FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Quantitative Analysis of Tourism Data using Text Mining

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Lisa Glatzer BSc
Matrikelnummer 0825034

an der Fakultät für Informatik
der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Mitwirkung: Mag.rer.nat. Dr.techn. Julia Neidhardt

Wien, 28. Juli 2017

Lisa Glatzer	Hannes Werthner

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Quantitative Analysis of Tourism Data using Text Mining

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Business Informatics

by

Lisa Glatzer BSc
Registration Number 0825034

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Assistance: Mag.rer.nat. Dr.techn. Julia Neidhardt

Vienna, 28th July, 2017

Lisa Glatzer	Hannes Werthner

Erklärung zur Verfassung der Arbeit

Lisa Glatzer BSc
Ignaz-Weigl-Gasse 3/4/11, 1110 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 28. Juli 2017

Lisa Glatzer

Acknowledgements

First of all I want to thank my thesis advisor Prof. Dr. Hannes Werthner for sharing his expertise, asking critical questions and giving me the chance to do research on this interesting topic. My sincere gratitude I want to share with my co-supervisor Dr. Julia Neidhardt who supported me in technical and organisational matters and was always very patient and motivating. Further, I want to thank the expert team from Pixtri for their participation during the dictionary analysis. Otherwise the evaluation of this approach could not have been done.

A special thanks I want to express to my family who were there for me during my whole life. My parents gave me the opportunity to study and even experience a semester abroad. Especially I want to mention my sister Birgit for reading through the complete thesis and improving my grammar. Further, as far as I remember, she was the one who encouraged me to study Business Informatics.

Many thanks to Moni, Alex, Flo, Richi, Sandra, Dominik, Lukas, Stefan ("The Hammer"), Seb, Ben, Franz, René, Manu and Matt. You've made it an unforgettable journey.

Finally I want to thank Daniel, who gave me strength and enthusiasm during the process of researching and writing this thesis. Without you this thesis could not have been accomplished.

Thank you.

Kurzfassung

Heutzutage nutzen immer mehr Personen Online-Buchungs-Portale, um ihren Urlaub zu planen. Daher verwenden viele Anbieter bereits automatisierte Mechanismen, so genannte Recommender Systeme, um ihren Kunden das passendste Hotel vorzuschlagen. In dieser Masterarbeit werden unterschiedliche Ansätze untersucht, welche es ermöglichen sollen, automatisierte Empfehlungen mittels Hotelbeschreibungen verschiedener Anbieter zu erstellen. Basierend auf ihren textuellen Beschreibungen werden Hotels bis zu sieben vordefinierten Benutzerprofilen zugewiesen, welche zentrale Charakterisika und Eigenschaften von Touristen wiederspiegeln. Um das zu erzielen, werden veschiedene Methoden des Natural Language Processing, darunter Tokenization, Stemming und Pruning eingesetzt. Weiters werden drei unterschiedliche Ansätze für die Zuweisung der Hotels zu den Touristenprofilen umgesetzt: Clustering, Klassifizierung sowie ein Wörterbuch-basierter Ansatz, bei dem Experten Keywords für die einzelnen Profile auswählen. Die Ergebnisse aller drei Vorgehensweisen werden vorgestellt und verglichen, wobei der beste Ansatz mit einem unabhängigen Testdatensatz final evaluiert wird. Die Resultate des Clusterings zeigen, dass ein rein automatisierter Algorithmus ohne manuelle Unterstützung für die Zuordnung von Hotels zu Profilen nicht geeignet ist. Stattdessen werden die Textbeschreibungen nach Anbieter in Cluster unterteilt. Die Methode Klassifizierung liefert die besten Ergebnisse für sechs der sieben Profile, während der Wörterbuch-Ansatz sich für ein Profil als geeignetste Lösung herausstellt. Grundsätzlich ist zwischen den Endresultaten der einzelnen Profile eine große Varianz zu erkennen. Dies ist einerseits auf die ungleiche Verteilung des Test-Datensatzes zurückzuführen. Andererseits haben die Charakteristika der einzelnen Profile signifikanten Einfluss auf die Ergebnisse. Die in dieser Arbeit vorgestellten Modelle dienen der Erstellung von Recommender Systemen auf Basis von Hotelbeschreibungen.

Abstract

The amount of people who use online booking platforms to select a travel accommodation has grown tremendously in the last years. Hence, many tour operator implement recommender systems in order to offer the most suitable hotels to their customers. In the context of this thesis, a method of using hotel descriptions collected by different tourist operators for recommendation is introduced. Based on the content of textual data samples, hotels are matched to seven predefined tourist roles ("The Seven Factors"), which represent general behaviours and preferences of tourists. To achieve this, a preprocessing of the unstructured hotel descriptions is done with different natural language processing methods including tokenizing, stemming and pruning. Further, three different approaches for the allocation of hotel descriptions to the tourist roles were implemented: unsupervised clustering, supervised classification and a dictionary-based approach, where keywords were identified by experts. The outcome of all three methods was compared and the best one was tested with an independent labelled data set of text description samples. The main results show that unsupervised clustering cannot be used to allocate hotels to tourist roles since the algorithm mostly relies on the operator-dependent structure which can be found in the descriptions. Further, it is identified that supervised classification achieves the highest precision for six of the seven tourist roles, whereas the dictionary approach is pointed out as the best solution for only one role. In general, the results for the different tourist roles vary due to the unequally distributed training and test data set as well as the various characteristics of the roles. The defined models are presented so that they can be used as an aid to design recommender systems based on hotel descriptions.

Contents

CHAPTER 1

Introduction

1.1 Problem Description and Motivation

In contrast to most human beings computational machines often face huge difficulties to analyse and understand text. However, the amount of unstructured text data is rising constantly which makes it inevitable to use automatic analysis instead of a manual one.

The approach of text mining is dealing with this problem. It is used to automatically obtain new and beneficial knowledge from text documents. In comparison to data mining, where structured and irreducible data is examined, text mining deals with unstructured data. Typical fields of application are text classification and clustering, the position and sentiment analysis, the extraction of concepts as well as information retrieval. Since the understanding of natural languages is a crucial part of text mining, methods like natural language processing play a major role for the evaluation of textual data [Hippner and Rentzmann, 2006].

During the last years data mining has become popular in many different information sectors including the tourism industry. In Europe, the majority of accommodations for vacations are booked via internet booking platforms [Eurostat]. To arouse the potential customer's interest it is important to present accommodations that the user is likely to book. Therefore, many travel operators want to provide automatic online recommendations for their customers. Some of them save data of visited hotels of their clients and on behalf of this recommend hotels in the same country or with a similar price. This collected information can easily be compared and analysed but might not be the only indicator for a useful recommendation.

The focus of this master thesis lies in the analysis of hotel descriptions generated by travel operators. The approach will be to analyse hotel descriptions using text mining methods and to classify these hotels. This will allow the classification beyond comparable

facts such as price, location and size of the hotel by considering information such as descriptions of the ambience in a hotel (exciting, relaxing, friendly, architecturally appealing,...) and its characteristics (modern, historical, glamorous, elegant,...). Finally, classes of hotels should then be mapped to seven tourist roles based on the results described in [Neidhardt et al., 2014a,b]. Yet the classification of hotels is even more complex than it seems since a hotel can appear in more than one class and the classes can potentially be allocated to more than one travel type.

1.2 Research Questions

The main aim of the work is to use machine learning algorithms in order to allocate a hotel to a tourist role by using textual hotel descriptions. When achieving this it should be possible to improve the quality of hotel recommendations in general and help both travel operators to advance their business processes and maximize their potential and also tourists to spend a more pleasant time on vacations.

Based on this defined statement the subsequent two **research questions** are treated in elaborating the thesis:

> *How can textual hotel descriptions be used to identify concepts or models to enable a classification of hotel descriptions?*
>
> *To what extent can the classified descriptions be allocated to different predefined tourist roles („The Seven Factors“) and a recommendation of hotels be enabled?*

1.3 Methodology

The following methodology was applied to achieve the expected goals. The Design Science approach by [Hevner et al., 2004] was taken as a reference.

Figure 1.1: Methodology of the master thesis

1) **Literature Review**
In order to gain information about the topic it was important to do accurate research for defining the state of the art. Especially the evolution of roles in tourism is discussed as well as the usage of recommender systems in the tourism domain. Further, the literature review was important for determining the appropriate text mining method for this research problem. Therefore, it was inevitable to execute

an analysis of the currently used text mining techniques as well as machine learning algorithms in order to select the appropriate tools and procedures.

2) **Data Acquisition and Analysis**
The data source for the evaluation of hotel recommendations is the GIATA XML Webservice [GIATA GmbH]. This service provides a rich amount of information about hotels including location, price, travel offers and on top of that also hotel descriptions of different travel operators from all over the world. The latter was used for the text mining analysis. As the data is mostly in German, the text mining methods were chosen due to their adequacy for German language. Further, the data had to be stored in a database in a feasible format to properly run different text mining methods. To become an insight on the structure and content of the descriptions a qualitative analysis of the data was performed.

3) **Data Preprocessing**
In order to use text mining methods in a beneficial way the data had to be prepared using natural language processing. The following approaches were applied: Tokenization, deletion of stopwords, stemming and pruning. To establish an appropriate data format for the appliance of machine learning algorithms, the data had to be converted to word vectors. Further, the training and test data set for building the models and evaluating them had to be established.

4) **Allocation of Hotel Descriptions**
With a predefined set of hotel descriptions, the training set, models were built to enable a grouping of the data samples. For allocating the hotel descriptions to seven tourist roles three different approaches were implemented and their results were compared:

- Clustering (unsupervised)
- Classification (supervised)
- Dictionary based approach

5) **Evaluation of Results**
After the establishment of the models based on the three approaches, a final evaluation was conducted. Therefore, the best approach was tested with an independent test set. Finally the results were discussed and the individual outcomes for all seven tourist roles were examined in the context of recommender systems for hotels.

1.4 Outline of the Work

The structure of the master thesis is organised as follows:
In chapter 2 literature studies regarding solutions related to the problem treated in this thesis are described and existing approaches are compared and summarised. Additionally

details about the used technologies and the tourist factors developed by [Neidhardt et al., 2014a,b] are introduced. Chapter 3 gives an overview of the acquisition of the data. Furthermore, the conclusions of a qualitative analysis of the textual descriptions are presented. The detailed preprocessing steps and the preparation of training and test set are described in chapter 4. In chapter 5 the three approaches for allocating the hotel descriptions to tourist factors are presented. Further, their performance is evaluated and the results are compared and critically reflected. The thesis will be concluded with a brief summary and a recommendation for future work in chapter 7.

CHAPTER 2

State of the Art

In this chapter an introduction to roles in the tourism domain is presented. Further, the basic principles of the Seven Tourist Factors are described. Moreover, an insight into existing approaches which perform recommendations in the tourism domain using text mining techniques is given.

2.1 Tourist Roles

Already in the 1970s tourism as well as its research area became more popular. One of the pioneers in tourism research was [Cohen, 1972] who sociologically studied the motives and reasons for people to travel. In his understanding, travelling is always a combination between getting to know new cultures and landscapes but still wanting to keep familiar parts, which remind them of home. Cohen describes this as the "environmental bubble", which many people are not willing to leave during their travels. Therefore, tourism can be seen as a trade-off between novelty and familiarity.

Based on this research he established four different tourist roles and described their behavioural pattern.

The **Organized Mass Tourist** is a very conservative tourist who wants to avoid all kind of unforeseen things during vacation. Therefore, he likes to travel in groups and have all activities planned by the tourist agency. In the best case, he gets a glimpse of the foreign country but does not have to step out of his environmental bubble at all.

Similar to the first one, the **Individual Mass Tourist** likes to have structured travelling but not every detail is planned beforehand. Thus, he prefers travelling alone to group tours, but lets his travels be arranged by a tourist operator.

The **Explorer** tries to get to know foreign people, cultures and their habits. He does not rely on tourist agencies to book his travel, he organises it himself. Although he leaves his environmental bubble, the Explorer still needs at least the possibility to return to it when necessary.

The most adventurous role is the **Drifter**. He is not only travelling but he wants to dive into the foreign places, its traditions and cultures. The least he wants is to appear like a normal tourist, therefore, he avoids common tourist attractions or accommodations. He is likely to stay in the foreign country and adapt the habits and lifestyle of the local people. In some cases he completely loses his sense for familiarity and his old home.

During their scientific research in the tourism domains [Yiannakis and Gibson, 1992, Gibson and Yiannakis, 2002] studied the correlation between age, sex, education and family backgrounds and touristic preferences. They conducted a questionnaire with over 1200 participants from the United States. The questionnaire consisted of three different parts. People were asked about their activities on vacation, the sanctification of their needs at their point of life as well as demographic characteristics (age, level of education, etc.).

Finally they came up with 17 different tourist roles which reflect the diverse touristic behaviours: Sun Lover, Action Seeker, Anthropologist, Archaeologist, Organized Mass Tourist, Thrill Seeker, Explorer, Jetsetter, Seeker, Independent Mass Tourist (I + II), High Class Tourist, Drifter, Escapist (I+II), Active Sports Lover and Educational Tourist. [Gibson and Yiannakis, 2002] could identify three different trends in tourist role patterns: The interest for certain roles decreases, increases or varies over lifetime. For example the Active Sports Tourist is among the roles which are not common tourist behaviours after a certain age. They are merely interesting for younger people who have better physical abilities than older ones. An example for an increase in preference over lifetime is the Anthropologist, who likes to get to know local people and culture. Some reasons are the raising need of company and the feeling of loosing the connection to one's roots in a higher age. The trend of the Independent Mass Tourist, for example, is variable and reaches its peak at the age of 46 to 49.

The **Seven Tourist Factors** presented in [Neidhardt et al., 2014a,b] play an important role in this thesis since they represent the user groups who the hotels shall be recommended to. The factors reflect different travel needs and characteristics of tourists. As stated in the previously mentioned papers it is somehow difficult for many people to clearly express their travelling expectations and desires. Therefore, Neidhardt et al. created a picture based recommender system, where users have to select pictures to identify their tourist roles. A profile could be either one of the Seven Factors or a combination of some of them. The algorithm has been integrated into [Pixtri OG, a]. For identifying the factors Neidhardt et al. established a questionnaire where the characteristics of the big five personality traits (extraversion, agreeableness, conscientiousness, neuroticism and openness to experience) by [Goldberg, 1999] as well as the 17 tourist roles by [Gibson and Yiannakis, 2002] have been taken into account. The questionnaire was distributed and in the end 997 valid questionnaires were identified and evaluated with a factor analysis. Based on the results, the number of factors could be reduced to seven.

The main characteristics of the Seven Factors were collected from [Neidhardt et al., 2014b] and [Pixtri OG, b] and summarised to enable a better understanding of the thesis.

For the sake of simplicity and recognition in further sections of the thesis, a general short name was defined for each factor.

Sunlover This factor is fond of sunbathing and warm weather whereas he avoids museums and crowded places. He is a mixture of tourist role *Sun Lover* and the personality trait *Neuroticism*. Since also strong correlations with *Openness* and *Conscientiousness* are found, it is also important for the factor to be connected to his friends, family or job via internet access and phone.

Educational During his vacation this factor wants to broaden his knowledge by travelling with groups or organised tours. The factor would rather discover or experience something than relax on the beach. He is highly related to the Big-Five-trait *Agreeableness* and to the roles *Organized Mass Tourist* and *Educational Tourist*.

Independent This factor is on a search for inspiration and the sense of life. He prefers independent travelling and wants to immerge into the history and tradition of famous locations. He is a combination of the tourist roles *Independent Mass Tourist I & II* and *Seeker*.

Cultural The main interests of the cultural tourists are the history of ancient civilisations, arts and culture. Further, he prefers first class hotels and premium restaurants as well as a modern interior design. The factor is a combination of the trait *Extraversion* and the tourist roles Archaeologist and *High Class Tourist*.

Sportive The sportive traveller is very active and wants to get to know the country and its traditions and meet local people during his journeys. He avoids ordinary touristic routes and areas of intense tourism. The factor is highly related to the roles *Anthropologist* and *Active Sports Lover* and to the Big-Five-trait *Extraversion*.

Riskseeker The main preferences of this factor are action, fun and adventures. He likes parties and wants to enjoy the night-life during his vacation. He is a mixture of the three tourist roles *Action Seeker*, *Explorer* and *Jetsetter*.

Escapist This factor highly enjoys silence and the peace of nature. His main aim during holidays is to escape the everyday life and clear his thoughts while relaxing on a deserted beach or in a tiny park. He avoids crowded places and large cities and combines the tourist roles *Escapist I & II*.

The main aim of the thesis is to identify hotels which would be interesting for one of these Seven Factors. With the knowledge of a person's tourist role and the hotels allocated for each corresponding factor, the recommendation system could be easily extended and a recommendation of suitable hotels for a user could be enabled.

2.2 Text Mining Based Recommender Systems in the Tourism Domain

A recommender system is a software which is able to predict the interest of a user in a certain item. This helps the user to select the most relevant item out of a large quantity of offered items. In the tourism domain this can be related to travel, hotel or destination recommendation.

[Burke and Ramezani, 2011] discuss different domains where recommender systems are applied and present main characteristics. They describe the different items that are offered in the tourism domain as rather homogeneous since they show similar characteristics e.g. hotels or transportation. Further, the value of the items has a rather long time span compared to news articles. However, they point out that recommendation in the tourism sector is still extremely difficult since the user preferences are very unstable. A traveller could go on a city trip and two weeks later he could prefer to go skiing. This sudden change in preferences cannot be detected by an automated system. Therefore many travel recommender systems request an explicit input of the user, which is required to understand his motives. Moreover, booking a trip is often expensive which means that the risk of a user to rely on recommendation is also higher than in other domains. Therefore, it can be helpful to give the user the information why a certain hotel or tour was recommended.

Currently, there is a great amount of literature available which deals with the topic of text mining in recommender systems. Therefore, the following literature is just a selection which is considered most influential and related to the topic of the master thesis. Since the process of text mining as well as the available text mining methods and techniques are steadily improving, the focus was on literature of the last five years.

The research of [Lahlou et al., 2013] investigates text mining methods in order to extract contextual features from hotel reviews created by users. The main aim is to enable context aware recommendation by implementing a Context Aware Recommender System (CARS). The data source used are online reviews of different hotels collected by the online platform Trip Advisor. Each of the collected reviews was assigned to one out of five different trip types ("Business", "Couple", "Family", "Friends" and "Solo"). By establishing a model for processing and classifying the reviews, it should be possible to automatically identify the trip type. In order to preprocess the unstructured data, methods like tokenization, stemming, stop words elimination, term frequency thresholding were applied. A frequency-based weighted vector space model was implemented to enable further processing with machine learning algorithms.

For the grouping to the five trip types, three different classification algorithms are used: Support Vector Machine (SVM), k-Nearest Neighbour (k-NN) and Multinomial Naïve Bayes. With the 10-fold cross validation method the models were trained and evaluated. The F1 measure, which is a combination of precision and recall (detailed

description in section 5.2.2), served as a performance measure. The best results were achieved with the Multinomial Naïve Bayes classifier with an average F1-measure of 60.1 percent. However, all three classifiers did not achieve promising results. The reasons for the performance of the classifiers are that many reviews do not refer explicitly to the trip type. Often they are created because the user wants to share his or her opinion about the hotel rather then explain the context of the trip.

The research purpose of this work is closely related to the one described in this master thesis. Although the data source were reviews, the used methods and algorithm can be applied for solving the problem of this thesis. The main challenge regarding text mining of reviews are the different authors and their writing styles as well as the purpose of the review, which is mostly sharing an opinion. It can be assumed that the hotel descriptions by travel operators are mostly more objective, or at least not negative since they should convince the reader to book the described project.

[Cosh, 2013] discusses an application of statistical natural language processing algorithms to a set of articles from Wikipedia about top tourist destinations. The author wants to automatically detect key features of destinations and then compare and cluster those destinations. The main aim of the work is the automated support for users to discover alternative destinations.

For achieving this standardised methods for natural language processing are implemented like the deletion of stop words as well as words with less than 2 letters and tokenizing. To extract the keywords and features of a destination the frequency of the tokens is calculated with the log likelihood comparison. This method combines the frequency value of each word with the frequency value of the word in a large standardised corpus. The goal is to identify words which are very frequent but do not have any information content regarding the topic itself. After calculation each article is then represented by a set of keywords and their log likelihood values. The log likelihood values of each word are then used to form a content cloud for each destination.

The next step is the comparison of the articles and the allocation to groups, which should further identify similar destinations. This is performed by calculating the similarity of all the articles using the RV coefficient introduced by [Robert and Escoufier, 1976].

For grouping the content clouds a bottom-up clustering approach was implemented. The advantage of this method is that the destinations can be ranked regarding the similarity. The main results that could be derived were that the destinations were mainly clustered by country, which was often a high ranked word in the content cloud for each article. To evaluate the results of his proposed solution Cosh performed a user survey where people had to assess how satisfied they were with the proposed alternative destinations. A majority of participants were pleased with the results.

The main limitations the author is emphasising are the subjectiveness of the evaluation method as well as the size of the data set, since only 100 different destinations were involved. Another point is that Wikipedia is a very unstructured data source, with many different authors and their individual writing styles. This makes standardised language processing very challenging.

hotelId	reviewId	food	room	location	facility	service	others
H_1	1	2.00	2.00	1.33	3.00	1.00	1.79
H_1	2	1.50	2.00	2.25	2.33	NULL	1.60
H_2	7	1.86	1.31	1.50	NULL	1.00	1.81

Table 2.1: Feature matrix: attributes (nouns) with corresponding polarity score

An approach for a multi-criteria recommender system for hotel recommendation is introduced by [Sharma et al., 2015]. The system combines rating parameters related to a hotel as well as the content of corresponding user reviews. The main aim is to create an algorithm which detects the most suitable hotel according to the user's preferences. For this purpose they collected 1000 reviews from booking.com whereas 800 were used as a training and 200 as a test set for their models. The following preprocessing steps were executed to extract the content from the unstructured review data. After loading the reviews into a database the sentences were tokenized and the stop words were removed. Further, it was necessary to do a language correction since writers of reviews often use abbreviations or create new words (e.g. "the room was gr8"). Finally the tokens were stemmed and a Part-Of-Speech-Tagging (POS-Tagging) was executed. After preprocessing the nouns that were occurring most were selected as the features and for each feature an individual polarity score was computed for each review. For the calculation the linked sentences where the noun is included were taken into account. Some examples of the feature matrix established by Sharma et al. is shown in table 2.1. To enable a multi-criteria rating, an aggregation function was implemented in order to combine the individual criteria. The authors used a collaborative approach to collect the user's preferences for recommendation. Therefore, they implemented an interface where the user has to state his favoured city, the trip type, his nationality and the personal interest in each of the defined criteria. The user preferences were compared to the database entries and based on the similarity in the criteria, more weight were given to them. Finally the weighted scores were aggregated and the top 5 hotels were recommended to the user. The results for the test data showed that an overall accuracy of 64% was reached.

CHAPTER 3

Data Acquisition and Analysis

Selecting one or more data sources and acquiring them is one of the most important steps during data mining. The quality and also the amount of available data strongly impact the outcome of the whole data mining process.

3.1 Description of the Data Source

The data source used for the establishment of algorithms and models in the context of this thesis are textual descriptions of hotels. The hotel descriptions are written by different tour operators and are mainly collected from websites or catalogues. Thus, the main aim of these descriptions is to give the customer an insight of the facilities and the location of the hotel in order to arouse his interest. The data set is provided by GIATA GmbH [GIATA GmbH], a German company which is specialised in collecting data from different travel operators e.g. TUI and FTI and providing a Global Distribution System (GDS). A GDS is a network which enables access to travel related information like hotel catalogues, bookings, geo-data, hotel descriptions, images and more. The database of GIATA provides more than 364.000 hotel entries and booking codes from more than 392 suppliers. GIATA does not only offer textual descriptions of hotels but also a database of attributes, the so called "GIATA Facts". These facts list different characteristics of a hotel including facilities, number of rooms, meals, locations and more.

Figure 3.1 shows a simple representation of the main entities in tourism and travel industry based on the descriptions in [Werthner and Klein, 1999]: Tourists, intermediaries and suppliers. Among the suppliers are companies which offer "products" like hotels or restaurants as well as airlines. Although it can be the case that consumers directly communicate with the suppliers, nowadays it's very common that the communication is done via an intermediary. Among them are e.g. travel agencies or, as in the context of this

Figure 3.1: Overview of the relationships in tourism industry focusing on hotels.

thesis, tour operators which combine products of suppliers to a new offer for the consumers.

The data source used for his project are textual descriptions created by different travel operators and published in their catalogues, websites or via travel portals. Therefore, this information source cannot be considered as completely unbiased. Hotel descriptions might be positively formulated and in some aspects communicating a better image for customers. The descriptions are meant to sell a product, which is in this case a hotel. In context of this work it was assumed that the factor of bias can be considered as similar for all different tour operators since they all intend to promote their hotels. However, the aim of this work is not to match a user profile with the hotel itself but only with its description. Furthermore, it is not in the context of this thesis to analyse the sentiment of the text.

In the following section the structure of the used data will be described as well as its significance in the context of this work. The data was provided by GIATA via an .xml-file with an amount of 30 311 025 hotel descriptions by different tourist operators. The following code shows an example of one of these dataset entries that would be received if searching for one random hotel description in the GIATA Database.

```
<result found="1">
  <data id="0">
    <GiataID isSpecialID="false">14181</GiataID>
    <Hotelkategorie>3,5</Hotelkategorie>
    <Stadtname>Vancouver</Stadtname>
    <Zielgebietsname>British Columbia</Zielgebietsname>
    <Stadtnummer>4216</Stadtnummer>
    <Zielgebietsnummer>519</Zielgebietsnummer>
    <Landname>Kanada</Landname>
    <Landcode>CA</Landcode>
    <Veranstaltercode>TOC</Veranstaltercode>
    <GeoData>
      <GiataID>14181</GiataID>
      <Latitude>49.278600393684</Latitude>
      <Longitude>-123.12598139048</Longitude>
      <Accuracy>1</Accuracy>
```

```
    </GeoData>
    <Text lang="de"><![CDATA[<br /><b>Lage: </b><br />
    Das Hotel liegt in der N&#228;he der Gesch&#228;fte und
    Restaurants von Yaletown, des Einkaufsviertels Robson
    Street und der Clubs von Gran Ville.<br /><br />
    <b>Ausstattung: </b><br />Das renovierte Hotel mit 245
    Zimmern bietet ein Restaurant, Bar, Hallenbad und Sauna.
    WLAN im Hotel ist kostenlos...]]>
    </Text>
    <Katalogname>Nordamerika 01.04.2014 - 31.03.2015
    </Katalogname>
    <Veranstaltername>Thomas Cook</Veranstaltername>
    <KatalogSaisonTyp>S</KatalogSaisonTyp>
  </data>
</result>
```

Besides the hotel description ("Text") itself there are several other attributes which provide information regarding the tour operators, the hotel and the hotel characteristics and facilities.

The **GiataID** is a unique identifier allocated by GIATA, which makes it possible to identify all hotels or group of hotels gathered in the GIATA database. The attribute *isSpecialID* is important to separate between a GiataID for a single hotel (*isSpecialID="false"*) or a hotel chain (*isSpecialID="true"*). For this thesis hotel chains were not considered since their hotels can be located in completely diverse countries and therefore have different characteristics. Further, the aim of this work is to match one or more tourist roles to a single hotel and later provide recommendations, which can be based on the position of the hotel. Considering this, hotel chains do not seem appropriate to perform a reasonable allocation between tourist factors and hotels.

The tag **Hotelkategorie** describes the category of a hotel based on a classification by hotel stars. However, the hotel category can hardly be used to compare different hotels since there are no specified international criteria to award hotel stars. The HotelStars Union [Hotelstars] is an association of fifteen countries in Europe, which provides a harmonised classification for hotels in the participating countries. Together they defined a hotel stars system with a range from one to five stars. However, there might be rating companies in other parts of the world who rank hotel with stars using different ranking criteria. Although the information regarding hotel stars is not completely reliable, it can help to identify the characteristics of a hotel.

The following tags provide information about the location of the hotel. **Stadtname** states the name of the city the hotel is located in whereas **Stadtnummer** is a digit which represents a unique identifier according to the city. The same properties apply for **Zielgebietsname** and **Zielgebietsnummer**, which specify the area of the hotel e.g. a state or

province. The tag **Landname** states the name of the country where the hotel is located and **Landcode**, a combination of two alphabetical characters, serves as a unique identifier.

Within the tag **GeoData** the longitude and magnitude of the hotel position as well as the accuracy of measurement are specified.

The **Text** includes the description which will be analysed in the course of this thesis. The attribute *lang* states the language the text is written in. In the context of this work only German texts were used for evaluation (*lang=de*). In the majority of cases the description is written in HTML-format due to the fact that GIATA mostly gathers its information from online resources.

Katalogname refers to the name of the catalogue the hotel description can be found in whereas **KatalogSaisonTyp** marks the type of season, the catalogue is written for.

The tag **Veranstaltername** gives information about the hotel operator who provided the description of the hotel. The **Veranstaltercode** is a unique identifier for the operator consisting of two to five numerical or alphabetical characters.

3.2 Implementation of the Database

Based on the information provided by GIATA via an .xml database dump, a MySQL database was implemented to allow a dynamic access and data normalisation. [Oracle Corporation, b]

A Java program was written to read the .xml file received by GIATA, extract the data and transmit the information into the database. The program was developed using the Java SE Development Kit Version 8, Update 45. [Oracle Corporation, a]

Figure 3.2 shows the entity-relationship model of the database. One data result of the .xml file was split to three different entities (hotel, description and provider) to eliminate the main data redundancies. The entity provider represents the tour operator. The entity description does not only contain the text but also some values which were calculated during the set up of the database. The attribute *length* is the number of words of the hotel description. The attribute *diversity* is a measure of lexical richness of a text and was calculated as the ratio of the number of unique words (vocabulary) in a text and its total number. The *ProviderID* and the *HotelID* are foreign keys which enable the allocation of one description to one hotel and one provider entity. How and in which context the additional information for a text was used, will be described in subsequent chapters.

The facts available in the database dump provided by GIATA were not completely transferred to the MySql database. The information corresponding to geodata as well as to the catalogue were not considered to be in the focus of this master thesis.

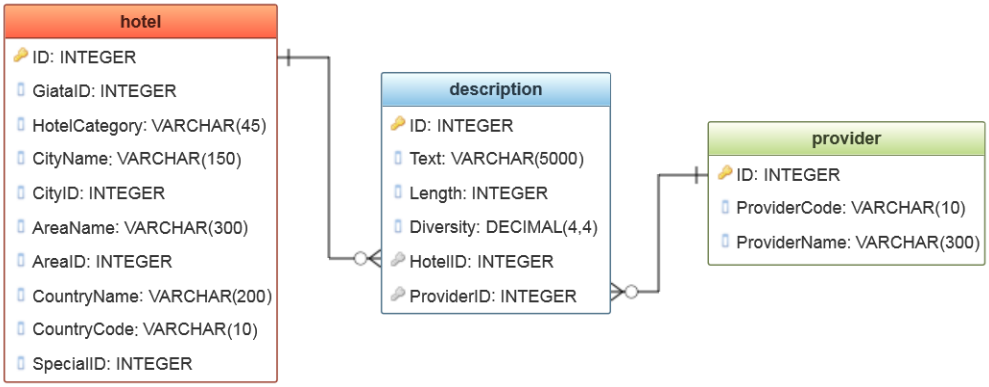


Figure 3.2: Entity-relationship model of the implemented MySQL database

Table	No. of database entries
hotel	48.963
description	210.802
provider (tour operator)	67

Table 3.1: Initial data sample

The number of entries for all three database entities is presented in table 3.1. 210.802 hotel descriptions and 48.963 hotels are presented by 67 different providers (tour operators). According to that, each provider writes an average of about 3.146 hotel descriptions and the mean value of hotel descriptions of one hotel is about 5.

3.3 Qualitative Analysis of Hotel Descriptions

A manually performed analysis is an important step leading to a better understanding of the text data. Since it was not possible to manually read through all the hotel descriptions due to resource constraints, a random sample of ten hotels was selected for the qualitative evaluation. Although ten hotels are not a critical mass of the whole data set and therefore not representative to draw a general conclusion, this step is crucial for further selection of proper methods for a fully automated analysis. The qualitative analysis was helpful in terms of getting an insight of the hotel descriptions, their structure, syntax and the content. It especially laid a foundation for further detailed evaluation of the results obtained by the established models and algorithms.

The ten analysed hotels had an average of 19,7 hotel descriptions per each. The evaluation covered texts of 39 different tour operators out of 68 available in the whole dataset. The goal of the analysis was to get a first impression of the quality and infor-

mation content of the text data and see if a manual allocation of hotel descriptions to tourist roles is feasible.

The hotel descriptions were evaluated regarding facilities, atmosphere and characteristics of the hotel. They were separated in three categories, based on the information content of the text:

1. **High**: Text including a detailed explanation of the ambience of the hotel which is important regarding the allocation of a hotel description to the Seven Factors.
2. **Medium**: Text including short information about the ambience of the hotel.
3. **Low**: Text including only an enumeration of GIATA facts and hardly valuable content.

The following conclusions could be drawn based on the outcome of the qualitative analysis:

More than one hotel description per tour operator
All of the selected hotel data sets involved more than one description of the same hotel operator. This is the case since the textual data was published via different media channels or in different travel catalogues. Therefore, more than one description per tour operator was collected in the GIATA database. All texts of one operator for one hotel were similar except for dates, prices or special offers which were not included in every catalogue or media platform. The longest texts were in all cases those which contained the most information since they included all the information of the shorter texts as well as additional offers. If texts were equal in length they often differed only in numbers for dates or prices which can be ignored in terms of further analysis. The conclusion can be drawn that **the information content of a text is proportional to its length**. This can be emphasised having a look at figure 3.3. Using boxplots, the distribution of the number of tokens (text length) of the evaluated hotel descriptions are compared. It can be seen that hotel descriptions with a higher information content contain a higher number of tokens as well. For that reason only the longest text of one tour provider was included in further analysis steps.

Several tour provider offer equal/similar descriptions
Not only texts of one hotel operator were equal but also those of different operators were identical. For example in all cases the descriptions of operator *Jahn Reisen* were equal to tour provider *ITS* and the descriptions of *12 Fly* correlated with *Discount Travel*. This fact had to be considered in the later evaluation process since a combined analysis of equal or similar texts can for example falsify values of frequency or word occurrences.

Serious differences in text quality
Many of the evaluated descriptions included no additional information compared

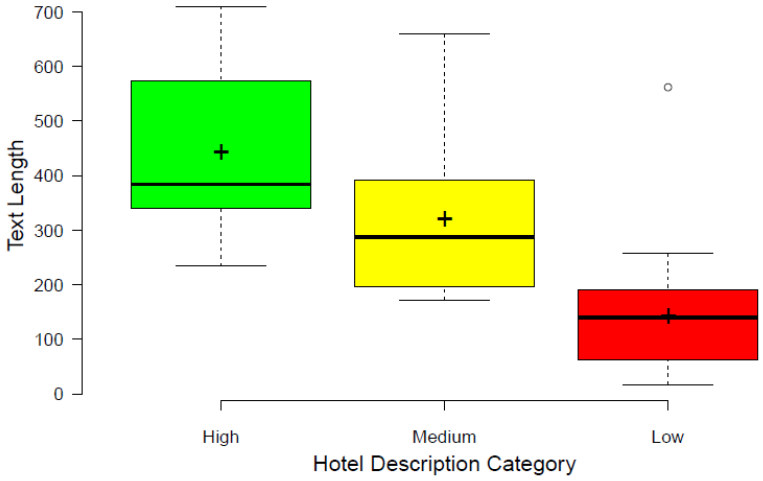


Figure 3.3: Distribution of text length of all hotel description within one category. The quality of the description correlates with the text length (= number of tokens)

to their GIATA Facts. The facts were formed to short sentences that had small or no coherence to each other. Some of them were just enumerations of GIATA Facts. One explanation could be that some travel operators use the GIATA Facts for an automated generation of their hotel descriptions and do not consider other information sources. This fully automated approach would reduce the costs and time of the production immensely. Not even half of the 39 analysed hotel operators offer texts with a higher literary quality and additional information e.g. impressions of the atmosphere and ambience of the hotel. For further evaluation those texts provide the best opportunity for an automated characterisation of a hotel. As a result a method should be presented which enables an approximate estimation of the data quality of a text.

Usage of templates

A majority of the evaluated tour operators uses predefined headline structures in their hotel descriptions. So every description follows a certain format. Beneath the headlines of these "templates" they embed some information gathered from a database or manually written text, depending on the operator itself. One reason for the usage of templates could be that a predefined structure makes it easier for a reader to compare different hotels especially if he or she is only interested in one special feature of a hotel.

"Unused" information for mapping to tourist factors

Some of the hotel descriptions provide information about the hotel's offer for children, students, families, singles or other kind of persons or communities. Using this knowledge a targeted hotel recommendation could be implemented. In case

of the tourist factors defined in [Neidhardt et al., 2014a,b] those elements are not considered since their establishment is based on characteristics of a single person. An integration of this additional context into the recommendation system could be an interesting part of future work.

Considering the outcome of this qualitative analysis, further automated evaluation methods can be applied to the data set. The complete result of the qualitative evaluation can be found in appendix in table A.1.

3.4 Lexical Diversity and Text Length

Lexical diversity is defined as the number of different words in a text (vocabulary) divided by the total number of words (tokens) [Johansson, 2008, Bird et al., 2009]. The result, which lies between a value of 0 and 1, is used as a factor for evaluating the quality of a text. However, lexical diversity alone is not an adequate indicator. Longer texts often have a lower lexical diversity compared to shorter ones. A reason is that the number of total words in a text can be infinite whereas the number of words in a vocabulary is finite, if it is assumed that only words available in any kind of dictionary are used. [Koizumi, 2012] Although the results of a calculation of the lexical diversity can be dependent on the length of the text, it can still be used as an approximate indicator for the further evaluation process. In the scope of this work an exact calculation is not needed. Hence, for a proper use of lexical diversity as a quality measurement, the length of a text has to be observed as well.

For a proper calculation of the lexical diversity and the text length of the descriptions, it was needed to extract the content of the .xml-structure and tokenize it. The methods and implementation behind this will be described in detail in the subsequent chapter "Data Preprocessing".

Figure 3.4 shows the trend of calculated lexical diversity over the whole data set. The values reach from one to zero and can be seen as approximately normally distributed with a mean of 0,72. This value means for example that a text includes 108 unique tokens (vocabulary), although in total there are 150 tokens available. However, there are two peaks which do not fit into the normal distribution. The reason for those peaks are templates used by some of the tour operator to establish hotel descriptions, as described in the previous section. For most of the hotel descriptions generated with a certain template, the same or a similar lexical diversity is calculated. Furthermore, we can identify peaks and outliers at a diversity of zero and one. This is possible because there are texts with a very small amount of words or even blank descriptions. For short texts it is more likely to reach a lexical diversity of one.

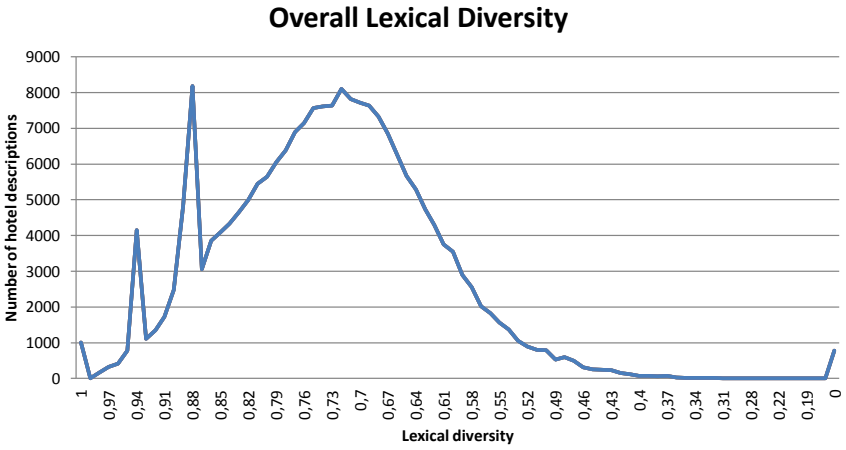


Figure 3.4: Distribution of lexical diversity over the complete data set. Over 8100 hotel descriptions have a lexical diversity of 0,88 or 0,72.

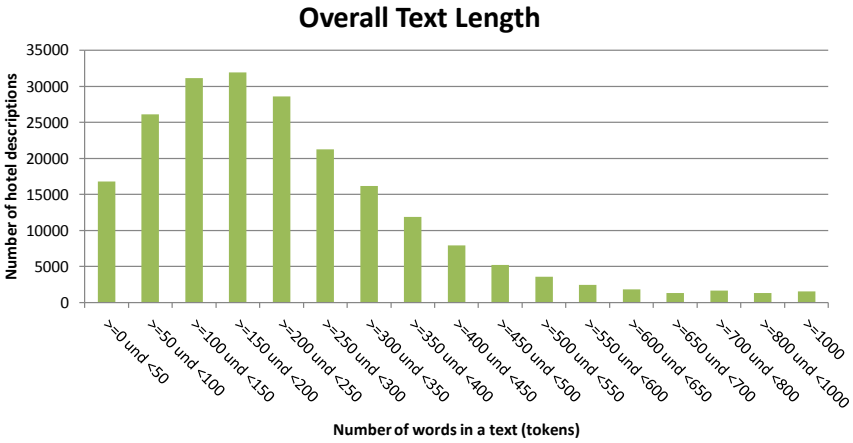


Figure 3.5: Distribution of the text length (number of tokens) over the complete data set. The text length is normally distributed. Around 32.000 hotel descriptions contain 150 to 200 tokens.

The distribution of the length of the available hotel descriptions is shown in figure 3.5. A normal distribution can be identified with a mean between 150 and 200 words per text. The importance of the length of a text depends in the text mining method used. Some methods like counting the word occurrences do not make sense for texts of a short length and a high lexical diversity.

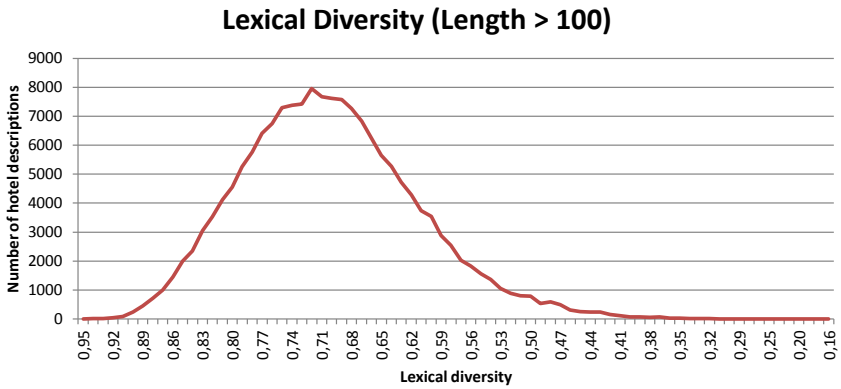


Figure 3.6: Distribution of lexical diversity for hotel descriptions with a text length of more than 100 tokens. The lexical diversity values are normally distributed.

Table	**No. of database entries**
hotel	48963
description	48963
provider (tour operator)	67

Table 3.2: Data sample which will be considered for further steps

The knowledge about these distributions is important to enable an understanding for the whole text data set and also helpful for selecting further evaluation methods.

In figure 3.6 the distribution of lexical diversity for texts with more than 100 words can be investigated. A widely homogeneous normal distribution is shown, where all previously identified peaks are missing. Therefore, texts with more than 100 words can be seen as a good sample in the further text mining process. Almost all outliers are eliminated and it is a good approach to converge closer to homogeneity of the data set. For all upcoming text mining steps, it is important that one hotel is represented by one text. Since texts with more words enable a more homogeneous data set and provide more information gain, for further procedures only the longest description was chosen to represent a hotel. An overview of the remaining text sample in the database is given in table 3.2. Compared to the initial sample described in table 3.1 there is now only one description per hotel left which matches the predefined requirements.

CHAPTER 4

Data Preprocessing

In order to enhance the results for algorithms grouping text documents, the raw texts have to go through certain preprocessing steps. The complete preprocessing was applied using the data science tool Rapidminer (Version 7.2.1) [RapidMiner, Inc., a] with the Text Processing (Version 7.2.1) as well as the Web Mining extension (Version 7.2.1) for handling the html content of the descriptions. Figure 4.1 shows the impact of the preprocessing steps on an example of a hotel description.

4.1 Extraction of Content and Text Encoding

Most of the raw hotel descriptions were collected from websites, which means they are stored in html format. Therefore, all the html tags had to be removed so that the textual content remains and can be processed further. Additionally the words were all transformed to lower case. Text encoding is especially relevant because of the German umlauts, which have to be dealt with. Therefore, the text descriptions had to be converted to utf8 for enabling an easy handling.

4.2 Tokenization

Tokenization is the base for the extraction of higher level information from a document as described in [Weiss et al., 2010]. The whole text is divided into individual units, in this case single words. A simple tokenizer where all non-letters represent separators was used. Numbers and punctuations were removed as they had no useful semantic meaning for the concept identification.

Raw Text	<![CDATA[Lage: Das Hotel liegt in der Nähe der Geschäfte und Restaurants von Yaletown, des Einkaufsviertels Robson Street und der Clubs von Gran Ville. Ausstattung: Das renovierte Hotel mit 245 Zimmern bietet ein Restaurant, Bar, Hallenbad und Sauna. WLAN im Hotel ist kostenlos.]]>
XML Extraction	Lage: Das Hotel liegt in der Nähe der Geschäfte und Restaurants von Yaletown, des Einkaufsviertels Robson Street und der Clubs von Gran Ville. Ausstattung: Das renovierte Hotel mit 245 Zimmern bietet ein Restaurant, Bar, Hallenbad und Sauna. WLAN im Hotel ist kostenlos.
Tokenization	Lage|Das|Hotel|liegt|in|der|Nähe|der|Geschäfte|und|Restaurants|von|Yaletown|des| Einkaufsviertels|Robson|Street|und|der|Clubs|von|Gran|Ville|Ausstattung|Das| renovierte|Hotel|mit|Zimmern|bietet|ein|Restaurant|Bar|Hallenbad|und|Sauna|WLAN| im|Hotel|ist|kostenlos
Stop Words Removal	Lage|~~Das~~|Hotel|liegt|~~in~~|~~der~~|Nähe|~~der~~|Geschäfte|~~und~~|Restaurants|~~von~~|Yaletown|~~des~~| Einkaufsviertels|Robson|Street|~~und~~|~~der~~|Clubs|~~von~~|Gran|Ville|Ausstattung|~~Das~~| renovierte|Hotel|~~mit~~|Zimmern|bietet|~~ein~~|Restaurant|Bar|Hallenbad|~~und~~|Sauna|WLAN| ~~im~~|Hotel|~~ist~~|kostenlos
Stemming	lag|hotel|lieg|nah|geschaf|restaura|yaletown|einkaufsviertel|robson|stree|club|gran|vill| ausstatt|renovier|zimm|biete|bar|hallenbad|sauna|wla|kostenlo
Pruning	lag|hotel|lieg|nah|geschaf|restaura|~~yaletown~~|einkaufsviertel|~~robson~~|stree|club|gran|vill| ausstatt|renovier|zimm|biete|bar|hallenbad|sauna|wla|kostenlo

Figure 4.1: Data preprocessing steps applied on an extract of a hotel description

4.3 Stop Words Removal

Words whose semantic content does not matter in case of classification or selection of a certain document are called stop words. In terms of this thesis the stop words were identified via the Natural Language Tool-kit German stop word corpus [NLTK Project]. Additionally words with less than two characters were deleted. The removing of those tokens enables a more efficient evaluation in terms of frequency of word occurrences.

4.4 Stemming

After preprocessing the text, the next step was to reduce the number of different tokens by grouping them semantically. Many tokens which are syntactically written in a different way belong together in a semantic way. Therefore the term *type* was introduced, which groups many instances of different tokens as described in [Weiss et al., 2010].

The grouping can be performed using different algorithms. As part of this thesis stemming has been assumed to be the most efficient one. Stemming is a text mining method to normalise the variety of words in the text, which may have the same principal part. Within the scope of this work, the German stemmer in the Text Processing plug-in of Rapidminer was implemented. The algorithm is described in detail in [Caumanns, 1999].

To determine the impact of stemming and the potential benefit for this specific data sample, a simple comparison of the number of token before and after stemming was executed. Stemming decreased the number of overall tokens in the data set by 20%. This reduction enhanced a faster execution of the grouping algorithms described in the subsequent chapter.

4.5 Pruning

Additionally to stop words, other words that do not have semantic value shall be removed from the document set. This decreased the number of attributes resulting from the conversion of text into the word vector as described in the next section. Further, it enhanced the speed and efficiency of subsequently used matching algorithms. It can be assumed that words which only occur in very few of the hotel descriptions do not have a high value for grouping data samples. Therefore, all tokens that were included in less than 5% of all the documents were removed. The example given in figure 4.1 shows that during this step two words were removed: "yaletown" and "robson". Both words refer to a specific location of the hotel. It can be assumed that no other hotel from the same region was present in the data set, hence these words were not included in any other hotel description. This was detected in the pruning step and therefore "yaletown" and "robson" were deleted and not considered for further steps.

4.6 Word Vector Generation

Most classification or clustering algorithms need to calculate distances between entities of the data set. They cannot handle nominal data as an input. Therefore, the preprocessed text has to be transformed to a numerical vector. The set of tokens can now be seen as the set of attributes which are available in the data sample. Every text will have a vector containing a numerical value for each attribute (token). There are different methods to create this word vector. One of the easiest examples is a binary vector, where a "1" indicates that a word or token occurs in a text and a "0" indicates that it is not present in the corresponding text. A different approach is to count the number of each token in the text and use the frequency of the word occurrences.

However, these methods have some disadvantages. It might not only be of importance to know how often a word is used in a text itself, but how often it is used in the whole dataset. With this information the actual information gain of a word can be evaluated. For example, if a word occurs in most of the texts in the dataset, it might not be of use if you want to classify or cluster the documents.

To solve this problem the word vector was generated by using the Term Frequency - Inverse Document Frequency (TF-IDF) method. Term frequency represents the frequency of a word occurrence while inverse document frequency is defined by dividing the total number of texts (N) by those which contain the word (df = document frequency). Furthermore, the idf is logarithmically scaled as described in [Weiss et al., 2010].

	ausstatt	bar	biete	club	geschaf	gran	hallenbad	hotel
Doc1	0,7	0,43	0,45	0,32	0,25	0,9	0,67	0,13
Doc2	0,32	0,54	0	0,12	0	0	0,4	0,04
.....								

Table 4.1: Example of a word vector for two hotel descriptions

The following equation 4.1 shows the formula.

$$\text{tf-ifd(i)} = tf(i) * log\frac{N}{df(i)} \tag{4.1}$$

To enable a better comparison for machine learning algorithms the TF-IDF score calculated in Rapidminer is normalised. After calculating the score it is divided by the square root of the sum of all TF-IDF values.

In table 4.1 an example of a word vector for two documents is given. The value is zero if a word does not occur in the text at all. It can be seen that "*hotel*" has a rather small lexical diversity since it occurs very often in the documents of the data sets and is therefore ranked lower. Whereas "*gran*" is rather infrequent and therefore the TF-IDF value is higher for Document 1.

4.7 Preparation of Training and Test Set

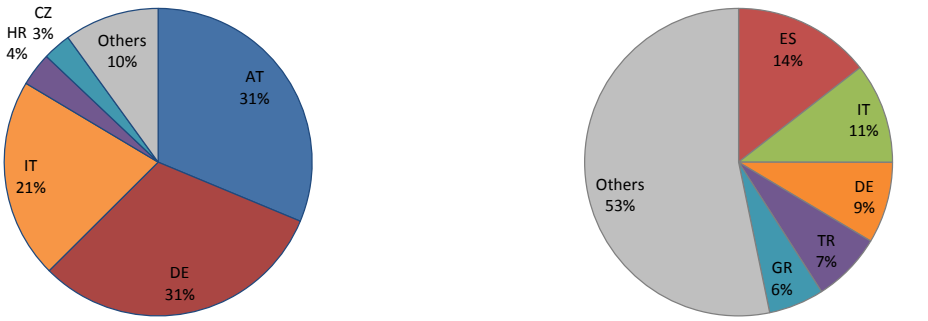
In the next chapter certain machine-learning methods to allocate the hotel descriptions to the factors are going to be presented. For those methods it is important to have one set of data which is already labelled, which means it consists of hotel descriptions which are already assigned to one or more factors. This data set can then be used to train supervised algorithms so that they can automatically classify documents, which are not included in the training sets. Further, the result of unsupervised clustering algorithms and the dictionary approach can be tested and evaluated with it.

For this thesis a labelled set was provided by experts from Eurotours International [Eurotours Ges.m.b.H.], a travel operator and incoming agency. It consists of 551 hotels that are part of the GIATA data set. One entry contains the GIATA ID, the name of the hotel and additionally a matching value for each factor expressed as a score between 0 and 100, as shown in figure 4.2. The higher the score, the more interesting a hotel is for a factor.

To give an insight into the diversity of hotels which are included in the labelled set, the distribution of countries is presented in figure 4.3a. The three most popular regions of the labelled data set are Austria, Germany and Italy followed by Croatia and Czech Republic. Compared to the complete dataset provided by GIATA, differences can be recognised. In the second pie chart in figure 4.3b the country which occurs most is Spain. In total there

GIATA	NAME	Sunlover	Escapist	Independent	Cultural	Sportive	Riskseeker	Escapist
883	Kreta	90	0	20	0	55	10	0
1482	Hotel Garda Bellevue	90	0	10	0	85	0	10
1486	Hotel Malcesine	66	0	18	0	45	0	8
1587	Giardini Naxos	100	21	40	8	22	0	48
5769	Zypern – Sternfahrt	0	100	30	10	20	0	0

Figure 4.2: Labelled data provided by Eurotours International



(a) Country distribution of labelled data set

(b) Country distribution of complete data set

Figure 4.3: Comparison of the distribution of countries in the labelled data set and in the complete GIATA database

are more than 150 countries covered in the GIATA database, whereas only 19 of them appear in the labelled data set. Therefore, the representativeness of the labelled data set might be not as good as possible. However, at least Germany and Italy are in the top three of the most popular countries for both data sets. Due to the fact that by now, no other labelled data set is available which can be used for this problem, the sample will be used as a training and test set in spite of certain drawbacks. To enable an independent training and testing of the established models the data set is split in two parts: The training set includes 371 hotel descriptions and the test set 180, which is a ratio of about 70/30.

Most classification or clustering methods require distinct labels for evaluation or training of the data. They cannot deal with data entries that can be in more than one class or are only allocated to a class with a certain score. Therefore, some preparation of the labelled sample was needed. A threshold had to be defined that indicates at which point a hotel is allocated to a factor. In order to detect this matching value the distributions of the hotels to the factors had to be evaluated. For this approach only the training set was used since this was already a pre-step for establishing a certain model for grouping and allocating the descriptions. Hence, the hotels in the test set were not

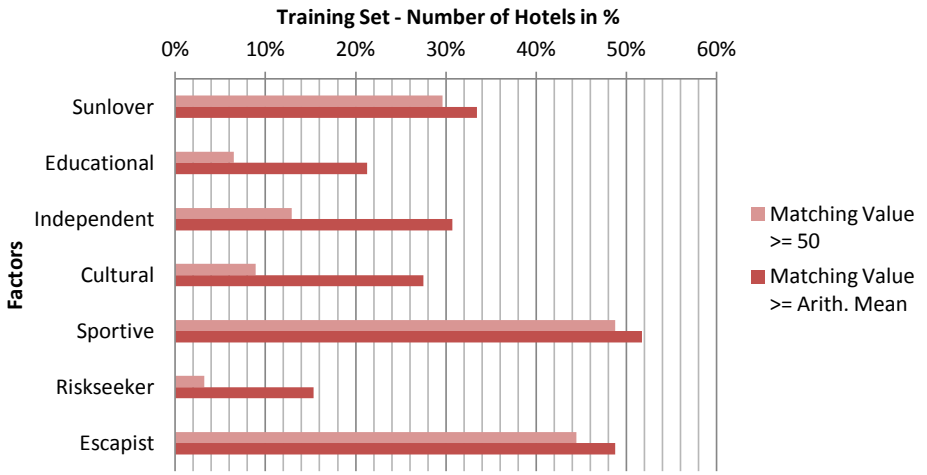


Figure 4.4: Distribution of hotels with threshold at the matching value greater than or equal to 50 and greater than or equal to the arithmetic mean of each factor

included into the establishment of a threshold.

In a first approach the training set was split at threshold of "50" meaning that a hotel was allocated to a factor if it had a matching value of greater than or equal to 50. The first bar for each factor in figure 4.4, marked in light red, indicates the number of hotels which match a particular factor. Educational, Independent, Cultural and Riskseeker had an extremely low share on hotels which are suitable for them. This is a problem for further supervised training algorithms since they need a certain amount of data to train a reliant model. So another method was to consider also lower matching values and apply a variable threshold, which was calculated out of the arithmetic mean of each factor. The dark red bars represent the percentage of assigned hotels over the factors using the arithmetic mean approach. It is shown that the data set looks more homogeneous and for some factors the amount of matching hotels more than doubled. Therefore, these thresholds were used to define the classes for the factors. A disadvantage though will be that also hotels with matching scores lower than 50 are still considered as being allocated to a tourist factor class. On the other hand even hotels with lower values might be of interest for the corresponding factor, otherwise they would have been marked with a matching value of 0.

The binary factor-labels for the independent test set were established with the arithmetic means calculated from the training set. Figure 4.5 shows an overview of the assigned hotels for each factor in training and test set. The distribution is similar, although for Independent, Cultural and Sportive the test set contains about 10% more relevant hotels.

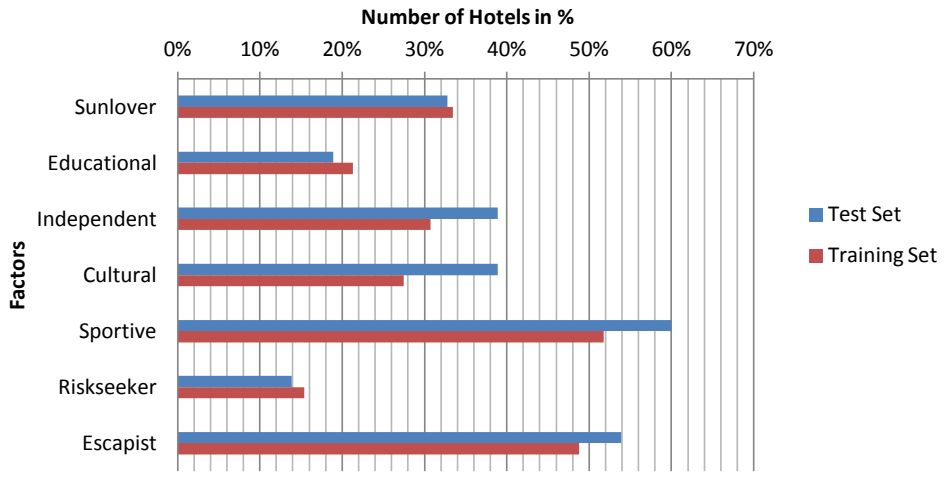


Figure 4.5: Comparison of training and test set. Distribution of hotels with threshold at arithmetic mean calculated for each factor.

	Training Set	**Test Set**
Overall hotels	371	180
Sunlover	124	59
Educational	79	34
Independent	114	70
Cultural	102	70
Sportive	192	108
Riskseeker	57	25
Escapist	181	97

Table 4.2: Overview of the hotels included into test and training set and their relation to the Seven Factors. Since the hotel allocation is not mutually exclusive to one factor, the sum of the hotels of all factors is higher than the overall number of hotels in the data sets.

Concluding this step the data entries in the labelled data set can be precisely allocated to classes of factors by splitting the data at the average mean matching value. This is needed as input for clustering and classification methodologies described in the next chapter. Table 4.2 summarises the results for the hotels corresponding to the established training and test set.

CHAPTER 5

Allocation of Hotel Descriptions to the Seven Factors

For allocating a hotel description to one of the seven tourist factors, three different methods were selected. The individual result for each factor was evaluated and compared to each other.

5.1 Unsupervised Cluster Analysis

For the first approach of matching hotel descriptions with the seven factors, a fully automated method was chosen: Cluster analysis. The set of documents are separated into a certain amount of clusters by an algorithm which calculates the similarity between the word vectors. One big advantage of this method is, that it is not necessary to have a labelled training data set. However, for the establishment and evaluation of the cluster model, the training data set with 372 hotels proposed in section 4.7 was used.

Many different clustering methods exist and it is not easy to select the one which solves the corresponding problem most efficiently. In the context of this thesis, the algorithm k-means was chosen. The algorithm can deal with a large amount of attributes and also data instances in an efficient way and is often used for solving text mining problems [Aggarwal and Zhai, 2012, Steinbach et al., 2000]. It divides the data set into disjoint sets of cluster - so each instance (in this case text-documents) will be present in one cluster only. This is one of the disadvantages since one hotel description can be interesting for more than one factor. Thus clustering can only solve one part of the problem the thesis is dealing with. However, the clustering results can be evaluated and may give insight to what extent the documents can be matched to the factors without using any human input. Further, it can provide information about the correlations of the factors among each other.

The distance based partitioning algorithm k-means computes a number of k clusters by calculating the so called centroid, the center of each cluster as described in [Aggarwal and Zhai, 2012]. The centroid represents the mean of all documents in the cluster which is calculated using the word vectors of the documents. New documents are allocated to the cluster which has the smallest distance between the mean of the document and the centroid. K-means uses a mathematical function to calculate these distances. One of the most efficient measurements for text clustering in information retrieval is the cosine similarity. This was used in the following experiments on clustering the hotel description data set.

To work with the algorithm the user has to specify certain entry parameters. The most important one is the number of clusters, k. This can be a drawback especially for problems similar to the one examined in this thesis, where the number of clusters is not clear from the beginning. Therefore, methods are available to estimate the number of clusters which generates the best information gain for the data sample. Some of these methods will be described in detail in subsequent sections.

Another entry parameter, which can be defined by the user of the algorithm, are the so-called seeds. These are k documents from the data sample whose means represent the starting centroids. Therefore, the efficiency, speed and the accuracy of the algorithm strongly depends on the initial seeds. If documents are chosen which are similar to each other, the algorithm might produce a weaker result than for centroids which have a poor similarity value. For the experiments of this thesis a different approach was chosen, where the seeds are neither selected randomly nor have to be selected by the user. The k-means++ heuristic determines only the first k centroids randomly but afterwards all other data points are weighted according to their squared distance before choosing the next centroid. Hence, the possibility of a selection of centroids which are all close to each other gets smaller.

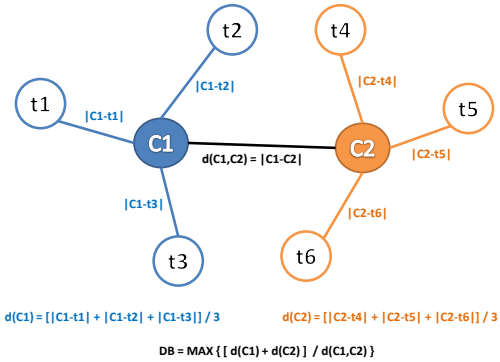
For a more detailed description of the k-means++ heuristic see [Arthur and Vassilvitskii, 2007]. An introduction to k-means clustering as well as the structure of the algorithm can be found in [Gupta, 2006] and [Leskovec et al., 2014].

All clustering experiments were executed with the data science tool Rapidminer [RapidMiner, Inc., a]

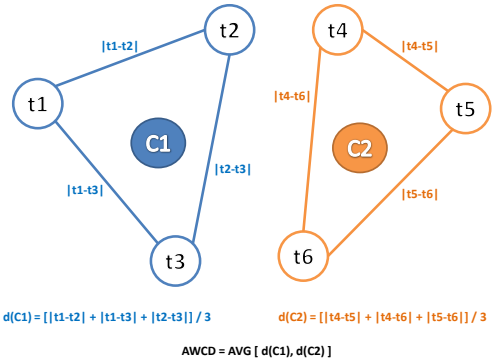
5.1.1 Determination of the Optimal Number of Clusters

For many data sets it is far from obvious what value of k shall be selected in order to receive the best information gain after clustering. Therefore, quality measures of clustering for different values of k can be evaluated and give hints which number of clusters are appropriate.

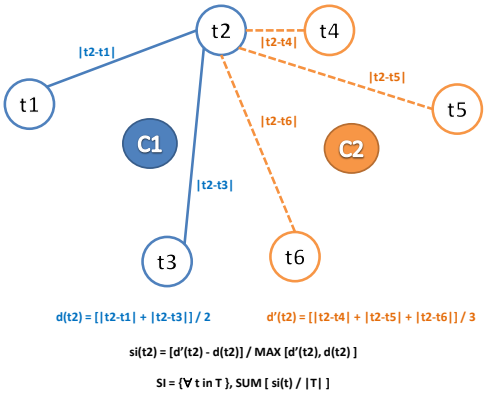
Three of the most common measures were selected, and the results were compared for a k of 2 to 10. Since the test data set consists of 371 documents, it was assumed that more than ten clusters are not meaningful.



(a) Davies Bouldin Index



(b) Average Within Cluster Distance



(c) Silhouette Coefficient

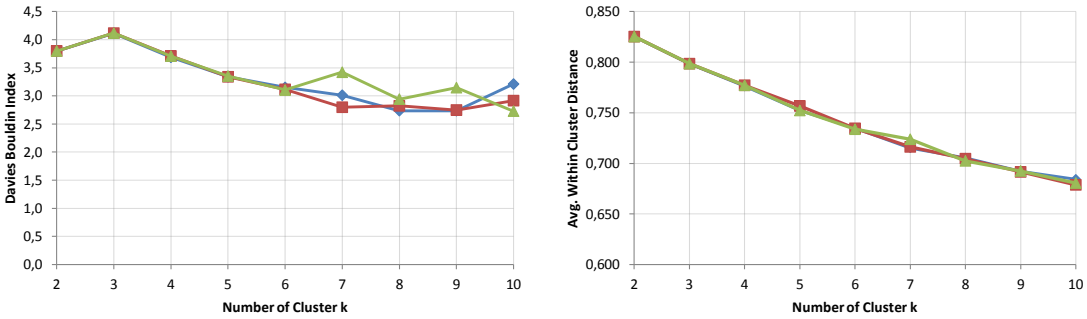
Figure 5.1: Graphical representation of the cluster indices

Davies Bouldin Index

This criteria measures the average similarity between each cluster and its most similar one. It can identify clusters which are compact and also far from each other [Davies and Bouldin, 1979]. Figure 5.1a gives an example of the calculation of the Davies Bouldin Index for two clusters.

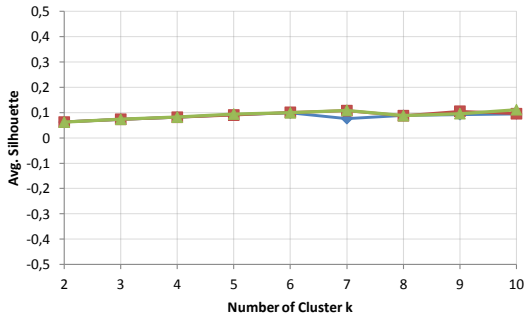
The clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies Bouldin Index. Generally a lower index indicates a better clustering result and also more information gain [Saitta et al., 2007].

Nevertheless, at a certain level of minimisation it is not meaningful to introduce more and more clusters just to lower the value of the cluster performance criteria. For example the minimum value of Davies Bouldin for a data set with ten documents is reached by implementing ten clusters whereas the information gain for the clustering is zero. Hence, the optimal number of clusters cannot be found by a



(a) Davies Bouldin Index

(b) Average Within Cluster Distance



(c) Silhouette Coefficient

Figure 5.2: Overview of three different indices to determine k. The figures show the index for a value of k from 2 to 10 for a dataset with 371 hotel descriptions.

simple minimisation of the cluster evaluation criteria but by using the so called elbow method. For most of the data sets it can be determined that the Davies Bouldin Index drops steadily with an increasing value of k. First the decline of the index can be very sharp, but at one point it will even out and adding one more cluster will not make sense anymore. A drawback of this method is that there is no guarantee that the elbow can be identified, and further there is no clear specification at which point exactly an elbow exists. It strongly depends on the person who analyses the data set.

In figure 5.2a three curves are shown representing the average Davies Bouldin indices for values of k from 2 to 10. The reason why there are three different curves is that for k-means the result strongly depends on the seeds and therefore it can vary to a certain point, which can also be seen in the figure. In this plot the elbow is not clearly visible but what can be determined is that with seven clusters the result starts to alternate stronger than for smaller values of k. Concerning this it will be safe to use six clusters since with seven it is not guaranteed that the information gain is clearly higher.

Average Within Cluster Distance

This parameter represents the average distance between the centroids and all other data points of the corresponding cluster and measures the compactness of a cluster, as described in figure 5.1b. For the best result the average within cluster distance should be minimised. The cluster validity measure can be used for the determination of k in combination with the elbow method, as the Davies Bouldin Index.

Unfortunately this parameter shows no elbow at all, the curve shown in 5.2b is approximately a linear slope. However, the same behaviour as before can be identified, since from cluster 7 on the results of the three curves differ.

Silhouette Coefficient

The Silhouette Coefficient measures how close each point in one cluster is to points in the neighbouring clusters as described in [Kaufman and Rousseeuw, 1990]. Figure 5.1c gives an example for the calculation of the Silhouette Coefficient for node t1. The range of the coefficient lies between the limits of -1 and 1, whereby 1 represents the optimum. If the value is smaller than zero it indicates that the cluster contains documents which should have been allocated to a different cluster. In 5.2c all the Silhouette Coefficients are close to zero. This points out that it is not explicit to which cluster a document should be assigned. However, it is again visible that six clusters have one of the highest average Silhouette Coefficient since higher numbers are dependent on the initial parameters for k-means.

According to the results of all three parameters it is assumed that six clusters are most appropriate for this data set and further evaluations will be executed with this parameter.

5.1.2 Cluster Evaluation

In this section the concrete results of the clustering will be analysed and described in detail. Considering the conclusions about the number of clusters described in the previous subsection, the value six was chosen as the optimal number of clusters for the labelled data set with 371 hotel descriptions.

In figure 5.3 the Silhouette Coefficient of the cluster experiment is drawn. As already described in the previous chapter, the average coefficient of the complete dataset is positive. In this figure though it can be examined that also for each single cluster, the average Silhouette value is slightly positive. There are only a few outliers with negative coefficients (e.g. in cluster 3 and 4), which means that those were allocated to the wrong cluster. Additionally the plot gives an insight on how the documents are distributed over the cluster. The larger the area of the cluster the more data samples are included.

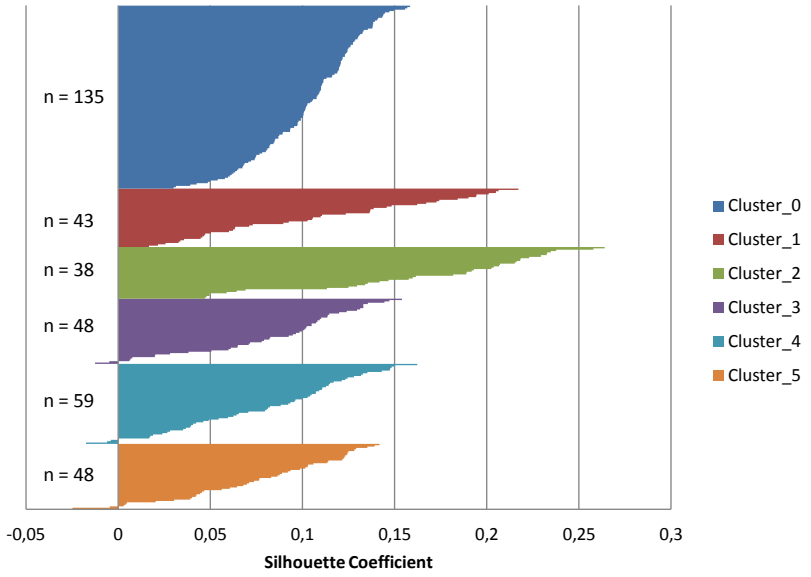
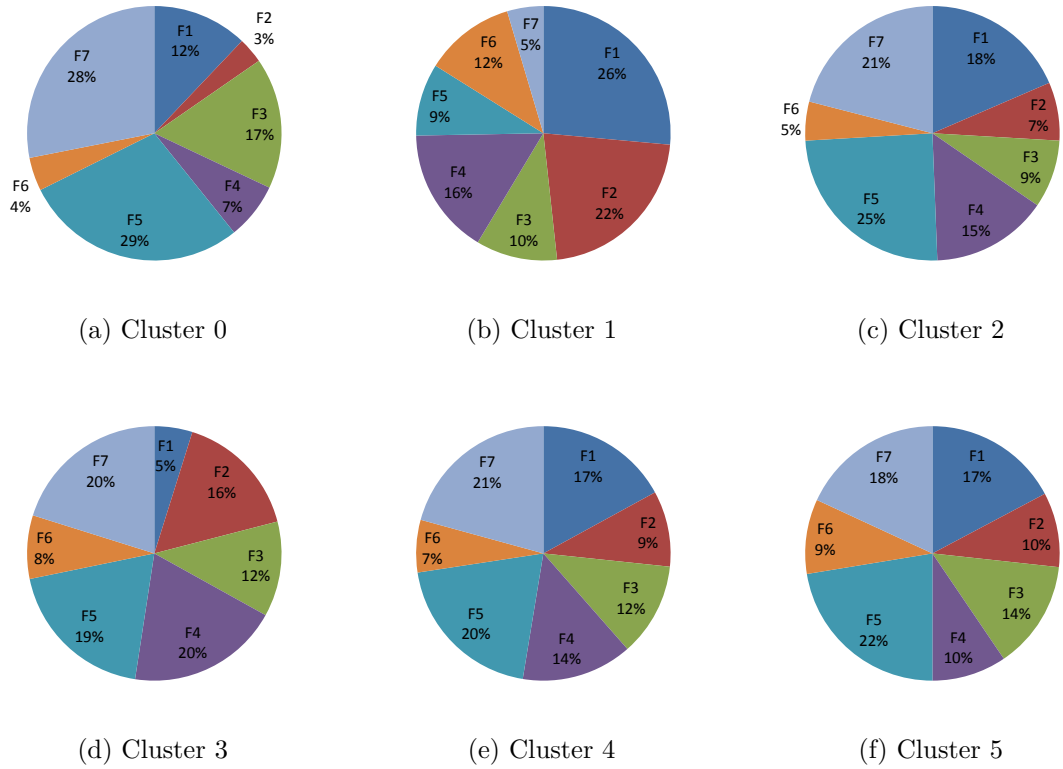


Figure 5.3: Silhouette Coefficient of all documents grouped to their allocated cluster. The variable n indicates the number of documents in the corresponding cluster. On the x-axis the Silhouette Coefficient for each document is plotted.

To find out if it is possible to use cluster analysis for allocation to the seven factors, the distribution of the factors over the six clusters was evaluated. The expected result was to identify cluster which reflect a factor in most instances. Unfortunately it could not be detected that one factor is represented by one cluster only. Figure 5.4 shows the the distribution of factors for all six clusters. Every cluster includes a certain share of every factor. Therefore, it can be identified that there is no factor which can be allocated exclusively to one cluster.

To become an insight on how the cluster algorithm works, the highest ranked words of the centroid had to be examined. In two of the clusters, the names of tour operators appeared in the top six of the most important words. Furthermore, one conclusion of the qualitative analysis of the hotel descriptions stated in section 3.3 was that most of the tour operators use templates to formulate or automatically generate their hotel descriptions. Therefore, the assumption was formed that, if the clustering does not represent the grouping into factors, it could be dependent on different tourist operators. After evaluating the distribution of operators the most remarkable outcome was that cluster 0 contained exclusively all available hotel descriptions of tour operator "Bucher Reisen". Further, in most of the other clusters, at least one dominant tour operator could be found, see figure 5.5. This confirms the expectation that the templates of the operators do have an impact on the hotel descriptions. A reason might be that the automated cluster algorithm does not implement any predefined model to identify which words are



(a) Cluster 0 (b) Cluster 1 (c) Cluster 2

(d) Cluster 3 (e) Cluster 4 (f) Cluster 5

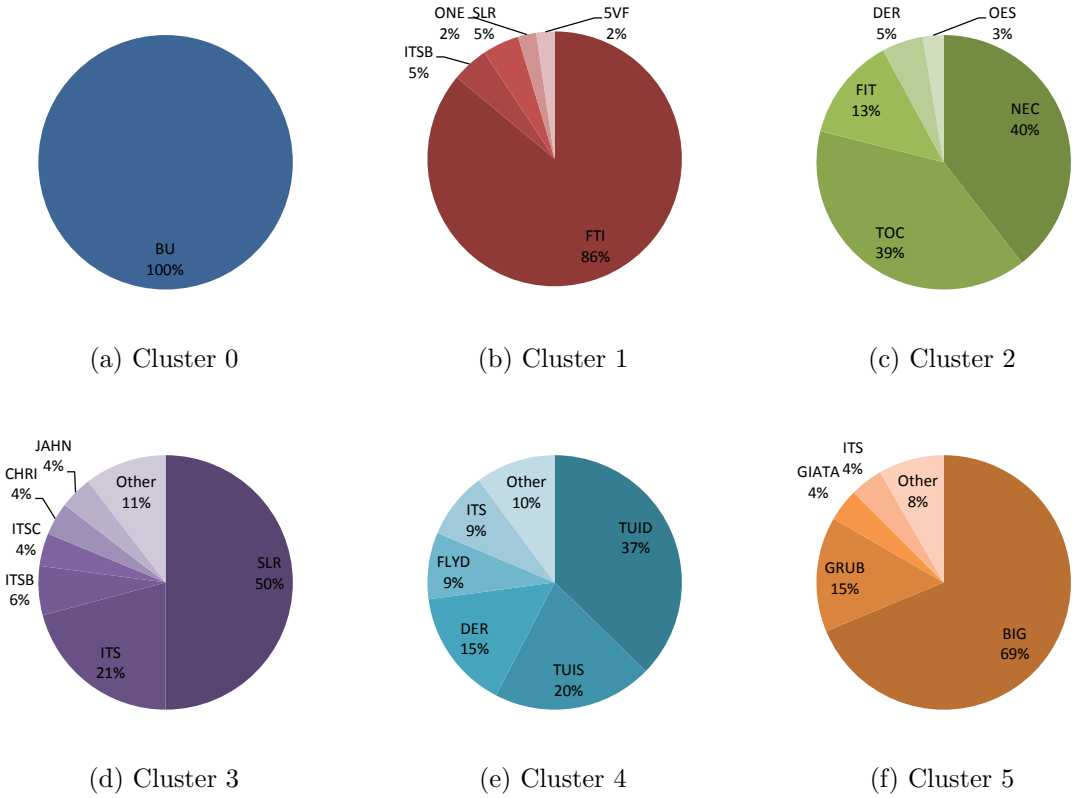
Figure 5.4: Distribution of factors over clusters

more important for defining the clusters suitable for the factors. It only relies on a simple similarity measure that groups documents together which use the same wording. More details about the distribution of hotel operators over the clusters can be found in the appendix (figure B.1).

Considering the output of the executed experiments, a clustering analysis with the k-means algorithm for matching hotel descriptions with tourist factors has turned out as not suitable.

5.2 Supervised Classification

For the results of unsupervised methods were not promising, the following approach deals with supervised classification. Documents are separated into classes based on the calculation of a trained model. In comparison to clustering, classification can be implemented only when a labelled training data set, for which the classes are already known, is available. This training set is needed in order to fit a model which can then be used to classify other data samples as well. The training set described in section 4.7 was used to implement the classification models.



(a) Cluster 0 (b) Cluster 1 (c) Cluster 2

(d) Cluster 3 (e) Cluster 4 (f) Cluster 5

Figure 5.5: Distribution of tour operators to clusters

5.2.1 Overview of the Classifier

In the context of this thesis, three different classification algorithms were used to divide the hotel descriptions into classes: k-Nearest Neighbour, Naïve Bayes and Decision Tree. Similar to the clustering approach described before all of those three classifiers allocate one document to only one single class. This causes a problem since one hotel description can be interesting for more than one factor. Therefore, it is necessary to generate one classification model for each factor. Each of these models divides the data set into two distinct classes. One class represents all documents which can be interesting for one factor, whereas the other class contains all data samples which are not interesting for the same factor. Finally there will be seven different models which correspond to the seven tourist factors. All classification approaches have been implemented using the tool Rapidminer [RapidMiner, Inc., a]. The following algorithms were used to classify the test data set.

k-Nearest Neighbour

This classifier calculates the closeness of each data object to all the other objects in the data set. Afterwards it builds a model which allocates an example to the class that occurs most frequently in the majority of its k nearest neighbours. The

input parameter k can be selected by the user of the classifier. For the experiments within the scope of this thesis a range of k from 1 to 10 was selected and evaluated. As a distance measure cosine similarity was used (see [Weiss et al., 2010]).

Naïve Bayes
Bayesian classification works with the hypothesis that a given example belongs to a certain class and further calculates the probability for this hypothesis with the Bayes' theorem described in [Gupta, 2006]. This algorithm can be trained very efficiently on a small amount of training data and it is quite robust since it delivers a correct classification as long as the correct class is more probable than the others.

Decision Tree
This algorithm can be seen as a tree where each node represents a decision on an attribute and each branch denotes an outcome of this decision [Gupta, 2006]. The tree can have one or more child nodes. This approach is very helpful to gain information about the relationship of the data set. Furthermore, it will illustrate clearly those attributes, in this case tokens, that make the difference between documents that might be interesting for a factor and the rest. To decrease the chance of generating an overfitting tree, pre-pruning was used in the subsequently described experiments. It was the preferred measure to pruning, since the latter did not improve the results of the classification. In order to increase the accuracy of the algorithm, the "minleave" parameter available in Rapidminer [RapidMiner, Inc., a] was varied between 1 and 15 and the best result was taken into account for further steps and comparison to the other classifier.

5.2.2 Validation and Performance Measure

The accuracy of the classification models is evaluated with **10-fold cross validation**. The data set is divided in nine training sets and one test set. The model is generated using the training sets and the evaluation is done with the test set. The subsets are stratified samples, which means that the class distribution for each subset is equal to the class distribution of the complete data set. The whole process is executed ten times with different samples. At the end an averaged performance measure can give insights on the goodness of the model.
For the subsequently described experiments the mean **accuracy** was used as well as **precision** (how many selected documents are relevant) as performance measures. Precision was selected since it is the most important performance measure when it comes to recommending hotels. Assuming that a person with a tourist role of a Sunlover wants to find the perfect hotel, then it is important that the majority of hotels which are suggested by a system are also relevant for him or her. Otherwise the user will not be satisfied with the result and won't use the recommender system again. Together with the precision also the **recall** of the class (how many of the relevant documents are selected) is evaluated. This measure must be considered as well since it shows how many percent of the hotels could have been interesting for a tourist but were not selected by the system. Precision

	Predicted Class		
True Class		**Sunlover**	**Not Sunlover**
	Sunlover	True Positive (TP)	False Negative (FN)
	Not Sunlover	False Positive (FP)	True Negative (TN)

Table 5.1: Confusion Matrix

and recall are always measured only for those documents which are relevant for the chosen factor.

The performance measures can be calculated using the confusion matrix, as presented in table 5.1. This matrix shows how the documents are distributed over the classes by the classifier. It gives a first impression on how many data samples got correctly and how many got misclassified. Based on the confusion matrix the values for prediction, recall and accuracy can be derived as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.2}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.3}$$

5.2.3 Evaluation of Training Set

To build the model for the classification process the 371 hotels of the training set described in 4.7 were used. Table 5.2 shows the results of all Seven Factors over the applied classifier.

In general, it can be identified that the outcome over all factors is widely spread. The values for accuracy reach from 85,7% to 67,38%. The results for precision and recall even vary over more than 50%. This indicates that the recommendation for some clusters might be easy whereas for others it may not even be possible. Considering each of the factors individually, most of the classifiers produce similar results. Since there is no classifier which in total performs better than the others, the best classifier for each factor was chosen individually to enable the best results. The classifier with the highest accuracy for each factor is marked as bold. For the factors Independent and Riskseeker it was not possible to create a model by using the classifier decision tree. The algorithm could not define specific decision rules to create a tree.

Factor	Classifier	Accuracy	Precision	Recall
Sunlover	k-NN	80,08%	76,04%	58,87%
	Naïve Bayes	80,85%	73,45%	66,94%
	Decision Tree	**85,70%**	**89,01%**	**65,32%**
Educational	**k-NN**	**83,56%**	**69,57%**	**40,51%**
	Naïve Bayes	79,22%	51,28%	50,63%
	Decision Tree	79,79%	56,25%	22,78%
Independent	**k-NN**	**67,38%**	**45,68%**	**32,46%**
	Naïve Bayes	62,28%	40,58%	49,12%
	Decision Tree	-	-	-
Cultural	k-NN	74,11%	56,25%	26,47%
	Naïve Bayes	72,50%	50,00%	54,90%
	Decision Tree	**76,55%**	**67,44%**	**28,43%**
Sportive	k-NN	67,38%	66,67%	73,96%
	Naïve Bayes	**69,80%**	**70,41%**	**71,88%**
	Decision Tree	66,84%	67,34%	69,79%
Riskseeker	**k-NN**	**84,37%**	**45,45%**	**8,77%**
	Naïve Bayes	74,37%	22,06%	26,32%
	Decision Tree	-	-	-
Escapist	k-NN	81,94%	82,76%	79,56%
	Naïve Bayes	81,68%	80,21%	82,87%
	Decision Tree	**82,21%**	**82,49%**	**80,66%**

Table 5.2: Results of classification experiments over the training set.

Further, the table shows that the accuracy for the four factors Sunlover, Educational, Riskseeker and Escapist exceeds 80%. At first sight, this looks like a good outcome of the classification model, but precision and recall have to be examined as well. Considering factor **Riskseeker** with classifier k-NN, the precision is at 45,45% and recall only at 8,77%, which means that not even every second hotel which would be recommended to a tourist is a relevant one. The gap between the rather high accuracy and the other performance measures precision and recall can be explained with the distribution of the training set for the Riskseeker. Only 57 of the given 371 data samples are assigned to this factor ("Class 1"), the other 314 are non-relevant ("Class 0"). The mean accuracy is calculated over both of the classes, which means that if all data samples are assigned to "Class 0", the overall accuracy is already at 84% whereas the precision for "Class 1" lies at 0%.

Because of this inhomogeneously distributed data set the building of the model is very difficult. The same problem can be detected for factor **Educational**, which has the second lowest amount of assigned hotels in the training data set. However, the model

for this factor has a precision of nearly 70%, which means that at least seven of ten recommended hotels could be interesting for an educational tourist. Consequently it is not enough to rely on accuracy alone, since the precision and recall of a certain class might give better information on the performance of a classification model.

The precision and recall for the factors **Sunlover** and especially **Escapist** are much higher and it can be assumed that a recommendation by using hotel descriptions would be possible. However, this has to be confirmed with the final evaluation using the test set.

The classification model for the tourist factor **Sportive** computes a precision of 70,41%. Although the number of hotels which can be allocated to this factor is the highest one compared to the other factors, the building of the model did not work as well as expected. The reason might be the complexity of the factor, since it represents not only the sportive tourist but also the anthropologist. The hotels which could be recommended to this factor are therefore quite diverse. However, it can be identified that recall and accuracy have a similar value as the precision, hence the three performance measures are balanced.

With a precision of 67,44% the factor **Cultural** ranks in the middle of all tourist roles. One interesting fact is the great gap between precision and recall. This means that most of the relevant documents were not correctly classified by the model.

The **Independent** classification model is, together with the Riskseeker, with a precision below 50% not adequate for a recommender system. Although the distribution of the training set is more balanced, it is not easy to identify a certain hotel type which would match the factor in general. This role wants to discover traditions and history and as well its own sense of life. Therefore, it can be very difficult to pick the right hotel for this person and even define an appropriate hotel.

Concluding the results, for the factors Sunlover, Educational, Cultural, Sportive and Escapist a recommendation using hotel descriptions and classification seems feasible. For the tourist roles Independent and Riskseeker, for whom a precision of about 45% could be computed, a recommendation is not meaningful without any other information than hotel descriptions alone.

5.3 Dictionary Approach

The goal of this method was to create an individual dictionary out of words or phrases for each of the seven tourist factors. Each of the dictionaries should contain tokens that describe the main characteristics of hotels which would be important for the corresponding tourist factor. To achieve this, the most common words were assigned to factors by a set of experts completely independent from one another. The experts work in the tourism domain and have already gained knowledge concerning the Seven Factors. Finally these dictionaries should be used to allocate hotel descriptions automatically to one or more tourist factors.

5.3.1 Establishment of Dictionaries

The first step to identify the relevant content for the dictionaries was to rank the different tokens based on their total occurrences in the GIATA data set. Tokens that occur more often were considered to be more significant for the generation of the dictionaries. This pre-selection had to be done since it was not possible to evaluate the huge amount of different tokens by the experts due to time and resource constraints. For the selection, 1000 hotels with their longest hotel description were randomly chosen from the data set and their tokens were ranked. The high frequently used words are not necessarily those which would also have the most occurrences in the training set. However, the dictionaries should not only work with the training set, but also with the test set and later, if the method proofs as successful, with the complete data set. Therefore, to keep it generalisable, the most common tokens from a randomly selected data set were chosen.

The word groups that were chosen to be included in the dictionaries were nouns and adjectives. Verbs were discarded since after a first qualitative evaluation it was obvious that they were not suitable for describing characteristics of hotels. The 260 top ranked nouns and adjectives were evaluated by a team of experts and assigned non-exclusively to the seven dictionaries. The objective was to add a token to a dictionary if more than 50% of the experts individually assigned it to the corresponding factor. Each dictionary should at least contain 10 different tokens.

Table 5.3 shows the tokens of the dictionary for each factor. For the factors Sunlover, Cultural and Sportive more than ten tokens could be identified by the majority of experts. For the remaining tourist roles the selection of tokens was more difficult. It was not possible to generate at least ten tokens by using only those words which were assigned by more than 50% of the experts. Therefore, also lower ranked words were taken into account in order to proceed with the implementation of the dictionary model for these factors. Those additionally added words are marked as italic in the table. At least half of the experts team have assigned those words to the corresponding factor. The establishment of dictionaries was not trivial, even for people who work in the tourism domain and deal with the Seven Factors. One reason could be that the tourist roles have been described with their characteristics, but it is not exactly clear which qualities a hotel needs to offer so that they are interested. Moreover, to allocate these words is a very subjective task, therefore the conformity between the experts was not extremely high. Out of 260 evaluated words, 99 tokens were at least selected from one expert. This means that a majority of the words were not suitable for describing hotel characteristics. According to the length of some of the dictionaries (the shortest one has only five tokens) the result is not expected to improve significantly for all factors.

Factor	Dictionary
Sunlover	Strand, Pool, Wlan, Liege, Swimmingpool, Sonnenschirm, Beach, Sandstrand, Meer, Internetzugang, Sonnenterrasse, Liegestuhl, Poolbar, Meerblick, Badetuch, WiFi, inclusive
Educational	Buffet, Buffetform, Show, *Resort, Club, Unterhaltungsmöglichkeit, Animation, Halbpension, Miniclub, Musik, inclusive*
Independent	Zentrum, *Eigenregie, gemütlich, lokal, individuell*
Cultural	Spa, Suite, Jacuzzi, superior, deluxe, elegant, Whirlpool, Bademantel, Wellnessbereich, Carte, Mietsafe, modern, geschmackvoll
Sportive	Fitness, Sport, Tischtennis, Tennis, Tennisplatz, Volleyball, Aerobic, Fitnesscenter, Golfplatz, Fitnessraum, Aktivität
Riskseeker	Club, *Unterhaltungsmöglichkeit, Show, Poolbar, Stadt, Animation, Unterhaltung, alkoholisch*
Escapist	ruhig, gemütlich, *Wellness, Park, Gartenanlage, Garten, Spa, Wellnessbereich*

Table 5.3: Final dictionaries for all Seven Factors

5.3.2 Evaluation of Training Set

For the evaluation of the dictionaries, a method needed to be defined in which case a hotel should be allocated to a factor. This is not a trivial task. One basic approach would have been the counting of words in a hotel description and based on the number of words which occur in a certain dictionary, the assignment to a tourist factor could be performed. However, since the dictionaries as well as the hotel descriptions differ in size, the calculation of a threshold would have been a difficult process. Therefore, a different procedure was taken into account. As the outcome of the classification method described in the previous section was satisfying, the same classifiers were also used for evaluating the dictionaries. Again one model for each of the seven factors and their dictionaries was established. The difference lies in the preprocessing of the hotel descriptions. Every attribute, which was not part of the dictionary which should be evaluated, was deleted from the data set. As a result the hotels were only defined by attributes which were in the dictionary. Again all in section 5.2 described classifiers were tested. The complete table with the results can be found in the appendix in table C.1. To enable a comparison with the results of the classification approach, figure 5.6 shows the precision computed by both methods for all seven factors. In general, the precision is higher with classification, apart from one factor: The Sunlover. Surprisingly, although for Sunlover an excellent precision value could already be reached in the simple classification, there was still an improvement possible with the dictionary extension. To understand this effect, a closer look at the structure of the

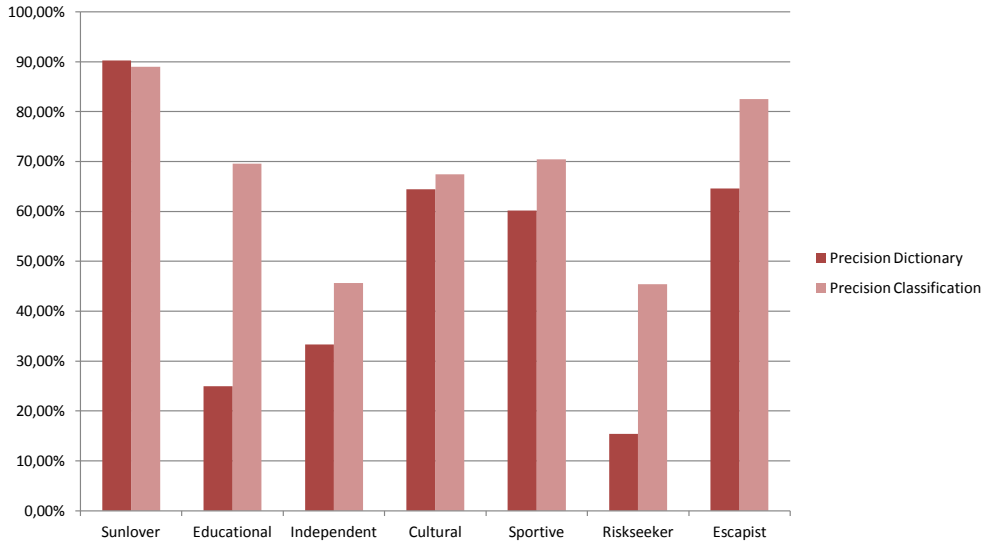
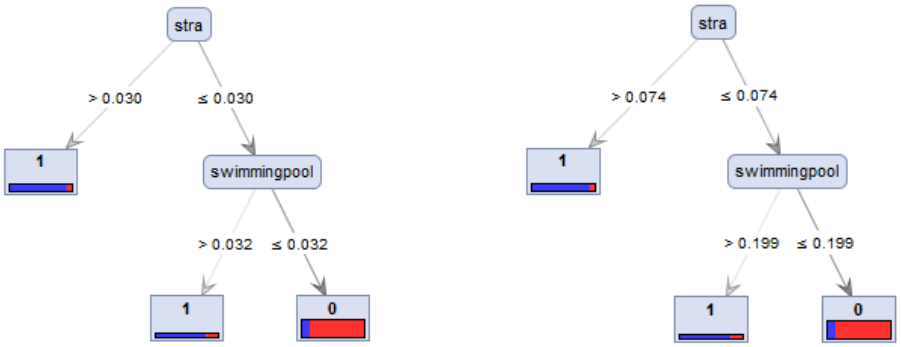


Figure 5.6: Comparison of precision of dictionaries and classification



(a) Classification (b) Dictionary

Figure 5.7: Structure of the decision trees of classification and dictionary approach for factor Sunlover.

decision tree model is helpful. The main rules for identifying the relevant hotels for factor Sunlover are based on the TF-IDF score for the tokens "strand" (refers to Strand, engl. beach) and "swimmingpool". Both words were also chosen as well by the team of experts. Since there are fewer attributes in the data sample generated with the dictionary, the TF-IDF scores vary, as shown in the two trees in figure 5.7. It can be supposed that the scores are more distinct with a fewer number of attributes, which enables a more accurate assignment to a class and further consequently higher values of the performance measures.

The dictionary precision of all other factors stays far behind the precision value of the simple classification. However, due to the better results of factor Sunlover it can

Factor	Top five tokens of χ^2 statistic
Sunlover	*Strand*, Klimaanlage, *Sonnenschirm*, *Meer*, *Swimmingpool*
Educational	Stadt, etwa, Hauptbahnhof, Balkon, Anreise
Independent	Bergbahn, Panoramablick, durchschnittlich, Betrag, möglich
Cultural	Sehenswürdigkeit, Balkon, Haustier, Stadt, Kreditkarte
Sportive	Gang, Anreise, Bahnhof, Bergbahn, Klimaanlage
Riskseeker	erleben, Erlebnis, *Stadt*, Nachmittag, gehen
Escapist	Klimaanlage, Bergbahn, Strand, Gang, Skigebiet

Table 5.4: Keywords for all Seven Factors derived by the χ^2 statistic

be concluded that the approach itself can be successful, but only if the right keywords are chosen. In order to understand which keywords would be most relevant for the differentiation of the factors, Pearson's chi-square (χ^2) test of goodness of fit was used [Pearson, 1900]. With this method it can be derived if a certain attribute of a population is dependent on a class value or equally distributed between all the classes.

The value of the χ^2 statistic can be derived as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E} \tag{5.4}$$

The variable O defines the observed value for an attribute i and E denotes the expected value. To derive the expected value, it is assumed that an attribute is equally distributed between all classes. The individual values are summed up for all n documents which are in a certain class. The higher the χ^2 statistic is, the bigger role the attribute plays for distinguishing one class from the other. [Li et al., 2004]

The χ^2 statistic was calculated for each token using the Rapidminer "Weight by Chi Squared Statistic" operator [RapidMiner, Inc., b]. With this method, the weight for each token is derived with respect to the class label. Hence, it shows which tokens are most important to identify if the text belongs to one class or the other. In table 5.4 the five tokens with the highest weights are presented for each factor. The italic words marked in blue are those which have already been identified by the experts and are present in the dictionary as well. Regarding factor Sunlover, four out of five words appear in both the dictionary and the χ^2 top five. This could be the reason why the dictionary approach worked so well for the Sunlover. Further, it can be identified that some tokens, like Bergbahn or Klimaanlage, are highly weighted for more than one dictionary. A reason could be that the appearance of those tokens is a clear indication for a certain hotel type. In general, the tokens detected with the chi-square statistic do not reflect the factor itself,

compared to the dictionary, but the tokens which are important to define if a hotel is relevant for a factor or not. Therefore, a highly weighted token not necessarily expresses a characteristic of a hotel for a certain factor. It could be even the case that the token mostly appears in descriptions of hotels which should not be recommended to the given factor. The values for the tokens can be found in appendix in table C.1.

An interesting part of future work could be the establishment of dictionaries or weights for tokens with feature selection techniques like the chi-squared statistic.

CHAPTER 6

Final Evaluation

In the previous section three different approaches for the allocation of hotels to the Seven Factors were discussed. A labelled training set with 371 hotels was used to build several models for each factor and further identify which one is the most accurate based on the trainings' precision computed with cross validation.

Table 6.1 gives a short overview on the models and classifiers which enabled the highest scores for accuracy and precision based on the trainings' evaluation. For Sunlover the dictionary approach was even more successful than the simple classification, in contrast to the other factors. Considering the classifier, both Decision Tree and k-NN worked best on three factors whereas Naïve Bayes on only one, the Sportive. Therefore, it could not be identified that one classifier is more suitable for solving this complete problem, it strongly depends on the factor itself.

To enable a final evaluation, these models were applied to the test set provided by Eurotours as well as a random sample of 1000 hotels extracted from the overall GIATA data set.

Factor	**Appraoch**	**Classifier**
Sunlover	Dictionary	Decision Tree
Educational	Classification	Knn
Independent	Classification	Knn
Cultural	Classification	Decision Tree
Sportive	Classification	Naïve Bayes
Riskseeker	Classification	Knn
Escapist	Classification	Decision Tree

Table 6.1: Best approaches on evaluation of training set

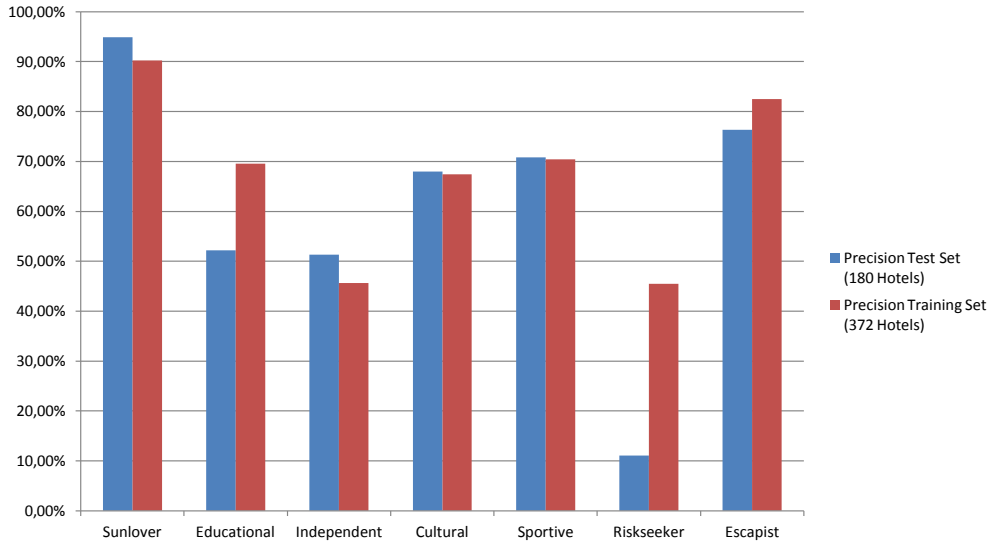
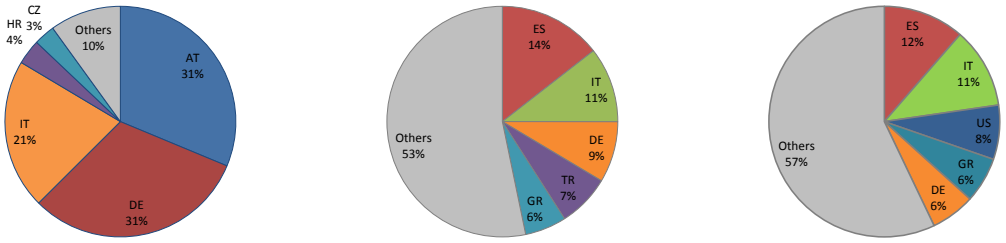


Figure 6.1: Comparison of precision of best approaches of the training data set and the test data set

6.1 Evaluation of Test Set

To enable a final evaluation of the best approaches, the test set from Eurotours described in section 4.7 was taken into account. It covers a scope of 180 hotels which have already been allocated to one or more factors. This step is important to ensure that the previously established models will also work on data that was not used during their generation.

The best models were applied on the test data set and the results were evaluated. The final outcome is introduced in figure 6.1. It shows the comparison of the precision of the training and the test set. For most of the factors the outcome for both data sets strongly correlates. Considering Sunlover and Independent the precision of the test set even improved about 5% which indicates that the models can handle new data very well. The same holds for Cultural and Sportive, where the precision slightly increased, as well as for Escapist, where the precision went down for about 5%, but still the second best result could be achieved. The two outliers are the factors Educational and Riskseeker. For the latter a precision value of only 11,11% could be accomplished. A reason can be that the Riskseeker had the lowest amount of assigned hotels in the test set and the performance measures for training set were also not high. The same holds for the Educational. Considering the distribution of the test set, those two factors are the ones with the smallest amount of related hotels. Due to the lack of training data the models are not as robust to new data samples, which explains the big fall of precision. The exact values for the evaluation can be found in appendix in table D.1.



(a) Eurotours data set (b) Complete data set (c) Random sample

Figure 6.2: Comparison of the distribution of countries in the labelled data set by Eurotours, in the complete GIATA database and in the randomly selected data sample including 1000 hotels.

6.2 Evaluation of Random Hotel Sample

In order to consolidate the results of the Eurotours test set with the overall data set provided by GIATA, the models were applied to a random sample of 1000 hotels. Since the GIATA data set is not labelled, the precision for the individual factors cannot be derived. However, this experiment is important to become an insight if the established models can be applied to different hotel samples as well.

Figure 6.2 gives an overview about the country distribution of the test set, the overall GIATA data set and the distribution of the random sample. It can be seen that the distribution of the randomly selected hotels is similar to the distribution of the complete data set. The countries Spain (ES), Italy (IT), Germany (GE) and Greece (GR) are most common in both data sets. Therefore the random sample of 1000 hotels can be considered as a representative sample of the complete data set.

The distribution of the hotels over the factors is presented in figure 6.3. The figure shows the percentage of hotels which were allocated by the implemented models to each of the Seven Factors. It does not compare the output of the random sample with the ground truth labels of the test set. For all factors besides the Riskseeker a deviation between the test set and the random sample can be detected. Based on the distribution, the random sample covers much more hotels for the Sunlover, Educational and Cultural tourist then the data provided by Eurotours. One explanation might be the country distribution. Spain, Italy and Greece are countries which have sea access and therefore offer a lot of beach hotels which could be more interesting for the Sunlover. Moreover, they are also quite popular for their historical sights and cultural heritage. Therefore it can be plausible that the hotels in the random sample can be more interesting for these three factors. Concentrating on the most common countries in the test set, which are Austria and Germany, it makes sense that hotels for the Escapist as well as the Sportive tourist are more represented. Especially Austria is known for its mountains,

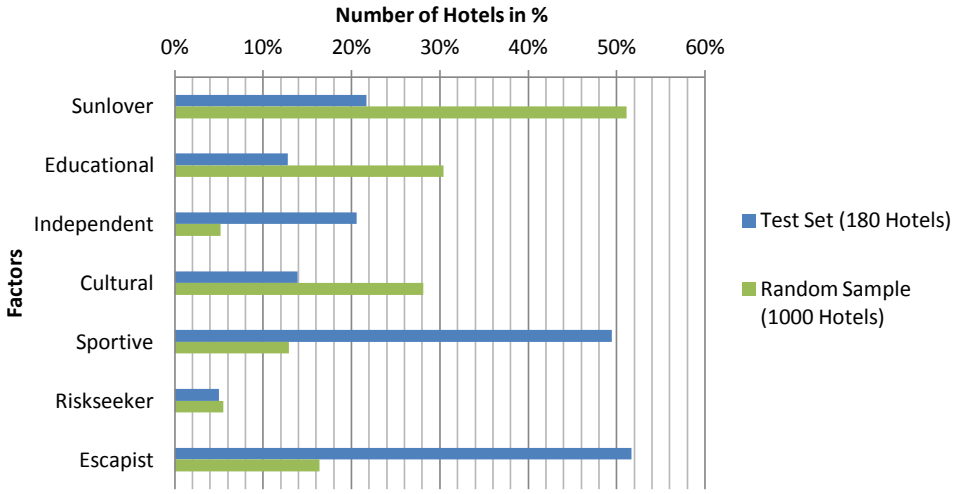


Figure 6.3: Comparison of the factor distribution of Eurotours' test set and the random selected hotel sample. It is shown how many hotels are relevant for which factor. The distribution is based on the output of the classification models. On average each hotel was assigned to 1,75 factors in the test set and 1,5 factors in the random sample.

which offer leisure areas as well as landscapes for sports like mountain biking, hiking and climbing. Also Germany is not known for its big amount of beach resorts, but for interesting cities and landscapes. This might also be favoured by Independent tourists.

Based on the distribution which is plausible for the data set it can be assumed that the models can be applied to the overall GIATA data set as well. However, an interesting part of further research would be the evaluation of the hotels by a user group, both for the Eurotours test set and the GIATA data set. This would be a first step to enhance the accuracy of the models.

6.3 Critical Reflection

This section critically analyses the used methods and evaluation techniques focusing on further enhancement of the proposed system.

Focusing on the final evaluation it can be identified that the established classification models work well for certain factors, but do not work at all for others. This gap can be explained with the distribution of the data set. The proposed training and test set are not suitable for all of the factors since some of them are very underrepresented and therefore, the model building is more challenging. This would hold for the Riskseeker, which is the factor with the lowest amount of suitable hotels in the data set and also the one which performs worst in the final evaluation. However, if we consider the Independent

traveller, the amount of related hotels is as high as the one of the Sunlover. Still the precision of the Sunlover is almost 45% better. So the distribution of the data set cannot be considered as the only dependency of the diverse results of the factors.

Further, it is not sure that a larger data set enables a more precise modelling for the underrepresented factors. In contrast, it can be expected that the distribution of factors in a larger data set is similar to the one described for this training and test set.

Focusing on the Riskseeker there are probable reasons why an allocation to hotels is difficult. Since this tourist factor likes adventures and experiencing crazy activities, the "standard" hotels one can find in a catalogue might not be the ones he is interested in. Hence, the majority of hotels in the GIATA data set can be considered as not suitable for this factor.

The model building for the Independent was also very difficult and did not yield the expected result. The precision on the test data set was only at about 45% although there were more than 30% of suitable hotels in both training and test set. In this case the personality of the factor itself is very diverse and special. He intends to go on a journey to inspiration and discover the sense of life. Therefore, it is a complex task to identify hotel features which would match this type of traveller. The size of the Independent's dictionary represents a clear indication for this. Only one word was identified by more than 50% of the experts which would describe a related hotel for this factor. It shows that even for experts in this field it was difficult to state the characteristics of a suitable hotel. This amplifies the assumption that also the allocation of hotels to this factor is challenging since there is no general understanding on how a hotel should be to inspire somebody. There might be a wide range of hotels which fit the criteria but the question itself is very individual and related to the users personality. For that reason a recommendation for this factor is complex and no satisfying solution could be found during the work on this master thesis.

However, observing these two factors, the Riskseeker and the Independent, the question can be raised if it actually makes sense to establish a model for those tourist roles in the context of an online recommender system. Both factors want to go on an individual journey, one searches for adventures, the other for inspiration. Probably none of them would use an online booking website for getting hotel recommendation for travelling.

Concentrating on the factor Educational a drastic decrease of more than 15% precision can be observed between training and test set although the distribution for the factor in the two data sets is similar. However, only about 20% of the available hotels in both sets were assigned to the Educational which might be the reason why the model building was not as robust as others. The data distribution in this case is somehow surprising since the Educational likes travelling in groups and broadening his knowledge. There are many travel operator like TUI or GRUBER-reisen which offer guided round trips for groups. However, an advertisement for a guided tour in a catalogue mostly focusses on the country and attractions, since these are the most important facts for an Educational

tourist. The hotels which are visited during the travel play a minor role. Since the hotel category and facilities are mostly not that important for an Educational traveller it can be discussed if it makes sense to recommend single hotels to this factor. A better solution would be the recommendation of a guided tour in a specific region. Unfortunately this cannot be modelled with a system using only hotel descriptions as an input.

The Cultural factor represents also the high class tourist and likes hotels with exceptional facilities. According to the hotel description it is often not easy do determine if the hotel is supreme since all hotel texts are very positively written in order to raise the customers attention. Furthermore, the performance of the described classifier, with a precision of around 68%, might be not sufficient for a hotel recommender system. For this factor it would make sense to include the hotel-star-rating into the model. Although there is no defined system for the rating, in most of the cases the high class hotels can be detected and it can be expected that the recommendation is more precise.

Examining the Sportive factor, the precision is similar to the Cultural at about 70%. The Sportive has the highest amount of assigned hotels compared to the other factors. The reason is that most of the hotels offer some kind of sports activities like a fitness room or a table tennis table. However, the Sportive tourist is not only a person who likes to be active, he also wants to meet local people and their culture. Therefore, it can be very complex to identify a hotel which satisfies both characteristics of this factor. In the dataset provided by Eurotours mostly hotels which offer sport activities were assigned to the factor whereas the characteristics referring to the anthropologist were hardly considered. Otherwise there would have been only a fraction of the hotels which could still be assigned to this factor. But again it can be discussed if the real anthropologist is willing to find a hotel via a booking website.

The best results for all factors were achieved for the Escapist and the Sunlover. Although the test results for the Escapist were below the training results, it can still be used for hotel recommendation. The typical hotels for an Escapist are wellness and spa resorts which offer stress relaxation and relief. This sort of hotels could also be found during the qualitative analysis. Further, the Escapist is one of the simpler built factors since he combines the very related tourist roles Escapist I and II. This eases the finding of suitable hotels manually as well as automatically compared to the other factors.

The same holds for the factor Sunlover. With his beach and sun loving features he is the typical tourist who wants to enjoy an all-inclusive vacation. As identified in the qualitative analysis the majority of hotels offer facilities for the Sunlover. Although the distribution of the Seven Factors themselves is not available, the conclusion can be drawn, considering the offer, that the Sunlover is one of the most widely spread tourist factors. Moreover, the Sunlover proves that the dictionary approach is a method which can be applied for online recommender systems. It shows that with the correct definition of the most prominent words the classification model for a factor can be enhanced. Unfortunately the definition of keywords with experts was in this case only suitable for one factor. A different approach could be the automatic detection of keywords with feature selection and extraction.

CHAPTER 7

Conclusion and Future Work

7.1 Summary

The main aim of this thesis was to identify concepts or models to enable an automated allocation of hotels to the Seven Factors defined by [Neidhardt et al., 2014b]. Therefore, state of the art literature in text mining and document classification and clustering was discussed and analysed. Further, three machine learning algorithms (clustering, classification and a dictionary based approach) for text grouping were designed, implemented and evaluated. The outcome of the thesis was expected to be a contribution for future recommender systems in the tourism domain.

After discussing the related work the data source itself as well as the data acquisition was described in chapter 3. The hotel descriptions, which served as textual input for the established algorithms, were collected from the Global Distribution System of GIATA. The first qualitative analysis of the data showed that on average there are 20 different hotel descriptions for one hotel, but some of them were equal or similar to each other. The reason was that many tour operators offer two or more descriptions for the same hotel, e.g. one for the winter and one for summer season. Even different operators offer similar or equal descriptions which means that tour operators work together or might even copy from each other. To avoid that the same description is used several times, only the longest description was selected to represent a hotel. This was based on the observation that text length correlates with the information gain of the descriptions. Furthermore, it could be identified that most tour operators use a predefined document structure and headlines for all of their descriptions, so called "templates". The impact of these templates could be noticed in the automated cluster analysis.

In order to modify the unstructured textual data so that it could be used as an input for a machine learning algorithm, different preprocessing steps were necessary. Since the data was provided in XML-format, the content had to be extracted. Further, the text was

separated into tokens and unessential words were removed using a standardised stop-word list. The stemmer of [Caumanns, 1999] was used to reduce the tokens to their word stem. Pruning was applied and the unstructured data was transformed to a numerical word-vector using the TF-IDF score. The labelled training and test data sets were established with support of Eurotours and consisted of 371 training and 180 test samples. These data sets served as the foundation for the established models described in this work.

The mapping of hotel descriptions with the Seven Factors was tested with three machine learning approaches: Clustering, classification and the dictionary based approach.

Unsupervised Clustering was implemented with the k-means algorithm and the cosine similarity measure. The main limitation of these method was that it could not be modelled that a hotel could be related to more than one factor, since only an exclusive assignment to a cluster was possible. However, the goal was to identify some cluster which would represent one or more tourist factors. The evaluation of the cluster analysis pointed out that the algorithm did not detect clusters representing a factor but representing a tour operator. The main reason was the previously described template structure of the descriptions. Hence, unsupervised clustering could not be used to identify hotels for tourist roles but for getting an input on which operators offer similar descriptions for their hotels.

For supervised learning three classifiers were implemented and compared: Naive Bayes, k-NN and Decision Tree. For each factor a model was established which should classify the factor related hotels and the residual hotels. The evaluation of the seven models were done with 10-fold cross validation of the training set. The outcome of this approach indicated that the allocation of hotels to factors strongly depends on the factor itself. It wasn't even possible to identify a most suitable classifier, since each of them performed best for at least one of the factors. The models for Sunlover and Escapist were the most promising ones regarding their training validation, whereas the related hotels of Independent and Riskseeker had turned out as very difficult to classify.

To enhance the simple classification models, a dictionary based approach was implemented. Experts in the tourism domain identified words which should describe the characteristics of hotels related to the Seven Factors. Based on this input, seven dictionaries were created and only their content was used with the three mentioned classifier. However, the input of the dictionary could enhance only the Sunlover's model. All others had lower precision values using the dictionary than with simple classification. The main problem was that even for the experts it was complex to define certain words for some of the factors' hotels.

The final evaluation of the best approaches, which were identified by means of their training set validation, was performed with an independent test set of 180 hotels. For the Sunlover the dictionary based method was chosen, for all others the simple classification was evaluated. Moreover, the models were applied to an unlabelled random sample of

1000 hotels extracted from the GIATA database and the outcoming hotel distribution over the factors was discussed and compared to the outcome of the test set.

With this evaluation the stated research questions which were introduced in section 1.2 could be individually answered for all seven factors. It was possible to design models and enable a classification of hotel descriptions, which could then be allocated to the seven factors. However, the outcome of the classification strongly depended on the factor itself.

The final result indicated that most of the previously defined models were capable of dealing with new hotel data. The Sunlover achieved the best precision value of over 90%, followed by the Escapist, Sportive and Cultural tourists with a precision around 70%. This states that a hotel recommendation based on hotel descriptions can be a promising approach for recommender systems, however, it depends on the targeted user group. For the three factors Educational, Independent and Riskseeker the precision value was below 55% which means that hardly every second proposed hotel is interesting for the tourist. Based on the results of this thesis it is not advisable to use hotel recommendations alone for those factors.

7.2 Limitations and Future Work

In order to enhance the proposed approach of the allocation of hotels to tourist roles, some features and extensions are proposed in this section.

One of the biggest limitations faced during the work on this thesis was the data set. Although it was labelled by experts in the field there was still no user based evaluation done. An improvement of the data quality could be the labelling of a critical amount of users e.g. via an online platform. There the users' factor could be identified (with the method proposed by [Neidhardt et al., 2014b]) and then he or she could select between different hotels. Better quality of the data set would also enhance the models and prevent users from receiving unsuitable recommendations. Further, the hotels included in the training and test set are mostly located in Central Europe. This deviates from the country distribution of the overall GIATA data base. Therefore, when applying the models to the complete GIATA data samples the quality of the models might be below the expected values. Still, the structure and design of the models could be used and the models could be easily adapted by executing another training phase, when more labelled data was available.

Another proposal for future work would be the extensions in the preprocessing of the textual data. Together with single tokens also multiwords like bi- or trigrams could be considered. With word combinations there is a chance that better results can be achieved.

In the described approaches a hotel is either allocated to a certain factor, or it is not. So there are only two possible classes for a hotel regarding a factor, "suitable" or "not suitable". This limits the recommender system because the hotels cannot be ranked. The goal would be to rank the hotels and show only the most favourable ones to the user. In the

systems described in this thesis, it is only possible to randomly select certain hotels from a pool of suitable ones, since the topic was treated as a classification problem. In order to improve this, regression models could be implemented. In contrast to classification, the regression model can handle labels with continuous values. Those labels could represent the percentage of congruence of a hotel to a factor. This step would enhance the model since the hotels could be ranked and the most suitable ones could be introduced to the user first.

Considering a real world scenario the behaviour of a tourist can hardly be mapped to only one factor. A user would most probably be represented by a mixture of certain factors. Since the proposed approaches are designed to assign hotels to just one of the Seven Factors, a combination of the presented models would be needed to cover this use case. For further development, it would be interesting to introduce a multi-class system which could handle factor combinations as well. With this extension the recommendations could be customised to the individual characteristics of a user.

APPENDIX A

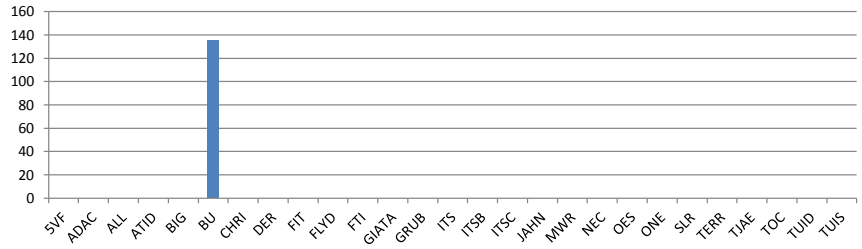
Qualitative Evaluation of Tour Operators

Provider details		GIATA Id of 10 selected hotels									
Provider name	**Provider code**	**3**	**111**	**242**	**333**	**770**	**999**	**1467**	**2345**	**3456**	**4567**
12 Fly	FLYD		18						19	18	17
5 vor Flug	5VF		234		536		244		186		248
Airtours	ATID	375									
Alltours	ALL			392			197				
Alltours XALL	XALL			392			197				
Big Xtra Touristik	BIG			447	457						
Bucher Reisen	BU		208		179	175	177				117
Bucher Reisen Schweiz	BUC		208		179	175	177				117
Byebye	BYE			159			197				
Christophorus Reiseveranstaltung	CHRI							257			
Dertour	DER				512						
Discount Travel	DIS		18		562	18			19	18	17
ECCO Reisen GmbH	ECC		266							161	
FIT Gesellschaft für gesundes Reisen	FIT							648			
FTI	FTI	304	371	473	551				160		172
FTI Schweiz	FTIS	304	371	473	551				160		172
GIATA	GIATA	179	179	710	232	120	112	118	76	69	56
Gruber Reisen	GRUB							451			
Hotelplan	HP		219		293						
ITS	ITS		339	390							
ITS Billa Reisen Austria	ITSB			390							
ITS Coop Travel	ITSC			390	616						
ITT Ferien Pur	ITT										326
Jahn Reisen	JAHN		339		616						
Kuoni	KUON				288						
Meier's Weltreisen	MWR				576						
Neckermann	NEC		79		124	186	218		237	155	
Neckermann Austria	OES						218		237		
Öger Tours GmbH	OGE	144									142
Öger Tours GmbH Austria	OGO	144									137
One Touristic	ONE										185
Schauinsland Reisen	SLR		350		393	267					
Sierramar	REBA9		258								
Sitalia Reisen	SIT							247			
Thomas Cook	TOC	570			315			231			
Touropa / touropa touristik GmbH	TOUR		660								
Trazoom gib Gas, mach Urlaub	ZOOM										326
TUI	TUID						325	424			
TUI Suisse	TUIS							430			

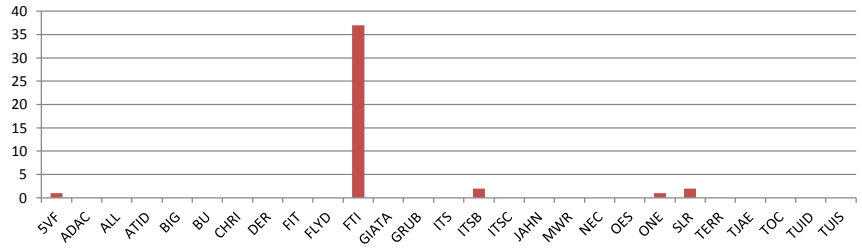
Table A.1: The table shows 10 hotels which were selected for a qualitative analysis as well as the corresponding tour operators (provider). The colors are based on the three categories, which visualise the quality of information content of the descriptions: High (green), Medium (yellow), Low (red). Further, the text length of each hotel description per tour operator is given.

APPENDIX B

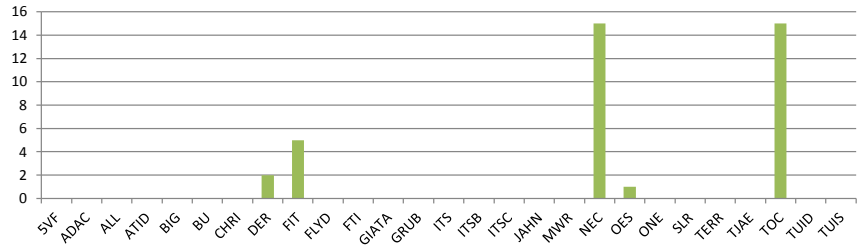
Cluster Distribution of Tour Operators



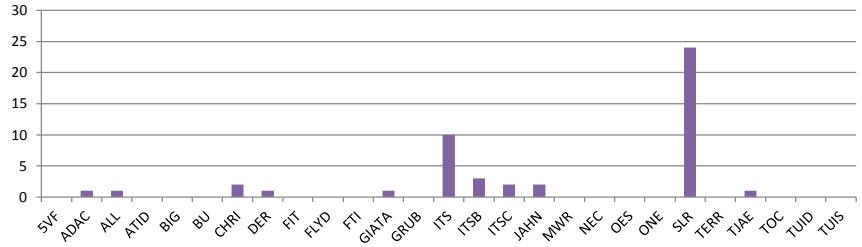
(a) Cluster 0



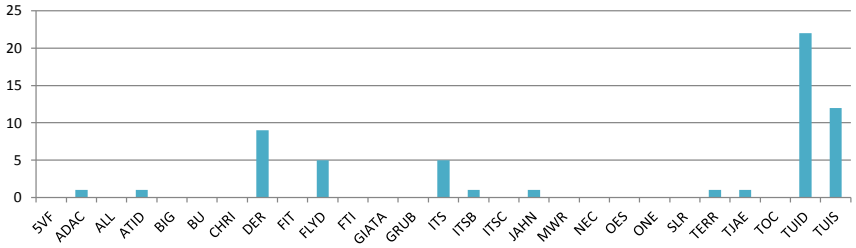
(b) Cluster 1



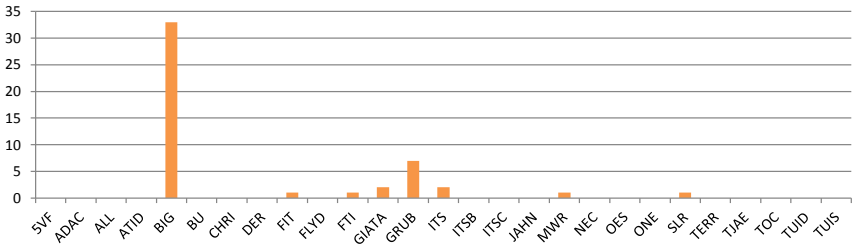
(c) Cluster 2



(d) Cluster 3



(e) Cluster 4



(f) Cluster 5

Figure B.1: Distribution of tour operators to clusters

APPENDIX C

Dictionary Approach: Results and χ^2 Statistic

Factor	Classifier	Accuracy	Precision	Recall
Sunlover	k-NN	84,08%	87,36%	61,29%
	Naïve Bayes	86,24%	89,25%	66,94%
	Decision Tree	**86,51%**	**90,22%**	**66,94%**
Educational	**k-NN**	**76,55%**	**25,00%**	**5,06%**
	Naïve Bayes	32,08%	22,36%	88,61%
	Decision Tree	78,71%	0,00%	0,00%
Independent	k-NN	68,73%	0,00%	0,00%
	Naïve Bayes	**67,12%**	**33,33%**	**7,02%**
	Decision Tree	69,27%	0,00%	0,00%
Cultural	**k-NN**	**76,04%**	**64,44%**	**28,43%**
	Naïve Bayes	70,65%	44,44%	27,45%
	Decision Tree	71,97%	0,00%	0,00%
Sportive	**k-NN**	**58,49%**	**59,79%**	**60,42%**
	Naïve Bayes	53,11%	52,76%	89,58%
	Decision Tree	60,67%	60,18%	70,83%
Riskseeker	**k-NN**	**83,83%**	**0,00%**	**0,00%**
	Naïve Bayes	19,13%	15,38%	94,74%
	Decision Tree	84,64%	0,00%	0,00%
Escapist	**k-NN**	**66,61%**	**64,62%**	**69,61%**
	Naïve Bayes	66,61%	63,13%	75,69%
	Decision Tree	68,50%	63,68%	82,32%

Table C.1: Results of dictionary-based approach with three classifiers

Factor	Token	χ^2 Statistic
Sunlover	stra	1
	klimaanlag	0,496244803
	sonnenschirm	0,395999707
	meer	0,354012563
	swimmingpool	0,31266444
Educational	stad	1
	etwa	0,92941106
	hauptbahnhof	0,760109942
	balko	0,751552397
	anrei	0,743313545
Independent	bergbah	1
	panoramablick	0,980563423
	durchschnittlich	0,969923322
	betrag	0,954863607
	moglich	0,924253919
Cultural	sehenswurdigkei	1
	balko	0,896724465
	haustier	0,786913206
	stad	0,732354621
	kreditkar	0,731225057
Sportive	gang	1
	anrei	0,948406119
	bahnhof	0,8814459
	bergbah	0,844627906
	klimaanlag	0,738741925
Riskseeker	erleb	1
	erlebni	0,68281526
	stad	0,616744154
	nachmittag	0,570166991
	geh	0,537518148
Escapist	klimaanlag	1
	bergbah	0,716602564
	stra	0,565460203
	gang	0,538525068
	skigebie	0,462297667

Table C.2: Results of χ^2 Statistic

APPENDIX D

Test Set Evaluation Details

Factor	Appraoch	Classifier	Accuracy	Precision	Recall
Sunlover	Dictionary	Decision Tree	86,67%	94,87%	62,71%
Educational	Classification	Knn	81,67%	52,17%	35,29%
Independent	Classification	Knn	61,67%	51,35%	27,14%
Cultural	Classification	Decision Tree	66,11%	68,00%	24,29%
Sportive	Classification	Naive Bayes	60,56%	70,79%	58,33%
Riskseeker	Classification	Knn	82,22%	11,11%	4,00%
Escapist	Classification	Decision Tree	73,33%	76,34%	73,20%

Table D.1: Results of evaluation of Eurotours test set (180 hotels)

List of Figures

List of Tables

Acronyms

CARS	Context Aware Recommender System
SVM	Support Vector Machine
k-NN	k-Nearest Neighbour
SVM	Support Vector Machine
POS-Tagging	Part-Of-Speech-Tagging
GDS	Global Distribution System
TF-IDF	Term Frequency - Inverse Document Frequency

Bibliography

Charu C. Aggarwal and ChengXiang Zhai. A Survey of Text Clustering Algorithms. In *Mining Text Data*, pages 77–128. Springer-Verlag New York, 2012.

David Arthur and Sergei Vassilvitskii. K-Means++: the Advantages of Careful Seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 8: 1027–1035, 2007.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009.

Robin Burke and Maryam Ramezani. Matching Recommendation Technologies and Domains. In *Recommender Systems Handbook*, pages 367–386. Springer US, 2011.

Jörg Caumanns. A Fast and Simple Stemming Algorithm for German Words. *Technical Reports of Department of Mathematics and Informatics, Free Univerity Berlin*, B 99/16, 1999.

Erik Cohen. Toward a sociology of international tourism. *Social Research*, 39(1):164–182, 1972.

Kenneth Cosh. Text mining Wikipedia to discover alternative destinations. *The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 43–48, 2013.

David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.

Eurostat. *Statistics on ICT use in tourism*. `http://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_ICT_use_in_tourism#Majority_of_tourist_accommodation_is_booked_online`. Accessed April 09, 2017.

Eurotours Ges.m.b.H. *Eurotours International*. `http://www.eurotours.at/de/`. Accessed October 23, 2016.

GIATA GmbH. *GIATA Company Profile*. `http://www.giata.com/company/company-profile`. Accessed May 8, 2015.

Heather Gibson and Andrew Yiannakis. Tourist roles: Needs and the lifecourse. *Annals of Tourism Research*, 29(2):358–383, 2002.

Lewis R. Goldberg. The Structure of Phenotypic Personality Traits. *American Psychologist*, 48(1):26–34, 1999.

G.K. Gupta. *Introduction to Data Mining with Case Studies*. Prentice-Hall of India Pvt.Ltd, 2006.

Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design Science in the Information Systems Research. *MIS Quarterly*, 28(1):75–105, 2004.

Hajo Hippner and René Rentzmann. Text mining. *Informatik-Spektrum*, 29(4):287–290, 2006.

Hotelstars. *Hotelstars Union Critera*. `http://www.hotelstars.eu/index.php?id=criteria`. Accessed May 25, 2015.

Victoria Johansson. Lexical diversity and lexical density in speech and writing : a developmental perspective. *Working Papers of Deptartment of Linguistics and Phonetics, Lund University*, 53:61–79, 2008.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.

Rie Koizumi. Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, 1(1): 60–69, 2012.

Fatima Zahra Lahlou, Asmaa Mountassir, Houda Benbrahim, and Ismail Kassou. A Text Classification Based Method for Context Extraction from Online Reviews. *8th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 2013.

Jure Leskovec, Anand Rajaraman, and Jeff Ullman. Clustering. In *Mining of Massive Datasets*, pages 240–280. Cambridge University Press, 2014.

Tao Li, Chengliang Zhang, and Mitsunori Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437, 2004.

Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. Eliciting the Users Unknown Preferences. *Proceedings of the 8th ACM Conference on Recommender systems*, pages 309–312, 2014a.

Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. A picture-based approach to recommender systems. *Information Technology & Tourism*, 15(1): 49–69, 2014b.

NLTK Project. *Natural Language Toolkit*. http://www.nltk.org. Accessed December 06, 2015.

Oracle Corporation. *Oracle Technology Network - Java*, a. http://www.oracle.com/technetwork/java/index.html. Accessed June 16, 2015.

Oracle Corporation. *MySQL*, b. http://www.mysql.com/. Accessed June 16, 2015.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

Pixtri OG. *PixMeAway*, a. http://www.pixmeaway.com. Accessed February 06, 2017.

Pixtri OG. *PixMeAway - Your Travel Profile*, b. https://pixmeaway.com/profile. Accessed February 18, 2017.

RapidMiner, Inc. *Rapidminer*, a. https://rapidminer.com/. Accessed October 16, 2016.

RapidMiner, Inc. *Weight by Chi Squared Statistic*, b. https://docs.rapidminer.com/studio/operators/modeling/feature_weights/weight_by_chi_squared_statistic.html. Accessed December 12, 2016.

P. Robert and Y. Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 25(3):257–265, 1976.

Sandro Saitta, Benny Raphael, and Ian F. C. Smith. A Bounded Index for Cluster Validity. *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)*, pages 174–187, 2007.

Yashvardhan Sharma, Jigar Bhatt, and Rachit Magon. A multi criteria review-based hotel recommendation system. *15th IEEE International Conference on Computer and Information Technology*, pages 687–691, 2015.

Michael Steinbach, George Karypis, and Vipin Kumar. A Comparison of Document Clustering Techniques. *KDD Workshop on Text Mining*, 2000.

Sholom M. Weiss, Nitin Indurkhya, and Tong Zhang. *Fundamentals of Predictive Text Mining*. Springer-Verlag London, 2010.

Hannes Werthner and Stefan Klein. *Information Technology and Tourism - A Challenging Relationship*. Springer-Verlag Wien New York, 1999.

Andrew Yiannakis and Heather Gibson. Roles tourists play. *Annals of Tourism Research*, 19(2):287–303, 1992.