ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

MANAGEMENT SCIENCE AND TECHNOLOGY

DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY

DISSERTATION

of

SPYRIDON GEORGIOU

TOPIC MODELING ON AIRBNB REVIEWS DURING THE PANDEMIC

Supervisor: Nikolaos Korfiatis - Associate Professor of Business Analytics University of East Anglia

Submitted as a part of the requirements for the acquisition of a Master's Degree (MSc) at Management Science and Technology

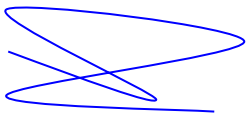
Athens, August 2021

This page is blank.

Βεβαίωση εκπόνησης Διπλωματικής εργασίας

«Δηλώνω υπεύθυνα ότι η συγκεκριμένη μεταπτυχιακή εργασία για τη λήψη του μεταπτυχιακού τίτλου σπουδών του ΠΜΣ στη Διοικητική Επιστήμη και Τεχνολογία του Τμήματος Διοικητικής Επιστήμης και Τεχνολογίας του Οικονομικού Πανεπιστημίου Αθηνών έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει εγκριθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών στην Ελλάδα ή το 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 διπλωματικής αναφέρονται στο σύνολό τους, δίνοντας πλήρεις αναφορές στους συγγραφείς, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο».

(Υπογραφή)



.....

< ΓΕΩΡΓΙΟΥ-ΣΠΥΡΙΔΩΝ>>

Φοιτητής MSc στη Διοικητική Επιστήμη και Τεχνολογία

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Acknowledgement

Thank you to my supervisor, Dr. Nikolaos Korfiatis, for your patience, guidance, and support. I have benefited greatly from your wealth of knowledge and meticulous editing. I am extremely grateful that you took me on as a student and continued to have faith in me.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Abstract

In the current dissertation, we are going to exploit LDA topic models to analyze AirBnB online comments and discover meaningful patterns. Topic modelling is a well-known and prevalent tool to extract concepts of small or large text corpora. These text collections often enclose hidden meta groups. Valuable information on online reviews is often ignored, therefore our study will concentrate on extracting important and profitable business insights. Moreover, this research project aims to provide a clear understanding of how COVID-19 pandemic has influenced the tourism industry and how AirBnB has dealt with this unique and unfamiliar phenomenon. To be more precise, this analysis consists of gathering data from Get the Data - Inside AirBnB. Adding data to the debate. We will handle data from spring of 2020, which was the initial period that CoVID-19 affected Greece. Before applying LDA algorithm, we are going to implement preprocessing techniques. Preprocessing is the process of bringing your text into a form that is predictable and analyzable for your task, fitting it to a certain schema. After that, we will implement and train our model so that we obtain the results that will be evaluated.

This page is blank.

Contents

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

1 Introduction

1.1 Big Data and Large Document Corpora

Big data tends to be voluminous, varied and scales up very fast with time because of internet. Large document corpora provide substantial opportunities for researchers and business managers. Data mining is the collective term for exploring large datasets using various techniques to find patterns in data. The aim of data mining is to analyse large datasets consisting of thousands to millions of attributes and data points (Zaki & Meira, 2014). Data mining uses six types of analysis: clustering, classification, regression, outer detection, sequential patterns and prediction (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Text mining is a specific area of data mining and it covers analysis of any type of text file. Topic modeling (TM) techniques discover latent abstract topics in a collection of documents and are applied to information retrieval, search, classification, clustering, natural language processing, machine learning and recommender systems. These techniques may be supervised, unsupervised or semi-supervised. They may be applied to fields like health, marine, transport/data network, agriculture, sociology for sentimental or opinion analysis and education.

1.2 Research Objectives

Our purpose is to investigate topic modelling with the implementation of Latent Dirichlet Allocation to provide insights to guests, hosts and AirBnB managers. The scope of this dissertation is to extract service quality dimensions and to instruct how Latent Dirichlet Allocation performs as it is a crucial tool for research.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ATHENS UNIVERSITY OF ECONOMICS & BUSINESS

The objectives of this thesis are:

1. To provide an example of a topic model analysis and to display how LDA can be handled to analyze large volumes of online reviews.
2. To contribute a simple interpretation of what LDA is and what is the logic behind it. Latent Dirichlet Allocation was chosen because it is the most common and popular method of topic modelling and it is easy to be understood and implemented.
3. To discern the impact of CoVID-19 on AirBnB and how it was handled from this huge organization.
4. To evaluate topic modelling as a method, to explore limitations, and to determine what is the importance of topic modelling to the analysis of online reviews.
5. To estimate if the results of topic modelling provide meaningful insights and can be exploited from guests, hosts, and managers.
6. To obtain a holistic view of how AirBnB reviews and scores function.
7. To calculate if the extracted topics can be correlated with the scores from reviews.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

2 *Definitions*

2.1 Text Analysis vs Text Analytics

In analyzing texts there are two broad terms in use, which are closely related but not exactly the same: Text analysis (or "text mining") and "text analytics" and their difference needs to be made clear (Grimes, 2017). Text Analysis is the term describing the very process of computational analysis of texts while Text Analytics involves a set of techniques and approaches towards bringing textual content to a point where it is represented as data and then mined for insights/trends/patterns. Case in point, Text Analysis helps translate a text in the language of data. And it is when Text Analysis "prepares" the content, that Text Analytics kicks in to help make sense of these data.

2.2 Key Terms of Text Analytics

In this section we describe some key terms that will be used in our study.

- Natural Language Processing (NLP): Natural Language Processing is a field of artificial intelligence that enables computers to analyze and understand human language.
- Document (D): any structured text (reviews, books, titles, tweets, etc.).
- Corpus: Collection of Documents.
- Vector: a quantity that has magnitude and direction and that is commonly represented by a directed line segment whose length represents the magnitude and whose orientation in space represents the direction broadly: an element of a vector space.
- Word Frequency: Word frequency can be used to list the most frequently occurring words or concepts in each text.

UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

2.3 Document-Term Matrix

Several times in text analysis we employ a particular tabular structure ("matrix") to represent words in the text as a table of numbers. The rows of the matrix represent the text responses to be analyzed, and the columns of the matrix represent the words from the text that are to be used in the analysis. Below, we will observe one simple example of two documents and how the document-term matrix would be:

- D1 = "Text analysis is useful".
- D2 = "Text analytics are useful".

	Text	Analitics	Analysis	Is	Useful	Are
D1	1	0	1	1	1	0
D2	1	1	0	0	1	1

Table 2.1: An example of a document-term matrix.

2.4 Topic Modelling

Topic modeling discovers abstract topics that occur in a collection of documents (corpus) using a probabilistic model. Pioneers on topic modelling are Papadimitriou, Tamaki, Raghavan, and Vempala (1998) with their publication Latent Semantic Indexing: A probabilistic analysis, and Hofmann (1999) with his publication probabilistic LSI. The technique was further developed by Blei, Ng, and Jordan (2003). There are varied methods for topic modelling, using diverse sampling algorithms for word selection and topic creation. There are many different methods of topic modelling like:

- Latent Dirichlet Allocation (LDA).
- Non-Negative Matrix Factorization (NMF).
- Latent Semantic Analysis (LSA).
- Parallel Latent Dirichlet Allocation (PLDA).
- Pachinko Allocation Model (PAM).

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

LSA looks at the frequency of words within a document and creates topics based on the frequencies of words occurring in each document. (Steyvers & Griffiths, 2007). Latent Dirichlet Allocation is one of the most well-known topic models. LDA groups words together based on how likely they are to appear in a document together (Blei et al., 2003). Correlated topic models explore the correlation of words to other words within a document. Topics are created based on the strength of correlations between words (Blei & Lafferty, 2007). Anyone using text documents in their research could have a potential use for topic modelling. Internet is providing an infinite amount of data that will only multiply over time. Topic modelling contributes to the easy and efficient processing of large amount of data. Topic modelling algorithms can produce similar topics to humans on an intuitive level, but on a metaphorical level there is not much interpretation. In other words, humans can apply emotions to create topics that are based on more of an emotional response rather than just words that relate together in a semantic way.

2.5 LDA

Large electronic documents can be structured, perceived, examined and recapped with the use of topic modelling. Latent Dirichlet Allocation (LDA) uses unsupervised learning, which provides a probabilistic Bayesian model for understanding text data and is one of the most prominent topic modelling techniques. Some assumptions have been deployed of LDA for Topic Modelling:

- Documents with similar topics use similar groups of words.
- Latent topics can then be found by searching for groups of words that frequently occur together in documents across the corpus.
- Documents are probability distributions over latent topics which signifies certain document will contain more words of a specific topic.
- Topics themselves are probability distribution over words.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

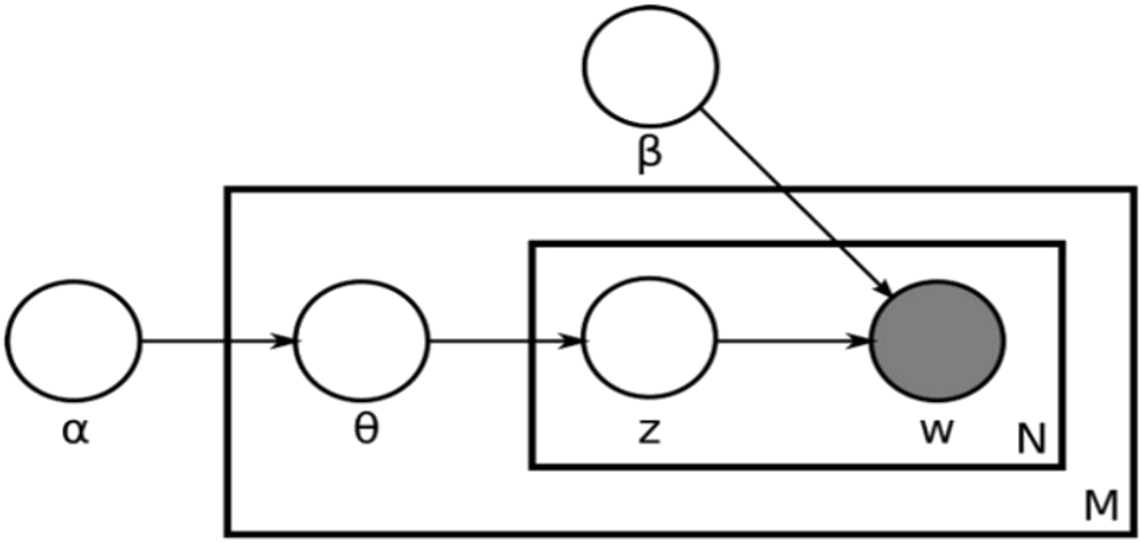


Figure 2.1: Plate Notation representing LDA model.

(Saura, José. (2019). A Three-Stage Methodological Process of Data Text Mining: A UGC Business Intelligence Analysis.)

- M denotes the number of documents.
- N is number of words in a given document.
- a is the parameter of the Dirichlet prior on the per-document topic distributions.
- b is the parameter of the Dirichlet prior on the per-topic word distribution.
- theta is the topic distribution for document i.
- varphi is the word distribution for topic k.
- z is the topic for the j-th word in document i.
- w is the specific word.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{i=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d)\, p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right)$$

Figure 2.2: Mathematical Formula.

An explanation with terms is that for every review there will be an outcome of a mixture of topics that will compose the document. We are going to choose the number of topics (Notation: K) and because of this choice a probability distribution is used to assign each word to a topic. Topics are proclaimed as probability distributions. LDA produces two kind of results:

- the words that belong to a review.
- the words that belong to a topic.

The logic behind the algorithm is that it is running behind each review and it is appointing the words to each one of the K topics at random. Furthermore, for every review (notation: R) it is going through each word (notation: W) and it is computing the probabilities p (topic t|review R) and p (word W|topic t). The probability p (topic t|review R) demonstrates how many words are an element of a topic t for a certain review R. The probability p (word W|topic t) exhibits the ratio of connections to a topic t from all reviews for the word W, thus this indicator aims to pick up how many reviews belong to a topic t as a result of the word W. Consequently, LDA depicts reviews as a concoction of topics. Finally, the algorithm is going to revise the probability of a word W that belongs to a topic t as shown below : p(word W with topic t) = p(topic t | review E) * p(word W | topic t).

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

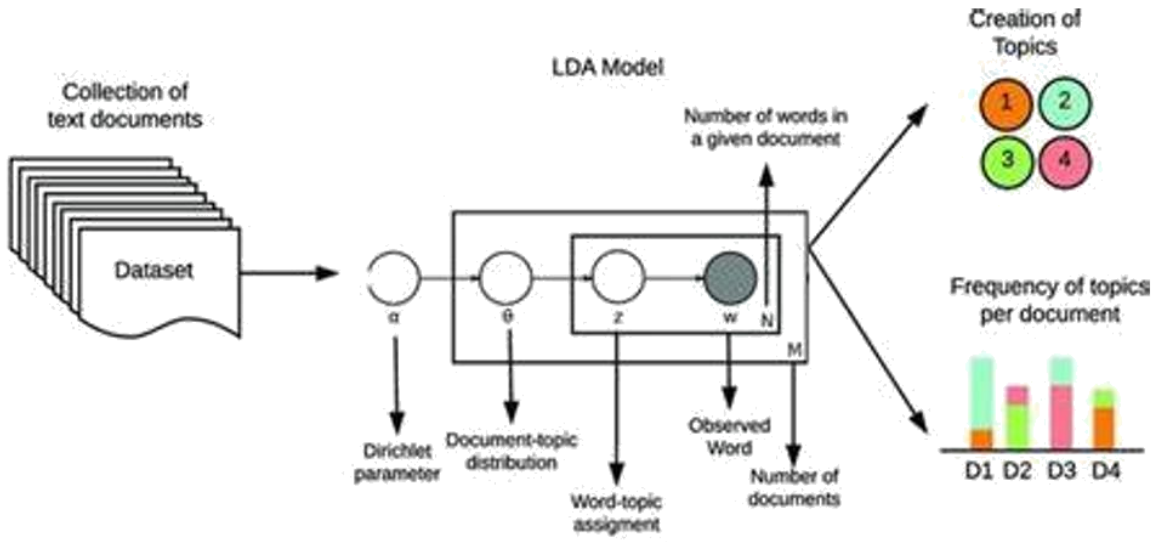


Figure 2.3: Schematic of LDA algorithm.

(Buenaño-Fernández, Diego & Gonzalez, Mario & Gil, David & Luján-Mora, Sergio. (2020) Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. IEEE Access. PP. 1-1.)

2.6 LDA "Hyperparameters"

The parameters of LDA are two and called Alpha and Beta. Alpha parameter is Dirichlet prior concentration parameter that represents document-topic density, with a higher alpha, documents are assumed to be made up of more topics and result in more specific topic distribution per document. While, Beta parameter is the same prior concentration parameter that represents topic-word density, with high beta, topics are assumed to made of up most of the words and result in a more specific word distribution per topic.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ UNIVERSITY OF ECONOMICS & BUSINESS

3 *Literature Review*

3.1 Online Reviews and Their Impact

Online reviews are utilized by organizations in order to gain valuable business insights as well as attract new customers and preserve the existing ones. They are the biggest source of social proof and they have a clear impact on sales and they are one of the most relied on sources of information for choosing holiday destinations (Murphy, Mascardo, & Benckendorff, 2007). On one hand, online reviews can help business managers find useful information, in order to build strategic plans and on the other hand, they can steer the decision-making process of the consumers. Business managers can find useful information that will help them build a strategic plan. Moreover, consumers think of them as an essential part that guide their decision-making process.

It is useful to refer to eWOM because it is a key concept that will clarify the importance of online reviews. Electronic word-of-mouth, or "eWOM" is "any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet" (Hennig-Thurau et al., 2004, p. 39). The importance of eWOM as a consumers' source for information gathering, and subsequently decision making in the tourism industry is highlighted by both the studies of Papathanassis and Knolle (2011) and Black and Kelley (2009). Researchers suggest that online reviews are the most important source of eWOM in the tourism industry (Gretzel & Yoo, 2008; Ip, Lee, & Law,2012; Parra-López et al., 2011).

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Thus, there are many reasons that lead consumers to consult online reviews, such as credibility and content. Therefore, it is common knowledge that hospitality and tourism heavily depend on online reviews, either positive or negative.

Consequently, social media marketing has surfaced as a productive and rewarding field that requires to be studied. The valence of online reviews appears to be huge and it refers to the average numerical rating, for example positive, negative or neutral. It was discovered that although the positive reviews contribute to a customer's choice of acquiring a product, negative reviews have greater impact and immensely trouble customers from purchasing. (Dellarocas, Zhang and Awad 2007; Floyd et al. 2014). This valuable finding resulted in many researches that struggle to explore the valence intensity. Some researchers found positive reviews to have minimum or no effect (Duan, Gu and Whinston 2008), while others deducted that negative reviews have a greater effect and are more influential. (Tsang and Prendergast 2009; Xie, Zhang and Zhan 2014; Chevalier and Mayzlin 2006). Nevertheless, these discoveries varied across different industries. We are going to focus on the hospitality industry as it is in the scope of our thesis. Xie, Zhang and Zhang (2014) revealed that certain hotel attributes, such as services, location, price, cleanliness, and room, have great impact on online reviews and are undoubtedly correlated with a hotel's performance. This discovery will be of great value to our dissertation and will be exploited in our research. More research must be dedicated to the impact that online reviews have on the hospitality industry.

3.2 AirBnB Insights

AirBnB is a peer-to-peer platform that is arguably the world's largest accommodation provider and is referred as part of the sharing economy. The statement about the sharing economy is ambiguous till this day, since not all characteristics of a sharing economy are represented to peer-to-peer accommodation. Nevertheless, this is not in the scope of our research and for us to criticize. To continue with AirBnB, it acts as the main component that connects the guests and the hosts and provides an alternative way of accommodation. It is not only vital for tourists that are searching for a "cheap" place to crash but also for the hosts, for the simple reason that it allows them to attain a new source of income. The company was founded very recently in 2008 and represents a new trend. When talking about AirBnB there are two consumer perspectives, one for the hosts and one for the guests that try to find apartments.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

The field of our study will focus on guests as we are going to explore online reviews and how they influence them. AirBnB is offering the aspect of online reviews so that any guest can optimize his/her travelling experience through searching, understanding and comparing this information. These reviews offer a deeper, richer and unexploited volume of information. In addition, they appear in an unstructured way that makes it challenge to be processed and to produce valuable intuitions.

3.3 COVID-19 And the Effect on Hospitality

COVID-19 is a new strain of SARS (SARS-COV-2) that has evolved into a global pandemic with serious implications on multiple fields such as tourism, economy, health system etc. Some of the most common impacts that derived from this pandemic includes job loss, revenues losses, decreasing market demand and many more. Specifically, a research from EY indicates that Greek Tourism has depicted 78 percent decline recorded in tourism receipts during the first nine months of the year, compared to the same period in 2019. Although the hospitality industry will identify a way to recover from such a devastating blow, organizations that associate with tourism are likely to perform massive changes to their operations. This fact will have therefore for several marketing researches to arise regarding the customer's attitudes and behaviors on the after-COVID era. Contemporary findings have indicated that sanitizing efforts, social distancing implementation and employee training of health and safety protocols will be key aspects that will determine the selection of customers regarding the hospitality and tourism industry.

3.4 Important Publications On LDA

First, we embarked on our research with the article of "David M. Blei, Andrew Y. Ng, Michael I. Jordan" by the name of "Latent Dirichlet Allocation". This scientific paper provided a new innovative meaning such as LDA that is used by researchers who are trying to extract information via unstructured text data. The aim of the Latent Dirichlet Allocation probabilistic model is to allow sets of observations to be grouped when they have similar data. These groups need to be calculated through complex mathematical expressions and are called topics. This article extensively describes all mathematic expressions, but we will comment and explain them in the methodology section below. The researchers compare LDA with three other simpler latent variable models for text such as unigram model, mixture of unigrams and probabilistic latent semantic indexing (pLSI). After this comparison they bring forth a geometric interpretation of all these models and conclude that LDA has the best functionality. Furthermore, De Finetti's representation theorem has a key contribution because it provides LDA the attribute of exchangeability.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

This theorem suggests that "An infinite sequence of random variables is infinitely exchangeable if every finite subsequence is exchangeable". Finally, there was a practical example of an LDA model on 16000 documents from a subset of the TREC AP corpus (Harman, 1992). This experiment used 10% of the data for test purposes and the rest for training models. On the data were applied except from the LDA model, a smoothed unigram model, a smoothed mixture of unigrams and a pLSI model. After retrieving the data LDA always functions greater than the other models. In conclusion, LDA is a model that theoretically and practically can enhance text mining and can have multiple extensions. After understanding the meaning and purpose of LDA, we thought useful to dive in the article of "Seshadri Tirunillai and Gerard J. Tellis" that is called "Mining Marketing Meaning from On-line Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation" to further comprehend the applications of LDA and how it affects the text mining of online reviews. This research focuses on getting worthy insights from unstructured text data, in order to form quality dimensions that will provide valence to an organization. The data that were used for this endeavor were collected by the researches without any help and accumulated to a number of 350000 consumer reviews. These data belonged to 15 firms that were distributed to mobile phones, computers, toys, footwear, and data storage markets. The results provided different number of topics for each market. For example, the optimal number of topics for a mobile phone market is 10 but for computer markets are 8. One of the limitations of this method is that the topics must be labelled by the researchers using the most frequent words of each topic and this can result in the loss of in-formation. After the extraction of topics, the researchers compared the result in dimensions with Consumer Reports, which is an extensive study for quality dimensions from (Mitra and Golder 2006; Tellis and Wernerfelt 1987). The overlapping was measured with the Jaccard coefficient and appeared to be high between the extracted dimensions and the already existing ones from Con-sumer Reports. Then the heterogeneity of dimensions was calculated in each firm. For assessing heterogeneity of dimensions, the dominance of dimensions within a brand in terms of reviews was used. Then the brand mapping was calculated with the assistance of the assigned dimensions. Concluding, this study was a pioneer in the field of text mining. It extracted useful insights regarding service quality dimensions, but also addressed some problems. The limitations existed because of the high computational power required in order to execute the computations and due to the fact that this research only focuses on reviews and not on online forums, microblogs, tweets etc. Then we move to the article "Measuring service quality from unstructured data: A topic modelling application on airline passenger’s online reviews" by "Nikolaos Korfiatis, Panagiotis Stamolampros, Panos Kourouthanassis, Vasileios Sagiadinos". This article aims to extract the dimensions of service quality for flight companies using structural topic modelling which is an extension of LDA. Structural topic modelling is considered as the most appropriate model to tackle the problems of LDA models, such as the definition of topics and the consideration of online reviews has no exogenous covariates.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

The data source was TripAdvisor and the online reviews that were gathered from it were from airline passengers and accumulated to a total of 557208. The reviews had an overall score for eight service aspects of the flight. The authors also computed the flight distance. After cleaning the data, a total of 184502 reviews were used for the topic estimation. The analysis was performed in R and generated a 20-topic solution that had the best relationship between exclusivity, likelihood, and semantic coherence. The topic labelling was achieved with the assistance of two airline customer service experience experts. The analysis that was performed included an ordinal logistic regression in order to add prognostic profit to the model and another problem that had to be dealt was the temporal variation of service levels and that happened by acknowledging the seasonal variations in the topic explanation. With these two techniques two main limitations of the model were tackled. Finally, correspondence analysis for service quality dimension extracted from the STM topic solution was executed by choosing seven of the most common topics from the corpus. The authors of this research find STM as a valuable asset that can be exploited from managers of big corporations, but still holds some constraints, such as biased responses that can influence and shift the meaning of the online revies, thus even more control variables should be added to erase any misinformation that may arise.

Following the above three articles that will be used extensively to shape our analysis and re-search , we are going to summarize the publication "Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia" of "I R Putri,R Kusumaningrum". This paper is outside the scope of my dissertation, but it will provide insights about the way LDA behaves and functions in a different kind of topic. The purpose of this research resides in the evaluation of text reviews and whether they are positive or negative. The text data were collected from TripAdvisor and the analysis was performed with Matlab programming. The scenario consisted of two experiments in order to obtain the best classification model for sentiment analysis. Moreover, there was a comparison of the first parameter which was the number of iterations and the second parameter which was the number of topics. The aim of the first scenario is to compare each combination and obtain the information of parameter affect for each accuracy in classification model. The second scenario aims to identify the best parameter combination that given the highest accuracy value. Explaining the results concluded that iterations and topics are parameters that have a vital meaning for a good accuracy in the model classification using LDA. Regarding sentiment analysis the more the number of topics and iterations are the better we receive. So, in this study LDA was used as a classification method for sentiment analysis successfully.

The next study we am going to analyze is the one named "How people reflect on the usage of cosmetic virtual goods: a structural topic modeling analysis of r/dota2 discussions" by "Denis Bulygin, Ilya Musabirov". This research is innovative and address a very modern topics that needs discussion, which is the analysis of discussions for cosmetic items that are bought for the online game Dota2.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

The data were gathered from comments on ReddiT and they were handled in a way that the researches would gain the knowledge of what dimensions emerge from discussions over those items, what dimensions prevail in the discussions and how is the change of the virtual items price connected with the dimensions. These intuitions are acquired via structural topic modelling which was rated by the authors as the most appropriate model. For this purpose, they handled three datasets, one consisted of the unique items, one consisted of the unique reviews and the last one referred to the market price of each item and how its price was altered. The researchers had to define which covariates should be used and the best number of topics. This task was completed with the use of R language package stm and 35 topics emerged for this field of study. Moreover, every topic that was created had 2 sets of words that described it, the words that have the highest probability to appear on this and the words with the highest FREX (frequency and exclusivity) score. The information gathered were very enlightening and useful about a field that has endless research potential.

The next article we have studied is called "Automatic Topic Discovery of Online Hospital Reviews Using an Improved LDA with Variational Gibbs Sampling" from "Richard de Groof and Haiping Xu". Before we comment on this study, it is a good opportunity to check what Marko-Chain Monte-Carlo methodology (Collapsed Gibbs Sampling) accomplishes and why it used. Gibbs sampler or Markov-Chain Monte-Carlo (MCMC) is an algorithm commonly used in statistics for the acquisition of observations from a certain multivariate probability distribution while the usage of an entire sample is troublesome. The authors discovered the inefficiency of this method when it is applied on multiple trials. Thus, to enhance the efficiency of the LDA model they utilized Varational Gibbs Sampling (VGS) which has a better performance. Collapsed Gibbs Sampling depends on random numbers while VGS aims to eliminate these random numbers. As a result, more consistent and accurate results emerge between trials. To prove this theory, online reviews for hospitals from Yelp.com were gathered and these two variations of LDA were assessed. The implementation of VGS methodology was better overall than the CGS and the most vital part of it was its consistency throughout the trials.

The essay composed by "D. Teja Santosha, K. Sudheer Babua, S.D.V. Prasada, A. Vivekanandab" with the title of "Opinion Mining of Online Product Reviews from Traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet" is another topic modelling research that manipulates LDA. These analysts adapted a Feature Ontology Tree to LDA clusters. The topics that were discovered were represented as features with the use of FOT. Additionally, SentiWordNet was used to compute the opinion bearing words. SentiWordNet is a lexical resource in which each WordNet syn-set is associated to three numerical scores Obj(s), Pos(s) and Neg(s), describing how objective, positive, and negative the terms contained in the syn-set are.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

The results were positive and product feature extraction can be conducted with the implementation of a FOT on LDA topics clusters. Besides that, traditional LDA could be transformed so that opinions about the features can be compared with SentiWordNet.

Another beneficial article for the comprehension of Latent Dirichlet Allocation topic modelling is named "Quantitative analysis of large amounts of journalistic texts using topic modelling" by "Carina Jacobi, Wouter van Atteveldt and Kasper Welbers". The authors manipulated the articles of New York Times that dealt with nuclear technology. They performed LDA topic modelling with the use of R statistical package and came to the conclusion that topic modelling is not yet up-to-date so that it will perform a full automatic analysis, but it can be employed for a semi-automatic analysis in which the researchers inspect the words of each topic and label them. From their findings they deducted that this model has two restraints. The first one indicates that not all topics can serve as substantive word clusters and the second one that the topics do not produce valence. Overall, this paper proved the theory that LDA topic modelling can be of great value to journalism research.

3.5 Applications Of LDA

Here are some applications of topic models, much of which are extensions or straightforward applications of LDA:

- Discovery of overlapping communities in social networks (Airoldi et al., 2008).
- Dynamic topic models (Blei and Lafferty, 2006), which have been used to analyze the progression of Science articles from 1880 to 2000.
- Collaborative topic models (Wang and Blei, 2011), which are used for content recommendation at the New York Times.
- Population analysis of 2 billion genetic measurements (Gopalan et al., 2016).
- Application of Latent Dirichlet Allocation in Online Content Generation (Yang, 2015).
- An Application of Latent Dirichlet Allocation to Analyzing Software Evolution (Erik Linstead, Cristina Lopes, Pierre Baldi, 2008).
- Applications of Latent Dirichlet Allocation Algorithm of Published Articles on Cyberbullying (Rommel L. Verecio, 2017).
- Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews (Korfiatis, Stamolampros, Kourouthanassis, Sagiadinos, 2018).

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

In addition, LDA has some essential applications on the fields of learning analytics, business analytics, bioinformatics, and sociology. In bioinformatics has been implemented to strengthen researchers' ability to comprehend the structure and decipher biological information. On the field of sociology, researchers suggest that they can benefit vastly from automated topic modelling and that it can be exploited on opinion analysis, media studies, etc. In learning analytics topic modelling has been extensively used to provide evidence of quality in content (text) produced by students (Lang, Siemens, Wise,& Gaševic,´ 2017), in cases like:

- Automated essay scoring: TM has been used to calculate the resemblance of an essay to a predefined set of essays and, based on those similarities, calculate a single, numeric measure of essay quality.
- Document cohesion: TM has been used to calculate measures indicating the quality of document writing by computing the cohesion of the text.
- Topic extraction from users posts in various media.

In business analytics topic modelling is applied on many purposes:

- Social Media Monitoring: to identify topics that people talk about in social media and provide added value to business understanding of customers' likes/dislikes.
- Customer Service: the application of topic modelling assists to detect topics and service quality in customer's posts and inquiries.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

4 *Methodology*

4.1 Definition of Business Objective

The business objective that needs to be addressed is clear and straightforward. Organizations must profit from online reviews so that they shape their brand reputation, affect sales, make their brand reliable, enhance their visibility on the internet, increment their search efficacy and obtain profitable information from customers. Business managers have developed the need to gain insights from critical information and to achieve this, application of Natural Language Processing is mandatory and vital. Hence, the potential added value of this procedure makes it compulsive to be utilized for mining practical insights. In addition, COVID-19 has increased the significance of research in the field of hospitality marketing. Every manager can potentially add value to his/her organization by utilizing modern techniques, such as deep and machine learning algorithms.

4.2 Dataset Description

Our data are sourced from http://insideairbnb.com/, which handles the publicly available information of AirBnB site and analyses, cleanses and aggregates them where appropriate to facilitate public discussion. Our analysis is based on the data that cover the months May to June of 2020 for the city of Athens. For the scope of our research, we have used 300000 columns, from which the comments column had 299854 non-null. After our text preprocessing only 207920 reviews are valuable for our analysis. Furthermore, we exploited two datasets which are named reviews and listings in accordance. For the scope of our research, we have used 300000 columns, from which the comments column had 299854 non-null.

After our text pre-processing only 207920 reviews are valuable for our analysis. The reviews dataset has 4018 unique listings and 5 columns; however, the listings dataset has 11314 unique listings. After our text preprocessing only 207920 reviews valuable for our analysis.

This is a logical outcome, considering that not every apartment on AirBnB can have an online review. We must take into consideration that some guests do not post a review and some apartments might not have high guest turnout.

We will explore the reviews dataset to discover, which words have the highest occurrence so we can rule them out in our analysis, by reason of not providing value. Some of the words that arise frequently are apartment, place, great, stay, Athens, location, host, nice and many more as depicted on the graph below. All these words have been present more than 33000 times in our review's dataset. It makes perfect sense that these words have high appearance, since we are literally investigating online reviews for rental apartments in the city of Athens.

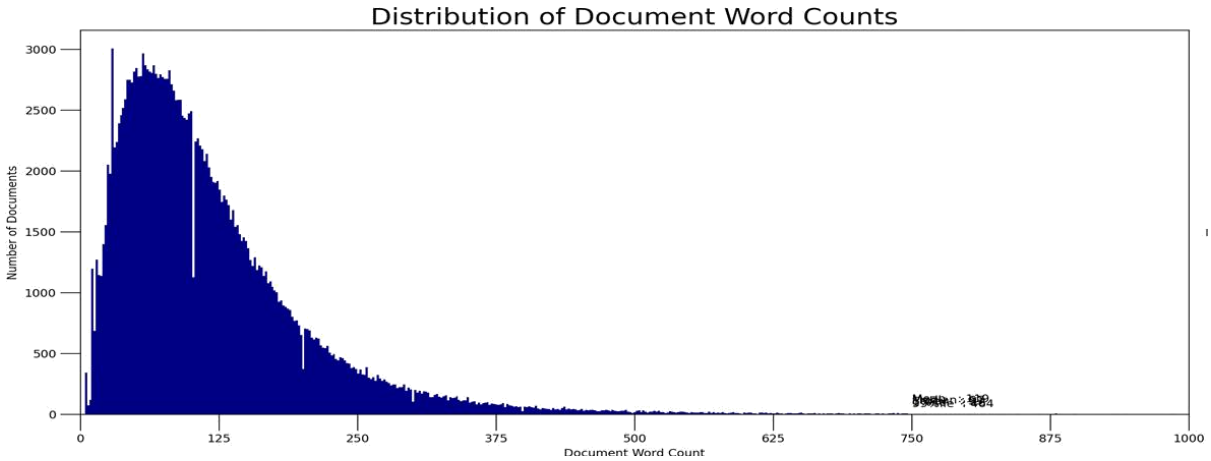


Figure 4.1: Distribution of words.

To boost our comprehension of the review's dataset, we should explore how many words are contained in each review and plot a frequency distribution histogram. We will obtain enlightening results; therefore, we can exclude comments that have extremely high volume of words and might upset our research.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ UNIVERSITY OF ECONOMICS & BUSINESS

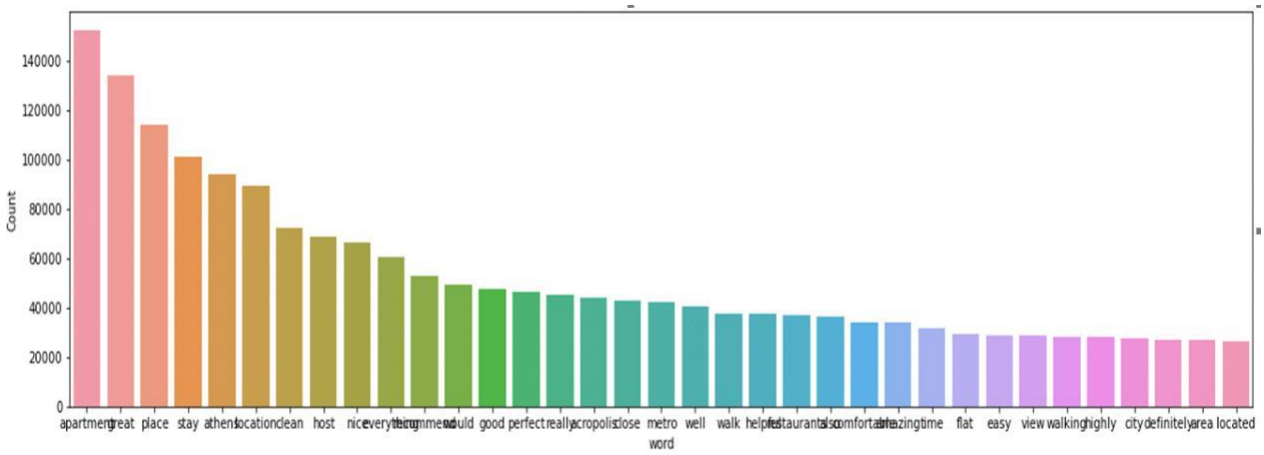


Figure 4.2: Distribution of document word counts

We can notice that the highest volume of reviews has an average of 100 words. A small amount of reviews has more than 250 words and we can observe that no reviews have more than 750 words. This is natural, because AirBnB reviews are limited to 1000 words. In our analysis, we are going to exploit the comments that have a word count between 25 and 280 after exploring the dataset in depth.

4.3 Online Reviews Preprocessing

Because text is unstructured data, we need to utilize special methods to transform the input in a flat form since statistical models often require structured, numerical inputs. Hence, building a pipeline that reads CSV files, extracts all relevant and meaningful words, and then transforms them in a ready for analysis state, was of great importance. We performed the following steps:

- We exploited Python's built-in Regular Expressions library to detect and remove undesirable words or patterns from our files. Certain words are considered undesirable due to the fact that they do not add any value to our analysis; specifically, emails, host names, locations and words with less than 3 letters that are frequently encountered in AirBnB reviews. The resulting regular expressions can be found below:
    - (ˆa zA Z#)";" ": This regex expression states that match the text string for any alphabets from small a to small z or capital A to capital Z, so it basically removes ever character that is not required for the purpose of our analysis.
    - (ˆnx00 n x7F)+ ", " ": This expression is used to replace non-ASCII characters with a single space.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ATHENS UNIVERSITY OF ECONOMICS & BUSINESS

- The Natural Language Toolkit (NLTK) library (NLTK,2020) was an integral part of the data preprocessing task. The built-in directory of English stop-words that the NLTK library has, was essential for the detection and removal of valueless words. To be more specific, words like this, that, these, and, but and many more do not add to this study.
- Except from unwanted stop-words list, another similar list was created containing common Greek names, well-known locations as well as any other name detected and was removed from the input.
- Another helpful library used was spaCy (spaCy,2017). SpaCy is a specialized library in the domain of text analytics frequently used for Named Entity Recognition tasks. Its beneficial properties were leveraged to identify dates and geopolitical areas such as cities and countries that are often present in online reviews. However, it did not perform on the expected level, because of the variety and complexity of Greek first names.
- Lemmatization is a method that transforms any kind of a word to its base root mode. This method is important to associate various inflected forms of words into the root form. Lemmatization was chosen over stemming because it returns a word that has a dictionary meaning, it has better accuracy compared to stemming and it is dictionary-based not rule-based as stemming. However, it has the disadvantage that it is slower in comparison to stemming. We applied lemmatization twice in our study, so that we get the expected results which will enhance our research.
- The NLTK library was further used to tokenize the given input texts. Tokenization is the process of splitting the entire text into separate string objects. Furthermore, all words were turned to lowercase, and one last filtering was performed. Words that contained only alphabetical characters were accepted for the subsequent analysis. To make it clear, tokenizing a document equals to creating a list of strings, where each string is an online review.
- Moreover, we applied Part-Of-Speech tagging using NLTK library. Part-of-speech (POS) tagging is a popular Natural Language Processing process which refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the defi-nition of the word and its context. For our study we only reserved nouns and adjectives.

- Finally, we performed language detection. The idea behind language detection is based on the detection of the character among the expression and words in the text. The main principle is to detect commonly used words in English. Python provides various modules for language detection, but we preferred langdetect as it provided the best results.

The outcome of data preparation process is our corpus, a list of lists, where each sub-list corresponds to a document and contains the acceptable tokenized words.

4.4 Detection of the optimal number of topics

We know probabilistic topic models, such as LDA, are popular tools for text analysis, providing both a predictive and latent topic representation of the corpus. However, there is a longstanding assumption that the latent space discovered by these models is generally meaningful and useful, and that evaluating such assumptions is challenging due to its unsupervised training process. Besides, there is a no-gold standard list of topics to compare against every corpus. It is of major importance to find the optimal number of topics, because then the extracted topics will provide worthy insights.

When determining how many topics to use, it is important to consider both qualitative and quantitative factors. Qualitatively, we should have domain knowledge of the data we are analyzing and be able to gauge a general ballpark of clusters our data will separate into. There should be enough topics to be able to distinguish between overarching themes in the text but not so many topics that they lose their interpretability. From a quantitative perspective, some data practitioners use perplexity or predictive likelihood to help determine the optimal number of topics and then evaluate the model fit. Perplexity is calculated by taking the log likelihood of unseen text documents given the topics defined by a topic model. A good model will have a high likelihood and as a result low perplexity. Usually, we would plot these measures over a spectrum of topics and choose the topic that best optimizes for our measure of choice. But sometimes these metrics are not correlated with human interpretability of the model, which can be impractical in a business setting.It must be noted that the nature of the LDA is subjective. It is possible that different people may reach different conclusions about choosing the most meaningful topic groups. We are looking for the most reasonable topic groups. Therefore, we will obtain different numbers of groups.

Then, we examine and compare topic modeling's, and decide which topic model makes more sense, most meaningful, have the clearest distinction within the model. Then, the group (model) that makes the most sense will be chosen among all topic groups.

To obtain the optimal model, we will utilize topic coherence as our evaluation metric. Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference. Especially, we will apply C_V as our topic coherence measure. C_V measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity.

We built an algorithm that evaluates every topic using the C_V measure. The algorithm starts computing the coherence for 2 topic models and it loops over until it reaches the point where the model has 40 topics. Furthermore, we have set the algorithm to have a step of 2 for better computing performance. When the loop is over, we are provided with a graph that will assist us in the choice of the optimal number of topics. This procedure will be accomplished with the elbow method. The following image shows us the C_V coherence for the topic models.

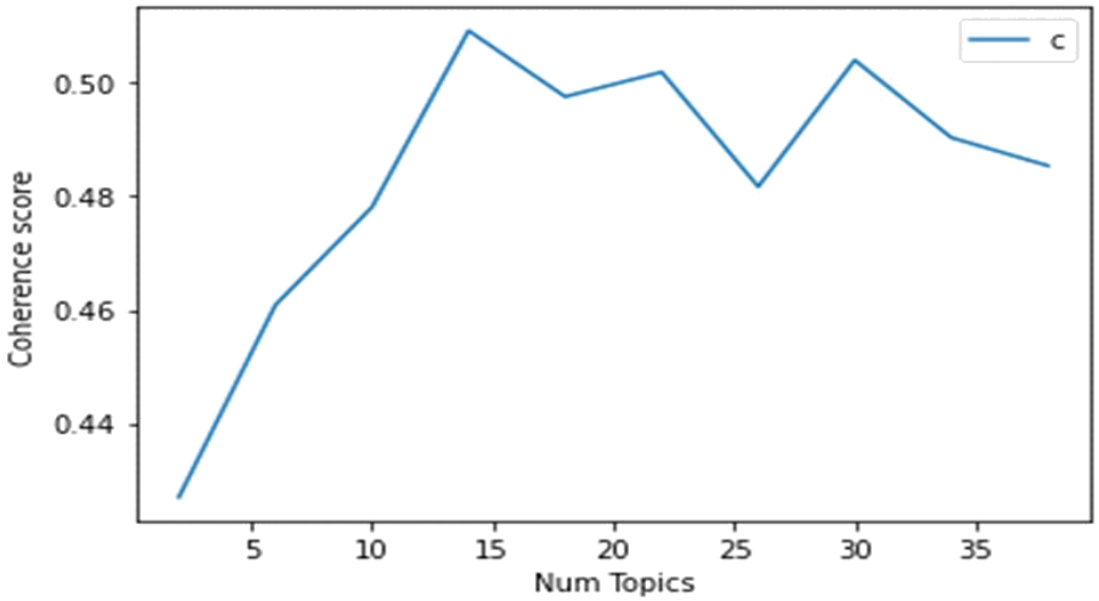


Figure 4.3: Coherence Score Per Number Of Topics.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

The elbow curve is a common heuristic that is used to determine the number of clusters, in our case topics. In mathematical optimization elbow curve is the optimization and the selection of the point where diminishing returns are no longer worth the additional cost. The intuition is that increasing the number of topics will naturally improve the fit (explain more of the variation), since there are more parameters (more topics) to exploit, but at some point this is over-fitting, and the elbow reflects this. In our case, we cannot detect a sharp elbow, however it is obvious that the optimal number of topics are 14. Below there are some exact results of coherence scores for certain topic models.

Number of Topics	Coherence Value
2	0.374
5	0.45
8	0.512
11	0.469
14	0.521
17	0.518
20	0.509
23	0.511
26	0.499
29	0.492
32	0.504

Table 4.1: Coherence Scores.

To be convinced for our finding, another method for obtaining the optimal number of topics will be deployed. This method is consisted of the creation of LDA models across different topic numbers and the check of the Jaccard similarity and coherence for each. Jaccard similarity is a statistic used for comparing the similarity and diversity of sample sets and it is explained by the following equation: J(A,B) = (A \B)=(A [B). This algorithm will build up to 30 topic models and will compare them in pairs. This technique has been found to have issues in practical applications, since it is time-consuming. Nevertheless, this technique is quite accurate because it will maximize the coherence score and minimize the topic overlap based on Jaccard similarity. This will contribute a strong intuition for the optimal number of topics. The following image depicts our findings with this method.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

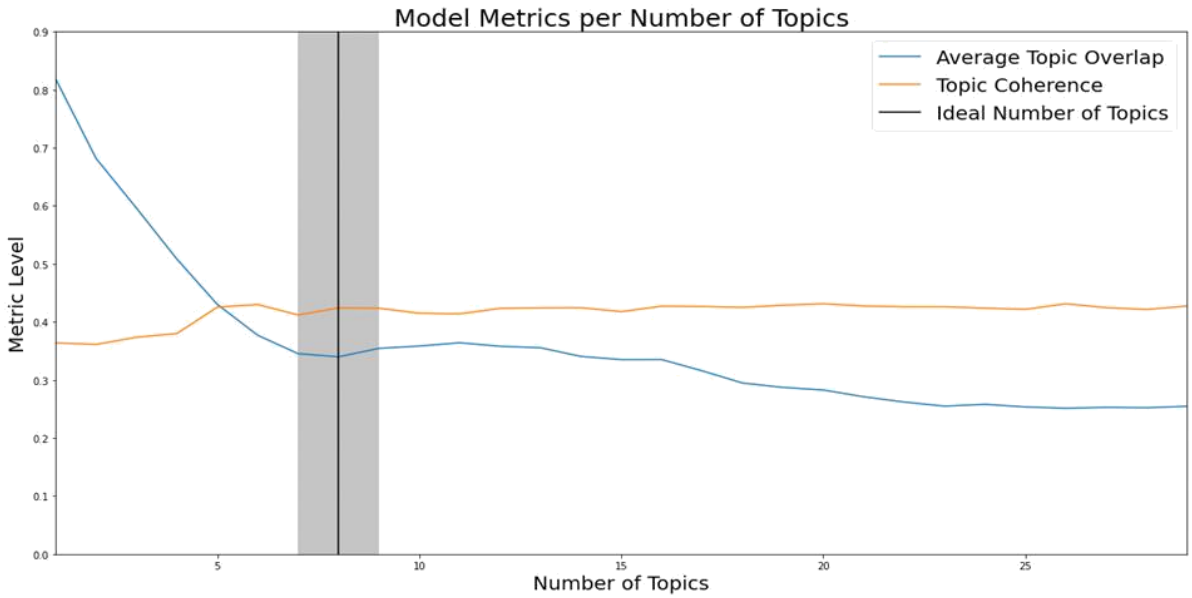


Figure 4.4: Topic Coherence and Topic Overlap per Number of Topics.

This graph depicts that the optimal number of topics is 8. This occurs as a result of the 2 metrics that were measured in this model. We can observe that average topic overlap decreases as the number of topics increase, but the topic coherence remains almost unaltered. That is the reasoning behind the election of 8 topics, considering that if we pick a higher number of topics, we might have more accuracy, but we won't be able to interpret and understand the individual topics and the relationship between them.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS & BUSINESS

4.5 Construction of LDA Model

We have everything required to train the LDA model. In addition to the corpus and dictionary, we need to provide the number of topics as well. Apart from that, alpha and beta are hyperparameters that affect sparsity of the topics, as already mentioned earlier. We will set alpha and beta to 'auto' for the optimal model. Chunksize controls how many documents are processed at a time in the training algorithm. Increasing chunksize will speed up training, at least if the chunk of documents easily fit into memory. We are going to take advantage of all our documents to build a model that is interpretable and exploitable. Finally, we will set the number of passes to 20, on the grounds that we need to ensure sufficient corpus topic distribution updates and the number of iterations to 100 for the sake of a good number of documents to reach convergence before moving on. Below, there is a visual representation of the extracted topics and their most representing words, that have the higher probability in the specific topic.

5 *Results*

5.1 Analysis of Models

First, we will try to understand and evaluate the topics that are extracted if their number is 14 and then if their number is 8. The topics that make more sense will be chosen to be analyzed and visualized. To be precise, both LDA models will have the same parameters to prevent any distinction.

To begin with, it has come to our attention that some words that should be excluded, like ‘athen, George, Maria’ exist in our analysis. Furthermore, apartment appears multiple times, which makes sense since AirBnB is a platform that rents apartments. Below, there is a visual representation of the extracted topics and their most representing words on a tabular form, that have the higher probability in the specific topic.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

5.2 14-Topic Model

1. (0, ’0.047*"apartment" + 0.034*"view" + 0.032*"area" + 0.026*"place" + ’ ’0.017*"balcony" + 0.015*"athen" + 0.014*"restaurant" + 0.014*"terrace" + ’ ’0.012*"comfortable" + 0.012*"space"’)

2. (1, ’0.052*"apartment" + 0.036*"airport" + 0.035*"place" + 0.023*"athen" + ’ ’0.020*"public" + 0.019*"taxi" + 0.018*"host" + 0.018*"transportation" + ’ ’0.017*"transport" + 0.017*"location"’)

3. (2, ’0.035*"apartment" + 0.027*"restaurant" + 0.023*"shop" + 0.023*"coffee" + ’ ’0.019*"acropolis" + 0.019*"minute" + 0.017*"museum" + 0.016*"walk" + ’ ’0.013*"location" + 0.013*"street"’)

4. (3, ’0.072*"apartment" + 0.052*"location" + 0.048*"host" + 0.040*"place" + ’ ’0.039*"athen" + 0.036*"view" + 0.034*"perfect" + 0.021*"fantastic" + ’ ’0.020*"beautiful" + 0.018*"wonderful"’)

5. ’0.099*"location" + 0.095*"great" + 0.047*"host" + 0.045*"place" + ’ ’0.042*"clean" + 0.038*"apartment" + 0.020*"helpful" + 0.020*"value" + ’ ’0.018*"excellent" + 0.018*"space"’)

6. (5, ’0.068*"place" + 0.065*"apartment" + 0.042*"metro" + 0.036*"clean" + ’ ’0.034*"station" + 0.033*"close" + 0.032*"host" + 0.026*"helpful" + ’ ’0.022*"city" + 0.016*"comfortable"’)

7. (6, ’0.023*"maria" + 0.021*"time" + 0.020*"home" + 0.019*"flat" + 0.019*"place" ’ ’+ 0.018*"athen" + 0.015*"wonderful" + 0.015*"welcome" + 0.014*"greek" + ”0.014*"warm"’)

8. (7, ’0.045*"george" + 0.043*"place" + 0.028*"touch" + 0.027*"apartment" + ’ ’0.022*"little" + 0.021*"nice" + 0.020*"clean" + 0.019*"comfortable" + ’ ’0.018*"night" + 0.014*"location"’)

9. (8, ’0.057*"apartment" + 0.037*"metro" + 0.033*"station" + 0.032*"restaurant" + ”0.031*"minute" + 0.030*"nice" + 0.030*"flat" + 0.023*"clean" + 0.017*"area" ”+ 0.016*"location"’)

10. (9, ’0.045*"place" + 0.043*"night" + 0.027*"flight" + 0.026*"late" + ”0.020*"check" + 0.020*"time" + 0.018*"athen" + 0.017*"early" + 0.015*"hour" ’ ’+ 0.015*"airbnb"’)

11. (10, ’0.071*"apartment" + 0.017*"water" + 0.017*"clean" + 0.016*"flat" + ”0.016*"nice" + 0.012*"time" + 0.011*"problem" + 0.011*"location" + ’ ’0.011*"metro" + 0.010*"night"’)

12. (11, ’0.098*"nice" + 0.072*"place" + 0.047*"clean" + 0.035*"host" + ”0.028*"apartment" + 0.027*"apartment" + 0.026*"center" + 0.024*"perfect" + ’ ’0.024*"athen" + 0.023*"city"’)

13. (12, ’0.101*"house" + 0.030*"good" + 0.026*"host" + 0.025*"place" + 0.024*"home" ”+ 0.024*"time" + 0.020*"experience" + 0.017*"location" + 0.016*"airbnb" + ”0.016*"perfect"’)

14. (13, ’0.026*"apartment" + 0.025*"room" + 0.019*"location" + 0.016*"place" + ”0.016*"small" + 0.016*"bathroom" + 0.012*"street" + 0.011*"night" + ’ ’0.011*"door" +0.011*"bedroom"’)

We are going to try to comprehend and interpret the results by assigning a distinctive name to each topic. This procedure is subjective and depends on the experiences of the author. We will try to have a holistic approach on the interpretation of the topics and appoint them the appropriate labels.

1. For the first topic, we spot words like ‘balcony, terrace and view’. We are going to nominate this topic “Apartment’s View”. This topic is related to the facilities an apartment can offer regarding outdoor spaces, which can provide a scenic view from it.

2. The second topic will be called “Apartment’s proximity to public transport”. This name is suitable, because we observe words such as “airport, transportation and public”. Every visitor pays a lot of attention to this issue, since it will smoothen their vacation and it will save valuable time.

3. The third topic refers to the location of the apartment and its proximity to tourist attractions, so we will name it “Tourist Attractions”. An apartment’s location regarding a restaurant, a shop or a tourist attraction is of vital importance for the choice of a guest.

UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

4. Regarding the fourth topic, we cannot detect something of value and the nomination is kind of complex, so we will go with the name of "Positive reviews", since there are lot of adjectives which describe a remarkable experience.

5. The fifth topic will be called "Excellent Experience" since all words are appraisal for either the host or the place.

6. The sixth topic is "Proximity to Metro Station" and it differs from topic 2, because these guests are searching for this specific feature, which is the distance from the apartment to metro stations.

7. The topic number seven can have the name of "Greek Hospitality", because the most common words are referring to the kind aspects of the host's character.

8. On the topic number eight we will ignore, the name of the host "George" and we will describe it us "Cozy Apartments", since words like "little, nice and comfortable "show up.

9. For topic nine, the name we will elect is "Neighborhood of the Apartment", because it is a general representation of the area that the apartment belongs

10. The next topic, number ten, will be named "Proximity to Airport" and it is obvious that this is the feature guests are searching for and commenting on for their AirBnB stay.

11. The topic eleven might be negative since it contains words as "problem, night, location and time". We will translate these words into a negative topic that will be nominated "Rough Neighborhood". This name might not be exactly representative of the topic; however it provides a comprehensible solution.

12. The topic numbered twelve will have the name of "Apartment in the center of Athens". One of the most basic and valuable attributes that an apartment can have is being in the center of the capital.

13. About topic thirteen, we could not conclude for a distinctive label, so we should just name it "Perfect Reviews", since no negative or neutral words emerge.

14. Finally, topic fourteen is clear that it refers to the areas of the apartment so a suitable label for it would be "Apartment spaces". Guests are discussing about the rooms of an apartment on this topic.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

5.3 8-Topic Model

1. (0,’0.117*"good" + 0.060*"place" + 0.054*"location" + 0.039*"nice" + ”0.037*"clean" + 0.033*"great" + 0.027*"host" + 0.022*"value" + ”0.017*"apartment" + 0.016*"room"’)

2. (1,’0.057*"apartment" + 0.048*"metro" + 0.043*"station" + 0.030*"minute" + ”0.018*"clean" + 0.017*"restaurant" + 0.015*"nice" + 0.015*"close" + ”0.015*"place" + 0.015*"area"’)

3. (2,’0.022*"place" + 0.015*"athen" + 0.014*"home" + 0.013*"apartment" + ”0.012*"coffee" + 0.011*"restaurant" + 0.011*"view" + 0.010*"local" + ”0.010*"wonderful" + 0.010*"host"’)

4. (3, ’0.093*"flat" + 0.043*"place" + 0.029*"clean" + 0.024*"host" + 0.023*"great" ”+ 0.020*"athen" +0.020*"check" + 0.019*"thank" + 0.017*"time" + ”0.017*"helpful"’)

5. (4, ’0.035*"apartment" + 0.018*"room" + 0.017*"night" + 0.015*"great" + ’ ’0.013*"location" + 0.012*"street" + 0.011*"bedroom" + 0.011*"bathroom" +”0.009*"floor" + 0.009*"small"’)

6. (5, ’0.056*"nice" + 0.033*"house" + 0.030*"host" + 0.025*"apartment" + ’ ’0.025*"clean" + 0.025*"apartment" + 0.024*"place" + 0.020*"good" + ”0.017*"time" + 0.014*"center"’)

7. (6, ’0.085*"apartment" + 0.043*"great" + 0.034*"host" + 0.034*"place" + ’ ’0.034*"athen" + 0.032*"location" + 0.025*"clean" + 0.021*"view" + ”0.019*"wonderful" + 0.019*"perfect"’)

8. (7, ’0.120*"great" + 0.059*"location" + 0.052*"place" + 0.038*"apartment" + ’ ’0.035*"host" + 0.029*"restaurant" + 0.022*"clean" + 0.017*"distance" + ’ ’0.017*"perfect" + 0.017*"helpful"’)

We can observe that the 8-topic model is not granular and did not capture a valid, useful and applicable result. Thus, it is not reasonable to waste time to nominate each individual topic, as it is undeniable that the 14-topic model can balance interpretability with other quantitative metrics (C_V) and will help us gauge our model’s performance.

After comparing both topic models, we conclude that we will evaluate the results of the first one. We plan on drilling into those topics and understanding the nuances of each of our guest’s individual requests.

5.4 WordCloud of The Model of 14 Topics

We will build a wordcloud of the most common words that appear in our topics to better understand and realize our discoveries. We can detect great, location, place, and apartment to be resurfacing multiple times since they exist in various topics.

Figure 5.1: Wordcloud of Chosen LDA Model

5.5 Exploratory Data Analysis On 14-Topic Model

In this section, we are going to discover in how many reviews each topic is the dominant one. It is clear, that a comment cannot consist only of one topic, but from a variety of them. For this reason, the dominant topic will be the identifier of a review. In the first place, we are interested in discovering the topic that appears most of the time in our dataset.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Index	Num_Documents	Perc_Documents
8.0	39453	0.1866
10.0	20466	0.0968
4.0	19436	0.0919
3.0	17447	0.0825
12.0	15297	0.0724
6.0	14451	0.0684
13.0	14084	0.0666
11.0	14073	0.0666
0.0	11407	0.054
5.0	10662	0.0504
2.0	10009	0.0473
1.0	8515	0.0403
7.0	8122	0.0384
9.0	7987	0.0378

Figure 5.2: Number of Documents and Their Percentage Per Topic.

The topic that shows up most of the times is the topic number 9, which is called “Neighborhood of the Apartment”. It has almost twice the appearances of topic number 11, which is named “Rough Neighborhood”. We can detect a pattern here that a lot of guests pay attention to the location of the apartment. Topics 11, 4 and 5 have almost the same number of appearances around 20 thousand. Moreover, topics 13, 7, 14 and 12 have similar number of appearances around 15 thousand. In addition, topics 2, 8 and 10 have the less appearances around 8 thousand.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Finally, topics 1, 6 and 3 can be encountered in approximately 10 thousand documents. Topic 10 "Proximity to Airport" is represented less on the comments and this is a conclusive indication that guests are not really interested if the apartment is close to the airport or if it has a connection to it through public transport.

From this analysis, we can draw the conclusion that most of the guests are attracted to a cozy, small, pretty, elegant and comfortable apartments and pay much attention to its surroundings. Furthermore, amenities that are important to guests are if an apartment has nice view, outdoor space and is near to tourist sites. Another finding is that guests consider an apartment's proximity to public transport or to an airport to be an essential requirement in order to book an apartment.

In our current research, we anticipated to encounter one topic referring to cancellations and another referring to CoVID-19. We did not come across any reviews that discussed these two present issues. We can imagine that people have not yet come into terms with this new reality. To boost our findings, we have searched the internet why AirBnB has a few reviews on cancellations and on CoVID-19. hey decided to institute a blanket refund policy offering last-minute cancellations and full refunds. In doing so, it provided relief for customers' apprehensions and convinced hesitant people to proceed with bookings. AirBnB's executives also understood customers' fears about health and safety due to the pandemic, so they introduced "Enhanced Cleaning" procedures and recommended hosts increase the time between guest stays. The changes were not mandatory but hosts who adopted them were given a badge to display on their listings, thus conveying transparency and reassuring concerned customers.

These two changes were of high importance to tackle the problems that have been introduced in the pandemic era and this is the reason why our online reviews do not discuss these issues, since AirBnB was flexible, adaptive and eager to encounter them.

We will explore the distribution of words that exist in each review per dominant topic so that we completely comprehend and retrieve insights from the length of reviews. We can deduct that except topics 13, 14 (no. 12, 13 on the images because of python's zero indexing), the rest of the topics had a similar word count across their documents.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

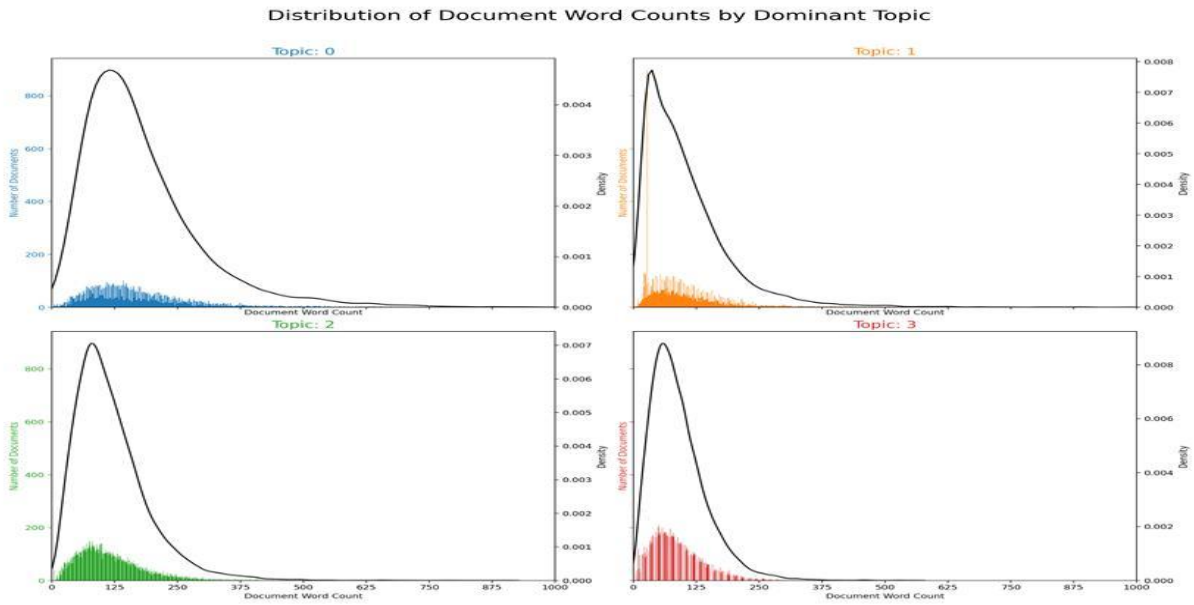


Figure 5.3: Distribution of Document Word Count For Topics 1 to 4.

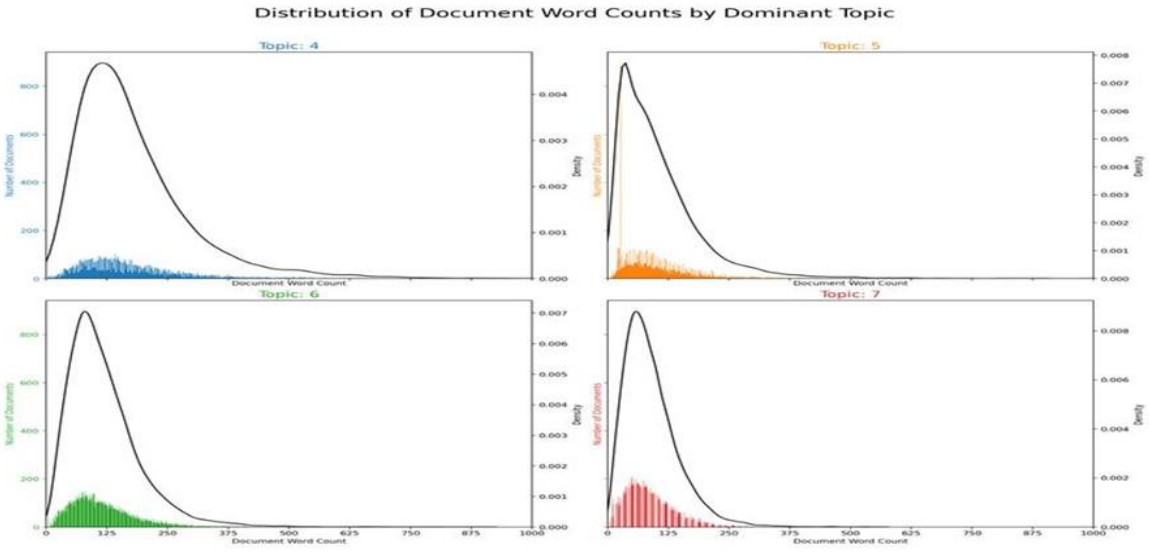


Figure 5.4: Distribution of Document Word Count For Topics 5 to 8.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

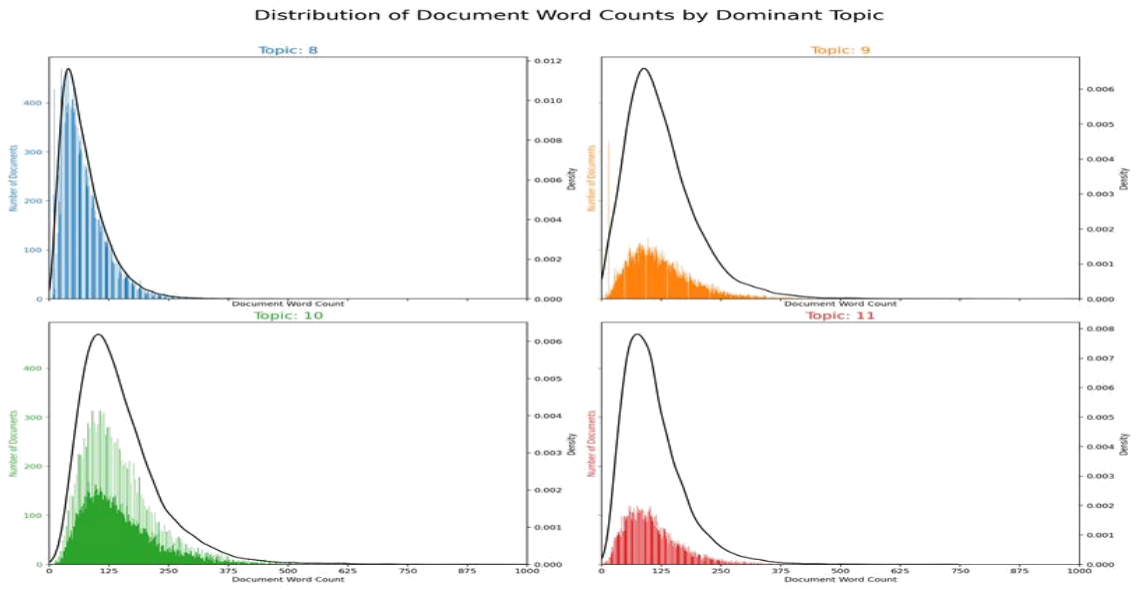
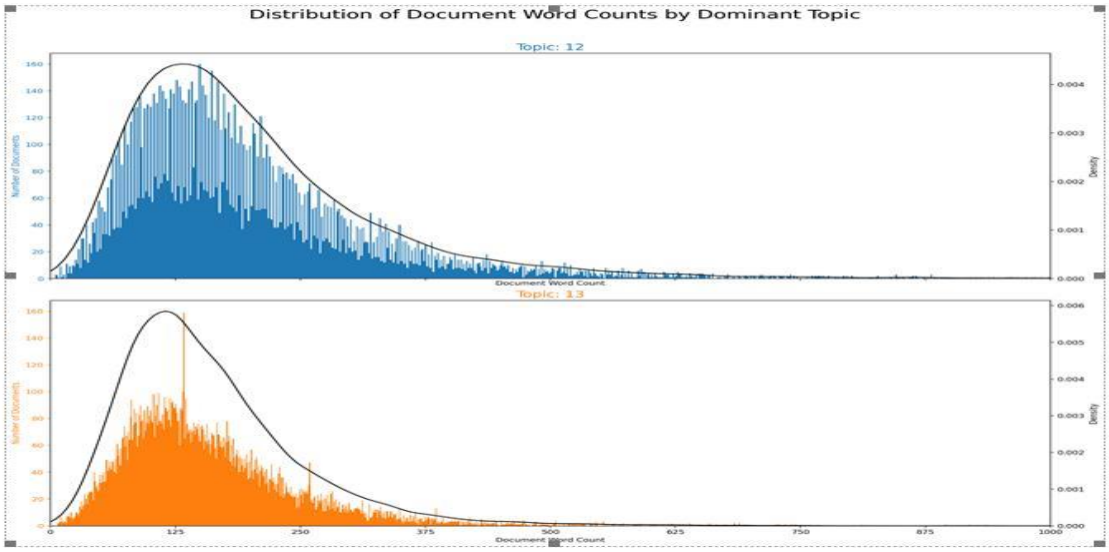


Figure 5.5: Distribution of Document Word Count For Topics 9 to 12.



.

Figure 5.6: Distribution of Document Word Count For Topics 13 and 14.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

After these visualizations, we are going to deploy pyLDAvis, which is a Python library for interactive topic model visualization. pyLDAvis is designed to help users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization. The visualization is intended to be used within an IPython notebook but can also be saved to a stand-alone HTML file for easy sharing.

The area of circle represents the importance of each topic over the entire corpus, the distance between the center of circles indicate the similarity between topics. For each topic, the histogram on the right side listed the top 30 most relevant terms. We are going to visualize topics 1, 8 and 9 because they are the most unique topics that do not overlap with the others. Apart from that, there is some overlapping between specific topics, but generally, the LDA topic model can help us grasp the trend. We can also detect that in the right side of the image the top-30 most salient terms in our documents are the same that have been extracted in our previous exploratory data analysis. This can reassure us that our prior analysis is valid.

Below we have a straightforward and transparent visualization of our extracted topics. We can recognize that our findings are accurate and solid since our topics do not have great overlapping and they differentiate from each other.

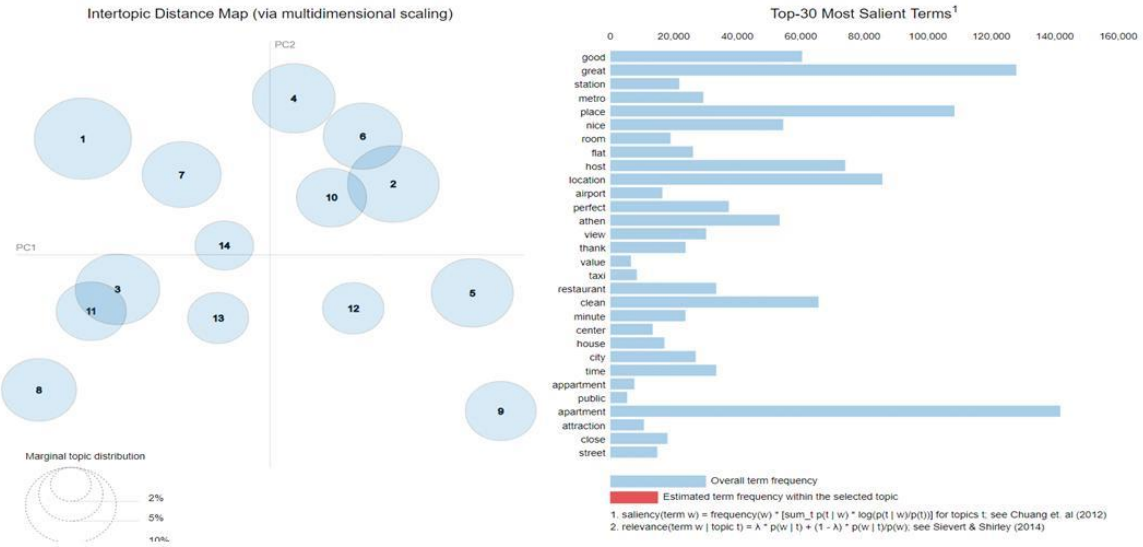


Figure 5.7: Holistic Overview of Topics.

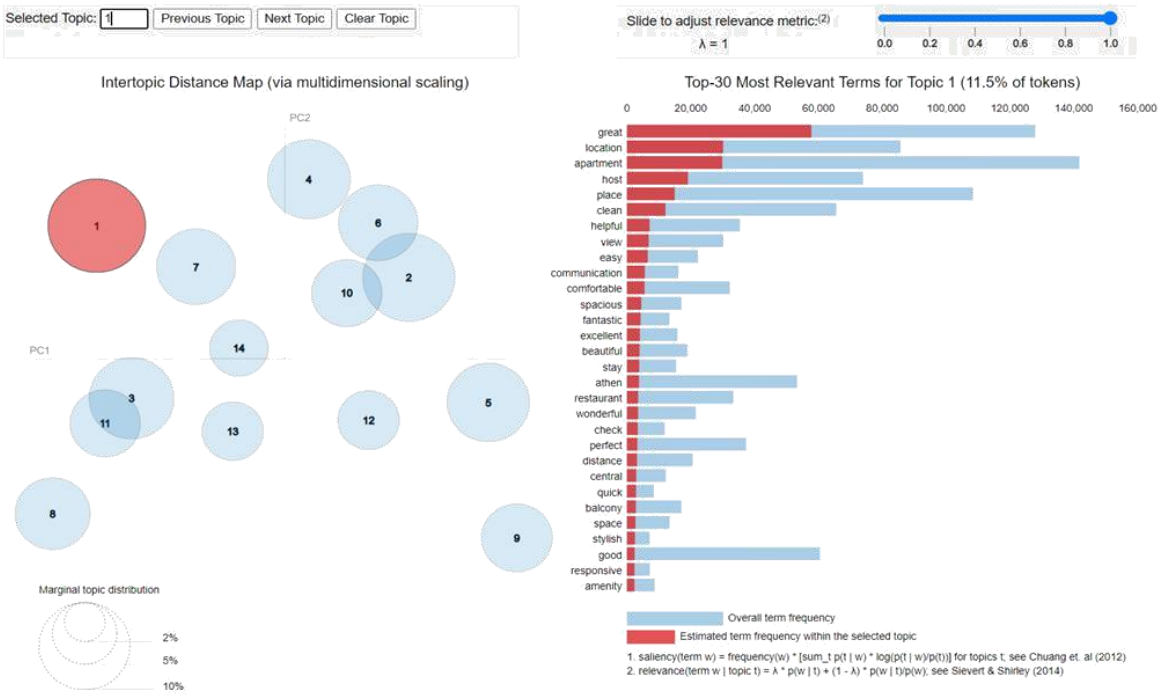


Figure 5.8: Visualization of Topic 1.

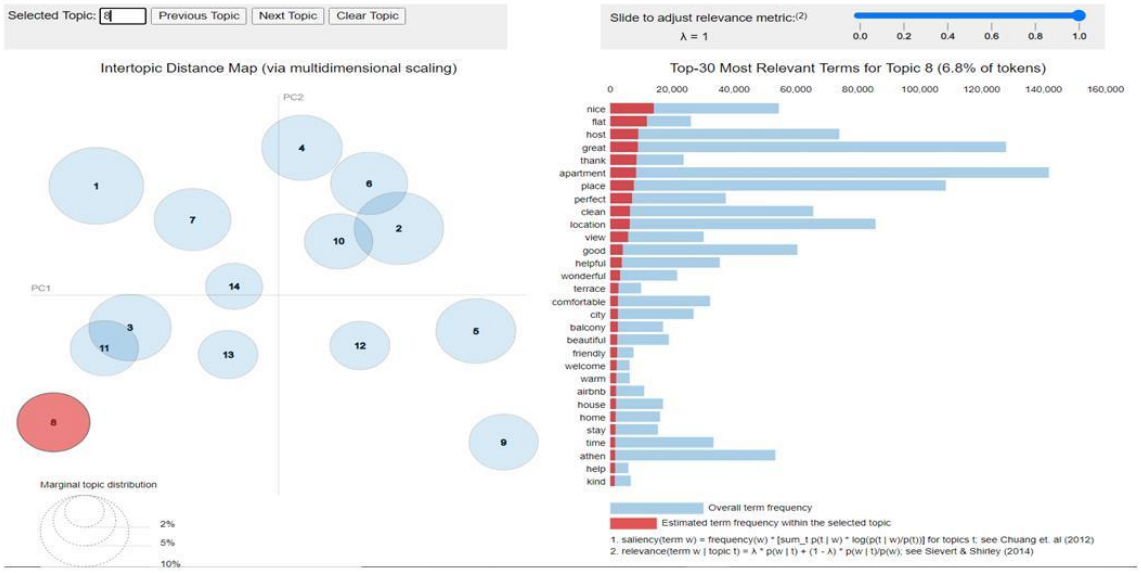


Figure 5.9: Visualization of Topic 8.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

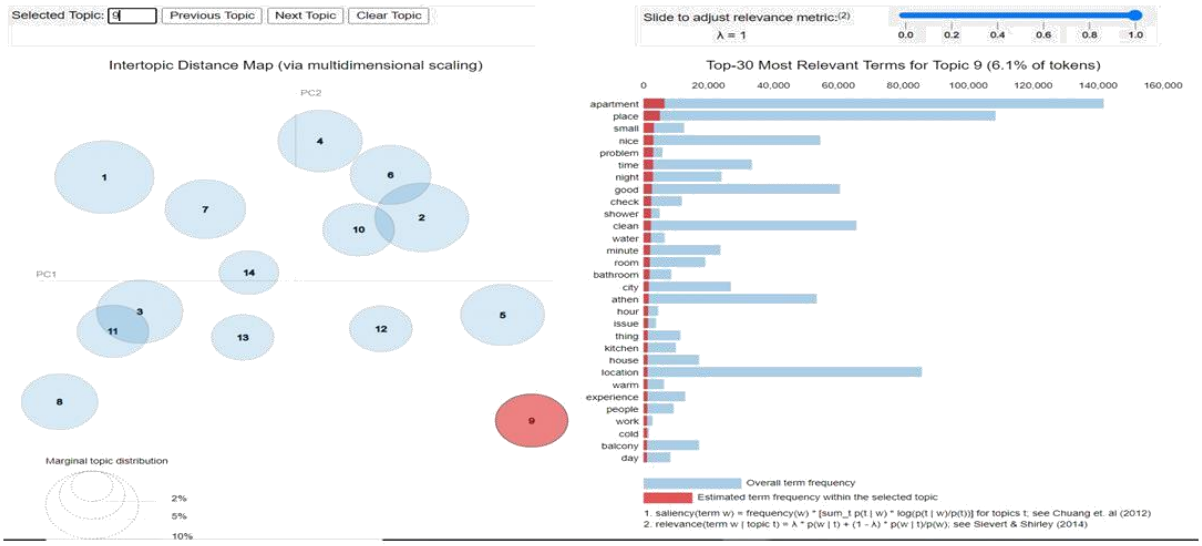


Figure 5.10: Visualization of Topic 9.

5.6 Exploratory Data Analysis On Reviews Scores

First, we should comprehend how the overall rating is calculated (reviews_scores_rating). When guests stay at an Airbnb, they can choose to leave a review and/or rate their stay across seven categories on a scale of 1 to 5 stars. Guests and hosts have 14 days after checkout to write a review. The reviews only show up when both parties submit their feedback or when the 14-day review period has ended, whichever comes first. This means that host reviews written by guests will eventually be made public, regardless of whether a host left a review of their guest. After a host receives at least three reviews, the average of each individual category rating will be displayed on their listings. The overall rating is not calculated using the average of ratings across the other six categories as one might think. AirBnB does prominently display summary statistics for each property, including the total number of reviews it has accumulated and its average rating rounded to the nearest half-star (provided the property has at least three reviews). Furthermore, we should address the fact that reviews and their scores cannot be deleted by hosts, if they are not constructive, but they can only be deleted if the violate AirBnB's policy, thus we can come to the conclusion that the scores and the reviews are impartial. We conducted an analysis on overall reviews rating and we can clearly detect that a huge proportion of the ratings are higher than 90.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
UNIVERSITY OF ECONOMICS & BUSINESS

Other than that, we extracted the sample statistics of reviews_scores_rating column.

Metrics Results	
mean	95.328980
std	7.039104
min	20.000000
25%	94.000000
50%	97.000000
75%	100.000000
max	100.000000

Table 5.1: Summary Statistics of Review_Scores.

We undoubtedly witness that the scores are extremely high, considering that the mean of all scores amounts to 95 and the lowest score is 20. This is not uncommon and can be easily clarified from simple facts/reasons. First, AirBnB is a distinctively human experience since we show up to a new place and meet an incredible host. Thus, we feel obligated to comment about the host's contribution as a fellow human and not pay attention to any minor issues. Additionally, usually uncomfortable guests do not say anything about their experience. It's less stressful to smudge the truth and say all was fine because you can avoid any sense of confrontation. Finally, inflated AirBnB star ratings come down to this most basic human instinct. Everyone wants to fit in and feel valued. As guests, we want to be compassionate and recognize all the time and effort and scrutiny our host has sunk into their beloved space. Below we have visualized the reviews_scores_rating of all listings.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

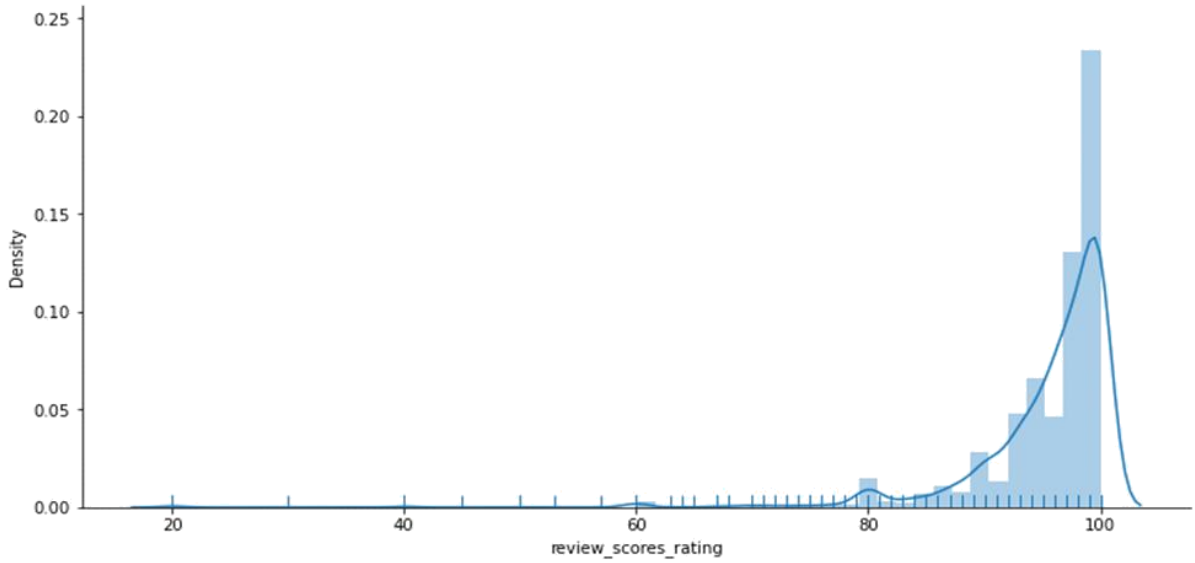


Figure 5.11: Rating of Reviews Scores.

5.7 Exploration of relationship between Reviews_Scores And Topics

In this section of our study, we are going to perform a simple linear regression between the columns dominant_topics, which is the independent variable and reviews_scores, which is the dependent one. Simple linear regression is a technique that we can use to understand the relationship between a single explanatory variable and a single response variable. This technique finds a line that best "fits" the data and takes on the following form:

- $\hat{y} = b_0 + b_1 x$
- $\hat{y}$: The estimated response value
- b_0: The intercept of the regression line
- b_1: The slope of the regression line

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

This equation can help us understand the relationship between the explanatory and response variable, and (assuming it's statistically significant) it can be used to predict the value of a response variable given the value of the explanatory variable. The idea of Simple Linear Regression is finding those parameters a and b for which the error term is minimized. We will deploy the OLS () function from the statsmodels library to fit the regression model. Statsmodels is a Python package for the estimation of many different statistical models, as well as for conducting statistical tests and statistical data exploration. To be more precise, the model will minimize the squared errors: indeed, we do not want our positive errors to be compensated by the negative ones, since they are equally penalizing for our model.

$$\hat{\alpha} = \min_{\alpha} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \min_{\alpha} \sum_{i=1}^{n} \varepsilon_i^2$$

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2$$

Figure 5.12: Least Squares Method.

This procedure is called Ordinary Least Squared error — OLS. Even though OLS is not the only optimization strategy, it is the most popular for this kind of tasks, since the outputs of the regression (that are, coefficients) are unbiased estimators of the real values of alpha and beta. Indeed, according to the Gauss-Markov Theorem, under some assumptions of the linear regression model (linearity in parameters, random sampling of observations, conditional mean equal to zero, absence of multicollinearity, homoskedasticity of errors), the OLS estimators a and b are the Best Linear Unbiased Estimators (BLUE) of the real values of a and b .

We can proceed to fit a simple linear regression model using dominant_topics as the explanatory variable and reviews_scores_rating as the response variable:

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

OLS Regression Results

Dep. Variable:	reviews_scores_rating	R-squared:	0.818
Model:	OLS	Adj. R-squared:	0.898
Method:	Least Squares	F-statistic:	63.91
Date:	Mon, 14 Jun 2021	Prob (F-statistic):	2.25e-06
Time:	15:51:45	Log-Likelihood:	-39.594
No. Observations:	11004	AIC:	83.19
Df Residuals:	13	BIC:	84.60
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	65.3340	2.106	31.023	0.000	60.784	69.884
dominant_topics	1.9824	0.248	7.995	0.000	1.447	2.518

Omnibus:	4.351	Durbin-Watson:	1.677
Prob (Omnibus):	0.114	Jarque-Bera (JB):	1.329
Skew:	0.092	Prob (JB):	0.515
Kurtosis:	1.554	Cond. No.	19.2

Figure 5.13: OLS Results.

We will only interpret two values of the model summary:

- P>|t|: This is the p-value associated with the model coefficients. Since the p-value for dominant_topics (0.000) is significantly less than .05, we can say that there is a statistically significant association between dominant_topics and reviews_scores_rating.
- R-squared: R-squared is possibly the most important measurement produced by this summary. R-squared is the measurement of how much of the independent variable is explained by changes in our dependent variables. This number tells us the percentage of the variation in the reviews scores can be explained by the dominant topics studied. In general, the larger the R-squared value of a regression model the better the explanatory variables can predict the value of the response variable. In this case, 83.1% of the variation in comments scores can be explained by the dominant topics studied.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

6 Discussion

6.1 Dissertation Findings

Topics are not always interpretable as already discussed, nevertheless there is ongoing research around the concept of topics in the context of LDA. Therefore, the instinct that this method will always produce topics identical to the traditional meaning of the concept is not accurate. When implemented on our online reviews, more than half of the topics were consistently found to be coherent. Even though we have achieved major improvements in topic coherence measured by both human judgement and a coherence metric, it has been straightforward that it provided a negligible impact on the accuracy of the model.

After manually inspecting some of the online reviews, insufficient information on most of the comments appeared. This is the reason that limitations existed while undertaking this task from the beginning. No matter how the parameters are set, matches with Covid-19, cancellation and hygiene related terms cannot be detected since they do not appear often on our reviews. This reason prevents LDA from assigning these words a high probability to topics. This may have had a negative impact in the model's ability to reproduce all the topics that appear on our AirBNB comments. Contextual terms that only occur in the context of the pandemic era could be gathered in one topic by LDA, but the lack of these kind of words has prohibited the model to include this kind of topic. Incorporating metadata is a possible solution that would improve the ability of the model to identify this kind of topics.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Regarding our research objectives, we reached some critical conclusions. The main findings of this dissertation are:

1. Topic modelling with the use of LDA provides topics that can be distinguished and labelled.

2. The prediction performance of topic models depends highly on the hyper-parameters that are set for the LDA model.

3. We did not obtain any indication that CoVID-19 has affected the AirBnB business plan.

4. Our extracted topics can be correlated with the reviews scores that are provided by the guests.

5. Preprocessing is a task that requires a lot of knowledge and a great deal of effort to be fruitful.

6. The interpretation of the extracted models is burdensome and subjective.

7. Gensim is a library that is noteworthy and of paramount important to anyone who is going to perform topic modelling using Python.

8. LDA is an established approach which has taken a prominent position in the landscape of text mining algorithms used in content matching and semantic classification of documents.

9. AirBnB has adopted a successful business plan to overcome the challenges of the pandemic.

10. The extracted topics demonstrated the guests interest on the apartment's proximity to public transport and on the apartment's location. These two were the mostly discussed aspects of an apartment.

11. Another crucial finding is that when a host is accessible, cheerful and kind-hearted, then there is a high possibility of getting a positive review.

12. Guests generally rate hosts with high scores, since they get to know them and for that reason they are not completely objective.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

6.2 Limitations

Common LDA limitations:

- Fixed K (The number of topics in the dataset are specified by the user (or based on some distribution (Poisson) by sampling) which is subjective and does not always highlight the true distribution of topics).
- Uncorrelated topics (LDA is unable to depict correlations which led to occurrence of uncorrelated topics).
- Non-hierarchical (in data-limited regimes hierarchical models allow sharing of data).
- Static (There is no development of topics over time)
- Bag of words (LDA assumes words are exchangeable, sentence structure is not modeled).
- Unsupervised (sometimes weak supervision is desirable, e.g. in sentiment analysis).
- Number of documents (it is practically unthinkable to guarantee identification of topics from a small number of documents, no matter how long.).
- Length of documents (LDA is expected to perform poorly if the documents are either too long or too short).
- Complex Structure of Topics (The topics are predicted based on the multinomial distribution and then the words are predicted based on another multinomial distribution trained specific to that topic. If the true structure is more complex than a multinomial distribution or if the data to train isn't enough, then it might underfit.).
- Topic Models are sensitive to the input data small changes (the stemming/tokenization algorithms can result in completely different topics).
- Manual categorization of topics (the user must arbitrarily identify the topics according to his/her own judgement).

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

- Limited computing power (The number of experiment online reviews could not reach over 300000. Even with this volume, the preprocessing and training time exceeded 16 hours, which is a time and effort consuming project).

- The LDA topic model is only working accurately in English reviews and this is the reason why we ruled out all non-English comments.

- The images and links in the documents are ignored in the preprocessing step. If combined with other methods to process all of the information, the results could be better.

A number of these limitations have been addressed in papers that followed the original LDA work. Despite its limitations, LDA is central to topic modeling and has really revolutionized the field.

6.3 Future Work

The field of topic modelling is so large that this thesis has only touched upon its surface. There are several other areas of topic modelling to be explored. Other opportunities of further research are abundant. Sentiment analysis techniques could be applied in parallel to topics to determine if a topic has negative, neutral, or positive impact to a certain review. Additionally, better filtering methods could be investigated to try to filter out the noise while preserving valuable information. Furthermore, innovative methods that allow for less subjective approach of the evaluation of topics would also immensely enhance the efficiency of these algorithms. From the perspective of this research, upgrading the computing power and enlarging the volume of comments with using datasets from multiple periods can be of great value. Finally, the application of data mining techniques to social network is of major importance to computer science, social science and business.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

7 *Bibliography*

[1] Bulygin, Denis and Musabirov, Ilya, How People Reflect On The Usage Of Cosmetic Virtual Goods: A Structural Topic Modeling Analysis Of R/Dota2 Discussions (February 20, 2020). Higher School of Economics Research Paper No. WP BRP 60/MAN/2020

[2] Bakar, N.A.; Rosbi, S. Effect of Coronavirus disease (COVID-19) to tourism industry. Int. J. Adv. Eng. Res. Sci. 2020, 7, 4.

[3] Korfiatis, Nikos & Stamolampros, Panagiotis & Kourouthanassis, Panos & Sagiadinos, Vasileios. (2018). Measuring service quality from unstructured data: A topic modeling appli-cation on airline passengers' online reviews. Expert Systems with Applications. Fortcoming.

[4] Blei, D. M., & Lafferty, J. D. (2006). Correlated Topic Models. In In Proceedings of the 23rd International Conference on Machine Learning (pp. 113–120). MIT Press.

[5] Blei, D. M., & Lafferty, J. D. (2007). A Correlated Topic Model of Science. The Annals of Applied Statistics, 1(1), 17–35

[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(Jan), 993–1022.

[7] Korfiatis, N., & Poulos, M. (2013). Using online consumer reviews as a source for demographic recommendations: A case study using online travel reviews. Expert Systems with Applications, 40(14), 5507–5515.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

[8] Kuhn, K. D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. Transportation Research Part C: Emerging Technologies, 87, 105–122.

[10] LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. MIT Sloan Management Review, 52(2), 21

[11] Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. Tourism Management, 68, 301–323

[12] Rossetti, M., Stella, F., & Zanker, M. (2016). Analyzing user reviews in tourism with topic models. Information Technology & Tourism, 16(1), 5–21.

[13] Sotiriadis, M. D., & Van Zyl, C. (2013). Electronic word-of-mouth and online reviews in tourism services: the use of twitter by tourists. Electronic Commerce Research, 13(1), 103–124

[14] Stamolampros, P., & Korfiatis, N. (2018). Exploring the Behavioral Drivers of Review Valence: The Direct and Indirect Effects of Multiple Psychological Distances. International Journal of Contemporary Hospitality Management, 30(8), Forthcoming.

[15] Wang, C., & Blei, D. M. (2013). Variational inference in nonconjugate models. Journal of Machine Learning Research, 14(Apr), 1005–1031

[16] Christine Pitt, Kirk Plangger & Theresa Eriksson (2021) Accommodation eWOM in the sharing economy: automated text comparisons from a large sample, Journal of Hospitality Marketing & Management, 30:2, 258-275

[17] Yen, C. L. A., & Tang, C. H. H. (2015). Hotel attribute performance, eWOM motivations, and media choice. International Journal of Hospitality Management, 46, 79-88.

[18] Zervas, G., Proserpio, D., & Byers, J. W. (2017). The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. Journal of Marketing Research, 54(5), 687-705.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

[19] Cheng-Hao Chen, Bang Nguyen, Philipp “Phil” Klaus & Meng-Shan Wu (2015) Exploring Electronic Word-of-Mouth (eWOM) in The Consumer Purchase Decision-Making Process: The Case of Online Holidays – Evidence from United Kingdom (UK) Consumers, Journal of Travel & Tourism Marketing, 32:8, 953-970

[20] R. de Groof and H. Xu, "Automatic topic discovery of online hospital reviews using an im-proved LDA with Variational Gibbs Sampling," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 4022-4029..

[20] Jaffe, S., Coles, P., Levitt, S., & Popov, I. (2017). Quality Externalities on Platforms: The Case of Airbnb .

[21] Carina Jacobi, Wouter van Atteveldt & Kasper Welbers (2015): Quantitative analysis of large amounts of journalistic texts using topic modelling, Digital Journalism

[22] IR Putri and R Kusumaningrum 2017 J. Phys.: Conf. Ser. 801 012073

[23] Dandibhotla, Teja Santosh & Babu, K. & Prasad, S.D.V. & Vivekananda, A.. (2016). Opinion Mining of Online Product Reviews from Traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet. International Journal of Education and Management Engineering. 6. 34-44.

[24] Calheiros, A. C., Moro, S. & Rita, P. (2017). Sentiment classification of consumer generated online reviews using topic modeling. Journal of Hospitality Marketing and Management. 26

(7), 675-693

[25] Browning, V. So, K. & Sparks, B. A. (2013). The Influence of Online Reviews on Consumers’ Attributions of Service Quality and Control for Service Standards in Hotels, Journal of Travel & Tourism Marketing, 30 (1-2) 23-40

[26] Tirunillai S, Tellis GJ. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. Journal of Marketing Research. 2014;51(4):463-479

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

[27] He L, Han D, Zhou X, Qu Z. The Voice of Drug Consumers: Online Textual Review Analysis Using Structural Topic Model. Int J Environ Res Public Health. 2020 May 22;17(10):3648

[28] Cheng, Xin & Shuai, Chuanmin & Liu, Jiali & Wang, Jing & Liu, Yue & Li, Wenjing & Shuai, Jing. (2018). Topic modelling of ecology, environment and poverty nexus: An integrated framework. Agriculture Ecosystems & Environment. 267. 1-14.

[30] Murphy, L., Mascardo, G., & Benckendorff, P. (2007). Exploring word-of-mouth influences on travel decisions: Friends and relatives vs other travelers. International Journal of Consumer Studies, 31,517-527

[31] Thorsten Hennig-Thurau, Kevin P. Gwinner, Gianfranco Walsh, Dwayne D. Gremler, Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?, Journal of Interactive Marketing, Volume 18, Issue 1, 2004, Pages 38-52, ISSN 1094-9968

[32] Beverley A. Sparks, Victoria Browning, The impact of online reviews on hotel booking intentions and perception of trust, Tourism Management, Volume 32, Issue 6, 2011, Pages 1310-1323, ISSN 0261-5177

[33] Eric W.T. Ngai, Spencer S.C. Tao, Karen K.L. Moon, Social media research: Theories, con-structs, and conceptual frameworks, International Journal of Information Management, Volume 35, Issue 1, 2015, Pages 33-44, ISSN 0268-4012

[34] Gretzel, U. and Yoo, K.H. (2008) Use and Impact of Online Travel Reviews. In: Information and Communication Technologies in Tourism 2008, Springer, Vienna, 35-46

[35] Phillips, Paul & Barnes, Stuart & Zigan, Krystin & Schegg, Roland. (2016). Understanding the Impact of Online Reviews on Hotel Performance: An Empirical Analysis. Journal of Travel Research. 56. 10.1177/0047287516636481.

[36] Duan, W.J., Gu, B. and Whinston, A.B. (2008) The Dynamics of Online Word-of-Mouth and Product Sales—An Empirical Investigation of the Movie Industry. Journal of Retailing, 84, 233-242.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

[37] Prendergast, G.P., Tsang, A.S.L. and Cheng, R. (2014) Predicting Handbill Avoidance in Hong Kong and the UK. European Journal of Marketing, 48, 132-146

[38] Xie, K. L., Zhang, Z., & Zhang, Z. (2014). The Business Value of Online Consumer Re-views and Management Response to Hotel Performance. International Journal of Hospitality Management, 43, 1-12

[39] Tellis, G.J. and Wernerfelt, B. (1987) Competitive Price and Quality under Asymmetric Information. Marketing Science, 6, 240-253.

[39] Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, Gresham D, Boer VM, Troyanskaya OG, Botstein D. Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. Mol Biol Cell. 2008 Jan;19(1):352-67.

[40] Gopalan, A., et al. (2016) Health-Specific Information and Communication Technology Use and Its Relationship to Obesity in High-Poverty, Urban Communities: Analysis of a Population-Based Biosocial Survey. Journal of Medical Internet Research, 18, e182.

[41] Linstead, Erik & Bajracharya, Sushil & Ngo, Trung & Rigor, Paul & Lopes, Cristina & Baldi, Pierre. (2009). Sourcerer: Mining and searching internet-scale software repositories. Data Min. Knowl. Discov.. 18. 300-336.

[42] Caluza, L.J., Verecio, R.L., Funcion, D.G., Quisumbin, L.A., Gotardo, M.A., Laurente, M., Cinco, J.C., & Marmita, V.A. (2017). An Assessment of ICT Competencies of Public School Teachers: Basis for Community Extension Program. IOSR Journal of Humanities and Social Science, 22, 01-13.

[43] Latent Dirichlet Allocation,

https://radimrehurek.com/gensim/models/ldamodel.html

[44] Gensim - Documents & LDA Model,

https://www.tutorialspoint.com/gensim/gensim_documents_and_lda_model.htm

[45] Topic Modeling with Gensim (Python),

https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

[46] Topic Modelling in Python with NLTK and Gensim,

https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and -gensim-4ef03213cd1

[48] Part 3: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn,

https://www.analyticsvidhya.com/blog/2021/06/part-3-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn

[49] Evaluate Topic Models: Latent Dirichlet Allocation (LDA),

https://towardsdatascience.com/evaluate-topic-model-in-python-latent-d irichlet-allocation-lda-7d57484bb5d0

[50] Part 2: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn,

https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and -latent-dirichlet-allocation-lda-using-gensim-and-sklearn

[51] NLP Gensim Tutorial – Complete Guide For Beginners,

https://www.geeksforgeeks.org/nlp-gensim-tutorial-complete-guide-for-b eginners

[52] Topic modeling visualization – How to present the results of LDA models?,

https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how -to-present-results-lda-models

[53] pyLDAvis: Topic Modelling Exploration Tool That Every NLP Data Scientist Should Know, https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that -every-nlp-data-scientist-should-know

[54] Hofmann, Thomas. (2013). Probabilistic Latent Semantic Analysis

[55] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. Knowledge discovery and data mining: towards a unifying framework. In Proceedings of the

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 82–88.

[56] Dreier, Jan & Kuinke, Philipp & Przybylski, Rafael & Reidl, Felix & Rossmanith, Peter & Sikdar, Somnath. (2014). Overlapping Communities in Social Networks.

[57] Yang, Y. (2015). Application of Latent Dirichlet Allocation in Online Content Generation. UCLA.

[58] E. Linstead, C. Lopes and P. Baldi, "An Application of Latent Dirichlet Allocation to Analyz-ing Software Evolution," 2008 Seventh International Conference on Machine Learning and Applications, 2008, pp. 813-818

[59] Verecio, Rommel. (2017). Applications of latent dirichlet allocation algorithm of published articles on cyberbullying. International Journal of Applied Engineering Research. 12. 10878-10884

[60] Lang, C., Siemens, G., Wise, A., & Gaševic,´ D. (Eds.) (2017). Handbook of Learning Analytics. (1st ed.) Society for Learning Analytics Research

[61] Hemmington, Nigel, and Lindsay Neill. "Hospitality Business Longevity under COVID-19: The Impact of COVID-19 on New Zealand's Hospitality Industry." Tourism and Hospitality Research, (February 2021).

[62] Airbnb: Advantages and Disadvantages

https://www.investopedia.com/articles/personal-finance/032814/pros-and-consusing-airbnb.asp

[63] https://news.airbnb.com/about-us

[64] gensim 4.0.1

https://pypi.org/project/gensm

[65] Natural Language Toolkit

https://www.nltk.org/#natural-language-toolkit

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

[66] Ip C, Lee H (Andy), Law R. Profiling the Users of Travel Websites for Planning and Online Experience Sharing. Journal of Hospitality Tourism Research. 2012;36(3):418-426

[67] Lopez, Eduardo Bulchand-Gidumal, Jacques Taño, Desiderio Armas, Ricardo J. (2011). Intentions to use social media in organizing and taking vacation trips. Computers in Human Behavior. 27. 640-654.

[68] Dellarocas, Chrysanthos Zhang, Xiaoquan Michael Awad, Neveen. (2007). Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures. Journal of Interactive Marketing. 21. 23 - 45.

[69] Floyd, Kristopher Freling, Ryan Alhoqail, Saad Cho, Hyun Young Freling, Traci. (2014). How Online Product Reviews Affect Retail Sales: A Meta-analysis. Journal of Retailing. 90.

[70] Robson, Karen Farshid, Mana Bredican, John Humphrey, Stephen. (2013). Making Sense of Online Consumer Reviews: A Methodology. International Journal of Market Research. 55. 521–537.

[71] Mayzlin, Dina Chevalier, Judith. (2003). The Effect of Word of Mouth on Sales: Online Book Reviews. Journal of Marketing Research. 43.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ