MASTER THESIS ECONOMICS AND BUSINESS ECONOMICS:

DATA SCIENCE AND MARKETING ANALYTICS

Developing and implementing a sentence-based LDA model to uncover topics in text data: an application to Disneyland theme park.

Author:

Regine Leito

388178rl

Supervisor:

Prof. dr. Bas A.C.D. Donkers

Second Assessor:

Dr. Kathrin Gruber

July 31, 2021

Abstract

In this research, an attempt is made to apply natural language processing techniques in an important yet overlooked industry when it comes to machine learning, namely the tourism industry. More specifically, hidden topics from Disneyland reviews are uncovered and then mapped to aspects. The sentiment of reviewers with regards to these aspects is then analyzed. To uncover topics, this research employs Latent Dirichlet Allocation (LDA) on both document and sentence level. The two models are compared in terms of interpretability of topics. Furthermore, Aspect Based Sentiment Analysis is performed to explore sentiment of reviewers towards aspects. The results shows that sentence-based LDA model provides more specific and interpretable topics compared to document-based LDA model. Moreover, popular aspects discussed by reviewers are rides, food, lines, staff, fireworks, prices and crowds.

Table of Contents

1 Introduction

“Disneyland is the happiest place on earth!”. The Impact of Disneyland on the entire world is undoubtedly big. Its three parks located in California, Paris, and Hong Kong; all appear on the top 25 list of largest amusement parks worldwide (Consultancy, 2016). Kids from all around the world happily sing along to Disney songs and want to meet the characters from Disney movies walking around the Disneyland parks. The impact of Disneyland is not only big on its consumers but also in the business world, as the entire Disney brand receives much respect for the way it runs it businesses. In fact, Fortune magazine, one of the top business magazines worldwide, ranked the Walt Disney Company at fourth place on the “World’s Most Admired Companies” list for the third consecutive year in 2021 (Shuler, 2021). Therefore, this research aims to analyze what visitors of the three Disneyland branches find important considering their experience at the parks using written online reviews. Surprisingly, not a lot of research about this has been done using machine learning techniques. Machine learning is the usage of computational algorithms and statistical models to draw inferences from patterns in data for decision-making (Jordan & Mitchell, 2015). In an age where one of the first things one does before purchasing a product or service online is read its reviews, it is particularly important for companies to know exactly what their customers feel and what may need improvement (Feldman, 2013). With the advancements in recent years in the natural language processing (NLP) area, analyzing large bodies of text has become more feasible.

Briefly, NLP is a field in machine learning that focuses on creating algorithms that will help computers understand and process written text (Chowdhury, 2003). It can be considered a powerful tool that filters large amounts of text and extracts high-value information. One example is the analysis of the words being used in data to discover topics (Büschken & Allenby,2016). Using NLP for online reviews, a good understanding of consumers perspective on product and service can be achieved. This, in turn, is very important for improving decision making, effective management and marketing.

The application of NLP techniques on online customer reviews is popular for businesses across a variety of domains. For example, countless analyses using reviews have been done by companies such as Netflix and Amazon but also different hotels, restaurants and beers brands (Bennett & Lanning, 2007; Ganu, et al., 2009; McAuley et al., 2012; Büschken & Allenby,2016). In the Tourism domain, however, specifically in terms of theme parks, the analysis of reviews has largely been overlooked. In this field, managers should understand how visitors have experienced the park and analyze the behavior they portray while

at the park to effectively run marketing campaigns and coordinate daily schedules (Kemperman, 2000; Fotiadis, 2016). Therefore, this study offers insight into the experience of Disneyland visitors by using machine learning techniques.

Most of the previous studies done on theme parks covered visitor behavior in terms of their expectations and participation and were conducted through interviewing or surveying visitors (Chen & Lamberti, 2013; Cabanas, 2020). These types of surveys limit respondents to preselected items, available response items and force them to use rating scales (Büschken & Allenby, 2016). Nowadays consumers voluntarily share their experiences in the form of an online review. These online reviews contain different information, as consumers are free to express their experiences without limitations/constriction, in contrast to surveys information where the topics tend to be predetermined. The challenge, however, lies in assessing what topics are being discussed in reviews. The statistical models used in natural language processing to uncover topics in written text are part of a class of models called topic models (Blei & Lafferty, 2007).

Although topic modeling has been around for quite some time already, it has only been employed in the tourism industry just recently and has been demonstrated to be a very capable way for exploring the thoughts and behavior of visitors of theme parks (Hofmann, 2001; Luo et al., 2020). Topic models are typically performed at document (review) level and try to find words that co-occur in each document (review). Some studies argue, however, that when people write reviews, they naturally bring structure to what they want to say by forming sentences (Bao & Datta, 2014; Büschken & Allenby,2016). This idea was formally introduced to topic modeling as a sentence-base topic model. The idea behind this model is that the lack of structure in regular topic modeling does not provide word combinations with good discriminating ability. According to Bao & Datta, using sentence-based topic modeling, more meaningful topics will be attained. As the goal of this research is to understand what topics summarize the experience of Disneyland visitors, topic interpretability is crucial. For this reason, it is relevant to see whether a sentence-base topic model provides more coherent, meaningful, and interpretable topics than a regular topic model.

Of course, it is one thing to discover which topics are discussed within the Disneyland reviews by visitors, but it is also useful to understand their sentiment towards these topics. Take iPhone X reviews for example. If customers mention the quality of the device often, it is good for Apple to know that customers attribute high value to this feature. It is even better for Apple to know, however, whether

customers feel positively or negatively towards this feature. This way, they know whether they should invest in improving the quality of the iPhone X or whether to allocate their funds elsewhere. The same goes for Disneyland and leads us to the following research question:

"What topics are discussed in the Disneyland reviews and how do visitors feel about these topics?"

To answer this question, the following sub-questions have been formulated:

1. What are the most popular topics discussed in Disneyland Reviews?
2. Does a sentence-based topic modeling provide better interpretation of topic discussed in Disneyland reviews as opposed to a regular topic model?
3. What sentiment do visitors feel towards the aspects expressed in the most popular topics?

The remainder of this research is structured as follows; chapter two covers the literature review surrounding theme parks, topic models and sentiment analysis and delineates the academic relevance of this study. This chapter will provide critical analysis of existing literature surrounding the employment of topic modeling in the context of online reviews as well. Chapter three covers the dataset used for the purpose of this research and includes exploratory data analysis. Chapter four covers the experimental setup and technical details of the methods that will be used in this research. The results of the analysis will be presented and discussed in chapter five and the conclusion and limitations of this research will be discussed in chapter six.

2 Literature review

It is widely accepted that 1966 is the starting year of the Theme Park Industry due to the opening of the first Disneyland in Anaheim, California in that year (Clavé, 2007, p. 3). This consideration however is arbitrary as the origins of theme parks can be traced back to European fairs of medieval origin. Additionally, other parks that fit the modern-day mold of being theme parks had already opened their doors by that time. An example would be Efteling, a Dutch Fairytale Forest with a playground, that opened its doors in 1952 (van Assendelft de Coningh, 1995). Despite all of this, upon its opening, Disneyland quickly captivated the United States, much of the Western world and even beyond and was met with great enthusiasm, emulation and envy by large corporations (Watts, 1995).

As previously mentioned in the introduction, studies using NLP on theme parks are scarce. Recently, a study was done showing the effectiveness of implementing topic modeling to Disneyland reviews to better understand the behavior of individuals who visit the park (Luo, Vu, Li & Law, 2020). The study found a wide range of topics and further focused on how visitors from different countries rate the Disneyland parks (Luo, Vu, Li & Law, 2020). In this study, it was found that Disneyland California tends to have the highest rating for all countries except for visitors from India. For India, Disneyland Hong Kong scored well above Disneyland California in terms of ratings. Remarkably, however, the Hong Kong traveling group rated Disneyland Hong Kong the lowest out of the three Disney parks. The result of the analysis suggests that satisfaction levels, studied through rating behavior, differs across visitor from different countries (Luo, Vu, Li & Law, 2020). In the study, LDA topic modeling is performed at document level to find topics discussed in the Disneyland reviews. Our research aims to expand the topic modeling analysis by comparing the topics found using document-based topic modeling with topics found through sentence-based topic modeling to see which model provides more interpretable topics. Furthermore, perception of different aspects of the park is also analyzed in the study by examining ratings. This approach, however, may not be the best as it assumes that whole reviews cover only one aspect of the parks, but this is often not the case. Ratings given by the reviewer cover different aspects at the same time, making it impossible to analyze perceptions towards each aspect separately. For this reason, our research employs another approach, explained in section 2.2.4, for the analysis of aspects.

2.1 Topic models

Topic modeling is a statistical tool that automatically extracts hidden topics, or latent variables, from large datasets. These topics are inferred from the way different words co-occur in a document,

irrespective of the order of words in the document. This is also known as the bag-of-words (BOW) method. BOW can also be modified to represent the co-occurrence of groups of words. This is referred to as an N-gram. Topic modeling is part of unsupervised learning, meaning that the input data does not contain labels nor is a target variable being predicted. The most common topic modeling techniques are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Non-negative Matrix Factorization. These four techniques will be discussed in the following sections.

2.1.1. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is one of the foundational approaches in topic modeling and it focuses on dimensionality reduction. This is done by first creating a document-term matrix (DTM) where rows represent documents of the data, columns represent unique words, and the values of the matrix are occurrence frequencies. Then, Singular Value Decomposition (SVD) is performed on this matrix (Deerwester et al., 1990). Text documents are represented as points in a Euclidean space, also known as vectors. SVD is performed on the DTM to reduce dimensionality by encoding it with latent features, which represent the topics in the original dataset. LSA received some critique due to its inability to capture the different meanings of polysemy, making topics difficult to interpret (Hofmann, 2001). This is the result of words being represented as a single point in a space. In fact, Hofmann found the idea to use Singular Value Decomposition to approximate matrices of word counts somewhat ad hoc.

2.1.2. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) stems directly from LSA. It is a generative model and the main difference with LSA is that it uses a probabilistic method instead of SVD, allowing for statistical techniques to be applied for decision making regarding the choice of models, fitting models, and complexity control. This type of modeling assumes that there is an interaction between observed and unobserved, or latent parameters, in a probabilistic way (Vayansky &Kumar, 2020). Furthermore, the method has the ability to handle polysemy as it associates a latent context variable with each word occurrence (Hofmann, 2001). Despite its ability to recognize that one document can contain multiple topics, it can only learn these topics on the data it has been trained on (Blei et al., 2003). In other words, document probabilities are fixed points in the data and the model does not recognize data points in unseen data, making PLSA less flexible to unseen data. Another critique is that this model is prone to overfitting (Popescul et al., 2013; Vayansky &Kumar, 2020).

2.1.3. Latent Dirichlet Allocation

LDA is considered a Bayesian version of PLSA. It is also a generative probability-based approach. The difference is that LDA probability distribution relies on a Dirichlet distribution. This allows LDA to generalize to unseen data more easily, making it perform better than PLSA often (Aznag et al, 2013;). The main idea of LDA is that documents contain latent topics and that single words are only indicators of these latent topics (Lau et al., 2014). These latent topics consist of multinomial distribution of words. This means latent topics are not comprised by a single word but rather a collection of words with high probability of being used (Büschken & Allenby,2016). The output of the model consists of the top-N words with highest probabilities. LDA is the most common topic model and has been adapted further to address different issues (Blei & Lafferty, 2007; Titov & McDonald, 2008; Qiu et al., 2013). One example would be the model developed by Blei & Lafferty named correlated topic model (CTM). The difference between LDA and CTM lies in the distribution used to portray the variability among topic distribution. Where LDA used the Dirichlet distribution to model this variability, CTM exhibits correlation between topics via a logistic normal distribution (Blei & Lafferty, 2007). Behavior-integrated topic model (B-LDA) is another variant of LDA where topic interest and interactions of a user with these topics are modelled at the same time (Qiu et al, 2013). These variants add to the original LDA methodology by either focusing on textual content as the subject of topic modeling or looking at the context in which text is generated (Qiu et al., 2013). Critiques on LDA topic models are that the Bayesian priors cannot be justified, and the number of topics cannot be chosen among other things (Gerlach et al., 2018).

2.1.4. Non-negative matrix factorization

The difference between Non-Negative Matrix Factorization and the other methods discussed is that it constraints the document-term and the topic-term matrices to be nonnegative (Paatero & Tapper, 1994). This constrain makes the vector in the Euclidean space purely additive, in contrast to the other methods such as LSA, PLSA and LDA. Because of this nonnegative constrain, NMF produces very sparse matrices allowing for few ‘active components’ which allow for easier interpretation of the output. The NMF has since been extended further in many ways, one being the inclusion of control over this sparseness (Hoyer, 2004). Another adaption was done by Lee & Seung (2001) as the initial NMF model was not able to find the global minimum. They included an iterative update algorithm to the model allowing NMF to find a local minimum of the objective function.

2.1.5 Sentence-based topic modeling

As mentioned in the introduction, it has been researched whether bringing structure to topic modeling by using sentences, rather than documents, as unit of analysis would provide more meaningful topics. Several topic modeling variants have been introduced to include this notion. One variant, called Multi-grain (MG-LDA) introduced the sentence-wide topic proportions that allow discovery of what they call local topics, next to the topics found using document-wide topic proportions with standard LDA (Titov & McDonald, 2008). In their paper, they wrote that simply splitting reviews into sentences would not allow for enough words to co-occur. They were wrong however, as one other variant of sentence-based LDA that does exactly that, called local-LDA, was found to outperform the MG-LDA model (Lu et al., 2011). Local-LDA was introduced by Brody & Elhadad (2010) where they extended the LDA model by simply treating each sentence as unit of analysis. The idea behind this is that using sentences as unit of analysis brings structure into the model by not allowing all words in a review to co-occur, as is the case with BOW, but only those in the same sentence. Bao & Datta (2016) later added to this model by introducing a one topic per sentence assumption. They believed that although standard unsupervised topic modeling did a good job at discovering topics within text, these topics may not be meaningful or specific enough to the context. The study in the paper was done on textual corporate risk disclosure data where they tried to use topic modeling to uncover risk types, but they claim that applying sentence-based topic modeling on customer reviews should provide improved topics. The results of all studies above showed that topics found with sentences as unit of analysis were more meaningful.

2.1.6 Evaluation

The aim of evaluation measures of topic models is to determine the interpretability of topics, also referred to as human-interpretability of topics (Lau et al., 2014). The two main ways of evaluating topic models are extrinsic or intrinsic in nature. Extrinsic methods are so-called real-life evaluations where the model is put in use and monitored for its performance. These types of methods usually require a longer period to be monitored and are therefore time-consuming. Intrinsic model evaluation uses a sample data set. This type of evaluation is most preferred especially in cases where time is limited. Some popular intrinsic methodologies will be discussed next.

The most popular way to evaluate topic models is the held-out likelihood method. In this method, the topic model is first trained on one part of the dataset and evaluated, typically using perplexity, on the other unseen part of the data (Wallach et al., 2009). This notion was challenged, however, and it was

found that topic interpretability correlates negatively with perplexity scores (Chang et al., 2009). In an attempt to evaluate the human-interpretability of topics, different approaches were explored. Topic coherence was introduced as a measure that focuses on the qualitative understanding of the semantic nature of the topics (Newman et al., 2010). The idea behind it is that top models can provide good, coherent topics, but they can also provide meaningless topics that are hard to interpret. A topic is generally assumed to be coherent if most of its top ten words are related. Various variations of topic coherence exist, including coherence based on pairwise pointwise mutual information (PMI), normalized PMI and log conditional probability (Mimno et al., 2011; Bouma, 2009; Newman et al., 2010).

Next to topic coherence, topic similarity is also used for topic model evaluation. The difference between word distribution of topics can be used to measure this topic similarity (Li & McCallum, 2006). Topic similarity is used in situations where finding similar topics is the goal, for example, when recommendations need to be made. It is also useful when finding distinguishable topics is the purpose of the research. Common measures of difference between two probability distributions are the Kullback-Leibler divergence (KLD) and the Jensen-Shannon divergence (JSD) where KLD is asymmetric and JDK is symmetric.

2.2 Sentiment Analysis

Sentiment analysis (SA) is the usage of techniques and methods to textual data to extract information about the opinion and emotional intent of an author (Kwartler, 2017, p. 85; Feldman, 2013). It is commonly used on reviews of consumer products and services and provides a sentiment score on review level, as well as analyzes the sentiment of individual aspects of the product or service (Feldman, 2012). This field in science has become increasingly more popular these last few years. The data shows that approximately 7000 papers on sentiment analysis have been published and 99% of this was published after the year 2004 (Mäntylä et al., 2018; Feldman, 2013). SA is typically used for analyzing reviews for products and services online. Companies like to know the opinion and sentiment their customers hold for their products. Consequently, SA is also referred to as opinion mining (Mäntylä et al., 2018). Using SA, companies can monitor their online reputation, the sentiments of their customers and receive timely feedback on their product and services (Feldman, 2013).

2.2.1 Approaches to Sentiment Analysis

In practice, sentiment analysis can be applied in a variety of ways, each with its benefits and shortcomings. It can be applied e.g., at document or sentence level, or it could be applied to specific

aspects/features in a corpus of documents. Additionally, there is also comparative sentiment analysis and sentiment lexicon acquisition (Feldman, 2013). At the core of all these types of implementations of SA, lies either one of two types of foundational techniques. These are BOW and syntactic parsing techniques. In SA, the BOW method treats every single word, or a group of words called n-grams, as a unique token of a document (Kwartler, 2017). In this approach, the order of words and grammatical words are disregarded. This immediately showcases some of the shortcomings of the BOW approach. Grammatical words i.e., signal the relationship between sentences amongst other things and without them, a bit of context is lost. Nonetheless, much research has been done on the accuracy levels of the BOW method and it was found that good accuracy levels are achieved even using the simplest SA approach (Feldman, 2013). The difference between BOW and syntactic parsing is that the latter can identify grammatical aspects of words such as nouns, articles, verbs, and adjectives (Kwartler, 2017). Studies have shown that accuracy levels are high using POS. The only downside to this method is that it is far more complex and currently demands high computational costs (Gómez-Rodríguez et al., 2019).

2.2.2 Sentiment Analysis at document level

The most basic application of SA is at the document level. In this type of analysis, sentiments scores are calculated using whole documents e.g., an entire review by a single author. Furthermore, document-level sentiment analysis can be approached through supervised or unsupervised learning. In the supervised learning approach, there are a set of classes, for example, positive or negative, to which each individual document in the corpus belongs to. Features, typically words, are then extracted from the data and fed into a classification algorithm such as Naïve Bayes or Logistic regression to tag documents to the correct sentiment class. In cases where reviews are rated based on the standard five-point rating scale, the reviews are first re-labeled as being either negative, positive or neutral.

In the unsupervised approach to SA, sentiment scores based on some predefined threshold are used to classify documents as being either positive or negative and sometimes even neutral. The assignment of sentiment scores to sentences or documents can be done using a predefined list of tagged words, called lexicon, or it can be done based on a set of predefined part of speech (POS) patterns. The words in a lexicon are tagged as a polarity or an emotion. Polarity indicates whether the orientation of an expressed sentiment is either positive, negative, or neutral. For example, a word that corresponds to positive sentiment, such as the word “good”, will increase the polarity of a document or a sentence, making it more positive. Likewise, a word that corresponds to a negative sentiment, such as “horrible”, will cause a decrease in the final sentiment score, making the polarity more negative. Next to positive,

negative, and neutral words, valence shifters influence the overall polarity of a document or sentence as well (Kwartler, 2017). The two types of valence shifters are negations and (d)amplifiers words. A negation word is a word that reverses the intent of a positive or negative word. For example, in the sentence “Popcorn with butter is not good”, the word “good” has a positive intent but the word “not” is a negation, turning the polarity of the sentence into a negative intent. Amplifiers and deamplifiers increase the intensity of the intent of a positive or negative word. For example, the sentence “Popcorn with butter is very good” contains the amplifier word “very”, which gives the sentence a stronger positive tone than if it were to use only the word “good”. The final sentiment score of a document or sentence is computed taking all scores assigned to said polarity, negation and (d)amplifier words into account. This final sentiment score indicates whether the document or sentence is positive, negative, or neutral. In the case of an emotion lexicon, the final sentiment scores indicate which emotions are present in the text.

2.2.3 Sentiment Analysis at sentence level

As previously mentioned, next to document-level, SA can be applied at a sentence level for a more fine-grained approach. This means that either the entire review gets a sentiment score or each sentence in a review gets its own sentiment score. In this approach, the BOW is constraint to sentences, also called Bag-of-Sentences (BOS). Studies have found that this approach improves accuracy levels compared to SA on document level, as BOS allows for focus on contextual information (Khan et al., 2010).

2.2.4 Aspect Based Sentiment Analysis

The two approaches to SA discussed above work best when the entire document or each sentence refers to a single aspect (Feldman, 2013). However, a quick look at a review will show that reviews consist of feedback on more than one aspect. Simply scoring reviews on a document or sentence level ignores valuable information on specific features. Therefore, document and sentence level sentiments do not provide valuable information for decision making (Peñalver-Martinez et al., 2014; Madhoushi, 2019; Mowlaei et al., 2020). This led to the introduction of aspects-based sentiment analysis (ABSA), where the sentiment related to an aspect within a document or sentence is explored (Hu & Liu, 2004; Liu, 2012). The idea behind ABSA is that even in cases where a document or a sentence covers one single entity, take restaurants at Disneyland as an example, the sentiment score of that document or sentence does not necessarily reflect how an author feels about aspects of that entity. Therefore, to perform

aspect-based sentiment analysis, aspects need to be determined. In this research, it is opted to perform ABSA. Further details on the implementation of this method will be discussed in chapter 4.

2.2.5 Sentiment Analysis domains

SA is used in several other domains, other than product reviews. One example is the implementation of sentiment analysis in the financial markets where the goal of the analysis is to use investors' opinion and sentiment towards stocks, for example, to forecast future changes in stock prices (Yu et al., 2013; Li et al., 2014). Another example is the political field. Nowadays, increasingly more people express their political opinions online. Studies have shown that thanks to this trend, and despite internet users not being a good representation of an entire population, SA applied to social media shows great potential to predict electoral results (Ceron et al., 2014).

2.2.6 Challenges of Sentiment Analysis

When performing SA, things such as sarcasm and irony are difficult to detect. Irony is one of the most subtle ways to deny what is being said and relies on opposition. The problem arises due to this opposition not being tagged in one way or another. In other words, there is no explicit negation word that is present in an ironic sentence (Reyes & Rosso, 2014). This makes it harder to be detected in written text and sometimes even in spoken dialogue. Though studies have tried to find solutions for this problem, detecting sarcasm and irony remain a difficult task. In their paper, Davidov et al. (2010) developed an algorithm to identify sarcastic sentences in product reviews that achieved a precision of 77% and recall of 83.1%. The method is not straightforward and has therefore not caught on. Despite the many attempts to study irony detection in text, it remains hard to detect due to its subjective nature (Reyes & Rosso, 2014). Another example of a challenge when performing SA is the ambiguity of words. The word "cheap" in "A cheap meal" for example, is considered a good thing while in the sentence "The material is cheap", it is not.

3 Methodology

In this study, we implement topic modeling and SA to answer the research question. To uncover topics in the Disneyland data, it was chosen to perform LDA at document and sentence level. LDA is chosen due to its positive track record and reliability. The topics uncovered using the two LDA models will be compared in terms of interpretability and coherence. Next, aspects will be mapped using the topics that are uncovered and to which ABSA is then performed on. ABSA is chosen as this type of SA is found to provide more useful information for decision-making purposes than a regular document- or sentence-based SA. To calculate sentiment scores, a polarity lexicon will be used because studies show that polarity scores are more accurate than SA performed on emotional frameworks (Kwartler, 2017).

In this section, technical details of methods and algorithms used for the analysis are described. Section 3.1 discusses the technical details of the LDA methods, section 3.2 presents the algorithm used to extract aspects from topics and lastly, section 3.3 discusses the implementation of SA.

3.1 LDA model

LDA is in principle a generative model based on probabilities and distributions and does topic generation based on word frequencies (Blei et al., 2003). The assumption behind LDA is that documents are made up of a mixture of topics, and the topic are made up of a collection of words. The documents are represented as BOW and high-order co-occurrence between words is captured. The idea is that words that often appear together within documents belong to the same topic. The output of the model includes probabilities of topics appearing in a document and probabilities of each word appearing in a topic. LDA is commonly explained through its generative process. At the basis of the model lies the idea that a collection of documents, or corpus, is made up of a random mixture of latent topics. The generative process for each document i in corpus C is summarized by the following formulas:

$$\theta_n \sim Dir(\alpha \dots \alpha),$$

where θ_n is a vector with topic probabilities for document n and α is the fixed parameter governing the prior distribution of θ_n.

$$\beta_k \sim Dir(\delta \dots \delta),$$

where β_k is a vector with word probabilities for topic k and δ is the fixed parameter governing the prior distribution of β_k.

For every word w in document i:

$$z_{wn} \sim Multinomial(\theta_n),$$

where a topic k is selected for each word i based on a multinomial distribution.

$$w_{wn} \sim Multinomial(\beta_{z_{wn}})$$

where the word w is selected given topic k and the word distribution belonging to topic k.

Hyperparameters are represented by α and δ, they need to be selected a priori and are usually both symmetric. This symmetry ensures that in the β_k distribution, all words have the same likelihood of appearing in a topic and in the θ_n distribution, all topics have the same likelihood of appearing in a document. α and δ control the sparseness of topic probabilities and word probabilities, respectfully. Low values of α and δ will produce documents that consist of one or very few topics and topics that consist of very few or one predominant word. The opposite is true for higher values of α and δ. Next to the hyperparameters discussed above, the number of topics K needs to be chosen a-priori as well.

The methods typically used to approximate the posterior multivariate distributions are variational Expectation-Maximization (VEM) (Minka & Lafferty, 2002) and collapsed Gibbs sampling (Griffiths & Steyvers, 2004). Using VEM, the hyperparameters are approximated through maximizing expectation while collapsed Gibbs sampling is a Markov Chain Monte Carlo technique that uses a stochastic process that computes and updates parameters (Griffiths & Steyvers, 2004). There have been studies done to determine which of the two methods work best. It was found that Gibbs sampling leads to stabler models with lower perplexity scores (Layman et al., 2016). In a recent study, it was found that VEM is faster and more efficient at performing tasks (Kim, 2020). In this research, however, it was opted for the more stable method Gibbs sampling.

Sentence-based LDA

In this research, sentence unit of analysis is introduced by first splitting al reviews in the dataset into sentences. We will name this model SLDA model while the standard LDA model will remain LDA. This approach to sentence based LDA is simple and straightforward yet has been proven to discover meaningful topics using custumer review data (Brody & Elhadad, 2010).

3.1.1 Evaluation

The hold-out method will be used to find the right value for the model parameters. The parameters are firstly trained using one part of the dataset and then the performance of the model will be evaluated on the other unseen part of the dataset. For this purpose, the dataset will be split into a training and validation set of 80% and 20%, respectively. To evaluate the estimated models, the perplexity metric is used. Although perplexity was found to focus more on the complexity of a model rather than the human-interpretability, it is still good to use it to assess the ability of the models to make predictions (Chang et al, 2009). Next to perplexity, the topic similarity measure will also be used as we are interested in finding distinguishable and interpretable topics. Lastly, human-interpretation will also be used to evaluate the models.

Perplexity Measurement

To determine how well the estimated models fit the data sample in this analysis, perplexity is used. In natural language processing, this goodness of fit is determined by using the estimated probabilities of all words appearing in a set of documents. In principle, a model will assign higher probabilities when the degree of certainty is higher. In other words, the perplexity measure will try to capture how well a model is able to assign high probabilities in the prediction of a test set. This is calculated by the following formula:

$$PP(W) = P(w_1, w_2 \dots w_N)^{-\frac{1}{N}},$$

where PP is the perplexity of model W, P is the probability distribution of all words w occurring in a set of documents with N number of words. P is calculated by taking the log of the probabilities of all the words appearing in a document and dividing it by the total number of words. Diving by N is done to normalize for the length of a document. Longer documents are more likely to be less probable as they are more difficult to predict. This normalization step is crucial to be able to compare test sets of different lengths.

The lower the value of perplexity, the better the model is at predicting the sample. This could be explained intuitively by looking at the meaning of the word “perplex”. This word means to be “confused”. The least confused you are, the better. The perplexity measure will be used for comparing different models with varying levels of K and α with each other.

Topic Similarity Visualization

In their study, Kim and Oh (2011), compared six measures of distribution similarity between words obtained from pairs of topics. They found that Jensen Shannon Divergence (JSD) performed best at estimating topic similarity. For this reason, JSD will be used to analyze topic similarity in this research. The JSD similarity measure uses the per-topic word distributions. To visualize this similarity, Principal Components Analysis (PCA) will be used for dimension reduction. The output will be a two-dimensional plot where the distance between topics and topic relevance will be visible.

JSD is calculated by averaging over the Kullback-Leibler (KL) divergence of each distribution to the average of the two distributions (Hall, Jurafsky & Manning, 2008). The JSD between two distributions is computed through the following formula:

$$D_{JS}(P \mid Q) = \frac{1}{2} D_{KL}(P \mid M) + \frac{1}{2} D_{KL}(Q \mid M)$$

$$D_{KL}(P \mid Q) = \sum P(i) log_2 \left(\frac{P(i)}{Q(i)}\right),$$

where D_{JS} is the distance between probability distribution $P$and Q, M is the mean distribution of $P$and Q and D_{KL} is the KL divergence. The D_{JS} is exactly zero when two distributions are identical and increases as the distance becomes bigger. With this JSD distance measure, the similarity between two topics is calculated using the following formula:

$$Sim_{D_{JS}}(k_i k_j) = \begin{cases} 1 - D_{JS}(k_i \mid k_j) & if\ D_{JS}(k_i \mid k_j) \neq 0 \\ 0 & \end{cases},$$

where $Sim_{JSD}(k_i k_j)$ is the similarity between topics k_i and k_j.

3.2 Topic Aspect mapping

Once topics are uncovered using LDA, each topic is mapped to the aspect(s) described inside of the topics. The Disneyland corpus then goes through Algorithm 1 where aspect specific sentences are extracted and stored into the concerned data frame. Algorithm 1 is described below.

Algorithm 1: Extracting aspect specific sentences from corpus

Data: Disneyland Reviews
Results: Aspect specific sentences
1 Initialization;
2 Pre-processed Review corpus ← Stopword Removal & Symbol removal
3 Topics ← LDA topics
4 Aspects ←Disneyland aspects (Pre-processed Review corpus)
5 **for** each "Topic" do
6 Topic Aspect Mapping
7 **end**
8 for each document in corpus do
9 | Splitting review text into sentences
10 **end**
11 **for** each "Aspect" do
12 | Scan the Review text lines
13 | **if** sentence contains *topic words* **then**
14 | | Data frame <--Write line;
15 | **else**
16 | | Skip line
17 | **end**
18 **end**

Steps 1 through 4 in Algorithm 1 cover data pre-processing and topic extraction using LDA topic modeling. Then, aspects are mapped in steps 5,6 and 7. After this step, the Disneyland review corpus is split into sentences in steps 8,9 and 10 in order to perform ABSA at sentence level. Next,

data frames are created for each aspect mapped. In steps 11 through 17, all sentences in the corpus are scanned for topic words that appear in the topics from which aspects were mapped and are assigned to the correct aspect data frame. Sentences that do not contain topic words are skipped and are not considered for ABSA.

The resulting data sets will each concern one aspect. Sentiment analysis is then performed on each data set individually. The resulting polarity scores per data set will indicate the sentiment towards that aspect. The next section will cover the process of calculating sentiments.

3.3 Sentiment Analysis

The polarity function used in this research is based on a subjectivity lexicon. In this case, the list consists of words that are linked to an emotional state. The emotional states can be either positive or negative states. The lexicon used in this research is by Bing Liu and contains 6789 words. These words have been inspected and have been compiled together through extensive academic research.

Computing the polarity of a document is a straightforward process of summation of positive words in a document and subtracting the negative ones. Positive words are a signed a value of one while negative words are assigned a value of negative one. Next to assigning value to positive and negative words corresponding to the subjectivity lexicon, words that shift the valence of a sentence also get assigned a value. As previously explained, the two types of valence shifters are negations and (d)amplifiers words. The impact of a negation word, a word that reverses the intent of a positive or negative word, on the final sentiment score will be explained through the following example. In the sentence “Popcorn with butter is not good”, the word “good” has a positive intent but the word “not” is a negation, turning the polarity of the sentence into a negative intent. The polarity score of the sentence is computed as: not (-0,8) + good (1) = -0,2. In de case of (d)amplifiers the following example is given. The sentence “Popcorn with butter is very good” contains the amplifier word “very”, which gives the sentence a stronger positive tone than if it were to use only the word “good”. The polarity of this sentence is computed as: good (1) + very (0,8) = 1,8.

Next to the subjectivity lexicon, the polarity function used in this research also contains lists of words for negators, amplifiers and deamplifiers. The two examples above showed a rather simplified version of computing the polarity scores with the inclusion of valence shifters.

The overall process of computing the polarity score for text sentences in this research is slightly more complex and is described in the five steps below:

1. The polarity algorithm searches for positive and negative words in the subjectivity lexicon that also appears in the text.
2. Once a word is found that that appears in both the subjectivity lexicon and the text being polarized, referred to as a polarized word, a cluster of terms is constructed. This cluster includes four words preceding and two words following the polarized word. The words can be considered as valence shifters.
3. Positive and negative words within the cluster are assigned values one and negative one, respectfully. Neutral words are assigned a value zero and the remaining words are considered to be valence shifters. The weight assigned to amplifiers and negations is 0,8. Meaning, an amplifier adds 0,8, while a deamplifier and a negation subtract 0,8.
4. The values assigned in the previous step are summarized.
5. The final polarity score is computed by dividing the summation in step four by all the square root of all words in the text to account for the density of the keywords.

4 Data

The dataset used in this research consists of reviews by visitors from across the world for all three Disneyland branches. The data is obtained from Trip Advisor and contains a total of 46,354 reviews after duplicates removal. Only reviews written in the English language will be used in this analysis due to language barrier. It must be noted however, that certain reviews written in English came from visitor who were born in countries where English is not the mother language. A total of 2,409 rows in the data contain missing values for the variable that represents the date of writing the review. This fact will not impact the analysis, as the analysis is focused on the textual data.

The reviews in the dataset were posted by Disneyland visitors between October 2010 till Feb 2021. Due to the worldwide pandemic of Corona, all the Disneyland parks closed their doors for visitor. Consequently, no review made after March first, 2021, will be considered for this analysis.

The dataset contains a total of 6 variables. These variables are Review_ID, Rating, Date, Reviewer_Location_Review and *Disneyland_Branch*. The *Review_ID* variable represents a unique identification number for each review posted, regardless of user. This means that the dataset may contain more than one review per reviewer, however this information cannot be traced. The R*ating* variables indicates how the reviewer has rated their visit to the respective Disneyland branch on a scale ranging from 1 to 5, with 1 being unsatisfactory and 5 being satisfactory. The variable *Date* indicates the month in which a review was posted while the variable *Reviewer_Location* indicates country of residency of the reviewer. The final two variables are *Review* and *Disneyland_Branch* and they contain the text review and the three Disneyland branches visited by the reviewer, respectively.

Originally, variable *Reviewer_Location* containsa wide range of demographic background, 162 different countries to be exact. To maintain a manageable overview, and because they amount to a combined share of 86% of all reviews, this research will focus only on the top 10 countries with most reviews. These countries will be called traveler groups moving forward and are United States, United Kingdom, Autralia, Canada, India, Philippines, Singapore, New Zealand, Malaysia, and Hong Kong. This leaves a total of 39,400 reviews in the dataset. Figure 1 displays the frequency distribution of reviews per traveler group for Disneyland California, Disneyland Hong Kong and Disneyland Paris, respectively. From this figure it is apparent that most reviews are posted by the residents of the United States and United Kingdom. The pie chart in Figure 2 gives an insight into the review shares per traveler group.

In terms of average rating, the Disneyland California seems to be rated highest by visitors with an average rate of 4.41, followed by Disneyland Hong Kong with an average rate of 4.24 and then finally Disneyland Paris with an average rate of 3.96.

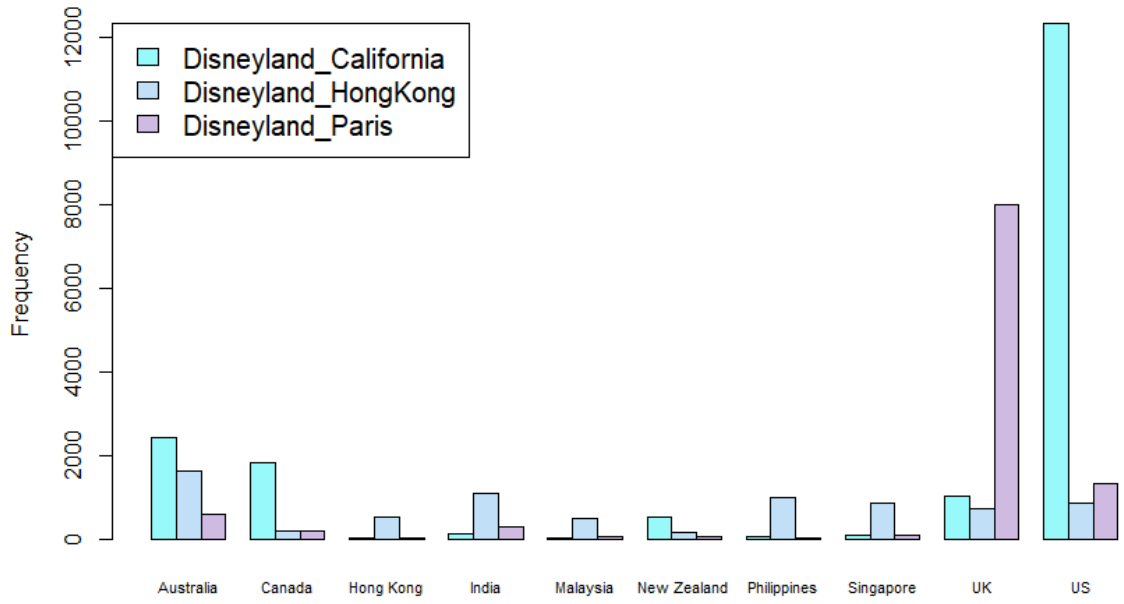

Figure 1: Review frequency per traveler group per Disneyland branch.

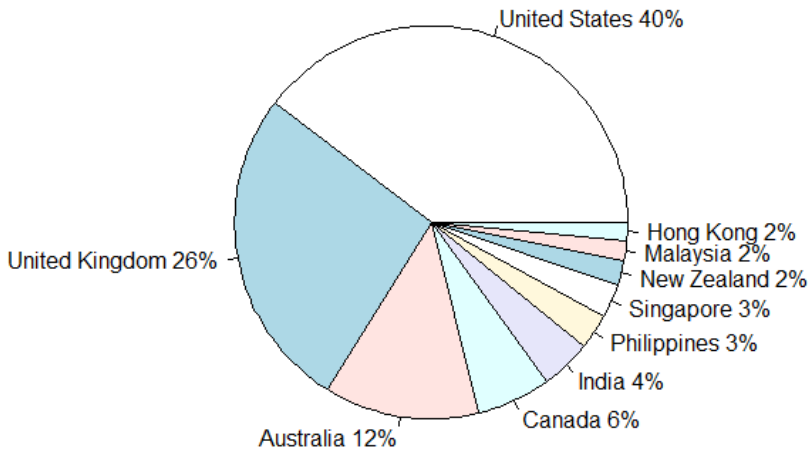

Figure 2: Distribution of reviews per traveler group.

Data pre-processing for LDA

To prepare the data for LDA analysis, the data needs to be cleaned and set up such that the algorithm used can properly analyze the data. This includes the process of removing unnecessary punctuations such as multiple dots or exclamation marks in a row, removing stop words, stemming and tokenizing named entities. The removal of stop words increases the quality of the output obtained from the model as it removes all common words that add little value, in terms of discriminatory power, from the analysis. Stemming is the process of returning words to their basic form so that words such as "beautifully" and "beautiful" are recognized as being the same word. As the LDA method is based on word count, it is important that words that have the same meaning be recognized and counted as one entity.

Next to stemming, tokenizing domain specific entities is an important part of pre-processing the data for LDA analysis. Specific named entities within the Disneyland data should be recognized as one token during the analysis. Take the named entity "Mickey Mouse" as an example. It consists of two words which will be recognized as to separate tokens if not properly tokenized. In this case, the two words are

combined as “mickeymouse” using regular expression such that they are recognized as one token. This is called tokenizing named entities and it significantly affects the results found with sentiment analysis (Laboreiro et al, 2010). In this research, rollercoaster, restaurant, hotel, Disney characters and show names were coded to be recognized as one single token, as well as other domain specific names. Some examples are the variety of passes available at Disneyland such as fast passes, day passes and annual passes; Disney character names such as “Minnie Mouse”, “Peter Pan”, “Winnie the Pooh”; ride names such as “It’s a Small World”, “Tower of Terror”, “Star Wars Hyperspace Mountain”; restaurant names such as “Harbour Galley”, “River Belle Terrace”,” Royal Street Veranda”.

Data pre-processing for sentiment analysis.

Nearly all steps above remain the same when pre-processing the data for sentiment analysis, except for the stemming stem. As previously explained, lexicons containing pre-defined lists of words are used for lexicon-based approach to sentiment analysis. The words in the lexicons are spelled out fully. The process of stemming words in the reviews would hinter the matching process between words in the lexicon and word in the reviews.

5 Results

LDA topic modeling

Using the perplexity measure, different values of k are tested to find the optimal number of topics to build an LDA model with. Figure 3 presents the perplexity scores for the different number of topics, ranging from 20 to 70 topics. It can be derived from Figure 3 that the perplexity score decreases as the number of topics increases. In other words, the model is less perplexed about piecing together the documents in the training set when the number of topic increases. Considering the computing time of over 9 hours and the number of topics to be interpreted, it was decided to not test the model for larger numbers of topics.

.

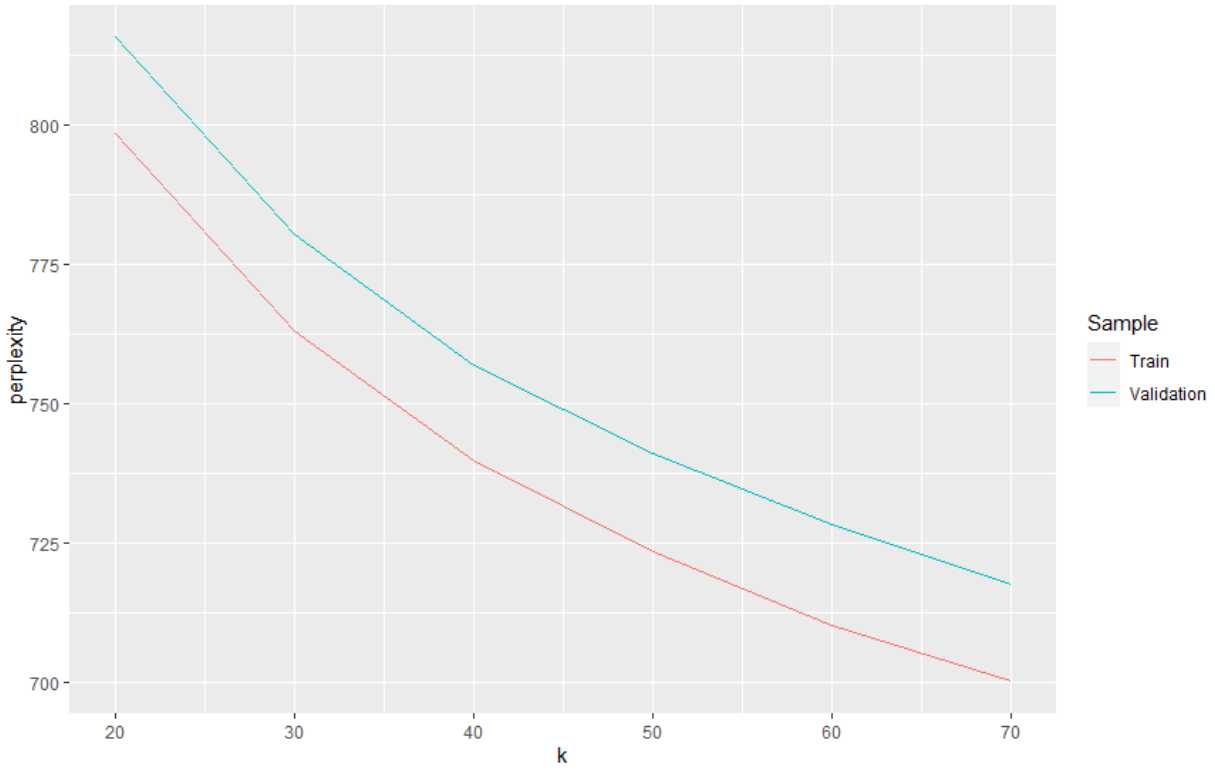

Figure 3: Perplexity plot for the LDA model with k being the number of topics

LDA-70

As was shown above, the perplexity score was the lowest for $k = 70$. Using this number of topics, the LDA-70 model was trained for different values of α. The resulting perplexity plot can be found in Figure 4. The best α value that fits the LDA-70 model is 0.1 because the perplexity score is the lowest for both the train and the validation set when α is 0.1.

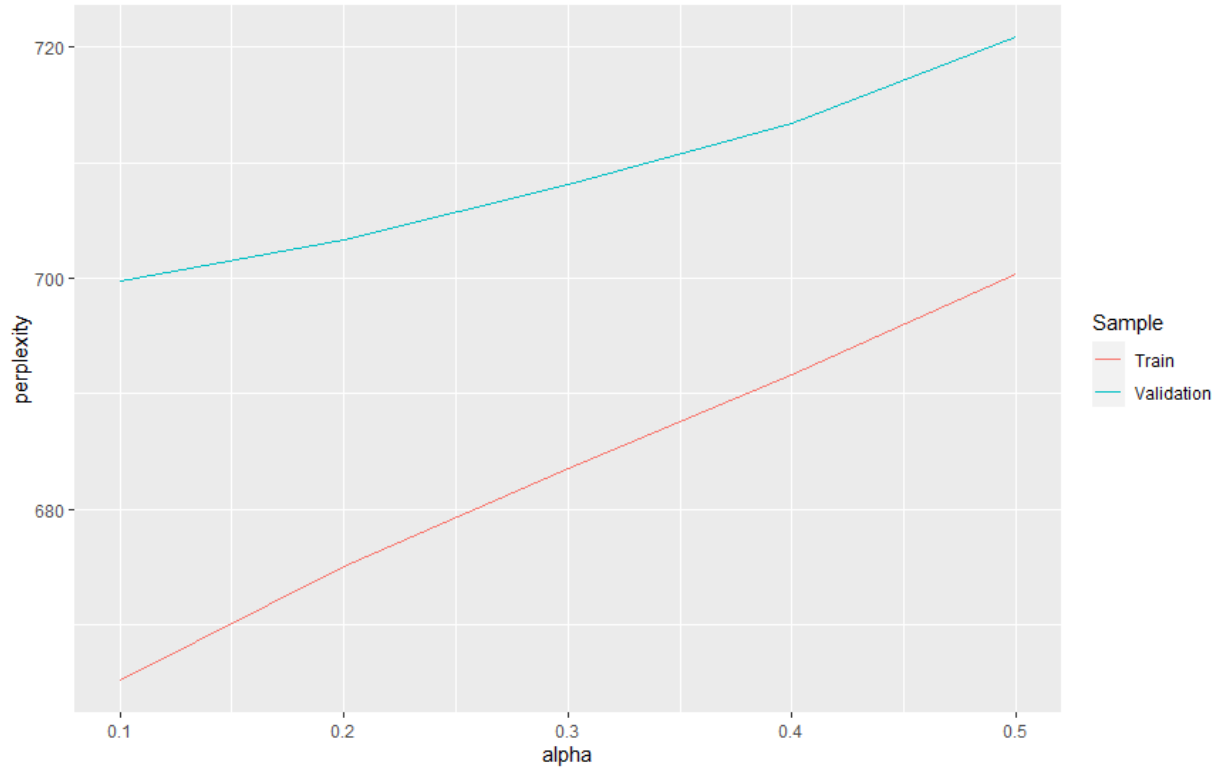

Figure 4: Alpha based Perplexity plot for the LDA-70

LDA-70 – Topic overlap

Figure 5 contains an Intertopic Distance Map (IDM) for the LDA-70 model and depicts topic similarity. Each circle in the map represents a topic and the distances between the circles represent the dissimilarity between the topics. This means that smaller distances between circles indicate more similarity between topics. Figure 5 shows that majority of the circles for the LDA-70 model contain little to no distance between one antoher, indicating quite some overlap between the topics. This notion was confirmed upon further inspection by looking at the top 20 topics. The terms for 6 different topics were

similar and represented long waiting lines for the rides. There terms can be found in Table 1 below. Considereing the aim of this study, which is to find the most popular topics within the dataset, it is desirable to find as many distinct topics possible to interpret. This can be achieved with a minor with little overlap between topics.

From Figure 5 it can be derived that there are approximately 30 overlapping topics. Consequently, a second LDA model is built with 40 topics.

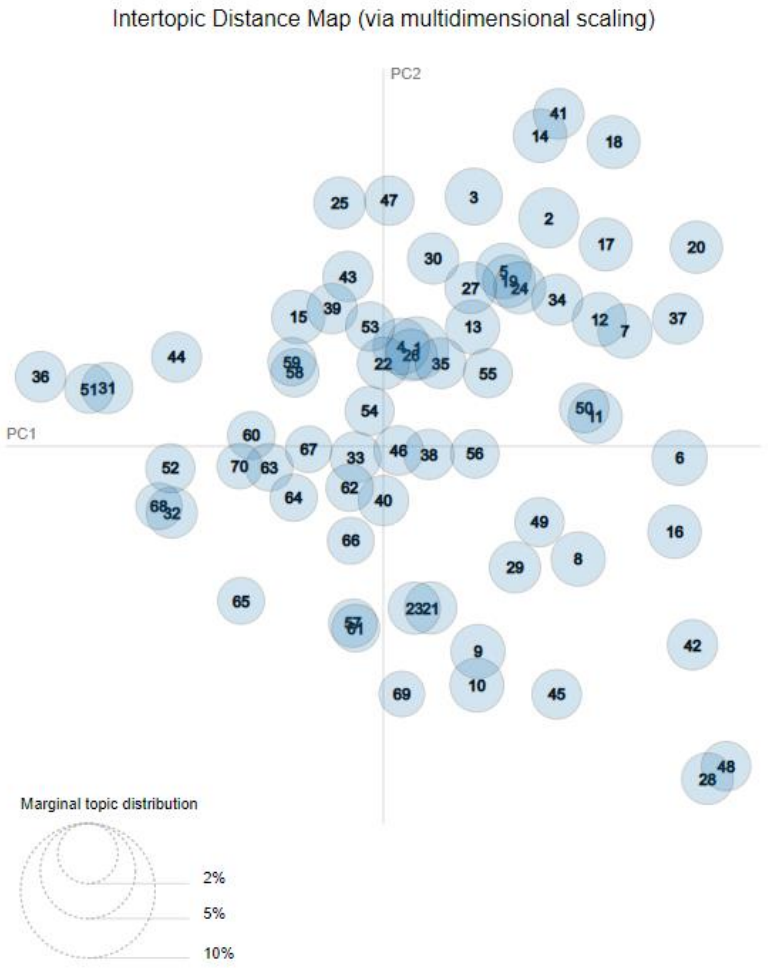

Figure 5: Visualization of topic proximities for LDA-70 model.

Topic number	Terms
1	queue, ride, time, long, wait, day, minutes, hour, park, fastpass, queue, fast, get, ticket
4	time, fastpass, plan, ride, app, day, use, can, get, park, wait, also, help, disneyland
22	line, wait, ride, long, hour, minutes, get, time, fastpass, people, even, crowd, stand, min
26	fastpass, ride, get, time, line, wait, use, can, one, day, singl, long, system, rider
35	queque, ride, hour, money, disney, food, get, euro, just, day, even disappoint, park, wait
53	wait, ride, time, went, day, minutess, line, park, minut, crowd, longest, disneyland, got, hour

Table 1: Topic with similar topic terms for the LDA_70 model

LDA-40

Next, an LDA model is trained for the optimal value of α using $k = 40$. The resulting perplexity plot can be found in Figure 6. The figure shows the perplexity for different values of α and shows that the training sample becomes increasingly more perplexed as the α increases. For the validation sample, however, it can be derived that the LDA-40 model is the least perplexed at α value 0.2.

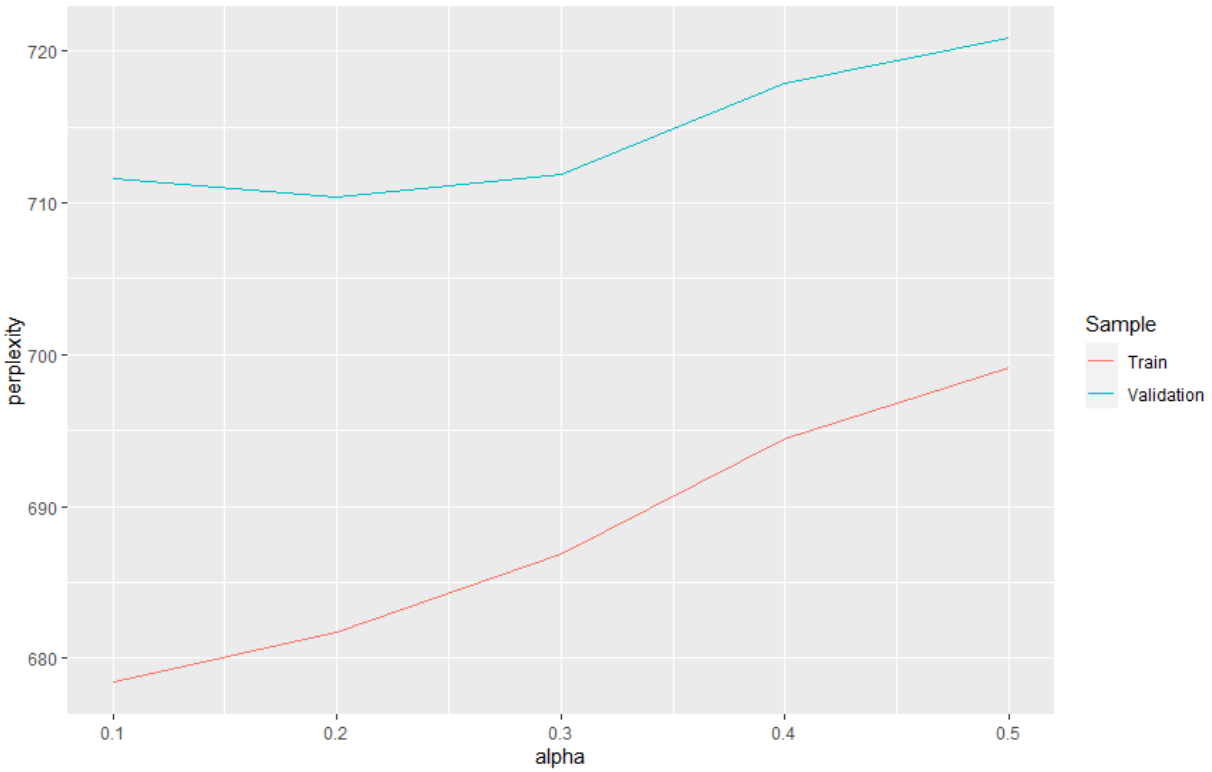

Figure 6: Alpha based Perplexity plot for the LDA-40

Topic overlap LDA-40

Figure 7 show the IDM for the LDA-40 model. A few things can be noted. The density of the plot has decreased compared to the LDA-70 IDM. This is favorable as more distance between topics indicates less overlap. The plot shows that there is still quite some overlap between certain topics. When looking at the top 20 topics for the model however, it can be seen that this overlap is minimal due to the topics being distinguishable from another. For this reason, the LDA-40 model is used for interpretation. The top 20 topics will be interpreted in the following section.

Figure 7: Visualization of topic proximities for LDA-40 model**.**

LDA-40 topic interpretation

A horizontal bar graph is presented in Figure 8 containing the top 20 topics from the LDA-40 model in descending order. This graph includes the the top 15 words that contribute to each topic, also in descending order. The interpretation for the topics was done using the top 20 terms, however.

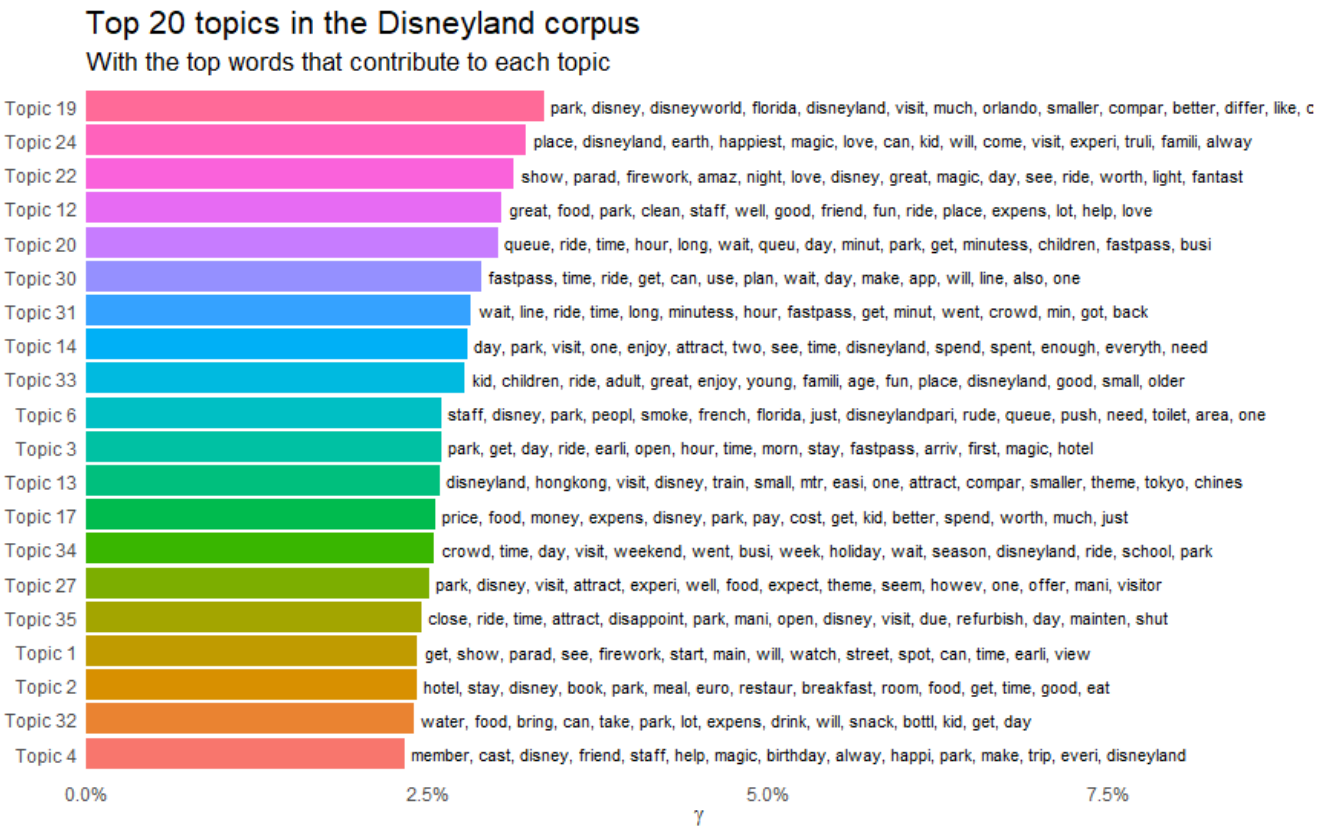

Figure 8: Top 20 topics in the Disneyland corpus for the LDA-40 model with γ indicating the probabilities that each document is generated by each topic.

The top 20 topics are: 19: Comparison of Disneyland California and Disney world in Orlando. 24: Disneyland is a great/ happiest and most magical place on earth and is fun for everyone. 22: Firework show at night is amazing and a must see. 12: The staff at the park being friendly, the park is clean, and the food is good, food price. 20: Waiting times while standing in queue for rides and usage of fast passes. 30. Saving time by planning your day using the Disney app or a fast pass. 31: How long they

waited in line for rides and the usage of the fast pass. 14: How many days are enough to see and enjoy everything in the park. 33: Rides at Disneyland being enjoyable and fun for everyone. 6: Unfavorable aspects of Disneyland Paris and Disneyland California such as smoking, rude staff and pushing in queues. 3: Arriving at the park early for rides, opening hours, staying at the hotel. 13: Transportation to Disneyland Hong Kong and it being smaller compared to other parks; food at Disneyland Hong Kong. 17: The food at Disneyland being expensive and overpriced. 34: Crowds at Disneyland during different times for the year/week. 27: Overall experience 35: Disappointment when rides are not open. 1: Claiming a spot before the parade and shows for a good view. 2: Booking hotel restaurants and discussing the food prices. 32: Bringing your own food, drink and snacks for the kids and prices of consumables at Disney being expensive. 4: Disney cast members/staff being friendly and helpful, birthdays at Disney are fun.

As expected, there are some topics that have some overlap. For example, topics 20 and 31. For the most part however, the top 20 topics are quite distinguishable from another.

Sentence based LDA model

In this section, a sentence based LDA model is trained for different values of k and the resulting perplexity plot can be found in Figure 9. According to the figure, model SLDA becomes less perplexed as the number of topics increases. In efforts of keeping the SLDA model as comparable as possible to the LDA-40 model, and to avoid topic overlap, k is set at 40.

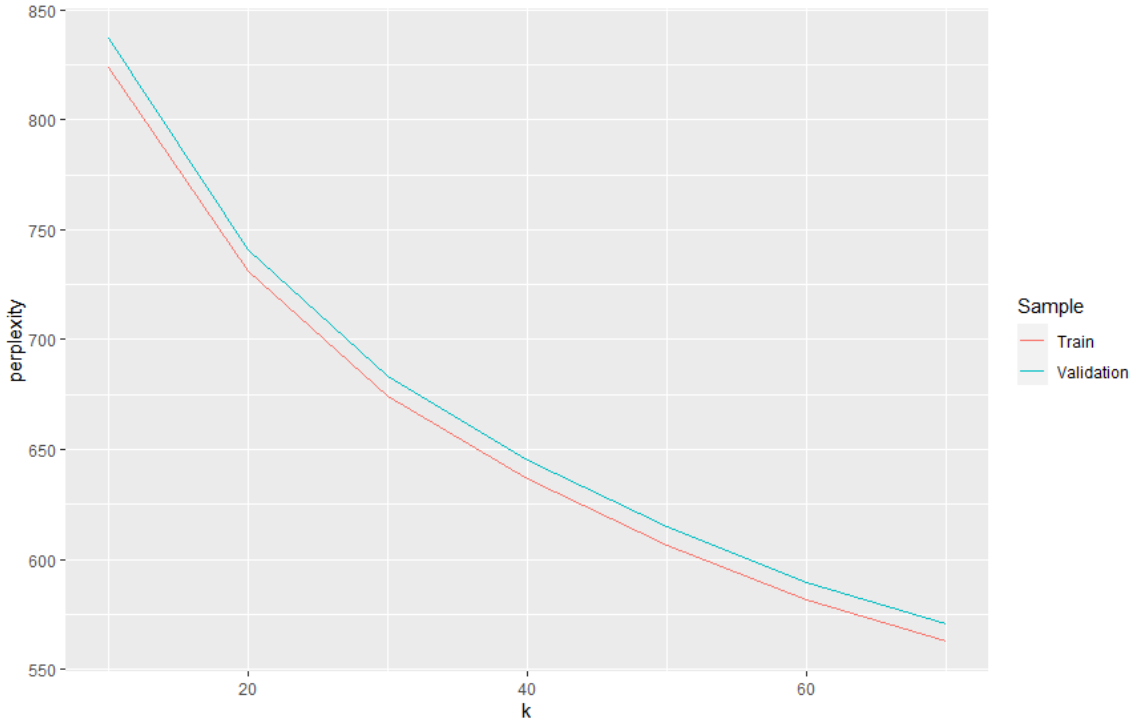

Figure 9: Perplexity plot for the SLDA model with k being the number of topics

Next, the optimal value of alpha is determined according to perplexity based cross validation and found to be 0.05. Figure 10 show the IDM for the SLDA-40 model. Again, the distances between the circles indicate some topic overlap.

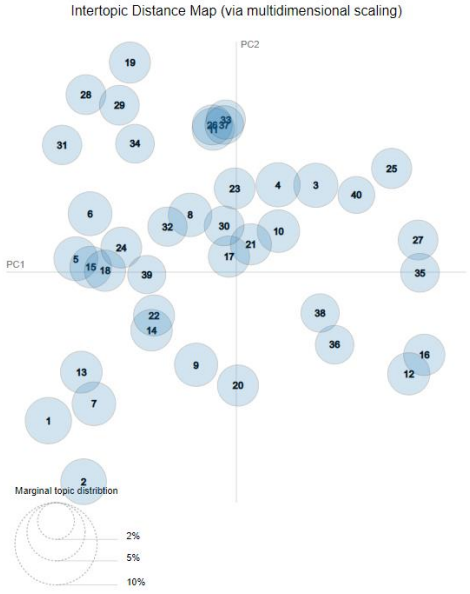

Figure 10: Visualization of topic proximities for SLDA-40 model.

Taking a closer look at the topics.

Figure 11 contains a graph with the top 20 topics from the SLDA-40 model in descending order.

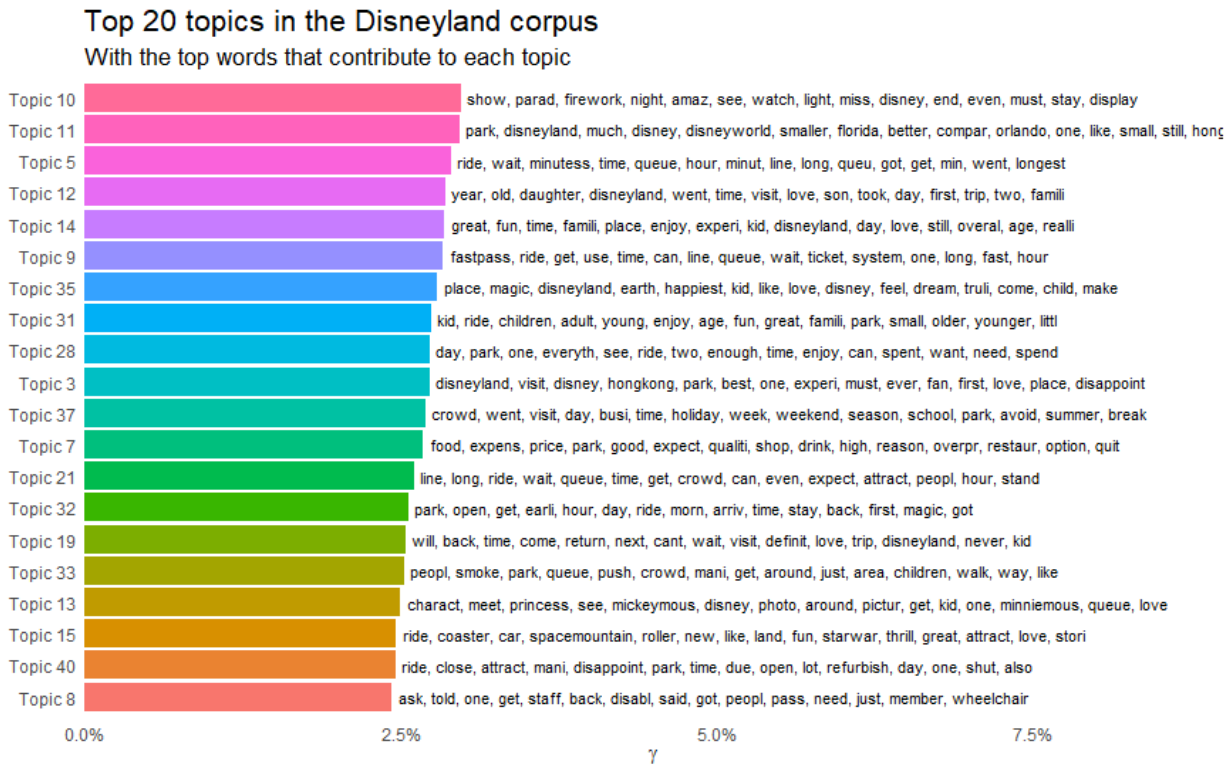

Figure 11: Top 20 topics in the Disneyland corpus for the SLDA-40_2 model with γ indicating the probabilities that each document is generated by each topic.

10: Firework show at night is amazing and a must see. 11: Comparison of Disneyland California and Disney world in Orlando. 5. Waiting times while standing in queue for rides. 12: First time visiting the park with family and them loving it. 14: Great experience with family at the park. 9: Waiting times while standing in queue for rides and usage of fast passes. 35: Disneyland is the happiest and most magical place on earth and is fun for everyone.31. Rides at Disneyland being enjoyable and fun for everyone. 28: How many days are enough to see and enjoy everything in the park. 3: Evaluating Disneyland Hong Kong. 37: Crowds at Disneyland during different times for the year/week. 7: Restaurants and shops at Disneyland being expensive, quality of food. 21: How long they waited in line. 32: Arriving early for the rides, opening hours 19: Cannot wait to return to Disney in the future. 33: Discussing negative sentiment towards smoking areas and pushing in crowds. 13: Meeting and taking pictures with the characters and

the queue. 15: Rides and roller coasters they liked. 40: Disappointment when rides are not open. 8. Communicating about disability and wheelchair with staff.

Comparing LDA-40 and SLDA-40

Both models agree that the following ten topics are in the top 20 topics discussed in the Disneyland reviews: Comparison of Disneyland California and Disney world in Orlando; Firework show at night is amazing and a must see; Waiting times while standing in queue for rides and usage of fast passes; Rides at Disneyland being enjoyable and fun for everyone; Disneyland is the happiest and most magical place on earth and is fun for everyone; How many days are enough to see and enjoy everything in the park; Crowds at Disneyland during different times for the year/week; Food at Disneyland being expensive and overpriced; Disappointment when rides are not open; Arriving early for the rides, opening hours. These topics are explained by both models using similar words and the difference between words are minimal. An example of this is topic 34 from the LDA-40 model where the crowds and busyness are discussed at different times of the visit and topic 37 from the SLDA-40 model where the topic mentions specifically to avoid visit during these times. Another example is topic 3 from the LDA-40 model compared to topic 32 from the SLDA-40 model. Both topics talk about Arriving times at Disney and opening hours but topic 3 is slightly more general as the topic includes discussion on the hotel as well. These two examples indicate a higher degree of coherence for the SLDA model.

There are also topics present in both models that cover the same broader theme with each model having a different focus. An example is topic 13 from the LDA model and topic 3 of the SLDA model. Both topics cover Disneyland Hong Kong but topic 13 focusses on the transportation and size of the park, while topic 3 focusses on the overall experience at the park. It must be noted that although it did not make it in the top 20 topics, the SLDA also contains a topic about transportation to Disneyland Hong Kong specifically by train from the MTR train station.

In terms of unique topics, it can be said that the topics from the SLDA are again found to more specific to one matter and are therefore easier to interpret. Some examples are topic 12, 19, 13, 15, and 8. For the LDA model, a topic often talk about more than one matter at the same time. For example, topic 12 talks about the staff at the park being friendly, the park being clean and discusses food at the same time. Another example is topic 2 which discusses booking a stay at the hotel, hotel rooms, prices, food at the hotel restaurants and topic.

Regarding the variety, both models found approximately equally wide variety of topics. Finally, it should be noted that LDA model found 3 advisory topics while the SLDA model found none. These topics are topics 30, 1 and 32. For the implementation of SA, only the topics that both models had in common will be used because they more reliably belong in the most popular topics discussed in the Disneyland reviews compared to topics that only appear in one model as a top popular topic.

Aspect based sentiment analysis.

In this section, topics are mapped as much as possible to the specific aspects of Disneyland they cover. By doing this, sentiment analysis is performed not on whole topics, but specific aspects of Disneyland that are described within the topics. To understand which aspects of Disneyland is covered by the topics found using LDA topic modeling, topic terms are inspected and analyzed. Table 1 in the appendix contains the topic to aspect mapping. It is important to note that not all topics cover a specific aspect of Disneyland. Topics that compare different parks with each other or topics that discuss the number of days that are enough to explore the parks cannot be assigned to one specific enough aspect of Disneyland. The aspects that were able to be mapped are rides, food, lines, staff, fireworks, prices and crowds. Next, sentences were assigned to each aspect using Algorithm 1 explained in the method section, and sentiments are calculated using polarity scores.

	Polarity		
	California	Hong Kong	Paris
Lines	0,081	0,088	0,046
Fireworks	0,208	0,219	0,257
Rides	0,157	0,190	0,123
Food	0,124	0,065	-0,017
Prices	0,070	0,148	0,019
Crowds	0,052	0,092	0,057
Staff	0,373	0,340	0,208

Table 2: Average polarity scores per aspect per Disneyland branch using polarity.

Table 2 contains the average polarity scores for all aspects per Disneyland branch. These scores indicate whether the overall sentiment per aspect of visitors is positive or negative. As can be seen in the table, visitors express on average positive sentiment towards all aspects for all Disneyland branches, except for the sentiment towards food for Disneyland Paris. The average polarity scores range between 0,373 and -0,017. This indicates that although visitors generally express positive sentiment for all aspects, except for food at Disneyland Paris, the overall sentiment on these aspects is only slightly positive.

Furthermore, visitors are in general most positive about the staff and the fireworks at the Disneyland branches. Next to this, the branches differ on the lowest positive aspects. Prices and crowds at Disneyland California receive the lowest sentiment value out of all aspects at Disneyland California, while for Disneyland Hong Kong is it food and lines that receive the lowest sentiment values. For Disneyland Paris, food and prices were the two aspects with the lowest sentiment values.

Disneyland Paris has the lowest sentiment scores for 5 out of the 7 aspects, indicating that visitors express less contentment with this branch overall. For the fireworks aspect however, this branch has the highest sentiment score out of all three branches. Disneyland Hong Kong scores the highest for aspects lines, rides, prices and crowds while Disneyland California scores the highest for staff and food.

Difference in opinions

In the next section, boxplots are created to analyze the difference in opinion of visitors by examining the spread of the polarity scores. Figure 12 shows that visitors disagree the most, in terms of sentiment, on aspects food, fireworks, and staff while they seem to agree the most on aspects lines, crowds and rides. Although the opinions on fireworks and staff varying the most among visitors, the overall sentiment towards these aspects were found to be the highest for all branches, as we could previously see. The same cannot be said for the food aspect.

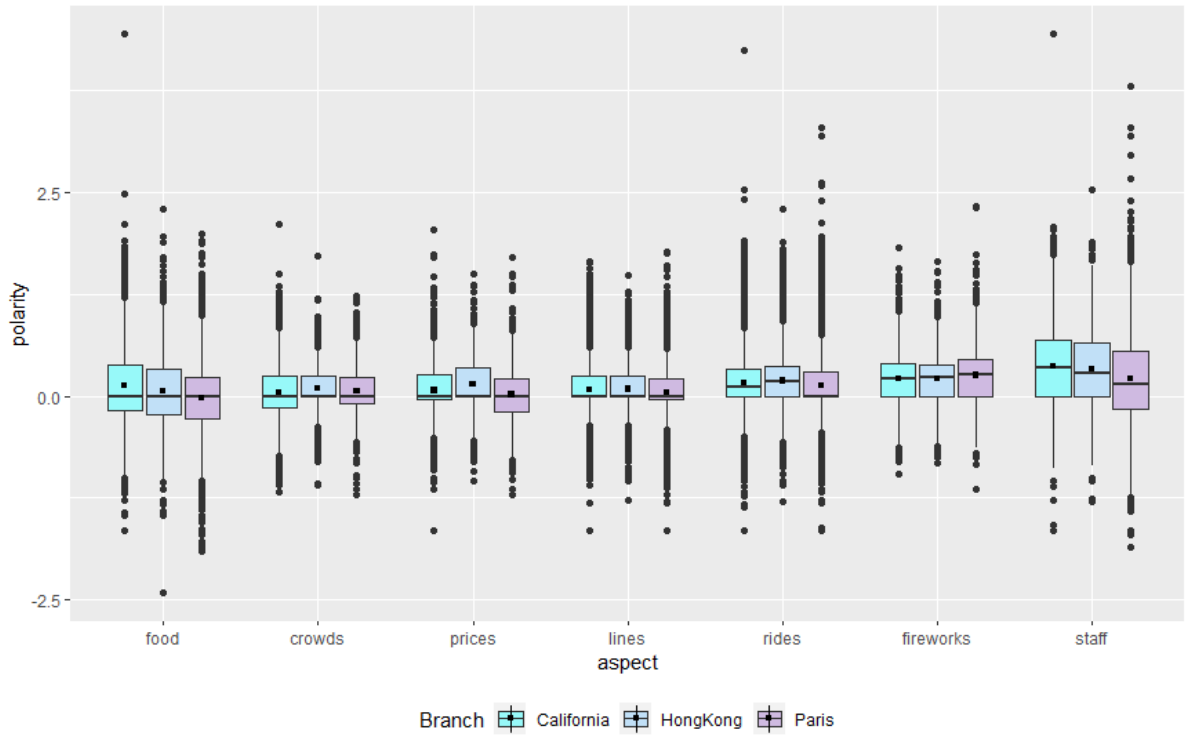

Figure 12: Polarity boxplot per aspect per branch with the black cube indicating the mean polarity score.

Sentiment intensity

Next, the intensity of the positive and negative sentences is analyzed for a more fine-grained analysis.

	Positive Polarity			Negative Polarity		
	California	Hong Kong	Paris	California	Hong Kong	Paris
Lines	0,352	0,346	0,336	-0,296	-0,296	-0,298
Fireworks	0,406	0,408	0,434	-0,272	-0,297	-0,283
Rides	0,398	0,413	0,390	-0,292	-0,280	-0,294
Food	0,477	0,447	0,406	-0,378	-0,388	-0,404
Prices	0,381	0,423	0,346	-0,320	-0,318	-0,332
Crowds	0,352	0,339	0,346	-0,326	-0,321	-0,300
Staff	0,622	0,626	0,609	-0,349	-0,324	-0,380

Table 3: Average polarity scores per aspect per Disneyland branch.

Table 3 presents the average polarity scores for sentences with a positive polarity value and sentences with a negative polarity value, separately per branch. By looking at the average polarity scores for positive and negative sentences separately, the intensity of the positive and negative sentiments can be explored.

Table 3 shows that the positive sentiments expressed by visitors are more intense that the negative sentiments for all aspects. The table also shows that staff at Disneyland is by far the aspect visitors feel the strongest positive sentiment towards, and it is the aspect with second strongest average negative polarity score. This is in line with the notion previously established that opinions are most divided for this aspect. The category with the second highest positive polarity scores for all branches is food. Notably, this category has the lowest polarity scores as well. The opinions are divided on this aspect and explains why the food is not one of the aspects with highest polarity scores overall.

It is also interesting to note that although Disneyland Hong Kong has the highest average polarity scores out of all three branches for aspects lines, rides, prices and crowds, it does not have the highest average scores for positive sentences for lines and rides. This means that for these two aspects, the negative and neutral polarity scores for Disneyland Hong Kong are less intense than they are for Disneyland California. The same goes for staff at Disneyland California, which has the highest average polarity overall but not the highest average polarity score for positive sentences. Disneyland Paris scores consistently lowest for both positive and negative polarity scores, with the exception of fireworks and crowds aspects.

Inspecting sentences

Next to knowing the intensity of sentiments visitors expresses with regards to an aspect, it is also interesting to know the rate at which they express these sentiments. Consequently, all sentences were classified as being either positive, neutral, or negative based on their polarity scores. The plots containing these ratios can be found in Figures 13 and 14.

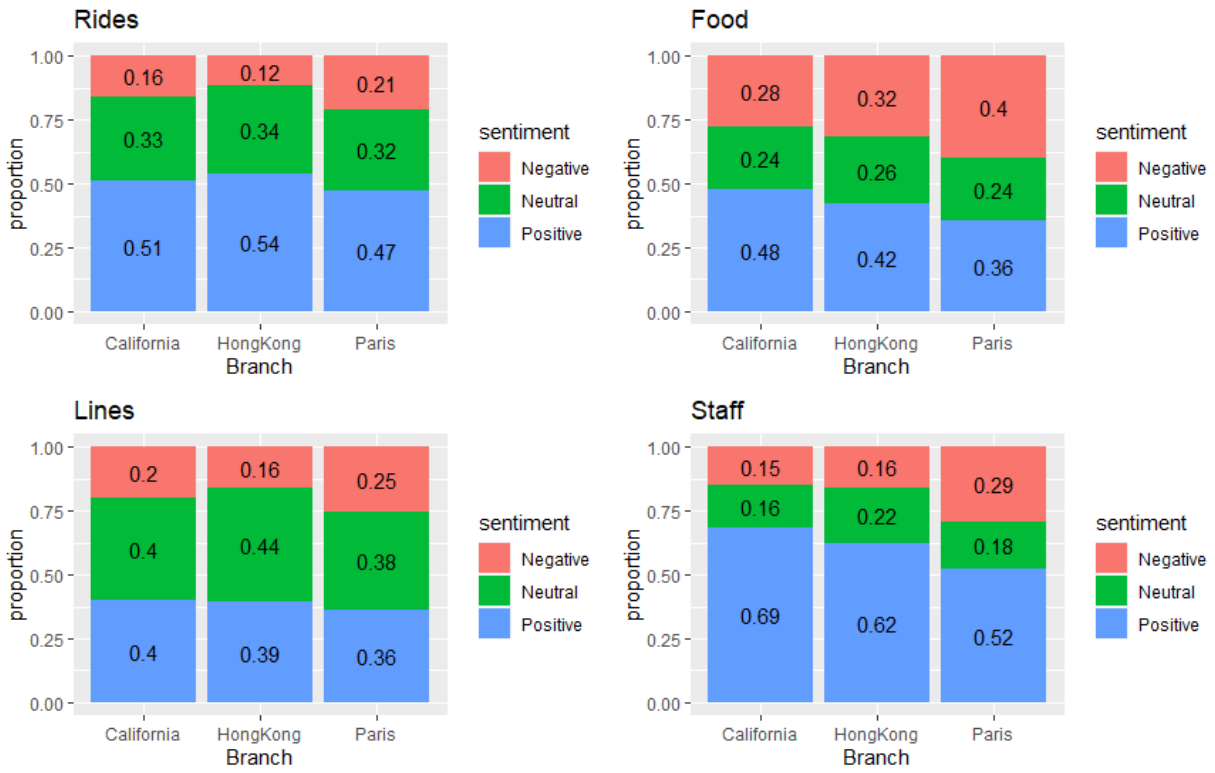

Figure 13: Ratio positive, neutral, and negative sentences for aspects Rides, Food, Lines and Staff.

Figure 14: Ratio positive, neutral, and negative sentences for aspects Fireworks, Prices and Crowd.

Visitors write mostly positive and neutral sentences for each aspect. The aspect with the highest positive rate at 0,69 is staff at Disneyland California, in other words, 69% of all sentences about staff at Disneyland California expresses positive sentiment. Only 15% of all sentences for staff at Disneyland California express negative sentiment. Considering that the average negative polarity rate for staff at Disneyland California was the second lowest out of all aspects, indicates that visitors express either a positive or a strong negative sentiment towards this aspect. Staff and fireworks contain the highest ratios of positive sentences for all branches and fireworks contains the lowest ratio of negative sentences for all branches.

The largest share of neutral sentences belongs to aspects lines and crowds for Disneyland Hong Kong. The figures show that visitors express mostly neutral sentiment when writing about lines for all three branches. Aspects rides, fireworks, prices and crowds also have high ratios of neutral sentences for all branches. With regards to negative sentences, aspects food, followed by prices then crowds contain the biggest largest share. The latter is valid only for Disneyland California and Disneyland Paris. Here again, it can be noted that although aspect food for Disneyland California and Disneyland Hong Kong is the aspect with biggest ratio for negative sentences, the average positive polarity scores are the second highest for both branches. This shows the effect of difference in opinions amongst visitors for this category. For Disneyland Paris however, a bigger portion of sentences are negative as opposed to positive for the food aspect. Lastly, it can be said that Disneyland Hong Kong has the highest portion of positive sentences for aspects rides, prices and crowds across all branches, while Disneyland California has the highest portion of positive sentences for food and staff and Disneyland Paris for fireworks.

Top 10 traveler groups

In this section, differences in sentiment across the 10 biggest traveler groups, the country of origin of residents with most reviews in the data, are analyzed to see if there are underlying differences in sentiment across traveler groups. Figure 2 in the data section contains the size of each traveler group in the data for more perspective.

Figure 15 contains boxplots for Disneyland California. When looking at average polarity scores, it becomes apparent that opinions across traveler group are quite varied for each aspect. This is especially

true for aspects prices, food and staff. Traveler groups agree the most on rides and fireworks at Disneyland California. India expresses the greatest positive sentiment out of all traveler groups for staff, fireworks and rides at Disneyland California, Malaysia for food and lines, Philippines for prices and Australia for crowds. According to the figure, Philippines expresses the lowest sentiment out of all traveler groups for most aspects, except for price.

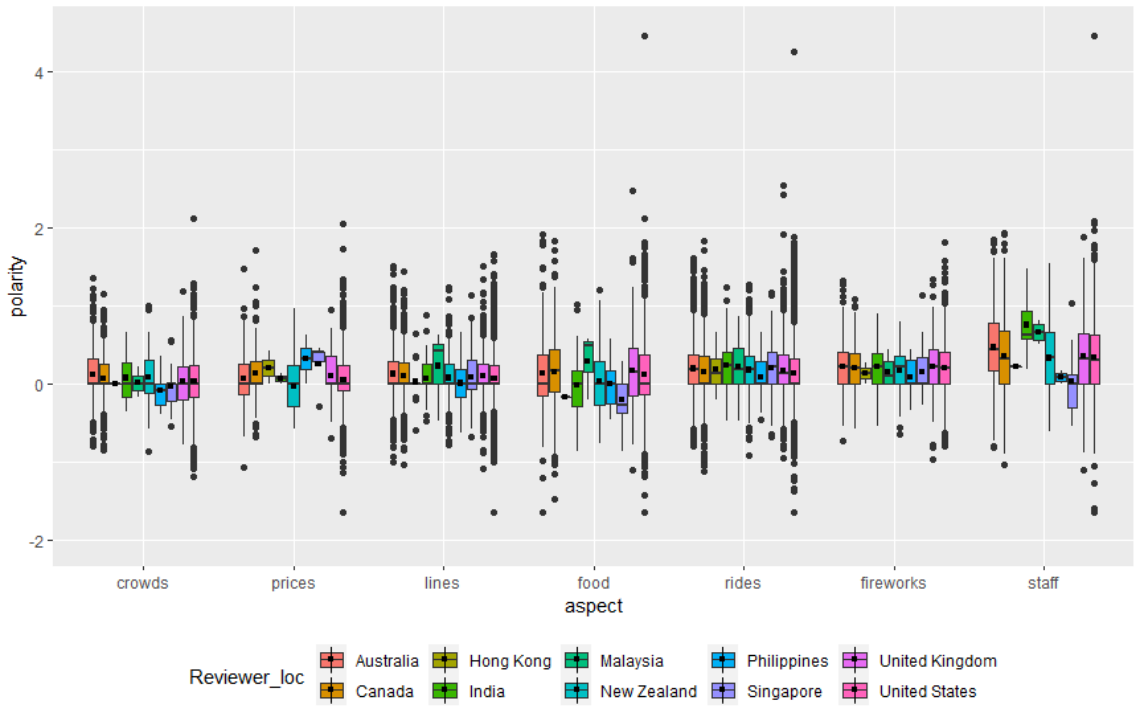

Figure 15: Boxplots per aspect showing polarity scores for Disneyland California for top 10 traveler groups

Figure 16 contains boxplots for Disneyland Hong Kong. The average polarity scores per aspect is similar across traveler group, making Hong Kong the branch with the most uniform opinions. The biggest difference of opinion across traveler groups is for aspect staff. United States expresses the greatest positive sentiment across traveler groups for food, prices, fireworks and staff at Disneyland Hong Kong, India for lines and rides and New Zealand for crowds. Hong Kong expresses the lowest sentiment for food, fireworks and staff for its own branch across all traveler groups, Malaysia scores the lowest for lines and rides, Philippines has the lowest sentiment towards crowds and India for prices.

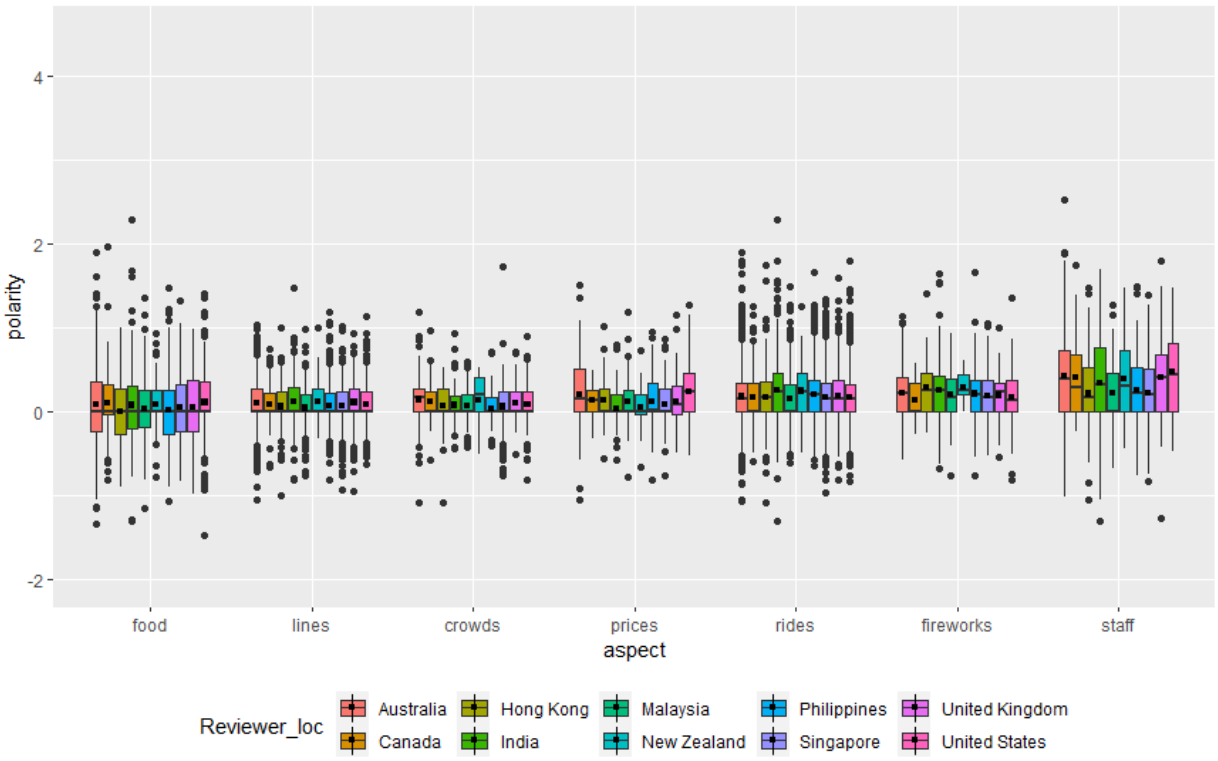

Figure 16: Boxplots per aspect showing polarity scores for Disneyland Hong Kong for top 10 traveler groups

Finally, Figure 17 contains boxplots for Disneyland Paris. For this branch, opinions across traveler group are quite varied for aspects crowds, staff and fireworks. Traveler groups agree the most on rides at Disneyland Paris and disagree the most on staff. Regarding average polarity scores, Malaysia expresses the highest sentiment towards food, crowds and staff across all traveler groups, Philippines towards prices, Hong Kong towards lines, India towards rides and New Zealand towards fireworks. Canada expresses the lowest sentiment towards food across all traveler groups, New Zealand towards prices, Philippines towards lines, Singapore towards crowds, Hong Kong towards staff and India towards fireworks.

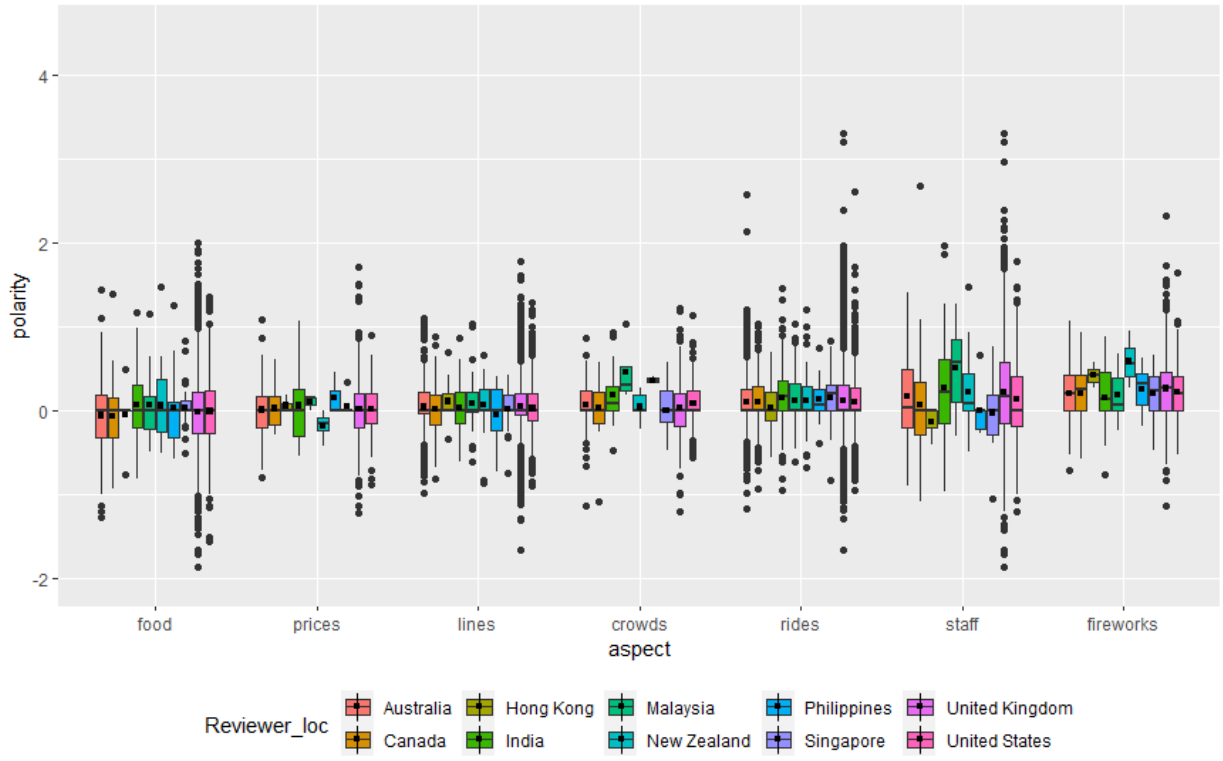

Figure 17: Boxplots per aspect showing polarity scores for Disneyland Paris for top 10 traveler groups

From this analysis, a few things become apparent. First, opinions of different traveler groups on Disneyland Hong Kong are quite uniform while for the other two branches they differ much more. Aspect staff belongs in the top aspects where opinions across traveler groups differ most for all three branches while aspect rides belongs to the top aspects where opinions are most similar across traveler groups for all three branches. Furthermore, Philippines scores out of all three branches the highest on aspect prices, indicating that prices at Disneyland are not as big of a concern for this traveler group compared to the others. Malaysia seems to have a liking towards food outside of Asia as scores food at Disneyland California and Disneyland Hong Kong the highest out of all country. Finally, India expresses the highest sentiment towards rides across all traveler groups for all three branches.

6 Conclusion

In this research, an attempt is made to apply NLP techniques in an important yet overlooked industry when it comes to machine learning, namely the tourism industry. More specifically, hidden topics from Disneyland reviews are uncovered, mapped to aspects, and then sentiment is computed. The topics in this research are uncovered using both a standard and a sentence based LDA model. We will start the conclusion section by answering the first sub-question, *"What are the most popular topics discussed in Disneyland Reviews?*". The results show that a wide variety of topics are discussed, ranging from a comparison of Disneyland California to Disney World in Orlando to expressing disappointment rides are closed down to the firework show is a great and must-see event. Some of the other topics being discussed concerned waiting times, visiting the park with family, expressing contentment with the park, rides, discussing when the park is the least crowded, treatment for disabilities at Disneyland, taking pictures with Disney characters, using fast passes to avoid long lines etc. In conclusion, it is found that visitors tend to complain, advise, compare, and show appreciation in the reviews they write.

In terms of the second sub-question, "*Does a sentence-based topic modeling provide better interpretation of topic discussed in Disneyland reviews as opposed to a regular topic model?*" it is found that sentence based LDA model provides more specific and coherent topics that are easier to interpret compared to the more general topics that are found using the standard LDA model. This is in line with previous studies. In terms of topic variety, both models found an approximately equally wide variety of topics. The implementation of sentence-based LDA in this research consisted of a simple procedure of splitting reviews into sentences and treating each sentence as its own document. As the results showed, the topics found using this approach provided more human-interpretable topics, despite this simplicity. Another benefit of this approach found in this research is that the SLDA model took less time to converge than the LDA model each time. Furthermore, the Disneyland review data covers a multitude of topics. As the topic modeling analysis showed, there are many aspects and topics to be discussed. This explains the lengthy reviews in the Disneyland data and could support the idea that reviewers need to naturally bring structure to the reviews they write through the usage of sentence. Sentence-based topic modeling proved to be a good method to perform in other to take this natural structure into consideration.

To answer sub-question three "*What sentiment do visitors feel towards the aspects expressed in the most popular topics?",* we first remind the viewer of the aspects that are mapped to topics. These are rides, food, lines, staff, fireworks, prices and crowds. In general, visitors express positive sentiment towards all aspects across all Disneyland branches on average, except towards food at Disneyland Paris. Managers at Disneyland Paris should invest in improving on this aspect. Furthermore, visitors express overall the greatest contentment with staff and the fireworks at the Disneyland branches. These two aspects, together with food, are also the aspects where opinions across visitors vary the most. It is also found that branches score lowest, in terms of sentiment, on different aspects. Visitors express the least contentment with crowds at Disneyland California, lines at Disneyland Hong Kong and Food at Disneyland Paris. Luckily, branch managers can learn from another branch that performs better on these aspects. This is especially applicable for the Paris branch because it earns the lowest sentiment scores for 5 out of the 7 aspects, indicating that visitors express less contentment towards this branch. Overall, Disneyland Hong Kong scores highest out of all branches in terms of polarity scores for these 7 aspects.

In terms of intensity of the sentiment of visitors, it is found that positive sentiments expressed are more intense than the negative sentiments for all aspects. Staff at Disneyland is by far the aspect visitors feel the strongest positive sentiment towards, and it is the aspect with the second strongest average negative polarity score. This is in line with the notion previously established that opinions are most divided for this aspect. The category with the second highest positive polarity scores for all branches is food. Notably, this category has the lowest polarity scores as well. The opinions are divided on this aspect, and this explains why the food is not one of the aspects with the highest sentiment overall.

It is found that visitors write mostly positive and neutral sentences for each aspect and that people write positive sentences about staff at Disneyland California 69% of the time. With regards to negative sentences, aspects food, followed by prices then crowds contain the biggest largest share. Lastly, it can be said that Disneyland Hong Kong has the highest portion of positive sentences for aspects rides, prices and crowds across all branches, while Disneyland California has the highest portion of positive sentences for food and staff and Disneyland Paris for fireworks.

Differences in sentiment across the top 10 biggest traveler groups are analyzed as well in this research. Firstly, the results show that opinions of different traveler groups on Disneyland Hong Kong are quite uniform while for the other two branches of Disneyland they differ much more. This indicates that all travelers from outside of Hong Kong seem to expect and experience similarly when visiting Disneyland

Hong Kong. Aspect staff belongs in the top aspects where opinions across traveler groups differ most for all three branches, which is in line with prior conclusions on this aspect, while aspect rides belongs to the top aspects where opinions are most similar across traveler groups for all three branches. The later indicates that country of origin does not form a big barrier for how visitors rate rides at Disneyland. Furthermore, Philippines scores out of all three branches the highest on aspect prices, indicating that prices at Disneyland are not as big of a concern for this traveler group compared to the others. Malaysia seems to have a liking towards food outside of Asia as scores food at Disneyland California and Disneyland Hong Kong the highest out of all country. Lastly, India expresses the highest sentiment towards rides across all traveler groups for all three branches.

Limitations

Textual data contains inherent limitations that are not accounted for in this research. For example, this research did not include detection of irony, sarcasm, or intensity expressed through things such as repetition of exclamation marks. Furthermore, it is no secret that people make spelling and grammar mistakes. Although these issues is dealt with for a big part in the data pre-processing stage through the usage of functions such as stemming and regular expression, some context is always lost due to the inability to clean text data completely. In terms of splitting the reviews into sentences, some sentences are not split properly due to misusage of punctuations. Another limitation of splitting reviews into sentences is the loss of context when words that refer to entities in previous sentences, such as “it”, are used. Furthermore, as mentioned in the data section, the Disneyland data contains many domain specific names that should be recognized by applications as one token. Although it is attempted to tokenize most of these names, chances are high that some have slipped through the crack. Finally, the LDA models in this research only considered single word tokens. Studies have shown that using n-grams will enrich the insights that can be extracted from data. It would be interesting to investigate n-gram tokenization for the Disney Review dataset in future research.

References

Aznag, M., Quafafou, M., Rochd, E. M., & Jarir, Z. (2013, September). Probabilistic topic models for web services clustering and discovery. In *European conference on service-oriented and cloud computing* (pp. 19-33). Springer, Berlin, Heidelberg.

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*(6), 1371-1391.

Bennett, J., & Lanning, S. (2007, August). The netflix prize. In *Proceedings of KDD cup and workshop* (Vol. 2007, p. 35).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The annals of applied statistics*, *1*(1), 17-35.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, *30*, 31-40.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, *35*(6), 953-975.

Cabanas, E. (2020). Experiencing designs and designing experiences: Emotions and theme parks from a symbolic interactionist perspective. *Journal of Destination Marketing & Management*, *16*, 100330.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New media & society*, *16*(2), 340-358.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288-296).

Chen, S., & Lamberti, L. (2013). Segmenting Chinese tourists by the expected experience at theme parks. *International Journal of Engineering Business Management*, *5*(Godište 2013), 5-22.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

Clavé, S. A. (2007). *The global theme park industry*. Cabi.

Consultancy.uk (2016, September 8). *The 25 biggest theme parks and water parks in the world***.** https://www.consultancy.uk/news/12519/the-25-biggest-theme-parks-and-water-parks-in-the-world

Davidov, D., Tsur, O., & Rappoport, A. (2010, July). Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 107-116).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.

Fotiadis, A. K. (2016). Modifying and applying time and cost blocks: The case of E-Da theme park, Kaohsiung, Taiwan. *Tourism Management*, *54*, 34-42.

Ganu, G., Elhadad, N., & Marian, A. (2009, June). Beyond the stars: improving rating predictions using review text content. In *WebDB* (Vol. 9, pp. 1-6).

Gómez-Rodríguez, C., Alonso-Alonso, I., & Vilares, D. (2019). How important is syntactic parsing accuracy? An empirical evaluation on rule-based sentiment analysis. *Artificial Intelligence Review*, *52*(3), 2081-2097.

Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science advances*, *4*(7), eaaq1360.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

Hall, D., Jurafsky, D., & Manning, C. D. (2008, October). Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 363-371).

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*(1), 177-196.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, *5*(9).

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Kemperman, A. (2000). Temporal aspects of theme park choice behavior. Modeling variety seeking, seasonality and diversification to support theme park planning. *Collection Bouwstenen*, *58*.

Khan, A., Baharudin, B., & Khan, K. (2010, December). Sentence based sentiment classification from online customer reviews. In *Proceedings of the 8th International Conference on Frontiers of Information Technology* (pp. 1-6).

Kim, D., & Oh, A. (2011, February). Topic chains for understanding a news corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 163-176). Springer, Berlin, Heidelberg.

Kim, J. N. (2020). Variational Expectation-Maximization Algorithm in Posterior Distribution of a Latent Dirichlet Allocation Model for Research Topic Analysis. *Journal of Korea Multimedia Society*, *23*(7), 883-890.

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.

Laboreiro, G., Sarmento, L., Teixeira, J., & Oliveira, E. (2010, October). Tokenizing micro-blogging messages using a text classification approach. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data* (pp. 81-88).

Lau, J. H., Newman, D., & Baldwin, T. (2014, April). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530-539).

Layman, L., Nikora, A. P., Meek, J., & Menzies, T. (2016, May). Topic modeling of NASA space system problem reports: research in practice. In *Proceedings of the 13th International Conference on Mining Software Repositories* (pp. 303-314).

Lee, D.D. & Seung, H.S. (2001). Algorithms for Non-Negative Matrix Factorization. *Advances in Neural Information Processing Systems*, *13*, 556-562

Li, W., & McCallum, A. (2006, June). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning* (pp. 577-584).

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*, 14-23.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011, December). Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th international conference on data mining workshops* (pp. 81-88). IEEE.

Luo, J. M., Vu, H. Q., Li, G., & Law, R. (2020). Topic modelling for theme park online reviews: analysis of Disneyland. *Journal of Travel & Tourism Marketing*, *37*(2), 272-285.

Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2019). Aspect-based sentiment analysis methods in recent years. *Asia-Pacific Journal Of Information Technology And Multimedia*, *7*(2), 79-96.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16-32.

McAuley, J., Leskovec, J., & Jurafsky, D. (2012, December). Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining* (pp. 1020-1025). IEEE.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).

Minka, T., & Lafferty, J. (2002). Expectation-propagation for the generative aspect model, Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence.

Mowlaei, M. E., Abadeh, M. S., & Keshavarz, H. (2020). Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Systems with Applications*, *148*, 113234.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 100-108).

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111-126.

Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodriguez-Garcia, M. A., Moreno, V., Fraga, A., & Sanchez-Cervantes, J. L. (2014). Feature-based opinion mining through ontologies. *Expert Systems with Applications*, *41*(13), 5995-6008.

Popescul, Alexandrin, Lyle H. Ungar, David M. Pennock, and Steve Lawrence. "Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments." *arXiv preprint arXiv:1301.2303* (2013).

Reyes, A., & Rosso, P. (2014). On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, *40*(3), 595-614.

Shuler, S. (2021, February 1). *The Walt Disney Company Ranks 4th on Fortune's 2021 List of "World's Most Admired Companies".* The Insider. https://thedisinsider.com/2021/02/01/the-walt-disney-company-ranks-4th-on-fortunes-2021-list-of-worlds-most-admired-companies/

Titov, I., & McDonald, R. (2008, April). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web* (pp. 111-120).

van Assendelft de Coningh, R. (1995). Marketing a theme park: Efteling. *Journal of Vacation Marketing*, *1*(2), 190-194.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009, June). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105-1112).

Watts, S. (1995). Walt Disney: Art and politics in the American century. *The Journal of American History*, *82*(1), 84-110.

Qiu, M., Zhu, F., & Jiang, J. (2013, May). It is not just what we say, but how we say them: Lda-based behavior-topic model. In *Proceedings of the 2013 SIAM international conference on data mining* (pp. 794-802). Society for Industrial and Applied Mathematics.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision support systems*, *55*(4), 919-926.

Appendix

Topic	Aspect
Comparison of Disneyland California and Disney world in Orlando	-
firework show at night is amazing and a must see	Firworks
waiting times while standing in queue for rides and usage of fast passes	Lines
rides at Disneyland being enjoyable and fun for everyone	Rides
Disneyland is the happiest and most magical place on earth and is fun for everyone	-
how many days are enough to see and enjoy everything in the park	-
crowds at Disneyland during different times for the year/week	Crowds
food at Disney being expensive and overpriced	Food-Price
disappointment when rides are not open	-
arriving early for the rides, opening hours.	-

Table 4: Topic aspect mapping.