Opinion Mining Approach to Improve Hotel Booking Process

By A.R.F Jawhara

149201C

Faculty of Information Technology
University of Moratuwa
May 2017

A.R.F Jawhara
149201C

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Master of Science in Information Technology

May 2017

Declaration

I declare that this thesis is my own work and has not submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from published or unpublished work of others has been acknowledged in the text and a list of list of references is given.

Name of Student Signature of student

A.F Jawhara Rahuman

Date: …… ……..

Supervised by

Name of Supervisor

S.C. Premarathne

Senior Lecture

Faculty of Information Technology

University of Moratuwa Signature of Supervisor

Date: …….. …..……

Acknowledgements

I would like to express my gratitude to my supervisor Mr. S.C Premarathne, Senior Lecturer in University of Moratuwa, whose expertise, understanding added considerably to my research experience. I appreciate his vast knowledge and his skills in many areas and his guidance to writing report.

I would like to thank the other Lecture of University of Moratuwa, Prof. Asoka Karunananda for the knowledge and assistance provided by all level of the research project.

Not only that my thank should go to all the lectures of M.Sc. in Information technology degree program of Faculty of IT, who sharpens our knowledge and ideas throughout this M.Sc. Program

Moreover I would also like to thank all the batch mates of the M.Sc. in IT degree program who gave their valuable feedback to improve the result of research and my family for the support they provided me. I must acknowledge my husband and my best friend for the encouragement.

Abstract

Customer Opinions play a very crucial role in daily life. When we have to take a decision, others opinion also considered. The e-commerce field has developed to the point that more and more hotel companies provide online booking services to travellers as an integral part of their business model. Increasing numbers of hotel companies now provide such services as an integral part of their business model and their guests' experiences with their hotel. Some third-party services allow customers to add comments on each hotel at the affiliated website. The proposed tool features at hotel websites are based on fixed properties, allowing companies to take advantage of the huge number of available customer reviews to provide relevant information to consumers considering new services.

In this thesis we are going to see how Sentiment analysis tool is working used for mining reviews from online reviews those are posted by customers. How Apriori frequent item set mining algorithm can be used for find associate hotel features. Our main theme is to create a system for analyzing opinions which implies judgment of different consumer products.

This research should improve online hotel booking by building a customized tool that utilizes available customer reviews at the Trip Advisor website. In this thesis we are going to see how

Contents

Declaration iii

Acknowledgements iv

Abstract v

List of Figures x

List of Tables xi

Chapter 1 1

Introduction 1

1.1 Prolegomena 1

1.2 Objective 2

1.3 Motivation 2

1.4 Proposed solution 3

1.5 Resource Required 4

1.6 Summary of chapters 4

1.7 Summary 5

Chapter 2 6

Development Challengers in Data Mining 6

2.1 Introduction 6

2.2 Early developments & Modern trends in Data mining 6

2.4 Future challenges 8

2.5 Summary 8

Chapter 3 9

Technology Adapted 9

3.1 Introduction 9

3.2 Tools & Technologies of Hotel Ranking system 9

3.3 Data Mining 9

3.3.1 Opinion Mining 10

3.3.2 Semantic Analysis 11

3.3.3 Sentiment Analysis or Sentiment Classification 11

3.4 Web Scraping 12

3.5. Xampp server 12
3.6 HTML 13
3.7 Hardware Requirements 14
3.8 Summary 14
Chapter 4 15
4.1 Introduction 15
4.1.1 Hypothesis 15
4.1.2 Users 15
4.1.3 Input 15
4.1.4 Output 16
4.1.6 Features 16
4.2 Methodology 16
4.2.1 Data Collection 17
4.2.2 Semantic Analysis 17
4.2.3 Aspect Oriented Customer Review Mining 18
4.2.4 Method of storing data to Database 20
4.2.5 Opinion Mining/Sentiment Analysis Tool 20
4.2.6 Natural Language Processing 21
4.2.7 Association Mining 21
4.3 System Architecture 22
4.4 Front End 22
4.5 Back End 23
4.9 Summary 23
Chapter 5 24
Design of IHBP 24
5.1 Introduction 24
5.2 General Architecture 24
5.2.1 Data Collection module 26
5.2.2 Semantic Analysis module 26
5.2.3 Review Data base module-Data set 1 27
5.2.4 Sentiment Analysis module 27

5.2.5 Data set-2 module 28
5.2.6 My Code module 28
5.2.7 Front End module 28
5.3 Summary 29
Chapter 6 30
Implementation of IHBP 30
6.1 Introduction 30
6.2 Overall Solution based 30
6.3 Data Collection 30
6.4 Feature Identification using Semantic Analysis 32
6.5 Implementation of Review Database 34
6.6 Sentiment Analysis Module 35
6.7 Implementation of MyCode Module 36
6.8 Implementation of the Frontend Module 38
6.9 Summary 39
Chapter 7 40
Evaluation 40
7.1 Introduction 40
7.2 Evaluation for Front End 40
7.3 Evaluation for Opinion Sentiment classifier - NLTK 2.0.4 41
7.4 Evaluate for Hotel Ranking 42
7.6 Summary 59
Chapter 8 60
Conclusion 60
8.1 Introduction 60
8.2 Achievement of Objective 60
8.3 Further work 61
8.4 Summary 61
References 62
Appendix A 63
Structure of Hotels Table in Database 64
Structure of Location Table in Database 64

Structure of Reviews Table in Database 65
After sentiment Analysis filtered only positive hotel features 65
Appendix B 66
Cording of Data base creation 67
Test script of efficiency of Sentiment Analysis 68
Apriori Test script 71
Hotel Ranking System/ MyCode PHP 73
Appendix C 76
Sample of User's survey feedback

List of Figures

Figure 1: Aspect Identification 19
Figure 2: Top Level Architecture 25
Figure 3: Architecture of My SQL Database 27
Figure 4: Architecture of Front End 28
Figure 5: Sample of collected Reviews- Unstructured Data 32
Figure 6:QDA Miner lite interface 33
Figure 7: Analyzed Data-Structured Data 34
Figure 8: -Structure of Database 34
Figure 9:Front End Interface 38
Figure 10:Front End- list of location 40
Figure 11:Front End-Hotel Feature List 41
Figure 12:-Result of test script ID-122 42
Figure 13:Hotel Rank based on selected feature 43
Figure 14:Kandy Earl's Regency hotel's TripAdvisor user review screen 44
Figure 15:Kandy Earl's Regency hotel's HRS in Ranking 45
Figure 16: Colombo Hilton Hotels TripAdvisor user review screen 46
Figure 17:Colombo Hilton hotels HRS in Ranking 47
Figure 18:Survey Sample 48
Figure 19:HRS's Overall performance chart 49
Figure 20: Apriori Test 58
Figure 21: Association Test 58
Figure 22: Test Result of Association 59
Figure 23: Structure of Hotels Table in Database 64
Figure 24: Structure of Location Table in Database 64
Figure 25: Structure of Reviews Table in Database 65
Figure 26: Structure of Reviews Table in Database 65
Figure 27: Survey of user 1 78
Figure 28: Survey of user 2 79
Figure 29: Survey of user 3 80
Figure 30: Survey of user 4 81
Figure 31: Survey of user 5 82
Figure 32: Survey of user 6 83
Figure 33: Survey of user 7 84
Figure 34: Survey of user 8 85

List of Tables

Table 1: User Feedback 49
Table 2: Transaction Table 51
Table 3: Table C1 51
Table 4: Table L1 52
Table 5: Table C2 53
Table 6: Table L2 54
Table 7: Table C3 55
Table 8: Table L3 56
Table 9: Table C4 56

Chapter 1

Introduction

In this chapter, there is a brief introduction about what Opinion Mining is and how it can be performed. Opinion Mining or Sentiment classification involves building a system to make use of reviews posted by the users and opinions that are expressed in Travel web sites as comments and reviews about the hotels.

1.1 Prolegomena

There has been a significant increase in the number of consumers using the Internet to find hotel rooms and companies are always seeking better services and products to satisfy their clients or customers. People are preferred to book hotels through the Internet, because method it is much easier and faster than traditional methods. Some of the most important information available on the web can be found in the opinions expressed by users, such as through customer product reviews. Online customer reviews provide the user with quick access to information related to hotels and help them to make faster decisions. Customer reviews are a bridge between sellers and buyers. Trust in online customer reviews is growing and often, positive customer reviews increase the trust others have in a hotel. Customer reviews are also important because they often reference hotel features that do not appear in hotel descriptions. The availability of clear and precise reviews is valuable for hotel companies.

The descriptions and categorizations of hotels on websites do not offer enough valuable information for all customers. Sometimes, a customer wants to know more details about a hotel's characteristics. Booking processes could therefore be more helpful for customers if they provided information on a lot of features, which would also improve their search results.

1.2 Objective

The overall objective of the research is to facilitate and strengthen the process of hotel searches and booking for customers by extracting elements of opinions from online customer reviews of particular hotels to determine the features commented on by consumers in order to match their preferences with the best hotels.

1.3 Motivation

The amount of available information nowadays makes the actual system each day more concerned about how to handle information overload, ensuring that the user will have access to the best sources with the least effort. On recent years, special attention has been giving also to the amount of produced user-generated content. The e-commerce sector is one of the most affected by the amount of data produced by customers, which increased dramatically during the phase known as Web 2.0 customer's opinions represent a valuable unique type of information which should not be mistreated or ignored by the research community. Thus, this work emphasizes the need of special mechanisms that aims to provide the community better ways to take full advantage from this data.

From the customer perspective, considering others opinions before purchasing a product is a common behavior long before the existence of Internet. In the era of the digital world, the difference is that a customer has access to thousands of opinions, which greatly improves decision making. Basically, customers want to find the best for the lowest price. In other words, they search for products that best fulfill their needs inside a price range that they are willing to pay.

It is important to emphasize that the benefit of analyzing other opinions, comes from their neutral nature, which are usually not linked to an organization or company. They represent the voice of ordinary consumers, and that differs greatly from ads (advertisements are biased and tend to favor the product, emphasizing the positives aspects and concealing the negatives ones).

From the e-commerce perspective, receiving consumers' feedback can greatly improve its strategies in order to increase profits of the sector.

It is common to find products with thousands of opinions; thus it could be a hard task for a customer to analyze all of them. Also, it could be a very tiresome work to find

opinions about just some features from a product, usually a requirement for an experienced customer.

An important difference makes the actual ranking mechanisms not so efficient to depict the information represented by opinions. This difference is mainly due to nature of textual information in the world. These information are either facts or opinions. The actual search systems are focused on facts (e.g ranking mechanisms used by search engine). One fact is usually equal to all other same facts. An opinion however is a belief or judgment of a subject. Therefore, one opinion from an object under discussion is usually different from multiple opinions for the same. In this sense, a summarization mechanism portrait betters the reality of opinions and thus provides better ways for users to draw conclusion out of them. This work presents ways for locating, extracting, classifying and summarizing opinions or reviews on the Internet. The proposed framework will combine several techniques to extract valuable information out of natural language text (user-generated content), in order to provide enrichment of the experience of users by taking advantage of the available content in a more intelligent and organized way. As a consequence of the employed techniques, data can be structured; this will also provide a necessary bridge for many applications to be able to fully interact with others in a Web 3.0 context.

1.4 Proposed solution

With this research we proposed a method using data mining. Researchers have identified data mining as a best solution for digging hidden patterns within large repositories of data using the support of different software tools with its ability to deal with large number of dynamic variables simultaneously.

As the initial step find out what and how hotels have been booked by online. It makes a big impact on hotel booking process. For this I found hotel reviews are the can be the data. Then data set obtained by Trip Advisor and prepared for further analysis. This data (opinion) used to build a Hotel Ranking System.

1.5 Resource Required

Access to common resources like books, journals, research papers and magazines about data mining on customer reviews. WebHarvy is used to scrap review from web site (www.TripAdvisor.com), QDA miner is used to semantic analysis for unstructured data to make structured well as to identify the hotel features. XAMPP used to create and develop my own application using Web server technologies. NLTK 2.0.4 tool kit is used to do sentiment analysis. Other than these resources PC or a Laptop with good processor speed will be used for the research purpose.

1.6 Summary of chapters

Including this chapter, this work is divided into eight parts as following**:**

Chapter 2 Development challengers in Data Mining

Development challengers in Data Mining will investigate some researches in the Opinion Mining field. These works mainly aim areas such as sematic analysis, sentiment analysis and feature identification. This chapter will demonstrate different strategies used for different works to cope with semantic analysis, sentiment analysis and feature identification in opinions, exposing the pros and cons of the employed methods.

Chapter 3 Technology Adapted

This chapter presents the technology used to the system, describe the theory and methodology.

Chapter 4 Approach

This chapter presents full picture of the entire solution, towards opinion mining from customer reviews.

Chapter 5 Design

Design presents the overall architecture of the opinion mining system and details about each subsystem task. The details about each basic subsystem are also introduced.

Chapter 6 Implementation

This chapter presents the technologies involved on the development of a platform. Also, all the aspects relevant to the chosen technology will be discussed and code demonstration.

Chapter 7 Evaluation

This chapter shows all the results achieved with new system. The evaluation of the system will be focused on accuracy.

Chapter 8 Conclusion and further work

The last chapter will discuss the results achieved with the system and the biggest shortcomings found during the development. Also, it will highlight the new directions expected for future works, where more efforts should be put on.

1.7 Summary

This chapter gave introduction about the research and brief idea about proposed solution created using data mining techniques. This chapter showed the aims and objective of this research work. Next chapter shows current development and similar works by others.

Chapter 2

Development Challengers in Data Mining

2.1 Introduction

Chapter 1 gave a comprehensive description of the overall project described in this thesis. This chapter provides a critical review of the literature in relation to developments and challenges in Data mining. For this purpose, the review of the past research has been presented under two major sections, namely early developments & Modern trends and future challenges. At the end of this chapter define the research problem as the "write problem" and identify the data mining technology that can be used to address the problem.

2.2 Early developments & Modern trends in Data mining

Walter Kasper [1] and others have done a research on Sentiment Analysis for Hotel Reviews. They presented a web based opinion mining system for hotel reviews and user comments that supports the hotel management in monitoring what is published on the web about their houses. The acquisition of reviews from the web is handled by a web crawler and they used Combining Statistical and IE Polarities for the process. The system is capable of detecting and retrieving reviews on the web, to classify and analyze them, as well as to generate comprehensive overviews. The system provides good performance for the analysis and the classification tasks. Mita K. Dalal [2] and others have done a research on Opinion Mining from Online User Reviews Using Fuzzy Linguistic Hedges. They have used Feature-based sentiment classification is a multistep process that involves preprocessing to remove noise, extraction of features and corresponding descriptors, and tagging their polarity. This technique extends the feature-based classification approach to incorporate the effect of various linguistic hedges by using fuzzy functions to emulate the effect of modifiers, concentrators and dilators. They propose an opinion mining system that can be used for both binary and fine-grained sentiment classifications of user reviews. Choochart Haruechaiyasak [3] and others have done a Case Study on Hotel Reviews and propose a framework for constructing Thai language resource for feature-based opinion mining; the used approach is for extracting features and polar words from opinionated texts is based on

syntactic pattern analysis. The proposed method first constructs a set of patterns from a tagged corpus. The extracted patterns are then used to automatically extract and collect more sub-features and polar words from an untagged corpus. The performance evaluation was done with a collection of hotel reviews obtained from a hotel reservation website.

Edison Marrese-Taylor [4] and others have done a research on extended Bing Liu's aspect-based opinion mining technique. They also offer an approach for considering a new alternative to discover consumer preferences. The proposed models to define and extract opinions from web documents present a simple, yet relatively effective manner of transforming the unstructured data about opinions available on the web. However, the algorithm for aspect expressions extraction, based on frequent nouns and NPs appearing in reviews, achieved a poor performance in the tourism domain. Larissa A. de Freitas [5] and Renata Vieira have done a research Ontology-based Feature Level Opinion Mining for Portuguese Reviews; experiments have considered movie and hotel ontologies Movie Ontology and Hontology. They have introduced state-of-the-art and preliminary results about ontology based feature level opinion mining. The algorithm to deal with Portuguese online review is presented. Also, this work intends to summarize opinions about objects and objects features in different levels for the ordinary or expert users. Vijayshri R. Ingale, [6] and others have done a research on Sentiment Analysis by Visual Inspection of User Data from Social Sites - A Review on Opinion Mining. The aim is to replace the human aspect in decision-making. Data conditioning is a hopeful key to the amounts of data that is on the rise, which is of need to be mined The proposed rising programmed data conditioning tools that provide a more effective dataset to be processed by the data mining.

Dr. S. Sagar Imambi [7] and others have done a research on Analyzing Customer Reviews Using Opinion Mining. They applied a feature weighting technique, which gives weights not only for the features but for the adjectives, used in opinion sentence. Individual categories also showed nearly the same precision and recall. But our procedure depends upon the common adjectives list and the phrases used in opinions. G. Vinothini [15] and others, have done a research on feature based opinion mining for product reviews. Their research tried to find out better product features from customer reviews. The performance evaluation results show that the precision was found to be 0.73, Recall was found to be 0.83 and F-measure was found to be 0.74.

Experimental results indicate that the proposed techniques are effective in performing their tasks.

2.4 Future challenges

Walter Kasper [1] and others further research will be necessary especially with respect to the demarcation of evaluative and neutral text as well as to the handling of multi-topic segments, especially for the user interface. Mita K. Dalal [2] and others future research, would be to build an advanced opinion mining system capable of rating the authenticity of a user review based on mining opinion threads of secondary reviewers. Choochart Haruechaiyasak [3] and others future work they plan to complete the construction of the corpus by considering the rest of main features. Another plan is to include the semantic analysis into the pattern extraction process. Edison Marrese-Taylor [4] and others further research, *Aspect expressions* that were extracted only represent a small percentage of the ones that were manually detected, and the method needs to be tested for all possible expressions on the topic of tourism in order to give a more conclusive analysis. Larissa A. de Freitas [5] and Renata Vieira's future study, to work on the improvement of the results reached by the algorithm. Plan to verify the inter-annotator agreement using the kappa coefficient. Moreover, they intend to use lemmatizer in pre-processing and properties, instances and hierarchies of ontologies in identification feature. Also, they plan to add list of verbs, list of adverbs and list of nouns in polarity identification. At last, we will apply a set of linguistic rules, such as: negatives, intensifiers and irony/sarcasm detection. Vijayshri R. Ingale, [6] and others Future work includes more challenges facing the development and wide spread use of data mining and visualization techniques. Dr. S. Sagar Imambi [7] and others, future they would like improve adjective phrases database, so that we can get better result.

2.5 Summary

This chapter presented a comprehensive literature review on the opinion mining research and identified the research problem as the inadequate attention to reliability of opinion mining algorithms. Next chapter will discuss the technology to be used for our solution.

Chapter 3

Technology Adapted

3.1 Introduction

In chapter 2 the review of other related studies and their importance as well as the improvements are described in a comprehensive manner. This chapter describes the technologies identified to conduct the research. How the Data collected, data Analysis Data Categorization, and opinion mining.

3.2 Tools & Technologies of Hotel Ranking system

- Programming language-PHP
- Scripting language-HTML
- Xampp server
- Data Base Management system-MYSQL
- Design Tool-Rational Rose
- Sentiment Analysis Tool- Python NLTK (Natural Language Toolkit)
- Semantic Analysis Tool-QDA Miner
- Web Scrapping Tool-WebHarvy

3.3 Data Mining

Generally, data mining (knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product

positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

3.3.1 Opinion Mining

Opinion Mining is a field of Web Content Mining that aims to find valuable information out of user's opinions. Mining opinions on the web is a fairly new subject, and its importance has grown significantly mainly due to the fast growth of e-commerce, blogs and forums.

From the business perspective getting important information out of opinions can represent a good source of advertisement or product feedback. For example, a web site specialized on electronics reviews could place advertisements on their pages based on consumer opinions. For instance, if the majority of user's express negative opinions about a given product, the web site could place ads from an alternative product from a competitor. Also, manufacturers can get the feedback of their products to improve their products or services.

An opinion is a personal belief or judgment of a subject, so it is important to differentiate between types of opinion holders. To achieve this differentiation, the term product review would refer to an expert opinion (or someone with a privileged status or higher knowledge on the subject) while user's opinions or user's reviews, refer to opinions given by common users (in this case, the consumers). An expert opinion is usually far superior in quality, richer in technical details, and goes through all the most relevant aspects of a product. Users and customers usually give opinions with less commitment, and generally what you see are just some pros and cons being discussed. In this work the terms user's opinions, users review and product opinions are used interchangeably which are different from product review as explained above.

This work will not deal with product reviews; thus it will focus only on mining ordinary customer's opinions on the Web where there is no different hierarchy level between opinion holders.

Better results could be obtained by using an opinion-mining system to systematically collect and analyse information from the external business environment to aid in improving the booking process for customers.

3.3.2 Semantic Analysis

Semantic Analysis is the study of the semantics or structure and meaning of speech.

It determines the meaning of given sentence and represents that meaning in an appropriate form. Semantics, as a part of linguistics, aims to study the meaning in language. The language demonstrates a meaningful message because of the semantic interaction with the different linguistic levels. The main objective of semantic analysis is to minimize the syntactic structures and provide the meaning, finding synonyms, word sense disambiguation, translating from one natural language to another and populating base of knowledge. [14]

3.3.3 Sentiment Analysis or Sentiment Classification

Sentiment Analysis or Sentiment Classification is the methodology to naturally focus the sentiments communicated in a bit of plain content utilizing some regular automated preparing systems. To be specific, term Sentiment is exceptionally wide and it constitutes feelings, opinions, dispositions, particular encounters, and so forth. In this theory, we speak just about the opinions communicated in writings which are composed in texts which are written in human readable natural language, in social media.

Sentiment analysis is a procedure for following the views of the clients around a specific item or subject. Sentiment analysis, which is likewise called opinion mining, includes in building a framework to gather and look at opinions about the item made in blog entries, remarks, audits or tweets. Sentiment analysis might be helpful in a few ways.

Case in point, in showcasing it helps in judging the achievement of a notice crusade or few item dispatch, figure out which forms of an item or administration are prominent and even recognize which demographics like or aversion specific attributes.

3.4 Web Scraping

Web Scraping is the process of information extraction from resources that are located on the World Wide Web and the classification of scraped and unstructured data (usually found in HTML pages) in a structured form like Spreadsheet or database

3.5. Xampp server

Xampp is software which acts like a local server in your system as an real time server how it works. This software is really helpful to have hands on experience before deploying your application on real server so that you get an idea how your application works.

XAMPP stands for Cross-Platform (X), Apache (A), MySQL (M), PHP (P) and Perl (P). It is a simple, lightweight Apache distribution that makes it extremely easy for developers to create a local web server for testing purposes. Everything you need to set up a web server – server application (Apache), database (MySQL), and scripting language (PHP) – is included in a simple extractable file. XAMPP is also cross-platform, which means it works equally well on Linux, Mac and Windows. Since most actual web server deployments use the same components as XAMPP, it makes transitioning from a local test server to a live server is extremely easy as well.

XAMPP has four primary components. These are:

1. Apache: Apache is the actual web server application that processes and delivers web content to a computer. Apache is the most popular web server online.

2. MySQL: Every web application, howsoever simple or complicated, requires a database for storing collected data. MySQL, which is open source, is the world's most popular database management system. It powers everything from hobbyist websites to professional platforms like WordPress. MySQL is a relational database management system (RDBMS) based on Structured Query Language (SQL). MySQL runs on virtually all platforms, including Linux, UNIX, and Windows. Although it can be used in a wide range of applications, MySQL is most often associated with web-based applications and online publishing

3. PHP: PHP stands for Hypertext Preprocessor. It is a server-side scripting language that powers some of the most popular websites in the world. It is open source, relatively easy to learn, and works perfectly with MySQL, making it a popular choice for web developers.

4. Perl: Perl is a high-level, dynamic programming language used extensively in network programming, system admin, etc. Although less popular for web development purposes, Perl has a lot of niche applications.

Different versions of XAMPP may have additional components such as phpMyAdmin, OpenSSL, etc. to create full-fledged web servers.

3.6 HTML

Hyper Text Markup Language is the standard markup language for creating web pages and web applications. With Cascading Style Sheets (CSS), and JavaScript, it forms a triad of cornerstone technologies for the World Wide Web. Web browsers receive HTML documents from a webserver or from local storage and render them into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.

3.7 Hardware Requirements

Below the list of minimum hardware requirements for the personal computer to develop the Hotel Ranking System: -

- Computer with 2GB RAM
- 2GHz or more processing power
- 500GB Hard disk space or higher
- 15’ Monitor (can support up to 1024*768 resolution)
- 56Kbps Modem
- Key board and Mouse

3.8 Summary

This described about the technology this research adapts to solve the problem and how these techniques are appropriate to create Hotel Ranking system. Next chapter will be discussing about novel approach of opinion mining and find the solution for hotel ranking.

Chapter 4

Novel Approach of opinion mining to Improve Hotel booking Process

4.1 Introduction

Here we describe our approach for performing Data mining and other techniques for getting useful information from online customer review to build HRS (Hotel Ranking System).

4.1.1 Hypothesis

My hypothesis is developing a Hotel Ranking System by analyzing online hotel review which can improve booking process of hotels.

4.1.2 Users

Number of users who can be benefited by the Ranking Systems in multiple ways. More importantly, Travelers, such as tourist, business travelers, researchers, hoteliers can be directly benefitted by this solution. Those who are interested in Data mining can also use this system for learning purpose.

4.1.3 Input

The system is Web base, the Location of the traveller would prefer to visit and Hotel Feature based the type of the journey would be the input for the system.

The system can accept from various devices include a computer connected to internet, smart phones, Tabs, Computer etc.

4.1.4 Output

The output of the system would be hotel names which are ranked and associate features based on investigating hotel reviews according to selected location and hotel feature.

4.1.6 Features

In connection with the input, output, users and process, the over features of the system include the following characteristics,

- online solution
- Single input forms
- high level of accuracy
- user friendly
- can use as research purposes
- Help the Travelers and hoteliers

4.2 Methodology

By this research the proposed solution, a Hotel Ranking System which provides users a quality feature-based summary, out of product opinions from ordinary customers on the web. Therefore, the system must achieve five important goals:

(1) Gather opinions from sources on the web.

(2) Do Semantic analysis to find the features.

(3) Prepare a feature-based summary from the opinions retrieved in.

(4) Import the data to Database

(5) Do Sentiment Analysis.

(6) Use Association rule.

(7) Get the Hotel Ranking based on selected hotel feature, with frequent searched hotel features.

4.2.1 Data Collection

Reviews Collected from Trip Advisor and Selected 4 cities from Sri Lanka, which are Colombo, Kandy, Nuwaraeliya and Hikkaduwa.

Using the Trip Advisor search engine, select a city and hotels, and searched a list of hotels will be displayed. Each Hotel link is used to collect user reviews.

Data collection has done using Web Scraping technology. Web Scraping (also termed Screen Scraping, Web Data Extraction, and Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.

A web scraping software will automatically load and extract data from multiple pages of websites based on your requirement. It is either custom built for a specific website or is one which can be configured to work with any website. With the click of a button you can easily save the data available in the website to a file in your computer.

4.2.2 Semantic Analysis

The process of semantic analysis is a complex one. Basically in semantic analysis, we try to obtain an initial representation of the meaning of the sentence from the possible parses of the syntactic analysis stage. This initial representation of the meaning of the sentence is known as the logical form.

Natural language is ambiguous by nature. A simple sentence can be interpreted in many different ways. In order for the computer to process the sentence, it needs to know the exact meaning of the sentence. Thus, the logical form is needed as an intermediate unambiguous representation of the meaning of the sentence.

At this stage of the analysis, only the context-independent meaning of the sentence is derived. By this, we mean that the sentence is treated as a stand-alone line by itself, disregarding any knowledge of previous sentences or the world. Because of this, certain sentences which are dependent on knowledge of previous sentences will not be fully interpreted. That is when the last stage comes in. Anyway, once the semantic

relationships are determined, certain words senses which are found to be impossible are then eliminated from consideration and the remaining choices are passed to the next stage.

Qualitative Data Analysis

To identify the hotel features and the expression about the particular features semantic analysis is done using a semantic analysis tool called QDA miner.

Collected reviews will be unstructured data which very heard to analyses; instead this unstructured data has to be categorized based of aspect base (Hotel features). To identify the hotel feature and to find the expression about particular feature a tool is used, which is QDA miner lite.

QDA Miner lite is a qualitative data analysis software package for coding textual data, annotating, retrieving, and reviewing coded data and documents. The program can manage complex projects involving large numbers of documents combined with numerical and categorical information. QDA Miner also provides a wide range of exploratory tools to identify patterns in coding and relationships between assigned codes and other numerical or categorical properties.

4.2.3 Aspect Oriented Customer Review Mining

Aspect-oriented review mining goes one step further and analyzes the customers' sentiment with regard to individual product aspects.

Aspect-oriented review mining involves the joined analysis of two dimensions. On one dimension we want to discover all relevant product aspects and on a second dimension we want to identify related expressions of sentiment and determine their polarity. Here transforms the unstructured information of a review text into a structured, aspect-oriented summary

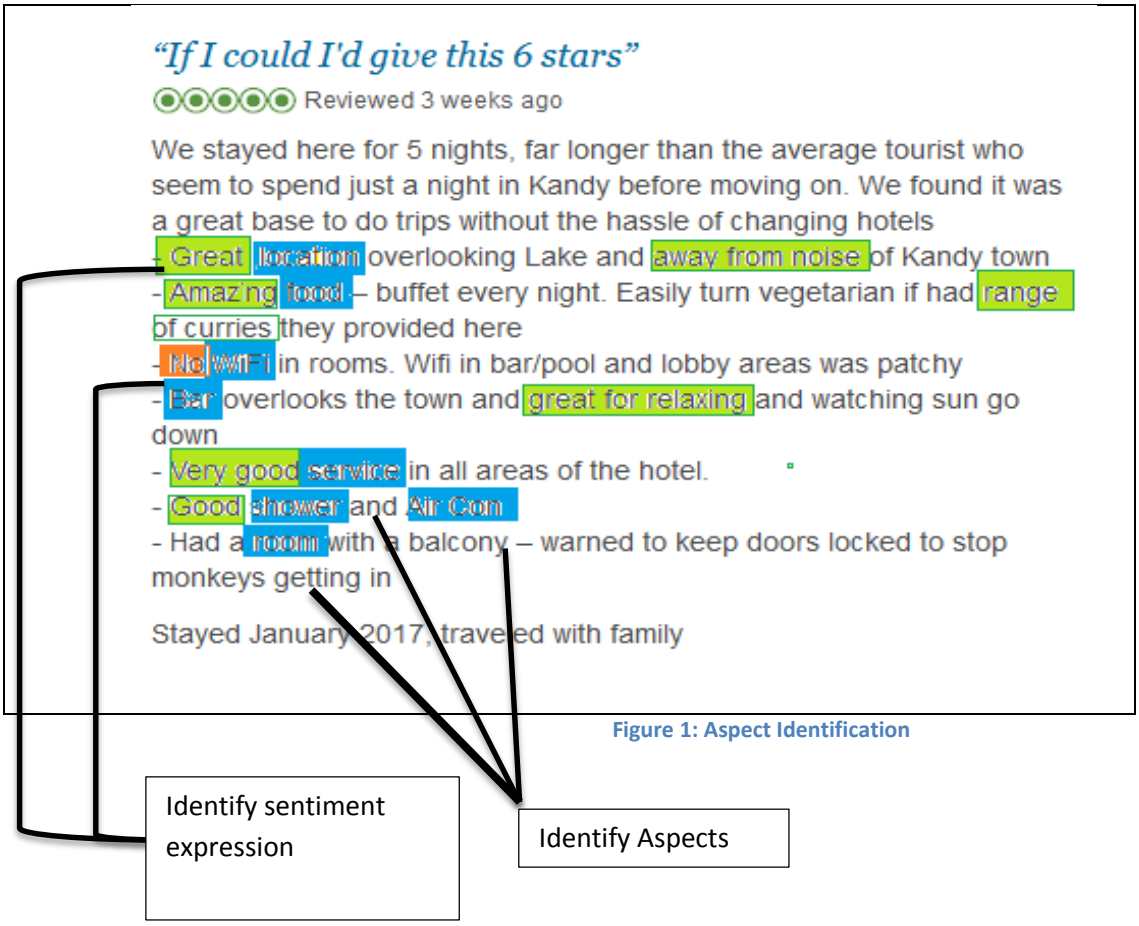


Figure 1: Aspect Identification

Aspect: An aspect refers to a distinct ratable facet of an entity (product), and it can be a semantic attribute, component, function, or feature of the entity, e.g., "Service" of a hotel.

Aspect Term: An aspect term refers to a particular term used to express an aspect. For example, the specific aspect terms "service", "Food", and "cleanness" are all indicative of the unique semantic aspect "screen" on cellphone review domain. Aspect term consists of explicit aspect term and implicit aspect term.

Explicit and Implicit Feature

When a feature f is readably available in a review R, f is called an explicit feature. There are cases where a feature f is not readably available, in R, therefore it is considered to be an implicit feature.

4.2.4 Method of storing data to Database

Collected Data analysed and changed in to structured format and uploaded to Database

Xampp server

XAMPP is a free and open source cross-platform web server solution stack package, consisting mainly of the Apache HTTP Server, MariaDB database, and interpreters for scripts written in the PHP and Perl programming languages [9]. XAMPP helps to create and develop your own applications using Web server technologies.

MY SQL:

MySQL is an open source relational database management system (RDBMS) based on Structured Query Language (SQL). MySQL runs on virtually all platforms, including Linux, UNIX, and Windows. Although it can be used in a wide range of applications, MySQL is most often associated with web-based applications and online publishing

4.2.5 Opinion Mining/Sentiment Analysis Tool

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker**.** A common use case for this technology is to discover how people feel about a particular topic.

Today's algorithm-based sentiment analysis tools can handle huge volumes of customer feedback consistently and accurately. Paired with text analytics, sentiment analysis reveals the customer's opinion about topics ranging from your products and services to your location, your advertisements, or even your competitors.

Python NLTK (Natural Language Toolkit) is a sentiment analysis tool which is used to do sentiment analysis for the data set to build the Hotel Ranker System. NLTK is a leading platform for building Python programs to work with human-language data. It provides easy-to-use interfaces to lexical resources, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an

active discussion forum [13]. NLTK 2.0.4 powered text classification process. It can tell you whether it thinks the text you enter given statement is expresses positive sentiment**,** negative sentiment**,** or if it's neutral**.** Using hierarchical classification,

4.2.6 Natural Language Processing

Natural Language Processing (NLP), also known as computational linguistics, is a field of computer science that studies interactions of human languages with computers. The main goal of NLP is to enable effective human-machine communication, which could be either as spoken or written form. Here, only the written form will be addressed. For many applications, is desirable to automatically process texts written in natural language. Computers can parse and automatically generate natural language texts, extract semantics from them and identify real world objects. As a consequence, many new applications could benefit from it, as well as existing ones which could become more human friendly. Some examples include search engines understanding natural language text queries and data information extraction applications which could interpret a large amount of text and store just the significant parts in a database**.**

4.2.7 Association Mining

Association analysis, which is useful for discovering interesting relationships hidden in large datasets the uncovered relationships, can be represented in the form of association rules or sets of frequent items. For Association mining Ariori algorithm is used. [17], [18]

Apriori Algorithm

The Apriori Algorithm is an influential algorithm for mining frequent item sets for boolean association rules.
Key Concepts:

- Frequent Item sets: The sets of item which has minimum support (denoted by Li for i^{th}-Itemset).
- Apriori Property: Any subset of frequent itemset must be frequent
- Join Operation: To find L k, a set of candidate k-itemsets is generated by joining Lk-1 with itself.

4.3 System Architecture

The Hotel Ranker system is a web base Application. The system has two main components: the front end and the back end. The backend consists of a database, an algorithm analyzer module, a request handler for the frontend, and a database connector. The frontend consists of one page.

4.4 Front End

The frontend is designed to be simple and elegant with minimalistic controls and settings to avoid confusions. When users first visit the site, they would be able select location they need to visit and based on their travel purpose select the Hotel Feature given in dropdown and search input that provides auto-complete functionalities to the natural language queries. After the search, the users are presented with a list of top rated hotels based on selected hotel feature and association features. Front end web pages created using HTML, CSS, and Java Script.

HTML: **Hyper Text Markup Language** is the standard markup language for creating web pages and web applications. With Cascading Style Sheets (CSS), and JavaScript, it forms a triad of cornerstone technologies for the World Wide Web. Web browsers receive HTML documents from a webserver or from local storage and render them into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.

4.5 Back End

PHP is a server scripting language, and a powerful tool for making dynamic and interactive Web pages. PHP scripting language is used to create the backend of Hotel Ranking System.

The backend dataflow is a straightforward one as seen in many web based applications. When a request for list of products comes in through the frontend, the frontend handler starts a search for the query terms into the database. The database is a central storage for all the user review and list of hotels. Through the database connector, Sentiment Analysis Tool (NLTK 2.0.4) should be connected and find opinion for particular feature is negative or positive or neutral and calculate the aspect rating scores used in ranking the list of products as the result.

Based on the result produced by Sentiment analysis will be taken as an input for the Association mining using Apriori Algorithm and create Association rule. The result of this will able to find in each hotel along with selected aspects what the other aspect that customers search about positive.

Example: Hotel Citrus

Whoever searching about better staff also search about good Service and good food.

The frontend handler then publishes this result to the incoming request. To program this PHP programming Language has been used.

4.9 Summary

In this chapter described novel approach to analyze customer feedback about hotels, in this pointed out the novel approach offers an efficient accurate solution for hotel ranking systems.

Chapter 5

Design of IHBP

5.1 Introduction

The previous chapter gave full picture of the entire solution. This chapter describes the design of the process presented in the approach. We design the solution as a client server system with a backend database. Here we describe the top level architecture of the design by elaborating the role of each component of the architecture.

5.2 General Architecture

The general architecture is composed by eight main components. These components are called: Data Collection, Semantic Analysis, Data Set 1, Sentiment Analysis, Associate Mining, Data Set 2, My Code and Front end.

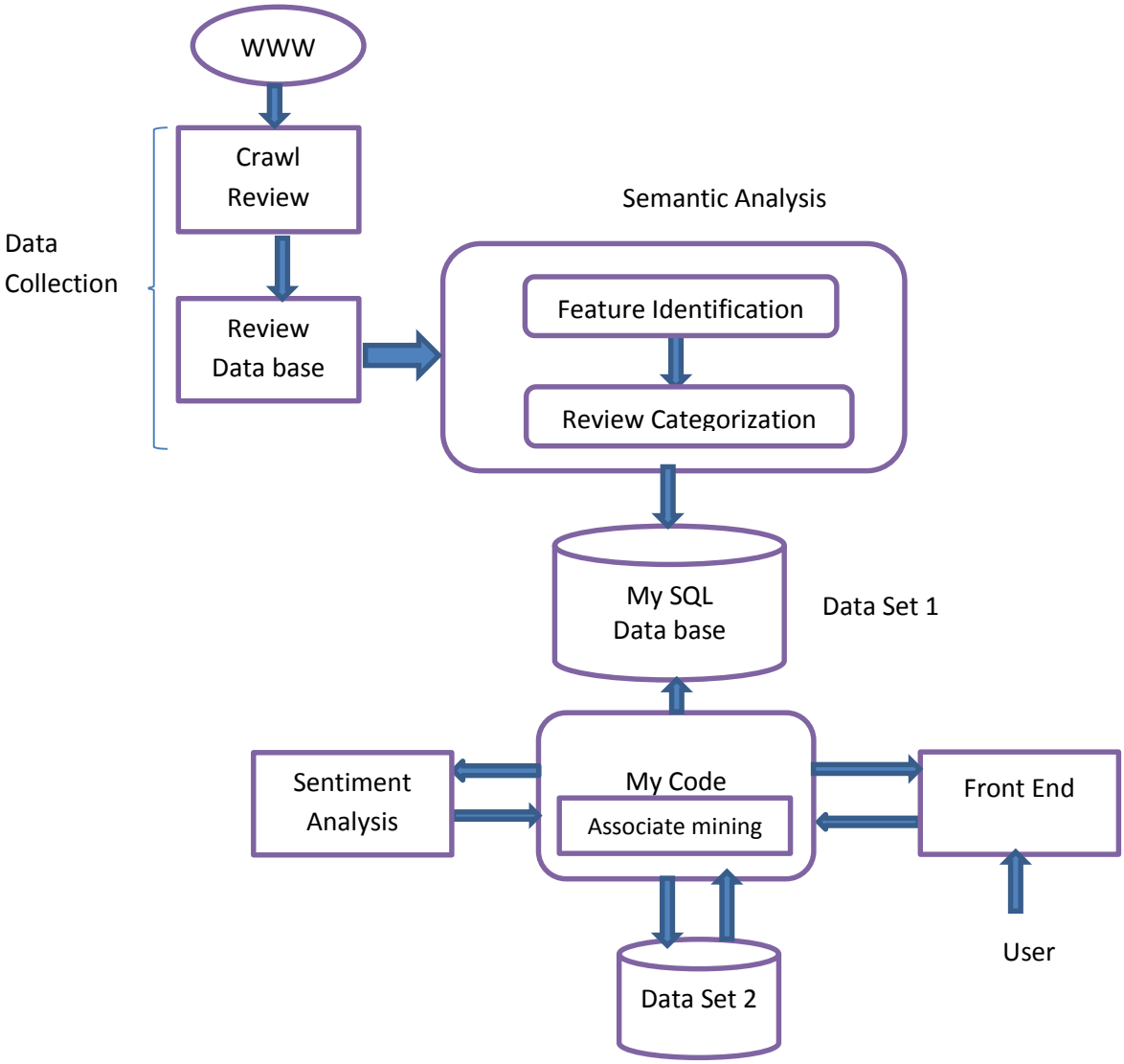


Figure 2: Top Level Architecture

5.2.1 Data Collection module

In this module there are 3 parts exist;

1) World Wide Web

The Web or World Wide Web is basically a system of Internet servers that support specially formatted documents. The documents are formatted in a markup language called HTML (Hyper Text Markup Language*)* that supports links to other documents, as well as graphics, audio, and video files.
This means you can jump from one document to another simply by clicking on hot spots. [16]
Trip Advisor is the source of data to this research, TripAdvisor website stores information about hotels at each page and determine the exact attribute name (hotel name, hotel address, reviews, customer name) used at the hotel pages.

2) Web Crawling

Trip advisor All hotels must have more than 500 reviews. From this around 50 reviews has been retrieved for each hotel using a web crawler. To collect data a web crawling software used which is WebHarvy.

3) Review Database

Collected reviews have been saved in csv file for easy import to database.

5.2.2 Semantic Analysis module

Collected reviews are unstructured. In this module Semantic Analysis performed to identify the hotel feature/Aspects from collected reviews and categorized the reviews according to identified hotel features. To perform Semantic Analysis a software tool is used which is called QDA Miner lite. In this module Hotel features identified, and categorized. Unstructured reviews convert in to structured format.

5.2.3 Review Database Module-Data set 1

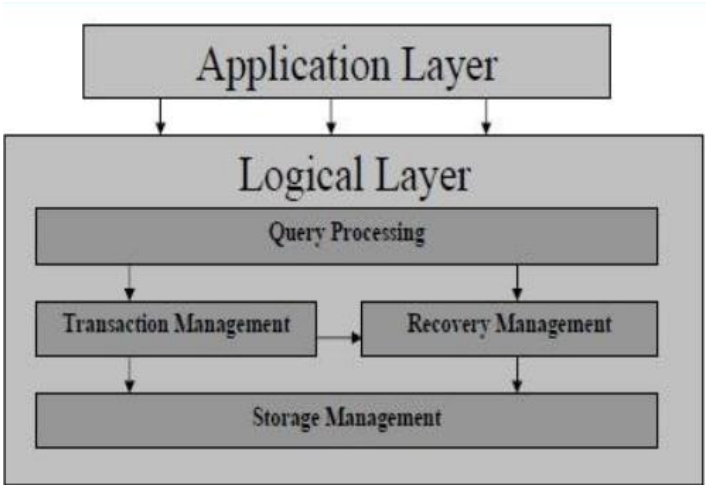


Figure 3: Architecture of My SQL Database

Structured reviews have been uploaded in to a database using the method data importing.

In the data set the following information are included

1. Four locations
2. Each location contains number of hotels
3. Each hotel got number of features
4. Each feature got number of reviews (data)

5.2.4 Sentiment Analysis module

Sentiment analysis is the measurement of positive and negative language. It is a way to evaluate written or spoken language to determine if the expression is favorable, unfavorable, or neutral, and to what degree.

This module, using the PHP My code, the Reviews which collected and imported to the database will be pass to the sentiment analysis tool which is Python NLTK 2.0.4 (Natural Language Toolkit). It can tell you whether the text you enter given statement is expresses positive sentiment**,** negative sentiment**,** or if it's neutral.

5.2.5 Data set-2 module

Based on the results produced by sentiment analysis, has been filtered positive reviews only to send to the next level which Associate mining using My Code module

5.2.6 My Code module

This module has been created using PHP scripting language which has coded to database connectivity to Python NLTK 2.0.4 and produced result is used to produce hotel ranking. Only positive reviews saved as dataset-2, this is used for next level which to calculate hotel ranking and Associate mining.

Data set-2 is used to find the frequent item pattern and association. The result is connected to Front end.

5.2.7 Front End module

User interact the system using front end. HTML page is connected with MyCode module to show the processed results. Front end is allowing user to select the location and the preferred hotel feature, and perform search. Results will be display the hotels in the particular city and Hotel Rank in descending order with associate hotel features along with selected hotel feature what are other features that customers reviewed positively.

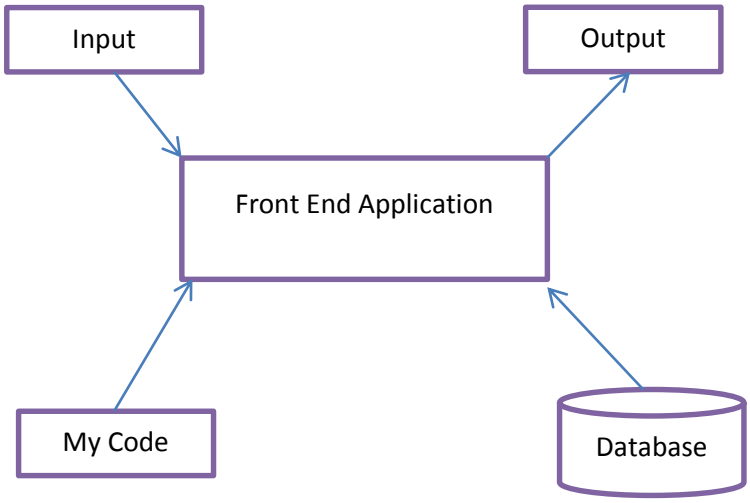


Figure 4: Architecture of Front End

5.3 Summary

This chapter provides details on research and applicability of selected research the research. Furthermore, this chapter focused on top level design and data mining design process also described in detail. The main level components of the systems also illustrated in the top level design diagram. What are the components interconnected together and how the system to build also explained.

Chapter 6

Implementation of IHBP

6.1 Introduction

In chapter 5 the design of the solution has been described in term of what each component does. This chapter will expose the implementation details of each internal module introduced on the last chapter. Moreover, this presents software and the algorithms used in each module.

6.2 Overall Solution based

Overall solution has been implemented as an open source application that can be accessed by any client running on any OS including Windows or Android. This is primarily client server architecture.

6.3 Data Collection

The current research focuses on the analysis of customer reviews, made about hotels in Sri Lanka, chooses from four locations which are Colombo, Kandy, Nuwara Eliya and Hikkaduwa.

To start of the research, reviews in English language were retrieved from TripAdvisor.com about hotels. Using the Trip Advisor search engine, select a city and hotels, and searched a list of hotels will be displayed. Each Hotel link is used to collect user reviews. The total of 40-50 English reviews from the period of 2016 – 2017 was collected for further analysis. All together 2465 reviews collected for the research. The reviews were saved in a csv format. To collect data web scraping technology is used using web scraping software called WebHarvy.

A **web scraping software** will automatically load and extract data from multiple pages of websites based on your requirement.

WebHarvy is a visual Web Scraper. You can configure WebHarvy to scrape data from websites using a point and click interface. There is no need to write any scripts/code or go through complicated setup procedures to scrape data. Just browse to the page

which contains the data to be extracted and select the required data by clicking on them [10].

- Open WebHarvy tool
- Load the relevant hotel webpage on Webharvy tool
- Click Record Button
- Select the Name of the Review and Collected as Name
  Extraction of the review provided
- Select reviewer Give a name as Reviewer
  To identify and avoid duplicate data
- Select Date, and collected as Year of Review
  Only collected the latest reviews (2016-2017)
- Select the reviews, Collected as Reviews
- Scroll Down and Select "Next" button to navigate to next page, Set this as next page
- Click Scrap button it collects Review from all the pages
- Save this Data into a Database directly or to Spreadsheet
- Collected around 50 Reviews for each hotel and this continued to all the hotels in the selected City
- This same repetition work done for another 3 cities and number of hotels
- Collected Data Save in Separate files in Spreadsheets format

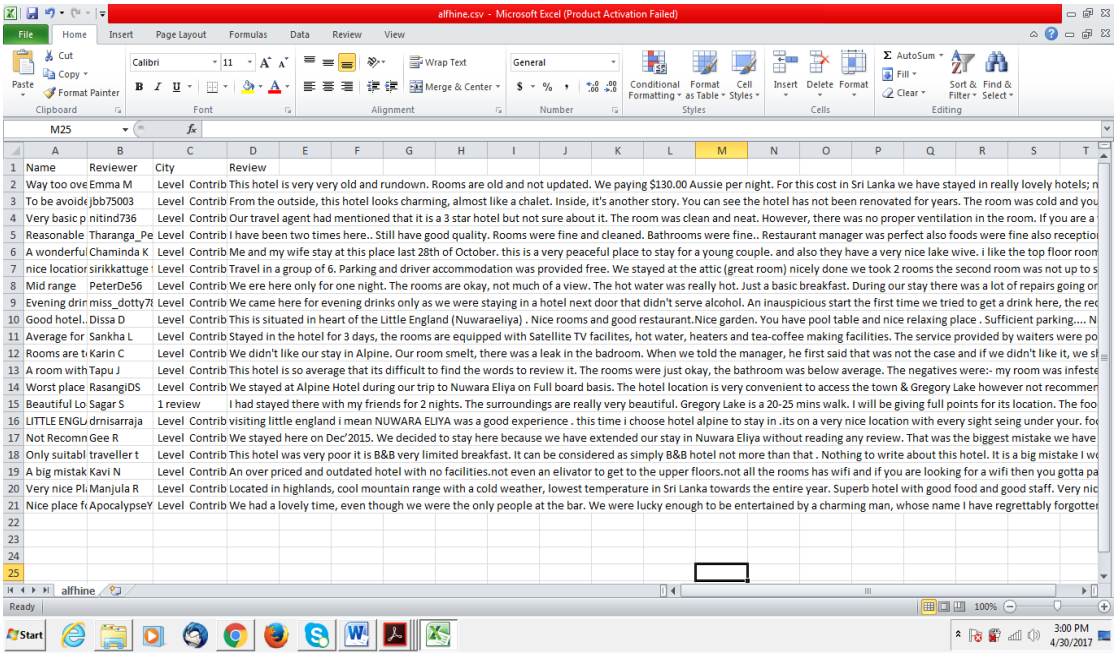


Figure 5: Sample of collected Reviews- Unstructured Data

6.4 Feature Identification using Semantic Analysis

Collected Reviews are in unstructured format. With reviews Reviewer, Reviewed date also collected to maintain the perfectness of data collection

Some additional configurations were made with obtained reviews in order to ensure the best results of the case study. First, information not connected to reviews was removed from text files, such as excessive hyperlinks, repetitions of the beginnings of reviews.

Unstructured data changed in to structured form. For this a semantic analysis tool QDA miner lite has been used

QDA Miner Lite is free and easy-to-use version of our popular computer assisted qualitative analysis software. It can be used for the analysis of textual data such as interview and news transcripts, open-ended responses, etc. as well as for the analysis of still images. QDA Miner Lite has been designed to meet the basic needs of researchers and analysts performing qualitative data analysis.

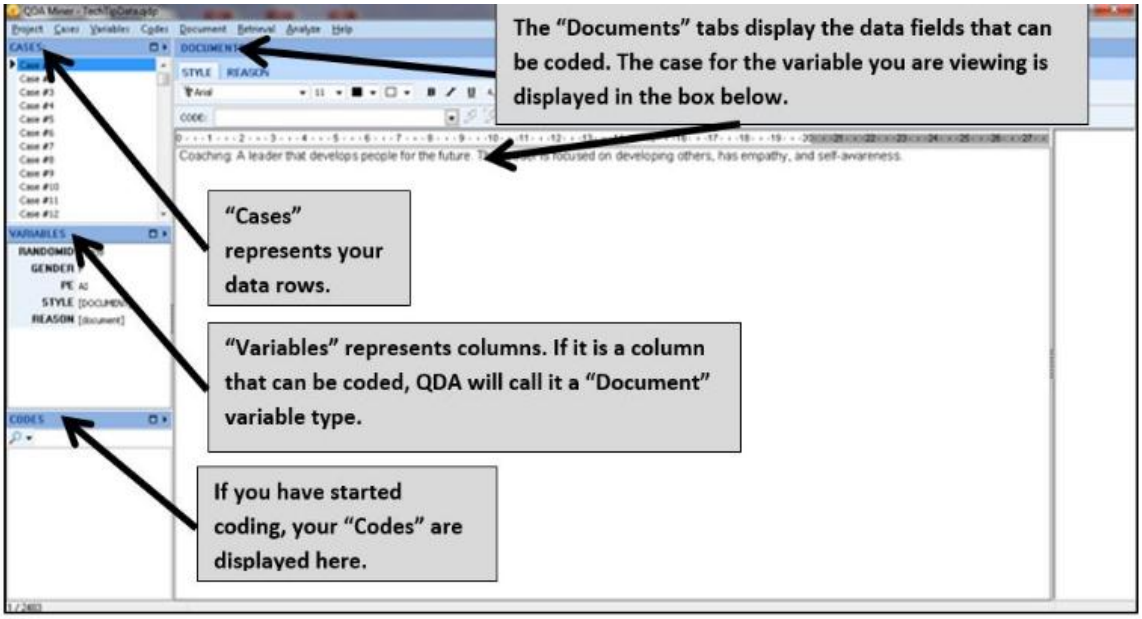


Figure 6:QDA Miner lite interface

The steps of using QDA Miner Tool for Semantic analysis of;

- Open QDA Miner
- Create a new Project
- Select 'Import existing data file' and import the unstructured data which done in data collection phase
- Select the relevant file and click to open
- Give a Name to the project and click to save
- And conform that uploading data is the correct one
- Create code using Codes tab
- Select the expression relevant to the created code, drag and drop
- This continues for all the reviews
- Save the file as csv file which can be directly upload in to the database

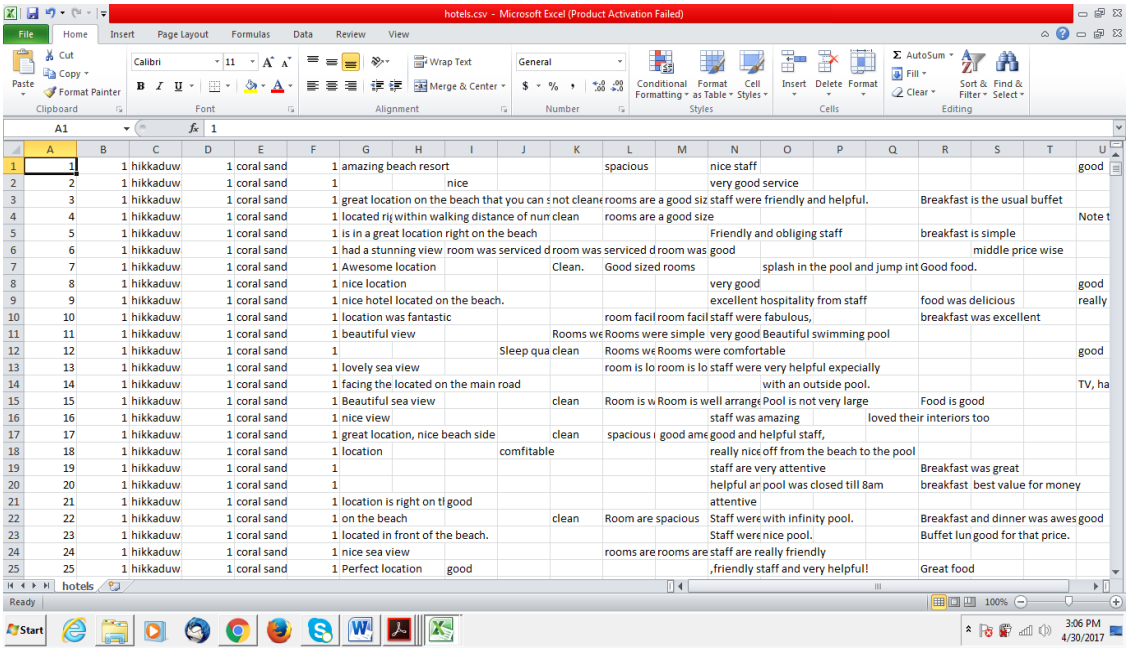


Figure 7: Analyzed Data-Structured Data

6.5 Implementation of Review Database

Data base created in My SQL, this database having 3 tables, which are hotels, location and reviews. Each table connected using foreign key.

Coding of database creation is included in Appendix A.

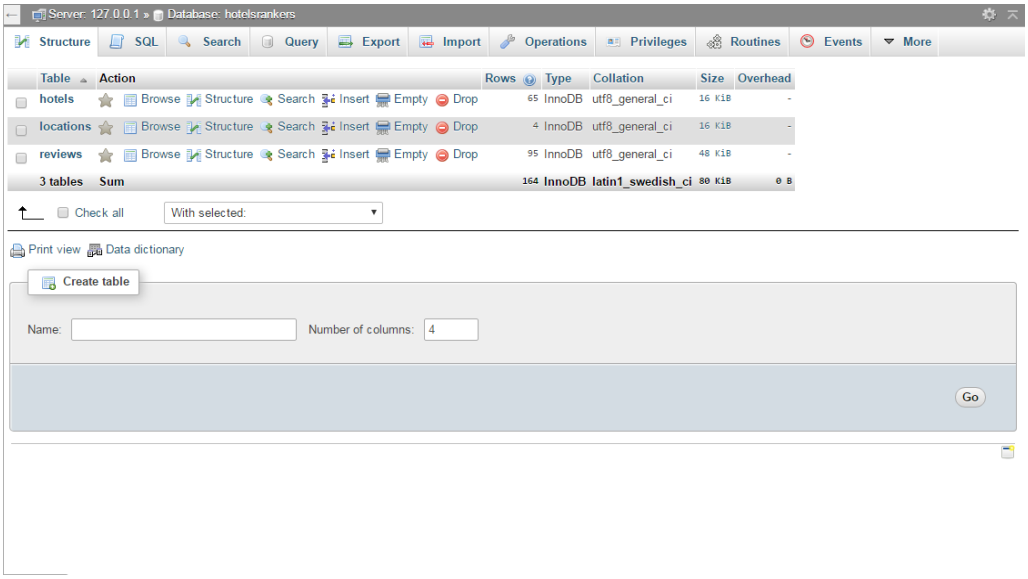


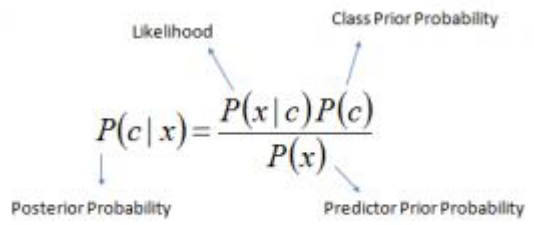
Figure 8: -Structure of Database

6.6 Sentiment Analysis Module

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. Saved reviews in the database should be analyzed which is connected to through API NLTK 2.0.4. From the result of sentiment analysis finds out whether written opinion is positive, negative or neutral. From this hotel ranking is calculated.

Sentiment analysis using a NLTK 2.0.4 powered text classification process. It can tell whether it thinks the text you entered expresses positive sentiment**,** negative sentiment, or if it's neutral. Using hierarchical classification**,** neutrality is determined first, and sentiment polarity is determined second, but only if the text is not neutral. NLTK 2.0.4 text processing software uses Naïve Bayes Classifier to perform sentiment analysis.

Bayes theorem provides a way of calculating posterior probability P (c|x) from P(c), P(x) and P (x|c). Look at the equation below. [21]



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P (x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

6.7 Implementation of MyCode Module

My code module is done by PHP language. Using code Database has been connected to a software tool NLTK 2.0.4 which does sentiment analysis based on the data set. For reviews result of hotel features and location, hotels have been save using arrays. A curl function is used to connect NLTK 2.0.4 API

```
{
        /* Curl request to http://text-processing.com API */
        $ch = curl_init();
        curl_setopt($ch, CURLOPT_URL,"http://text-processing.com/api/sentiment/");
        curl_setopt($ch, CURLOPT_POST, TRUE);
        curl_setopt($ch, CURLOPT_POSTFIELDS, 'text='.$row["feature"]);
        curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
        /* Process Curl response */
        $response = curl_exec ($ch);
        $res = json_decode($response, true);
        /* save the result */
        $results[$row["hname"]][] = $res['label'];
        curl_close ($ch);
}
```

Above code will connect database and the sentiment analysis tool, the result of sentiment analysis will be forward back to database. Using this result can be able to find out how many reviews are positive, based on this hotel ranking has been calculated. Only positive hotel features filtered out and used for two different products. One is to calculate hotel ranking and second is to perform association.
Hotel Ranking displays in Descending order from front end

$$\text{Hotel Rank} = \frac{\text{No. of Positive Reviews}}{\text{Total No. of Reviews}} * 100$$

Second is performing Associate Mining, for this Apriori Algorithm is used to identify the frequent item pattern and to generate association rules.

Association rule mining

To perform association mining two types of thresholds decided
Minimum sup=2
Minimum confident= 60%

Apriori Algorithm

The Apriori Algorithm is an influential algorithm for mining frequent item sets for boolean association rules.
Key Concepts:
• Frequent Item sets: The sets of item which has minimum support (denoted by Li for i^{th}-Itemset).
• Apriori Property: Any subset of frequent itemset must be frequent.
• Join Operation: To find L k, a set of candidate k-itemsets is generated by joining Lk-1 with itself.

The following are the steps involved in Apriori Algorithm:

Assume for Ck and Lk
Ck denotes candidate itemset of k size
Lk denotes frequent itemset of k size
Important steps of algorithm are:
1) Initially get frequent set Lk−1
2) Join step: get Ck by doing cartesian product of Lk−1 with itself
3) Those itemsets which are of size (k-1) and those are not frequent should not be a subset
of a frequent itemset of size k, so those should be removed
4) Finally frequent set Lk has been achieved

Result of frequent item set, Association rule will generate.

Association rule mining is defined as:

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short itemsets) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively.

To perform Association Mining Apriori library has been called to MyCode module [22].

6.8 Implementation of the Frontend Module

User interact the system using front end has been designed with single page to this module is developed using HTML, CSS, Java scripts. This interface is integrated with the MyCode module.

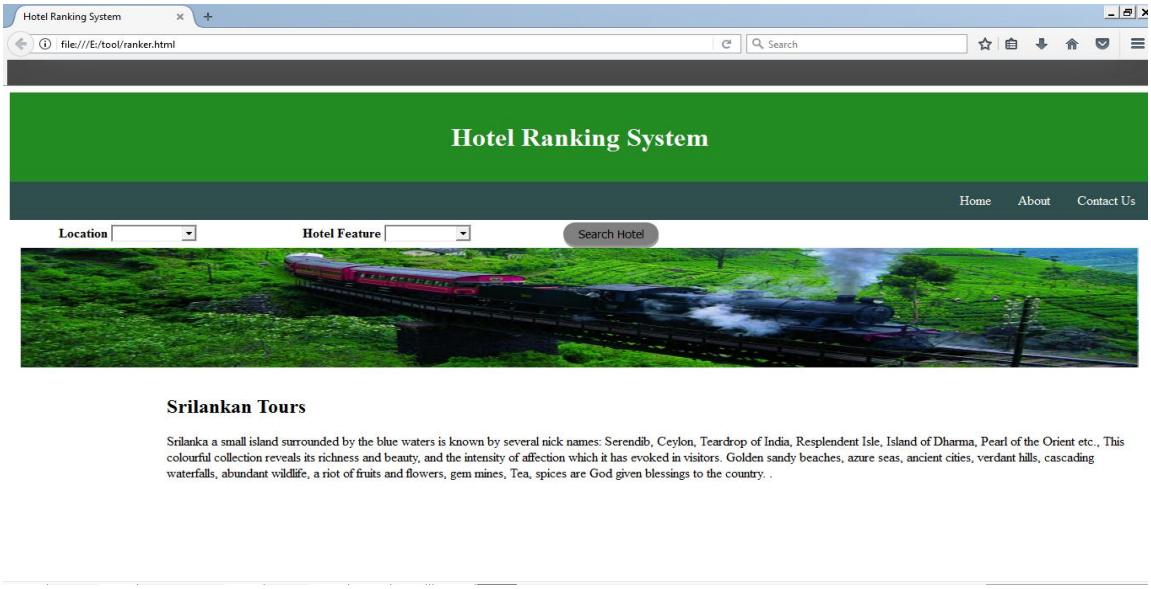


Figure 9: Front End Interface

6.9 Summary

This chapter provides implementation details of each module of proposed solution. Moreover, it mentioned software, algorithm and resulting models of each module in the design. Next chapter evaluates all the modules implemented in the solution.

Chapter 7

Evaluation

7.1 Introduction

This chapter presents how the software solution can be tested with respect to different aspects such as functionality, reliability, efficiency, maintainability and portability.

Testing is a process of evaluating a system and its components with the intent to find whether it satisfies the given requirements. Testing is running a system in order to identify any gaps, errors, bugs or missing requirements with respect to actual requirements.

7.2 Evaluation for Front End

Front end design is manually tested with center alignment and the with a background image. There should be two drop down menus. One is to select location and the other is to select Hotel features.

Four locations available in drop down list which are Hikkaduwa, Kandy, Nuwara Elliya, and Colombo.

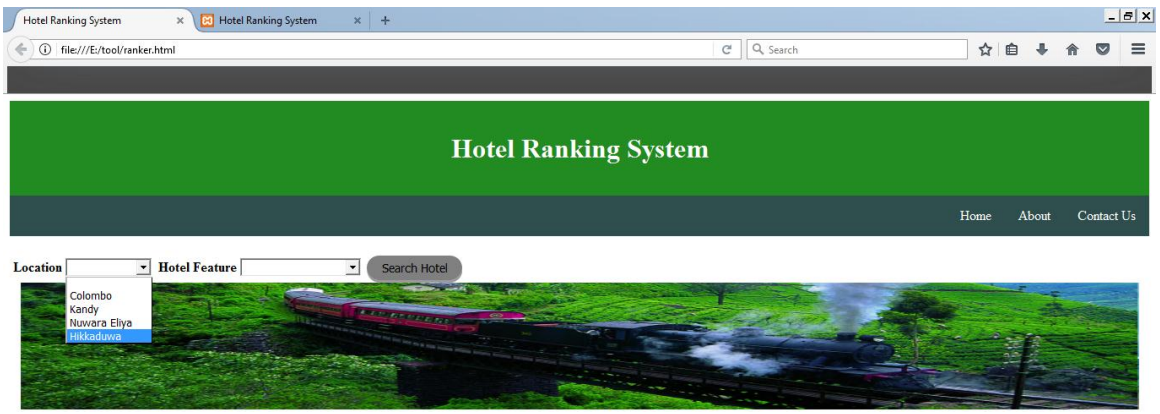


Figure 10:Front End- list of location

There are 15 variables in hotel features drop down list from the front end interface

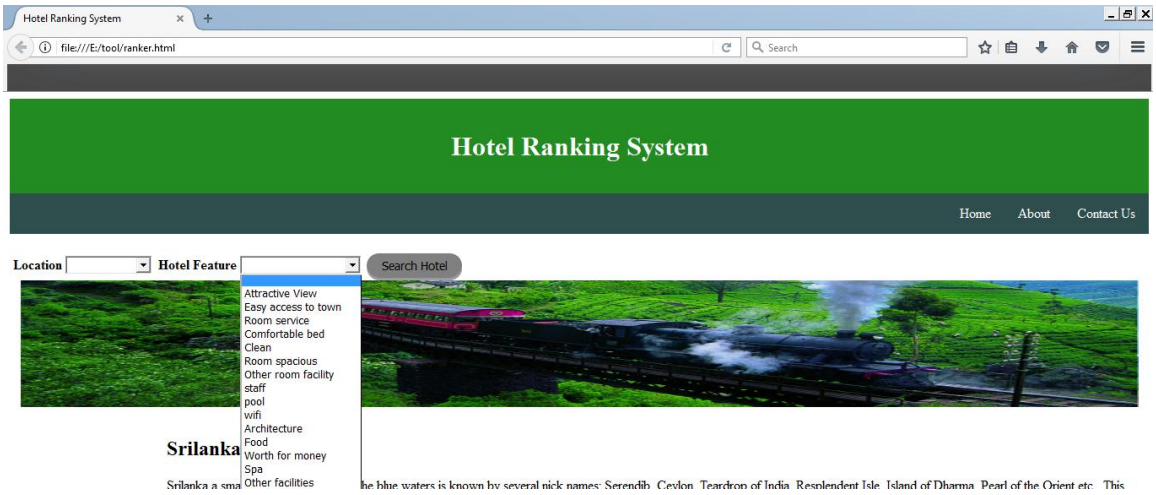


Figure 11:Front End-Hotel Feature List

7.3 Evaluation for Opinion Sentiment classifier - NLTK 2.0.4

To evaluate the effectiveness of the sentiment classifier, the orientation associated with each feature in a sentence was analyzed manually in order to achieve a high degree of confidence.

A correctly classified opinion is either a negative or positive opinion for a given feature, which was correctly identified by the system.

Test script is showed in Appendix B

Example 1:

Review ID-122

1. Access to Galle road
2. Staffs are efficient
3. Breakfast were extraordinary

Sentence 1 is difficult to say whether negative or positive. In sentence 2 “Staffs are efficient” also very heard to say positive or negative. But sentence3 “Breakfast were extraordinary” being analyzed, which definitely positive behavior.

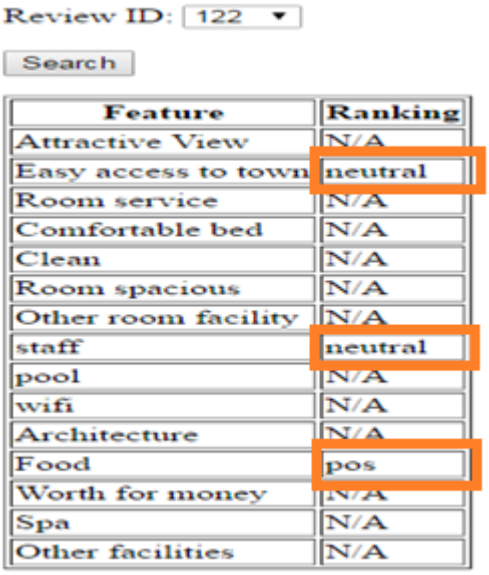


Figure 12:-Result of test script ID-122

7.4 Evaluate for Hotel Ranking

Two evaluation tests for hotel ranking and display in descending order hotel rank

```
/* Get the positive rating for each hotel */
        $finalResult = array ();
        foreach($results as $hotel => $reviews)
        {
                /* Total number of reviews */
                $totalReviews = count($reviews);
                $posReviews = 0;
                /* Get the number of positive reviews */
                foreach($reviews as $review)
                {
                        if($review=='pos')
                        {
                                $posReviews++;
                        }
                }
                /* Calculate the positive rating */
                $finalResult[$hotel] = $posReviews*100/$totalReviews;
```

Get the positive counts per hotel; this count is divided by total number of reviews; result is multiply by 100 to make it percentage.

Example: Avenra Beach Hotel

Hotel ID: 4

Hotel feature: Attractive view

Manually counted the number of reviews (no of people commented about 'Attractive view' feature) = 20 reviews

From data set-2 found how many positive reviews on the same feature=4

$$\text{Hotel Rank} = \frac{\text{No. of Positive Reviews}}{\text{Total No. of Reviews}} * 100$$

Hotel Rank = 4/20*100

= 20%

Hotel Ranking System

Location: Hikkaduwa

Feature: Attractive View

Search

Hotel	Ranking	Association
Hikkaduwa Beach Hotel	75%	
Hotel Ocean View Cottage	70%	
kalla bongo lake resort	68%	
Lavanga Resort _ Spa	65%	
Coralrock	58.823529411765%	
Hotel lanka super coral	58.823529411765%	
Sapphire seas	54.545454545455%	
Main Reef Hotel	53.571428571429%	
Coral sand	50%	
Citrus	47.368421052632%	
Sunbeach hotel	40.740740740741%	
Drifters Hotel	38.461538461538%	staff
Hikka Tranz by Cinnamon	32%	
Adithya	31.25%	
Avenra Beach Hotel	20%	

Figure 13:Hotel Rank based on selected feature

Evaluation of Hotel Ranking Based on Real User Reviews:

From TripAdvisor website user reviews base on selected hotel features.

- Earl's Regency : Kandy→ Swimming pool(Hotel Feature)

TripAdvisor data:

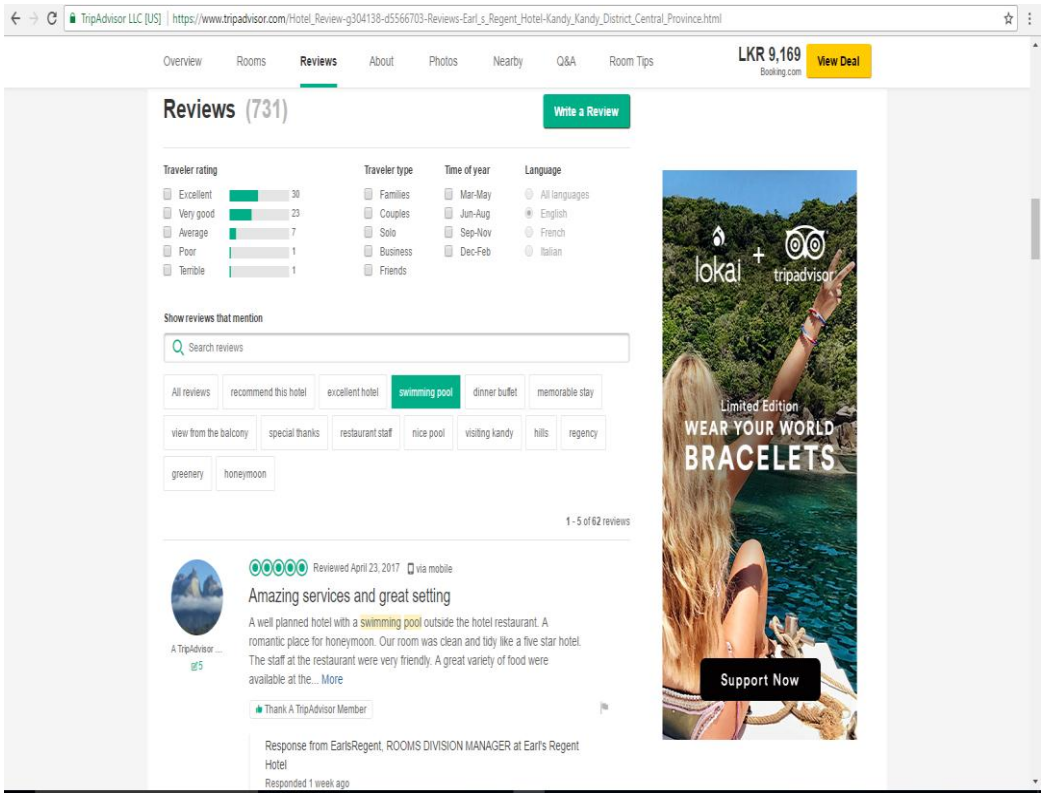


Figure 14:Kandy Earl's Regency hotel's TripAdvisor user review screen

Total no of reviews: 62(30+23+7+1+1)

Positive reviews: 53(30+23)

Ranking: 53*100/62=85.48%

Hotel Ranking System's Result

Figure 15:Kandy Earl's Regency hotel's HRS in Ranking

- Hilton: Colombo→ Food (Hotel Feature)

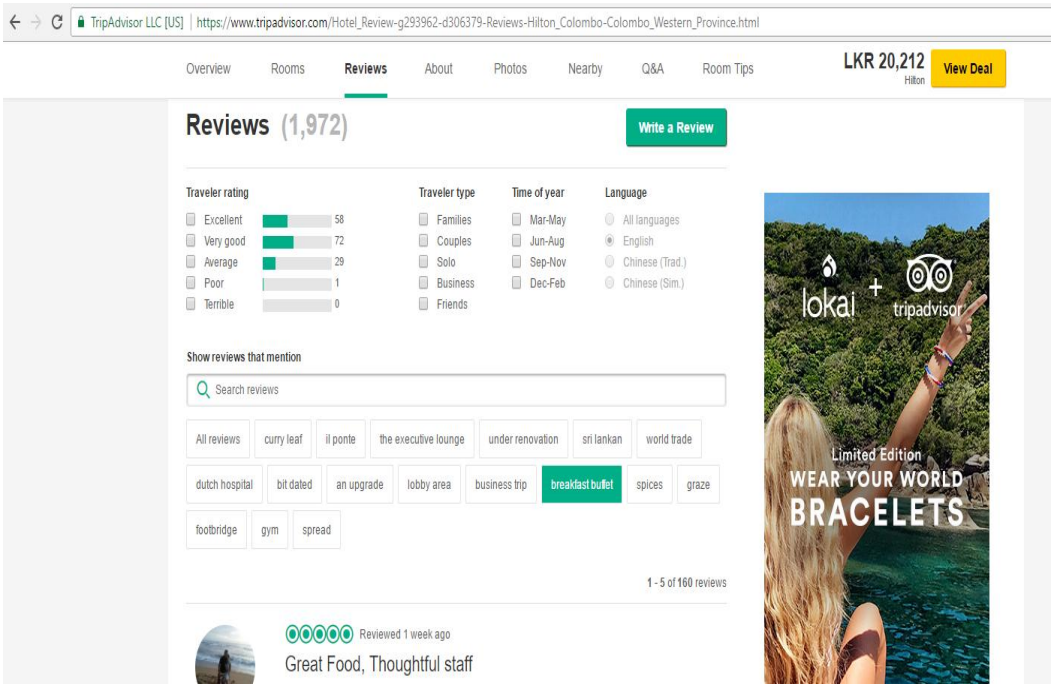


Figure 16: Colombo Hilton Hotels TripAdvisor user review screen

Total no of reviews: 170(68+72+29+1+0)

Positive reviews: 140(68+72)

Ranking: 140*100/170=82.35%

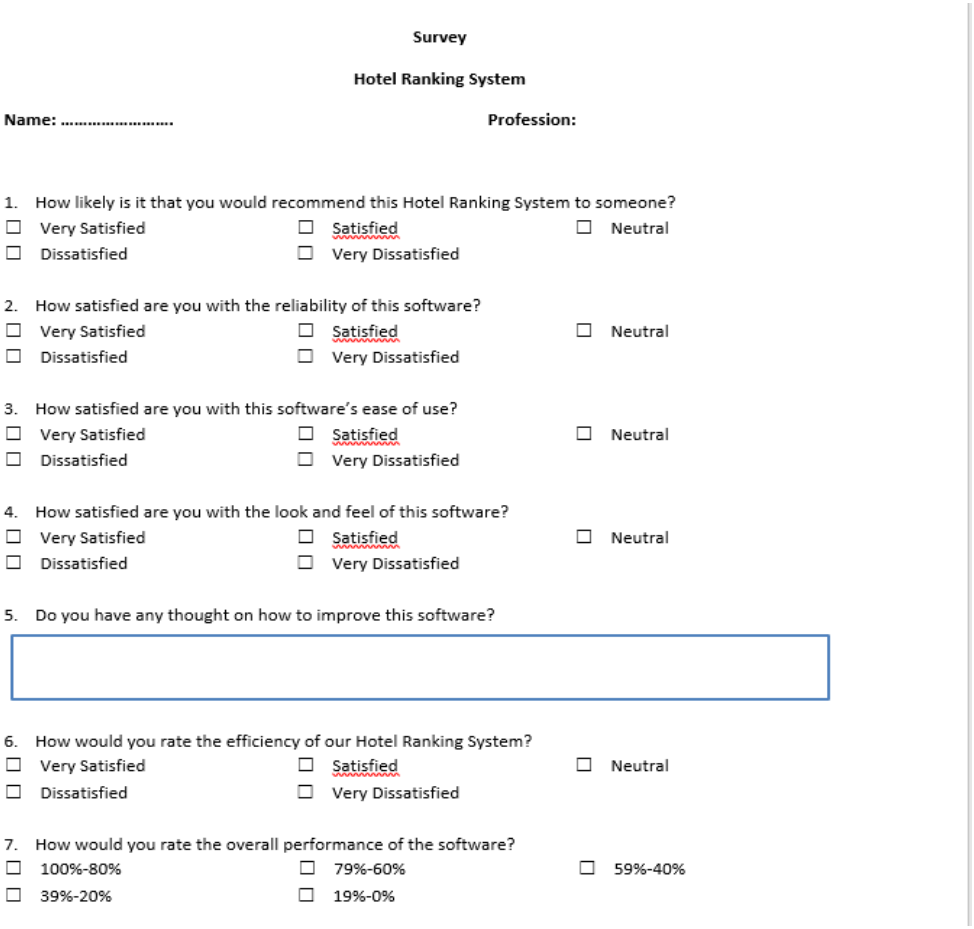
Hotel Ranking System's Result

Hotel	Ranking	Association
Ramada	83.333333333333%	
Fair view	80%	Room service
Chapelton House	80%	staff
Hilton	78.5714285714 29%	
Galadari	76.470588235294%	Room service
Country yard	75%	

Figure 17:Colombo Hilton hotels HRS in Ranking

Evaluation of Overall performance of Hotel Ranking System:

To evaluate the proposed system a survey produced for 8-10 people.



Survey

Hotel Ranking System

Name: …….. **Profession:**

1. How likely is it that you would recommend this Hotel Ranking System to someone?
   ☐ Very Satisfied ☐ Satisfied ☐ Neutral
   ☐ Dissatisfied ☐ Very Dissatisfied
2. How satisfied are you with the reliability of this software?
   ☐ Very Satisfied ☐ Satisfied ☐ Neutral
   ☐ Dissatisfied ☐ Very Dissatisfied
3. How satisfied are you with this software's ease of use?
   ☐ Very Satisfied ☐ Satisfied ☐ Neutral
   ☐ Dissatisfied ☐ Very Dissatisfied
4. How satisfied are you with the look and feel of this software?
   ☐ Very Satisfied ☐ Satisfied ☐ Neutral
   ☐ Dissatisfied ☐ Very Dissatisfied
5. Do you have any thought on how to improve this software?
6. How would you rate the efficiency of our Hotel Ranking System?
   ☐ Very Satisfied ☐ Satisfied ☐ Neutral
   ☐ Dissatisfied ☐ Very Dissatisfied
7. How would you rate the overall performance of the software?
   ☐ 100%-80% ☐ 79%-60% ☐ 59%-40%
   ☐ 39%-20% ☐ 19%-0%

Figure 18:Survey Sample

Collected Results based on user's feedback (10 users)

User	Performance
1	59-40
2	59-40
3	59-40
4	79-60
5	79-60
6	79-60
7	100-80
8	59-40
9	100-80
10	79-60

Table 1: User Feedback

Performance	Count of User
100-80	2
59-40	4
79-60	4

Table 2: Summary of user's feedback

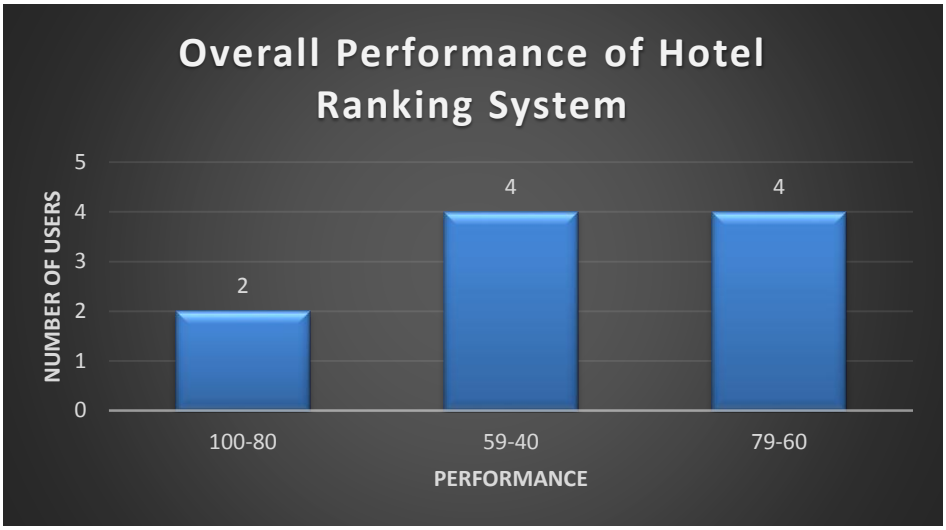


Figure 19:HRS's Overall performance chart

7.5 Evaluation for Associate Mining

To evaluate Associate mining using Apriori algorithm will find frequent hotel feature list. Based on the result of frequent feature list, generate association rule.

Results will be based on selected hotel feature what are other features people have commented positively.

Code is showed in Appendix B

Ariori Algorithm work of Hotel Citrus

Example: Citrus hotel reviews

Review ID	List of Features
1	Other facilities
2	Room spacious, Other room facility
3	Food
4	Attractive View, staff, Food
5	Room spacious, Other room facility, staff
6	Other facilities
7	Attractive View, staff, pool
8	Comfortable bed, Room spacious, Other room facility, Other facilities
9	Attractive View, Room spacious, Other room facility
10	Attractive View, Room spacious, Other room facility, Food
11	Attractive View, Architecture
12	Other room facility
13	Staff
14	staff, Food
15	Room service, staff
16	Food, Other facilities
17	pool, Food, Worth for money
18	Attractive View, staff
19	Attractive View, Room service, Food
20	Attractive View, staff
21	Easy access to town, staff, Worth for money
22	Easy access to town
23	Attractive View, staff
24	Room spacious, Other room facility

25	Food
26	staff, Food
27	Room spacious, Other room facility, staff, Food, Other facilities
28	Other facilities
29	Other facilities

Table 2: Transaction Table

Step-1 Generating 1-itemset frequent pattern

Scan data for each candidate

Feature set	Sup. Count
Attractive View	9
Easy access to town	2
Room service	2
~~Comfortable bed~~	~~1~~
~~Clean~~	~~0~~
Room spacious	7
Other room facility	8
Staff	12
~~Pool~~	~~0~~
~~Wi-Fi~~	~~0~~
~~Architecture~~	~~1~~
Food	10
Worth for money	2
~~Spa~~	~~0~~
Other facilities	7

Table 3: Table C1

L1

Feature set	Sup. Count
Attractive View	9
Easy access to town	2
Room service	2
Room spacious	7
Other room facility	8
Staff	12
Food	10
Worth for money	2
Other facilities	7

Table 4: Table L1

Step-2 Generating 2-item set frequent pattern

~~Attractive View, Easy access to town~~	~~0~~
~~Attractive View, Room service~~	~~1~~
Attractive View, Room spacious	2
Attractive View, Other room facility	2
Attractive View, staff	5
Attractive View, Food	3
~~Attractive View, Worth for money~~	~~0~~
Attractive View, Other facilities	2
~~Easy access to town, Room service~~	~~0~~
~~Easy access to town ,Room spacious~~	~~0~~
~~Easy access to town, Other room facility~~	~~0~~
~~Easy access to town , staff~~	~~1~~
~~Easy access to town, Food~~	~~0~~
~~Easy access to town, Worth for money~~	~~1~~
~~Easy access to town, Other facilities~~	~~0~~
~~Room service, Room spacious~~	~~0~~
~~Room service, Other room facility~~	~~0~~

~~Room service, staff~~	~~1~~
~~Room service, Food~~	~~1~~
~~Room service, Worth for money~~	~~0~~
~~Room service, Other facilities~~	~~0~~
Room spacious, Other room facility	7
Room spacious, staff	2
Room spacious, Food	2
~~Room spacious, Worth for money~~	~~0~~
Room spacious, Other facilities	2
Other room facility, staff	2
Other room facility, Food	2
~~Other room facility, Worth for money~~	~~0~~
Other room facility, Other facilities	2
staff, Food	4
~~staff, Worth for money~~	~~1~~
~~staff, Other facilities~~	~~1~~
~~Food, Worth for money~~	~~1~~
Food, Other facilities	2
~~Worth for money, Other facilities~~	~~0~~

Table 5: Table C2

L2

Attractive View, Room spacious	2
Attractive View, Other room facility	2
Attractive View, staff	5
Attractive View, Food	3
Attractive View, Other facilities	2
Room spacious, Other room facility	7
Room spacious, staff	2
Room spacious, Food	2

Room spacious, Other facilities	2
Other room facility, staff	2
Other room facility, Food	2
Other room facility, Other facilities	2
staff, Food	4
Food, Other facilities	2

Table 6: Table L2

Step-3 Generating 3-itemset frequent pattern

C3=L2 join L2

Example: -

{Attractive View, Room spacious, Other room facility}, sub set {Attractive View, Room spacious}, {Attractive View, Other room facility} and {Room spacious, Other room facility}

This is allowed according to Apriori pruning

{Attractive View, Room spacious, staff} sub set {Attractive View, Room spacious}, {Attractive View, staff} and {Room spacious, staff}-This allowed

{Attractive View, Room spacious, Food} sub set {Attractive View, Room spacious}, {Attractive View, Food} and {Room spacious, Food}-This is allowed

{Attractive View, Room spacious, Other facilities} sub set {Attractive View, Room spacious}, {Room spacious, Other facilities}and {Attractive View, Other facilities}-This is allowed

{Room spacious, Other room facility, staff} sub set {Room spacious, Other room facility}, {Room spacious, staff} and {Other room facility, staff}- This is allowed

{Room spacious, Other room facility, Food} sub set {Room spacious, Other room facility}, {Room spacious, Food} and {Other room facility, Food}-This is allowed

{Room spacious, Other room facility, Other facilities} sub set {Room spacious, Other room facility}, {Room spacious, Other facilities} and {Other room facility, Other facilities}-this is allowed

{Other room facility, staff, Food} sub set {Other room facility, staff}, {staff, Food} and {Other room facility, Food}-This is allowed etc.

{staff, other facilities, Attractive view} sub set {staff, other facilities}-not a sub set of L_2, {other facilities, Attractive view} and {staff, Attractive view}- Not allowed

L3

Attractive View, Room spacious, Other room facility	2
~~Attractive View, Room spacious, staff~~	~~0~~
~~Attractive View, Room spacious, Food~~	~~1~~
~~Attractive View, Room spacious, Other facilities~~	~~0~~
~~Room spacious, Other room facility, staff~~	~~1~~
Room spacious, Other room facility, food	2
Room spacious, Other room facility, Other facilities	2
~~Room spacious, staff, food~~	~~1~~
~~Room spacious, food, Other facilities~~	~~1~~
~~Other room facility, staff, Attractive view~~	~~0~~
~~Other room facility, staff, food~~	~~1~~
~~Other room facility, food, Attractive view~~	~~1~~
~~Other room facility, food, Other facility~~	~~1~~
~~Other room facility, Other facility, Attractive view~~	~~0~~
~~Other room facility, Other facility, staff~~	~~1~~
~~Food, staff, Attractive view~~	~~1~~
~~Food, Other facilities, Attractive view~~	~~0~~

Table 7: Table C3

Attractive View, Room spacious, Other room facility	2
Room spacious, Other room facility, food	2
Room spacious, Other room facility, Other facilities	2

Table 8: Table L3

L3 join L3

By Pruned frequent item set is

{Attractive View, Room spacious, other room facility}

Attractive View, Room spacious, Other room facility, food	1

Table 9: Table C4

Associations

R1: Attractive View ^ Room spacious → other room facility

- Confidence= {Attractive View, Room spacious, other room facility}/{ Attractive View, Room spacious} => 2/2=>100%
- R1 is selected

R2: Attractive View ^ other room facility → Room spacious

- Confident = {Attractive View, Room spacious, other room facility}/{ Attractive View ^ other room facility}=> 2/2=> 100%
- R2 is selected

R3: other room facility ^ Room spacious→ Attractive View

- Confident = {Attractive View, Room spacious, other room facility}/{ other room facility ^ Room spacious}=>2/7=>28%
- R3 is rejected

R4: Attractive View→ Room spacious ^ other room facility

- Confidence = {Attractive View, Room spacious, other room facility}/{ Attractive View}=> 2/8 => 25%
- R4 is rejected

R5: Room spacious→ Attractive View ^ other room facility

- Confidence= {Attractive View, Room spacious, other room facility}/{ Room spacious }=>2/7=>28%
- R5 is rejected

R6: other room facility→ Room spacious ^ Attractive View

- Confidence=={Attractive View, Room spacious, other room facility}/{ other room facility }=>2/8=>25%
- R6 is rejected

  In this way found two strong association

Using test script:

Hotel ID: Citrus

Search

Frequent Itemsets

```
Time: 0 second(s)
================================================================================
{Food,Attractive View} = 3
{Food,staff} = 4
{Attractive View,staff} = 5
{Room spacious,Other room facility,Food} = 2
{Room spacious,Other room facility,Attractive View} = 2
{Room spacious,Other room facility,staff} = 2
```

Frequent Itemsets Array

```
Array ( [0] => Array ( [sup] => 3 [0] => Food [1] => Attractive View ) [1] => Array ( [sup] => 4 [0] => Food [1] => staff ) [2] => Array ( [sup] => 5 [0] => Attractive View [1] => staff
) [3] => Array ( [sup] => 2 [0] => Room spacious [1] => Other room facility [2] => Food ) [4] => Array ( [sup] => 2 [0] => Room spacious [1] => Other room facility [2] => Attractive View
) [5] => Array ( [sup] => 2 [0] => Room spacious [1] => Other room facility [2] => staff ) )
```

Figure 20: Apriori Test

Association Rules

```
Time: 0 second(s)
================================================================================
Attractive View => Food = 33.33%
Attractive View => staff = 55.56%
Attractive View => Other room facility = 22.22%
Attractive View => Room spacious = 22.22%
Attractive View => Room spacious,Other room facility = 22.22%
Food => Attractive View = 30%
Food => staff = 40%
Food => Other room facility = 20%
Food => Room spacious = 20%
Food => Room spacious,Other room facility = 20%
staff => Food = 33.33%
staff => Attractive View = 41.67%
staff => Other room facility = 16.67%
staff => Room spacious = 16.67%
staff => Room spacious,Other room facility = 16.67%
Other room facility => Food = 25%
Other room facility => Room spacious = 87.5%
Other room facility => Room spacious,Food = 25%
Other room facility => Attractive View = 25%
Other room facility => Room spacious,Attractive View = 25%
Other room facility => staff = 25%
Other room facility => Room spacious,staff = 25%
Room spacious => Food = 28.57%
Room spacious => Other room facility = 100%
Room spacious => Other room facility,Food = 28.57%
Room spacious => Attractive View = 28.57%
Room spacious => Other room facility,Attractive View = 28.57%
Room spacious => staff = 28.57%
Room spacious => Other room facility,staff = 28.57%
Room spacious,Other room facility => Food = 28.57%
Room spacious,Other room facility => Attractive View = 28.57%
Room spacious,Other room facility => staff = 28.57%
Room spacious,Food => Other room facility = 100%
Other room facility,Food => Room spacious = 100%
Room spacious,Attractive View => Other room facility = 100%
Other room facility,Attractive View => Room spacious = 100%
Room spacious,staff => Other room facility = 100%
```

Figure 21: Association Test

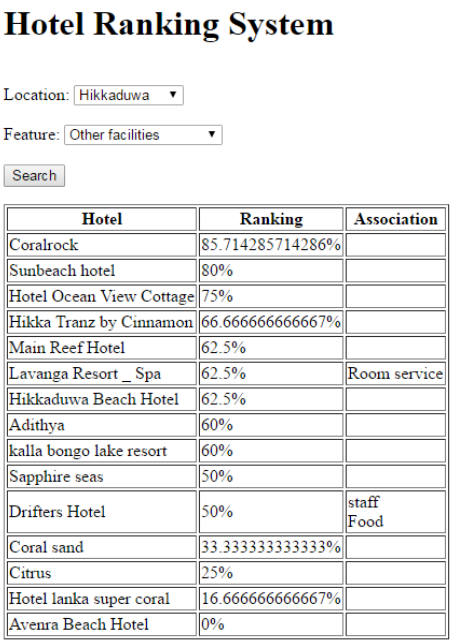


Figure 22: Test Result of Association

7.6 Summary

This chapter evaluated the mythologies and the result discussed in the implementation chapter. Next chapter will summarize overall research, some limitation and future improvements for the proposed solution.

Chapter 8

Conclusion

8.1 Introduction

This chapter contains a discussion on the result obtain from the research and how the objectives are met. The future works can be done to continue this research for make the broader environment also discussed.

8.2 Achievement of Objective

Opinion mining has become a fascinating research area due to the availability of a huge volume of user-generated content in web, forums and blogs. Opinion mining has applications in a variety of fields ranging from market research to decision making to advertising. With the help of opinion mining, companies can estimate the extent of product acceptance and can devise strategies to improve their product. Individuals can also use opinion mining tools to make decisions on their buying by comparing competitive products not just based on specifications but also based on user experience and public opinions.

In this thesis we have shown how sentiment analysis performed using software tools and we have seen how normalized polarity varies for the values taken from customer reviews.

This thesis shown semantic analysis to identify the hotel features, sentiment analysis process to identify the given review is positive, negative or neutral. And Apriori algorithm and association rules used to find frequent data item to conclude with selected hotel feature, what are other features that customers seeking for positive behavior. Using this information, hoteliers can improve their booking process.

8.3 Further work

Here, we see a few but important challenges for text analytics tasks like opinion mining a good direction would be methods which have global knowledge of opinions dependent on complex contexts, which can use this information later to help solving context problems in any local analysis. And automated feature extraction from unstructured reviews is also very important challenge in semantic analysis. Most of the times we see many spam blogs and spam reviews posted by the users. If we consider these reviews for performing Opinion Mining, we may get deviated from our desired results. So, lot of work has to be done in this field for identifying spam blogs, considering spelling mistakes and for other challenges.

8.4 Summary

This chapter presents the conclusion and further work of this research. And also it contains explanation on problem encountered during implementation and further work to be done.

References

1. https://www.dfki.de/web/research/publications/renameFileForDownload?...25...
2. MK Dalal, MA Zaveri - Applied computational intelligence and soft …, 2014 - dl.acm.org
3. http://www.aclweb.org/anthology/W10-3209
4. http://www.sciencedirect.com/science/article/pii/S1877050913008879
5. http://dl.acm.org/citation.cfm?id=2487944
6. http://www.ijsr.net/archive/v3i12/U1VCMTQxMDI2.pdf
7. https://www.ijarcsse.com/docs/papers/Volume_5/11_November2015/V5I11-0349.pdf
8. www.Tripadvisor.com
9. https://www.apachefriends.org/download.html
10. https://www.**webharvy**.com
11. https://en.wikipedia.org/
12. https://www.**mysql**.com/about/
13. http://www.nltk.org/
14. https://www.cs.helsinki.fi/u/myllymak/Teaching/2004/Fall/Seminar/poroshin.pdf.
15. http://serialsjournals.com/serialjournalmanager/pdf/1345547827.pdf
16. http://www.webopedia.com/TERM/W/World_Wide_Web.html
17. http://cse.iitkgp.ac.in/~bivasm/uc_notes/07apriori.pdf
18. https://en.wikipedia.org/wiki/Apriori_algorithm
19. http://www.vskills.in/certification/tutorial/data-mining-and-warehousing/association-rules-mining/
20. https://www.slideshare.net/TuyenVuongDinh/mysql-architectures
21. https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/
22. https://github.com/VTwo-Group/Apriori-Algorithm

Appendix A

Structure of Hotels Table in Database

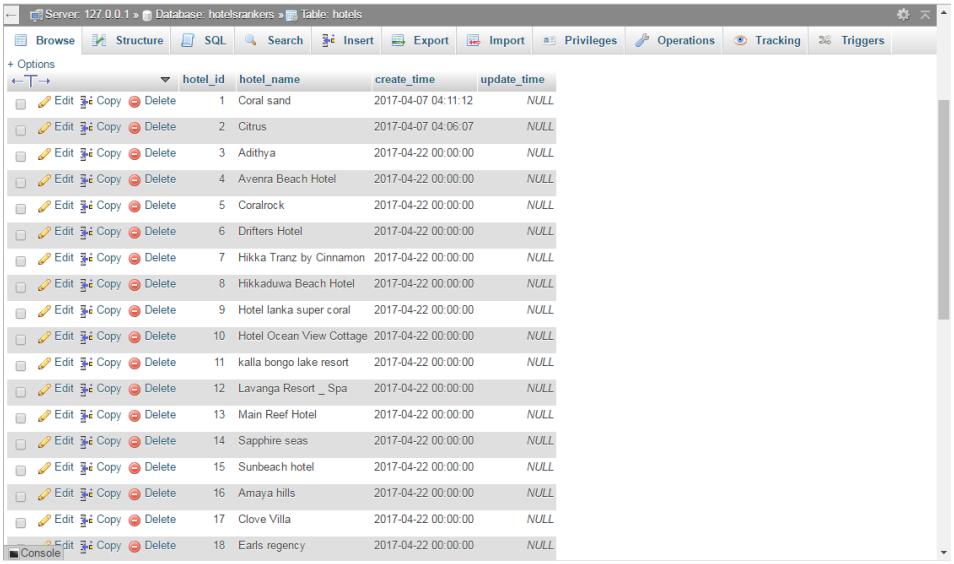


Figure 23: Structure of Hotels Table in Database

Structure of Location Table in Database

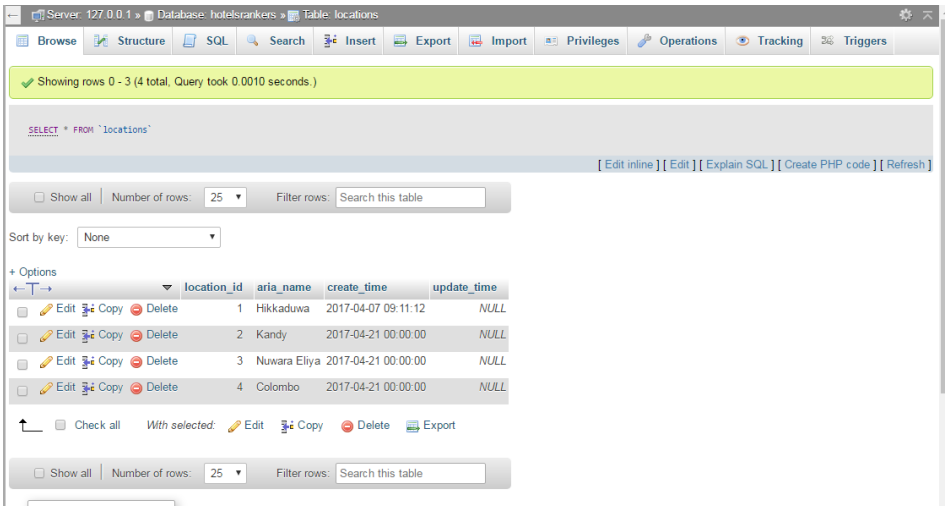
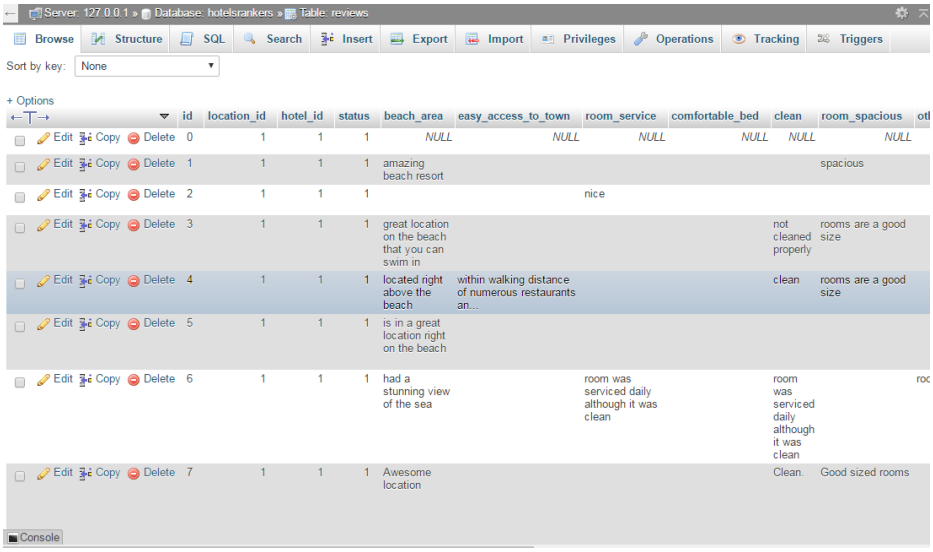


Figure 24: Structure of Location Table in Database

Structure of Reviews Table in Database



	id	location_id	hotel_id	status	beach_area	easy_access_to_town	room_service	comfortable_bed	clean	room_spacious	otl
Edit Copy Delete	0	1	1	1	NULL	NULL	NULL	NULL	NULL	NULL	
Edit Copy Delete	1	1	1	1	amazing beach resort					spacious	
Edit Copy Delete	2	1	1	1			nice				
Edit Copy Delete	3	1	1	1	great location on the beach that you can swim in				not cleaned properly	rooms are a good size	
Edit Copy Delete	4	1	1	1	located right above the beach	within walking distance of numerous restaurants an...			clean	rooms are a good size	
Edit Copy Delete	5	1	1	1	is in a great location right on the beach						
Edit Copy Delete	6	1	1	1	had a stunning view of the sea		room was serviced daily although it was clean		room was serviced daily although it was clean		roc
Edit Copy Delete	7	1	1	1	Awesome location				Clean.	Good sized rooms	

Console

Figure 25: Structure of Reviews Table in Database

After sentiment Analysis filtered only positive hotel features

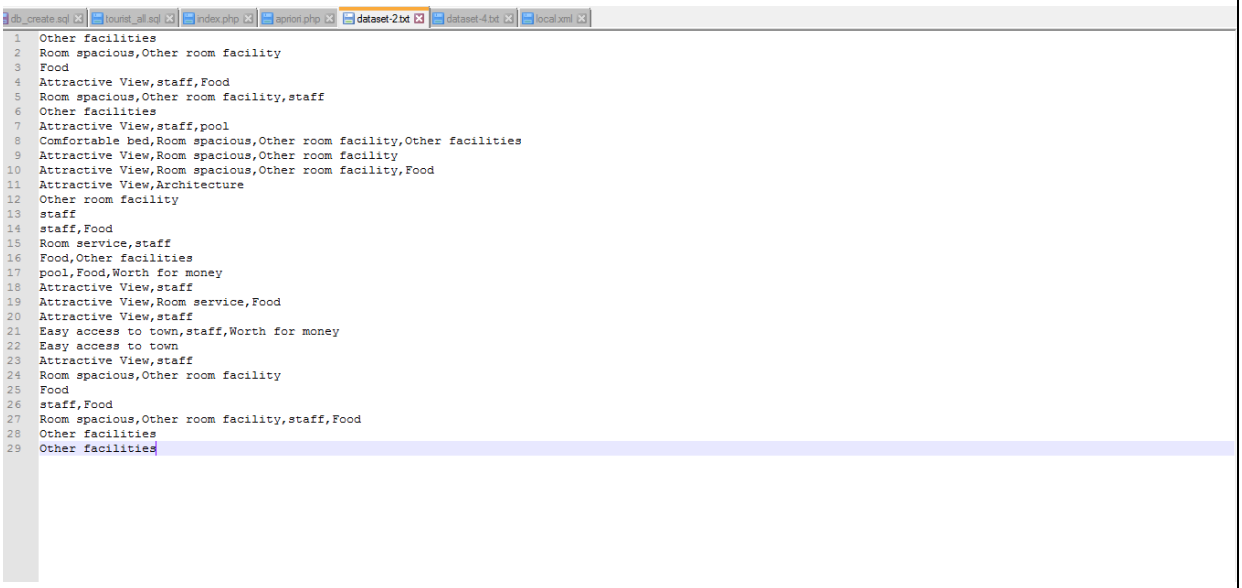


Figure 26: Structure of Reviews Table in Database

Appendix B

Cording of Data base creation

```
CREATE TABLE IF NOT EXISTS `hotels` (
 `hotel_id` INT,
  `hotel_name` varchar(200),
`create_time` datetime DEFAULT NULL,
 `update_time` datetime DEFAULT NULL,
  PRIMARY KEY (`hotel_id`)
) DEFAULT CHARSET=utf8;

CREATE TABLE IF NOT EXISTS `locations` (
  `location_id` INT,
`aria_name` varchar(200),
`create_time` datetime DEFAULT NULL,
  `update_time` datetime DEFAULT NULL,
 PRIMARY KEY (`location_id`)
) DEFAULT CHARSET=utf8;


CREATE TABLE `reviews` (
  `id` bigint(20) NOT NULL,
 `location_id` int(11) DEFAULT NULL,
  `hotel_id` int(11) DEFAULT NULL,
  `status` int(11) DEFAULT '1',
  `beach_area` varchar(200) DEFAULT NULL,
 `easy_access_to_town` varchar(200) DEFAULT NULL,
 `room_service` varchar(200) DEFAULT NULL,
 `comfortable_bed` varchar(200) DEFAULT NULL,
 `clean` varchar(200) DEFAULT NULL,
  `room_spacious` varchar(200) DEFAULT NULL,
  `other_room_facility` varchar(200) DEFAULT NULL,
 `staff` varchar(200) DEFAULT NULL,
 `pool` varchar(200) DEFAULT NULL,
  `wifi` varchar(200) DEFAULT NULL,
  `architecture` varchar(200) DEFAULT NULL,

 `food` varchar(200) DEFAULT NULL,

`worth_for_money` varchar(200) DEFAULT NULL,

`spa` varchar(200) DEFAULT NULL,

 PRIMARY KEY (`id`),

  FOREIGN KEY (location_id) REFERENCES locations(location_id),

 FOREIGN KEY (hotel_id) REFERENCES hotels(hotel_id)

) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

Test script of efficiency of Sentiment Analysis

```
<?php

include_once "dbcon.php";
include 'class.apriori.php';
/* Get all locations */
$sql = "SELECT DISTINCT aria_id, aria_name, hotel_id, hotel_name FROM reviews";
$result = $conn->query($sql);
$locations = array();
$hotels = array();
while($row = $result->fetch_assoc())
{
	$hotels[$row['hotel_name']] = $row['hotel_id'];
	$locations[$row['aria_id']] = $row;
}
/* Get all features */
$features = array(
			"attractive_view" => "Attractive View",
			"easy_access_to_town" => "Easy access to town",
			"room_service" => "Room service",
			"comfortable_bed" => "Comfortable bed",
			"clean" => "Clean",
			"room_spacious" => "Room spacious",
			"other_room_facility" => "Other room facility",
			"staff" => "staff",
			"pool" => "pool",
			"wifi" => "wifi",
			"architecture" => "Architecture",
			"food" => "Food",
			"worth_for_money" => "Worth for money",
			"spa" => "Spa",
			"other_facilities" => "Other facilities"
	);

/* Print the page header */
print "<h1>Hotel Ranking System</h1><br/>";
/* Print the form */
print '<form action="index.php" method="post">';

/* Print Locations */
print '<lable for="aria">Location: </lable>';
print '<select name="aria">';
/* Foreach location */
foreach($locations as $location)
{
	$seleted = '';
	/* Check if submit was click before and the aria_id is set */
	if(isset($_POST['aria']) && $_POST['aria'] == $location['aria_id'])
	{
		$seleted = 'selected';
	}
	/* Ingore values where 'aria_name' is not set */
	if(!$location['aria_name'])
	{
		continue;
	}
	print '<option value="'.$location['aria_id'].'" '.$seleted.'> '.$location['aria_name'].'</option>';
}
print '</select>';
print '<br/><br/>';

/* Print Feature */
print '<lable for="feature">Feature: </lable>';
print '<select name="feature">';
/* Foreach feature */
foreach($features as $key => $value)
```

```
{
		$seleted = '';
		/* Check if submit was click before and the aria_id is set */
		if(isset($_POST['feature']) && $_POST['feature'] == $key)
		{
			$seleted = 'selected';
		}
		print '<option value="'.$key.'" '.$seleted.'> '.$value.'</option>';
}
print '</select>';

/* Print submit button */
print	'<br/><br/><input type="submit" value="Search">';

print '</form>';

/* check if the submit button was click */
if(isset($_POST['aria']) && isset($_POST['feature']))
{

		$aria_id = $_POST['aria'];
		$feature_col = $_POST['feature'];
		/* Get all reviews for the given hotel */
		$sql = "SELECT DISTINCT hotel_id AS hid, hotel_name AS hname, ".$feature_col." AS feature
FROM reviews WHERE aria_id=".$aria_id;
		$result = $conn->query($sql);
		$results = array();
		/* Foreach Review */
		while($row = $result->fetch_assoc())
		{
			/* Curl request to http://text-processing.com API */
			$ch = curl_init();
			curl_setopt($ch, CURLOPT_URL,"http://text-processing.com/api/sentiment/");
			curl_setopt($ch, CURLOPT_POST, TRUE);
			curl_setopt($ch, CURLOPT_POSTFIELDS, 'text='.$row["feature"]);
			curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
			/* Process Curl response */
			$response = curl_exec ($ch);
			$res = json_decode($response, true);
			/* save the result */
			$results[$row["hname"]][] = $res['label'];
			curl_close ($ch);
		}

		/* Get the positive rating for each hotel */
		$finalResult = array();
		foreach($results as $hotel => $reviews)
		{
			/* Total number of reviews */
			$totalReviews = count($reviews);
			$posReviews = 0;
			/* Get the number of positive reviews */
			foreach($reviews as $review)
			{
				if($review=='pos')
				{
					$posReviews++;
				}
			}
			/* Calculate the positive rating */
			$finalResult[$hotel] = $posReviews*100/$totalReviews;
		}
		/* Sort using the rating */
		arsort($finalResult);
		/* Print the values */
		print '<table>';
		print '<table border=1px><th>Hotel</th><th>Ranking</th><th>Association</th>';
```

```
		foreach($finalResult as $hotel => $data)
		{
			$file = 'dataset/dataset-'.$hotels[$hotel].'.txt';
			$Apriori = new Apriori();

			$Apriori->setMaxScan(20);	//Scan 2, 3, ...
			$Apriori->setMinSup(2);		//Minimum support 1, 2, 3, ...
			$Apriori->setMinConf(60);	//Minimum confidence - Percent 1, 2, ..., 100
			$Apriori->setDelimiter(',');	//Delimiter
			$Apriori->process($file);
			$ruleString = '';
			$rule = $Apriori->getAssociationRules();
			$ruleDatas = array();
			if(isset($rule[$features[$feature_col]]))
			{
				$ruleDatas = $rule[$features[$feature_col]];
			}
			foreach($ruleDatas as $key => $value)
			{
				$ruleString .= $key.'<br/>';
			}
			print '<tr>';
				print '<td>'.$hotel.'</td>';
				print '<td>'.$data.'%</td>';
				print '<td>'.$ruleString.'</td>';
			print '</tr>';
		}
		print '</table>';
	}
```

Apriori Test script

```
<!DOCTYPE HTML>
<html>
<head>
   <meta http-equiv="content-type" content="text/html; charset=utf-8" />
        <title>Apriori Alghoritm</title>
</head>
<body style="font-family: monospace;">
<?php
include 'class.apriori.php';
include_once "dbcon.php";
$rows = array();
$sql = "SELECT * FROM reviews";
$result = $conn->query($sql);
while($row = $result->fetch_assoc())
{
        $rows[$row['hotel_id']] = $row;
}
print '<form action="apriori.php" method="post">';
print '<lable for="hotel_id">Hotel ID: </lable>';
print '<select name="hotel_id">';
foreach($rows as $row)
{
        $seleted = '';
        if(isset($_POST['hotel_id']) && $_POST['hotel_id'] == $row['hotel_id'])
        {
                $seleted = 'selected';
        }
        print '<option value="'.$row['hotel_id'].'" '.$seleted.>
'.$row['hotel_name'].'</option>';
}
print '</select>';
print '<br/><br/>';
print    '<input type="submit" name="submit" value="Search">';
print '</form>';
if(isset($_POST['submit']))
{
        $file = 'dataset/dataset-'.$_POST['hotel_id'].'.txt';
        $Apriori = new Apriori();

        $Apriori->setMaxScan(20);      //Scan 2, 3, ...
        $Apriori->setMinSup(2);        //Minimum support 1, 2, 3, ...
        $Apriori->setMinConf(1);       //Minimum confidence - Percent 1, 2, ..., 100
        $Apriori->setDelimiter(',');   //Delimiter

        /*
        Use Array:
        $dataset   = array();
        $dataset[] = array('A', 'B', 'C', 'D');
        $dataset[] = array('A', 'D', 'C');
        $dataset[] = array('B', 'C');
```

```
        $dataset[] = array('A', 'E', 'C');
        $Apriori->process($dataset);
        */
        $Apriori->process($file);

        //Frequent Itemsets
        echo '<h1>Frequent Itemsets</h1>';
        $Apriori->printFreqItemsets();

        echo '<h3>Frequent Itemsets Array</h3>';
        print_r($Apriori->getFreqItemsets());

        //Association Rules
        echo '<h1>Association Rules</h1>';
        $Apriori->printAssociationRules();

        echo '<h3>Association Rules Array</h3>';
        print_r($Apriori->getAssociationRules());

        //Save to file
        //$Apriori->saveFreqItemsets('freqItemsets.txt');
        //$Apriori->saveAssociationRules('associationRules.txt');
}
?>
</body>
</html>
```

Hotel Ranking System/ MyCode PHP

```
<?php

include_once "dbcon.php";
include 'class.apriori.php';
/* Get all locations */
$sql = "SELECT DISTINCT aria_id, aria_name, hotel_id, hotel_name FROM reviews";
$result = $conn->query($sql);
$locations = array();
$hotels = array();
while($row = $result->fetch_assoc())
{
	$hotels[$row['hotel_name']] = $row['hotel_id'];
	$locations[$row['aria_id']] = $row;
}
/* Get all features */
$features = array(
			"attractive_view" => "Attractive View",
			"easy_access_to_town" => "Easy access to town",
			"room_service" => "Room service",
			"comfortable_bed" => "Comfortable bed",
			"clean" => "Clean",
			"room_spacious" => "Room spacious",
			"other_room_facility" => "Other room facility",
			"staff" => "staff",
			"pool" => "pool",
			"wifi" => "wifi",
			"architecture" => "Architecture",
			"food" => "Food",
			"worth_for_money" => "Worth for money",
			"spa" => "Spa",
			"other_facilities" => "Other facilities"
	);

/* Print the page header */
print "<h1>Hotel Ranking System</h1><br/>";
/* Print the form */
print '<form action="index.php" method="post">';

/* Print Locations */
print '<lable for="aria">Location: </lable>';
print '<select name="aria">';
/* Foreach location */
foreach($locations as $location)
{
	$seleted = '';
	/* Check if submit was click before and the aria_id is set */
	if(isset($_POST['aria']) && $_POST['aria'] == $location['aria_id'])
	{
		$seleted = 'selected';
```

```
		}
		/* Ingore values where 'aria_name' is not set */
		if(!$location['aria_name'])
		{
			continue;
		}
		print '<option value="'.$location['aria_id'].'" '.$seleted.>
'.$location['aria_name'].'</option>';
}
print '</select>';
print '<br/><br/>';

/* Print Feature */
print '<lable for="feature">Feature: </lable>';
print '<select name="feature">';
/* Foreach feature */
foreach($features as $key => $value)
{
		$seleted = '';
		/* Check if submit was click before and the aria_id is set */
		if(isset($_POST['feature']) && $_POST['feature'] == $key)
		{
			$seleted = 'selected';
		}
		print '<option value="'.$key.'" '.$seleted.'> '.$value.'</option>';
}
print '</select>';

/* Print submit button */
print	'<br/><br/><input type="submit" value="Search">';

print '</form>';

/* check if the submit button was click */
if(isset($_POST['aria']) && isset($_POST['feature']))
{

		$aria_id = $_POST['aria'];
		$feature_col = $_POST['feature'];
		/* Get all reviews for the given hotel */
		$sql = "SELECT DISTINCT hotel_id AS hid, hotel_name AS hname, ".$feature_col." AS
feature FROM reviews WHERE aria_id=".$aria_id;
		$result = $conn->query($sql);
		$results = array();
		/* Foreach Review */
		while($row = $result->fetch_assoc())
		{
			/* Curl request to http://text-processing.com API */
			$ch = curl_init();
			curl_setopt($ch, CURLOPT_URL,"http://text-
processing.com/api/sentiment/");
```

```
		curl_setopt($ch, CURLOPT_POST, TRUE);
		curl_setopt($ch, CURLOPT_POSTFIELDS, 'text='.$row["feature"]);
		curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
		/* Process Curl response */
		$response = curl_exec ($ch);
		$res = json_decode($response, true);
		/* save the result */
		$results[$row["hname"]][] = $res['label'];
		curl_close ($ch);
}

/* Get the positive rating for each hotel */
$finalResult = array();
foreach($results as $hotel => $reviews)
{
		/* Total number of reviews */
		$totalReviews = count($reviews);
		$posReviews = 0;
		/* Get the number of positive reviews */
		foreach($reviews as $review)
		{
			if($review=='pos')
			{
				$posReviews++;
			}
		}
		/* Calculate the positive rating */
		$finalResult[$hotel] = $posReviews*100/$totalReviews;
}
/* Sort using the rating */
arsort($finalResult);
/* Print the values */
print '<table>';
print '<table border=1px><th>Hotel</th><th>Ranking</th><th>Association</th>';
foreach($finalResult as $hotel => $data)
{
		$file = 'dataset/dataset-'.$hotels[$hotel].'.txt';
		$Apriori = new Apriori();

		$Apriori->setMaxScan(20);      //Scan 2, 3, ...
		$Apriori->setMinSup(2);        //Minimum support 1, 2, 3, ...
		$Apriori->setMinConf(60);      //Minimum confidence - Percent 1, 2, ..., 100
		$Apriori->setDelimiter(',');   //Delimiter
		$Apriori->process($file);
		$ruleString = '';
		$rule = $Apriori->getAssociationRules();
		$ruleDatas = array();
		if(isset($rule[$features[$feature_col]]))
		{
			$ruleDatas = $rule[$features[$feature_col]];
		}
```

```
                foreach($ruleDatas as $key => $value)
                {
                        $ruleString .= $key.'<br/>';
                }
                print '<tr>';
                        print '<td>'.$hotel.'</td>';
                        print '<td>'.$data.'%</td>';
                        print '<td>'.$ruleString.'</td>';
                print '</tr>';
        }
        print '</table>';
}
```

Appendix C

Evaluation

User1:

Survey

Hotel Ranking System

Name: *Dewmi* **Profession:** *Teacher*

1. How likely is it that you would recommend this Hotel Ranking System to someone?
   - ☐ Very Satisfied ☐ Satisfied ☒ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied

2. How satisfied are you with the reliability of this software?
   - ☐ Very Satisfied ☐ Satisfied ☐ Neutral
   - ☒ Dissatisfied ☐ Very Dissatisfied

3. How satisfied are you with this software's ease of use?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied

4. How satisfied are you with the look and feel of this software?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied

5. Do you have any thought on how to improve this software?

   None

6. How would you rate the efficiency of our Hotel Ranking System?
   - ☐ Very Satisfied ☐ Satisfied ☒ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied

7. How would you rate the overall performance of the software?
   - ☐ 100%-80% ☐ 79%-60% ☒ 59%-40%
   - ☐ 39%-20% ☐ 19%-0%

Figure 27: Survey of user 1

User 2:

Survey

Hotel Ranking System

Name: *Archana* **Profession:** *House wife*

1. How likely is it that you would recommend this Hotel Ranking System to someone?
☐ Very Satisfied ☐ Satisfied ☒ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

2. How satisfied are you with the reliability of this software?
☐ Very Satisfied ☐ Satisfied ☒ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

3. How satisfied are you with this software's ease of use?
☒ Very Satisfied ☐ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

4. How satisfied are you with the look and feel of this software?
☒ Very Satisfied ☐ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

5. Do you have any thought on how to improve this software?

Speed should improve and should have some more information for a traveler

6. How would you rate the efficiency of our Hotel Ranking System?
☐ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

7. How would you rate the overall performance of the software?
☐ 100%-80% ☐ 79%-60% ☒ 59%-40%
☐ 39%-20% ☐ 19%-0%

Figure 28: Survey of user 2

User 3:

Survey

Hotel Ranking System

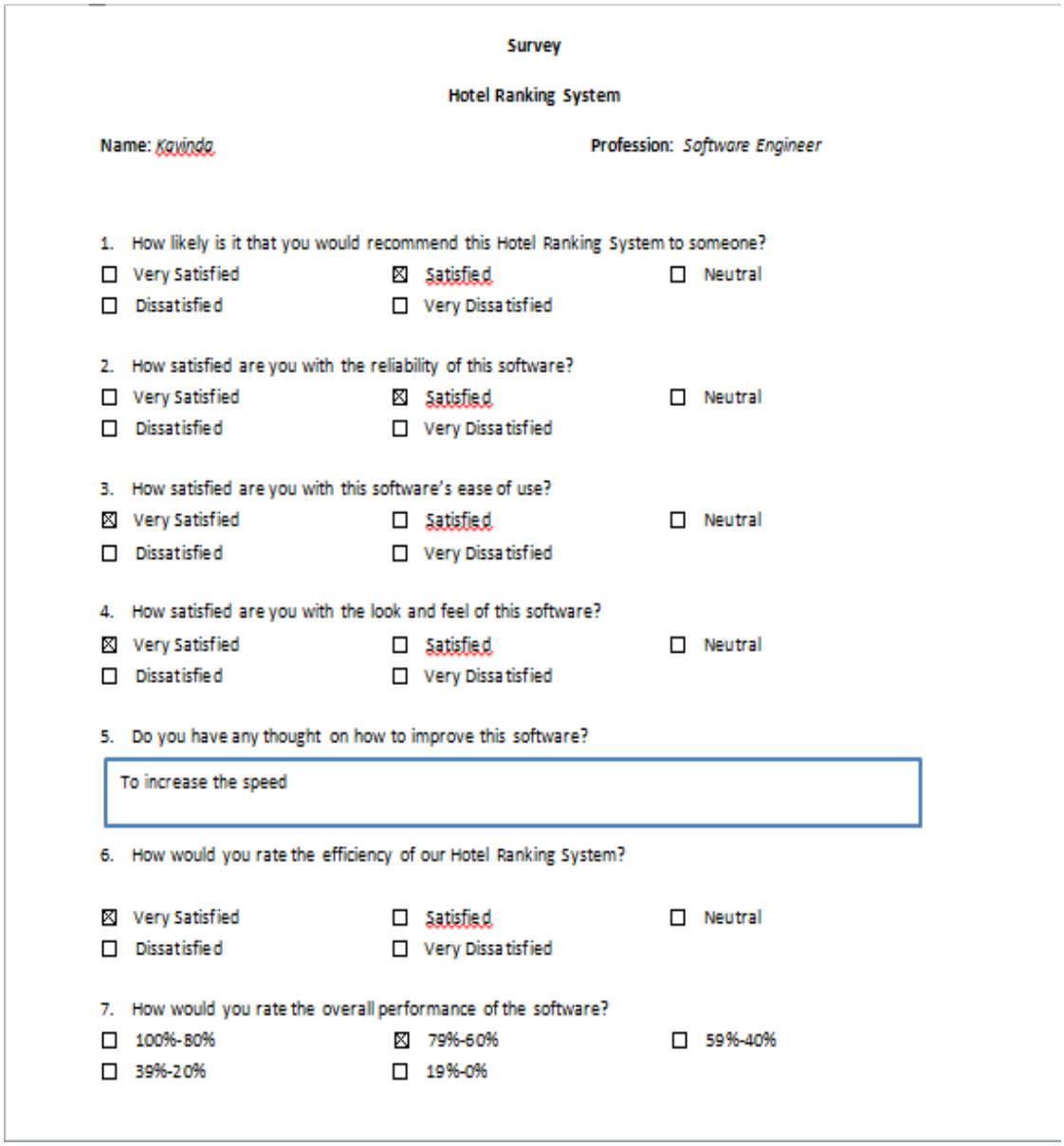
Name: *Gunathilaka* **Profession:** *Associate QA Engineer*

1. How likely is it that you would recommend this Hotel Ranking System to someone?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
2. How satisfied are you with the reliability of this software?
   - ☐ Very Satisfied ☐ Satisfied ☒ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
3. How satisfied are you with this software's ease of use?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
4. How satisfied are you with the look and feel of this software?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
5. Do you have any thought on how to improve this software?

   Speed should improve
6. How would you rate the efficiency of our Hotel Ranking System?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
7. How would you rate the overall performance of the software?
   - ☐ 100%-80% ☐ 79%-60% ☒ 59%-40%
   - ☐ 39%-20% ☐ 19%-0%

Figure 29: Survey of user 3

User 4:



Survey

Hotel Ranking System

Name: *Kavinda* **Profession:** *Software Engineer*

1. How likely is it that you would recommend this Hotel Ranking System to someone?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
2. How satisfied are you with the reliability of this software?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
3. How satisfied are you with this software's ease of use?
   - ☒ Very Satisfied ☐ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
4. How satisfied are you with the look and feel of this software?
   - ☒ Very Satisfied ☐ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
5. Do you have any thought on how to improve this software?

   To increase the speed
6. How would you rate the efficiency of our Hotel Ranking System?
   - ☒ Very Satisfied ☐ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
7. How would you rate the overall performance of the software?
   - ☐ 100%-80% ☒ 79%-60% ☐ 59%-40%
   - ☐ 39%-20% ☐ 19%-0%

Figure 30: Survey of user 4

User 5:

Survey

Hotel Ranking System

Name: *Aqqil* **Profession:** *QA Engineer*

1. How likely is it that you would recommend this Hotel Ranking System to someone?
   - ☒ Very Satisfied ☐ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
2. How satisfied are you with the reliability of this software?
   - ☐ Very Satisfied ☐ Satisfied ☒ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
3. How satisfied are you with this software's ease of use?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
4. How satisfied are you with the look and feel of this software?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
5. Do you have any thought on how to improve this software?

   Speed should improve
6. How would you rate the efficiency of our Hotel Ranking System?
   - ☐ Very Satisfied ☒ Satisfied ☐ Neutral
   - ☐ Dissatisfied ☐ Very Dissatisfied
7. How would you rate the overall performance of the software?
   - ☐ 100%-80% ☒ 79%-60% ☐ 59%-40%
   - ☐ 39%-20% ☐ 19%-0%

Figure 31: Survey of user 5

User 6:

Survey

Hotel Ranking System

Name: *Udara* **Profession:** *Associate Software Engineer*

1. How likely is it that you would recommend this Hotel Ranking System to someone?

☒ Very Satisfied ☐ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

2. How satisfied are you with the reliability of this software?

☐ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

3. How satisfied are you with this software's ease of use?

☒ Very Satisfied ☐ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

4. How satisfied are you with the look and feel of this software?

☒ Very Satisfied ☐ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

5. Do you have any thought on how to improve this software?

Can add some more features which travelers looking for

6. How would you rate the efficiency of our Hotel Ranking System?

☒ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

7. How would you rate the overall performance of the software?

☒ 100%-80% ☒ 79%-60% ☐ 59%-40%
☐ 39%-20% ☐ 19%-0%

Figure 32: Survey of user 6

User 7:

Survey

Hotel Ranking System

Name: *Ranjani* **Profession:** *Hotel Manager*

1. How likely is it that you would recommend this Hotel Ranking System to someone?

☐ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

2. How satisfied are you with the reliability of this software?

☐ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

3. How satisfied are you with this software's ease of use?

☒ Very Satisfied ☐ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

4. How satisfied are you with the look and feel of this software?

☐ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

5. Do you have any thought on how to improve this software?

Identify some hotel feature that customer looking for because current system is limited for 15 features only

6. How would you rate the efficiency of our Hotel Ranking System?

☐ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

7. How would you rate the overall performance of the software?

☒ 100%-80% ☐ 79%-60% ☐ 59%-40%
☐ 39%-20% ☐ 19%-0%

Figure 33: Survey of user 7

User 8:

Survey

Hotel Ranking System

Name: *Ushani* **Profession:** *Tutor*

1. How likely is it that you would recommend this Hotel Ranking System to someone?

☒ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

2. How satisfied are you with the reliability of this software?

☐ Very Satisfied ☐ Satisfied ☒ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

3. How satisfied are you with this software's ease of use?

☐ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

4. How satisfied are you with the look and feel of this software?

☒ Very Satisfied ☐ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

5. Do you have any thought on how to improve this software?

None

6. How would you rate the efficiency of our Hotel Ranking System?

☐ Very Satisfied ☒ Satisfied ☐ Neutral
☐ Dissatisfied ☐ Very Dissatisfied

7. How would you rate the overall performance of the software?

☐ 100%-80% ☐ 79%-60% ☒ 59%-40%
☐ 39%-20% ☐ 19%-0%

Figure 34: Survey of user 8