2019

Travel destinations and route prediction tool: case of dynamic and personalized ecosystem

Phillis W. Kiragu
Faculty of Information Technology (FIT)
Strathmore University

Travel Destinations and Route Prediction Tool: Case of Dynamic and Personalized Ecosystem

Kiragu Phillis Wacera

A Research Thesis Submitted in Partial Fulfilment of the Requirements for the Award of the Degree of Master of Science in Information Technology Degree at Strathmore University

Faculty of Information Technology

Strathmore University

Nairobi, Kenya

June, 2019

Declaration and Approval

I Phillis Wacera Kiragu declare that this research has not been submitted to any other University for the award of a Degree in Master of Science in Information Technology

Student Name: PHILLIS KIRAGU

Student Number: 062464

Sign: _____ Date: _____

Supervisor's Name: DR. VINCENT OMWENGA

Sign: _____ Date: _____

Abstract

Travelling is part of every individual's life today. Travelling has become part of people's life goals. These travels range from short vacations to yearlong round the world trips. There are many reasons why people travel, most of which fall in the broader categories of either business or leisure. People have been travelling for ages and each day attempt to come up with better ways of planning. However, there are many issues that come up when travelling or planning to travel to new places. One of the major issues is not knowing which destinations fits them best and the most optimal route to use. As such most people end up going to popular destinations that are traditionally known. Most systems today are very good at guiding the user on what to do when they get to a certain destination but they do little in identifying the destinations in the first place. This study proposes a natural language processing model that takes in a user's preferences in terms of factors such a budget and weather and returns a list of predicted destinations for that user. Natural Language Processing is carried out by training a neural network model that detects similarities based on word vectors. The end result is a prediction of destinations suitable for a user based on their personal preferences.

Keywords: Travelling, Destinations, Routes, Artificial Neural Networks, Natural Language Processing, Prediction

Table of Contents

List of Figures

List of Tables

List of Abbreviations/Acronyms

ANN	–	Artificial Neural Networks
NLP	–	Natural Language Processing
HMM	–	Hidden Markov Models
CDMA	–	Code Division Multiple Access
Tf-idf	–	Term Frequency – Inverse Document Frequency
Word2Vec	–	Word to Vector
CBOW	–	Continuous Bag of Words
KDD	–	Knowledge Discovery in Databases
CRISP-DM	–	Cross-Industry Standard Process for Data Mining
API	–	Application Programming Interface
EOM	–	Object Exchange Model

Acknowledgements

I would like to sincerely thank my supervisor Doctor Vincent Omwenga for his support, patience and dedication in the course of my research study. His commitment and guidance allowed me to complete my research and achieve the set objectives. I would also like to thank the Faculty of Information technology members for their support in my master's program as well as suggestions and recommendations given throughout the study.

Dedication

This research work is dedicated to my parents and siblings for their support through my academic life. I would also like to send a special dedication to my grandmother who always supported and encouraged me to pursue graduate studies. I also thank God for his immense blessings in my academic life.

Chapter 1. Introduction

1.1 Background

Travelling has become part of life today. People travel for different reasons such as business, education or leisure. In all cases, there is the need to plan ahead of time. There are many factors to consider such as season, weather, safety, budget and cost, and distance. Many people rely on online information and platforms such as google maps to plan their trips. However, this raises a question as to whether the recommended routes are the best for everyone. According to a research done by Ceikute et al as cited in Su, K. Zheng, B. Zheng, & Zhou (2013), there is major difference between popular routes and recommended routes. This means that existing systems still have a gap they have not filled. While these systems have continuously improved mapping of places by using tools such as satellite imagery, they have a general look out that does not always provide a user with the best options. This causes users to end up getting information from the numerous travel blogs and sites which often have conflicting information.

The main reason for this is that most recommendation systems in place today only use distance or time between two destinations which usually end up recommending either the fastest of the shortest routes (Chidlovskii, 2017). The recommendation systems are also very generic. They do not put in place dynamic factors based on personal preferences of travellers. Furthermore, most recommendations consider the actual road networks and geographical map (Su, K. Zheng, B. Zheng, & Zhou, 2013). Other popular travel websites either depend on other users' reviews or historically common destinations (Seyidov and Adomaitiene, 2016). There is a clear need to provide a real time recommender algorithm that takes in dynamic factors as mentioned above and provides predictions based on a user's personal preferences as opposed to a destination's popularity. The real time data can be obtained using crowdsourcing algorithm.

The solution proposed in this study looks at dynamic factors. These factors are used to determine if the destination is viable at the time of travel despite the historical data. For instance, according to Mayaka and Prasad (2012), Kenya is a top tourism attraction. If a person interested in visiting Africa for its wildlife and game, Kenya is one of the top choices. This remained the same even in August to October 2017. During that period, there was a lot of political tension causing many people, residents and otherwise to live in fear. Even at that point, Kenya was still one of the most recommended tourism destinations in Africa. This recommendation clearly went

against one of the factors suggested in this research, security. Most tourists then sought for information from social media platforms in order to determine whether or not it was safe to visit the country. As much as social media sites may be informative, they are also filled with trolls and fake accounts created to spread false information and fear. The algorithm suggested in this research aims at using crowdsourcing data from relevant sources such as government websites to determine the environment in a certain destination. These sources include travel advisories sent out by various governments to warn their citizens against visiting some countries.

The factors considered are based on research done and the data collected in the course of the research. The factors that most people consider when planning a trip include: one, scenery (Alivand, Hochmair & Srinivasan, n.d). Most travellers choose paths that have more water bodies, forests, mountains and parks. Two, travelling time and traffic situation (Lin, Liu, & Gong, n.d). This usually come up in short distance travel and whenever there are scheduled transport modes such as flights or trains. Three, purpose of travel (Amirgholya, Golshanib, Schneiderc, Gonzales & Gao, 2017). This could be one of the most important factors. For instance, if a person is travelling for business, they look for the fastest and shortest possible route. In contrast, a person travelling for leisure may take a longer route if it has more amenities than the shorter route. Four, security for the people and their belongings. In many cases, especially when a person is travelling with small children, security is more important than all the other factors. Therefore, it is important to have measures of how secure a destination is. The final and also very important factor is budget (Amirgholya, Golshanib, Schneiderc, Gonzales & Gao, 2017). The budget is what determines the mode of transport, for instance; whether a person will take a flight or a train or a bus; whether a person will stay in a city or in the outskirts, etc. The budget also differs when a person is travelling alone and when they are travelling as a group. It is important to come up with an algorithm that takes in these and other factors before predicting the best route.

The real-time data is obtained from crowdsourcing. Crowdsourcing attempts to move from the traditional forms of collecting information from a specified group of people. The advantage is that the data in crowdsourcing is obtained from a large group of people from different walks of life, who usually are not biased, making the data more viable (Buecheler, Sieg, Füchslin, & Pfeifer, 2010). The data used is from social media platforms and travel sites.

The algorithm also provides alternative routes for each user. There is a feature where the routes the user prefers will be saved over time. Machine learning will be applied to optimize the algorithm in order to improve the results it provides as more data is collected.

1.2 Problem Statement

Travelling is a big part of everyday life. Therefore, it is paramount for researchers to keep providing new and adaptive ways to make travelling planning efficient and less stressful. Travel planning has many aspects such as budgeting, packing and obtaining insurance. All these aspects vary widely in preparation, but they all depend on the planned destination(s) and routes that they will follow. With the internet age, many people use the online resources to assist in making choices on where to go to and what route to use, or which activities to undertake. However, this information is scattered and unorganized. This leaves many users confused as there is often contradiction opinions. In addition, traditionally, most preferred destinations are well defined geographical areas (Blasco, Guia and Prats, 2014), that may not always be the goto first choice for all individuals.

The internet has made it easier for travellers to navigate their way around the world. There are many applications and websites that make travelling easier. They include maps applications such as Google Maps and Bing Maps. However, these more organized resources use only distance to advice. These applications map all the areas on the globe so that one can know how to move from one point to the other. Currently, Google maps uses satellites to map the world in real time. While it usually helps to know which route to use based on distance and traffic situation, it does not provide any feature to assist a user in choosing which city to go to first.

The other category is the destination advisers, they include popular sites such as TripAdvisor, Airbnb, Booking.com and Expedia. They mainly use other people's review of destinations to give a rating. They also mainly give predictions based on popular destinations that may not work for all users.

Another category is the trip planners such as Google Trips. The android only application allows a traveller to input all the destinations they are planning to go to beforehand and then gives them advise on things like what to do when they are there and places to eat and visit and the like.

Many other existing systems uses time and distance as the major criteria (Chidlovskii, 2017). Those that use other factors such as popularity use static information such previous experiences of other travellers (Su, K. Zheng, B. Zheng, & Zhou 2013). Furthermore, these

systems utilize a general criterial as opposed to personal preferences per user (Ravi, & Vairavasundaram, 2016). The algorithm provides a customized solution for each user.

As it is evident from the above explanation, among all these, there is none that explicitly helps the user determine which routes to follow. For instance, if one wanted to go to a back-packing trip to Europe, he or she would have to visit too many websites to determine the cities to travel to and in what order. There is not a single site that exists today that does that. The only solution that comes close is the use of travel agencies who are often pricy and, in most cases, offer sharing and group activities that are inflexible. These travel agencies also give fixed destinations and timelines making them rigid and unadaptable. There is one fundamental question that all these solutions fail to answer and that is "Where to go based on personal preference at that point in time".

To solve this problem, this study proposes a dynamic algorithm that allows a user to input several factors that they would like to consider when travelling. The algorithm uses some data obtained from a crowd source to determine the most suitable destinations for a user. The algorithm also utilizes a lot of data from crowdsourcing so as to get the most optimal route possible. The algorithm should also be able to provide the attributes of the chosen destinations such as weather, security information, events in the area, etc. The algorithm may be used as a standalone or can be attached to any mapping application. In this study, algorithm is applied in a web application.

1.3 Aim

The aim of this research is to develop a travel destinations and route prediction algorithm that uses dynamic and personalised factors to provide the most fit destination and routes for a user.

1.4 Research Objectives

i. To examine the effects of personal and dynamic factors in the determination of travel destinations and routes
ii. To review existing travel destinations and route determination techniques
iii. To design and develop travel destinations and route prediction tool for a highly dynamic and personalized ecosystem
iv. To verify efficiency of the prediction tool.

1.5 Research Questions

i. What are the effects of personal and dynamic factors in the determination of travel destinations and routes?
ii. What are the existing travel destinations and route determination techniques?
iii. How can a travel destinations and route prediction tool for a highly dynamic and personalized ecosystem be developed?
iv. Does the prediction tool work efficiently?

1.6 Justification

Travel is part of life. Every day, researchers and big companies such as google attempt to come up with more efficient ways to make travelling as seamless as possible. However, most of the existing research and systems have focused on distance, time and popularity. In light of this, it is important to provide a new algorithm that focused on a more customized and real-time solution.

This study deals with how to develop an algorithm that enables each user to get the best recommendations based on their personal preferences. Moreover, the research examines factors that most travellers consider while planning a trip in order to determine which factors to include in the algorithm.

The data used in this research is based on crowdsourcing on the web. This ensures that the algorithm is not biased based on a certain region or class of people. The crowd sourced data includes data from all over the world.

The main beneficiary of the study is any traveller not bound by geography. The solution will allow travellers to identify locations suitable to them based on their own personal preferences matched against destination characteristics. The system will save time and money for the user by allowing them to visit destinations that are suitable for them.

1.7 Scope and Limitation

This research focuses on developing an algorithm that predicts most suitable destinations per user. It also involves developing an algorithm that utilises dynamic factors and finally provides an optimal route. The research does not focus on a particular geographical area. It is designed for users from all over the world. The prediction is solely based on the search parameters a user inputs in the system.

One of the major limitations is viability of the crowd sourced data. The problem is that there is a lot of fake information on social media. The scope of this paper covers how to determine truthfulness of data, but it is not the main focus of the research hence making it an issue. Another major limitation is the diversity of travelling industry. There are many factors that people consider when travelling or planning a trip. This study only covers dynamic factors such as security, weather, socioeconomic status and personal preference.

Another major limitation is that due to the expansive nature of the data that can be applied to the study, it is only possible to train any models with only a few data sources for testing purposes.

Chapter 2. Literature Review

2.1 Introduction

This section covers the literature available from research done prior. It covers the following issues: the effects of personal and dynamic factors when determining travel destinations and route choices, existing travel destinations and route determination techniques and algorithms, existing crowdsourcing and data mining techniques and how the truthfulness of data obtained from a crowd be guaranteed, as well as Natural Language Processing (NLP) and predictive analytics. When making travel plans there are many factors that each person considers. These factors vary depending on the individual. Researchers have conducted multiple studies in an attempt to determine which factors these are. This study considers two categories of factors: personal and dynamic.

2.2 Effects of Personal Factors in Travel Destinations and Route Determination

Personalised factors are those factors that individuals have based on preference and reasons for travel. These factors vary from one individual to another. They further changed based on the number of people who are travelling. For instance, the needs of a person travelling alone are different from the needs of a couple. They also change when there is a group of people travelling. Researchers since 1980s have come up with ways to make travel as personalised as possible. This study has been referred to as Personalised Travel Planning (PTP) or Individualised Travel Marketing (Bartle and Avineri, 2013). These studies show the importance of making travel as close to a person as possible in order to improve the experience.

While personalisation is at the core of the service, hospitality and tourism industry, travel industry is not an exception. Personalisation is about giving customers what they need and feel. The best way to provide a personalised touch is to use current technology such as data mining to determine what people prefer. According to a study done by Google (2018), 57% of travellers agreed that brands should make sure that their information is based on personal preference and past experiences.

2.2.1 Personal Factors in Determining Travel Destinations

People travel for different reasons, therefore the decision to choose the destination will vary from one person to the next. Some of these factors include personal characteristics such as

age, country of origin, cultural, historical, psychological, religion, shopping, gastronomic, events, activities, facilities, and past experiences, among others.

Personal characteristics such as age, stage in the life cycle, occupation, and lifestyle, will have a huge impact on the decision-making process for travel destinations (Fred, 2015). According to a research by Sarma (n.d) young people will look for destinations with high energy activities and ample night life while older people prefer quieter destinations. The stage in life style for instance, young unmarried people tend to travel for longer periods compared to family people, the same is also noticed for older retirees whose kids are all grown up. Occupation and Lifestyle also influences where people go. Most people will travel to destinations that have some connection to their lines of work. For instance, People in technology would choose Silicon Valley while Musicians would prefer Ibiza, Italy.

The country of origin, nationality, is another very important factor. People with nationalities such as Japan, Singapore and South Korea, which ranked number one and two in the Henley Passport Index (2019), have more freedom in choosing the destinations since they can trave1 to 190 and 189 countries respectively, without a visa. In contrast, people from nationalities such as Somalia and Iraq, can only access 30 countries, limiting their mobility. Nationality also affects the attitude of the people at the destinations towards the travellers which may affect the experience (Jonsson and Devonish, 2008).

Cultural experience is another major personal factor when choosing destinations (Seyidov and Adomaitiene, 2016). Travellers will have different expectations. Some people have interests in environmental conservation and will prefer to travel to destinations with flora, fauna, landscapes and amazing natural structures (Holloway and Humphreys, 2016). Others travel looking for appreciation of local culture and hence travel to destinations with deep cultural roots and heritage. Others are drawn to destinations that may have unique cultural practices.

Religion is yet another personal factor that affect travel destination choice. Religious travel involves moving from one destination to another in search of some spiritual awareness or in fulfilment of a religious obligation (Holloway and Humphreys, 2016), for instance Christian Missionaries and Pilgrims. On the other hand, some people will choose destinations based on the popular religion either to study it or to avoid discrimination.

Shopping is another crucial factor in choosing destinations. Some people prefer destinations where they will be free to buy anything and take back home. As such the destinations they choose will be dependent on what type of shopping they would like (Jansen-Verbeke, 1986).

Psychological factors such as perception, motivation, beliefs and attitude also affect the destination choices. For instance, there is a tourism type called "dark tourism", where travellers visit places associated with pain, suffering, and death such as haunted houses and abandoned murder scenes (Holloway and Humphreys, 2016). Others will visit places based on their beliefs or attitude towards that particular place. Others will have different attitudes towards amenities in destinations, therefore looking down on destinations that do not provide the same amenities they are used to, back home.

Past experiences play a huge role in destination choice. While some people prefer to go to places similar to where they have been in the past, others want to experience new environments every time they travel (Seyidov and Adomaitiene, 2016).

Other personal factors that people consider are related to their reason for travelling. For instance, gastronomical travellers move from one destination to another to experience different cuisines and food. Others travel based on events such as music festivals and concerts, and sports. Such people follow events such as a music starts tour to various cities or sports seasons.

2.2.2 Personal Factors affecting Route Choice

When people travel between different destinations, route choice is important. The following are some of the personal factors that influence the route choice from one destination to another.

Personal characteristics such as age and occupation. According to a research done by Eby and Molnar, (2001), people aged 65 and above are more likely to enjoy a long overnight drive as opposed to younger people. In addition, younger people are more likely to choose flights with stop overs as opposed to direct flights.

Infrastructure and Transport modes available are also critical factors. Some travellers will want to enjoy scenery and local places hence prefer ground transport, therefore choosing a route that gives them the best option (Alivand, Hochmair, Srinivasan, n.d). Others will use water

transport to move from one place to the others while others will only want to use air travel. The mode of transport will therefore influence the order of destinations.

The current season of weather is another factor. For instance, if someone is travelling to multiple destinations and want to experience warm weather, then they will choose a route such that they follow the destinations that are warmer.

2.3 Effects of Dynamic Factors in Travel Destinations and Route Determination

Dynamic factors are those factors that change often with time. This means that a traveller needs to have real time up to date information on these factors in regard to destination and routes.

2.3.1 Dynamic Factors affecting Destination and Route Choice

Dynamic factors are those factors that are ever changing based on the season and the time of travel. One of the most considered dynamic factors is budget and cost (Amirgholya, Golshanib, Schneiderc, Gonzales & Gao, 2017). The budget and costs for destinations vary based on the season. Most destinations have high seasons and low seasons. Prices in high seasons tend to go up due to the high demand of services while during the low season prices go down because the demand for the services has declined. There are many angles to this consideration. One of them is when travelling on a low budget. People who have financial constraints will tend to favour cheaper routes than faster or shorter routes. On the other hand, those people who finances are not an issue will want to make the best use of their money. This means either going with cheaper options to save more money or more expensive options which usually save time. No matter what budget you are operating on, one must consider costs in advance (Tian, 2013).

Another important factor is distance and time. These two factors are the most used in today's recommender systems (Lin, Liu, & Gong, n.d; Chidlovskii, 2017). In most cases people choose destinations based on the shortest distance from where they are and routes that take the shortest time. When travelling for leisure however, the shortest distance and time does not always guarantee maximum experience. Research show that these two factors should be considered without giving them priority over dynamic factors.

When travelling for leisure especially, another factor is scenery (Alivand, Hochmair, Srinivasan, n.d). The scenery means the types of views and things travellers get to see along the routes. It usually matters most to people travelling for leisure rather than on business. In this case,

they usually prefer to use longer routes and slower modes of transport such as buses where applicable if the route passes through water bodies, forests, mountains, parks, etc. This means that when choosing destinations and routes, it is important to consider the scenery between the destinations.

Another factor is purpose of travel (Amirgholya, Golshanib, Schneiderc, Gonzales & Gao, 2017). Travel requirements differ depending on the reason for travel. Business travellers will often choose the fastest or shortest routes to get them to their destinations. Leisure travellers will always choose the path that will give them satisfaction based on their hobbies and preferences.

Another one of the most dynamic factors is weather (Barrosa, Martínez, & Viegas, 2015). Weather plays a very important factor when choosing destinations. Most people usually prefer to travel to areas that are sunny and warm as opposed to rainy and cold. Weather patterns sometime change drastically and hence the need to rely on real-time data as opposed to historical and past experiences. For instance, at the beginning of March 2018, weather in the UK changed drastically where a snow storm lead to a red alert. This disrupted every aspect of life especially transportation. In this case, if a person were to rely on past data, they would decide to visit UK as it normally has bearable weather. However, they would be disappointed to find such bad weather which could even be fatal. Weather dictates a lot of things starting from what to pack, what to wear to what activities to undertake. It is therefore necessary to include weather as one of the most important factors to consider when choosing the destinations.

Over the years another factor that has taken precedence is safety and security of destinations (Dijkstra & Drolenga 2008). When travelling to a new country, travellers usually consider safety. It includes safety of themselves and their belongings. In fact, some people even offer to pay more if their safety is guaranteed. One of the factors affecting safety is political environment. In most countries that are usually politically charged, travellers tend to be cautious or even avoid them all together. Security of a particular destinations is dependent on the current situations making it very dynamic.

Lastly, a factor that is slowly cropping up the chain is internet availability (Pel & Nicholson, 2013). In the recent years, travellers have cited internet availability as a factor to consider when planning a trip. This is because most travellers require internet either for social

media or for official business. There are many reasons that travellers cite for the need of a stable internet connections. To mention a few, posting photos on social media, blogging, joining in skype calls for work purposes, etc.

2.4 Recommendation models vs Prediction models

Recommendation models utilises previous behaviour and patterns in order to provide suitable matches. Therefore, for a model to be called a recommender model, there needs to be an existing pattern of behaviour from which conclusions can be drawn. Prediction on the other hand is the based on leveraging large data volumes to provide data on unknown and future events (Ravi and Vairavasundaram, 2016).

This study provides a list of destinations by running a user's preference through a large amount of data to determine which destinations are most suitable, hence it its predictive and not recommendation.

2.5 Natural Language Processing and Predictive Algorithms

Predictive analytics combines data mining, predictive modelling and machine learning to utilise historical data in order to make predictions for a future event. There are various algorithms that have been used in various applications ranging from medicine, to law in the recent years. Natural Language Processing (NLP) is the field that attempts to use natural language as well as humans can despite its form, that is written, or spoken (Weischedel, et. Al.,2003). NLP has been used in a wide variety of applications such as translation, parts of speech tagging, sentiment analysis among others.

2.5.1 General Predictive Algorithms

Over the years there has been a lot of research done on prediction. Many researchers have come up with new algorithms and techniques while others have improved existing ones. Three algorithms that can be used for prediction include Hidden Markov Models (HMMs), Viterbi, and Naïve Bayes. The following section gives a brief description of the three algorithms.

2.5.1.1 Hidden Markov Models

Hidden Markov Models are probabilistic models that contains a finite set of possible states where each of those states has a probability distribution (Crowder, 2011). HMMs are built from Markov chains, which means that the Markov Property holds. HMMs are mainly used for

reinforcement learning and temporal pattern recognition such as parts of speech tagging, speech recognition, handwriting and gesture recognition and bioinformatics.

In predictive analysis, HMMs have been used in areas such as system failure prediction (Salfner, 2005), where the model was used to predict the possibility of system failure in the future. It was used to detect suspicious patterns that could lead to system failure. Other HMM applications include biological studies such as Krogh, Larsson, Heijne, and Sonnhammer (2001), Coast, Stern, Cano, and Briller (1990), and Shihab et al. (2013). Additional research has been done in transportation such as Mathew, Raposo and Martins (2012) and in Customer Relationship Management such as Rothenbuehler, Runge and Garcin (2015).

2.5.1.2 Viterbi Algorithm

The Viterbi algorithm is a dynamic programming algorithm that is used to find the most likely sequence of hidden states (Viterbi, 1967). It was developed as a solution to solve three of the main problems with HMMs: the evaluation problem, the decoding problem, and the training problem. Its main application area is the decoding of convolutional codes used in communications technologies such as CDMA, dial up modems and deep space communications.

In predictive analysis, the Viterbi algorithm has been used in various applications. Livani, Jafarzadeh, Fadali, and Evrenosoglu (2014), used the Viterbi Algorithm for forecasting in electrical power systems. Another application of the algorithm is in crime location prediction (Hussein, Croock, and Al- Qaraawi, 2019).

2.5.1.3 Naïve Bayes Algorithm

Naïve Bayes is a probabilistic classifier based on the Bayesian theorem. It operates on the assumption that all features are independent of each other hence the name ‘naïve’ (Xu, 2018). It is one of the most known and used algorithm since its introduction in the early 1960s (Maron, 1961). It is mainly used in supervised learning, classification and regression.

In predictive analysis, it has a myriad of applications from estimating loan risk (Krichene, 2017) to biological applications such as predicting heart disease (Pattekari and Prveen, 2012), and in internet traffic prediction (Zhang, Chen, Xiang, Zhou, and Xiang, 2013). The naïve Bayes is widely used in many other fields for predictive analytics.

2.5.2 Natural Language Processing Algorithms and Techniques

There are many techniques and algorithms that are used for natural language processing. Each new or improved techniques aims at understanding the natural language better. Many researchers in the past focused on NLP in English but over time other critical languages such as Spanish and Mandarin have started gaining traction (Weischedel, et. Al., 2003).

In the earlier years, most Natural Language Techniques were rule based which proved to be too robust (Winograd, 1971). Later on, researchers adopted Statistical models to solve that challenge. The major tasks for NLP can be broadly grouped into four categories: Syntax, which deals with understanding the structure; Semantics, which deals with understanding the meaning; Disclosure, which deals with summarisation; and Speech, which aims at understanding spoken language. The following section outlines techniques that have been adopted for Natural Language Processing.

2.5.2.1 N-Gram Modelling

N-gram modelling is a widely used Natural language technique that finds the probability of a word occurring after another word or a sequence of words (Brown, et al., 1992). An N-gram is connected sequence of n words in a given set of text. N can be any number starting 1, therefore forming various types of n-gram models. The most basic model is Unigram which provides the probability of a word occurring in a given text. The next level is the Bigram which provides the probability that a word follows another word in a given sequence of text. The next level is the Trigram, and they can move up to N. The computations for the statistics utilise either the chain rule or the Markov assumptions in order to obtain the final result.

N-gram modelling has been used in many applications such as machine translation, text generation from speech, relative sentiment analysis and automatic spelling detection (Sookocheff, 2015).

2.5.2.2 tf–idf Weighting

Tf-idf stands for term frequency–inverse document frequency and it is a numerical statistic that is used to determine how important a word is to a document or a corpus (Wu, Luk, Wong and Kwok, 2008). It has been widely used in search engines to determine the best results to return to a user. It is basically divided into two parts. The term frequency checks the number of times the

word appears in the text while the Inverse document frequency is used to determine how important that word is to the text.

There are many applications of tf-idf, with major search engines adopting customised versions of the algorithm. Other researchers have coupled the algorithm with other techniques such as K-means and cosine similarity (Zong, 2013).

2.5.2.3 Artificial Neural Networks

Artificial Neural Networks (ANN), are advanced computing systems that mimics the operations of a human brain (Luong, Socher, and Manning, 2013). They contain a large number of interconnected neurons that work together to solve a problem. The basic structure of an ANN contains three main layers, the input layer, the hidden layer and the output layer. The input layer is where the data is provided. The hidden layer is where the computation occurs and can contain as many layers as needed. The output layer presents the final results to the user. When the hidden layer contains more than one hidden layer, it is referred to as deep learning.

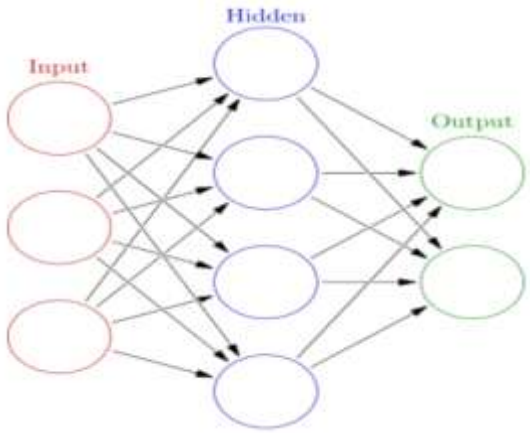

Figure 2.1: Artificial Neural Network

Source (Luong, Socher and Manning, 2013)

Natural language processing using traditional means requires a lot of engineering, that is why researchers started looking into ANN and deep learning to solve this problem. When deep learning is applied, the learns features on its own and all it requires is the pre-processed data (Luong, Socher, and Manning, 2013).

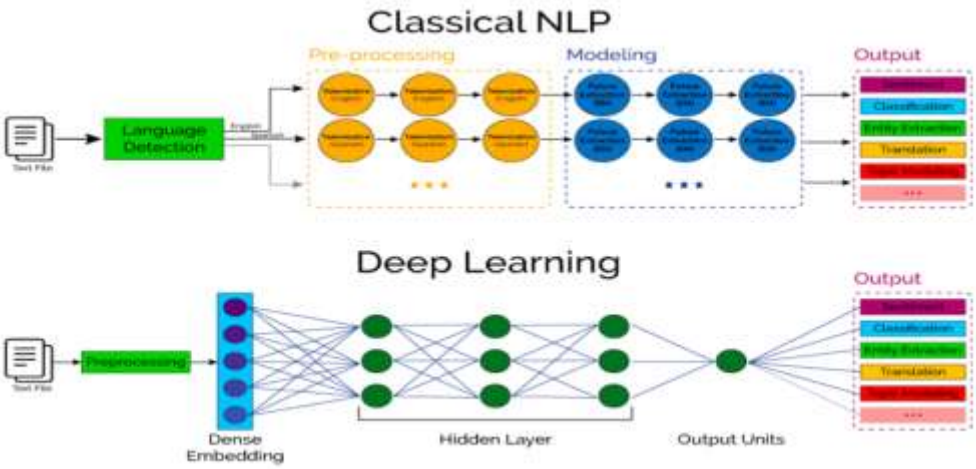

Figure 2.2: Deep Learning

Source (Luong, Socher, and Manning, 2013)

The common approach to conduct Natural Language processing using ANN is to convert the words into vectors so that computations can be worked out mathematically (Mikolov, Chen, Corrado, and Dean, 2013). One of the most common models is the Word2Vec that coverts text to vectors and uses two main methods to CBOW, continuous bag of words, and Skip-grams to predict a word based on context and predict context based on a word respectively (Goldberg and Levy, 2014).

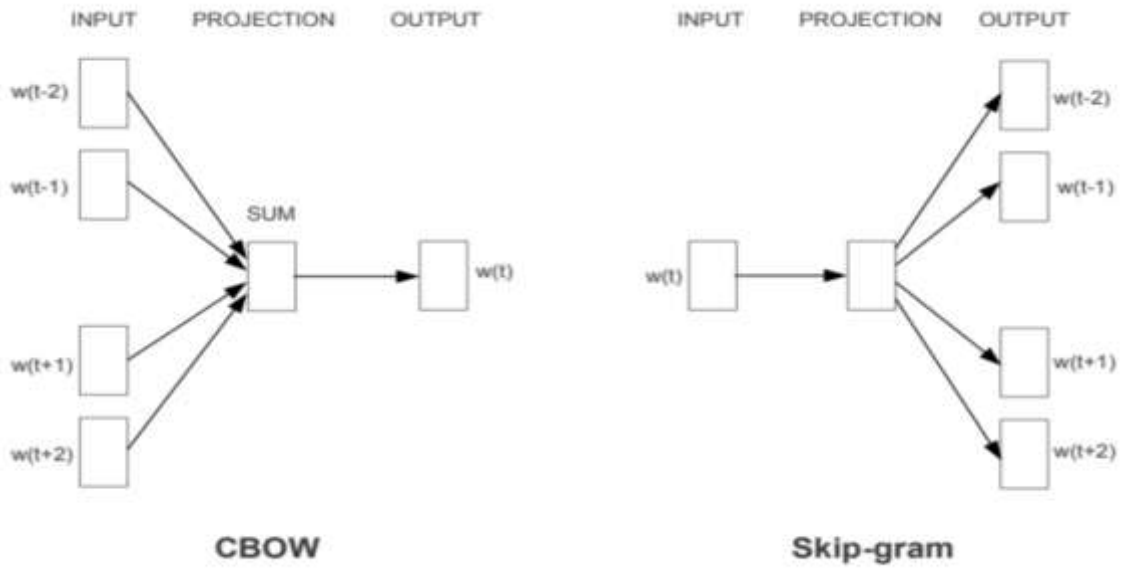

Figure 2.3: Word2Vec Model

Source (Goldberg and Levy, 2014)

Applications of ANNs in NLP span from named entity recognition, sentiment analyses and recommendation engines for varying fields such as scientific research, legal discovery, and e-commerce (Luong, Socher, and Manning, 2013).

2.6 Existing Travel Destination and route prediction techniques and algorithms

There has been a lot of research done in the travel industry in the recent years. Some of the most commonly used systems and well-defined research are defined in the following paragraphs. They include algorithms and techniques for both destinations and route prediction.

2.6.1 Route Prediction Algorithms

The most common route recommender system today is Google maps (Google, 2018). Google maps uses satellite information to determine the distance between two places. They also use satellite to determine real-time traffic situation on the road to determine the fastest route. The advantage with google maps is that it works great for local destinations. However, it uses road networks to recommend routes making international route recommendation a nightmare. For example, the figure below shows a search from a location in Kenya to London, UK on Google

Maps. For destination search, Google maps only provides places reviews under a platform called

Figure 2.4: Map showing google maps route

Google places. It focuses on places such as malls or restaurants and not destinations such as a city.

Another recommendation system is the "TripPlanner: Personalised Trip Planning Leveraging Heterogeneous Crowdsourced Digital Footprints" (Chen et al, 2015). The research proposes a system that applies a heuristics algorithm the routes users prefer to candidate routes and determine the most efficient and effective route. This system fails to consider factors such as weather and safety.

The next recommendation system is a research by Su, K. Zheng, B. Zheng, & Zhou, (2013). The research is titled "Landmark-Based Route Recommendation with Crowd Intelligence". They propose a system that uses crowd-based information to determine the best routes. This system may work efficiently is the factors remained static. For instance, is a user recommends a city based on their trip that happened three years ago, it will be used to rate that city. However, it is possible that the factors have changed since then

The final research tackled in this paper was conducted by Chen et al, (2014), titled "R3: A Real-Time Route Recommendation System". This research concentrates on the traffic situation on the road. It also poses the same challenge as Google maps where only road transport can be recommended.

In conclusion, there is a lot of research conducted on route recommendation. However, most of the research concentrates on specific areas such as running paths (Long, Jia & Xu, 2017), travel costs and fuel consumption (Dai, Ding, Guo & Yang, 2015), location-based social planning (Chen et al, 2015) and public transit (Bajaj, et. al., n.d) among many others.

2.6.2 Existing Destination Prediction Techniques and Algorithms

Over the years there has been many destinations prediction techniques and systems. Most destination prediction systems are moving to a point where they are able to recommend important destinations for a user (Yang and Hwang, 2013). Others have used the concept of decision-making theory to understand how travellers conclude on destinations (Hsu, Tsai and Wu, 2012; Huang and Bian, 2009). The following are some of the travel destinations recommendation and prediction systems that have been covered in research.

iTravel (Yang and Hwang, 2013), is a destinations recommendation system in a mobile peer to peer environment. This application recommends attractions and destinations on a rating basis. Users with iTravel are required to rate the places they have visited where other users can see the best rated systems. Users can also pose questions to other users through the system.

Hsu, Tsai and Wu (2012), proposed a 4-level AHP model that uses preferences to establish important factors that tourists consider when choosing Taiwan as a destination choice. The 4th level consisted of 22 attributes. The study then used fuzzy set theory to categorise and evaluate the factors with visiting friends/family and personal safety ending up as the top two factors that Tourists considered.

Sebastia, García, Onaindia, and Alvarez, (2009), proposed a system called e-Tourism that involves two steps. First, the system provides a list of destinations based on the demographics of the user, and former trips that they have taken. The system them plans a schedule based on some metrics. The system uses Artificial Intelligence planning to come up with the schedule. It is also adaptive relying on user feedback to enhance the effectiveness.

Ravi, and Vairavasundaram, (2016), proposed a location-based system that was based on social pertinent trust walker (SPTW) for a group of user recommendations. They based their research on the fact that there is a lot of data provided on the internet by users that is underutilised.

ATRS, a system proposed by Etaati and Sundaram, (2015), is an adaptive tourist recommendation system that was built on top of the existing systems to improve their adaptiveness and growth. The new system tries to achieve adaptiveness through changing a user's preference when it changes.

2.7 Data Mining on the Web

Data mining is the process of obtaining patterns from large data sets it is a confluence of various fields such as artificial intelligence, statistics, and database technology (Han, Kamber and Pei, 2012). It stemmed from the fact that there is so much data in the world especially in the new information age. Its main purpose is to turn tonnes of otherwise meaningless data into knowledge. The data sources are numerous ranging from databases, data warehouses, repositories and the web.

There are various techniques used in data mining. They all follow a process called the Knowledge Discovery in Databases process (KDD), that has five main stages (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). These steps are: Selection, Pre-processing, Transformation, Data Mining and Interpretation/Evaluation. Selection is the process of identifying the data sets and their sources. Pre-processing also sometimes called data cleaning is the process of removing unnecessary noise, handling any missing parts of the data and establishing any known changes. The end product of the pre-processing stage is processed data that can be feed into models. The third stage, transformation, involves finding the most useful features that can represent the data depending on the reason for mining. The fourth step is the actual pattern recognition, data mining. It involves choosing the data mining algorithms and methods to use and searching for patterns in the transformed data. Once the patterns are identified, they need to be interpreted and evaluated to give a conclusive finding. This is achieved through various methods such as visualisation. The results are then presented as knowledge. Another additional step is the processes of acting on the knowledge. The steps can be performed iteratively as required.

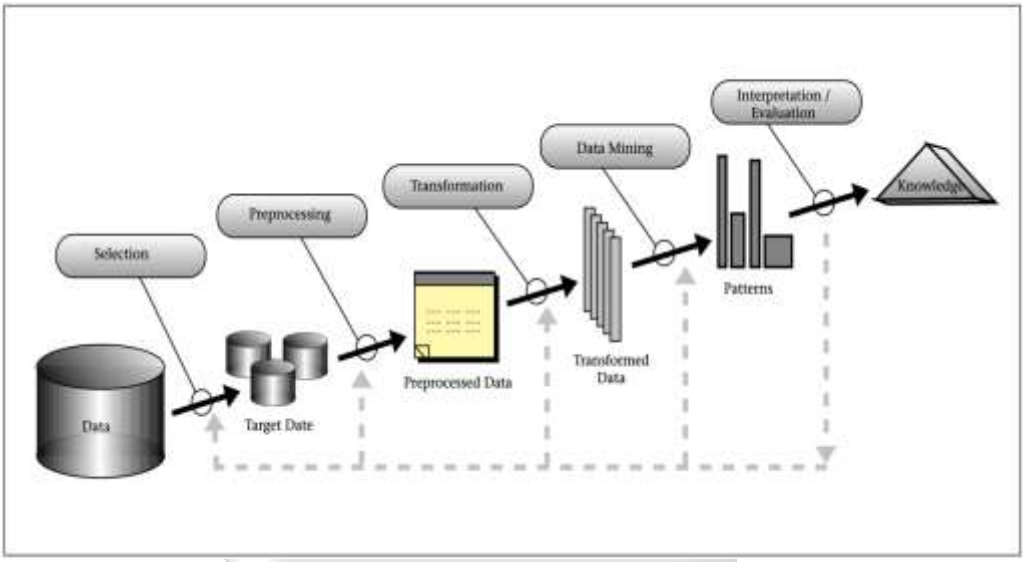

Figure 2.5: KDD Process

Source (Fayyad, Piatetsky-Shapiro, and Smyth, 1996)

Over the years, there have been many themes and variations of the KDD process. However, they all follow some generic steps from the raw data from the source to the knowledge obtained from them. One of the most commonly used models is the Cross-industry standard process for data mining (CRISP-DM). The CRISP-DM model has six steps that are an extension of the original KDD and has been largely accepted by data miners (Shearer, 2000). These steps are, Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment. Business Understanding involves determining the business objectives and outlining the business goals. Data Understanding involves collecting initial data, describing and exploring it while examining its quality. Data Preparation involves selecting, cleaning and transforming the data. Modelling involves selecting the modelling techniques, building and assessing the model. Evaluations involves analysing the results and reviewing the process. The last step is Deployment that provides the final knowledge as well as maintenance and overall project review.

Data mining on the web widely known as web mining is the process of extracting documents and data from the world wide web and extracting patterns from them. It can be divided into three major categories: web content mining, web structure mining and web usage mining (Jokar, Honarvar, Aghamirzadeh, and Esfandiari, 2016). Web content mining faces various challenges such as complexity of web pages, the web being too large, dynamicity and diversity of

information of users and how to figure out the relevance and reliability of that information (Jawad, 2018). There are various techniques used for web mining which are broadly classified into four categories: unstructured, structured, semi-structured and web mining (Kumar and Singh, 2016). These techniques have been developed into tools such as SKICAT, Web Crawler, Multimedia Miner and Object Exchange Model (EOM), among others. In the recent years, there has been more and more organisations that provide data for public use in form of Application Programming Interfaces (APIs). These platforms have a lot of data in almost all research fields. As such, web mining for individuals and small organisations has become easier and less challenging.

2.8 Crowdsourcing

Crowdsourcing is the use of information based on the intelligence of a group of people. In the recent years, the concept has gained traction and attention in both commercial and academic perspectives (Amrollahi, 2016). The main reason for the attention is that crowds sourcing offers flexibility in terms of time and location. It also takes advantage of the information age where people are willing to share a lot of information online.

Crowdsourcing usually improve the quality, volume and timeliness of data (Garcia-Molina, Joglekar, Marcus, Parameswaran & Verroios, n.d). However, it also has its challenges. One of the most common challenges is uncertainty. This is the possibility that the data obtained from crowdsourcing is incorrect due to the fact that anyone can post anything on the internet.

Another challenge is that even when the data is correct, it is in natural language which may be easily misinterpreted (Garcia-Molina, Joglekar, Marcus, Parameswaran, & Verroios, n.d). Sometimes humans use language to mean something different that the literal meaning of the word.

Despite its challenges, crowdsourcing still remain an efficient way to obtain data from all corners of the globe. It allows the algorithm to get all the real-time data it requires to produce the best results.

2.9 Conceptual framework

The figure below shows the conceptual framework for the algorithm. The independent variables in the model are the user preferences provided from the web application. The factors are fed into the pretrained model. Data from the data source, in this case, Wikivoyage, is taken through

a rigorous cleaning to return it in a list format. The list is used to train a neural network model. The trained model is then used to provide the list of final destinations for a user.

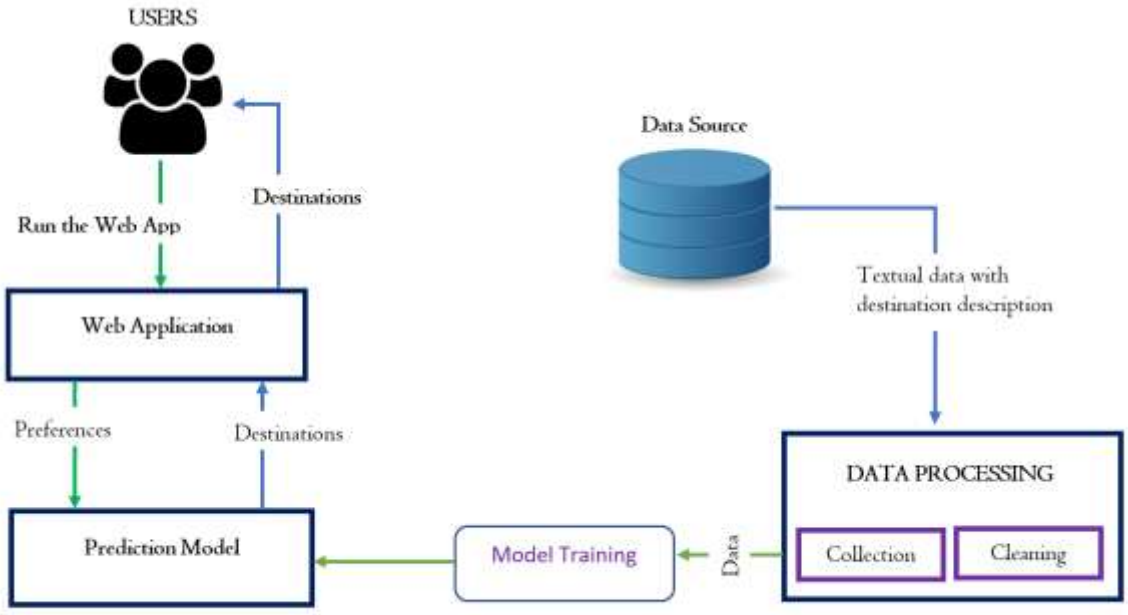

Figure 2.6:Conceptual framework for the study

Chapter 3. Research Methodology

3.1 Introduction

The aim of this research is to find out the factors that people consider when travelling and using these factors to provide a personalised algorithm that can be used to provide the most suitable destinations and most optimal routes for travellers. This chapter describes the methods that used for conducting the research and their viability. In addition, the chapter contains the data collection techniques and analysis that are used as well as the target population. Moreover, this section introduces the approaches that used for system analysis, system architecture, system design, system development, and implementation and testing.

3.2 Agile Software Development Methodology

Agile development is a software development method that allows for faster and iterative development (Abrahamsson, Salo, Ronkainen, & Warsta, 2002). This means that each step of development can be iterated as many times as needed. Its main features include modularity, iterating of short cycles, people-oriented, adaptive and incremental (Abrahamsson, Salo, Ronkainen, & Warsta, 2002). The figure below illustrates the main cycles of agile development which include requirements analysis, design and implementation, testing and evaluation (Boyer, 2015).

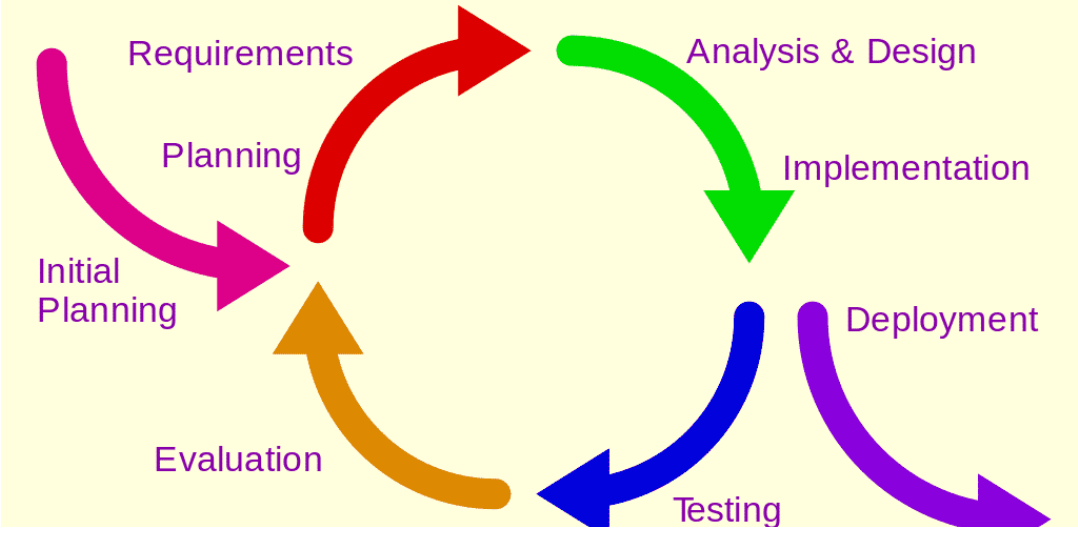

Figure 3.1: Agile development
Source (Boyer, 2015)

Agile methodology is the most suitable for this research because it is iterative in nature. The research focuses on using real-time data for recommendation. This means that the dynamic factors used by the algorithm may change from time to time. In this case it is most suitable to use a methodology that is adaptive and incremental. Further, the more the data fed into the algorithm, the better results it is bound to give.

In the initial planning, the scope of the algorithm was determined. In this phase it was important to outline the boundaries of the algorithm. In the requirements analysis phase, intensive requirement analysis will be carried out to identify the personalised and dynamic factors that affect destinations and route choice. In the analysis and design phase, the models and information flow models will be produced to help structure the requirements. This also includes designing the system that was used for the proof of concept. The implementation phase involves writing the code and tests for the algorithm and the application. The testing and evaluation phases were used for user acceptance tests and checking the efficiency of the algorithm. These phases were carried out iteratively as needed. Lastly, the algorithm was packaged in a way that it can be used in system.

3.3 Research Design

Research design is the systematic approach that a researcher uses to conduct their study (Creswell et al, 2013). The research design is informed by the methodology, the literature and the research questions.

3.4 Target Population

The target population is people on the internet. These include travellers, tour guides, tour operators, travel companies and other tourists. Therefore, it means that this study worked with an unknown population size. The amount of data gathered could span to very large numbers to infinity on any particular cycle of the algorithm hence it is necessary to find a way to derive a working sample size.

3.5 Data Collection

The data used in the study was obtained from an online public source called Wikivoyage. The site is updated publicly and contains details of many destination around the world. This enhances the crowd sourcing the research aimed to utilise. The site provides data in a format called xml. This data is in a raw form that contains many unnecessary details and therefore has to go

through a rigorous cleaning process for it to fit the model. The method of data collection used was document analysis. The data set has no specific labels or features. After the cleaning of the data, it was used in the training of a neural network whose hidden layer learns the data on its own.

3.6 Data Analysis

The process of analysing and evaluating qualitative data usually involve interpreting and trying to understand the data based on the perspective of the participant.

The data acquired underwent data preparation. The data requires to be cleansed and formatted in a way that it is can be consumed by the model. The word2vec model used in the study is trained using a corpus in a list format. The dump articles from the source are passed through various functions to ensure that the end result is a list of word. The first step was to remove special characters that were used for punctuation and formatting. The data is then transformed into lists within lists that are then provided in a json format for model training.

3.7 Model Development

The process of developing the model involved various steps. Natural language processing involves processing of text-based data so as to understand the meaning of the text. The data obtained in the study is text based and therefore has to be processed in order to provide predictions for the user.

This research performs natural language processing through a method known as word embeddings. This is where similarity of text is determined by comparing the word vectors. Specifically, the research focus on using continuous skip gram. Given text, skip gram aims to predict the context in which that text is used in (Mikolov, Chen, Corrado, and Dean, 2013). The weight obtained from the text is used to determine where that text falls.

When a user adds the parameters of the destinations they are interested in, that text is converted into its vector space. The text is then run through the pretrained model to determine the destinations where the text fit in.

3.8 Research Quality Aspects

Research quality can be checked using four main aspects: relevance, credibility, legitimacy and effectiveness (Belcher, Rasmussen, Kemshaw and Zornes, 2016). These aspects try to check

that the research done is relevant to the field it is based in, credibility that can be proven through scientific means, legitimate in its claims in that they are no false declarations and truth in data collection and finally effectiveness, which means that the proposed solutions work better than previous works. These aspects can be represented by asserting the validity and reliability of the research. These are the aspects that were checked in this research.

3.8.1 Validity

Validity is important because it confirms that the conclusions made accurately reflects the research. Because the research is qualitative in nature, the validity was checked based on its dependability.

The research quality was determined by the validity of the algorithm and by checking if all the research questions have been answered. The validity of the algorithm can be determined by getting a survey from the intended users. The users will be asked to determine if the destinations and routes recommended by the algorithm was their best option or not. This will also determine how much a user can depend on the model for their destinations and routes determination.

3.8.2 Reliability

Reliability is a way of assessing the measurement procedure used in data collection. Reliability of a study guaranties validity, therefore, for any study to be considered valid it must be proved that it is reliable. One of the main ways to measure reliability is by using the error component (Cronbach, 1997). This is achieved by calculating the mean squared error.

Reliability of the research was determined by checking the ability of the algorithm to give the same results if the factors provided remain the same. This is because, when crowdsourcing, the algorithm is bound to select different sources that vary slightly. Reliability tests ensures duplication of results.

3.9 Ethical Considerations

The ethical consideration in this research was determining the truthfulness of the data obtained from the crowd. This research notes that the data obtained from the internet is not always true. The main issue was to try and obtain valid data. As such, the algorithm does promise 100% correct recommendations.

Chapter 4. System Design and Architecture

4.1 Introduction

This chapter explores the design and architecture of the proposed travel destinations and route prediction algorithm. The propose tool aims at predicting a list of destinations for a user based on their own preferences and then proposing an optimum route that can be followed by the user.

This section explains the structure and characteristics of the algorithm as well as diagrammatic representation of the resulting tool. It shows the connection between the system components and the interactions between objects illustrated through the use of various UML diagrams.

4.2 Requirement Analysis

Requirement analysis of a system involves determining the users of the system and their expectations of the end result. The section below outlines the functional requirements and non-functional requirements for the proposed prediction model. The requirements are in line with the objectives in chapter one.

4.2.1 Functional Requirements

Functional requirements identify what the system must do to fulfil user functions and activities. This includes any behaviour, inputs and outputs that supports these functions. The functional requirements for the travel destinations and route prediction model include:

i. The system should accept data from the crowdsource, clean the data and train the Neural Network model.

ii. The system should allow a user to enter the search preferences they want the algorithm to consider such as personal preference and other dynamic factors.

iii. The system should clearly display the recommended destinations and routes.

iv. The system should allow the user to select the destinations and access more information on them.

4.2.2 Non-Functional Requirements

Non-functional requirements are qualities that the system should have to ensure smooth operations. These requirements may include attributes such as accuracy and availability. While these requirements are not tangible and cannot be directly represented in a system like with functional requirements, they very critical as they inform the user experience. The non-functional requirements for the proposed system include:

i. **Availability**: The system should be available 24/7. The systems can be used by any user anywhere in the world. This means that the system needs to run efficiently at all times without being affected by any time zone.

ii. **Reliability**: The system provides data about sensitive issued such as security and weather. It is therefore essential that the results are reliable so as to serve users well.

iii. **Scalability**: One of the aspects of the proposed model is machine learning. The system needs to be flexible enough to grow on its own without affecting efficiency.

iv. **Responsiveness**: The system should be as fast as possible. The algorithm should be optimised to reduce response time.

v. **Adaptability**: The algorithm is expected to fit into any application. These include web applications, desktop applications and mobile applications.

vi. **Security**: The system should ensure that user data is secure.

4.3 Proposed Tool

The proposed tool has two components, the algorithm and the web application. The structure of the algorithm is dependent on the data structure and the sources. The design therefore takes into consideration of these factors as it outlines the stages that exist in the algorithm right from the input to the output.

4.3.1 Algorithm Characteristics

The following are the characteristics of the proposed algorithm:

I. Data Source

This is where the data being used is collected from. The research employs crowdsourcing to obtain destinations data. Therefore, data sources vary from travel websites such as trip advisors to APIs by private companies. For the purpose of this study, the data is obtained from Wikivoyage, which is a publicly edited and verified wiki site.

II. Data

This is the actual data obtained from the data source. The data includes destinations, their characteristics and geographical markers for those destinations. The data is in raw text format and therefore had to undergo some preparation and cleaning for it to be usable in the algorithm.

III. Algorithm

This describes the actual steps that are involved from the input to the output. The algorithm has two major stages. These stages can be summed up as follows:

i. Training the word2vec model using data from Wikivoyage

ii. Use the trained model to predict destinations based on the user's input.

IV. Algorithm Structure

The structure diagram in figure 4.1 below broadly describes the major steps of the algorithm. The structure does not include the web application that is used to obtain the user input and display the results.

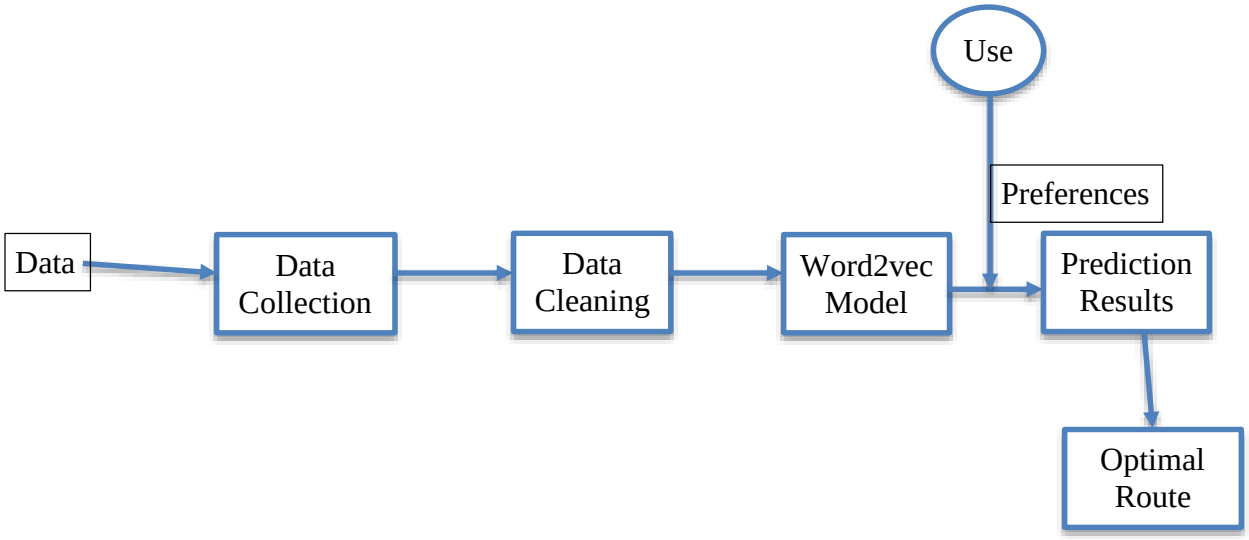

Figure 4.1: Proposed Algorithm Structure

Raw data is obtained from the source in whichever format it is provided. The data collection stage is used to transform this data into the required format. In the data cleaning stage, the data goes through a process that attempts to remove all the noise from it such as links, special characters and unwanted spaces. The word2Vec Model is then trained using the cleaned data so as to give predictions. At this point the user inputs their preferences and gets the prediction results. The final result is for the algorithm to provide an optimal route for the user.

4.4 System Architecture

The system architecture explains the proposed model design and user requirements. The section below outlines the different components of the system and how they work together to achieve the requirements.

Figure 4.2 below shows the various components. The main components in the model are data collection, data preparation, model training, prediction output, and route optimisation. They all work together to provide the most suitable destinations for a user.

The first step is to obtain the data from the source. In this case the data was obtained from the public site, Wikivoyage. This is done in the data collection phase. Next, the data has to be cleaned and formatted in the correct format usable by the model. The data is first transformed into JSON format then cleaned to remove unnecessary characters and spaces. The data is then put

through the model for training. A user then enters their preference and gets the most suitable destinations and the optimum route.

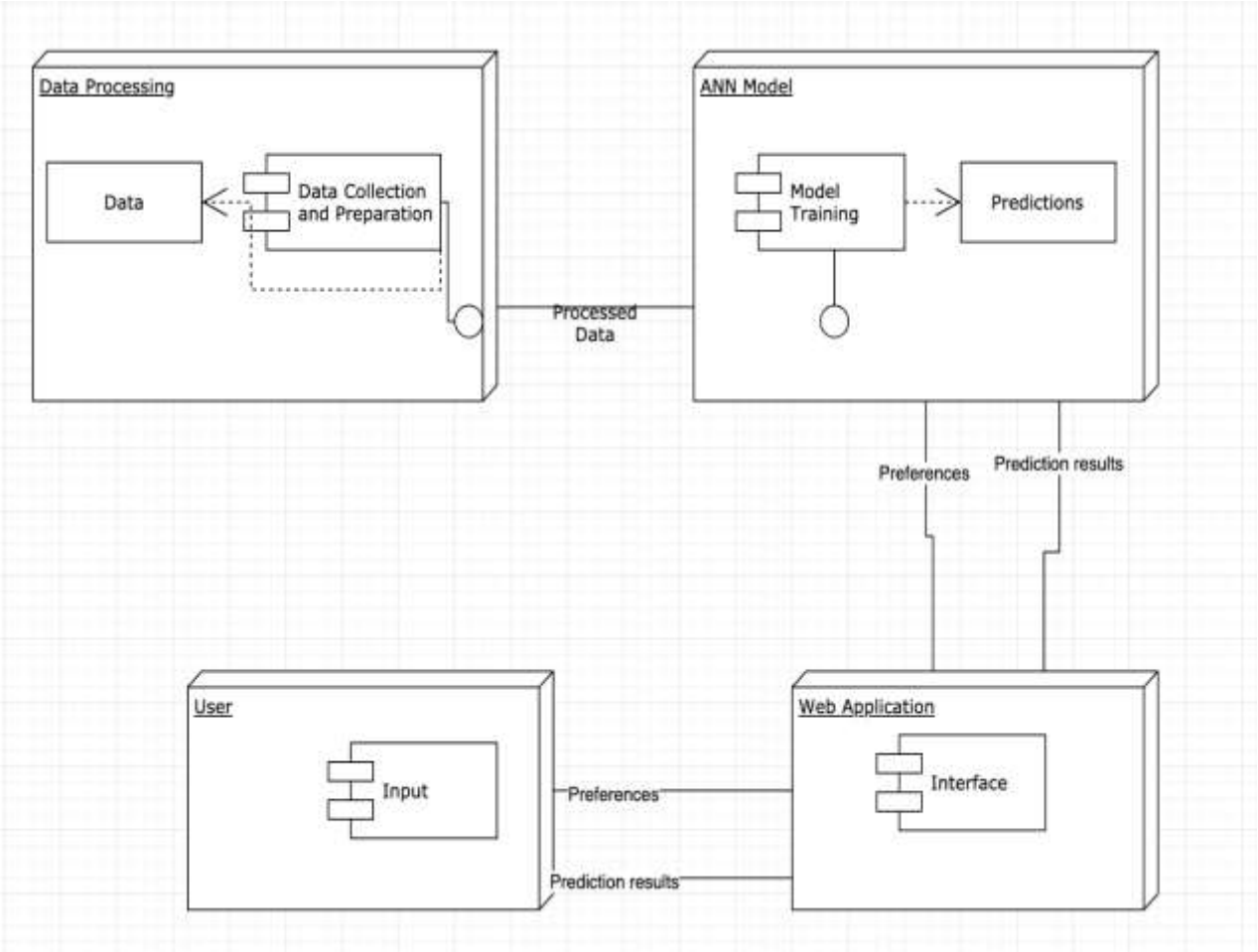

Figure 4.2: System Architecture

4.5 System Behaviour Modelling

Behaviour modelling using use case diagrams and sequence diagrams is one of the most efficient ways to show the interaction between the actors and the system.

4.5.1 Use Case Diagram

The actors in the system are users, developer and the crowd source. The developer trains the model first to allow it to give proper predictions. They are also involved in the data collection and cleaning. The user has to enter the search preferences they require the model to apply. The model uses the user input to provide a list of travel destinations and routes. The algorithm then outputs the end results where the user can view the suggested destinations and routes to use. Figure

4.3 below show the use case diagram for the model while table 4.1, table 4.2 and table 4.3 show the main use cases with their success scenarios.

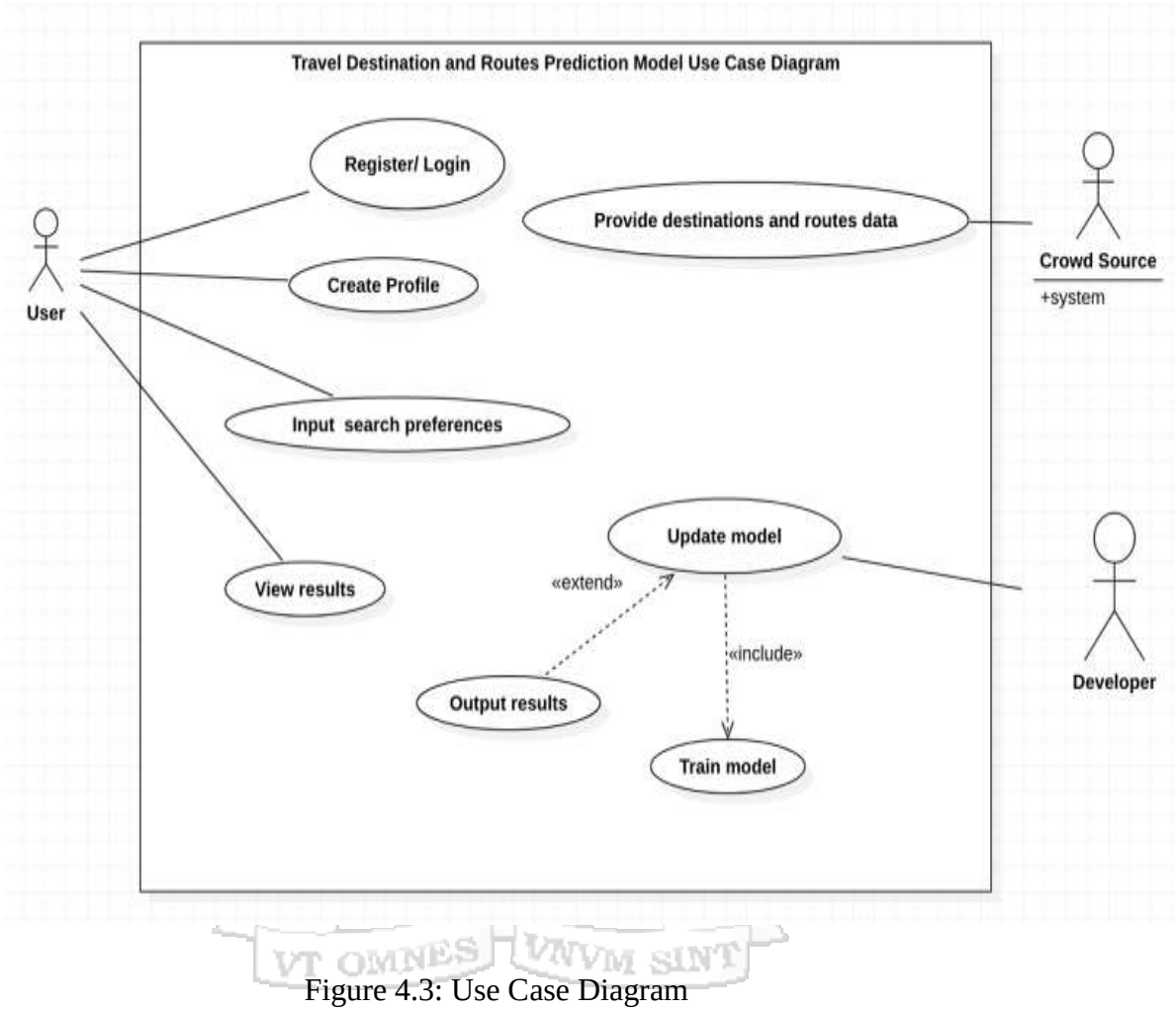

Figure 4.3: Use Case Diagram

Table 4:1 Profile Creation and User Input

Use Case: Profile creation and User Input
Primary Actors: User
Precondition: Input the search preferences
Postcondition: Obtain all the factors that the user requires their destinations and routes to contain

Main Success Scenarios	
Actor Intention	**System Responsibility**
1. User loads the web application	
2. User inputs parameters important to them	
	3. System saves the user input to their history
	4. System sends the parameters to the algorithm

Table 4:2 Data collection

Use Case: Data mining to obtain travel destinations and routes	
Primary Actors: Crowd Sources	
Precondition: None	
Postcondition: Provide a list of all destinations and routes	
Main Success Scenarios	
Actor Intention	**System Responsibility**
1. Crowdsources provide all destinations	
	2. Filters the data for destinations

Table 4:3 Data cleaning and model training.

Use Case: Data collection, cleaning and model training.

Primary Actors: Developer	
Precondition: Obtain a list of destinations and routes, user input, and train the model.	
Postcondition: Provide the most optimum prediction	
Main Success Scenarios	
Actor Intention	**System Responsibility**
1. The algorithm obtains the list of destinations through data mining	
	2. The system provides user profile and input
3. Algorithm uses the Word2Vec Neural Network to produce results	
	4. The system displays the top result to the user

4.5.2 Sequence Diagram

Figure 4.4 shows the sequence diagram. The data collection involves getting the data from the source. The data is then cleaned and formatted in data preparation. The model is trained using this data. A user loads the web application to allow them to input their search parameters. The user then has to input the parameters that she wants to be considered. These will be in terms of the personal and dynamic factors mention in chapter 2. The data is then sent to the trained model which will get the possible list of destinations and routes. Once the destinations have been received, the analyser will then pick the most suitable recommendations and display them to the user.

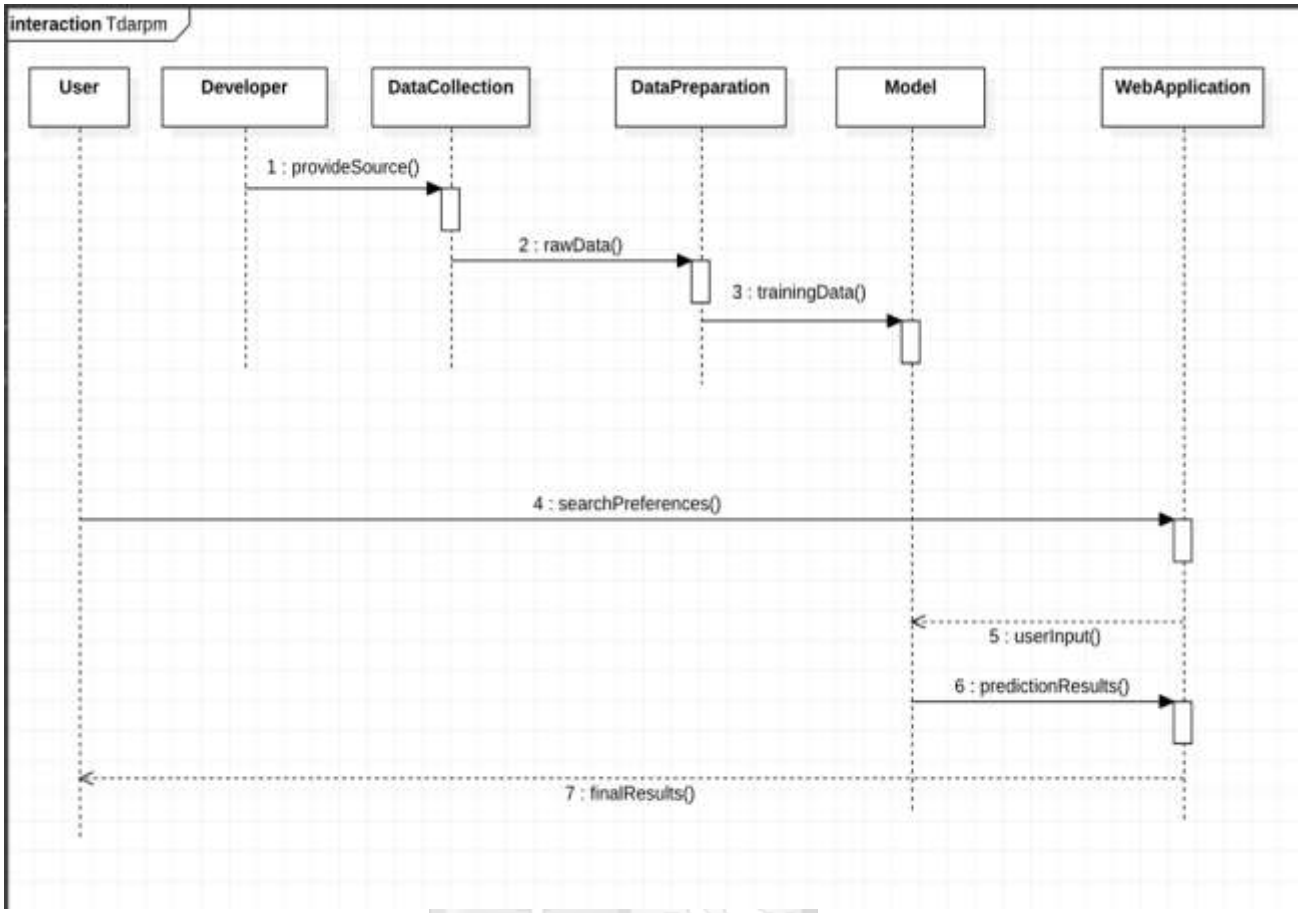

Figure 4.4: Sequence Diagram

4.6 System Process Modelling

Process modelling is used to show how the data flows through the system. This is demonstrated through the use of a context diagram, a level 0 data flow diagram and a flow chart diagram.

4.6.1 Context Diagram

Figure 4.5 below shows the context diagram. There are three major external entities: the users, the crowdsource and the developer. The user sends the parameters and web application which in turn sends the proposed destinations and routes. The crowdsources provides the list of destinations in raw form. The developer prepares the data and trains the model.

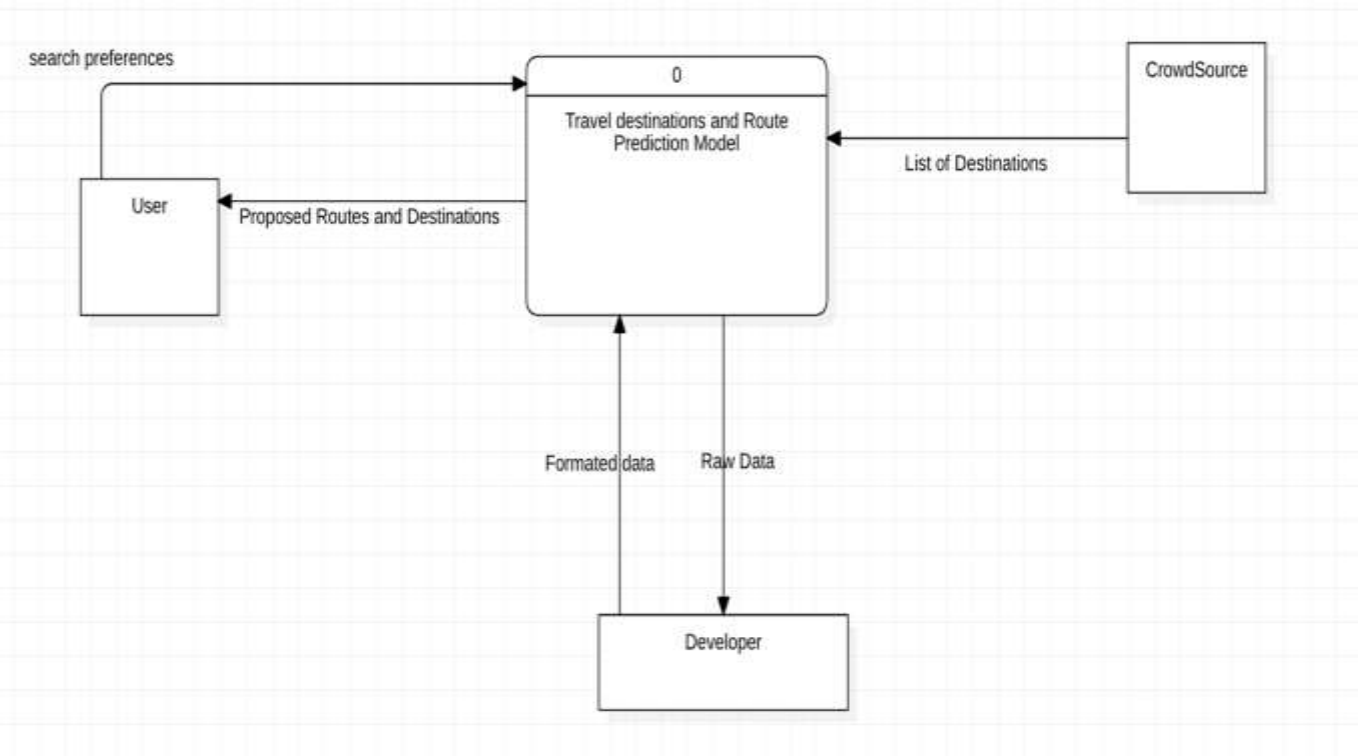

Figure 4.5: Context Diagram

4.6.2 Level 0 Data Flow Diagram

Figure 4.6 displays the DFD Level 0. The figure is an extension of the context diagram above. Figure 4.6 contains the data stores, data flows between the processes as well as the main processes. There are four major processes that have been identified. The first process is the collecting of destinations data from the crowdsource. The second process cleans and prepares the data. The third process involves training of the model. The last process is analysing and recommending the suitable destinations and routes. This process the user input and the trained model, to give the final list of destinations to a user.

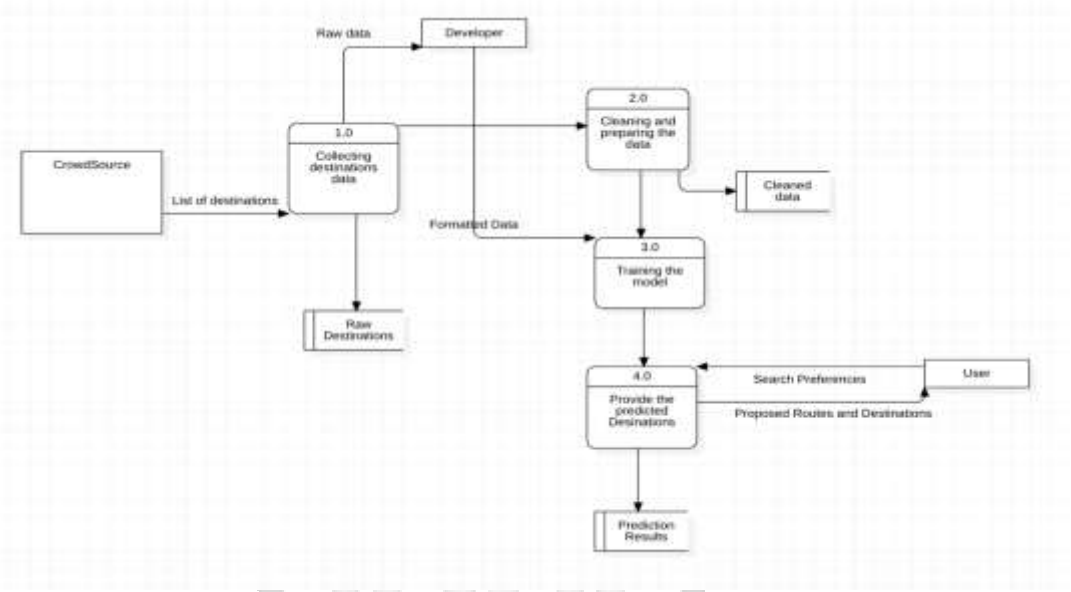

Figure 4.6: DFD Level 0 Diagram

4.6.3 Flow Chart

Figure 4.7 below represents the flow chart diagram. The process starts when a developer obtains data from the source. The data is then cleaned and prepared for the model. The model is then trained with the data to allow for it to give predictions. After that, a user inputs the search preferences they want considered in the mining of the destinations. After that the list of suitable destinations is displayed to the user.

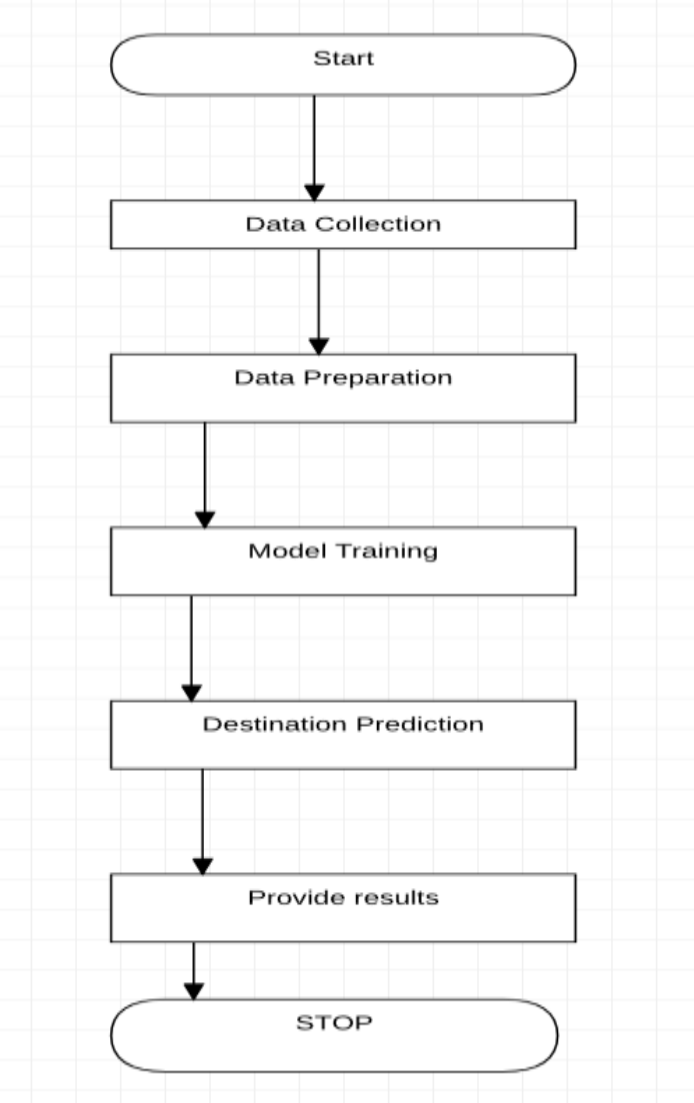

Figure 4.7: Flowchart Diagram

Chapter 5. System Implementation and Validation

5.1 Introduction

The implementation of the proposed tool includes all the steps right from obtaining the travel destinations data, cleaning and preparation, training of the word2vec model, and the code for the web application that will be used to proof that it works. Data collected is transformed into JSON format and then taken through the cleaning step. The following section explains all the steps involved in the implementation all the way to the final output that id displayed to the user.

5.2 Development Environment

The implementation is in python programming language. The model has been trained using the Jupiter Notebook platform. User input and display of results is achieved through a web application in Flask framework combined with AJAX to send user requests. The IDE used was IntelliJ's PyCharm.

5.3 Tool Implementation

The implementation involves many distinct steps as follows: the data collection steps mines the data from the source. The data preparation step is where data is cleaned and formatted. The next step is the model training using the cleaned data and filtering locations giving them geographical markers. After that a user inputs their preferences and they can view the predictions and more details about them. The algorithm applied in this research is an Artificial Neural Network with one hidden layer. The merits of using this algorithm is that it uses vectors for Natural Language Processing making it more efficient.

5.3.1 Data Collection

Travel destinations data was obtained from the Wikivoyage API. It is in eXtensible Markup Language (XML) form. Since the required data format is Json, a simple python script was written to convert this data. The data is then saved into a file that will be used later in the system.

```
xml_str = open(filename).read()
o = xmltodict.parse(xml_str)

json_str = json.dumps(o)
json_d = json.loads(json_str)

with open('../data/wikivoyage/wikivoyage.json', 'w') as j:
    j.write(json.dumps(json_d))

return json_d
```

Figure 5.1: Data Collection

The second set of data is geo data for the pages that is provided in sql format. In order to query this data, a MySql database was created. This data will be able to differentiate the geographical locations from results that match but are not locations.

5.3.2 Data Preparation and Cleaning

The data obtained from the collection process in JSON format contains a lot of noise and unnecessary characters. Therefore, it needs to go through a data cleaning process. The first step is to load the JSON file. After it is loaded, the content is modified so that it is represented in python dictionaries for easier processing. The following is an example of dictionary keys that represent article titles.

```
an_(Texas)', 'McLean_(Virginia)', 'Mclean_County', 'McLean_County', 'McLean_County_(North_Dakota)', 'McMinnville', 'M
cMinnville_(Oregon)', 'McPherson', 'Mdina', 'Mdumbi', 'Meads_Bay', 'Meadville', 'Mealhada', 'Meath', 'Medugorje', 'Me
canhelas', 'Mecca', 'Mecca_(disambiguation)', 'Mechanicsburg', 'Mechanicsburg_(Pennsylvania)', 'Mechanicsburg,_pa',
 'Mechelen', 'Mechi_(zone)', 'Mechi_Zone', 'Mecklenburg_County', 'Mecklenburg_County_(North_Carolina)', 'Mecklenburg_
County,_NC', 'Mecklenburgische_Seenplatte', 'Mecklenburg-Western_Pomerania', 'Medan', 'Medanos', 'Medellin', 'Medelli
n_(Colombia)', 'Medelpad', 'Medewi_beach', 'Medewi', 'Medfield', 'Medford', 'Medford_(Massachusetts)', 'Medford_(New_
Jersey)', 'Medford_(Oregon)', 'Media', 'Medias', 'Mediawiki_templates', 'Medical_tourism', 'Medicine_Bow', 'Medicine_
Hat', 'Medicine_Park', 'Medicine_Park,_Oklahoma', 'Medina', 'Medina_(disambiguation)', 'Medina_(Ohio)', 'Medina_count
y', 'Medina_County_(Ohio)', 'Medina_County_(disambiguation)', 'Medina_County_(Texas)', 'Medina_de_pomar', 'Medina_de_
Pomar', 'Meditation_in_Japan', 'Meditation_in_Thailand', 'Mediterranean', 'Mediterranean_Europe', 'Mediterranean_Moro
cco', 'Mediterranean_Sea', 'Mediterranean_Turkey', 'Medjugorje', 'Medjumbe', 'Medlow_Bath', 'Medora', 'Medway_(Ohi
o)', 'Medway_(Massachusetts)', 'Meerbusch', 'Meerlanden', 'Meersburg', 'Meerut', 'Meetjesland', 'Megabus', 'Megalithi
c_temples_of_Malta', 'Megalochori', 'Megeve', 'Meghalaya', 'Meghri', 'Megiddo', 'Mehedinti_County', 'Meidum', 'Meie
n', 'Meizhou', 'Meizhou_Island', 'Mekedatu', 'Mekele', 'Meknes', 'Mekong_Delta', 'Mekong_Delta_(Vietnam)', 'Mekong_Lo
wlands_and_Central_Plains', 'Mela', 'Melacauvery', 'Melaka', 'Melanesia', 'Melbourne', 'Melbourne_(disambiguation)',
 'Melbourne (Florida)', 'Melbourne, Australia', 'Melbourne/Albert Park', 'Melbourne/Brunswick and Coburg', 'Melbourn
```

Figure 5.2: Dictionary Keys From Cleaned Data

From this it is clear that some of the keys are not destinations, but categories, files, etc., example the Mediawiki_templates file is clearly not a destination. More cleaning is involved to

remove such articles so as to obtain only travel destinations. An example of a travel destination is as follows:

```
In [25]: articles_dict['Abuja']

Out[25]: {'title': 'Abuja',
          'ns': '0',
          'id': '134',
          'revision': {'id': '3720833',
           'parentid': '3720826',
           'timestamp': '2019-02-05T11:58:02Z',
           'contributor': {'username': 'Ibaman', 'id': '195012'},
           'comment': 'Undo revision 3720826 by [[Special:Contributions/41.217.19.109|41.217.19.109]] ([[User talk:41.217.19.1
         09|talk]]) [[ttcf|useless for the traveller]]',
           'model': 'wikitext',
           'format': 'text/x-wiki',
           'text': {'@xml:space': 'preserve',
            '#text': "{{pagebanner|Abuja banner Aso Rock.jpg|caption=Aso Rock}}\n[[File:Moschee in der Hauptstadt Abuja.jpg|th
         umb|300px|right| Abuja National Mosque]]\n'''Abuja''' is the capital of [[Nigeria]]. Since most Nigerian government a
         gencies are now headquartered in Abuja and most other countries' embassies have been relocated from Lagos to Abuja, i
         t is a surprisingly expensive city.\n\n==Understand==\nAbuja is very beautiful. One of few purpose-built cities in th
         e world planned and built from scratch, it has an excellent road network, a beautiful rolling terrain and modern Nige
         rian architecture. However, power is often erratic.\n\n==Get in==\n\n===By plane===\n* {{listing | type=go\n| name=Nn
         amdi Azikiwe Airport | alt={{IATA|ABV}} | url=http://abuja.airport-authority.com/ | email=\n| address= | lat=9.006806
```

Figure 5.3: A Travel Destination

From the data above, there are still many unwanted characters that are not required in the model, therefore more steps are carried out to remove them. The process utilises python inbuilt functions such as regex, eg:

```
# get rid of https:
m = re.sub(r'(https?://\S+\ +\S+)', '', string)
```

Figure 5.4: Regex to remove noise from data

The final output is then transformed into a list of words. For instance, the article in figure 5.3 above is transformed to:

```
In [42]: print(convert_article_into_list_of_words(final_articles.get('Abuja', None)))

[['abuja', 'is', 'very', 'beautiful'], ['one', 'of', 'few', 'purpose', 'built', 'cities', 'in', 'the', 'world', 'plan
ned', 'and', 'built', 'from', 'scratch', 'it', 'has', 'an', 'excellent', 'road', 'network', 'a', 'beautiful', 'rollin
g', 'terrain', 'and', 'modern', 'nigerian', 'architecture'], ['however', 'power', 'is', 'often', 'erratic'], ['whil
e', 'the', 'industry', 'is', 'being', 'overhauled', 'and', 'aviation', 'safety', 'is', 'being', 'upgraded', 'only',
'few', 'local', 'airlines', 'are', 'reliable', 'aerocontractors', 'arik', 'air', 'and', 'chanchangi'], ['arik', 'ai
r', 'has', 'embarked', 'on', 'an', 'ambitious', 'programme', 'to', 'add', 'several', 'new', 'jetliners', 'including',
'the', 'new', 'boeing', 'dreamliner', 'to', 'its', 'fleet'], ['from', 'the', 'airport', 'you', 'are', 'best', 'advise
d', 'to', 'take', 'the', 'official', 'green', 'cab'], ['board', 'the', 'cab', 'with', 'only', 'people', 'you', 'kno
w'], ['uber', 'operates', 'here', 'as', 'well'], ['the', 'light', 'rail', 'line', 'when', 'change', 'opened', 'in',
'july', 'connects', 'the', 'airport', 'to', 'the', 'city', 'centre', 'at', 'the', 'abuja', 'metro', 'station'], ['yo
u', 'can', 'travel', 'to', 'abuja', 'by', 'bus', 'from', 'major', 'cities', 'like', 'lagos', 'benin', 'kano', 'and',
'port', 'harcourt'], ['reliable', 'services', 'include', 'abc', 'transport', 'with', 'air', 'conditioned', 'luxuriou
s', 'bus', 'rides', 'ekene', 'dili', 'chukwu', 'chisco', 'transport', 'ctn', 'and', 'young', 'shall', 'grow'], ['th
e', 'preferred', 'bus', 'service', 'would', 'be', 'abc', 'transport'], ['take', 'only', 'day', 'trip', 'buses', 'fro
m', 'lagos', 'or', 'kano'], ['a', 'few', 'buses', 'have', 'been', 'attacked', 'by', 'robbers'], ['if', 'youre', 'unfa
miliar', 'with', 'the', 'country', 'do', 'not', 'take', 'a', 'bus', 'without', 'an', 'escort'], ['buses', 'allow', 'y
ou', 'to', 'appreciate', 'the', 'terrain', 'the', 'towns', 'and', 'cities', 'and', 'the', 'subtle', 'changes', 'in',
'these', 'and', 'culture', 'as', 'you', 'drive', 'towards', 'the', 'capital', 'either', 'from', 'the', 'south', 'or',
'the', 'north'], ['car', 'or', 'taxi', 'is', 'the', 'main', 'mode', 'of', 'getting', 'around', 'abuja', 'public', 'tr
```

Figure 5.5: Formatted destination Data

This format is now ready to be input in the model for training and is therefore output into a file using the python library pickle.

5.3.3 Model Training

The algorithm utilises the Word2Vec model. The Word2Vec model was developed at google and it turns words into vectors in an attempt to learn their meaning. It is based on the assumption that it is possible to know the meaning of a word based on their surroundings. It uses the same concepts used by the n-gram NLP model.

Word2Vec is a two-layer neural network that processes text (Word2Vec, 2019). It contains a hidden layer weighting matrix that is also the vector representation of the words. The output of the model is a vocabulary where each word has been assigned to the vector. This output can be fed into deep learning models for further analysis or be used to establish the relationships between the words.

In this study, the file obtained from the data cleaning process is fed into the model for training. The geolocation data queried from the database mention earlier is also loaded. The model is trained using the data and upon completion, a search like the one below displays the data as shown.

```
In [*]: wordvec_list

In [*]: ms = model_bigrams.most_similar(positive=['paris','london','sevilla'], negative = [], topn=20)

In [31]: ms

Out[31]: [('lyon', 0.8221684098243713),
          ('frankfurt', 0.7805337905883789),
          ('lausanne', 0.7772169709205627),
          ('dijon', 0.7735536098480225),
          ('brussels', 0.7724005579948425),
          ('strasbourg', 0.770388126373291),
          ('toulouse', 0.7703643441200256),
          ('basel', 0.770146369934082),
          ('dusseldorf', 0.7695063352584839),
          ('madrid', 0.7682887315750122),
          ('lille', 0.7659573554992676),
          ('milan', 0.763250470161438),
          ('munchen', 0.7561508417129517),
          ('berne', 0.7543441653251648),
          ('geneva', 0.7535152435302734),
          ('bonn', 0.753292441368103),
          ('hamburg', 0.753288745880127),
          ('hannover', 0.7530144453048706),
          ('london_paddington', 0.7529534101486206),
          ('st_pancras', 0.7514403462409973)]
```

Figure 5.6: Results from the model

The result above represents the prediction of locations based on a user's search.

5.3.4 GeoLocation Mapping

The results from the model are then passed through a function to give them geo markers as follows:

```
[{'type': 'Feature',
  'geometry': {'type': 'Point',
   'coordinates': [Decimal('-0.21645000'), Decimal('5.54967000')]},
  'properties': {'title': 'ACCRA',
   'marker-color': '#9c89cc',
   'marker-size': 'large',
   'marker-symbol': 'rocket'}},
 {'type': 'Feature',
  'geometry': {'type': 'Point',
   'coordinates': [Decimal('38.76140000'), Decimal('9.01039000')]},
  'properties': {'title': 'ADDIS_ABABA',
   'marker-color': '#548cba',
   'marker-size': 'large',
   'marker-symbol': 'rocket'}},
 {'type': 'Feature',
  'geometry': {'type': 'Point',
   'coordinates': [Decimal('30.06055560'), Decimal('-1.96861110')]},
  'properties': {'title': 'KIGALI',
   'marker-color': '#63b6e5',
   'marker-size': 'large',
   'marker-symbol': 'rocket'}},
 {'type': 'Feature',
  'geometry': {'type': 'Point',
   'coordinates': [Decimal('28.28333300'), Decimal('-15.41666700')]},
  'properties': {'title': 'LUSAKA',
   'marker-color': '#b7ddf3',
   'marker-size': 'large',
   'marker-symbol': 'rocket'}},
 {'type': 'Feature',
  'geometry': {'type': 'Point',
   'coordinates': [Decimal('32.63333300'), Decimal('0.30000000')]},
  'properties': {'title': 'KAMPALA',
   'marker-color': '#c091e6',
```

Figure 5.7: Locations with Geographical markers

5.3.5 User Input

The web application is used to obtain the user input with their preferences as well as display the results. The user input represents the dynamic factors used to determine the destinations predicted for a user. These factors include weather, budget, languages spoken in the destination, and the preferred mode of transport within the destination at a given time. The input also takes into consideration other user preferences and anything the user would like to exclude from the search. These factors were selected based on research done in chapter 2 on what factors determine destination selection. As discussed in chapter 3, this was achieved through the use of the flask framework which is a light weight python framework. The figure below shows the simple form that takes the user input:

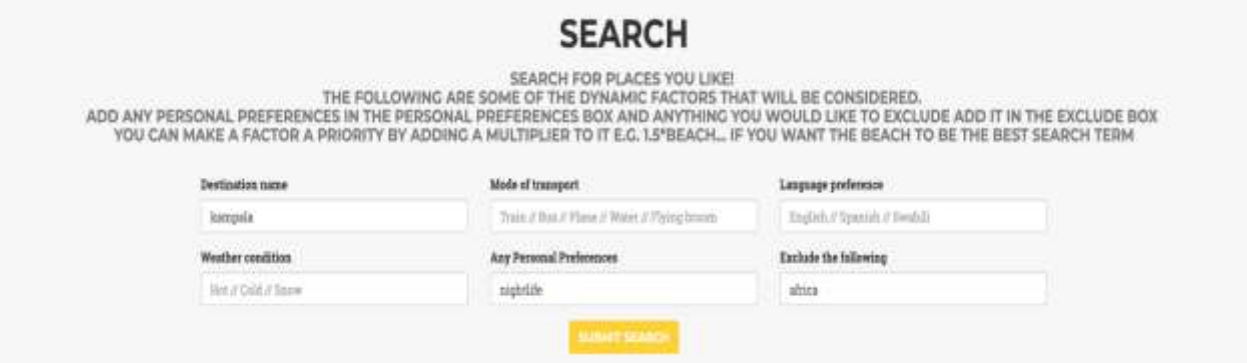

Figure 5.8: Web application interface to accept user input

The results are then displayed in a map with the location markers as follows:

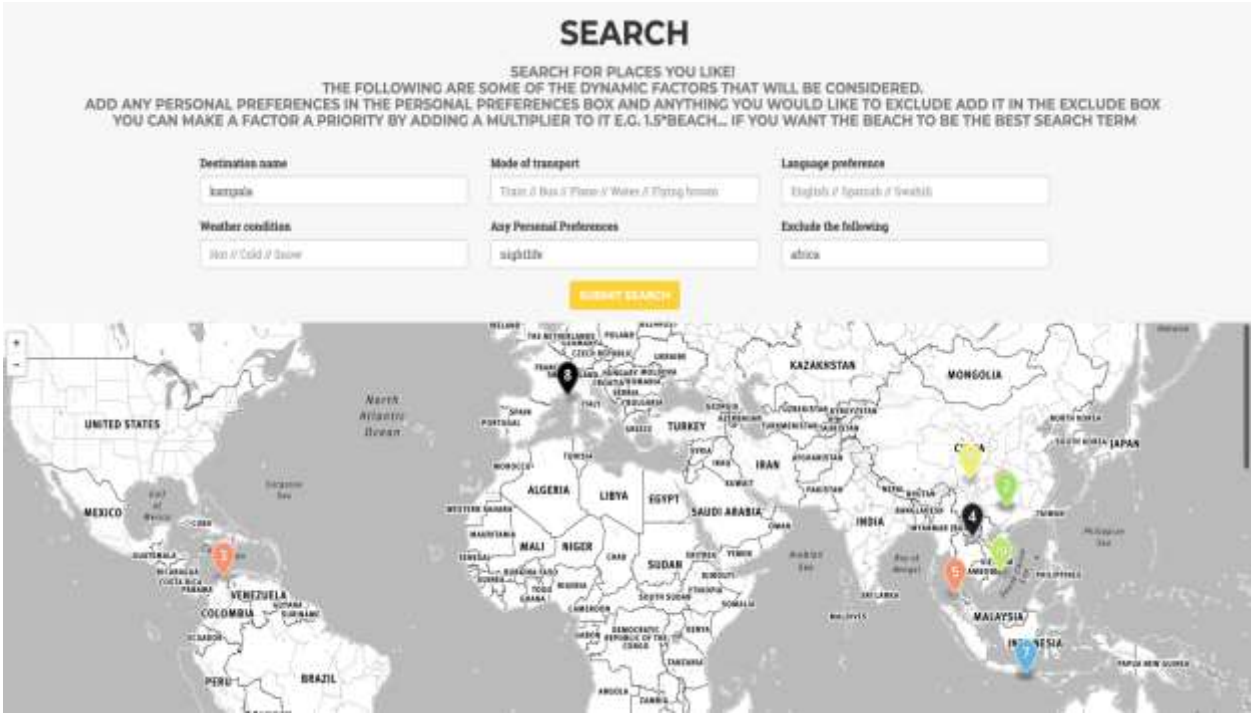

Figure 5.9: Map with search results

Each of the destination is ranked based on the cosine difference from the search made.

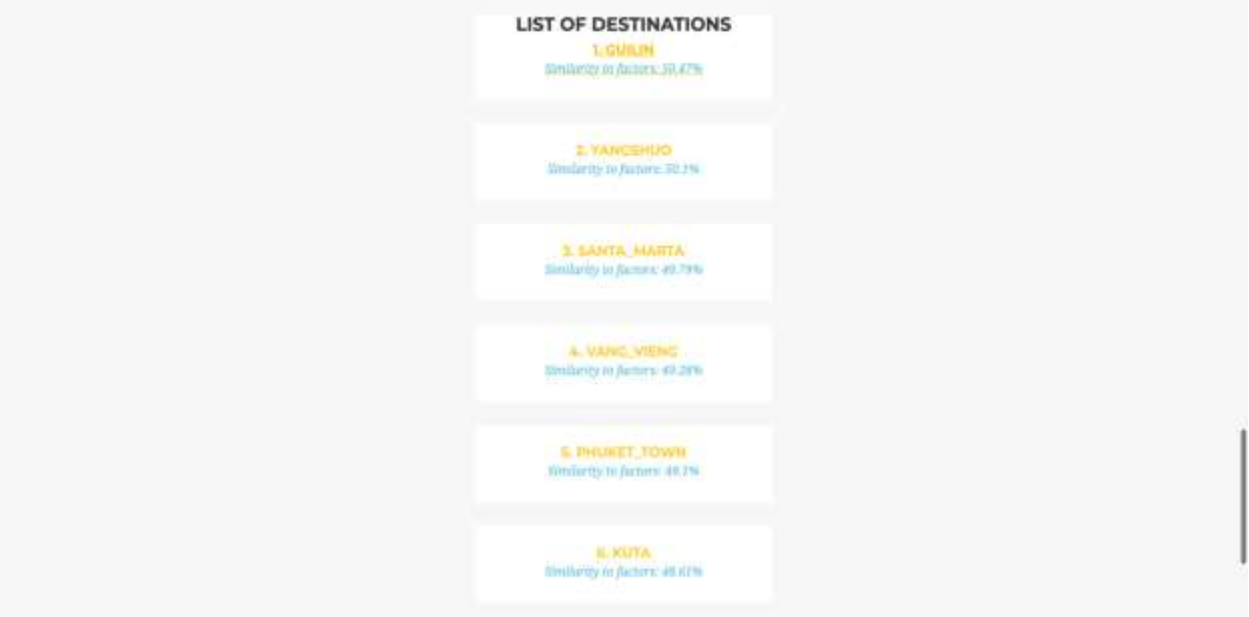

Figure 5.10: List of Predicted Destinations

The ranking is based on the understanding of the users keywords and the amount of data thaat was used to train the model.

A user can also view real-time additional information about the weather and the cost to stay in a place per day, among other details. Details such as racial tolerance and internet speed are available based on the city. Some remote cities may lack this details due to the scarcity of the data available.

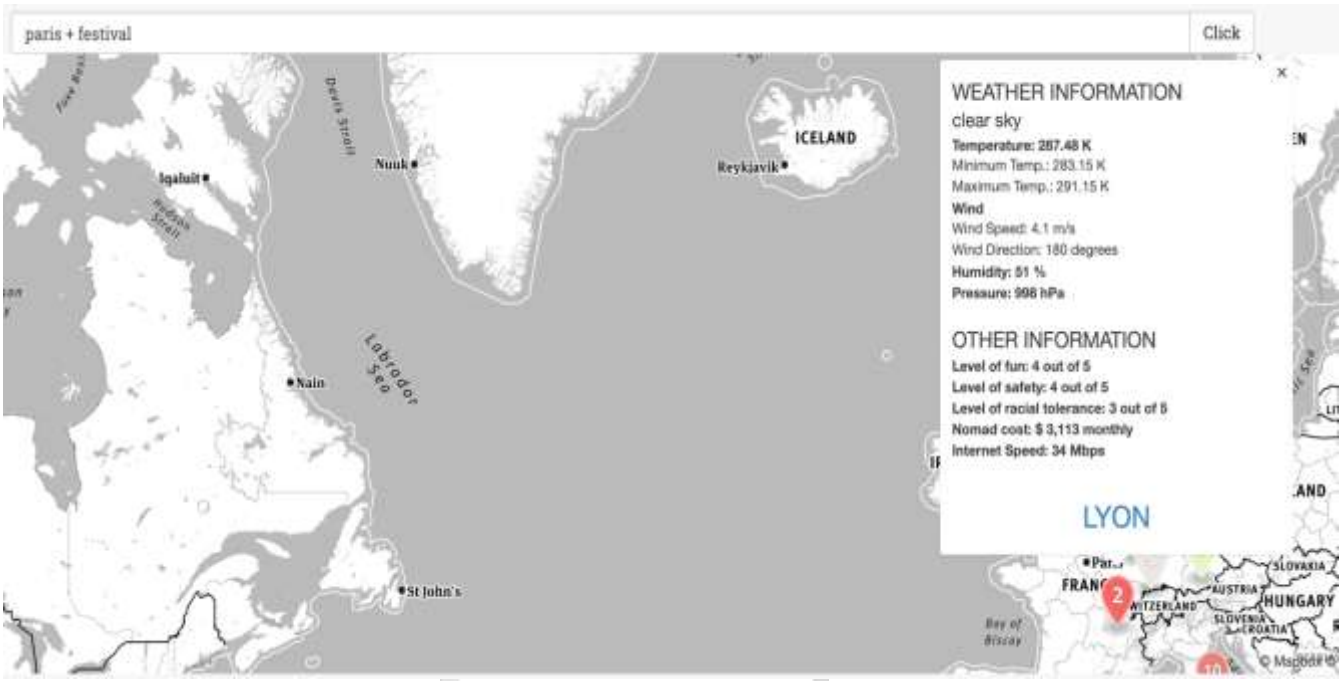

Figure 5.11: Selected Location with further information

When a user clicks on a city name, they are able to explore more details about a destination as shown below.

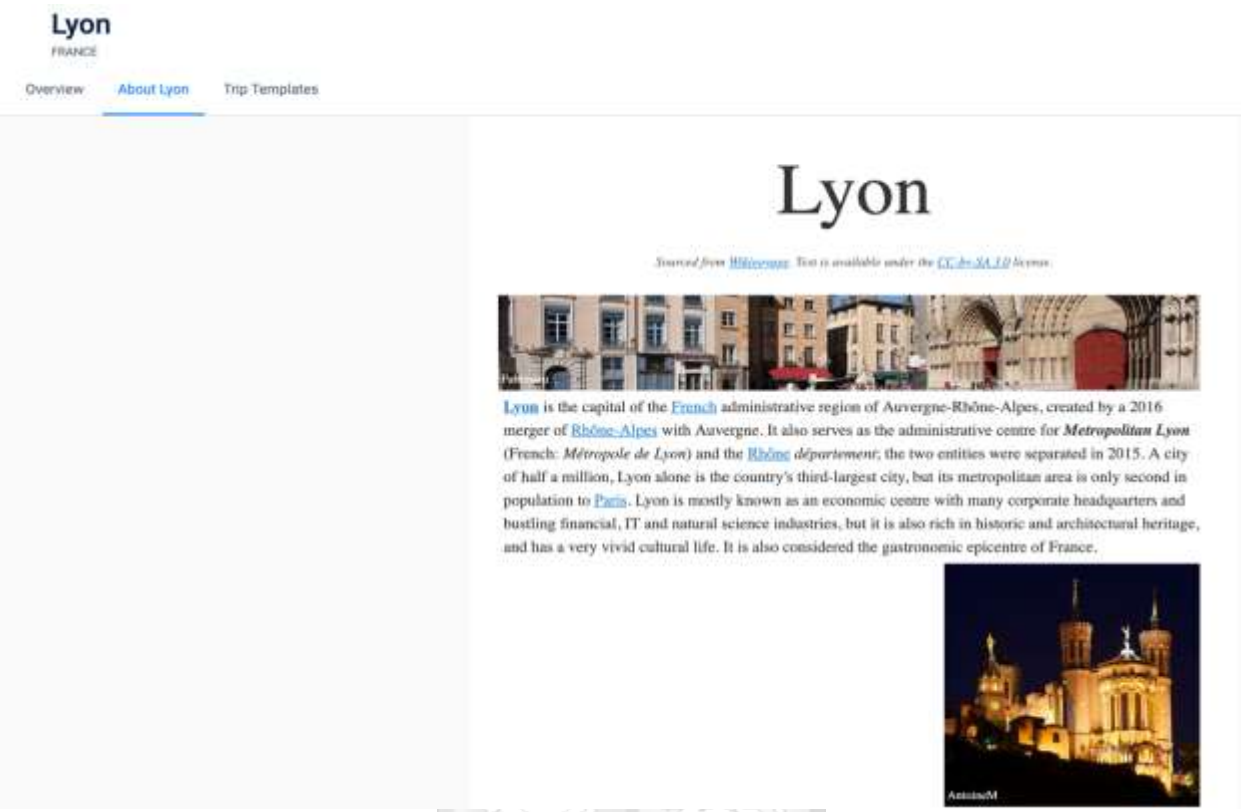

Figure 5.12: More details about a location

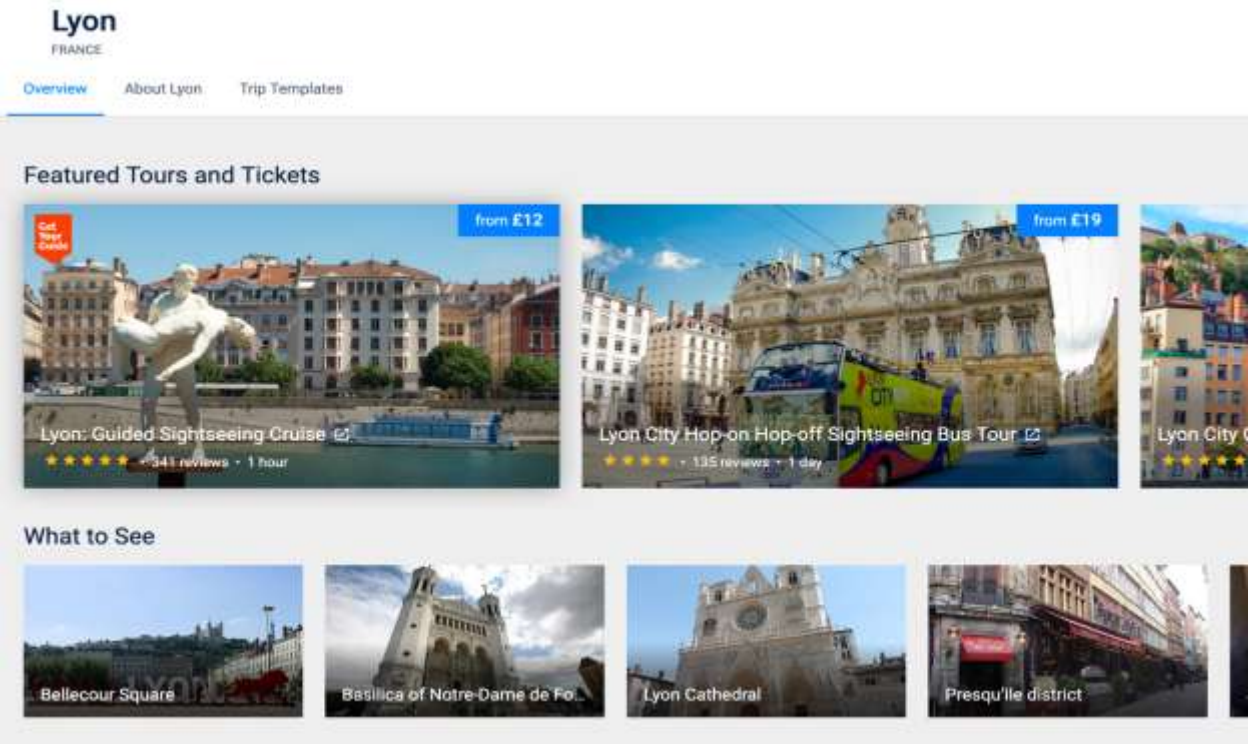

Figure 5.13: What to do or see in a location

5.4 Testing and Validation

The algorithm was tested by running multiple searches to determine the accuracy of the predictions. The output provides vectors from 0.0 to 1 to determine how close to the search each result is. For instance, a search for destinations similar to Nairobi but not in Europe like this:

```
In [13]: ms = model_bigrams.most_similar(positive=['nairobi'], negative = ['europe'], topn=20)
```

Figure 5.14: Testing the model

Returns the following results:

```
Out[15]: [('entebbe', 0.6250103712081909),
          ('kisumu', 0.6188884973526001),
          ('surat_thani', 0.6072408556938171),
          ('mwanza', 0.602759599685669),
          ('salaam', 0.5950684547424316),
          ('nakhon_ratchasima', 0.5912175178527832),
          ('arusha', 0.5898104906082153),
          ('nakuru', 0.5864733457565308),
          ('to/from', 0.5800447463989258),
          ('hat_yai', 0.5750609636306763),
          ('kangding', 0.5748545527458191),
          ('jerantut', 0.5747544765472412),
          ('catania', 0.5729742050170898),
          ('surigao', 0.5704734921455383),
          ('chumphon', 0.5688413381576538),
          ('luton', 0.568634033203125),
          ('maputo', 0.5670652389526367),
          ('from/to', 0.5658801794052124),
          ('oruro', 0.5649811029434204),
          ('luton_airport', 0.564949095249176)]
```

Figure 5.15: Test Results

The geo marking function filters out the non-locations and part of the final output is:

```
  'coordinates': [Decimal('36.69107000'), Decimal('-3.36692000')]},
 'properties': {'title': 'ARUSHA',
  'marker-color': '#000000',
  'marker-size': 'large',
  'marker-symbol': 'rocket'}},
{'type': 'Feature',
 'geometry': {'type': 'Point',
  'coordinates': [Decimal('36.06670000'), Decimal('-0.28330000')]},
 'properties': {'title': 'NAKURU',
  'marker-color': '#a3e46b',
  'marker-size': 'large',
  'marker-symbol': 'rocket'}},
{'type': 'Feature',
 'geometry': {'type': 'Point',
  'coordinates': [Decimal('100.46670000'), Decimal('7.01670000')]},
 'properties': {'title': 'HAT_YAI',
  'marker-color': '#eaf7ca',
  'marker-size': 'large',
  'marker-symbol': 'rocket'}},
```

Figure 5.16: Locations only with geo markers

The algorithm clearly identifies that the “to/from” is not a geographical location. This data is then fed into the web application and the user gets the output inform of a map. The figure below illustrated the results as viewed from a web browser.

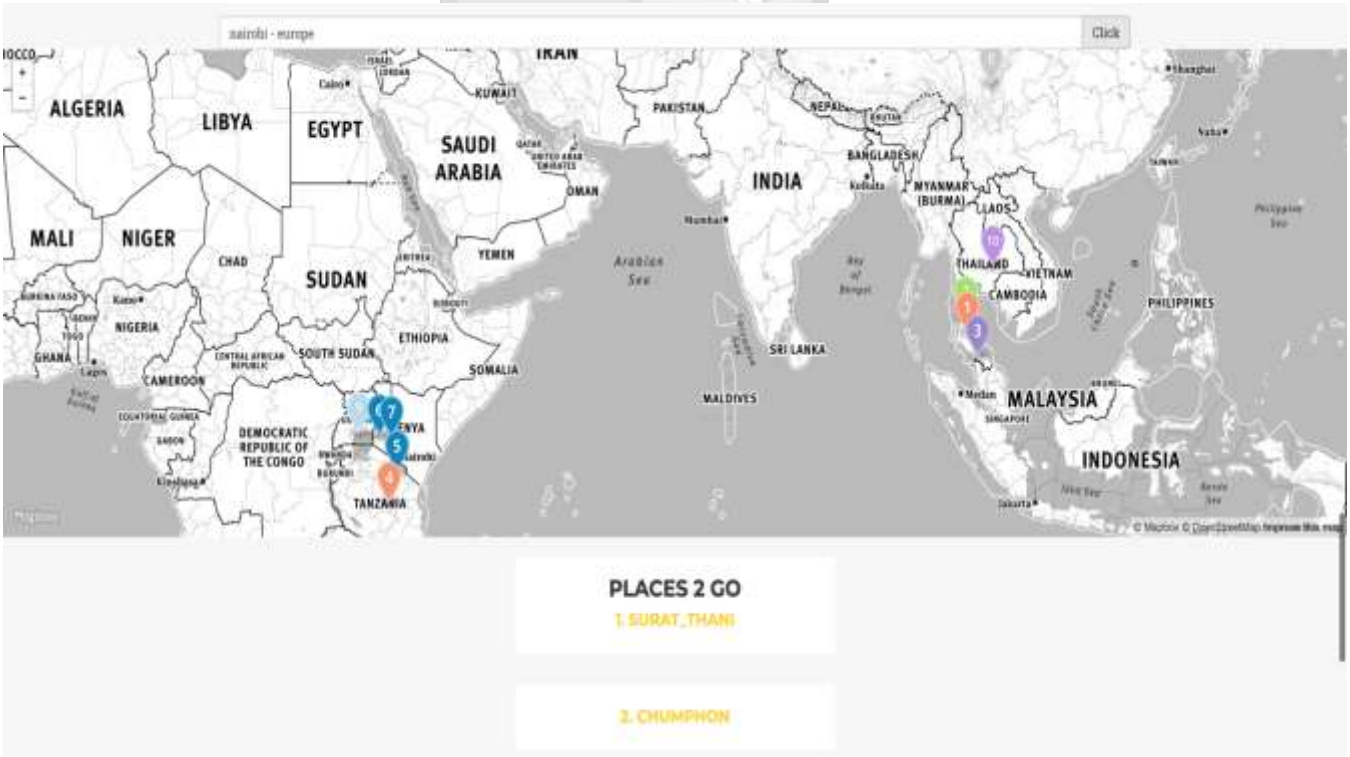

Figure 5.17: Final prediction result display

The aim of this research was to predict a list of destinations for a user based on their preferences. The section above details how that can be achieved through the use of an Artificial Neural Network model. The accuracy of the prediction can be improved over time by training the model using more data.

Chapter 6. Discussion

6.1 Introduction

This chapter discusses the results of the research in light of the research objectives, questions and aim discussed in chapter one. The aim of this study was to develop a travel destinations and route prediction algorithm that uses dynamic and personalised factors to provide a user with the most suitable destinations based on their preferences as well as the most optimal route they can use to navigate through those destinations.

The solution provided in the study as explained in chapter five, is an algorithm that takes in travel destinations and routes data from a crowd source, in this case, Wikivoyage, trains a model using the Word2Vec neural network model, to produce a list of destinations that matches a user's input.

6.2 Results of the study

The results of the study can be viewed in two ways. The first is the result of the model that provides a list of destinations for a user together with the geo-location markers. The second is the part that calculates the most optimum route for the user to follow. The study has been successful in providing a list of destinations to the user based on their personal preferences.

6.3 Challenges associated with the study

The major challenge with the study is the lack of a user's historical data. One of the major aspects of the study is to provide personalised predictions. In order to achieve this, there is the need to have a deep analysis of a user's likes and behaviour in a field called behaviour theory. Unfortunately, this was past the scope of the study and the only data that was relied on is what a user supplies on request.

6.4 Prediction Confidence

The predictions are provided through word2vec, a neural network model. This model has been proven to work great with text processing because it detects similarity in vector space. Its mathematical nature allows the model to achieve a certain level of accuracy that other NLP algorithms struggle to attain. It returns the results in a vector of cosine difference. For instance, when the model was used to detect the similar searches to "Sweden" it produced the following results. This means that the cosine distance between Norway and Sweden is 0.760124.

Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

Figure 6.1: Model showing vectors

Utilising this model for the study guarantees that destinations are better predicted instead of using plain text matching.

Comparison Between this Model and other NLP Models

There are many methods used in text processing as seen in chapter 2. However, they fail to capture the true meaning of a word. The following table shows various methods of text processing and why the study chose neural-network based embeddings as opposed to the other methods.

Table 6:1 NLP methods comparisons

Method	**How it works**	**Weakness**
Neural-network based embeddings	The neural network develops a deep vector representation of text, enough to know it meaning in different contexts.	The quality of the prediction depends on the corpus used to train the model.

TF-IDF	It measures the importance of a word in a particular document.	It is mainly used by search engines to know what results to give back in a search.
Bag of Words (BoW)	It analyses the word vector and checks the frequency it appears in a document.	It is mainly useful for document classification and not prediction.
One-hot encoded vectors	This method checks whether a word exists in a corpus or not.	It does not detect anything other than the presence or absence of a word

6.5 Research Shortfalls

The proposed study has a few limitations that may affect its validity.

i. The model utilises only data from Wikivoyage, while the data has thousands of articles about destinations, it does not provide the complete overview of a destination. In addition, the data maybe too formal lacking user descriptions that may shed more light to a destination.

ii. The study lacks historical user data which can infinitely improve the predictions. Unfortunately, this can only be achieved over time.

iii. The data from Wikivoyage does not cover many African destinations hence affecting accuracy of results based on African cities. This can be rectified by using a data source with as many destinations as possible.

Chapter 7. Conclusions and Recommendation

7.1 Conclusions

Over the years, travelling has become part of people's lives. From the one-time vacation, to the few weeks exploring, and the one year travelling sabbatical, many people have adopted the idea of travelling. As such, many solutions have cropped up to make the planning process and the travelling as seamless as possible. These solutions have provided maps that show people where they are and directions all around, recommendations for popular places often labelled "Top 10/100 destinations", or even user reviews about certain places.

However, many of these solutions either fail to answer the question "Where to go?" and those that answer usually answer in the form "Where have many people liked or been to before?". This study has proposed a solution that will answer the questions "Where do I go now based on my preferences?" by employing personalised and dynamic factors in the prediction. In a drive to answer those to key questions, this study proposed four objectives that stem from the aim. This research seeks to answer four main questions derived from the objectives.

This section will review the study, showing how each of the objectives were met in the study, and highlighting any recommendations and future work that were observed.

The first objective aimed at examining the different personalised and dynamic factors that affect destinations and route choices amongst travellers and highlight the effect of each factor in the final decision. The factors were outlined and discussed in depth in chapter two. The research also highlighted the need for more personalised solutions.

The second objective aimed and analysing the existing algorithms and techniques so as to understand the current situation as well as proof that there is indeed a gap that this research fill as its contribution.

The third objective was to develop the proposed solution with the factors proposed. The Travel Destinations and Routes Prediction Model was developed as a result utilising Artificial Neural Networks for text analysis. The solution proved to work as the results are provided purely based on a user's preferences.

The final objective was to prove that the proposed solution was valid. The discussions in chapter six above show why the solution was viable as well as highlighting a few shortfalls of the final solution.

In conclusion, this research has proven that there was a niche in the intelligent travelling solutions available and attempted to fill one of the gaps.

7.2 Recommendations

Based on the findings of this study, the following recommendations can be made:

i. The model should be trained with as much data as possible to allow for a broader field of prediction. The data should especially contain descriptions of the destinations from a regularly updated public source.

ii. The model can be extended to mine social media profiles so as to analyse a user's trends, likes and dislikes to allow more suitable recommendation.

7.3 Future Work

The scope of this research allowed the use of the neural network model in examining relationships between the search query and the provided data. In the future, deep learning can be employed to allow a more in-depth analysis of prediction. Other additions could include more parameters to provide accuracy. Factors such as population density could improve the reliability of the system.

References

Abrahamsson, P., Salo, O., Ronkainen, J. & Warsta, J. (2002). *Agile software development methods: Review and analysis*, VTT publication 478, Espoo, Finland.

Alivand, M., Hochmair, H., Srinivasan, S., (n.d). *Analysing how travellers choose scenic routes using route choice models.* Retrieved March 17, 2018 from https://pdfs.semanticscholar.org/2f5b/9019022c890a41b95542bb2f04a389fe5ef1.pdf

Amirgholya, M., Golshanib, N., Schneiderc, C., Gonzales E. J., Gao, O. H., (2017). An advanced traveller navigation system adapted to route choice preferences of the individual users. *International Journal of Transportation Science and Technology*, 6(4), 240-254.

Amrollahi, A. (2016). *A Process Model for Crowdsourcing: Insights from the Literature on Implementation*. Australasian Conference on Information Systems. Adelaide.

Bajaj, G., Agarwal, R., Bouloukakis, G., Singh, P., Georgantas, N., Izssarny, V., et. al. (n.d). *Towards Building Real-Time, Convenient Route Recommendation System for Public Transit.* Retrieved March 17, 2018 from https://itra.medialabasia.in/data/Documents/HumanSense/publications/Towards%20Building%20Real-Time,%20Convenient%20Route%20Recommendation%20System%20for%20Public%20Transit.pdf

Barrosa, A., P., Martínez, L., M. & Viegas, J., M. (2015). A new approach to understand modal and pedestrians route in Portugal. *Transportation Research Procedia*, 10, 860 - 869.

Bartle, C., Avineri, E. and Chatterjee, K. (2013). Personalised Travel Plans in the Workplace: a Case-Study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 16, 60-72.

Belcher, B., Rasmussen, K., Kemshaw, M., & Zornes, D. (2016). Defining and assessing research quality in a transdisciplinary context. *Research Evaluation*, 25(1), 1-17. DOI: 10.1093/reseval/rvv025

Blasco, D., Guia, J. & Prats, L. (2014). Tourism destination zoning in mountain regions: a consumer-based approach. *Tourism Geographies*, 16(3), 512-528, DOI: 10.1080/14616688.2013.851267.

Boyer, S. (2015). *5 Agile development best practices to Dodge Project Failures.* Retrieved on April 3, 2018 from http://www.nutcache.com/blog/how-to-dodge-project-failures-with-agile-practices/

Brown, P., deSouza, P., Mercer, R., Della Pietra, V, & Lai, J. (1992). Class-based n-gram models of natural language. *Comput. Linguist*. 18(4), 467-479.

Buecheler T., Sieg, J. H., Füchslin, R. M., & Pfeifer, R. (2010). *Crowdsourcing, Open Innovation and Collective Intelligence in the Scientific Method: A Research Agenda and Operational Framework.* Retrieved March 25, 2018 from https://mitpress.mit.edu/sites/default/files/titles/alife/0262290758chap123.pdf

Chen, C., Guo, B., Ma, X., Pan, G., Wu, Z., & Zhang, D. (2015). TripPlanner: Personalized Trip Planning Leveraging Heterogeneous Crowdsourced Digital Footprints. *IEEE Transactions on Intelligent Transportation Systems*, 16, 1259-1273.

Chen, S., Hu, H., Li, G., Li, W., Shen, B., Tan, K., Wang, H., & Wu, H. (2014). R3: A Real-Time Route Recommendation System. *PVLDB*, 7, 1549-1552.

Chidlovskii, B. (2017). Mining Smart Card Data for Travellers' Mini Activities. *Artificial Intelligence (cs.AI).* arXiv:1712.06935v1 [cs.AI].

Coast, D. A., Stern, R. M., Cano, G. G., & Briller, S. A. (1990). An approach to cardiac arrhythmia analysis using hidden Markov models, *IEEE Transactions on Biomedical Engineering*, 37(9), 826-836, doi: 10.1109/10.58593

Cronbach, L. J. (1997). Test "reliability": Its meaning and determination. Psychometrika, 12(1): 1-16.

Creswell, J., et al. (2003). Advance Mixed methods Research Designs. Handbook of mixed methods in social and behavioural research. 209-240.

Crowder, M. (2011). *Hidden Markov Models for Time Series: An Introduction Using R*, Second Edition by Walter Z, Iain L. MacDonald. International Statistical Review. 79. 132-133. 10.2307/41306185.

Dai, J., Ding, Z., Guo, C., & Yang, B. (2015). Personalized route recommendation using big trajectory data. *2015 IEEE 31st International Conference on Data Engineering*, 543-554.

Dijkstra, A. & Drolenga, H. (2008). Safety effects of route choice in a road network: Simulation of changing route choice. *Research in the framework of the European research programme In Safety,* R-2008-10, SWOV, Leidschendam, 2008.

Eby, D. & Molnar, L. (2001). Age-Related Decision Factors in Destination Choice for United States Driving Tourists. *Journal of Hospitality & Leisure Marketing.* 9. 97-111. 10.1300/J150v09n01_07.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence.* AI Magazine

Fred, B. B. (2015). *The Factors Influencing the Choice Of Tourist Destinations And Attraction Sites Among International Tourists.* MBA/CM/MZC/027/T.13. Retrieved November 10th, 2018 from http://scholar.mzumbe.ac.tz/bitstream/handle/11192/1144/MSc_MBA-CM_Brian%20B%20Fred_2015.pdf?sequence=1

Garcia-Molina, H., Joglekar, M., Marcus, A., Parameswaran, A. & Verroios, V. (n.d*). Challenges in Data Crowdsourcing.* Retrieved April 3, 2018 from https://web.stanford.edu/~verroios/papers/challengesDataCrowdsourcing.pdf

Goldberg, Y & Levy, O. (2014). word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. *Computation and Language. arXiv:1402.3722v1*

Google, (2018). The future of travel: New consumer behavior and the technology giving it flight. Retrieved on Thursday 4th October 2018 from https://www.thinkwithgoogle.com/marketing-resources/new-consumer-travel-assistance/

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* (3rd ed.). Morgan Kaufmann. ISBN 978-0-12-381479-1.

Henley Passport Index, (2019). Retrieved on January 16th, 2019 from https://www.henleypassportindex.com/passport-index

Holloway, J. C. & Humphreys, C. (2016). *The Business of Tourism 10th edn* (10th ed). Pearson Education Limited, Harlow, United Kingdom

Hsu, T. K., Tsai, Y. F. & Wu, H. H. (2009). The preference analysis for tourist choice of destination: A case study of Taiwan. *Tourism Management,* 30(2), 288–297. https://doi.org/10.1016/j.tourman.2008.07.011

Huang, Y.X. & Bian, L. (2009). A Bayesian network and analytic hierarchy process based

personalized recommendations for tourist attractions over the Internet. *Expert Systems with Applications* 36, 933–943.

Hussein, A., Croock, M., & Al-Qaraawi, S. (2019). Developed Crime Location Prediction Using Latent Markov Model. *Journal of Theoretical and Applied Information Technology*. 96(1).

Jansen-Verbeke, M. (1986). Inner-city tourism: resources, tourists and promoters. *Annals of Tourism Research*, 13(1), 79-100. DOI: 10.1016/0160-7383(86)90058-7.

Jawad, H. M. M. (2018). Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview. *International Journal of Advanced Computer Science and Applications.* 9(6). 10.14569/IJACSA.2018.090630.

Jokar, N., Honarvar, A. R., Aghamirzadeh, S., & Esfandiari, K. (2016). Web mining and Web usage mining techniques. *Bulletin de la Société des Sciences de Liège*. 85.

Jönsson, C. & Devonish, D. (2008) Does Nationality, Gender, and Age Affect Travel Motivation? a Case of Visitors to The Caribbean Island of Barbados, *Journal of Travel & Tourism Marketing*, 25(3-4), 398-408, DOI: 10.1080/10548400802508499

Jonathan, Z. (2013). K Means Clustering with Tf-idf Weights. *Advanced Computer Science and Applications.*

Kaplan B., Maxwell J.A. (2005) Qualitative Research Methods for Evaluating Computer Information Systems. Evaluating the Organizational Impact of Healthcare Information Systems. Health Informatics. Springer, New York, NY

Krichene, A. (2017). Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank, *Journal of Economics, Finance and Administrative Science*, 22(42), 3-24. https://doi.org/10.1108/JEFAS-02-2017-0039

Krogh, A., Larsson, B., Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. Edited by F. Cohen, Journal of Molecular Biology, 305(3), 567-580, ISSN 0022-2836, https://doi.org/10.1006/jmbi.2000.4315.

Kumar, A., & Singh, R. K. (2016). Web Mining Overview, Techniques, Tools and Applications: A Survey. *International Research Journal of Engineering and Technology (IRJET)*, 3(12).

Lin, N., Liu, H., & Gong, C. (n.d). *Research and Simulation on Drivers' Route Choice Behavior Cognition Model.* Retrieved March 17, 2018 from https://arxiv.org/pdf/1303.2764.pdf.

Livani, H., Jafarzadeh, S., Fadali, S., & Evrenosoglu, C. Y. (2014). Power system state forecasting using fuzzy-Viterbi Algorithm. *IEEE Power and Energy Society General Meeting*. 10.1109/PESGM.2014.6938837.

Long, J., Jia, J., & Xu, H., (2017). SenseRun: Real-Time Running Routes Recommendation toward Providing Pleasant Running Experiences. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17*).

Louangrath, P.T.I., (2014) Sample Size Determination for Non-Finite Population. *International Conference on Discrete Mathematics and Applied Science. University of Thai Chamber of Commerce (UTCC). Conference Proceedings*. Applied Science Section, Article No. 2.

Luong, M., Socher, R. & Manning, C. (2013). Better Word Representations with Recursive Neural Networks for Morphology. Computer Science Department Stanford University, Stanford.

Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. *Journal ACM*, 8(3), 404-417. DOI=http://dx.doi.org/10.1145/321075.321084

Mathew, W., Raposo, R., & Martins, B. (2012). Predicting future locations with hidden Markov models. *UbiComp '12 Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 911-918.

Mayaka, M., & Prasad, H. (2012). Tourism in Kenya: An analysis of strategic issues and challenges. *Tourism Management Perspectives*, 1(1), 48-56.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Computation and Language, v3. arXiv:1301.3781v3*

Pattekari, S.A., & Parveen, A. (2012). Prediction system for heart disease using naive bayes. *International Journal of Advanced Computer and Mathematical Sciences*, 3(3). 290-294.

Pel, A. J., & Nicholson, A. J. (2013). Network effects of percentile-based route choice behaviour for stochastic travel times under exogenous capacity variations. *Intelligent Transportation Systems - (ITSC), 2013 16th International IEEE Conference.*

Ravi, L., & Vairavasundaram, S. (2016). A Collaborative Location Based Travel Recommendation System through Enhanced Rating Prediction for the Group of Users. *Computational Intelligence and Neuroscience*, 2016, 1291358. http://doi.org/10.1155/2016/1291358.

Rothenbuehler, P., Runge, J., Garcin, F., & Faltings, B. (2015). Hidden Markov models for churn prediction. *2015 SAI Intelligent Systems Conference (IntelliSys)*. https://doi.org/10.1109/INTELLISYS.2015.7361220

Salfner, F. (2005). Predicting Failures with Hidden Markov Models. Research Paper from Humboldt University Berlin. Retrieved March 9th, 2017 from https://citemaster.net/get/8bd1acc0-f04b-11e3-bbaf-00163e009cc7/salfner05predicting.pdf

Sarma, M. K. (n.d). Destination Choice Pattern and Tourist Segments. *Tourist Behaviour: A Psychological Perspective* (Ed. Anita Raj), 137-149, Kanishka: New Delhi

Sebastia, L., García, I., Onaindia, E., & Alvarez, G. C. (2009). E-Tourism: A tourist recommendation and planning application. *International Journal of Artificial Intelligence Tools*, 18(5), 717-738. 10.1142/S0218213009000378.

Seyidov, J. & Adomaitiene, R. (2016). Factors Influencing Local Tourists' Decision-Making On Choosing A Destination: A Case Of Azerbaijan. *ekonomika 2016*, 95(3), Online ISSN 2424-6166. DOI: https://doi.org/10.15388/Ekon.2016.3.10332

Shearer C. (2000). The CRISP-DM model: the new blueprint for data mining, *Journal of Data Warehousing*, 4(5), 13—22.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., et al. (2013), Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, 34, 57-65. doi:10.1002/humu.22225

Sookocheff, K. (2015) Modeling Natural Language with N-Gram Models. Retrieved 14th March 2019 from https://sookocheff.com/post/nlp/n-gram-modeling/

Su, H., Zheng, K., Zheng, B., & Zhou, X. (2016). Landmark-Based Route Recommendation with Crowd Intelligence. *Data Science and Engineering*, 1(2), 86 - 100.

Tian, H., (2013). Travelers' Route Choice Behaviour in Risky Networks. *Open Access Dissertations*, 821.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2), 260-269, doi: 10.1109/TIT.1967.1054010

Weisberg, J. A. & Bowen, B. D. (1971*). Introduction to data Analysis*. 41.

Weischedel, et. Al. (2003). White paper on natural language processing. Speech and Natural Language: *Proceedings of a Workshop Held at Cape Cod Massachusetts*, 4, 481-493. 10.3115/1075434.1075526.

Winograd, T. (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. http://hci.stanford.edu/winograd/shrdlu/

Word2Vec. (2019). Word2Vec Archives. Retrieved 15th March 2019 from https://code.google.com/archive/p/word2vec/

Wu, H., Luk, R., Wong., K, & Kwok, K. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26 (3).

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, 44(1), 48–59. https://doi.org/10.1177/0165551516677946

Yang, W. & Hwang, S., (2013). iTravel: A recommender system in mobile peer-to-peer environment, *Journal of Systems and Software*, 86(1), 12-20, ISSN 0164-1212, https://doi.org/10.1016/j.jss.2012.06.041

Zhang, J., Chen, C., Xiang, Y., Zhou, W., & Xiang, Y. (2013). Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions. *IEEE Transactions on Information Forensics and Security*, 8(1), 5-15. doi: 10.1109/TIFS.2012.2223675

Appendix A: Plagiarism Report