 | 0,21 | 0,18 | 0,18 | 0,25 | 0,35 | 1,17 | 0,700854701 | 0,328 |
| 24 | 0,21 | 0,18 | 0,18 | 0,25 | 0,29 | 1,11 | 0,738738739 | 0,328 |
| 25 | 0,23 | 0,31 | 0,39 | 0,46 | 0,46 | 1,85 | 0,637837838 | 0,472 |
| 26 | 0,23 | 0,31 | 0,37 | 0,36 | 0,29 | 1,56 | 0,743589744 | 0,464 |
| 27 | 0,23 | 0,31 | 0,39 | 0,39 | 0,35 | 1,67 | 0,706586826 | 0,472 |
| 28 | 0,23 | 0,31 | 0,37 | 0,36 | 0,29 | 1,56 | 0,743589744 | 0,464 |
| 29 | 0,6 | 0,54 | 0,57 | 0,54 | 0,46 | 2,71 | 0,697416974 | 0,756 |
| 30 | 0,39 | 0,39 | 0,37 | 0,36 | 0,29 | 1,8 | 0,777777778 | 0,56 |
| 31 | 0,39 | 0,46 | 0,43 | 0,39 | 0,35 | 2,02 | 0,757425743 | 0,612 |
| 32 | 0,39 | 0,39 | 0,37 | 0,36 | 0,29 | 1,8 | 0,777777778 | 0,56 |

$\sum \min(X,Y) =$ min{min(0,21;1)+min(0,18;0,75)+min(0,18;0,50)+min(0,25;0,25) +min (0,45;0) = min (0,21+0,18+0,18+0,25+0) =0,82

$\sum X = (0{,}21+0{,}18+0{,}18+0{,}25+0{,}45) = 1{,}27$

The consistency for combination no5 = $\frac{0{,}82}{1{,}27}$=0,64

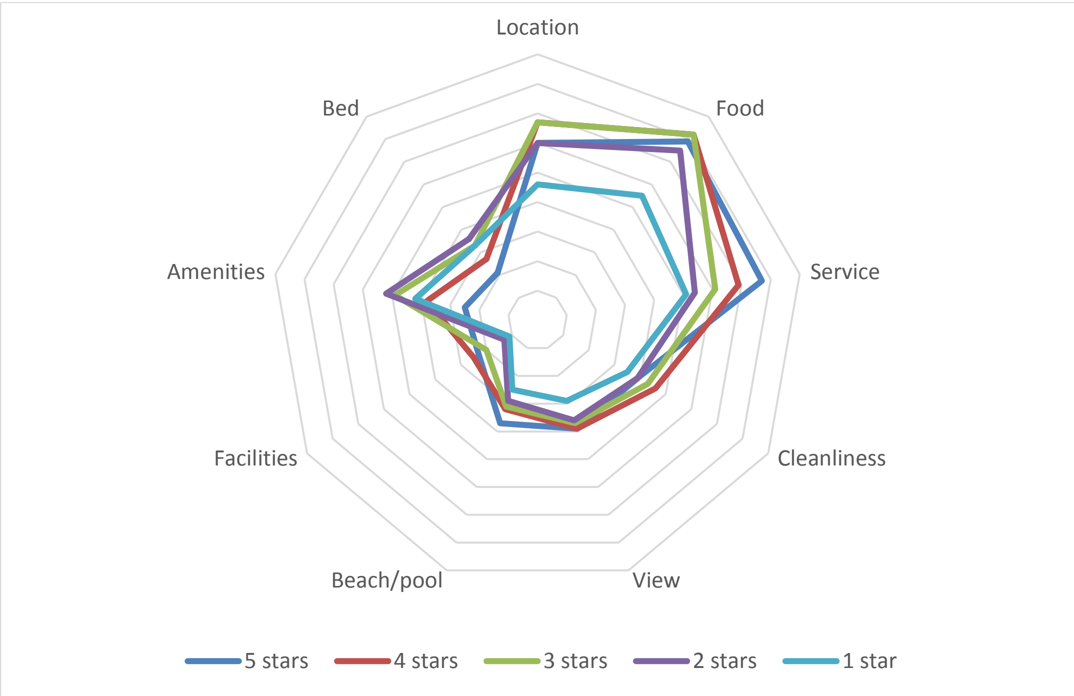
The coverage is $\sum \min(X,Y) = 0{,}82$ and $\sum Y = 2{,}5$, coverage=0,328.

The analysis results in two casual permutations, no29 and no31.

| Causal Permutation | Location | Food | Service | Cleanliness | View |
|---|---|---|---|---|---|
| 29 | 1 | 1 | 1 | 0 | 0 |
| 31 | 1 | 1 | 1 | 1 | 0 |

The analysis suggests that cleanliness and view may not be necessary services for customers. As the consistency and the coverage of the casual permutations is low, more terms should be taken into account in order to identify the best combination. The analysis should consider also the terms beach/pool, facilities, amenities and bed for better results.

The graph below shows user's interests based on their ratings.




The graph indicates that the term beach/pool may be a possible alternative for the analysis. If the term beach/pool replace view, the results change significantly.

Now the best casual permutations are:

| Location | Food | Service | Cleanliness | Beach/Pool | Consistency | Coverage |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 0,8117 | 0,5 |
| 1 | 1 | 1 | 1 | 1 | 0,8117 | 0,5 |

Consistency has increased but still more terms should be considered for more meaningful results.

### 4.11 Customer experience

Customer experience (CX) is the product of an interaction between an organization and a customer over the duration of their relationship. The following definition will be the basis of this analysis in this chapter.

“The Customer Experience originates from a set of interactions between a customer and a product, a company, or part of its organization, which provoke a reaction (LaSalle and Britton, 2003; Shaw and Ivens, 2005). This experience is strictly personal and implies the customer’s involvement at different levels (rational, emotional, sensorial physical and spiritual) (LaSalleand Britton, 2003; Schmitt, 1999).”

Classical economic theory regards the consumer as a logical thinker whose purchasing decisions are based on rational problem solving. In the present day, differentiating solely in traditional elements, like price and quality is no longer a sustainable advantage. Providing emotionally positive experience to the customers is a vital strategy for all businesses that are facing competition.

The collected user’s data from TripAdvisor are a valuable source for multiple purposes. A competitive analysis can be conducted based not only on the features of the company’s competitors but also how customers interact with its competitors. For example, let us assume that a hotel manager in Crete makes a competitive analysis. The first step is to determine who the existing and potential competitors are. This can be concluded by the features of the other hotels in the area and by how costumers evaluate these features. Then profile for each major competitor is created and their key strengths and weaknesses are determined. From the data we know how customers have rated each hotel in general and in every subcategory. Moreover, we are aware of customers profiles and what their expectations are. Customer experience, revealed by the reviews, showcases potential opportunities and threats. In addition, through the reviews the target user is identified. For instance, if the location of the hotel is negatively assessed by the majority of the reviewers because it is far from Point of Interest but positively for its pool, the hotel manager should emphasize and target customers accordingly. The data can also inform in what other services does the target customer care about and in what else is the target customer interested beyond the hotel. Finally, as user's expectations are set in part from their previous experiences with the hotel, we can estimate how perception has changed over the past few years and how it will change in the future.

The tourist sector with its low market entry barriers is attractive for SMEs, since many type of tourism require low capital investments and operating costs. On the other hand, Tourism SMEs are hence confronted with competitive disadvantages, such as poor economies of scale and scope, minimum potential for diversification and limited access to capital markets. These weaknesses can be confronted with a flexible structure and the offer of a unique customer experience. The constant improvement of the perceived experience provides the opportunity for the businesses to gain and preserve competitive advantage. For most leisure tourists, their holidays are of superior value, due to the temporarily limited time period per year and the investment of financial resources. In order to fulfill and if possible exceed customers' expectations, tourism businesses need to provide an exceptional customer experience.

As the majority of the customers book hotels though travel intermediates, like Booking which implement recommendation systems, hotels should utilize the available information. Hotels can rely on the user profile given by the travel intermediates to personalize the customer experience and make it memorable. For example, if the user profile indicates that the customer who has booked the room is particularly interested in cleanliness, the hotel manager should exploit the information and inform the staff accordingly.

# Conclusions

In this thesis we presented a tourism recommendation system based on user generated content through multiple approaches. Users' profiles were formed with keyword extraction. Applying collaborative filtering we were able to recommend to the target user hotels, based on ratings provided by other users with similar interests. Content based technique was used to generate suggestions based on the similarity of the hotels' attributes. Moreover, sentiment analysis was conducted in order to identify the polarity of the reviews and for that purpose a domain specific lexicon was built. The implementation of the lexicon based model was successful (F-score=0,84) and exceeded our initial expectations. In addition, utilizing Fuzzy set Qualitative Comparative Analysis we examined which set of users' interests leads to higher rating of the hotels.

Future research can focus on the improvement of the lexicon based model. With the addition of weights in every word the accuracy of the results is expected to increase. Furthermore, calculating the distance of the extracted keywords from words with known polarity we will be able to determine user's attitude towards every set of interests.

# Περίληψη


## Μεθοδολογία

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η δημιουργία συστήματος συστάσεων για τουριστικές υπηρεσίες και κυρίως για ξενοδοχεία, βασισμένο σε δεδομένα που έχουν δημιουργηθεί από χρήστες (User generated content).Τα συγκεκριμένα δεδομένα είναι ενδεχομένως πιο χρήσιμα από άλλους τύπους δεδομένων καθώς το 86% των χρηστών διαβάζει κριτικές για εταιρείες, προϊόντα και υπηρεσίες που έχουν δημοσιευθεί από άλλους χρήστες .Με την χρήση της εφαρμογής Scrapy, 10276 κριτικές χρηστών εξήχθησαν από το TripAdvisor. Η συγκεκριμένη πηγή επιλέχθηκε επειδή το TripAdvisor είναι πιθανά η πιο εύχρηστη και πιο διαδεδομένη εφαρμογή, που υποστηρίζει δημιουργία περιεχομένου από χρήστες.

Τα συλλεχθέντα δεδομένα περιλαμβάνουν 10276 κριτικές δημοσιευμένες από μεμονωμένους χρήστες για 4153 ξενοδοχεία στην Αθήνα, Θεσσαλονίκη, Κρήτη, Μύκονο και Ρώμη. Το πρώτο βήμα για την δημιουργία του προφίλ του κάθε χρήστη είναι η εξαγωγή λέξεων-κλειδιών (keyword) από τις κριτικές , που περιγράφουν τα ενδιαφέροντα του χρήστη. Με την εφαρμογή ανάλυσης κειμένου online-utility.org, οι πιο συχνές λέξεις καταμετρούνται και ταξινομούνται σε κατηγορίες. Εννέα κατηγορίες σχηματίζονται με βάση τα ενδιαφέροντα των χρηστών. Στη συνέχεια καταμετράται η εμφάνιση των λέξεων-κλειδιών σε κάθε κριτική και σχηματίζεται το προφίλ ενδιαφερόντων του κάθε χρήστη. Με αυτόν τον τρόπο , γνωρίζοντας το προφίλ των χρηστών μπορούμε να κάνουμε συστάσεις σε κάθε μεμονωμένο χρήστη, βάσει των προτιμήσεων που ο ίδιος έχει καταγράψει και δημοσιεύσει.

Με την αξιοποίηση ενός συστήματος σύστασης βασισμένο στην συνεργασία (collaborative filtering recommendation system) μπορούμε να προτείνουμε στον χρήστη ξενοδοχεία που έχουν αξιολογηθεί θετικά από άλλους χρήστες με παρόμοια ενδιαφέροντα και προτιμήσεις. Μέσω της συσχέτισης Pearson (Pearson correlation) θα υπολογίσουμε τη γραμμική συσχέτιση των ενδιαφερόντων των χρηστών και θα επιλέξουμε ποιοι χρήστες θα είναι η βάση για την δημιουργία συστάσεων.

Με την χρήση ανάλυσης συναισθήματος (sentiment analysis) των κριτικών μπορούμε να κατανοήσουμε την στάση του κάθε χρήστη έναντι του ξενοδοχείου και πως εκφράζεται για την εμπειρία διαμονής του σε αυτό. Η ανάλυση συναισθήματος είναι χρήσιμη στην περίπτωση που οι χρήστες δεν έχουν βαθμολογήσει τα ξενοδοχεία. Μέσω των εφαρμογών MonkeyLearn και LEXALYTICS μπορούμε να εξάγουμε και να

καταγράψουμε την συναισθηματική πολικότητα (polarity) των κριτικών. Παράλληλα χρησιμοποιείται μια εφαρμογή συναισθηματικής ανάλυσης βασισμένη σε λεξικό που δημιουργήθηκε για τον σκοπό αυτής της εργασίας 75 θετικά και 35 αρνητικά εννοιολογικά φορτισμένες λέξεις που εξήχθησαν από τις κριτικές , σύμφωνα με την συχνότητα εμφάνισής τους , σχηματίζουν ένα λεξικό , βάσει του οποίου θα γίνει η ανάλυση. Προκειμένου να αναλύσουμε και να κατηγοριοποιήσουμε τις κριτικές πολλαπλοί κανόνες δοκιμάζονται. Οι κανόνες βασίζονται στην υπόθεση, ότι εάν αφαιρέσουμε το άθροισμα των αρνητικών λέξεων από το άθροισμα των θετικών λέξεων και το αποτέλεσμα είναι πάνω από ένα όριο μπορούμε να συμπεράνουμε κατά πόσο η κριτική είναι θετική ή αρνητική.

Στη συνέχεια εφαρμόζεται ένα σύστημα συστάσεων βασισμένο στο περιεχόμενο (content based recommendation system) . Ο σκοπός αυτής της προσέγγισης είναι να προτείνουμε στον χρήστη ξενοδοχεία με παρόμοια χαρακτηριστικά με το ξενοδοχείο που έχει ήδη διαμείνει. Η ομοιότητα υπολογίζεται μέσω της συσχέτισης Pearson. Τα χαρακτηριστικά των ξενοδοχείων μπορούν να βρεθούν στο Booking και στο TripAdvisor. Συνδυάζοντας τα αποτελέσματα των δυο συστημάτων μπορούμε να παρέχουμε συστάσεις με βάση τα ενδιαφέροντα του χρήστη αλλά και τα χαρακτηριστικά του κάθε ξενοδοχείου.

Μια ενναλακτική προσέγγιση κατηγοριοποίησης των ενδιαφερόντων το χρηστών είναι μέσω της εφαρμογής Fuzzy set Qualitative Comparative Analysis (FsQCA). Η συγκεκριμένη ανάλυση εξετάζει ομοιότητες και διαφορές μεταξύ διαφορετικών εναλλακτικών καταστάσεων που μπορούν να οδηγήσουν σε ένα αποτέλεσμα. Μέσω της ανάλυσης θα προσπαθήσουμε να διερευνήσουμε η ύπαρξη ποιων ομάδων ενδιαφερόντων των χρηστών οδηγεί σε υψηλότερη βαθμολογία των ξενοδοχείων. Τέλος θα εξετασθούν διαφορετικές πιθανές εφαρμογές των δεδομένων που έχουν δημιουργηθεί από χρήστες , εκτός των συστημάτων σύστασης. Καθώς τα δεδομένα παρέχουν πλήθος πληροφοριών σχετικές με τον τρόπο αξιολόγησης των ξενοδοχείων από τους χρήστες , είναι δυνατή η πραγματοποίηση ανάλυσης ανταγωνισμού καθώς και η βελτιστοποίηση της εμπειρίας πελάτη.

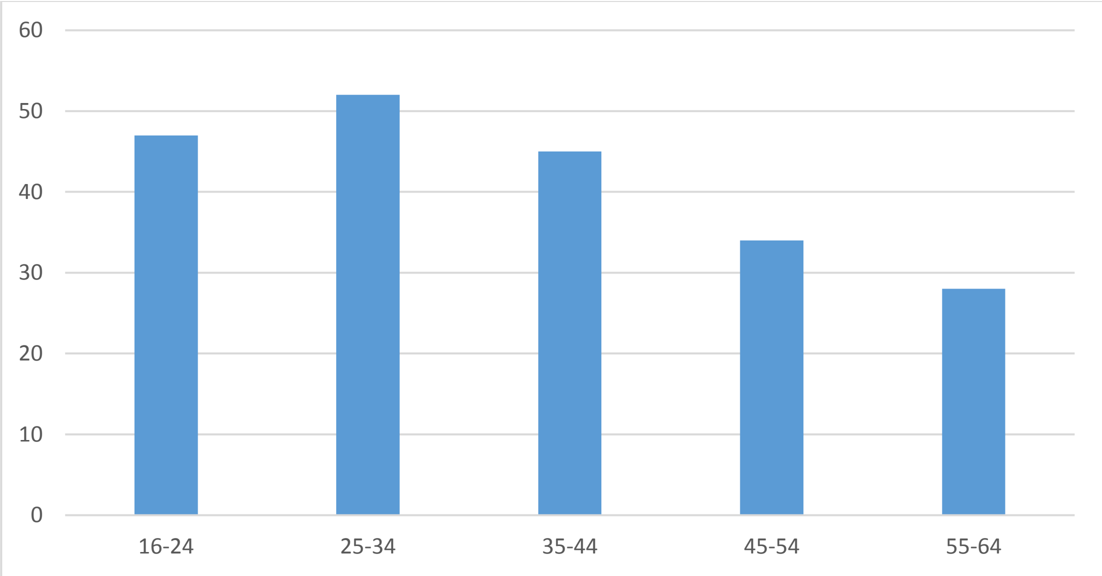
## Δημιουργία συστήματος συστάσεων

Τα δυνατά αποτελέσματα που προσφέρει μια μηχανή αναζήτησης στο διαδίκτυο σχετικά με ταξιδιωτικούς προορισμούς, ξενοδοχεία, εστιατόρια, μουσεία είναι σίγουρα πολύ χρήσιμα για τον χρήστη αλλά ενίοτε δυσκολεύουν την διαδικασία λήψης μιας απόφασης λόγω του τεράστιου όγκου δεδομένων. Τα συστήματα σύστασης είναι ιδιαίτερα χρήσιμα στην αντιμετώπιση του συγκεκριμένου προβλήματος καθώς μπορούν να παρέχουν προτάσεις στον χρήστη , βάσει των προσωπικών του προτιμήσεων και ενδιαφερόντων. Τα συστήματα σύστασης με εφαρμογή στον τουρισμό στοχεύουν να αντιστοιχήσουν τις ανάγκες των χρηστών με εναλλακτικές δυνατές προτάσεις , αξιοποιώντας την ανατροφοδότηση που έχει δοθεί στο σύστημα από προγενέστερη ταξιδιωτική εμπειρία του χρήστη. Έρευνες έχουν δείξει ότι ο τρόπος επιλογής προορισμού και οι προτιμήσεις των χρηστών αναφορικά με το ταξίδι τους βασίζονται σε πολλαπλούς παράγοντες , όπως την προσωπικότητα του χρήστη και προηγούμενες εμπειρίες.

Περιεχόμενο που έχει δημιουργηθεί από χρήστες είναι κάθε μορφής περιεχόμενο όπως κείμενο, εικόνες, βίντεο το οποίο έχει δημοσιευθεί από τους χρήστες στο διαδίκτυο. Η ανάλυση και αξιοποίηση του περιεχομένου αυτού μπορεί να δημιουργήσει νέα εργαλεία κατανόησης των αναγκών των καταναλωτών καθώς και νέους διόδους επικοινωνίας με τους καταναλωτές. Η σημασία αξιοποίησης του περιεχομένου διαφαίνεται ξεκάθαρα στα παρακάτω διαγράμματα.


(source:statista.com, USA, 2017)

Ποσοστό χρηστών διαδικτύου που δημοσιεύουν περιεχόμενο ανά ηλικιακή ομάδα




(source:statista.com, USA, 2017)

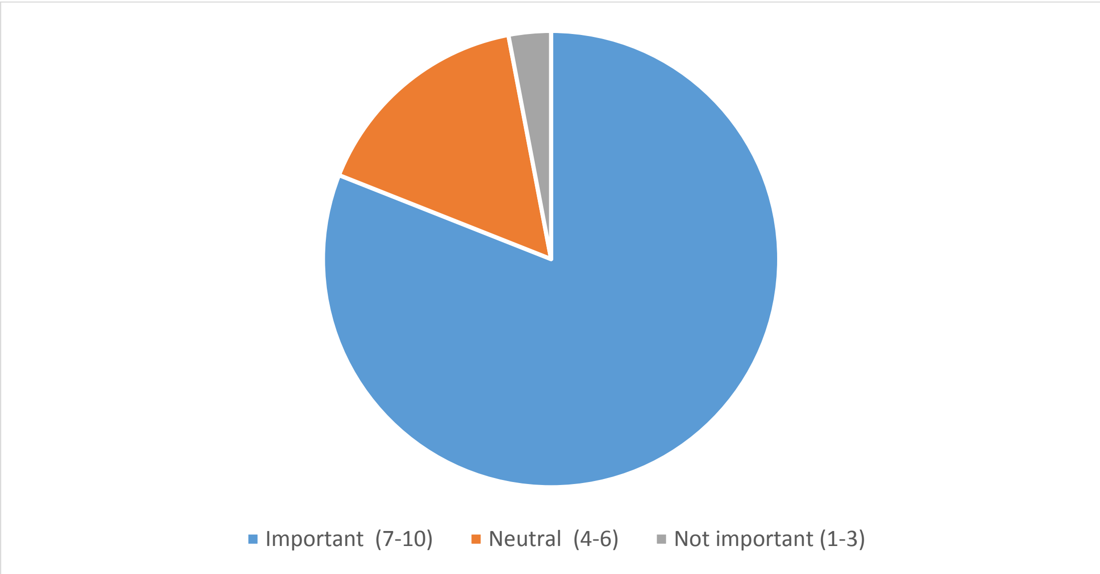
Το 2018 το 55,1% του παγκόσμιου πληθυσμού είχε πρόσβαση στο διαδίκτυο και το 86% των καταναλωτών διαβάζει κριτικές χρηστών για επιχειρήσεις και προϊόντα . Παράλληλα το 86% των καταναλωτών εμπιστεύεται, σε κάποιο βαθμό, τις κριτικές χρηστών σαν προσωπικές συστάσεις, με το 45% εξ αυτών να τις εμπιστεύεται έντονα. Το 2021 αναμένεται περισσότεροι από 2,14 δισεκατομμύρια άνθρωποι να αγοράσουν προϊόντα και υπηρεσίες μέσω διαδίκτυού. Το ενδιαφέρον στοιχείο που καθιστά το περιεχόμενο δημιουργημένο από χρήστες βάση της συγκεκριμένης ανάλυσης και ιδανικό για αξιοποίηση σε συστήματα συστάσεων είναι ότι αυξάνει την εμπιστοσύνη των καταναλωτών για την αγορά ενός προϊόντος κατά 73% και βελτιώνει την ανατροφοδότησή τους κατά 71% . Τέλος το 73% των καταναλωτών δηλώνει ότι εμπιστεύεται περισσότερο μια επιχείρηση εάν διαβάσει θετικές κριτικές ενώ το 50% ότι εμπιστεύεται λιγότερο μια επιχείρηση εάν διαβάσει αρνητικές κριτικές. Σαν συμπέρασμα το περιεχόμενο που δημιουργείται από χρήστες στο διαδίκτυο είναι πρωτότυπο, ενδιαφέρον και επιδρά πολλαπλώς στους καταναλωτές.

## Μελέτη περίπτωσης

Ο σκοπός της εργασίας είναι η ανάλυση και αξιοποίηση των ενδιαφερόντων των χρηστών μέσω του περιεχομένου που οι ίδιοι έχουν δημοσιεύσει στο διαδίκτυο για την δημιουργία συστήματος συστάσεων που θα παρέχει προτάσεις για ξενοδοχεία . Η πηγή του περιεχομένου είναι το TripAdvisor. Η εφαρμογή επιλέχθηκε λόγω αξιοπιστίας καθώς διαθέτει 661 εκατομμύρια κριτικές και 456 εκατομμύρια μηνιαίους μοναδικούς χρήστες. Παράλληλα το TripAdvisor ασκεί έντονη επίδραση στους καταναλωτές καθώς το 37% των καταναλωτών παγκοσμίως αναζητά πιθανούς προορισμούς στην συγκεκριμένη εφαρμογή και το 10% των καταναλωτών τελικά επιλέγει προορισμό μέσω αυτής της περιήγησης.

Η διαδικασία επιλογής ξενοδοχείου επηρεάζεται από προσωπικούς, ψυχολογικούς και κοινωνικούς παράγοντες. Το 22% των καταναλωτών επιλέγει προορισμό ταξιδιού βάσει των διαθέσιμων ξενοδοχείων και για το 57% είναι σημαντικό να διαμείνουν σε γνωστά και αξιόπιστα ξενοδοχεία.

Πόσο σημαντικές είναι οι κριτικές χρηστών όταν επιλέγετε ξενοδοχείο




(source: TripBarometer 2017/18, global report)

Πιο συγκεκριμένα θα αναλυθούν κριτικές χρηστών από το TripAdvisor σχετικές με την διαμονή σε 5 γεωγραφικές περιοχές. Με την χρήση το Scrapy έγινε εφικτή η εξαγωγή 10276 κριτικών δημοσιευθέντων από χρήστες. Από τα αρχικά δεδομένα ήταν γνωστά ο τίτλος της κριτικής, το όνομα χρήστη, η τοποθεσία του χρήστη, η κριτική, η επωνυμία του ξενοδοχείου, η τοποθεσία του ξενοδοχείου και η βαθμολογία που είχε δώσει ο χρήστης.

Τοποθεσία χρήστη: 35 χώρες, κυρίως από την Ευρώπη

Αριθμός χρηστών: 10276

Αριθμός ξενοδοχείων:4153

Τοποθεσία ξενοδοχείων : Αθήνα, Μύκονος, Κρήτη, Θεσσαλονίκη, Ρώμη

Κατηγορία ξενοδοχείου: 4 και 5 κατηγορίας

Βαθμολόγηση: πενταβάθμια κλίμακα Likert

Το πρώτο βήμα της ανάλυσης είναι η εξαγωγή λέξεων-κλειδιών από τις κριτικές. Οι λέξεις-κλειδιά αναγνωρίζουν τις βασικές πληροφορίες κάθε κριτικής, βοηθούν στην σύγκριση των κριτικών και καθιστούν εφικτή την δημιουργία του προφίλ των χρηστών. Με την χρήση της εφαρμογής online-utility.org μπορούμε να αναλύσουμε την συχνότητα εμφάνισης λέξεων στο σύνολο των κριτικών , οι οποίες εξετάζονται σαν ένα ενιαίο κείμενο. Ο παρακάτω πίνακας δείχνει την συχνότητα εμφάνισης κάποιων ενδεικτικών λέξεων στις κριτικές.

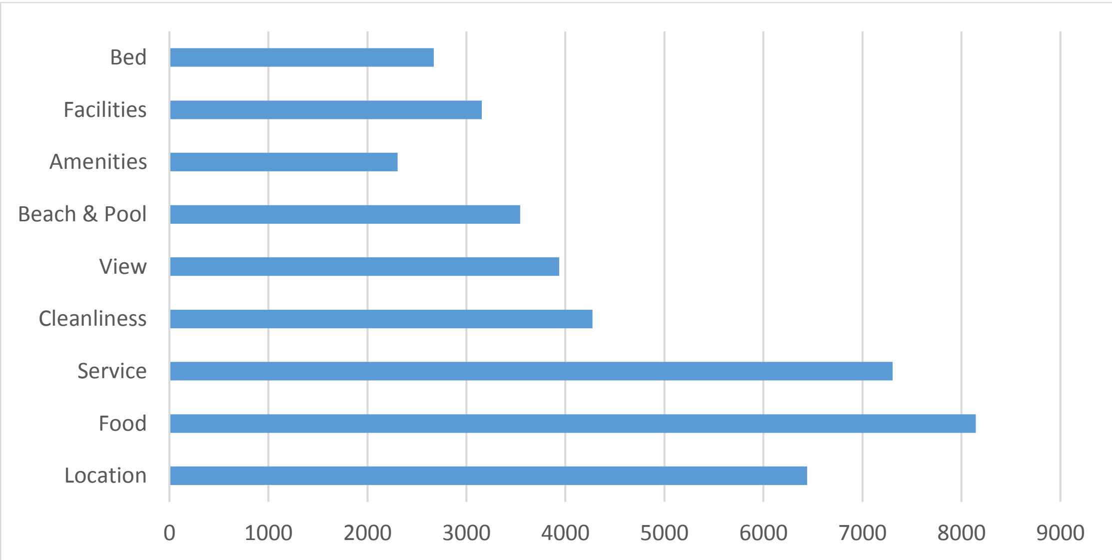
| Word | Total number |
|---|---|
| Staff | 6233 |
| Breakfast | 5103 |
| Food | 2877 |
| Pool | 2703 |
| Clean | 4171 |
| Beach | 2067 |
| Restaurant | 3190 |
| Service | 2468 |
| Area | 2156 |
| Location | 3176 |
| Bar | 2418 |
| Walk | 3273 |
| Reception | 1710 |
| Sea | 1514 |
| Bathroom | 1556 |
| Bus | 1861 |
| Located | 1018 |
| Dinner | 917 |
| View | 3005 |
| Balcony | 843 |

Παραδείγματος χάριν από την παρακάτω κριτική μπορούμε να εξάγουμε τις λέξεις-κλειδιά Location, bed, sleep, terrace, drink, breakfast, staff.

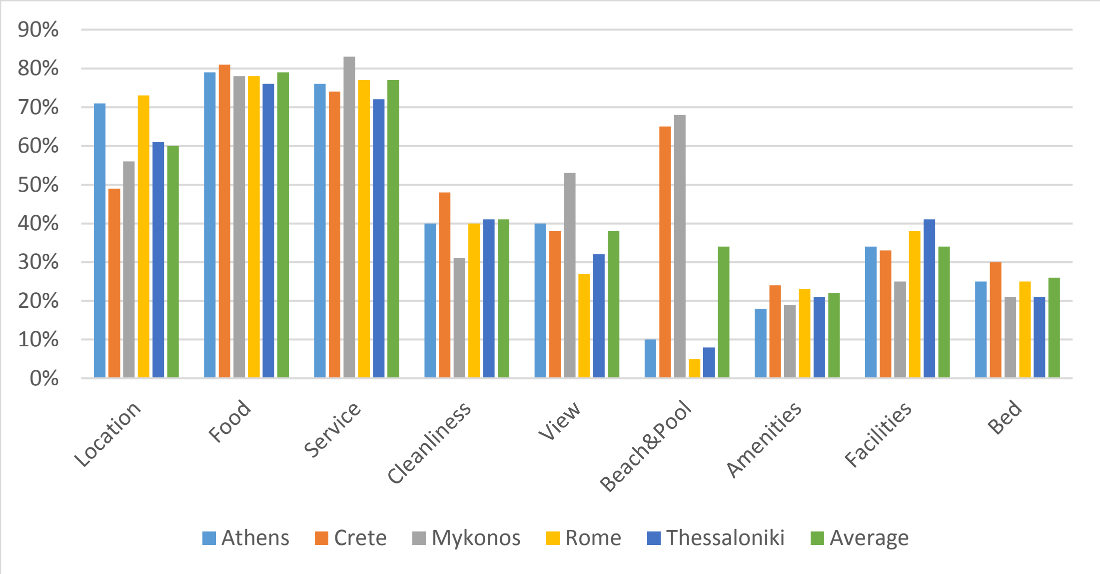
Στη συνέχεια οι λέξεις-κλειδιά ομαδοποιούνται και σχηματίζονται οι εξής 9 κατηγορίες, που περιγράφουν τα ενδιαφέροντα των χρηστών.

10) Location: location, area, located, walking, walk, metro, car, airport, bus
11) Food: breakfast, dinner, menu, food, restaurant, bar, drinks
12) Service: service, staff, reception
13) Cleanliness: clean, cleanliness, dirty
14) View: view, balcony, window
15) Beach & Pool: beach, pool
16) Amenities: spa, gym
17) Facilities: tv, wifi, Wi-Fi, bathroom, parking, elevator, lift, air condition, kitchen, facilities
18) Bed: bed, sleep, mattress, pillow

Αναζητώντας τις λέξεις–κλειδιά σε κάθε κριτική μπορούμε να κατανοήσουμε το προφίλ του κάθε μεμονωμένου χρήστη. Εφαρμόζοντας αυτήν την προσέγγιση ο χρήστης , του οποίου η κριτική αναφέρθηκε νωρίτερα , ενδιαφέρεται για τις κατηγορίες location, food, service, view, facilities. Αντίστοιχα, αναλύονται και οι 10276 κριτικές και κατηγοριοποιούνται ανάλογα. Το παρακάτω διάγραμμα δείχνει πόσες κριτικές ανήκουν σε κάθε κατηγορία.



Μέσω της στατιστικής ανάλυσης των στοιχείων προέκυψαν σημαντικές διαφοροποιήσεις τόσο στην εμφάνιση κάθε κατηγορίας ενδιαφερόντων ανά κατηγορία ξενοδοχείου, όσο και ανά προορισμό.




Για την δημιουργία του μοντέλου ενδιαφερόντων των χρηστών θα πρέπει να υπολογισθεί η ομοιότητα των ενδιαφερόντων. Ο υπολογισμός βασίζεται στον συντελεστή συσχέτισης Pearson. Ο συντελεστής συσχέτισης Pearson είναι +1 σε περίπτωση μίας τέλειας άμεσης γραμμικής σχέσης, -1 σε περίπτωση μίας τέλειας φθίνουσας (αντίστροφης) γραμμικής σχέσης και κάποια τιμή μεταξύ -1 και 1 σε όλες τις άλλες περιπτώσεις, που δείχνει το βαθμό της γραμμικής εξάρτησης μεταξύ των μεταβλητών. Όσο πιο κοντά είναι ο συντελεστής στο 1, τόσο ισχυρότερη είναι η συσχέτιση μεταξύ των μεταβλητών.

$$
\begin{aligned}
r_{XY} &= \frac{\text{Degree to which } X \text{ and } Y \text{ vary together}}{\text{Degree to which } X \text{ and } Y \text{ vary separately}} \\
&= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \\
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2\ \sum_{i=1}^{n}(y_i-\bar{y})^2},
\end{aligned}
$$

Υπολογίζοντας τον συντελεστή συσχέτισης Pearson του χρήστη $x_y$ με όλους τους υπόλοιπους χρήστες προκύπτουν τα εξής αποτελέσματα. 55 χρήστες έχουν τέλεια γραμμική σχέση (r=1) , έχουν δηλαδή ακριβώς τα ίδια ενδιαφέροντα με τον χρήστη $x_y$. Παράλληλα υπάρχουν 2262 χρήστες με συντελεστή συσχέτισης μεταξύ 0,5 και 0,7 και

358 χρήστες με συντελεστή συσχέτισης μεταξύ 0,7 και 1. Είναι σημαντικό να επισημανθεί ότι η συσχέτιση των ενδιαφερόντων υπολογίσθηκε χωρίς να λάβουμε υπόψιν τον προορισμό, τον σκοπό του ταξιδιού η τις βαθμολογίες των ξενοδοχείων. Με τον ίδιο τρόπο μπορούμε να εξετάσουμε τις κριτικές όλων των χρηστών και να ομαδοποιήσουμε τους χρήστες ανάλογα.

Σε αυτό το σημείο θα πρέπει να εξετάσουμε τις βαθμολογίες των ξενοδοχείων που έχουν δοθεί από χρήστες με υψηλό συντελεστή συσχέτισης (r>0,7) για να καταλήξουμε εάν αποτελούν επιτυχημένες προτάσεις προς τον χρήστη. Όταν ένας άνθρωπος βαθμολογεί ένα αντικείμενο , όπως ένα ξενοδοχείο, υποκειμενικοί παράγοντες υπεισέρχονται και δεν είναι πάντοτε ξεκάθαρο εάν έχει βαθμολογήσει θετικά , αρνητικά ή ουδέτερα. Προκειμένου να αξιολογήσουμε την βαθμολογία του κάθε ξενοδοχείου θα την συγκρίνουμε με την μέση βαθμολογία. Η μέση βαθμολογία των ξενοδοχείων από τους 412 χρήστες με υψηλό συντελεστή συσχέτισης (r>0,7) είναι 3,93 ενώ από τους χρήστες με τέλεια γραμμική σχέση (r=1) 3,53.

| Review | Full Revi | Rating | Hotel's | Locatio | Food | Service | Cleanli | View | Beach8 | Amenit | Faciliti | Bed | correl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| themisb | Nice. Brillia | 5 of 5 bubb | The Zillers | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | TRUE |
| holidayfar | My husbanc | 3 of 5 bubb | SENTIDO F | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0,790569 |
| etrev28 | I am finding | 5 of 5 bubb | Mykonos I | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0,755929 |
| Arjay R | You will be | 5 of 5 bubb | Kouros Ho | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0,790569 |
| Ellen F | This was on | 5 of 5 bubb | Hermes M | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| maggie19 | Spent 3 nigh | 4 of 5 bubb | Areos Hot | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0,790569 |
| | Stayed here | 3 of 5 bubb | Manoulas | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0,755929 |
| hannahpra | Currently st | 5 of 5 bubb | Berg Luxu | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0,755929 |

Με αυτήν την προσέγγιση μπορούμε να προτείνουμε ξενοδοχεία στο χρήστη , βάσει των ενδιαφερόντων του , σε 5 περιοχές . Παραδείγματος χάριν, στην Θεσσαλονίκη υπάρχουν 10 ξενοδοχεία που ταιριάζουν στις ανάγκες του και έχουν αξιολογηθεί πάνω από 3,93. Φυσικά εάν υπήρχαν περισσότερα δεδομένα για κάθε χρήστη το προφίλ του θα ήταν πιο ολοκληρωμένο και οι συστάσεις πιο ακριβείς και χρήσιμες.

**Ανάλυση συναισθήματος**

Ο όρος ανάλυση συναισθήματος (sentiment analysis) αναφέρεται στην εξαγωγή συναισθημάτων, απόψεων και στάσεων από έγγραφα κειμένου. Ο βασικός στόχος της ανάλυσης είναι η εύρεση και χαρακτηρισμός της πολικότητας του κειμένου για την κατανόηση της στάσης του συγγραφέα έναντι του αντικειμένου που περιγράφεται εντός του κειμένου. Η συγκεκριμένη ανάλυση είναι ιδιαιτέρως χρήσιμη στην περίπτωση δεδομένων που απουσιάζουν οι βαθμολογίες των χρηστών για τα ξενοδοχεία, όπως για παράδειγμα μια αξιολόγηση στο Facebook. Κάθε πληροφορία εντός ενός κειμένου μπορεί να κατηγοριοποιηθεί σε γεγονότα ή απόψεις. Γεγονός είναι μια δήλωση που μπορεί να αποδειχθεί σωστή ή λανθασμένη. Αντίθετα άποψη είναι η έκφραση των συναισθημάτων ενός ανθρώπου και δεν επιδέχεται απόδειξη. Η γνώμη ενός ανθρώπου μπορεί να βασίζεται σε γεγονότα ή συναισθήματα και είναι συνήθως υποκειμενική.

Η ανάλυση συναισθήματος έχει δυο προβλήματα που καλείται να επιλύσει. Το πρώτο είναι η ταξινόμηση του κειμένου με βάση την ύπαρξη ή απουσία υποκειμενικότητας από τον συγγραφέα του κειμένου και το δεύτερο η ταξινόμηση βάσει της πολικότητας του κειμένου. Στην συγκεκριμένη εργασία θα επικεντρωθούμε στην ταξινόμηση βάσει της πολικότητας των κριτικών. Καθώς η πολικότητα συσχετίζεται με συναισθήματα οι κριτικές θα κατηγοριοποιηθούν με βάση τα συναισθήματα που εκφράζουν οι χρήστες στις κριτικές τους. Ένα συνηθισμένο πρόβλημα που ανακύπτει στη ανάλυση είναι ότι οι λέξεις μπορεί να έχουν πολλαπλές ερμηνείες ανάλογα με την χρήση τους και τους σκοπούς του συγγραφέα.

Δυο εφαρμογές θα χρησιμοποιηθούν για τους σκοπούς της εργασίας , το MonkeyLearn και το LEXALYTICS. Το MonkeyLearn διαχωρίζει τις κριτικές σε θετικές και αρνητικές , ενώ το LEXALYTICS τις διαχωρίζει σε θετικές, αρνητικές και ουδέτερες. Ο παρακάτω πίνακας δείχνει την κατηγοριοποίηση των κριτικών μέσω της χρήσης των εφαρμογών.

| Full Review | Rating | Classification | Confidence | Classificatioı | Confidence |
|---|---|---|---|---|---|
| Nice. Brilliant location opposite the | 5 of 5 bubbles | Positive | 0.999 | positive | 0,535 |
| The upscale hotel Daios has much | 4 of 5 bubbles | Positive | 0.901 | positive | 0.277 |
| Nice hotel with friendly staff and free | 4 of 5 bubbles | Positive | 0.992 | neutral | 0,204 |
| I love this hotel, stayed here last ye | 5 of 5 bubbles | Positive | 0.999 | positive | 0,556 |
| Good Hospitality & Friendly Recepc | 5 of 5 bubbles | Positive | 0.978 | positive | 0,603 |
| If you want a hotel walking distance | 3 of 5 bubbles | Positive | 0.873 | positive | 0,465 |
| We stayed at San Antonio Summer | 2 of 5 bubbles | Negative | 0.922 | neutral | 0,033 |
| We stayed in a Panoramic Double I | 3 of 5 bubbles | Negative | 0.999 | neutral | 0,051 |
| It was unbelievable experience!! Ver | 5 of 5 bubbles | Positive | 0.976 | positive | 0,985 |
| We've just spent a week here and c | 5 of 5 bubbles | Positive | 0.997 | positive | 0,332 |
| Wow.......what can we say to give y | 5 of 5 bubbles | Positive | 0.997 | positive | 0,617 |
| A nice hotel with lovely interior and | 4 of 5 bubbles | Negative | 0.639 | positive | 0,315 |
| Mykonos should be so proud for ha | 5 of 5 bubbles | Positive | 1 | positive | 0,548 |

Παράλληλα δημιουργήθηκε ένα λεξικό από 75 θετικές και 35 αρνητικές λέξεις που προέκυψαν από την επεξεργασία των κριτικών. Το λεξικό εμπεριέχει λέξεις που έχουν γενικά καθορισμένη πολικότητα αλλά και λέξεις που προέκυψαν από το συγκεκριμένο σύνολο δεδομένων. Η ιδέα πίσω από αυτήν την προσέγγιση είναι να χρησιμοποιηθούν οι πιο συχνές λέξεις που υπάρχουν στις κριτικές και εκφράζουν πολικότητα για να κατηγοριοποιηθούν οι κριτικές.

Με την χρήση της εφαρμογής online-utility.org καταμετρούνται οι λέξεις και υπολογίζεται η συχνότητα εμφάνισής τους. Στην συνέχεια χωρίζονται σε 2 κατηγορίες. Η πρώτη κατηγορία έχει 75 θετικές λέξεις και η δεύτερη 35 αρνητικές. Ο παρακάτω πίνακας δείχνει τις πιο συχνές λέξεις ανά κατηγορία.

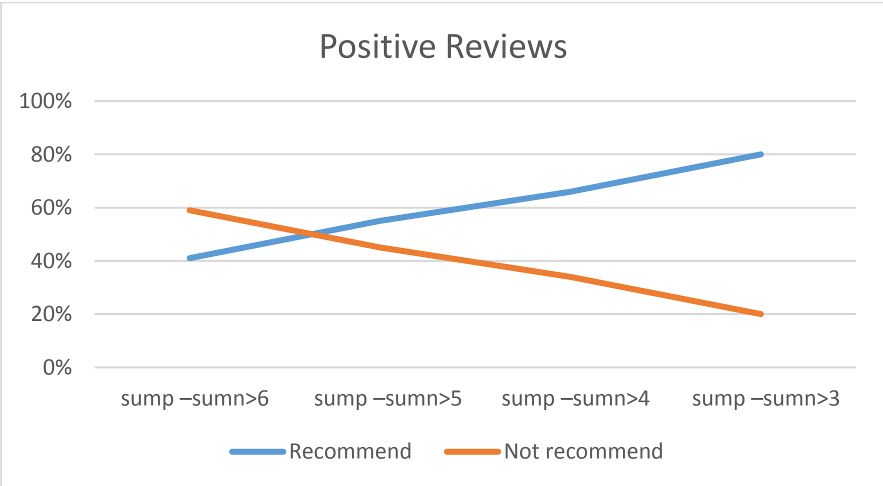
| Positive words | Negative words |
|---|---|
| Good | Small |
| Great | Problem |
| Nice | Noise |
| Lovely | Busy |
| Friendly | Wait |
| Well | Poor |
| Helpful | Noise |
| Amazing | Disappointed |
| Excellent | Unfortunately |
| Best | Fault |

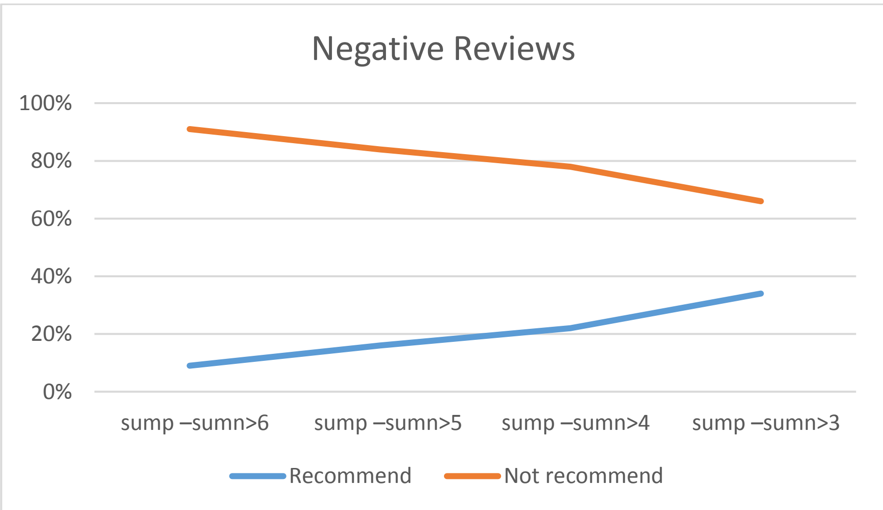
Καταμετράται η εμφάνιση των λέξεων κάθε κατηγορίας σε όλες τις κριτικές και υπολογίζεται το άθροισμα κάθε κατηγορίας ανά κριτική. Ο πίνακας δείχνει την εμφάνιση ενδεικτικών θετικών λέξεων σε μέρος των κριτικών.

| Full Review | good | great | nice | clean | lovely | friendl | well | helpful | amazin |
|---|---|---|---|---|---|---|---|---|---|
| Nice. Brilliant location opp | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| The upscale hotel Daios ha | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| Nice hotel with friendly sta | FALSE | FALSE | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| I love this hotel, stayed he | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE |
| Good Hospitality & Friendl | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | TRUE | TRUE |
| If you want a hotel walking | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| We stayed at San Antonio S | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| We stayed in a Panoramic I | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| It was unbelievable experi | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE | FALSE |
| We've just spent a week he | FALSE | TRUE | TRUE | FALSE | TRUE | FALSE | FALSE | TRUE | TRUE |
| Wow.......what can we say | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE |

Οι χρήστες διάκεινται θετικά έναντι των ξενοδοχείων, καθώς το 81% εξ αυτών έχει δημοσιεύσει θετική κριτική.1000 τυχαίες κριτικές θα χρησιμοποιηθούν δοκιμαστικά για να εξετάσουμε ποιος είναι ο βέλτιστος κανόνας. Από τις 1000 κριτικές 11 δεν έχουν κάποια θετική λέξη και 431 δεν έχουν κάποια αρνητική. Από τις 431 κριτικές χωρίς αρνητική λέξη μόνο στις 41 έχουν αξιολογήσει οι χρήστες τα ξενοδοχεία αρνητικά. Η υπόθεση που εξετάζεται είναι εάν αφαιρέσουμε το άθροισμα των αρνητικών λέξεων από το άθροισμα των θετικών και συγκρίνουμε το αποτέλεσμα με ένα κατώφλι (threshold), μπορούμε να κατηγοριοποιήσουμε την κριτική θετικά ή αρνητικά. Για να σχηματιστεί ο πιο αποδοτικός κανόνας θα εξετασθούν 4 κατώφλια και θα γίνει αντιπαραβολή με τις βαθμολογίες που έχουν δώσει οι χρήστες στα ξενοδοχεία.

- $sum_p - sum_n > 6$
- $sum_p - sum_n > 5$
- $sum_p - sum_n > 4$
- $sum_p - sum_n > 3$



Η πρώτη παρατήρηση είναι ότι ο πιο αποδοτικός κανόνας για τις θετικές κριτικές είναι ο $sum_p - sum_n>3$ επειδή επιστρέφει το 80% των θετικών κριτικών. Αντίστοιχα ο πιο αποδοτικός κανόνας για αρνητικές αξιολογήσεις είναι ο $sum_p - sum_n>6$ καθώς επιστρέφει 9% των κριτικών.

| | Recommended | Rejected |
|---|---|---|
| MonkeyLearn | 891 | 109 |
| $sum_p$-$sum_n$ >3 | 708 | 292 |
| $sum_p$ –$sum_n$ >4 | 572 | 428 |
| $sum_p$ –$sum_n$ >5 | 472 | 528 |
| $sum_p$ –$sum_n$ >6 | 347 | 653 |

**Αξιολόγηση αποτελεσμάτων**

Προκειμένου να αξιολογήσουμε τα συστήματα συστάσεων και την επιτυχία των προβλέψεων θα χρησιμοποιηθούν οι μετρικές ποιότητας των αποτελεσμάτων ανάκληση (recall) ,ακρίβεια (precision) και F-score.

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

| | Precision | Recall | F1-score |
|---|---|---|---|
| MonkeyLearn | 0,84 | 0,94 | 0,88 |
| $sum_p$-$sum_n$ >3 | 0,90 | 0,80 | 0,84 |
| $sum_p$ –$sum_n$ >4 | 0,92 | 0,66 | 0,77 |
| $sum_p$ –$sum_n$ >5 | 0,93 | 0,55 | 0,69 |
| $sum_p$ –$sum_n$ >6 | 0,94 | 0,41 | 0,57 |

Το MonkeyLearn έχει το υψηλότερο F-score(0,88), ακολουθούμενο από το κανόνα $sum_{p}$-$sum_{n}$ >3 (0,84), αλλά η ακρίβεια του είναι η χαμηλότερη από όλους του κανόνες που δοκιμάσθηκαν. Εξετάζοντας αποκλειστικά το F-score θα έπρεπε να επιλεγεί η εφαρμογή MonkeyLearn.Ο σκοπός όμως των συστημάτων σύστασης είναι να παρέχουν στους χρήστες όσο το δυνατόν πιο ακριβείς προτάσεις .

Προκειμένου να εκμεταλλευτούμε τα πλεονεκτήματα των δυο καλύτερων προσεγγίσεων , ένα νέο υπόδειγμα συναισθηματικής ανάλυσης υλοποιείται , το οποίο είναι σύνθεση των αποτελεσμάτων του MonkeyLearn και του κανόνα $sum_{p}$-$sum_{n}$ >3. Μέσω αυτής της σύνθεσης ο χρήστης θα μπορεί να επιλέγει ανάμεσα σε περισσότερα αποτελέσματα ή μεγαλύτερη ακρίβεια των αποτελεσμάτων.

Η συναισθηματική ανάλυση με την χρήση του λεξικού είχε καλύτερα αποτελέσματα από το αναμενόμενο, ταξινομώντας τις κριτικές με μεγάλη ακρίβεια. Η συγκεκριμένη προσέγγιση όμως ταξινομεί την κριτική σαν σύνολο και δεν μπορεί να διακρίνει επιμέρους στοιχεία της. Παραδείγματος χάριν ένας χρήστης μπορεί να έχει εκφραστεί πολύ θετικά για την εμπειρία διαμονής του στο ξενοδοχείο αλλά να είχε σχολιάσει ότι το πρωινό ήταν κατώτερο των προσδοκιών του. Με την παρούσα εφαρμογή του λεξικού δεν είναι εφικτή η αξιολόγηση της πολικότητας ανά κατηγορία. Ένας πιθανός τρόπος επίλυσης της αδυναμίας αυτής είναι με την εφαρμογή ενός νέου κανόνα που θα μετράει την απόσταση λέξεων με γνωστή πολικότητα από λέξεις-κλειδιά. Ο κανόνας θα μπορούσε ενδεικτικά να είναι:

(location) near (excellent, good, great)

(food) near (bad, horrible, awful)

Με αυτόν τον τρόπο θα μπορούσαμε να κατανοήσουμε την στάση του χρήστη για τις επιμέρους κατηγορίες. Παράλληλα μια ακόμη προσέγγιση που πιθανά θα βελτίωνε τα αποτελέσματα εφαρμογής του λεξικού είναι να προσθέσουμε βάρη στις λέξεις. Παραδείγματος χάριν , οι λέξεις καλός και εξαιρετικός περιγράφουν και οι δυο θετικά συναισθήματα, αλλά η λέξη εξαιρετικός είναι πολύ πιο έντονα θετική. Εάν και οι δυο λέξεις έχουν την ίδια αξία υποτιμάται η σημασία της έντασης των λέξεων .Με την χρήση των κατάλληλων βαρών σε κάθε λέξη αναμένεται να βελτιωθεί η ακρίβεια του συστήματος συστάσεων.

### Συστήματα συστάσεων με βάση το περιεχόμενο

Τα συστήματα προτάσεων με βάση το περιεχόμενο αξιοποιούν το περιεχόμενο των αντικειμένων για να προβλέψουν την σχέση τους με το προφίλ του χρήστη. Κάθε υποψήφιο αντικείμενο σε μια βάση δεδομένων διαθέτει κάποια χαρακτηριστικά που του περιγράφουν. Για παράδειγμα σε ένα σύστημα συστάσεων για βιβλία τα χαρακτηριστικά που μπορούν να χρησιμοποιηθούν είναι ο συγγραφές ,το έτος έκδοσης, το είδος του βιβλίου. Στη συνέχεια το σύστημα προτείνει βιβλία με παρόμοια χαρακτηριστικά με βιβλία που ο χρήστης έχει ήδη αξιολογήσει. Η ομοιότητα των χαρακτηριστικών υπολογίζεται με τον συντελεστή συσχέτισης Pearson.

Η χρήση της συγκεκριμένης προσέγγισης είναι για να προτείνουμε στον χρήστη ξενοδοχεία με παρόμοια χαρακτηριστικά με αυτά που έχει ήδη διαμείνει και αξιολογήσει θετικά. Ένα σύνολο 200 ξενοδοχείων θα εξετασθεί σαν παράδειγμα για το πως δομείται ένα σύστημα προτάσεων με βάση το περιεχόμενο. Η εύρεση των χαρακτηριστικών των ξενοδοχείων έγινε μέσω των εφαρμογών TripAdvisor και Booking.Τα συγκρινόμενα χαρακτηριστικά είναι Breakfast included, Location, kitchen facilities, Air-conditioning, Airport shuttle, Parking, Front desk 24/7, Restaurant, Double Bed. Εάν το ξενοδοχείο διαθέτει το εξεταζόμενο χαρακτηριστικό δίδεται η τιμή 1 στη κατηγορία, αλλιώς η τιμή 0. Ο χρήστης $x_y$ έχει διαμείνει σε ένα ξενοδοχείο 4 αστέρων στην Αθήνα. Επομένως τα διαθέσιμα ξενοδοχεία θα ταξινομηθούν αναλόγως. Βάσει των χαρακτηριστικών του ξενοδοχείου $x_1$ , το οποίο έχει αξιολογηθεί από τον χρήστη τα ξενοδοχεία με τα πιο όμοια χαρακτηριστικά είναι στον παρακάτω πίνακα.

| Hotel's Name | Hotel's Location | Hotel's Class | Correlation |
|---|---|---|---|
| X1 | Athens | 4 Stars | |
| X51 | Athens | 4 Stars | 0,661437828 |
| X134 | Athens | 4 Stars | 0,661437828 |
| X135 | Athens | 4 Stars | 0,661437828 |
| X140 | Athens | 4 Stars | 1 |
| X195 | Athens | 4 Stars | 1 |
| X199 | Athens | 4 Stars | 0,661437828 |

Βάσει της ομοιότητας των χαρακτηριστικών , 7 ξενοδοχεία μπορούν να προταθούν στον χρήστη. Ενοποιώντας τα αποτελέσματα των συστημάτων συστάσεων με βάση τη συνεργασία και το περιεχόμενο, μπορούμε να προτείνουμε στον χρήστη ξενοδοχεία σύμφωνα με τα προσωπικά του ενδιαφέροντα αλλά και βάσει των χαρακτηριστικών των ξενοδοχείων. Μέσω του νέου υβριδικού συστήματος η πλέον κατάλληλη πρόταση είναι το ξενοδοχείο $x_{195}$ καθώς τα χαρακτηριστικά των ξενοδοχείων έχουν τέλεια συσχέτιση (r=1) και τα προφίλ των χρηστών υψηλή συσχέτιση (r=0,86).

# Συμπεράσματα

Στην παρούσα εργασία δημιουργήθηκε μέσω εναλλακτικών προσεγγίσεων ένα σύστημα συστάσεων με βάση περιεχόμενο που έχει δημοσιευθεί από χρήστες στο διαδίκτυο , με σκοπό την παροχή προτάσεων για ξενοδοχεία. Μέσω της εξαγωγής λέξεων κλειδιών από τις κριτικές καταγράφηκαν και κατηγοριοποιήθηκαν τα προφίλ των χρηστών. Με την εφαρμογή συστήματος σύστασης βασισμένο στην συνεργασία έγινε δυνατή η παροχή προτάσεων στους χρήστες, βάσει των αξιολογήσεων που είχαν κάνει άλλοι χρήστες με παρόμοια ενδιαφέροντα. Με την εφαρμογή συστήματος σύστασης βάσει περιεχομένου μπορέσαμε να προτείνουμε στους χρήστες παρόμοιων χαρακτηριστικών ξενοδοχεία με εκείνα που είχαν ήδη διαμείνει και αξιολογήσεις. Στην συνέχεια πραγματοποιήθηκε ανάλυση πολικότητας των κριτικών , προκειμένου να ταξινομηθούν και δημιουργήθηκε ένα λεξικό από το συγκεκριμένο σύνολο δεδομένων. Αξιολογώντας την εφαρμογή του λεξικού κρίνεται επιτυχημένη (F-score=0,84). Τέλος μέσω ποιοτικής και ποσοτικής ανάλυσης εξετάσθηκε η αιτιώδης σχέση ανάμεσα σε μια ομάδα ενδιαφερόντων των χρηστών και την βαθμολογία που έδωσαν στα ξενοδοχεία.

Μελλοντική έρευνα μπορεί να επικεντρωθεί στην βελτίωση χρήσης του λεξικού. Μέσω της προσθήκης βαρών σε κάθε λέξη τα αποτελέσματα του λεξικού αναμένεται να βελτιωθούν. Παράλληλα υπολογίζοντας την απόσταση των λέξεων κλειδιών από λέξεις με γνωστή πολικότητα μπορούμε να κατανοήσουμε την στάση του χρήστη για κάθε κατηγορία ενδιαφέροντος.

# Bibliography


1. Agrawal R. (1999) Fast algorithms for mining association rules. 20th International Conference VLDB. pp.487-499.
2. Ahujaa R., Sawhneya A., Arifb M. (2017) Driving lean and green project outcomes using BIM: A qualitative comparative analysis. International Journal of Sustainable Built Environment. Volume 6, issue 1, pp. 69-80. Elsevier.
3. Anitha L., Kavitha Devi M. K. and Anjali Devi P. (2013) A Review on Recommender System. International Journal of Computer Applications. Volume 82, No3.
4. Baccianella S., Esuli A. and Sebastiani F. (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. LREC, pp. 2200-2204.
5. Baka V. (2016) The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. International journal of tourism management. Volume 53, pp. 148-162.
6. Bellogín Kouki A. (2012) Recommender system performance evaluation and prediction: An information retrieval perspective. Universidad Autónoma de Madrid.
7. Bennett J. and Lanning S. (2007) The Netflix prize. KDD Cup and Workshop, California.
8. Burke R. (2000) Knowledge-based recommender systems. In: Kent A. (ed.) Encyclopedia of library and information system. Volume 69, pp. 175–186. CRC Press.
9. Burke R., (2007) Hybrid web recommender systems. In: Brusilovsky P., Kobsa A. and Nejdl W. (Eds.), The Adaptive Web. LNCS 4321, (pp. 377 – 408) Springer.
10. Borràs J., Moreno A., Valls A. (2014) Intelligent tourism recommender systems: A survey. Expert Systems with Applications. Volume 41, Issue 16, pp.7370-7389, Elsevier.
11. Cantador I. and Cremonesi P. (2014) Cross-domain recommender systems. 8th ACM Conference on Recommender Systems. Silicon Valley USA, ACM.
12. Chandramouli B., Levandoski J.J., Eldawy A., Mokbe lM.F. (2011) Streamrec: a real-time recommender system. ACM SIGMOD. Greece.
13. Chua A. and Banerjee S. (2013) Reliability of reviews on the internet: the case of TripAdvisor. World Congress on Engineering and Computer Science. San Francisco, USA, WCECS.

14. Collomb A., Costea C., Joyeux D., Hasan O. and Brunie L. (2015) A study and comparison of sentiment analysis methods for reputation evaluation. University of Lyon.
15. Δαγλάς Σ. (2012) Ανάπτυξη συστήματος εμπλουτισμού αρχείου καθιερωμένων εγγραφών με κοινωνικές επισημειώσεις. Οικονομικό Πανεπιστήμιο Αθηνών.
16. Dalal A. (2017) Tourist destination recommendation system based on user facebook profile. Cloud computing research project. National college of Ireland.
17. Ekstrand M.D. (2014) Towards recommender engineering tools and experiments for identifying recommender differences. University of Minnesota.
18. Elliott T. Fuzzy set qualitative comparative analysis. Διαθέσιμο στο: http://thomaselliott.me/teaching.html [Πρόσβαση 2018]
19. European Parliament (2018) Online consumer reviews: The case of misleading or fake reviews. Think Tank. Διαθέσιμο στο: http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2015)571301
20. Felfernig A., Jeran M., Ninaus G., Reinfrank F. and Reiterer S. (2014) Toward the next generation of recommender systems: Applications and research challenges. Graz University of Technology, Austria.
21. Fritzner G. E. H. (2017) Automated information extraction in natural language. Norwegian University of Science and Technology.
22. Good N., Schafer J. B., Konstan J. A., Borchers A., Sarwar B., Herlocker J. and Riedl J. (1999) Combining collaborative filtering with personal agents for better recommendations. Group Lens Research Project, University of Minnesota.
23. Gorripati S. K. and Vatsavayi V. K. (2017) A community based content recommender systems. International journal of applied engineering research. Volume 12, Number 22. Research India Publications.
24. Herrada O.C., (2009) Music recommendation and discovery in the long tail. Universitat Pompeu Fabra.
25. Herlocker J. L. and Konstan J. A., Terveen L.G. and Riedl J. T. (2004) Evaluating collaborative filtering recommender systems. Publications ACM, New York.
26. Hosseini-Pozveh M, Zamanifar K., Naghsh-Nilchi A.R. (2016) A community-based approach to identify the most influential nodes in social networks. Journals of information science. Vol. 42, Issue 3, pp.204-220. Sage Journals.
27. Kardaras D.K., Kaperonis S., Barbounaki S., Petrounias I., Bithas K. (2018) An Approach to Modelling User Interests Using TF-IDF and Fuzzy Sets Qualitative Comparative Analysis. In: Iliadis L., Maglogiannis I., Plagianakos V. (eds)

Artificial Intelligence Applications and Innovations. pp. 606-615. AIAI 2018. IFIP Advances in Information and Communication Technology, vol. 519. Springer, Cham.

28. Kent R. (2008) Using fsQCA- A brief guide and workshop for fuzzy-set qualitative comparative analysis. Διαθέσιμο στο: http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2008/2008-10-teaching-paper-fsqca.pdf
29. Kim H. R. and Chan P. K., (2008) Learning implicit user interest hierarchy for context in personalization. Applied Intelligence. Volume 28, Issue 2, pp. 153–166. Springer Link.
30. Kumar M., Yadav D.K., Singh A. M. and Gupta V. C. (2015) A Movie Recommender System: MOVREC. International Journal of Computer Applications. Volume 124, No.3, pp.7-11.
31. Kumar V. P. N. and Reddy R.V. (2014) A survey on recommender systems (RSS) and its applications. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 2, Issue 8, pp.5254-5260.
32. Lee J. (2015) Local approaches for collaborative filtering. Georgia Institute of Technology.
33. Mahedi H., Falguni S., Dipankar C., Haider A. (2017) An empirical study of important keyword extraction techniques from documents. BRAC University, Bangladesh.
34. Mingang C. and Pan L. (2017) Performance evaluation of recommender systems. International Journal of Perfomability Engineering. (vol. 13- no. 8) pp. 1246-1256.
35. Mobasher B., Burke R., Bhaumik R. and Williams C. (2007) Towards trustworthy recommender systems: an analysis of attack models and algorithm robustness. ACM Transactions on Internet Technology. Vol. 7, No. 2, pp.1-41.
36. Montaner M. (2003) Collaborative recommender agents based on case-based reasoning and trust. Universitat de Girona.
37. Μπαλτζή Β., (2015) Γραφοθεωρητικές μεθοδολογίες για την δημιουργία συστημάτων σύστασης. Πανεπιστήμιο Πειραιώς.
38. Ozlem O., Gulla J.A. and Erdur R.C. (2013) A survey on challenges and methods in news recommendation. Universitaet Zuerich.
39. Park D.H., Kim H. K., Choi Y., Kim J. K. (2012) A literature review and classification of recommender systems research. Expert Systems with Applications. Volume 39. Issue 11. pp. 10059-10072. Elsevier.

40. Rafsanjani A.H.N., Salim N., Aghdam A.R. and Fard K.B. (2013) National Journal of Computational Engineering Research, Vol 03, Issue 5, Universiti Teknologi Malaysia.
41. Ragin C. C., (2009) Qualitative comparative analysis using fuzzy sets (fsQCA), In book Rihoux B. and Ragin C.C. (eds) Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques., pp.87-122, Sage Publications, USA.
42. Real L. and Rademaker A. (2015) HAREM and Klue: how to compare two tagsets fornamed entities annotation. Proceedings of the Fifth Named Entity Workshop, joint with 53rd ACL and the 7th IJCNLP, pages 43–49, Beijing, China, July 26-31, 2015 Association for Computational Linguistics.
43. Ricci F., Rokach L., Shapira B. and Kantor B.P. (2011) Recommender systems handbook. New York. Springer: Science and Business Media.
44. Rihoux B., Ragin C.C. (2009) Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques. Sage Publications.
45. Sarwar B., Karypis G., Konstan J. and Riedl J. (2001) Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th international conference on World Wide Web, pp.285-295, May 01-05, 2001, Hong Kong.
46. Schafer B.J., Konstan J.A. and Riedl J., (1999), Recommender systems in e-commerce, 1st ACM Conference on Electronic Commerce, pp.158-166, Denver USA, ACM.
47. Schneider C.Q. and Wagemann C. (2012) Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis. Cambridge University Press.
48. Smirnov A., Kashevnik A., Ponomarev A., Shilov N, Schekotov M., Teslya N. (2013) Recommendation system for tourist attraction information service. Proceedings of the 14th Conference of Open Innovations Association FRUCT. pp.141-147, September 11-15, 2013, Helsinki, Finland.
49. Taboada M., Brooke J., Tofiloski M. and Voll K. (2011) Lexicon-based methods for sentiment analysis. pp.267-307, Volume 37, No 2. Association for Computational Linguistics.
50. Thiengburanathum P., (2018) An intelligent destination recommendation system for tourists. Bournemouth University.
51. Trattner C. and Elsweiler D., (2018) Food recommender systems important contributions, challenges and future research directions. In: Berkovsky S., Cantador I. and Tikk D. (Eds), Collaborative recommendations: algorithms,

practical challenges and applications. (pp.653-685) World Scientific Publishing Co. Pte. Ltd.

52. TripAdvisor (2018). TripBarometer Global Report: 2017-2018. Διαθέσιμο στο: https://www.tripadvisor.com/TripAdvisorInsights/w4594 [ Πρόσβαση 2019].
53. Τσεχελίδου Ε., (2010) Συστήματα σύστασης βασισμένα σε τεχνικές νοημοσύνης σμήνους. Πανεπιστήμιο Πειραιώς.
54. Villegas N.M., Sánchez C., Díaz- Cely J., Tamura G. (2018) Characterizing context-aware recommender systems: A systematic literature review. Knowledge-Based Systems, Volume 140, pp.173-200. Elsevier.
55. Yun Y., Hooshyar D., Jaechoon J., Lim H. (2018) Developing a hybrid collaborative filtering recommendation system with opinion mining on purchase review. Journals of information science. Vol. 44, Issue 3, pp. 331-344. Sage Journals.
56. Yize L., Jiazhong N., Yi Z., Bingqing W., Baoshi Y. and Fuliang W., (2010) Contextual recommendation based on text mining. Beijing.
57. Zheng G., Zhang F., Zheng Z., Xiang Y., Yuan N.J., Xie X. and Li Z., DRN: a deep reinforcement learning framework for news recommendation. Ανάκτηση 2018 από: http://www.personal.psu.edu/~gjz5038/paper/www2018_reinforceRec/www2018_reinforceRec.pdf