Student Thesis

Second-cycle

Value Creation From User Generated Content for Smart Tourism Destinations

Author: Mustafa Celen & Maximiliano Rojas
Supervisor: Azadeh Sarkheyli
Examiner: Moudud Alam
Subject/main field of study: Business Intelligence
Course code: MI4002
Higher education credits: 15
Date of examination: 2020-06-08

At Dalarna University it is possible to publish the student thesis in full text in DiVA. The publishing is open access, which means the work will be freely accessible to read and download on the internet. This will significantly increase the dissemination and visibility of the student thesis.

Open access is becoming the standard route for spreading scientific and academic information on the internet. Dalarna University recommends that both researchers as well as students publish their work open access.

I give my/we give our consent for full text publishing (freely accessible on the internet, open access):

Yes ☒  No ☐

Dalarna University – SE-791 88 Falun – Phone +4623-77 80 00
Abstract:

This paper aims to show how User Generated Content can create value for Smart Tourism Destinations. Applying the analysis on 5 different cases in the region of Stockholm to derive patterns and opportunities of value creation generated by UGC in tourism. Findings of this paper is also discussed in terms of improving decision making, possibilities of new business models and importance of technological improvements on STD’s. Finally, thoughts on models are presented for researchers and practitioners that might be interested in exploitation of UGC in the context of information-intensive industries and mainly in Tourism.

Keywords:

Goggle Trends
Tripadvisor
Smart Tourism Destinations
User Generated Content
NLP
Text Mining
Topic Analysis
Sentiment Analysis
# TABLE OF CONTENTS

**ABSTRACT** ................................................................. I

**LIST OF TABLES & GRAPHS** .................................. II

**LIST OF ABBREVIATIONS** ..................................... III

1 INTRODUCTION ......................................................................................................................... 1
   1.1 Research Background and Problem ................................................................. 1
   1.2 Research questions ......................................................................................... 2
   1.3 Objectives of the Study .............................................................................. 2
   1.4 Significance of the Study ........................................................................ 2

2 LITERATURE REVIEW .................................................................................................................. 3
   2.1 Smart Tourism ............................................................................................... 3
   2.2 Smart Tourism Destination ......................................................................... 3
   2.3 Value Creation for Smart Tourism Destination ............................................ 4

3. METHODOLOGY ......................................................................................................................... 5
   3.1 Data Collection Methodology ....................................................................... 5
       3.1.1 Tripadvisor as a source .................................................................... 5
       3.1.2 Google Trends as a source ............................................................... 6
       3.1.3 Scraping techniques ......................................................................... 6
   3.2 Analysis Methodology .................................................................................. 7
       3.2.1 Pre Process ...................................................................................... 7
       3.2.2 NLP and Machine Learning Models ................................................ 8

4 RESULTS ........................................................................................................................................... 11
   4.1 Case Djurgarden .......................................................................................... 11
   4.2 Case ABBA Museum ................................................................................... 18
   4.3 Case City Hall ............................................................................................. 23
   4.4 Case Gamla Stan ......................................................................................... 29
   4.5 Case Vasa Museum ..................................................................................... 35

5. DISCUSSION ................................................................................................................................... 41

6. CONCLUSION ................................................................................................................................... 42
   6.1 Limitations and Future Studies ................................................................. 43
LIST OF TABLES & FIGURES

List of Tables

Table 1. Bi-grams Djurgarden ................................................................. 12
Table 2. Corpus with predicted sentiment for Djurgarden Reviews ........................................... 13
Table 3. Topics of interests for Google’s users that searches the query Djurgarden (2013-2020) ........ 17
Table 4. Bi-grams ABBA Museum ........................................................................ 19
Table 5. Corpus with predicted sentiment for ABBA museum’s Reviews ................................... 19
Table 6. Topics of interests for Google’s users that searches the query ‘ABBA museum’ (2013-2020) ... 23
Table 7. Bi-grams City Hall ............................................................................ 24
Table 8. Corpus with predicted sentiment for City Hall’s Reviews ........................................... 25
Table 9. Topics of interests for Google’s users that searches the query ‘City Hall’ (2013-2020) ........ 29
Table 10. Bi-grams Gamla Stan ......................................................................... 30
Table 11. Corpus with predicted sentiment for Gamla Stan’s Reviews ................................. 31
Table 12. Topics of interests for Google’s users that searches the query ‘Gamla Stan’ (2013-2020) .... 35
Table 13. Bi-grams Gamla Stan ......................................................................... 36
Table 14. Corpus with predicted sentiment for the Vasa museum’s Reviews ........................... 37
Table 15. Topics of interests for Google’s users that searches the query ‘Vasa museum’ (2013-2020) ... 41

List Of Figures

Figure 1. Word Cloud Djurgarden reviews on TripAdvisor .......................................................... 12
Figure 2. Overall Results of the Sentiment Analysis Model ......................................................... 13
Figure 3. Topic modelling LSA for Djurgarden positive Reviews ................................................. 14
Figure 4. LSA clustering for Djurgarden positive Reviews .......................................................... 14
Figure 5. Topic modelling LDA for Djurgarden positive Reviews ................................................. 15
Figure 6. LDA clustering for Djurgarden positive Reviews ........................................................ 15
Figure 7. Number of experiences reported per month in Djurgarden (2013-2020) ...................... 16
Figure 8. Interest over time in Google for the query the Djurgarden (2013-2020) ....................... 17
Figure 9. Word Cloud ABBA museum’s reviews on TripAdvisor ............................................. 18
Figure 10. Topic modelling LSA for ABBA museum’s positive Reviews .................................... 20
Figure 11. LSA clustering for ABBA museum’s positive Reviews ............................................... 20
Figure 12. Topic modelling LDA for ABBA museum’s positive Reviews .................................... 21
Figure 13. LDA clustering for ABBA museum’s positive Reviews .............................................. 21
Figure 14. Number of experiences reported per month in ABBA museum (2013-2020) ............. 22
Figure 15. Interest over time in Google for the query the ABBA museum (2013-2020) ............... 22
Figure 16. Word Cloud City Hall’s reviews on TripAdvisor ....................................................... 24
Figure 17. Topic modelling LSA for City Hall’s positive Reviews ............................................... 25
Figure 18. LSA clustering for City Hall’s positive Reviews ........................................................ 26
Figure 19. Topic modelling LDA for City Hall’s positive Reviews ............................................... 26
Figure 20. LDA clustering for City Hall’s positive Reviews ........................................................ 27
Figure 21. Number of experiences reported per month in the City Hall (2013-2020) ............... 27
Figure 22. Interest over time in Google for the query at the City Hall (2013-2020) .................... 28
Figure 23. Word Cloud City Hall’s reviews on TripAdvisor ....................................................... 30
Figure 24. Topic modelling LSA for Gamla Stan’s positive Reviews ......................................... 31
Figure 25. LSA clustering for Gamla Stan’s positive Reviews ................................................................. 32
Figure 26. Topic modelling LDA for Gamla Stan’s positive Reviews ...................................................... 32
Figure 27. LDA clustering for Gamla Stan’s positive Reviews ............................................................... 33
Figure 28. Number of experiences reported per month in Gamla Stan (2013-2020) ................................. 34
Figure 29. Interest over time in Google for the query the Gamla Stan(2013-2020) ................................. 34
Figure 30. Word Cloud Vasa museum’s reviews on TripAdvisor ............................................................ 36
Figure 31. Topic modelling LSA for the Vasa museum's positive Reviews .............................................. 37
Figure 32. LSA clustering for the Vasa museum’s positive Reviews ....................................................... 38
Figure 33. Topic modelling LDA for Gamla Stan’s positive Reviews ...................................................... 38
Figure 34. LDA clustering for the Vasa museum's positive Reviews ....................................................... 39
Figure 35. Number of experiences reported per month in the Vasa museum (2013-2020) ....................... 39
Figure 36. Interest over time in Google for the query 'Vasa museum' (2013-2020) ................................. 40
LIST OF ABBREVIATIONS

UGC : User Generated Content
ICT : Information and Communications Technology
STD : Smart Tourism Destinations
NLP : Natural Language Processing
LSA : Latent Semantic Analysis
LDA : Latent Dirichlet Allocation

Abstract

This paper aims to show how User Generated Content can create value for Smart Tourism Destinations. Applying the analysis on 5 different cases in the region of Stockholm to derive patterns and opportunities of value creation generated users in tourism. Findings of this paper is also discussed in terms of improving decision making, possibilities of new business models and importance of technological improvements on STD’s. Finally, thoughts on models are presented for researchers and practitioners that might be interested in exploitation of UGC in the context of information-intensive industries and mainly in Tourism.
1. Introduction

1.1 Research background and problem

The term 'big data' applies to the vast collection of knowledge and the systems processing these large data sets. Gandomi and Haider (2015) illustrate the three Vs (volume, variety, and velocity) that define big data, and argue that conventional data management systems are inadequate to handle it, resulting in big data technologies capable of generating real-time information from large quantities of various data. In this regard, Sanz (2013) claims, among other sources, that cities with a suitable operating system can store, analyze and produce near-real-time business intelligence (BI) with big data collected from social media feeds.

Although the advent of Big Data (Gandomi & Haider, 2015; Laney, 2001) is a general phenomenon for all industries, now it also becomes more important for the tourism industry (Koo, Gretzel, Hunter, & Chung, 2015, Werthner & Klein, 1999). From an information-based perspective, tourism is a dynamic trend in which data, information and knowledge, from and about tourists, are the fundamental foundations for the competitiveness and developments of destinations (Hjalager & Nordin, 2011; Jafari, 2001) and how tourists play a significant role as a key source of knowledge (Hall and Williams, 2008).

The big data created on the web is becoming more important for the generation of insights. Tourism is one of the leading growth sectors, so methods and technologies that used for creating insights with Big Data can also be applied in this sector to simplify and empower the selection, processing and analysis of web content are becoming increasingly relevant in this application field.

Despite awareness of the opportunities provided by Big Data in the tourism context Smart Tourism's use of data for value creation is still premature (Gretzel et al., 2015). In addition, to the best of our knowledge, there has been little work on how tourism organizations can use the big data that tourists produce on their travel experiences for a more effective value creation process; this calls for a more in-depth analysis.

1.2 Research questions

Considering the mentioned gaps, this paper aims to show that valuable information from UGC can be extracted for Smart Tourism Destinations. In this research the following research question will find answer: What information can be revealed from the attractions of Stockholm considering their reviews in tripadvisor and google trends using text analysis that can let tourism operators of the city understand what is relevant for the tourists of the international market to support their businesses decisions. To prove this, a several case study methodology is used to assess social big data generated by the visitors of the well
known attractions related to destination Stockholm, the Swedish capital. The aim is to provide context for the process of creating value to derive implications for the agendas of practitioners and researchers dealing with UGC and Smart Tourism Destinations.

1.3 Objectives of the Study

It is important to understand what will be achieved when the research question is answered in this study. Given that people are the main interest for the tourism industry, the aim of this thesis is helping to understand the people behaviour and needs to support the businesses decisions in this sector. The main objective of this thesis is to perform a descriptive analysis of the reviews for the 5 attractions extracted to demonstrate the potential usefulness of applying social media analytics to create relevant knowledge that can be used in market intelligence in the tourism industry fulfilling the gap of understanding the tourists and the destination relationship by using the user generated content on the web.

For the specific objectives of this study we can mention couple checkpoints to answer the research question, which are:
- Review the literature related to UGC and STD relations
- Selection of the cases that fits for our Thesis
- Collection of the reviews from the top attractions from Stockholm in Tripadvisor
- Collection of data of the queries for the same attractions from Google Trends.
- Analysis of the extracted data –
- Interpretation of the results
- Discussion about the finding of this study and limitations of this Thesis.

1.4 Significance of the Study

Research is valued by the variety of stakeholders that might be interested in it. As well as the academic contribution of this paper, it is possible to mention some other stakeholders that might create value from this conducted study in many levels.

Analysis on the UGC for tourism destinations can be considered as an application under the concept of Smart Cities, Smart Tourism and Smart Tourism Destinations. For such concepts, especially Smart Tourism and Smart Tourism Destinations, we know that in actuality, more work is done theoretically rather than the applications of it, therefore this Thesis and its results carries out important implications for these frameworks with the purpose to expand the levels of study that are conducted considering as a trustful source the user generated content in the travel internet information that. This thesis is intended to contribute in the applied case studies to incentive more applications in different regions and markets that can be performed in the future.

As it is obvious the thesis itself is providing models and possibilities for the tourism industry that can be personalized for different smart tourism destinations and tourists spots. The insights that can be obtained from such analysis can be used for development of new products/services and improvement in decision-making processes for the tourism providers and operators.

The following sections of the paper includes: the literature background summarized the current developments in the debate on Big Data for Smart Tourism Destination, and also opportunities to generate value from Big Data in Tourism; the methodology describes the research approach consisting of data collection and analysis sections, case studies; findings and discussions present the results in terms of Big Data value creation.
2. Literature Review

2.1 Smart Tourism

Smart tourism within the new technological framework refers to the competitive advantage that comes from using Smart technologies such as sensors, beacons, mobile phone apps, radio frequency identification (RFID), near-field communication (NFC), smart meters, the Internet-of-Things (IoT), cloud computing, relational databases, etc., that together form a smart digital ecosystem that fosters data-driven innovations and supports new business models (Gretzel, 2018, p.173).

A new way of practicing tourism that enables tourists to access the services and information regarding their tour more conveniently thanks to some advanced technologies and interactive/participative management. Smart tourism does not only deal with tourists but also with residents. Tourists’ and residents' wishes and needs can be understood more accurately in a smart setting due to advanced technology and interactive/participative management.

Smart Tourism has focused on the use of advanced technologies to transform data into efficient new business models by using and evaluating data collected through physical infrastructures and social connections. It relies on new technologies such as ICT, mobile communication, cloud computing, artificial intelligence, and virtual reality in order to provide better tourism experiences.

2.2 Smart Tourism Destinations

Smart Tourism Destinations stored in massive tourism resources within data centres, supported by Internet of Things and Cloud Computing, focused on enhancing tourists experience through intelligent identification and monitoring. The real sense of Smart Tourism Destinations is to focus on tourists' needs by combining the ICT with casual culture and tourist innovation industry in order to promote tourism service quality, improve tourism management and enlarge industry scale to a broader extent (Huang et al. 2012).

According to Buhalis to define a Smart Tourism Destination the integration of ICT within a destination solely will not be enough. Buhalis suggested 4 key concepts. These concepts, human capital, leadership, social capital, and innovation must be considered and added that advanced ICT infrastructures such as Cloud Computing and the Internet of Things will provide the essential infrastructure for developing a Smart Tourism Destination (Boes & Buhalis, 2015)

The priorities of Smart Tourism Destinations construction are to enhance tourists’ travel experience; to provide more intelligent platform both to gather and distribute information within destinations; to facilitate efficient allocation of tourism resources; and to integrates tourism suppliers at both micro and macro level aiming to ensure that benefit from this sector is well distributed to local society (Rong, 2012).
2.3 Value Creation for Smart Tourism Destination

Most of the time, tourists only have limited knowledge and low awareness on destinations they visit. Each tourist has different needs and preferences. Developing crowdsourced applications by using tourists’ input could give valuable insight to destinations in capturing tourists’ demand and tourist complaints in a timely manner (Haubensak 2011).

According to Del Vecchio analysing digital local experiences enables a better understanding of the potential offered by digital technologies for a tourism destination’s smart configuration. The findings of their research showed the need for combining the experiences and ability to develop a regional offering of knowledge intensive business services to sustain the smart tourism destination development. This moves towards the implementation of data-driven business models able to combine the benefits deriving from the domain of Big Data, both in terms of methodologies and technologies, with the demand of more personalized and co-created tourism experiences (Vecchio & Ndou & Secundo, 2018).

“ICTs will provide the “info-structure” for the entire industry and will overtake all mechanistic aspects of tourism transactions.” said by Buhalis and Law. They stated that the future of e-Tourism will be focused on consumer centric technologies that will support organisations to interact with their customers dynamically. Their study revealed that the consumers are evolved with ICTS and able to determine elements of their tourism products. This evolution makes the consumers more sophisticated and experienced, thus, being much more difficult to please. A solution suggested as development of ICT applications that empowers suppliers and destinations efficiency and reconstructs their communication strategies (Buhalis & Law, 2008).

Buhalis and Amaranggana stated that applying smartness concepts within destinations is deemed necessary to potentially enhance tourism experience through advance feedback loop, enhanced access to real-time information and advanced customer service through Internet of Things to address factors that potentially shape negative experiences. They also discovered the personalised services expected by tourists to be offered within Smart Tourism Destinations in order to enhance their tourism experience which characterised as; Before, During and After the trip. They also suggested that to offer personalised services to the customers the STD’s should create the environment that they can access real-time users data, collect instant feedback loop about services and a platform that stakeholders exchange data to promote service integration and ability to precisely predict what visitor wants through historical data (Buhalis & Amaranggana, 2015).

Brandt and Bendler collected more than 600,000 geo-referenced Twitter messages and demonstrated the potential value of spatial and semantic analytics to the tourism sector by combining methods that analyze position, textual content, and photo attachments of tweets, their results showed that the information contained in social media data provides insights into the presence, environmental engagement, and topical engagement of users across the city. (Brandt & Bendler & Neumann, 2017)

“UGC analytics should be seen as an important asset in destination smartness, as it is useful to make ‘smarter’ decisions in several areas such as destination planning,strategy, destination branding and imaging, and multiple territory brand architecture” said by Estela Marine-Roig Salvador Anton Clavé. They stated that with UGC analytics it is possible to gather information about territorial brands and destinations, post dates, user languages and hometowns, and post topics, prior to content analysis itself (Marine-Roig & Clave, 2015).

Kun Kim stated that applying sentiment analysis on destination marketing research could reduce research cost and time. The results showed that with an automated customized algorithm it is possible to analyze the data almost real time. They also revealed the hybrid-
method they used could provide practical and detailed insights about reasons why travellers feel negatively about certain travel destinations. It is also stated that the insights gained from the analysis could be shared and utilized by various stakeholders, such as hospitality service providers, policy makers, and educational institutes (Kim & Park & Yun, 2017).

UGC makes up an increasing share of overall Internet content (Liu, 2007). Regarding the tourism domain, UGC comes in different forms, but the most relevant in this study is the one that comes as product reviews on platforms like TripAdvisor. Especially in tourism, social media gain more and more attention and play an increasingly important role in customers’ decision making process (Lexhagen, Kuttainen, Fuchs, & Höpken, 2012). On one side, the tourists use review platforms in order to express their opinion on the tourism services and products that they have consumed and, on the other side, they use it to inform themselves about relevant products and services and their characteristics such as quality and suitability before the consumption takes place. UGC is also a valuable knowledge for tourism service providers in order to understand and learn about what customers are saying about the tourism products and services in a specific place. Usually product reviews are unbiased and concrete feedback from the customers giving information about the product quality and suitability for a certain segment of customers, which can be interpreted as a valuable input in order to optimize the products and services that are offered to the tourists and improve their customer relationship.

When UGC is analyzed, it is usually done to find out if a user’s opinion is positive or negative regarding a specific topic or tourism service. When a user posts a review in one of these platforms, it is commonly described by basic demographic characteristics and travel motives in a well-structured format, while the review itself can lead to the extraction of opinions and topics from the text. As the number of reviews is increasing drastically, manual methods of analysis are no longer practical, for this reason, automatic methods of extraction of knowledge from these texts are gaining high attention in the research field during the last years. These approaches of natural language processing (NLP), include analysis like N-grams, Sentiment Analysis and Topic Modelling where the process is intended to be done automatically by statistical and/or machine learning methods. These approaches are intended to discover relevant patterns regarding the sentiment or polarity of the feedback and the topic that the feedback is about (Liu, 2017).

2.4 Tourism Psychology Theory

Tourism psychology is a new discipline from the psychological perspective to study travel and tourism industry, which studies the rule of human behavior in tourism activities, mainly including tourism consumer behavior, tourism service behavior and consumer psychology tendency (P. Zheng, Y. Ma, and T. Li, 2009.). Foreign scholars on the research of this field began in the late 1970s, and Chinese scholars also began in the mid-80s then tourism psychology teaching and scientific research. Tourism psychology research has two ideas; one is based on travel consumer object, and studies the general rule of tourists spending behavior. Another side, it is to study the interaction between tourism workers and tourists. The latter is based on the theory of interaction analysis, profound analysis was carried out on the interpersonal relationships in the tourism industry. This paper draws on the two kinds of thoughts to tourists as object, the research of the general regularity of tourist behavior and tourist motives influence on travel behaviour. (Liu, Yan & Yang, Qizhi & Pu, Bo., 2015.)

2.5 Consumer Decision-Making Factors Theory

Consumers purchase choice is mainly by four psychological factors: motivation, perception, learning, beliefs and attitudes. Need to determine motivation, the demand for real commodity use value as a direct result of purchasing motivation, based on social needs, the need for respect, can also lead to the motivation of online purchase. Perception is the objective things by sense organs directly reflected in the human brain. Consumers have the evening after purchase demand, whether to take action, influenced by their perception of objective things. Learning is to point to by experience caused by the changes in individual behavior. A person's learning is suggested by driving force,
stimulating, and reaction and strengthen the influence of interaction. At the same time, consumers through practice and learning beliefs and attitudes. (Liu, Yan & Yang, Qizhi & Pu., Bo., 2015.)

2.6 The Decision-Making Behavior of Tourists

Tourism is refers to the tourists' decision-making behavior of tourism destination choice behavior. From the review of tourists' behavior research, it is not difficult to find the decision-making behavior of tourists is a key component of tourist behavior. Behavior relative to the tourists, tourism demand forecasting, tourism spatial structure research content, research the decision-making behavior of tourists is still relatively weak. Only in recent years, domestic scholars begin to pay close attention to this area. Scholars from the tourist decision-making process, the factors influencing the decision of the decision-making behavior of tourists and tourism enterprises market segment aspects of study, for the tourism administrative department, and provide theoretical help and guidance. (Liu, Yan & Yang, Qizhi & Pu., Bo., 2015.)

2.7 Text Analytics

There are similar applications related to text analytics method. For the data used by these similar applications are from social media sources such as Facebook, Twitter and some online review data from online travel agencies and online tourism product sales like TripAdvisor and Yelp. All of these studies are made using data taken from online sources and they require the process of identifying the required data form, given that the data extracted from a specific source differs from the others. Therefore, each similar application using different tools and methods in obtaining data extracted from the online source platform.

Referring to the related works, there are two types of Natural Language Processing used by similar applications previously called Sentiment Analysis and Text Analytics. Regarding sentiment analysis, there are studies that identify social sentiment through social media using Twitter post data. This suggests that, sentiment analysis and text analytics are one of the ways to process opinions that are voiced by people in order to gain relevant information. (Chiu, S. I., et al., 2018)

This method is widely used because through research that has been made, areas that use sentiment analysis and text analytics include healthcare, politics, tourism and business. These are the major areas that involve direct contact with humans. From the research conducted was possible to see that there are various studies conducted on sentiment analysis, but no for text analytics and most of them are theoretical studies and not applied to real life problems. (Kadir, Nasibah & Aliman, Sharifah. 2020) Through this observation, it can be concluded that there is still a gap between the use of these methods in theoretical studies and in real life problems.

In addition, all of the above studies have features that are used in displaying analytical results from sentiment and text analysis works that have been done. As observed, the extraction and term frequency data process are mandatory processes that has been done by all of the above studies. Next are the features analysis results used by these studies are Word Cloud, N-grams and topic analysis. Lastly, the understanding of this thesis can give insights that can be further expanded to gain the breadth of generating information and upgrade the quality of an insight in this industry.

3. Research Methodology

In this section it is explained all the methods used during the analysis process including the choice of the city Stockholm, Data collection in TripAdvisor and Google Trends, Scraping techniques, Data pre-processing and NLP and sentiment analysis that were utilized in this thesis.

Choice of the smart tourism destination: Stockholm
The city of Stockholm is the principal city of Sweden in terms of economic and social importance, is one of the oldest cities in Scandinavia and also, it is considered as the capital and heart of Scandinavia. Stockholm is Scandinavia's mid-point, a leading cultural city and a financial center. It is the place in Scandinavia that attracts the most visitors and ranks as one of Europe's leading tourism destinations, a trend that has been growing during the last years especially by international travelers who have had a faster growth than domestic travellers. According to Stockholm Business Region, in 2018, last edition of their annual report of the tourism industry in Stockholm, there were close to 14.6 million overnight stays in lodging establishments located within Stockholm County with over a 30 percent of all foreign bednights in Sweden.

The City Council of Stockholm, in 2017, adopted the strategy to become a Smart and Connected City with a focus on sustainability. This included different approaches in areas like Climate, Energy, Mobility, Urban Planning, Social Sustainability, among others.

In this road to become a Smart City, Stockholm, in the scope of Smart Tourism Destination, started collecting real-time information from scattered sensors in the city and processes them in order to provide accurate city information through end-user devices; which reflect the use of ICT as a predictive tool to implement a smarter way of managing Tourism Destinations (Achaerandio et al. 2011). It is here where the concept of Smart Tourism Destination plays an important role in the actuality, given that cities are using the technologies that are available in order to offer a better overall experience to the tourists and therefore, having more individuals every season.
3.1 Data Collection

3.1.1 Tripadvisor as a source

In this Study TripAdvisor was the main source of information for the Data Collection process, which is considered one of the most relevant travel review platforms. This platform helps 463 million travelers each month. TripAdvisor has more than 860 million reviews and opinions of 8.7 million accommodations, restaurants, experiences, airlines and cruises. This platform is available in 49 markets and 28 different languages.

The main difference between social media platforms and travel review platforms, regarding tourism, is that social media is used to inform their social peers about their holidays, meanwhile review platforms, like TripAdvisor, are used specifically to qualify the quality of the products and services that they have used during their holiday experience (Murphy, Gil, & Schegg, 2010). UGC and product reviews have a strong influence on the decisions regarding the travel of the tourists. 90 % of travelers consider other consumers' comments during their trip planning, 87 % of travelers say that reviews impacted their choices during the travel experience, and 70 % of travelers trust online recommendations. It is considered that UGC is more up-to-date, reliable and enjoyable compared to the regular information given by the traditional travel services providers. (Gretzel, Yoo, & Purifoy, 2007)

3.1.2 Google Trends as a source

Google has the information about every search that has been conducted for a specific query since 2004, being able to filter the search for industry, regions and timeframe. The displayed data is scaled to have a more visually simple way to describe data, where it does not show a specific number of queries during time but a comparison between the number of queries in the specific timeframe that it is used as input.

One of the many APIs from google, Google Trends, is a search trend that shows the frequency of a given search query entered into Google’s search engine. It gives information such as the volume over a given time frame, comparative keyword research, related queries to the keyword that it has been analyzed, search volume index and geographical information about the search engine users.

Data from the Google Trends platform are normalized over the selected period and Google Trends explain this as follows: “Search results are normalized to the time and location of a query by the following process: Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic’s proportion to all searches on all topics. Different regions that show the same search interest for a term don’t always have the same total search volumes”
3.1.3 Scraping techniques

In order to extract the thousands of reviews from TripAdvisor that were intended to be used in this Thesis, scraper scripts were developed using the libraries Selenium and BeautifulSoup from python with a model that was able to extract all the data needed. Scraping is a technique for automated data collection, in order to extract large volumes of data in the same way a human would do it and parse its content into structured data that can be used for text mining after. This process consists in getting the whole webpage and its static information to get the raw HTML. However, today, most websites that uses UGC display their content as you swipe, scroll or tap, which means that it is dynamic information, cannot be accessed through the raw html. For these situations, Selenium helps simulate these actions in order to fetch the desired information of the website to extract. After this information is extracted it needs to be parsed, so the library BeautifulSoup helps interpreting the RAW HTML to transform it into structured data. With these tools it is possible to collect data that other methods do not allow.

3.2 Data Analysis

3.2.1 Pre Process

Preprocessing the data is one of the process that consumes more time in the whole process of analyzing data and can be crucial in order to obtain the desired results in the machine-learning algorithms that are intended to use in the analysis given that if the data preprocessing is not handle correctly, the algorithms will not run properly or can give misleading results (John Wiley & Sons, 2014). In social media, UGC data are rich with annotations and free-style handwriting given that in social media people can write whatever they want in contrast to organizations, where both the data and the hierarchy of knowledge are well-organized (Zuber, 2014).

In this process there are four mandatory steps involved: data cleaning, data integration, data transformation, and data reduction. These processes are integral components of data preparation. If any of these steps are not performed as planned, data mining algorithms will not run and will probably give unexpected results. It is very important to have good and robust data, even more important than having efficient algorithms that can be applied to large data sets with poor quality.

Data cleaning

This is the first process that was intended to remove noise, fill out missing values, fix inconsistencies in data and identification and removal of outliers. Dirty data in the database can be the result of wrong data entry, update, or transmission(Springer, 2015) . In social networks, this process is not always straightforward given that there are many factors that can affect the text itself (P. Gupta and V. Bhatnagar, 2015). It requires an understanding of the data given that can include different elements like: HTML tags, Images, audio, video, URLs and white spaces. Machine learning methods have been commonly used to perform data cleaning in social data(S. S. De and S. Dehuri, 2014). In this case the performed data cleaning also included: trimming out white spaces, punctionations, special characters, hashtags and stop words of the reviews.

Data integration
Inconsistency and redundancy in social data are very likely because users have different perspectives and show a different behavior (L. Wenxue and G. Sun, 2013). Data integration aims at combining data from several sources into coherent data storage (J. Han, et al., 2011). This task requires achieving a match between different schema types and inefficient or incomplete data integration can lead to redundancy and/or inconsistency of the data while a proper implementation will enhance the speed and accuracy in the processes that are upcoming. In this study techniques of data integration were used like: entity identification, redundancy and correlation analysis, as well as data value conflict detection and resolution.

**Data transformation**

Data transformation is the process to transform data into an usable and understandable format in order to optimize the data mining operations (J. Han, et al., 2011). Data transformation in this study includes techniques such as data smoothing, feature construction and normalization. These processes require human supervision and are highly dependent on the data being preprocessed (M. Al-Taie and S. Kadry, 2012). In social media data preprocessing, also consider data stemming to transform all the terms from the extracted reviews into their morphological roots, in this case using Porter Stemmer (T. H. Wen, et al.) to reduce the number of words to analyze.

**Data reduction**

Data reduction aims at reducing the size of data while keeping information loss at a minimum (J. Han, et al., 2011). Some of the techniques used in this study includes dimensionality reduction to reduce the number of attributes or random variables, data cleaning, which can also be considered as a technique to remove undesired content, numerosity reduction to achieve smaller data representations using techniques such as nonparametric models (sampling, histograms and clustering). The most important one used in this study is data compression which consists in compressing data to obtain a reduced copy of the original data. This process can result in losing quality of data if it is not done carefully resulting in a loss of relevant information. Both dimensionality reduction and numerosity reduction can be applied to social data to have a smaller data volume while at the same time producing the same (or near the same) mining results (Taie, Mohammed & Kadry, Seifiedine & Lucas, Joel, 2019). In this case some keywords that might make the results of the analysis biased like the name of the case were removed from the reviews to have a more reliable results, stopwords and using pos tagging only relevant elements from each review were analyzed in order to have more meaningful and precise results in the text mining.

### 3.2.2 NLP and Machine Learning Models

In the study, UGC as reviews in free text form are vast on social media and user review platforms, to the point that it is almost impossible to analyze all this content by humans or extremely costly and slow. Thus, in order to analyze this content, the approach in this conducted study is called sentiment analysis which has been gaining popularity among practitioners and researchers during the last years.

Sentiment analysis can be divided into two sub tasks which are sentiment detection and topic detection. The first one aims to determine the polarity (positive or negative) per each review given a model that it is previously prepared to identify the relevant factors and classify...
this text according to its characteristics. The second one, aims to identify the topic or the idea of what an user is providing in the review that was written in these platforms. For example, if an attraction in Stockholm is considered, there are some topics discovered that could help understanding what the tourists are thinking at the moment they are writing a review.

In this study there were used different models and techniques in order to have a broader view of what was the tourist behaviour in their UGC including: N-gram, LSA, LDA, Naive Bayes Classifier.

*N-grams*

N-grams are a contiguous sequence of $n$ items from a given sequence of text, in this case reviews. The items used in this study were the reviews pre processed before to only have relevant words for the analysis which were collected from the created corpus (P. D. Turney). In this case there were created bi-grams ($n = 2$) and 3-grams ($n = 3$) in order to understand per attraction what were the most relevant concepts mentioned in the reviews which can lead to acknowledge what the tourists care for in these places. Even though N-gram analysis is a simple method it could be extremely useful to knowledge and can lead to hidden trends that show crucial information that can help in the understanding of the behaviour of the tourist in the city.

*Latent Semantic Analysis (LSA)*

Any text can be described by the semantic content it includes. There has been years of development of computational models to create semantic representations for the words in a text. One of these models is Latent Semantic Analysis which consists in a model that works under the principle that words with similar meanings appear in similar contexts (Landauer TK, Foltz PW, Laham D, 1998). This model creates semantic representations of the words analyzing the patterns with which words occur together analyzing a large number of texts that is determined in the training corpus. Then after analyzing when words co-occur or not in the corpus, the model estimates which words should occur in similar texts (i.e, context). In this case LSA was used to try to find similarities in what are the tourists considering the most in the top 5 attractions of the city of Stockholm According to Tripadvisor.

*Latent Dirichlet Allocation (LDA)*

This model is one of the most frequently used in topic-modelling, according to Blei (2008) LDA is a generative probabilistic model for collections of discrete data such as text corpora. LDA consists of a model that considers that a text has multiple themes. This model is a three-level hierarchical bayesian model where every item of the given corpus is modeled as a finite mix over and underlying topics and each one of these topics is modeled as an infinite mixture over and underlying groups of topic probabilities. In text modeling, these probabilities give an explicit representation of a document (Blei DM, 2012). This topic modelling generates groups of words in discrete probability topics. In simple words, LDA algorithms automatically identify words that occur in similar contexts and group them to find relevant topics. In this study it is used to compare the results of the LSA model in order to identify with more precision the relevant information about what tourists are considering in their visits to Stockholm given that it use another method to determinate the topics and therefore it is possible to compare both of the topic modeling methods that were conducted in this thesis.
Naive Bayes Classifier

Naive Bayes is a classification method based on statistical and probabilistic approaches in order to predict future events based on previous experience, which is known as the bayes theorem (Feldman, R., & Sanger, J, 2007). The theorem is combined with "naive" where it is assumed that the conditions between attributes are mutually independent (Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R., 2003). In a dataset, each row / document is assumed to be a vector of attribute values where each value becomes a review of the attributes of Xi (iЄ [1, n])). Each line has a ci class label Є {c1, c2, ..., ck} as the value of the class C variable, so the classification can be calculated the probability value p (C = ci | X = xj), because at Naïve Bayes each attribute is assumed to be free, then the equation obtained is as follows: Opportunity p (C = ci | X = xj) shows the opportunity for the attribute Xi with value xi given class c, where in Naïve Bayes, class C is of qualitative type while attribute Xi can be qualitative or quantitative(Sarifah, Monalisa, 2019). In this case, this model was used to determine if a review was positive or negative given a specific training set which includes more than 10.000 reviews, from the library NLTK of python, previously manually tagged with its polarity.

4. Findings

In this section the results obtained from the analysis will be represented case by case. For each case you will see the results for the techniques which are explained in the section 3.2.2. In terms of the choice of cases, we followed a simple logic by scraping the reviews of the Top 5 attractions in Stockholm in Terms of the number of reviews.

Case Djurgarden

Djurgården is a gorgeous island in the middle of Stockholm. It collects many of the city’s most famous museums and cultural attractions like the Vasa Museum, Gröna Lund, the Abba museum and Skansen. Djurgarden and Gamla stan are different from the other cases due to their nature as a public place. The results that are obtained can be useful for all the attractions that are located inside or close to them. For this attraction there were 1022 reviews collected.

Wordcloud

For this part of study, a Word Cloud was generated from the analysis, after pre-processing the dataset, to have a first approach of what has been discussed most in the reviews disregarding its polarity. In the Word Cloud of Figure 1, it is possible to see that the words ‘Ferry’, ‘Beautiful’ and ‘island’ are the most prominents words highlighted and the less prominters are ‘cold’, ‘Slussen’ and ‘attractive’. This group of words let us understand, as a first and fast analysis, that these factors might be relevant for the tourists that visit Djurgarden during the year disregarding if it was with a positive or negative polarity.
**N-gram Analysis**

The n-gram analysis is used as a second approach for our topic extraction idea in this study, given that this method checks frequency of words without any condition. For this study there were considered bi-grams as the standard rule to compare. The results of the 15 most common values can be seen in Table 1:

<table>
<thead>
<tr>
<th>N</th>
<th>Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>amusement park</td>
</tr>
<tr>
<td>2</td>
<td>vasa museum</td>
</tr>
<tr>
<td>3</td>
<td>abba museum</td>
</tr>
<tr>
<td>4</td>
<td>walk around</td>
</tr>
<tr>
<td>5</td>
<td>gruna lund</td>
</tr>
<tr>
<td>6</td>
<td>open air</td>
</tr>
<tr>
<td>7</td>
<td>gamla stan</td>
</tr>
<tr>
<td>8</td>
<td>great place</td>
</tr>
<tr>
<td>9</td>
<td>around island</td>
</tr>
<tr>
<td>10</td>
<td>spend day</td>
</tr>
<tr>
<td>11</td>
<td>along water</td>
</tr>
<tr>
<td>12</td>
<td>nice place</td>
</tr>
<tr>
<td>13</td>
<td>place visit</td>
</tr>
<tr>
<td>14</td>
<td>museum skansen</td>
</tr>
<tr>
<td>15</td>
<td>lund amusement</td>
</tr>
</tbody>
</table>

Table 1. Bi-grams Djurgarden.

As it is possible to see when the bigrams are created it is easier to understand what the tourists are talking about regarding the Djurgarden. In this case, the tourists are mentioning other attractions like ‘amusement park’, ‘Vasa Museum’, ‘Abba museum’ and ‘Gamla Stan’ which means that the tourists that visit Djurgarden also mention these attractions. Regarding the attraction itself, they are talking about ‘walk around’, ‘open air’, ‘great place’, ‘spend day’, ‘around island’ and ‘along water’ which gives an idea of what the tourists like the most about this place.
Sentiment analysis

To complement the text analysis it was conducted a sentiment analysis using Naive Bayes Method. In this part of the process all the reviews from all the attractions were merged into an unique corpus in order to have better results and then splitted by attraction and polarity. The results of the model can be seen in Figure 2:

![Confusion Matrix](image)

Figure 2. Overall Results of the Sentiment Analysis Model

Seen above, the average accuracy of predicting the polarity of the reviews was 0.80, the average recall was 0.79 and also the average f1-score. It is also important to see that the precision was better in the prediction of class 1 which is related to the positive reviews with a 0.83 compared to a 0.76 in the negative class that it was labeled as 0.

After this process was executed, the corpus were separated by attraction and polarity, there were 1022 reviews collected and the results of the sentiment analysis shows that 890 of them have positive and 132 of them have negative polarity. The representation for the results of sentiment analysis for the case Djurgarden can be seen in Table 2.

![Corpus Table](image)

Table 2. Corpus with predicted sentiment for Djurgarden Reviews.
Continuing with the topic modelling process, a Latent Semantic Analysis was conducted per polarity and attraction, to go one step further in understanding what are the tourists actually talking about Djugarden in TripAdvisor. The results of the principle cluster of words for the positive reviews can be seen in Figure 3.

![Figure 3. Topic modelling LSA for Djurgarden positive Reviews.](image)

It is possible to see that there are some concepts that are repeated in the topics like ‘Museum’, ‘Vasa’ and ‘walk’, but also that there are different concepts that might give the idea of what these groups of tourists’ reviews are talking about like for example topic 1 and 2 talks about close attractions, topic 3 and 8 are talking more about the characteristics of Djugarden, topic 4, 5, 6 and 7 seems to describe what are some activities to do during the stay in this place.

In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics we can see from Figure 4 that the blue points which represents the 1st and the most dense cluster was not well separated from the rest of the clusters by using LSA model.

![Figure 4. LSA clustering for Djurgarden positive Reviews.](image)
In order to have more comprehensive results in the topic modelling process, a Latent Dirichlet Allocation was conducted to complement the LSA analysis and compare to see if it’s possible to have better results. These conditions were found to be the same. Therefore, the analysis was performed per polarity about Djurgarden’s reviews in TripAdvisor, for purposes of ratios, only the LDA for positive reviews will be displayed. The results of the principle cluster of words for the positive reviews can be seen in Figure 5.

When the positive reviews are analysed, figure 4 shows the 4 groups of words that were created, with similar characteristics than the LSA topic modelling. Thus, we believe that these are the most relevant concepts that the tourists are considering at the moment of writing their reviews which are probably the things that caught their attention during the visit.

In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics.

**Date of experience**

For each review that a user of TripAdvisor posts, they can choose if they want to put the date of experience or not. For this analysis, the reviews were grouped by dates in order to understand when do the tourists actually go to visit Djurgarden and try to identify if there is a seasonality for this attraction itself to understand when are the tourists coming. Figure 7 shows the months with more visitors according to their experience reported.
Figure 7 shows that the most with more experiences reported during this period of time were July, August, June, September and May which are the months during the end of spring and summer time revealing a seasonality among the TripAdvisor’s users visiting this attraction.

**Interest over time in Google**

For this part of the analysis, google trend’s information was retrieved for the same timeframe (2013 - 2020) than the reviews that were collected for the Djugarden in order to identify if there is a relation between when does people search in google about this attraction and when they actually visit it. Figure 8 shows the interest over time of the query ‘Djugarden’ in the category of travel.
In the previous series it is possible to identify some peaks per month were usually happened between May and September, which are the same months that people reported in TripAdvisor that they had their experience in the Djurgarden, but also, there is possible to see other seasonality between January and March, giving another perspective of how people behave on the Internet regarding this specific Attraction while they are interested on visiting it.

**Related Topics in Google**

Google trend also has information about the related topics to Djurgarden that people is searching in google, for this analysis, there was retrieved this information for the same timeframe (2013 - 2020) in order to understand what does also is interesting for the potential tourists that are looking Djurgarden as a potential place for visiting during their stay in Stockholm.

<table>
<thead>
<tr>
<th>N</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Museum</td>
</tr>
<tr>
<td>2</td>
<td>Gamla stan</td>
</tr>
<tr>
<td>3</td>
<td>Stockholm metro</td>
</tr>
<tr>
<td>4</td>
<td>Swedes</td>
</tr>
<tr>
<td>5</td>
<td>Stockholm Arlanda Airport</td>
</tr>
<tr>
<td>6</td>
<td>Vasa Museum</td>
</tr>
<tr>
<td>7</td>
<td>ABBA The Museum</td>
</tr>
<tr>
<td>8</td>
<td>Reykjavík</td>
</tr>
<tr>
<td>9</td>
<td>Stockholm City Hall</td>
</tr>
<tr>
<td>10</td>
<td>Breiðhemblik men’s football</td>
</tr>
</tbody>
</table>

*Table 3. Topics of interests for Google’s users that searches the query Djurgarden (2013-2020).*

As is possible to see from Table 3, the most relevant topics for this users are related to Museums, Gamla Stan, Stockholm Metro, Swedish people, The airport, Vasa Museum and ABBA museum, among others which give a complementary view for the topic analysis performed before proving that there is a relation between what people is talking about on the reviews for Djurgarden and the related queries that they search in Google showing that is very likely that they have a big interest in these topics also.

**Case ABBA:**

ABBA The Museum is a Swedish interactive exhibition about the pop band ABBA that has been active since May 2013. Unlike the other cases ABBA museum fits the definition of Smart Tourism Destination very well. There is audio guided tours inside the museum and they also offer a VR helicopter tour of Stockholm with ABBA songs. The website has a lot of information that can be needed, it is possible to buy tickets online to the events, Website is also connects you to the related social media accounts of the Museum.

There are 4553 reviews scrapped from Tripadvisor for the ABBA museum.

**Wordcloud**

For this part of study, a Word Cloud was generated from the analysis, after pre-processing the dataset, to have a first approach of what has been discussed most in the reviews disregarding its polarity. First, in the Word Cloud of Figure 8, it is possible to see that the words ‘must’, ‘fan’ and ‘recommendable’ are the most prominents words highlighted without considering ‘ABBA’ and ‘Stockholm’ and the less prominent are ‘name’, ‘Official’ and ‘Object’. This group of words let us understand, as a first and fast analysis, that these factors
might be relevant for the tourists that visit the ABBA Museum during the year disregarding if it was with a positive or negative polarity.

N-gram Analysis

Figure 9. Word Cloud ABBA museum’s reviews on TripAdvisor

The n-gram analysis is used as a second approach for our topic extraction idea in this study, given that this method checks frequency of words without any condition. For this study there were considered bi-grams as the standard rule to compare. The results of the 15 most common values can be seen in Table 4.

<table>
<thead>
<tr>
<th>N</th>
<th>Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>audio guide</td>
</tr>
<tr>
<td>2</td>
<td>abba fan</td>
</tr>
<tr>
<td>3</td>
<td>gift shop</td>
</tr>
<tr>
<td>4</td>
<td>worth visit</td>
</tr>
<tr>
<td>5</td>
<td>sing dance</td>
</tr>
<tr>
<td>6</td>
<td>sing along</td>
</tr>
<tr>
<td>7</td>
<td>swedish music</td>
</tr>
<tr>
<td>8</td>
<td>great fun</td>
</tr>
<tr>
<td>9</td>
<td>well worth</td>
</tr>
<tr>
<td>10</td>
<td>interactive exhibit</td>
</tr>
<tr>
<td>11</td>
<td>lot interactive</td>
</tr>
<tr>
<td>12</td>
<td>singing dancing</td>
</tr>
<tr>
<td>13</td>
<td>hall fame</td>
</tr>
<tr>
<td>14</td>
<td>fun interactive</td>
</tr>
<tr>
<td>15</td>
<td>mamma mia</td>
</tr>
</tbody>
</table>

Table 4. Bi-grams ABBA Museum.

As it is possible to see when the bigrams are created it is easier to understand what the tourists are talking about regarding the ABBA Museum. In this case, the tourists are mentioning concepts like ‘audio guide’, ‘sing dance’, ‘great fun’ and ‘interactive exhibit’ which reveals that the visitors of this museum are grateful with the interactive museum and it can be said that they have fun at this museum.
Sentiment analysis

For this analysis it was used the same model that it was explained in the previous case. There were 4553 reviews collected and the results of the sentiment analysis shows a number of positive and negative reviews are 4107 and 546 respectively. The representation for the results of sentiment analysis for the case Djurgarden can be seen in Table 5.

<table>
<thead>
<tr>
<th>Text</th>
<th>pred_sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>the is cool place to visit if interested in th...</td>
<td>1</td>
</tr>
<tr>
<td>If you are or were an fan, you must go and see...</td>
<td>1</td>
</tr>
<tr>
<td>This is a must for Tourists to Stockholm. I wa...</td>
<td>1</td>
</tr>
<tr>
<td>Walk in Dance out........they aren't joking. N...</td>
<td>1</td>
</tr>
<tr>
<td>Words cannot express how much this means to so...</td>
<td>1</td>
</tr>
<tr>
<td>We have always been a bit fan of and had a lot...</td>
<td>1</td>
</tr>
<tr>
<td>it has just opened. Easy to reach with public ...</td>
<td>1</td>
</tr>
<tr>
<td>Highly recommendable. Fantastic. We spent 1,5 ...</td>
<td>1</td>
</tr>
<tr>
<td>It took some time until finally got an museum...</td>
<td>1</td>
</tr>
<tr>
<td>For the first time there is an official museum...</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5. Corpus with predicted sentiment for ABBA museum’s Reviews.

LSA

Continuing with the topic modelling process, a Latent Semantic Analysis was conducted per polarity and attraction, to go one step further in understanding what are the tourists actually talking about the ABBA museum in TripAdvisor, given that the vast majority of reviews were positive, only this ones were considered for the LSA model. The results of the principle cluster of words for the positive reviews can be seen in Figure 10.

![Figure 10. Topic modelling LSA for ABBA museum’s positive Reviews.](image)

It is possible to see that there are some concepts that are repeated in the topics like ‘Museum’, ‘Interactive’ and ‘fun’, but also that there are different concepts that might give the idea of what these groups of tourists’ reviews are talking about like for example topic 1, 2, 3 and 7 are related to the music characteristics of the museum, topic 4 and 5 are talking more about the characteristics of the museum, topic 6 mention other attractions and activities and topic 8 seems to be a mix of the other topics.

In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics:
In order to have more comprehensive results in the topic modelling process, a Latent Dirichlet Allocation was conducted to complement the LSA analysis and compare to see if it’s possible to have better results. The conditions were the same, so the analysis was performed per polarity about ABBA Museums’s reviews in TripAdvisor, for purposes of ratios, only the LDA for positive reviews will be displayed. The results of the principle cluster of words for the positive reviews can be seen in Figure 12.

As it is possible to see, when the positive reviews are analysed, there are only 4 groups of words created, with similar characteristics than the LSA topic modelling but more condensed, which give us an idea that these are the most relevant concepts that the tourists are considering at the moment of writing their reviews which are probably the things that caught their attention during the visit.

In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics:
Date of experience

For each review that a user of TripAdvisor posts, they can choose if they want to put the date of experience or not. For this analysis, the reviews were grouped by dates in order to understand when do the tourists actually go to visit Djurgarden and try to identify if there is a seasonality for this attraction itself to understand when are the tourists coming. In Figure 14 it is possible to see which are the months with more visitors according to their experience reported.

*Figure 14. Number of experiences reported per month in ABBA museum (2013-2020).*

*Figure 13. LDA clustering for ABBA museum’s positive Reviews.*
Figure 14 shows that the most with more experiences reported during this period of time were July, August, June, September and May which are the months during the end of spring and summer time revealing a seasonality among the TripAdvisor’s users visiting this attraction which collaborates with the prior knowledge about the seasonality for this destination.

Interest over time in Google

For this part of the analysis, google trend’s information was retrieved for the same timeframe (2013 - 2020) than the reviews that were collected for the ABBA museum in order to identify if there is a relation between when does people search in google about this attraction and when they actually visit it. Figure 15 shows the interest over time of the query ‘ABBA museum’ in the category of travel.

In Figure 15, it is possible to identify clear peaks per month between May and September, which are the same months that people reported in TripAdvisor that they had their experience in the ABBA museum. Furthermore, there is another seasonality between january and march, giving another perspective of how people behave on the Internet regarding this specific Attraction while they are interested in visiting it.

Related Topics in Google

Google trend also has information about the related topics to the ABBA museum that people is searching in google, for this analysis, this information was retrieved for the same timeframe (2013 - 2020) in order to understand what does also is interesting for the potential tourists that are looking the ABBA museum as a potential place for visiting during their stay in Stockholm.

<table>
<thead>
<tr>
<th>N</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Museum</td>
</tr>
<tr>
<td>2</td>
<td>Stockholm</td>
</tr>
<tr>
<td>3</td>
<td>ABBA</td>
</tr>
<tr>
<td>4</td>
<td>Swedish krona</td>
</tr>
<tr>
<td>5</td>
<td>Skansen</td>
</tr>
<tr>
<td>6</td>
<td>Swedes</td>
</tr>
<tr>
<td>7</td>
<td>Gamla stan</td>
</tr>
<tr>
<td>8</td>
<td>Gröna Lund</td>
</tr>
<tr>
<td>9</td>
<td>The Royal Palace</td>
</tr>
<tr>
<td>10</td>
<td>Djurgården</td>
</tr>
</tbody>
</table>

Table 6. Topics of interests for Google’s users that searches the query ‘ABBA museum’ (2013-2020).
Table 6 shows the most relevant topics for this users are related to Museums, Stockholm, ABBA, The airport, Gamla Stan, Grona lund, The Royal Palace, Djurgarden and swedish people, among others which give a complementary view for the topic analysis performed before proving that there is a relation between what people is talking about on the reviews for the ABBA museum and the related queries that they search in Google showing that is very likely that they also have a big interest in these topics.

Case City Hall

The Stockholm City Hall is the building of the Municipal Council for the City of Stockholm. It is one of the most visited attractions for the city, however the steps for becoming a STD is not taken very well. The website is mediocre, it is not possible to buy tickets online and the only way of booking before visit is by email. 1817 reviews were scrapped from Tripadvisor.

Wordcloud

For this part of study, a Word Cloud was generated from the analysis, after preprocessing the dataset, to have a first approach of what has been discussed most in the reviews disregarding its polarity. In the Word Cloud of Figure 11, it is possible to see that the words 'Nobel', 'Impressive' and 'Building' are the most prominents words highlighted and the less prominents are 'Unexpectedly', 'Length' and 'Breathtaking'. This group of words let us understand, as a first and fast analysis, that these factors might be relevant for the tourists that visit the City Hall during the year disregarding if it was with a positive or negative polarity.

\[ \text{Figure 16. Word Cloud City Hall's reviews on TripAdvisor} \]

N-gram Analysis

The n-gram analysis is used as a second approach for our topic extraction idea in this study, given that this method checks frequency of words without any condition. For this study there were considered bi-grams as the standard rule to compare. The results of the 15 most common values can be seen in Table 7.
Table 7. Bi-grams City Hall.

Table 6 displays the bigrams. This table shows an easier understanding of what the tourists are talking about regarding the City Hall. In this case, the tourists are mentioning concepts like ‘nobel prize’, ‘guided tour’, ‘nobel banquet’ and ‘beautiful building’ which reveals that the visitors care about those mentioned words about the museum.

**Sentiment analysis**

For this analysis it was used the same model that it was explained in the first case. There were 1817 reviews collected and the results of the sentiment analysis shows a number of positive and negative reviews are 1582 and 235 respectively. The representation for the results of sentiment analysis for the case Djurgarden can be seen in Table 8.

<table>
<thead>
<tr>
<th>Text</th>
<th>pred_sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0  Impressive building – if you like architecture...</td>
<td>1</td>
</tr>
<tr>
<td>1  It is where Nobel Laureates have walked. This ...</td>
<td>1</td>
</tr>
<tr>
<td>2  We arrived at the in the afternoon and their l...</td>
<td>1</td>
</tr>
<tr>
<td>3  It was fun to get caught up in the Nobel aura ...</td>
<td>1</td>
</tr>
<tr>
<td>4  The tour guides working at Statshuset are my fa...</td>
<td>1</td>
</tr>
<tr>
<td>1812 I didn’t think the would be all that interesting...</td>
<td>1</td>
</tr>
<tr>
<td>1813 Nice building, impressive to see where the ban...</td>
<td>1</td>
</tr>
<tr>
<td>1814 My beloved place, each time visiting Sto the f...</td>
<td>1</td>
</tr>
<tr>
<td>1815 Breathtaking rooms and fascinating history in ...</td>
<td>0</td>
</tr>
<tr>
<td>1816 Unexpectedly wonderful. The golden room is exqu...</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 8. Corpus with predicted sentiment for City Hall’s Reviews.

**LSA**

Continuing with the topic modelling process, a Latent Semantic Analysis was conducted per polarity and attraction, to go one step further in understanding what are the tourists actually talking about the City Hall in TripAdvisor, given that the vast majority of reviews were positive, only this ones were considered for the LSA model. The results of the principle cluster of words for the positive reviews can be seen Figure 17.
It is possible to see that there are some concepts that are repeated in the topics like ‘Tour’, ‘Nobel’, ‘Tower’ and ‘beautiful’, but also that there are different concepts that might give the idea of what these groups of tourists’ reviews are talking about, in most of the cases they are talking about relatively the same, mention about the nobel, the hall, the tower, tours, place worth visiting. The main factor that can be seen in the group of topics is that some of them focus more in the architecture and views of the building and in the nobel prize related topics.

In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics:

**LSA**

In order to have more comprehensive results in the topic modelling process, a Latent Dirichlet Allocation was conducted to complement the LSA analysis and compare to see if it’s possible to have better results. The conditions were the same, so the analysis was performed per polarity about City Hall’s reviews in TripAdvisor, for purposes of ratios, only the LDA for positive reviews will be displayed. The results of the principle cluster of words for the positive reviews can be seen in Figure 19.
These figures 19 and 20, show that when the positive reviews are analysed, there are only 4 groups of words created, with different characteristics than the LSA topic modelling for example, in topics 1, 2 and 4 they talk about the description of the City Hall considering different factors, but in topic 3 they talk about the concepts more related to the Nobel prize. These topics give us an idea that these are the most relevant concepts that the tourists are considering at the moment of writing their reviews which are probably the things that caught their attention during the visit.

In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics:

**Date of experience**
For each review that a user of TripAdvisor posts, they can choose if they want to put the date of experience or not. For this analysis, the reviews were grouped by dates in order to understand when do the tourists actually go to visit the City Hall of Stockholm and try to identify if there is a seasonality for this attraction itself to understand when are the tourists coming. In Figure 21 it is possible to see which are the months with more visitors according to their experience reported:
Figure 21 shows that the most with more experiences reported during this period of time were August, July, June, September and May which are the months during the end of spring and summer time revealing a seasonality among the TripAdvisor’s users visiting this attraction. It is interesting to notice that in August is the month with more visits.

**Interest over time in Google**

For this part of the analysis, google trend’s information was retrieved for the same timeframe (2013 - 2020) than the reviews that were collected for the ABBA museum in order to identify if there is a relation between when does people search in google about this attraction and when they actually visit it. Figure 22 shows the interest over time of the query ‘ABBA museum’ in the category of travel.

Figure 22 has some peaks between May and September, which are the same months that people reported in TripAdvisor that they had their experience in the City Hall. However, it is possible to see other seasonality between November and January, probably because the nobel prize ceremony is held during December which is an important world wide event. This
gives another perspective of how people behave on the Internet regarding this specific Attraction while they are interested in visiting it.

**Related Topics in Google**

Google trend also has information about the related topics to the City Hall that people is searching in google, for this analysis, there was retrieved this information for the same timeframe (2013 - 2020) in order to understand what does also is interesting for the potential tourists that are looking City Hall as a potential place for visiting during their stay in Stockholm.

<table>
<thead>
<tr>
<th>N</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Museum</td>
</tr>
<tr>
<td>2</td>
<td>Vasa Museum</td>
</tr>
<tr>
<td>3</td>
<td>The Royal Palace</td>
</tr>
<tr>
<td>4</td>
<td>Wedding</td>
</tr>
<tr>
<td>5</td>
<td>Palace</td>
</tr>
<tr>
<td>6</td>
<td>Swedes</td>
</tr>
<tr>
<td>7</td>
<td>Skansen</td>
</tr>
<tr>
<td>8</td>
<td>Nobel Prize</td>
</tr>
<tr>
<td>9</td>
<td>Gamla stan</td>
</tr>
<tr>
<td>10</td>
<td>Alfred Nobel</td>
</tr>
</tbody>
</table>

*Table 9. Topics of interests for Google’s users that searches the query ‘City Hall’ (2013-2020).*

Table 9 shows the most relevant topics for this users to be to Museums, Vasa Museum, The Royal Palace, Nobel prize, Gamla Stan and swedish people, among others which give a complementary view for the topic analysis performed before proving that there is a relation between what people is talking about on the reviews for the City Hall and the related queries that they search in Google showing that is very likely that they have a big interest in these topics also.

**Gamla Stan**

Gamla Stan, the Old Town, is one of the largest and best preserved medieval city centers in Europe, and one of the foremost attractions in Stockholm. As mentioned before in the Djurgarden case, Gamla stan is also a public place and the interpreted results can create value for many stakeholders. The dataset scrapped from Tripadvisor contains 6925 reviews makes it the most reviewed attraction for Stockholm.

**Wordcloud**

For this part of study, a Word Cloud was generated from the analysis, after pre-processing the dataset, to have a first approach of what has been discussed most in the reviews disregarding its polarity. First,in the Word Cloud of Figure 11, it is possible to see that the words ‘streets’, ‘Old’ and ‘Time’ are the most prominents words highlighted and the less prominents are ‘Quirky’, ‘Port’ and ‘Wonderful’. This group of words let us understand, as a first and fast analysis, that these factors might be relevant for the tourists that visit Gamla Stan during the year disregarding if it was with a positive or negative polarity.
N-gram Analysis

The n-gram analysis is used as a second approach for our topic extraction idea in this study, given that this method checks frequency of words without any condition. For this study there were considered bi-grams as the standard rule to compare. The results of the 15 most common values can be seen in Table 10.

<table>
<thead>
<tr>
<th>N</th>
<th>Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>old town</td>
</tr>
<tr>
<td>2</td>
<td>royal palace</td>
</tr>
<tr>
<td>3</td>
<td>narrow street</td>
</tr>
<tr>
<td>4</td>
<td>walk around</td>
</tr>
<tr>
<td>5</td>
<td>souvenir shop</td>
</tr>
<tr>
<td>6</td>
<td>shop restaurant</td>
</tr>
<tr>
<td>7</td>
<td>cobbled street</td>
</tr>
<tr>
<td>8</td>
<td>must see</td>
</tr>
<tr>
<td>9</td>
<td>cafe restaurant</td>
</tr>
<tr>
<td>10</td>
<td>changing guard</td>
</tr>
<tr>
<td>11</td>
<td>nobel museum</td>
</tr>
<tr>
<td>12</td>
<td>ice cream</td>
</tr>
<tr>
<td>13</td>
<td>beautiful building</td>
</tr>
<tr>
<td>14</td>
<td>tourist trap</td>
</tr>
<tr>
<td>15</td>
<td>christmas market</td>
</tr>
</tbody>
</table>

Table 10. Bi-grams Gamla Stan.

Table 10 bigrams shows what the tourists are talking about regarding Gamla Stan. In this case, the tourists are mentioning concepts like ‘Old Town’, ‘Royal Palace’, ‘Souvenir Shop’, ‘Cafe Restaurant’, ‘Tourist Trap’ and ‘beautiful building’ which reveals what visitors care about Gamla Stan. It is interesting to notice that given there are more reviews, the bigrams are more descriptive.

Sentiment analysis

For this analysis, the same model used in the first case was used. There were 6925 reviews collected and the results were 5531 positive reviews and 1394 Negative reviews, also the attraction that has the highest ratio of negative over positive reviews. The representation for the results of sentiment analysis for the case Djurgarden can be seen in Table 11.
Table 11. Corpus with predicted sentiment for Gamla Stan’s Reviews.

<table>
<thead>
<tr>
<th>Text</th>
<th>pred_sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Unwinding in the old streets and buildings was...</td>
<td>1</td>
</tr>
<tr>
<td>1 A very unique old town in the heard of Stockholm...</td>
<td>1</td>
</tr>
<tr>
<td>2 We had a great time walking down the path, eve...</td>
<td>0</td>
</tr>
<tr>
<td>3 Narrow streets, boutique shops and hotels. Tru...</td>
<td>1</td>
</tr>
<tr>
<td>4 This is a beautiful place to walk around. Bril...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6918 Lots of interesting old buildings and quaint n...</td>
<td>1</td>
</tr>
<tr>
<td>6919 A nice historic old town, with lots of quirky ...</td>
<td>1</td>
</tr>
<tr>
<td>6920 We visited Gamla Stan, the old town whilst on ...</td>
<td>1</td>
</tr>
<tr>
<td>6921 We had a wonderful time in Stockholm with our ...</td>
<td>1</td>
</tr>
<tr>
<td>6922 Pulled into port. Weather will be 73 F. This i...</td>
<td>1</td>
</tr>
</tbody>
</table>

LSA

Continuing with the topic modelling process, a Latent Semantic Analysis was conducted per polarity and attraction, to go one step further in understanding what are the tourists actually talking about Gamla Stan in TripAdvisor, given that the vast majority of reviews were positive, only this ones were considered for the LSA model. The results of the principle cluster of words for the positive reviews can be seen in Figure 24.

Figure 24. Topic modelling LSA for Gamla Stan’s positive Reviews.

In Figure 24 there are some concepts that are repeated in the topics such as ‘Old Town’, ‘Great’, ‘Walk’ and ‘beautiful’, but also that there are different concepts that might give the idea of what these groups of tourists’ reviews are talking about. In most of the cases, they are talking about relatively the same thing, walking around, a lot of shops, restaurants, about the streets and the architecture.

In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics:
In order to have more comprehensive results in the topic modelling process, a Latent Dirichlet Allocation was conducted to complement the LSA analysis and compare to see if it's possible to have better results. The conditions were the same, so the analysis was performed per polarity about Gamla Stan's reviews in TripAdvisor. For simplicity, only the LDA for positive reviews will be displayed.

The results of the principle cluster of words for the positive reviews can be seen Figure 26.

Figure 26 shows that when the positive reviews are analysed, there are 8 groups of words created, with similar characteristics than the LSA topic modelling mention a lot about the shops, restaurants, tourists and the streets. Furthermore, this analysis shows that topics are related to close attractions and topic 1 that also mention kids and play which can give us an idea that there are families going to visit this place. These topics give us an idea that these are the most relevant concepts that the tourists are considering at the moment of writing their reviews which are probably the things that caught their attention during the visit.
In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics:

**Date of experience**

For each review that a user of TripAdvisor posts, they can choose if they want to put the date of experience or not. For this analysis, the reviews were grouped by dates in order to understand when do the tourists actually go to visit Gamla Stan and try to identify if there is a seasonality for this attraction itself to understand when are the tourists coming. In Figure 28 it is possible to see which are the months with more visitors according to their experience reported:

![Figure 27. LDA clustering for Gamla Stan’s positive Reviews.](image)

**Figure 27. LDA clustering for Gamla Stan’s positive Reviews.**

![Number of visits per month of Gamla Stan (2013-2020).](image)

**Figure 28. Number of experiences reported per month in Gamla Stan (2013-2020).**
Figure 28 shows that the most with more experiences reported during this period of time were July, August, June, September and May which are the months during the end of spring and summer time revealing a seasonality among the TripAdvisor’s users visiting this attraction. It is interesting to notice that in October and December there are also a considerable number of visits which are off-season compared to the other attractions.

**Interest over time in Google**

For this part of the analysis, google trend’s information was retrieved for the same timeframe (2013 - 2020) than the reviews that were collected for Gamla Stan in order to identify if there is a relation between when does people search in google about this attraction and when they actually visit it. The following graph shows the interest over time of the query ‘Gamla Stan’ in the category of travel.

![Interest over time Gamla Stan(2013-2020)](image)

Figure 29 displays peaks over the months. The first clear peak is in May every year, but also in September and December, which are the same months that people reported in TripAdvisor that they had an experience in Gamla Stan. This gives another perspective of how people behave on the Internet regarding this specific attraction while they are interested in visiting it.

**Related Topics in Google**

Google trend also has information about the related topics to Gamla Stan that people is searching in google, for this analysis, there was retrieved this information for the same timeframe (2013 - 2020) in order to understand what does also is interesting for the potential tourists that are looking Gamla Stan as a potential place for visiting during their stay in Stockholm.
Table 12. Topics of interests for Google’s users that searches the query ‘Gamla Stan’ (2013-2020).

<table>
<thead>
<tr>
<th></th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gamla stan</td>
</tr>
<tr>
<td>2</td>
<td>Restaurant</td>
</tr>
<tr>
<td>3</td>
<td>Hotel</td>
</tr>
<tr>
<td>4</td>
<td>Museum</td>
</tr>
<tr>
<td>5</td>
<td>Södermalm</td>
</tr>
<tr>
<td>6</td>
<td>Bar</td>
</tr>
<tr>
<td>7</td>
<td>Christmas market</td>
</tr>
<tr>
<td>8</td>
<td>Scandic Hotels</td>
</tr>
<tr>
<td>9</td>
<td>Cafe</td>
</tr>
<tr>
<td>10</td>
<td>Swedes</td>
</tr>
</tbody>
</table>

Seen in table 12, the most relevant topics for this users regarding Gamla Stan are Museums, Restaurants, Hotels, Bars, Cafes and swedish people, among others, which give a complementary view for the topic analysis performed before proving that there is a relation between what the people are talking about on their reviews for Gamla Stan and the related queries in Google showing that is very likely that they have a big interest in these topics also. It is noteworthy that ‘Christmas Market’ is also one of top 10 related topics because it reveals that it is very likely that the large number of visits and web searches in december might be related to this.

Vasa Museum

The Vasa Museum is a maritime museum on the island of Djurgården, the museum displays the only almost fully intact 17th century ship that has ever been salvaged, the 64-gun warship Vasa sank on its maiden voyage in 1628. Vasa Museum also shows some of the characteristics of a STD. It is possible to buy tickets online for entrance, it is also possible to visit the museum online with the videos uploaded which is similar to a guided tour in the museum. There are 2480 reviews collected from Tripadvisor.

Wordcloud

For this part of study, a Word Cloud was generated from the analysis, after pre-processing the dataset, to have a first approach of what has been discussed most in the reviews disregarding its polarity. First, in the Word Cloud of Figure 11, it is possible to see that the words ‘Museum’, ‘Ship’ and ‘Interesting’ are the most prominents words highlighted and the less prominents are ‘seen’, ‘Remarkable’ and ‘Shipwreck’. This group of words let us understand, as a first and fast analysis, that these factors might be relevant for the tourists that visit the Vasa museum during the year disregarding if it was with a positive or negative polarity.
N-gram Analysis

The n-gram analysis is used as a second approach for our topic extraction idea in this study, given that this method checks frequency of words without any condition. For this study there were considered bi-grams as the standard rule to compare. The results of the 15 most common values can be seen in Table 13.

<table>
<thead>
<tr>
<th>N</th>
<th>Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>must see</td>
</tr>
<tr>
<td>2</td>
<td>guided tour</td>
</tr>
<tr>
<td>3</td>
<td>worth visit</td>
</tr>
<tr>
<td>4</td>
<td>maiden voyage</td>
</tr>
<tr>
<td>5</td>
<td>well worth</td>
</tr>
<tr>
<td>6</td>
<td>well preserved</td>
</tr>
<tr>
<td>7</td>
<td>see stockholm</td>
</tr>
<tr>
<td>8</td>
<td>old ship</td>
</tr>
<tr>
<td>9</td>
<td>history ship</td>
</tr>
<tr>
<td>10</td>
<td>around ship</td>
</tr>
<tr>
<td>11</td>
<td>highly recommend</td>
</tr>
<tr>
<td>12</td>
<td>ship sank</td>
</tr>
<tr>
<td>13</td>
<td>free guided</td>
</tr>
<tr>
<td>14</td>
<td>viking ship</td>
</tr>
<tr>
<td>15</td>
<td>djurgarden</td>
</tr>
</tbody>
</table>

Table 13 bigrams shows what the tourists are talking about regarding the Vasa museum. In this case, the tourists are mentioning mostly about the ship itself that is held in this museum, but also there are other concepts like ‘Guided Tour’, ‘Must See’, ‘See Stockholm’ and ‘Djurgarden’ among others which reveals that what does the visitors care about the Vasa Museum.

Sentiment analysis

For this analysis it was used the same model that it was explained in the first case. There were 2480 reviews collected and the results were 2049 positive reviews and 431 Negative reviews. The representation for the results of sentiment analysis for the case Djurgarden can be seen in Table 14.
Continuing with the topic modelling process, a Latent Semantic Analysis was conducted per polarity and attraction, to go one step further in understanding what are the tourists actually talking about the Vasa museum in TripAdvisor, given that the vast majority of reviews were positive, only this ones were considered for the LSA model. The results of the principle cluster of words for the positive reviews can be seen in Figure 31.

Some concepts are repeated in the topics such as: ‘Ship’, ‘interesting’, ‘museum’ and ‘recommend’. In this particular case it is possible to see that all the topics are related to the same, the good experience in the museum and how interesting it is regarding the ship itself and its history, which gives us a perfect idea of what is what the tourists like the most about this attraction.

In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics:
In order to have more comprehensive results in the topic modelling process, a Latent Dirichlet Allocation was conducted to complement the LSA analysis and compare to see if it’s possible to have better results. The conditions were the same, so the analysis was performed per polarity about the Vasa museum’s reviews in TripAdvisor, for simplicity, only the LDA for positive reviews will be displayed.

The results of the principle cluster of words for the positive reviews can be seen Figure 33.

As seen on figures 32 and 33, when the positive reviews are analysed, there are only 3 groups of words created. This is most likely due to the topics created with the LSA model are really close to each other. In this case, topic 1 has descriptions of how good and interesting the museum is. Topic 2 is talking more about what is suitable for childrens and young people and also about the information and the film. Lastly, topic 3 they talk about the tickets of the attraction itself and the staff members. These topics give us an idea that these are the most relevant concepts that the tourists are considering at the moment of writing their reviews which are probably the things that caught their attention during the visit.

In order to have a better understanding about the distribution of this topics, the following cluster analysis was performed using trigrams to differentiate better the topics:
Date of experience

For each review that a user of TripAdvisor posts, they can choose if they want to put the date of experience or not. For this analysis, the reviews were grouped by dates in order to understand when do the tourists actually go to visit the Vasa museum and try to identify if there is a seasonality for this attraction itself to understand when are the tourists coming. In the following graph it is possible to see which are the months with more visitors according to their experience reported:

Figure 35 shows that the most with more experiences reported during this period of time were September, August, July and June which are the months during the end of spring
and summer time revealing a seasonality among the TripAdvisor’s users visiting this attraction. It is interesting to notice that in October there are more visitors than in May which is a considerable number of visits during off-season compared to the other attractions.

**Interest over time in Google**

For this part of the analysis, google trend’s information was retrieved for the same timeframe (2013 - 2020) than the reviews that were collected for the Vasa museum in order to identify if there is a relation between when does people search in google about this attraction and when they actually visit it. Figure 36 shows the interest over time of the query ‘Vasa museum’ in the category of travel.

![Figure 36. Interest over time in Google for the query ‘Vasa museum’ (2013-2020).](image)

In the previous series it is possible to identify seasonality every year between April and November, a period of time that is bigger compared to the other attractions previously discussed. July is the month with the peak each year. Also it is possible to see a growing trend each year, which means that every year there are more people interested in searching about this museum. The months that reveal more searches are the same that people reported in TripAdvisor that they had their experience in the Vasa museum, with the difference that September is the one month with more experiences reported in TripAdvisor instead of July that is the month with more searches. This gives another perspective of how people behave on the Internet regarding this specific attraction while they are interested in visiting it.

**Related Topics in Google**

Google trend also has information about the related topics to the Vasa museum that people is searching in google, for this analysis, there was retrieved this information for the same timeframe (2013 - 2020) in order to understand what does also is interesting for the potential tourists that are looking the Vasa museum as a potential place for visiting during their stay in Stockholm.
Table 15 shows the most relevant topics for the users regarding Vasa Museum are Skansen, Ship, ABBA museum, The Royal Palace, Gamla Stan and swedish people, among others which give a complementary view for the topic analysis performed before proving that there is a relation between what people is talking about on the reviews for the Vasa museum and the related queries that they search in Google showing that is very likely that they have a big interest in these topics also, in this case most of the related queries are other attractions of the city of Stockholm, which gives the impression that people tend to be more keen to visit more attractions when they think about the Vasa museum.

<table>
<thead>
<tr>
<th>N</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Skansen</td>
</tr>
<tr>
<td>2</td>
<td>Ship</td>
</tr>
<tr>
<td>3</td>
<td>Swedes</td>
</tr>
<tr>
<td>4</td>
<td>ABBA The Museum</td>
</tr>
<tr>
<td>5</td>
<td>The Royal Palace</td>
</tr>
<tr>
<td>6</td>
<td>Gamla stan</td>
</tr>
<tr>
<td>7</td>
<td>Stockholm City Hall</td>
</tr>
<tr>
<td>8</td>
<td>Nordiska museet</td>
</tr>
<tr>
<td>9</td>
<td>Djurgården</td>
</tr>
<tr>
<td>10</td>
<td>Gröna Lund</td>
</tr>
</tbody>
</table>

Table 15. Topics of interests for Google’s users that searches the query ‘Vasa museum’ (2013-2020).

5. Discussion

Considering the analysis it is possible to discuss the results in some aspects like data preprocessing, descriptive analytics, sentiment analysis and models applied.

Data preparation is significantly important in such datasets due to the nature of unstructured UGC content. We observe that the structure, data types of the dimensions can be differ in and between datasets, in such datasets which are obtained with scraping preprocessing requires higher attention due to exploiting meaningful results, one simple example that can be mentioned that is to removing some of the frequent words (disregarding stopwords) can change the distribution and density of the clusters in topic analysis. For each dataset that was used in this study the data was standardized and prepared to obtain the best results possible given that a crucial part in any analysis is to have good data preparation otherwise the results can be not relevant, biased or with many outliers.

Descriptive analysis applied to the datasets can be used as a proof of how value can be also generated from UGC. From the datasets scrapped from tripadvisor it was possible to see the distribution of reviewers and their locations, however we know that these outcomes are biased due to the language barrier which forced us to scrap only English reviews. Also it was possible to create a time series and observe the high season and low season for the destinations which can be very useful information for such destinations.

In the cases we saw that most of the reviews have the positive sentiment that leads us to conclude that in such attractions tourists tend to review when they have a satisfied experience. It was interesting to see that when the analysis is done by separating the datasets regarding the sentiments we were not able to see any major differences in any aspect. Even in the negative reviews we were able to see the similar results for words, phrases and topic.
analysis, negative insights from the reviews can be captured just by analysing the complete datasets without separating them with polarity.

Word frequency, N-gram and topic analysis is applied to the datasets as NLP applications. Results of the analysis showed us that in such destinations we can create more insights with more complex analysis. With word frequency analysis we can explain the majority of the dataset in a simplified way, however for more meaningful results N-gram and Topic analysis should be used. LSA and LDA topic analysis models are used for comparison in the analysis, we can say that the LSA model was sufficient for the cases ABBA Museum and Vasa Museum. However for the cases Djurgarden, Gamla stan and City Hall LDA model was able to capture the clusters of the topic with better distribution, we believe that as a public place for Djurgarden and Gamla Stan might be the reason for showing multiple Clusters with a higher density, for the case of City Hall we saw that there are a lot of Nobel Prize related reviews in the word frequency and bi-gram analysis, we believe that this is the reason for resulting more dense clusters in the Topic analysis for this case.

As it mentioned in the section 3 it is possible to accept the results obtained from Google Trends as user generated data regarding the provided information from Google. The results from Google Trends were used as a comparison between the information regarding the queries in google about the attractions and the topics that were extracted in the analysis of this Thesis and it was possible to see that the insights obtained from the text analysis conducted are quite similar than the related queries present in Google Trends, which means that it is indeed possible to gain insights with both platforms as a first approach to understand what is relevant in any touristic point from a STD for tourists and therefore, potentially conduct a more extensive reasearch to support the decisions that might be taken regarding the tourists’ preferences of an specific destination improving the overall tourism experience.

6. Conclusion

The recent debate on Smart Tourism has highlighted the challenges and opportunities related to the management of UGC for destinations’ competitiveness. The wide adoption of social media and mobile technologies, smart devices and sensors, has contributed to the exponential growth of data. Despite their proliferation, the challenges for smart tourism destinations are linked to the capability of exploiting them for creating new value propositions and remaining competitive (Gretzel et al, 2015, Sigala et al., 2012). Thus, Big Data resulting from user-generated content, their aggregation, analysis, integration, real time synchronization and intelligent use of data has become the major drivers for the value creation process in a smart destination.

Analysis on the case studies revealed some interesting results per each attraction and more importantly showed how value can be created from UGC from the web. We were able to detect that most of the reviews in the cases has positive sentiments regarding the sentiment analysis, by looking at the date of experiences of the reviews it is possible to determine the high and low seasons for the destinations. For the cases it is crucial to detect words with high density but not stopwords, results of the analysis proved that keywords such as the name of the case or the city which in this case is Stockholm were dominating clusters in topic analysis which is making it harder to create insights. Some cases also revealed how tourists are interested in STD applications and appreciate it by reviewing and talking about it like in the case of ABBA museum, City Hall and Vasa Museum. The analysis on the case of Gamla Stan had interesting results that can be generalized for Stockholm, one of the highlighted topic in that case was about how expensive is the city.

With respect to the improvement in decision-making, the analysis of cases allows managers and promoters of the destination makers to gain insights. Monitoring of social media data can enable us to identify critical features and take immediate corrective actions for
removing problems that may decrease the value of a destination.

As for the new business models and new products/services development cross-case analysis demonstrates that digital experiences can be considered drivers for the smart development of a destination. Relevant information for the industry that can be used to planify different strategies to foment the tourism given that the UGC gives a broader view of what is really interesting for the tourists. Some of the findings of this research can be shown as a proof of appreciation of the technology in tourism destinations when it exists.

For researchers, implications in this research can be used as an entry point for collecting and analyzing Big Data, developing integrated and analytical frameworks allows extracting the useful information from Big Data coherently with the different value dimensions. Findings offered evidence about the different dimensions of value that can be created from social media Big Data.

As a final conclusion, it is possible to say that there is a vast amount of relevant information that can be extracted from User Generated Content from sources like TripAdvisor and Google Trends like date of experience of the tourists in an attraction, what are the most interesting topics for the tourists regarding a place to visit, what else does the tourists that goes to one attraction show interest in the web, which are the seasons that the tourists visit an specific touristic place, what are the most relevant topics per polarity, among others. All this information can be crucial for the tourism industry given that it gives real information of what the visitors are actually talking about regarding an STD and their touristic points, which leads to the possibility of use this new relevant information to create insights that can help them to understand better their customers and reduce the uncertainty during the decision-making process when they have to make important decisions in order to create value. As it mentioned in section 2 there are little work done in order to generate insights from the feedbacks of the tourists and more importantly those studies did not use many different NLP techniques. What we provide in this research was a possible model and techniques to achive this goal with an acceptable results to gain insights. We believe that those models can be used generally in the field of Tourism and it is open for improvements.

6.1 Limitations and Future Research of The Study

With regard to the complexity and the broadness of this research, limitations of this study should be mentioned to fill some of the gaps.

Tourism is such an industry that the relationship with languages is one of the first things to consider. Therefore, the combination of NLP and the tourism industry comes with a big challenge. Even though it is possible to explain more than 50% of the existing information just by analysing the English reviews, regarding the cultural differences exists and can only be exploited by analysing the different languages is still unpredictable for us. Thus, a research into different languages on this topic could reveal more interesting results.

The choice of destinations in this case study research is focused on the most famous attractions in Stockholm. Even though attractions are very important extensions of the tourism industry, the characteristics of different modules such as restaurants, hotels and etc. of tourism can be different. We believe that future research considering the other factors and comparison of them can explain more about the tourist and their behaviour in a more complex way.
REFERENCES


Wen, T. H., Kung, P. H., & Su, P. H. Analysis of combining Topic model, Sentiment, Geolocation information approaches on Social Network. Massachusetts Institute of Technology, Cambridge, MA

Wenxue, L. I. U., & Sun, G. (2013). A trust-based information propagation model in online social networks. The FARMS Review Volume 8, Number 8, August 2013, 8(8), 1767-1773.