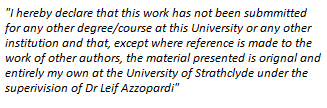


A Cost Model for Query Auto Completion in Information Retrieval.





Samantha Coll

CES Final Year Project

Supervisor: Dr Leif Azzopardi

Second Assessor: Dr David Harle

Date: 31/03/2017

# Abstract

In Information retrieval, query auto completion (QAC), denotes the resulting functionality: for any given prefix of characters entered into a search box, a list of projected alternative ways of extending the prefix into a full query, must be suggested by the user interface. [1] Most approaches to this issue require query logs to produce meaningful completion proposals. These proposals are ranked in order of importance to match a certain criteria, e.g. popularity, time, location, and the top *k* proposals are offered to the user as a completion suggestion. In this report, we explore how the ranking effects these query completions by developing a cost model that will show how users interact with the search bar on a website like Booking[[1]](#footnote-1). By analysing different datasets and exploring the different ways of ranking the QACs, shows the role that different parameters have on the algorithm and indicate a high quality of completion suggestions.

The cost model developed throughout the project can be found at:   
[www.github.com/samanthacoll22/PlacesProject](http://www.github.com/samanthacoll22/PlacesProject)

# Acknowledgment

*I would like to thank Dr Leif Azzopardi for enabling me to work on this interesting topic. A big thank-you also, for your guidance and extensive knowledge of the problem. Also, an extended thank you to Dr David Harle for his useful and insightful comments during the poster session.*

*Finally, my deep gratitude towards my family. Thank you for your encouragement throughout my studies, and especially this project.*

# Report Summary

Chapter 1: Introduction – A summary of aims to be discussed throughout report.

Chapter 2: Project Planning and Organisation – An analysis of the overall organisation of the project and the time management it was carried out in.

Chapter 3: Project Context – A look into the background of the project and the different aspects it contains.

Chapter 3: Design – A technical look at the programmatic design of the application.

Chapter 4: Testing Strategies – A technical look into whether the projects outcomes were as expected.

Chapter 5: Performance Analysis & Evaluation – A study into the consistency of the project as the parameters are changed and an evaluation of the aims.

Chapter 6: Future Work – Discussion and research into ways the project could be extended and the effects this would have on the project

Chapter 7: Conclusions - A mirror to Chapter 1, covering which of the aims the project has met and which aims remain to be completed.

Contents

[Abstract i](#_Toc478451864)

[Acknowledgment ii](#_Toc478451865)

[Report Summary iii](#_Toc478451866)

[1. Introduction 1](#_Toc478451867)

[1.1. QAC Problem Description 1](#_Toc478451868)

[1.2 Aims and Objectives 2](#_Toc478451869)

[2. Project Planning and Organisation 3](#_Toc478451870)

[2.1. Original Project Plan 3](#_Toc478451871)

[2.1.1. Research 3](#_Toc478451872)

[2.1.2. Implementation 3](#_Toc478451873)

[2.1.3. Testing 4](#_Toc478451874)

[2.1.4. Documentation 4](#_Toc478451875)

[2.2. Final Project Plan 4](#_Toc478451876)

[2.2.1. Research 4](#_Toc478451877)

[2.2.2. Design 5](#_Toc478451878)

[2.2.3. Implementation 5](#_Toc478451879)

[2.2.4. Testing 5](#_Toc478451880)

[2.2.5. Write-Up 5](#_Toc478451881)

[2.3. Organisational Tools 5](#_Toc478451882)

[2.3.2. Version Control 6](#_Toc478451883)

[2.3.3. Documentation 6](#_Toc478451884)

[3. Project Context 7](#_Toc478451885)

[3.1. Background 7](#_Toc478451886)

[3.2. Query Auto Completion Using Query Logs 9](#_Toc478451887)

[3.3. Query Auto Completion without Using Query Logs 9](#_Toc478451888)

[3.4. Cost Models 10](#_Toc478451889)

[3.4.1. Finding the Cost of a Place 10](#_Toc478451890)

[3.4.2. Determining the Ranking Cost. 11](#_Toc478451891)

[3.5. State Diagrams 11](#_Toc478451892)

[3.6. Probability Distribution 12](#_Toc478451893)

[3.7. Sampling 13](#_Toc478451894)

[3.7.1. Monte Carlo Simulations 14](#_Toc478451895)

[3.8. Ranking of Suggestions 15](#_Toc478451896)

[3.9. Software Engineering 16](#_Toc478451897)

[3.9.1. Architectural Style 16](#_Toc478451898)

[3.9.2. Design Patterns 17](#_Toc478451899)

[4. Design 21](#_Toc478451900)

[4.1. State Machine 21](#_Toc478451901)

[4.1.1. State Diagram 23](#_Toc478451902)

[4.1.2. Determining Probabilities 24](#_Toc478451903)

[4.1.3. Determining Costs of Each State 27](#_Toc478451904)

[4.2. Java Application 29](#_Toc478451905)

[5. Testing Strategies 34](#_Toc478451906)

[5.1. Full System Testing 34](#_Toc478451907)

[5.2. Unit Testing 35](#_Toc478451908)

[5.3. Sub-System Testing 35](#_Toc478451909)

[5.4. Parametric Simulation 35](#_Toc478451910)

[5.5. Black Box Testing 36](#_Toc478451911)

[6. Performance Analysis and Evaluation 37](#_Toc478451912)

[6.1. Ranking of Suggestions 37](#_Toc478451913)

[6.2. Instance of Suggestions Displayed 40](#_Toc478451914)

[6.3. Amount of Suggestions Displayed 42](#_Toc478451915)

[6.4. Typing Cost 44](#_Toc478451916)

[6.5. Running Cost Model over Different Processors 45](#_Toc478451917)

[7. Future Work 48](#_Toc478451918)

[7.1. Trie Structure 48](#_Toc478451919)

[7.2. Developing a Native Application 50](#_Toc478451920)

[7.2.1. Android vs iOS 50](#_Toc478451921)

[7.2.2. Location Services 51](#_Toc478451922)

[8. Future Applications 55](#_Toc478451923)

[8.1. Without Extension 55](#_Toc478451924)

[8.2. With Extention 55](#_Toc478451925)

[8. Conclusions 56](#_Toc478451926)

[8.1. Overall Reflection of the Project 56](#_Toc478451927)

[Appendices 57](#_Toc478451928)

[Appendix 1: Original Project Plan 57](#_Toc478451929)

[References 59](#_Toc478451930)

# Introduction

Web searching is a tedious and omnipresent activity which is carried out by millions of people on a daily basis. However, with the aid of QAC, an element which tries to predict the user’s query and offers a list of proposed suggestions, this time-consuming task could be reduced by minimising the amount of input needed from the user. This is possible by keeping the amount of keystrokes to a minimum by helping form a query. The primary objective for effective QAC is to provide the user with their intended query after the fewest possible keystrokes, and at the highest possible rank in the list of completion suggestions. A brief look into the problem faced with when using QAC can be seen in section 1.1, whilst section 1.2 outline the aims and deliverables of the project.

## QAC Problem Description

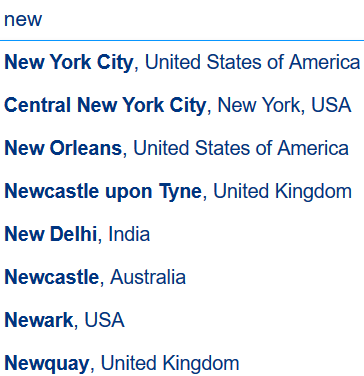
QAC is an algorithm that takes a list of possible query completion suggestions, say Qs, and then ranks these suggestions in order of importance before showing the top *k* completions to the user. This can only be done by taken the array of characters that the user has input, say qi, to form this list of proposed completions. Then the user will get to decide whether a suggestion (preferably the highest ranked), is a match. Some approaches, like Google[[2]](#footnote-2) and other major search engines, can account for spelling mistakes and suggest what they believe the query was meant to be. For instance, if the string “David Attenbur” is inputted by a user as their qi, then the top query completion presented would be the intended string “David Attenborough”. Other QAC approaches support mid-string completions, where the query suggestion “Kurt Cobain” would be returned if only the string “Cobain” was entered. For this report however, spelling correction and mid-string completion are not necessary as the main aim is to build a cost model which shows how a user handles QAC. Therefore, the full query will be coded into the simulation and checked with query logs before proceeding, meaning it is compulsory that qi is a prefix of all Qs.

## 1.2 Aims and Objectives

The main aims and deliverables of the project are:

* To create a cost model of how people interact with the search box on a website like Booking, which can be seen in Figure 2, where people start to type in letters and the system predicts where they might want to go.
* To evaluate the impact of ranking on query autocomplete and how it influences the goal of finding the minimum cost.
* To evaluate the number of suggestions the user is presented with and the effect it has on the cost model.
* To evaluate the outcome of generating suggestions at different instances and how they assist to keep cost to a minimum.

An additional aim of the project is to create a Java application that will display various destinations that the user will type in at their own speed. The time taken for the user to type the destinations will be noted and an average time for each letter will be generated. This will then be implemented into the cost model for each user, giving a more realistic, personal result.



***Figure 2: Output from user typing “new” into Booking.***

# Project Planning and Organisation

To guarantee the project could be completed by the final deadline, knowledge of the status of the project was key. To guarantee success, the current status of the project at all points was confirmed by the creation of a project plan. The project plan meant that if the project was destined to be incomplete for a particular reason, then this would become apparent with enough time to allow changes to rectify the problem.

## Original Project Plan

The original plan was created by splitting the project into four main sections: Research, Implementation, Testing and Documentation. Each of these sections were then sub-divided into smaller, more manageable tasks, which could be tackled individually. Two of these sub-tasks came with strict deadlines: Poster Creation (27th November 2016) and Report Write-up (31st March 2016), which dictated deadlines to support the sections. The creation of the poster required the research to have been exhausted, which confirmed that section complete. The report covers every aspect of the project, guaranteeing each section has been fully covered by that date. The full original plan is located in Appendix 1: Original Project Plan.

### Research

Research was undertaken into the way query auto completion works. A detailed look into how the user jumps between typing a letter, to looking at the list of suggestions, to then picking from the list, and the probability of each state occurring. A decision was made that seven days was a suitable amount of time to create the poster which assisted in creating a deadline for the end of December for the research section to be competed. The time available was then split amongst the sub tasks within research.

### Implementation

The Implementation of the system was to be completed by the end of February, thus allowing a carefully allocated, suitable amount of time for the final overall system to be tested and the report to be written before the final deadline. The time available between creating the poster and the deadline set for Implementation gave the overall time available. This time was taken and divided amongst these main tasks based on the perceived level of difficulty they carry.

### Testing

Four main aspects of testing were anticipated to be employed throughout the project: Unit Testing, Sub-System Testing, Full Testing and User Trials. Ideally, the testing portion of the report was to be carried out in parallel with the Implementation section. Unit Testing would be performed as each small section of the cost model was created, with Sub-Section testing being carried out at the end of each new feature and User Trials carried out just before writing the final report. Leaving only Full System Testing as the only part of testing carried out independently of the other sections. It was decided that two weeks would be a sufficient amount of time to satisfy this.

### Documentation

The documentation section consists solely on two set deliverables for the project. Strategically, the timescales that are allocated to these deliverables have been greatly exaggerated to ensure the project’s completion before the deadline. Creating a buffer like this allows expansion for the other sections, if necessary, whilst still ensuring the deliverables completion before their respective deadlines.

## Final Project Plan

As the project progressed the initial project plan became impossible to follow precisely. As problems occurred, and the aims of the project evolved, the project plan was continually updated to reflect the current state of the project. The final project plan is available in Appendix 2 Final Project Plan.

### Research

The research section was largely unchanged from the original plan. The only extra research that was carried out was to source a significantly larger amount of places/cities. This makes the cost model much more accurate, as originally, there were only 125 cities being loaded into the simulations. Therefore, after the third letter was typed the user was frequently presented with their destination as the first suggestion. This resulted in them selecting it the majority of the time, when in reality they may have been forced type again or look further down the list. There are now currently around 3000 places/cities which result in far better simulations.

### Design

The design of the report was largely unchanged from the original plan. The state machine was derived as anticipated. The only main significant change in design was that originally a webpage was to be created to find the users personal typing speed. This was then simplified to a Java application, reducing the need to employ a new programming language into the project. The design and research sections of the project were largely dependable on one another, meaning that at parts they were carried out simultaneously.

### Implementation

### 2.2.4. Testing

Testing was completely different from first anticipated. As development began it seemed logical to perform Full System testing as the implementation was undergoing. The Unit and Sub-System testing were carried out before the completion of the cost model, although most of the code was completed. Finally, the decision to carry out Parametric Simulation testing was made, to confirm expected outputs from changing a single parameter whilst keeping all others constant. The addition of black box testing came as a last minute decision at the end to test certain aspects of the JavaApp. The testing section fit within the timescale outlined in the original plan despite the re-arrangement and addition of tests.

### 2.2.5. Write-Up

The time taken to complete both the report and poster was largely as anticipated during the original project plan. The only change occurring in this section was to allow other unfinished work to run alongside it. Therefore, stages of implementation and testing ran in parallel to writing the report. This was greatly expected when the original plan was drawn up and was the sole reason for awarding the write-up so much time.

## Organisational Tools

In order to complete the project, various organisations tools were employed to keep within the timescale outlined in Chapter 3.2 Final Project Plan.

### 2.3.2. Version Control

Version Control is an important aspect of any project with a heavily weighted software component. For this project specifically, Git was utilised. Git offers a variety of benefits to a project, including the ability to work on different features, or sections, in parallel whilst on separate branches. This ability allows the changes on an individual feature to have no negative consequences or effects on the other. However, the most important and most relied on feature of Git, if the key functionality breaks, is the ability to revert to a previous version of the software. Allowing access to a functional version at a later date, as a form of back-up copy.

Git operates on a central server, where work can be uploaded to and downloaded from. The Git server used for this project was provided by Github[[3]](#footnote-3), which holds a copy of all work on their servers. This not only grants access to be worked on by multiple machines, but as long as the software is being pushed on a regular basis to the server, creates a safety net as large amounts of work cannot be lost if something were to happen to the work station.

### 2.3.3. Documentation

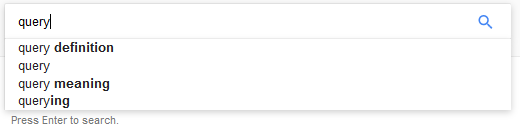
When dealing with a project of significant size, it is important to maintain a detailed record of work completed throughout. As the project was carried out, design and implementation notes have been documented in a logbook. Giving an insight to the thought processes throughout. Additionally, when software is pushed to Git, it is a requirement to detail any changes that have been made. These messages are known as a Git log and they show what has been altered and more importantly, why.

# Project Context

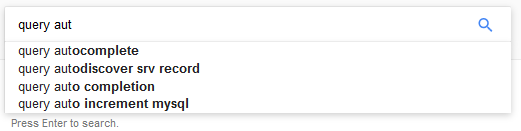
Initially, before any work could begin to take place, several key questions had to be answered about how the project should proceed. To answers these questions, research into the different approaches of QAC was initiated. Furthermore, investigations into several different aspects of the project were explored further.

## Background

The time-consuming and error-prone process of formulating and physically typing search queries is what can make query input laborious. In response to this problem, search engines have adopted the concept of query auto-complete (QAC) as a way or reducing the effort it takes to submit a query. QAC provides the user with a list of suggested queries as they begin to enter their query in the search box. The incomplete input of the user is often referred to as *query prefix* and the suggested queries are often called *query completions*. As shown in Figure 1, query completions often start with the characters that the user has entered into the search box (see Figure 1a) and are updated as the user continues to type (see Figure 1b). QAC helps the user to formulate their query when they have an intent in mind but not a clear way of expressing it in a query. It helps to avoid possible spelling mistakes, especially on devices which have smaller screens. The most common approach to QAC is to extract past queries with each prefix from a query log, and rank them by their past popularity [2]. Although this approach provides satisfactory result on average, it is hardly most favourable as it fails to take in clues such as time or user context which quite often influences the queries most likely to be typed. This report will focus on the development of a cost model that shows how a user interacts with a ‘search bar’ on a website. It will then explore these different approaches of QAC and show how each keep cost to a minimum. In addition to a minimal cost, cutting down the search time duration of users implies a lower load on the search engine, which results in savings in machine resources and maintenance [3].



**a) *A list of four query completions for the prefix “query”.***



1. **b) *An updated list of four query completions for the prefix “query aut”.  
   Figure 1: Examples of query auto completion.***

As the web continuously becomes a platform for real-time news and media, time consumption takes a vital role on the retrieval of information. Users turning to search engines for information about new or ongoing events, takes up the majority of daily queries [4]. However, importantly the use of QAC is not limited to the use or search engines, such as Bing, Google or Yahoo!. QAC has been integrated into other common services for search tasks. Facebook[[4]](#footnote-4) provides a QAC for finding friends by taking in mutual friends the user may share with others; it also assists users to formulate requests for searching through tweets on the social medium Twitter[[5]](#footnote-5); QAC is also used for product searches in online shopping stores, such as Amazon[[6]](#footnote-6) which will often provide up to 10 queries. In the concept of web search, research indicates that using the suggested queries can vastly improve user satisfaction, especially for informational queries [5]. The main focus for this report is how QAC is integrated into a travel metasearch engine like Booking, where user’s search for the broadest selection of accommodation in every corner or the world.

## 3.2. Query Auto Completion Using Query Logs

Most approaches to QAC are based on the exploitation of information provided by query logs. A common approach is to suggest queries on the basis of how often they have been observed in a query log. Context- or time-related information are then added to improve on the QAC results.

Whiting and Jose [12], for example, describe an approach in which they address the problem of providing query suggestions for both consistently and recently popular queries. To tackle this issue, they propose a variety of different algorithms. These include a sliding window approach where they only consider the popularity of queries over the last N days, a similar approach in which they vary the length N of the sliding window depending on how common the entered prefix is, and an approach in which they use short-range popularity trends in order to predict the current popularity of a query.

Bar-Yossef and Kraus [1] use recent queries of a user to produce more relevant completion suggestions. Their proposed algorithm suggests those queries from a query-log as completions for a user input, which are most similar to the user’s previously typed queries. In order to measure similarity, they expand each query by adding related terms. Queries are then represented as term vectors and their similarity is computed as the cosine similarity.

## 3.3. Query Auto Completion without Using Query Logs

Relatively few researchers so far have explored the possibilities of query-log-independent QAC. Among them are Bast and Weber [3, 4] who developed a search engine that features instant auto-completion. Completion suggestions are offered for the last incomplete word (as opposed to offering completion suggestions for the entire query the user might have in mind). The completion suggestions are terms occurring in the documents that their search algorithm yields, when searching for the preceding complete words of the query.

Bhatia et al. [5] propose a method that relies on an n-gram model built from the documents of the corpus that is being searched. N-grams extracted from these documents are used to complete a query prefix typed by a user. By using an n-gram model, this approach comes closest to what we are proposing in this thesis. A key difference is, however, that we use an abstract language model in which we insert specific entities. Moreover, the objective of the auto-completion is a different one: both approaches mentioned last aim mainly at finding certain information rather than predicting the user’s input, as is the case in this work.

## Cost Models

Cost estimation models are mathematical or parametric equations used to estimate the cost of a product or project. These estimations will never be exact as there are many variables involved in calculating a cost estimate, such as technical or human. Moreover, any process that is significantly human based will never be exactly accurate as humans are far too unpredictable and complex to be exact. Cost models function through the input of parameters that describe the attributes of the product or project in question. In the case of this report, the cost model is formed by generating the sum of the places. The equation for the cost model can be seen in “(1)”.

*Total cost = ∑ [cost (place) \* popularity (place)]* (1)

### 3.4.1. Finding the Cost of a Place

To generate the cost of a place 3 main factors must be known: the cost it takes for the user to type a single letter, then the cost of a user to look at the list of suggestions, and finally, the cost for the user to pick a suggestion. However, when the user is looking at the list suggestions there is a degree of uncertainty of what they might do next. This leads to a probability that the user might type and a probability that they may pick from the list. For the purpose of the cost model, it has been decided the user will only be presented with a list of suggestions after typing in a minimum of three letters.

Therefore, if example “Paris” is taken, it is known that it is necessary for the user to type “P”, “A”, “R”, which each carry a cost to type letter. So, right now our total cost for our destination is (3 \* typeCost). However, when the user is presented with the retrieved list of suggestions, will the user look at the list and pick? Or will they simply type another letter? This is where the system will have to prepare for all outcomes. The process for finding the fourth letter can be seen in “(2)”.

*(3 \* typeCost) + (probabilityType \* typeCost)   
 + ((1 – probabilityType) \* lookCost)* (2) *+ (lookCost + (probilityPick \* pickCost))*

### 3.4.2. Determining the Ranking Cost.

On the other hand, the popularity of a place can be determined as the United Nations Statistical Division (UNSD), publishing an annual book on world statistics helped us construct a more precise idea of countries popularity, based on its geographical size. A simple representation of the cost model can be seen in Figure 3.2.

Cost (Place)

Looking Cost

Typing Cost

Picking Cost

Ranking (Place)

Total Cost

Probability Pick

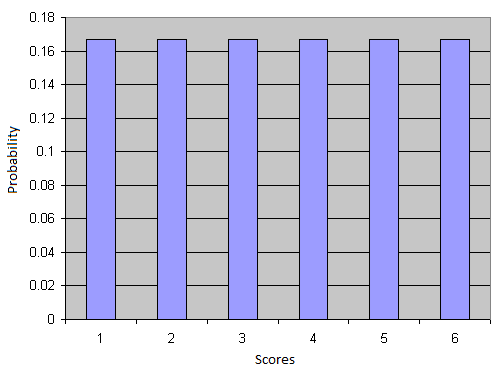
Probability Type

***Figure 3.2: A Simple Diagram of Key Parameters for Cost Model.***

## State Diagrams

A state diagram is used in computer science and fellow related fields to give an abstract description of the behaviour of a system. Specifically, a state diagram analyses the behaviour of a single object in response to a series of events occurring and takes into consideration all possible states that object could find itself in. The cost model will generate four states: a type state, a look state, a pick state and a search/end state. These will each be explore in detail in the design section of the report.

## Probability Distribution

Probability distribution in probability theory and statistics, is a mathematical function which provides the probability of occurrence of various possible outcomes in an experiment. If the notion of a coin toss is taken, the probability distribution for the outcome would be, 0.5 for ‘heads’, and 0.5 for ‘tails’. Probability distribution is normally separated in two types; continuous probability distribution and discrete probability distribution. The simulation developed invokes the latter, discrete probability distribution, where a random variable was exactly equal to some encoded value, which is known as a probability mass function. An example of the discrete probability of a dice roll can be seen in Figure 3.4, where since there is six possible outcomes then the chances of each individual, unique, outcome happening would be.

***Figure3.4. Probability Distribution of a Dice Roll***

Thus, relating to the problem at hand and the user is searching a destination. Values have to be set to the likelihood that they may type another letter into the search bar and the likelihood of them picking from the list of suggestions. A detailed explanation of how the value of these probabilities are elected can be seen in the design section of the report.

## Sampling

The process of sampling takes a predetermined number of observations from a larger population and carries out analysis on the smaller group. To take every city/place in the world and run the cost model could take days, or maybe even weeks to carry out. It is then far more time-efficient and economical to take a sample of the data. The simulation to run the cost model has around 3000 cities/places loaded into it. These are the cities in the world that have a population of over 100,000 people. Therefore, sampling is already being carried out before generating any total cost. Originally, 319,000 cities/places were used for the simulations which would give more accurate results, but the waiting time to load was around 10 minutes before the simulator proceeded. For that reason, a smaller amount of cities was managed. The advantages and disadvantages of sampling can be seen in Table 1.

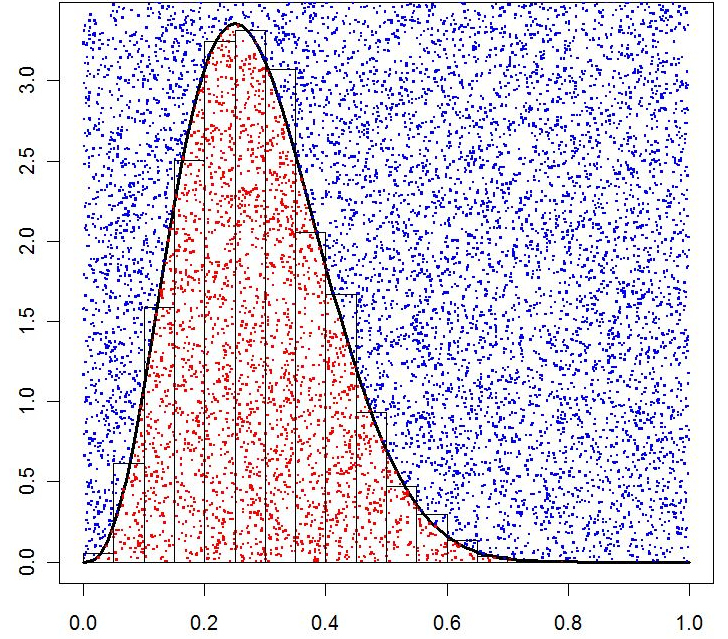
***Table 3.4: The Advantages and Disadvantages of Sampling***

|  |  |
| --- | --- |
| Advantages of Sampling | Disadvantages of Sampling |
| * Low cost of sampling. If data was being collected for the every place in the world, the cost would be extremely high. | * **Chance of bias.** Since specific data is predetermined, this data could favour some places more than others. This could draw erroneous conclusions, e.g. more well-known places are perhaps more likely to get selected. |
| * Less time consuming. Could take days to run the simulation for all places. Sampling gives us an understanding, in a healthier amount of time. | * **Difficult to select truly representative sample.** For instance, favouring higher ranked places. This will give a lower cost than lower ranked places. Meaning the lower ranked destinations are not represented truly and the total cost may be lower than it should be. |
| * Organisation of convenience. By taking a selection of places, it makes it easier to handle and less chance of error along the way. |  |

### Monte Carlo Simulations

Monte Carlo methods are a wide-ranging class of computational algorithms that depends solely on recurring random sampling to acquire a range of possible numerical results. Their fundamental purpose is to use randomness to find a solution to a problem that might be deterministic. These methods are normally used in three distinctive problem classes:optimisation, numerical integration, and in the case of this project, generating draws from a probability distribution.

Monte Carlo simulations take an action choice and display the user with a wide range of results and the probability of them occurring. It shows the extreme possibilities (a user typing every letter of the desired place), whilst highlighting all middle-of-the-road decisions (the user picking their destination once presented in the suggestions). Monte Carlo simulation of each place will allow a mean and standard deviation to be calculated which will be compared to simulation with different parameters. An example of a Monte Carlo simulation can be seen in Figure 3.4.1.



***Figure 3.4.1: A Monte Carlo Simulation***

## Ranking of Suggestions

In essence, query completions are ranked according to predetermined criterion. Therefore, when a prefix is provided, some of this criterion is then returned to the user. Normally, a precomputed auto-completion system is necessary for determining QAC which correspond to each individual prefix in advance; storing these associations in data structures, such as prefix trees. Figure 3.5 shows a simple QAC framework.



***Figure 3.5: Simple QAC Framework***

When a user enters a prefix in the search box, it is shown in Figure 3.5, that a basic list of suggestions is retrieved. Based on further re-ranking using signals determined at query time, e.g. location, season, time, the user will then receive a finalised list of query completions. Mainly these completions that are returned have a limited length, for instance, 5 for Booking and 4 for Google. For this project a list of cities/places will be searched and the ones that match the prefix of the desired place will be placed into a separate list. This list will have no form of structure and will simply be outputting the first suggestion it comes across. One of the parameters for the cost model will allow the user to decide how they would like their suggestions to be ranked. Either, alphabetically, reverse-alphabetically or by population, depending on their desired destination.

## Software Engineering

To develop and ensure an easy maintainable project it is vital to try conduct good software engineering practices throughout. By using well-structured software engineering practices, they should assist in making testing easier as the software should be more flexible and subject to change. The CS409 Software Architecture and Design at Strathclyde University helped pin point various designs and patterns used in the project.

### Architectural Style

An architectural style improves partitioning and promotes design reuse by providing solutions to frequently recurring problems, making this the highest level programming choice to be made in a project. They are thought of as sets of principles that shape an application and they determine how much data and functionality should be separated. Since the decision was made for the project is to be developed using Java, as this was the most familiar programming language, the architectural design that best suits would be Object-Orientated architectural style. Therefore, this is the primary style used throughout.

#### Object-Oriented

Object-oriented architecture style is based on dividing up tasks a system may have into individual, independent, reusable objects, which hold the data and behaviour relevant to that object. This style interprets the system as a series of banding objects, instead of a set of routines or procedural calls. This approach evokes a high level of abstraction, which means if one section of code is altered then it would not dramatically affect the programs other sections. Objects are discrete, self-sufficient and loosely coupled. They communicate with one another through interfaces, by calling methods or accessing properties in other objects. Creating an interface gives access to publically available method calls that additional classes may use to interconnect with any class implementing that interface. The key advantages of the object-oriented architectural style are:

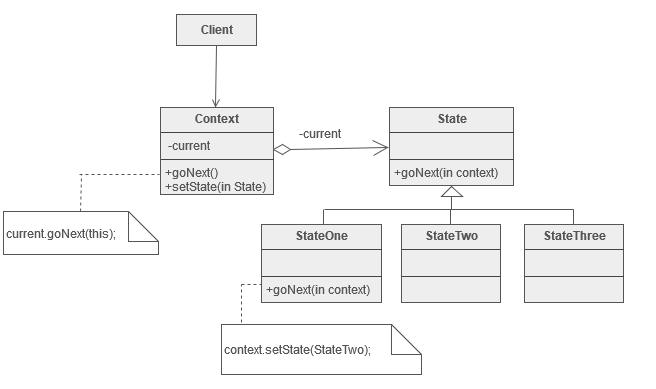
* + **Understandable**. It maps the application more closely to the real world objects, making it more understandable.
  + **Reusable**. It provides for reusability through polymorphism and abstraction.
  + **Testable**. It provides for improved testability through encapsulation.
  + **Extensible**. Encapsulation, polymorphism, and abstraction ensure that a change in the representation of data does not affect the interfaces that the object exposes, which would limit the capability to communicate and interact with other objects.
  + **Highly Cohesive**. By locating only related methods and features in an object, and using different objects for different sets of features, you can achieve a high level of cohesion

### Design Patterns

Design patterns are one of the most valuable tools for developers. They illustrate the best design solutions that others have encountered, and allow you to apply the same principle to your own designs. More importantly, knowing design patterns gives a common vocabulary for software developers to use when talking about their designs. Design Patterns are essentially proven solutions to commonly recurring problems. The use of design patterns provides the means to deal efficiently with such problems. Below is an overview of each of the design patterns which have been used throughout the project.

#### 3.7.2.1 State Pattern

The state pattern is a behavioural software design pattern that implements a state machine in an object-oriented way. With the state pattern, a state machine is implemented by implementing each individual state as a derived class of the state pattern interface, and implementing state transitions by invoking methods defined by the pattern's superclass. The state pattern can be interpreted as a strategy pattern which is able to switch the current strategy through invocations of methods defined in the pattern's interface. This pattern is used in computer programming to encapsulate varying behaviour for the same object based on its internal state. This can be a cleaner way for an object to change its behaviour at runtime without resorting to large monolithic conditional statementsand thus improve maintainability. The UML diagram of a state pattern cab be seen in Figure 3.6.2.1.



***Figure 3.6.2.1: UML Diagram of State Pattern***

The main advantage in the State design pattern, an object's behaviour is the result of the function of its state, and the behaviour gets changed at runtime depending on the state. This removes the dependency on the if/else or switch/case conditional logic. With State pattern, the benefits of implementing polymorphic behavior are evident, and it is also easier to add states to support additional behavior. The State design pattern also improves Cohesion since state-specific behaviors are aggregated into ConcreteState classes, which are placed in one location in the code. With the State design pattern, it is very easy to add a behavior by just adding one more ConcreteState class. State pattern thus improves the flexibility to extend the behavior of the application and overall improves code maintenance.

#### 3.7.2.2 Observer Pattern

The Observer Design Pattern is essentially, when one component allows others implementing a specific interface to register for updates. Whenever this first object recognises a change of some sort it notifies all of those listening for updates. Using the Observer Pattern allows for two components which hold similar data to remain consistent while also staying loosely coupled. This design pattern can be seen most clearly in the sample UML diagram shown in Figure 3.6.2.2.

Observer

*Update()***Update(State)**

Concrete Subject

*getState()*

Subject

addObserver(Observer)  
notify()

Concrete Observer

***Figure 3.6.2.2: Observer Design Pattern UML***

The main advantage to this design pattern is that it very strongly decouples the concrete subject from the concrete observer allowing any changes to the concrete observer to be made easily. The main disadvantage, however, is that debugging potential problems may become more difficult as the flow of code becomes less obvious. There are two versions of the Observer pattern: the push model and the pull model. They both follow the same basic structure however with one major difference. The push model gives the Observer the information it requires when it is notified. By comparison the pull model simply alerts the Observer that something has happened. This means the Concrete Observer requires a reference to the Concrete Subject in order to determine what has changed and react accordingly. These differences can be seen clearly in Figure 3.6.2.2 with the aspects only in the push model shown in **blue** and those only in the pull model shown in *red*. The push model has the added advantage of also cleanly decoupling the Concrete Observer from the Concrete Subject.

The Observer pattern is used in the creation of the JavaApp that allows the user to login and find their typing speed. This application depends greatly on the use of listeners, which the Observer pattern is based on. The event listener was used a great deal, where an ActionListener was registered to a button UI control. In this case the ActionListener is the observer and the button is the subject. As the button changes state, the next method would be called.

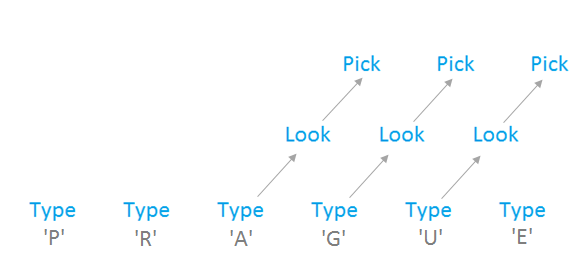
# Design

Whilst developing a cost model that has interchangeable parameters, a high quality software design is crucial to allow the user to swap out these parameters, whilst ensuring an easily maintainable program. This chapter will look at the algorithms developed and how they were broken down to solve the problem at hand. It will also discuss the User Interface for the JavaApp, whilst outlining other notable features. The formation of a strong design will result in the software being easily understood by an outside party experienced in that programming field. If the design can be understood simply, the option of extending the project at a later time becomes far more probable.

## State Machine

The fundamental algorithm necessary for developing the cost model was to design and build a state machine that could handle all possible scenarios of a user searching a destination. A state machine is an abstract machine that can be in precisely one of a finite number of states at any given point in time. The four states in the development of the state machine for the cost model are: when a user is typing a letter into the search bar, this is known as ‘type’ state. When the user is looking at the list of suggestions to check if their desired destination is there, this is known as the ‘look’ state. When the user has selected their destination from the list then this is known as the ‘pick’ state. Finally when a user is finished their query, then the will be a final end state to show that the state machine has come to a halt. This I known throughout the program as the ‘search’ state.

However, these four different states can only jump between each other depending on their response to some external inputs; the process of changing from one state to another is called a *transition*. These transitions will rely solely on a list of probabilities, which will determine what the next state will become. Time was taken to analyse all possible ways a user may reach their end result to help aid in the designing of the state machine. Figure 4.1 shows the though process behind searching ‘Prague’.



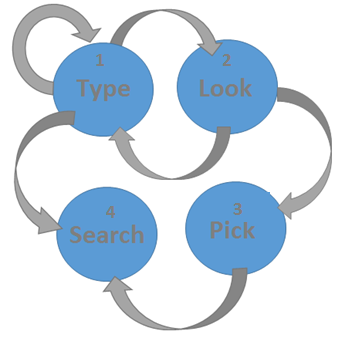
***Figure 4.1. Thought Process behind ‘Prague’***

In Figure 4.1, it is shown that whilst the user will remain in the ‘type’ state until the length of the prefix if greater than three, the user then has the option to jump to the ‘look’ state or remain in the ‘type’ state. Remaining in the ‘type’ state is still a viable option as in this example, ‘Prague’ is a city made up by a small amount of letters. Therefore, by the time the list of suggestions will appear to the user, they will have already inputted half the query. Meaning they could potentially just type in the last three letters to complete the search without jumping state. However, whilst in the ‘type’ state, they can transition to the ‘look’ as this can be a valid option if the user is unsure of spelling or know there is a high likelihood that the desired destination will be on the list. Whilst in the ‘look’ state, the user now has they are presented with two options: ‘type’ and ‘pick’. By entering the ‘type’ state it confirms that the destination was not presented in the list of suggestions. Resulting in them returning to the initial state of typing a letter, where the process will begin again whether or not to enter the ‘look’ state for a second time or just type another letter. However, at any point the user is in the ‘look’ state (first time, second time, third time etc.), if there desired place is visible on the list of suggestions they will find themselves transitioning to the ‘pick’ state. Although it is not visible in Figure 4.1, once the user reaches this ‘pick’ state, the only next state possible to transition to will be the end state, ‘Search’.

When typing ‘Par’ into Booking, ‘Prague’ is there number one listed suggestion, so the order of states for this query would be: *Type 🡪 Type 🡪 Type 🡪 Look 🡪 Pick 🡪 Search*¸ unless the user remained in the ‘type’ state throughout or switched to the ‘look’ state at a later time.

### State Diagram

The thought process model representing the user’s search, which is found in Figure 4.1, was then transformed into a visual representation of the system that could be used for all queries, not just specific to ‘Prague’. A state diagram is one of the five UML diagram which are used to model dynamic nature of a system. The state diagram created for the reactive system being developed can be found in Figure 4.1.1.

******

***Figure 4.1.1: State Diagram for cost model.***

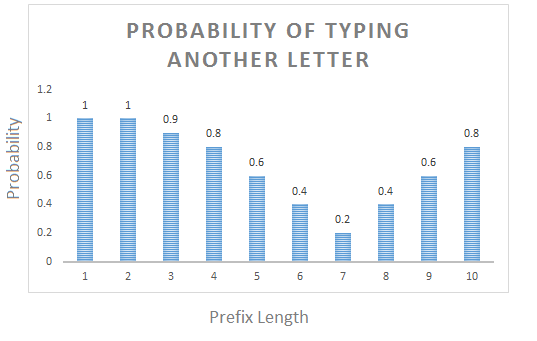
In Figure 4.1.1, the flow of control from one to state to the next is clearly demonstrated. The states can be shown clearly in blue circles, whilst the transitions can be seen in grey. The states are numbered in order of which state is most likely to be visited first, with exception to the transition from ‘type’ to ‘search’ state which could occur without the other states being visited at all. Normally in state machines, the first state is an idle one where the object waits on an event to trigger whether or not the object transitions into the next state. However, with the program being developed, the user must already be in the ‘type’ state for the state transition to proceed. Therefore, this represents the initial state, (State 1), whist the end state remains the ‘search’ state, (State 4). Figure 4.1.1 satisfies that the most important purpose of a state diagram is to model the transition of an object from creation to termination.

State diagrams are also used for forward and reverse engineering, by doing this a basic state model was created for the system. At this point, the user had to physically tell the system which state they wanted to transition to because there was no way of determining what state to transition to next.

### 4.1.2. Determining Probabilities

It is already known that the transition from one state to another, is dependent on its inputs from the original state. In the development of the cost model, it is a necessity for the state machine to know what state to enter next to be able to satisfy the user’s query. It is at this point that the idea of setting probabilities take form.

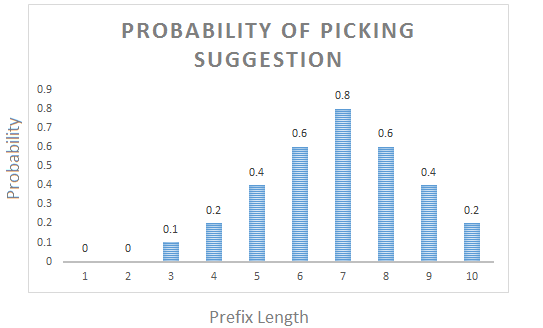
When the user finds themselves in either the ‘type’ state or the ‘look’ state they find themselves answering the question of whether or not to type another letter. In the ‘look’ state the answer is simple, if their chosen place is not part of the list of suggestions, then it makes it impossible to select it, so they must revert to the ‘type’ state. However, whilst in the ‘type’ state the choice of staying there or moving to the ‘look’ state are both feasible options. It is in this case that the probability for each option must be encoded into the state machine to show which state should be visited next. Figure 4.1.2 shows the probability of typing another letter, depending on the prefix length.



***Figure 4.1.1: Probability of Typing another Letter***

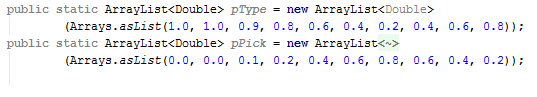
Since the list of suggestions does not appear until the user there is at least three letters as the prefix, Figure 4.1.2, shows the accurate assumption that the user will definitely be obligated to type another letter. However, once the prefix becomes three letters long, the probability of the user typing another letter begins to slowly decrease. This happens because the greater the amount of letters that is inserted by the user, the higher chance that the suggestions list will display their desired place, allowing the user to select it. Nevertheless, Figure 4.1.1 shows that after the user has typed in seven letters of their desired place, then the likeliness of them typing another begins to increase fairly rapidly. This is due to high probability that if a user’s anticipated suggestion is not yet revealed then the user may get jaded and finish of the query without the aid of QAC.

The same principle works for when the user is in the ‘look’ state and they find themselves answering the question of whether or not to pick an option from the list. Ideally, if the anticipated destination is uncovered in the list of suggestions, then the chances of the user selecting it is extremely high. However, this may not always be the case. The user may accidently fail to spot their destination and begin typing again but the chances of this happening are slim. The probabilities that a user will pick from the suggestions are shown in Figure 4.2.1.



***Figure 4.2.1: Probability of Picking Suggestion***

The probability that the user will pick from the suggestions made, works in reverse of the probability that a user will type another letter. Figure 4.2.1, shows that whilst the prefix is less than three that it is impossible to pick a place from a list of suggestions which does not yet exist. Once the user has typed in their third letter then the chances of them picking begins to increase quickly, this is because the higher the prefix then the more likely that the looked-for place will become visible. After a seven-lettered prefix, the reason for the decrease in probability is the increase in typing probability that the user may just finish typing the destination without the aid of QAC. Logically, as the probability of picking decreased, then the probability of typing must increase, and vice-versa. The probabilities for picking and typing are then stored in ArrayLists within the program, which can be seen in Figure 4.2.1. Therefore, the prefix length can be used to find the position number in the ArrayList and the probability can be determined.



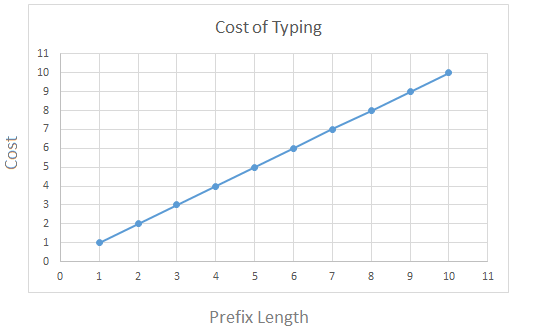
***Figure 4.2.1. Code Snippet of Pick and Type ArrayLists***

For instance, if a user was searching the destination ‘Edinburgh’ and for this example it is assumed they have already typed in ‘EDINBU’ and that they are currently in the ‘look’ state. Then by taking the length of the prefix, Figure 4.2.1 is used to determine that the probability they will type is 0.4, and the probability they will pick is 0.6. Therefore, given that ‘Edinburgh’ has been suggested, then the user will pick. This feature was then integrated into the state machine, which now carried the capability of running a place from start to finish without any physical input from the user.

### 4.1.3. Determining Costs of Each State

Once the probabilities were determined and the state machine was in fully working order. The next protocol was to generate the cost associated with changing from one state to another. It is the sum of these costs that calculates the overall cost associated with each individual city/place in the model.

Whilst the user finds themselves in the ‘type’ state, a fixed cost will be placed in addition to the total cost of that time. This cost will never be changed throughout the simulation and the linear graph can be seen in Figure 4.1.3, where the typing speed was set to 1 second per letter.

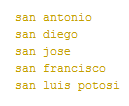


***Figure 4.1.2: Cost of Typing***

Originally the cost associated with typing a letter was hard coded within the program at 1 second per letter typed. However, this assumption was presumptuous by believing the speed of all human interaction to be the same. For instance, the time taken for a computer programmer who spends most of their day, with years of experience using a keyboard, will be vastly faster that an elderly person who may not have had much interaction with a keyboard at all. Therefore, it was at this point that the idea of creating the JavaApp took form. Regardless, it is plausible to run the simulation without determining their speed but then a typing speed of 1 second will be implemented, which could be far-fetched or an injustice to their actual speed.

Whilst the user finds themselves in the ‘look’ state, there are two possible costs that can be the outcome: the cost of looking at the list of suggesting with picking one, and the cost of looking at the list of suggestions without picking one. The cost of picking itself, is kept in an ArrayList like the probabilities from before, where the larger the prefix when entering the ‘pick’ state, the larger the cost. Therefore, whilst looking at the list of suggestions, the user does not find their desired place, **the cost for this will be the average of this ArrayList.**

Determining how long a person takes to read a list of suggestions, is a hard task to pursue. One way of finding this would be to give a user a sentence and record how long it takes them to read that sentence, then calculate an average speed for each word. Or better, give a group of users the same sentence and calculate an overall average. Instead, with some research, the average adult in the United Kingdom reads roughly 3 words per second [5]. However, all the adults this figure represents are educated. The query log stored is made up of places ranging between 1 and 5 words, with the majority being 1, we make the assumption it will take any user on average 1 second per suggestion presented. A list of possible suggestion can be seen in Figure 4.1.3



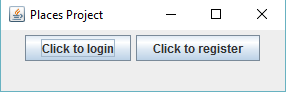
***Figure 4.1.3 List of Suggestions for prefix “san”.***

As previously stated, to read every suggested in Figure 4.1.3 should take 5 seconds (1 second per suggestion), to complete. Pursuing this logic, if the user’s intended destination was “san antonio”, then it is assumed that they would not carry on reading the rest of the suggestions, and that they would just select their desired place. Therefore, the cost in this instance would be 1 second, because it is in position 1 in the list. Likewise, if the user was searching for “san francisco”, then the cost for this looking would be 4 seconds, and so on.

The only state an object could be in that does not already have a cost assigned to it is the search cost, this is because this is being treated as an end/terminated state. Therefore, the object will stop here and simply calculate the different costs it’s gathered from the fellow states. Now the developed cost model, had all necessary parameter to generate the total cost for each place visited.

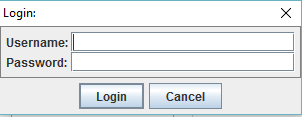
## Java Application

As the JavaApp was an extended aim, no unnecessary time was spent making a highly detailed user interface. Instead, a basic login system was created using Swing, which is a GUI widget toolkit for Java. The JavaApp consisted of a main interface called Login which called each new class, if an event was triggered to say that current class was succeeded. The initial look of the Login page can be seen in Figure 6.2.



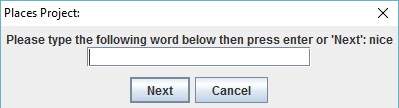
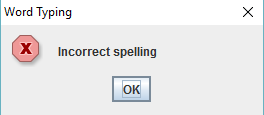
***Figure 6.2: GUI Login Home Page***

Figure 6.2, represent what can be described as the JavaApp home page. This is the first window that will appear to the user when they run the application. From here, a previously registered user can press the “Login” button to calculate a new typing speed, or a new user can sign up by pressing the “Register” button.



***Figure 6.2: Login Page GUI for JavaApp***

For either the “Login” button or the “Register” button, the next window the user will be presented with will look like the representation in Figure 6.1. However, this figure is specifically for the “Login” button. Here is where the user will enter their pre-registered username or password. At this point if the user has realised that their details are not in the system, then there is a “Cancel” button that will return them to the home page. Nevertheless, there is check on the system that ensures if the details are not already registered, an alert notifies the user giving them the option to register them. Likewise, on the Register GUI page, if a user enters already known credentials, this alerts the user and asks them if they would like login with these details. If the users is able to log in successfully, then the next window is the Input class, which can be seen in Figure 6.2a.

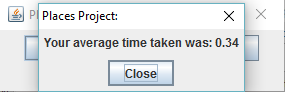


1. ***User Input of Several Places b) Error Handling Message***

***Figure 6.2: User Input GUI for JavaApp***

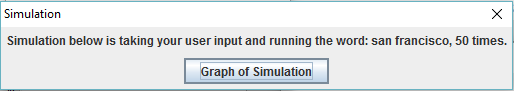
Figure 6.2a, represents where the user will be presented with a variety of different places, which they will type into the given input field. The dictionary used to generate random places was composed of the 125 most popular cities in the world. This happens because the original dictionary of all 3000 places was generating places that the user may not have heard of or knew how to spell. This resulted in inaccurate results for the average typing speed at the user was constantly having to double check the place for confirmation of the spelling. The amount of places presented to the user is one of the parameters that can be changed in the simulation. Where, the higher the amount, the more accurate the result.

A timer will start when the user presses their first key, with exception to the Enter key, which is a shortcut to submit their word and move onto the next awaiting word. So, pressing the enter key would cause an alert to tell the user they have not entered a place. Then again, when the first accepted key is pressed the timer will begin and will only finish when the last letter of the same word is typed. This time is then saved in an ArrayList and once all they words have been attempted, the average for each letter is then taken. If the user happens to make a spelling mistake whilst submitting their input, then an error message like the one in Figure 6.2b will be displayed. At this point the timer will be stopped and once the user presses “Ok” they will be able to have another attempt. This feature was made simpler by calculating the time of each word separately, meaning the user could have as many attempts on the one word without hindering the results from the previous words. Once all the word were completed, the calculated average for each letter was displayed to the user, this can be seen in Figure 6.2.

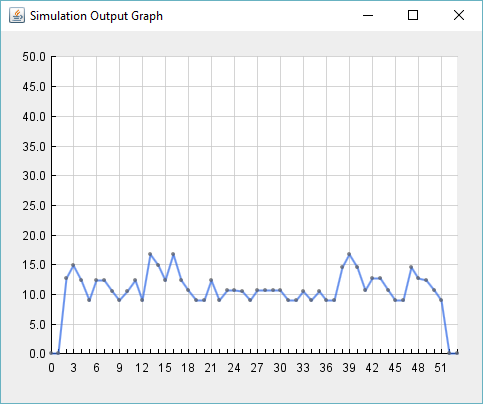


***Figure 6.2: Pop-up of Average Letter Type***

The pop-up window in Figure 6.2, shows the recently generated average time to type a single letter for that specific user. At this point their user information will be updated to include this new result. However, a drawback, if the users previous personal time was automatically be overwritten at this time, even if it was a faster more accurate time. This could be easily adapted in the code but was found to be as it is expected that most users using the JavaApp will be newcomers. Once the user presses the “Close” button, the average time to type a letter will be taken and run in the simulation to find their personal cost of searching a specific place. The simulation will continue to run in the console of the development environment, but the results will be outputted in a graph made for the purpose of the JavaApp. This can be seen in Figure 6.2.



1. ***Option for user to generate the graph.***

******

1. ***Graph for running ‘san francisco’ 50 times.***

***Figure 6.2: Graph of Simulation***

In Figure 6.2a, we can see that the user is made aware of what place is being simulated and by how many times. Then, they have the option to generate the graph which can be seen in Figure 6.2b, this was made an option to represent a visual representation of the simulation as leaving the user jumping between the JavaApp and the console is not very user friendly. For a user wanting to see how each individual place was found and the cost for each, this JavaApp is inadequate. Nevertheless, for the purpose this JavaApp was made, finding the users average letter speed, it works perfectly and is simple enough to navigate.

# Testing Strategies

Testing a simulation based greatly on probability can sustain a difficult task, as there an array of acceptable outcomes. Therefore, there can be an expected output through the assumptions made, through the parameters inputted.

Once the initial state transition model was created, it allowed basic testing with a text file containing 125 places to be tested. Whilst output was expected, the amount of places did not deem sufficient to test a simulation of its complexity. After the user had completed their prefix of three, the desired destination was often top of the suggestions list, causing the user to automatically pick if they had glanced the suggestions. Where this is ideal, realistically it would not happen unless it was the top of the popularity ranking. Therefore, after the addition of 3000 place, results became far more accurate and realistic. Once the JavaApp was created, it allowed the state transition functionality to be tested with a changing input using Full System testing. Sub-System testing was then used as a means of providing a deeper assurance that each segment of the system was working as expected. Finally, parametric simulation would be carried to ensure each parameter was in full working order, and they were not getting altered throughout the simulation. The cost model was also run on different processors, to analyse if the results differed and the time taken to compute them was compared.

The overall system was tested using a database of 300,000 places. However, in order to operate in an adequate amount of time, all the places could not be loaded into the system at the beginning of the simulation. Instead, the first letter of the desired destination was taken and all the places beginning with that same letter were then fed into the simulation, where it where it proceeded as normal.

## 5.1. Full System Testing

The Full System tests that were performed evolved the more the project began to progress. During the final stages of the project, several tests would be carried out if any part of the system was altered. The tests were to ensure that any changes made to a section had not effected any other areas of the program. The tests carried out after each alteration were:

1. Generate the cost with user just typing letters.
2. Test a String (place), so ranker will find test String in list.
3. Never pick, just look and type until end.
4. Set probabilities. Every time a suggestion is shown, look.

A further look into the Full System tests and the results they produced can be found in Appendix **BLAH** Full System Test Results.

## Unit Testing

To confirm that each individual component worked as expected, Unit testing was employed. To perform the Unit testing, the Junit framework was implemented. The Junit framework operates by creating instances of the class that is being tested, where methods can be called on the newly formed class and the results are tested to confirm they meet expectations. The routinely ran tests, which return either pass or fail, are a confirmation that no changes in code cause the functionality to break. However, to create the instances necessary to test the program, other classes need to be made available. Thus, if one of the available classes does not work as expected, then it can cause the test to fail although the class being tested functions correctly.

## Sub-System Testing

Sub-System Testing is fairly similar to unit testing. The main difference is that with Sub-System Testing, instead of test being performed on one class, it is performed on a cluster of classes that are tested simultaneously. The Sub-System that was tested within this project was the cost model on its own, followed by the JavaApp, to show that both entities could operate independently. The cost model was also split into separate sections, the state transition and ranking system, where Sub-Testing was also carried out.

## Parametric Simulation

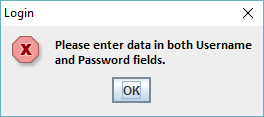
Parametric simulation is used to improve the overall performance of a system. This method will change a single variable being inputted to the system and hold the others constant, then the effect on the design will be closely monitored. Depending on the scale of the project this process can be time-consuming whilst improving the performance partially. The Parametric Simulation carried out within the project were:

1. Changing the cost of typing to see there is a continuous increase/decrease in total cost.
2. Changing when the user was able to see the list of suggestions. For example, after typing the first letter and after typing the third. Therefore, showing if users were able to select their destination fast, the difference this will make on the total cost.

A deeper look into the results of these test can be found in the Performance Analysis and Evaluation section of the report.

## 5.5. Black Box Testing

Black-box testing is a method of software testing that takes the functionality and examines it without peering into its internal structures or workings. Black Box Testing was used when developing the JavaApp for alerting the user when their intended action did not met the criteria needed to carry it out. For example, if the user attempted to register or login with a blank username or password. Then an alert would pop up saying that there has been no data entered. An example of one of these alerts can be seen in Figure 5.5.



***Figure 5.5: Example of Black Box Testing***

# Performance Analysis and Evaluation

This section of the report focuses on taking the cost model developed and logging the different effects that changing the parameters, which formulate the simulation, have on the cost model. There are four main parameters that make up the simulation; ranking of suggestions, when suggestions are shown to the user, the amount of suggestions shown and the typing cost. Each of these factors will be evaluated and the sum of all places is generated. It can be assumed that unless the parameter is being evaluated, the following figures are constant:

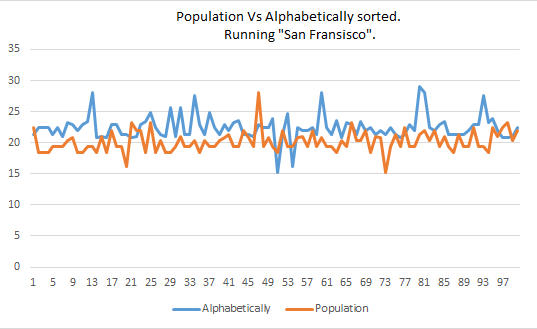
*Ranking of Suggestions: Population  
Instance of Suggestions to User: After prefix of 3.*

*Amount of Suggestions Displayed: 5*

*Typing Cost: 2s*

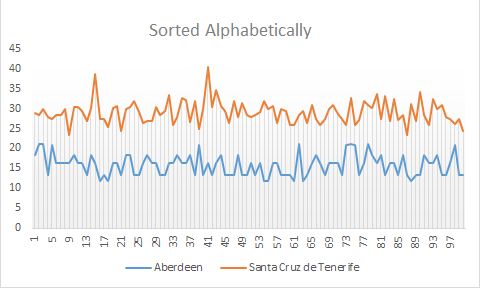
## 6.1. Ranking of Suggestions

As previously stated numerous times, the ranking of suggestions is one of most important factors in QAC. How accurate the ranking is in reference to the anticipated query, plays a major role on how quickly the query can be suggested and therefore, completed. For instance, if the ranking of suggestions was set to alphabetical, and the users prefix was ‘Saint‘, then ‘Saint Denis’ would appear before ‘Saint Petersburg’. This ranking would be ideal if the user intended query was in fact ‘Saint Denis’, however, it was not. Therefore, by changing the ranking to population instead, the intended ‘Saint Petersburg’, would be found in a shorter amount of time. Figure 6.1, show’s a representation of how these rankings effected the total cost.



***Figure 6.1: Alphabetical Vs Population***

Figure 6.1, represents the city ‘San Francisco’ being simulated 100 times alphabetically sorted, and 100 times by population sorting. Where the total cost of each simulation is plotted on the graph, the blue line represents alphabetical, and the orange line shows the population simulation. As can be seen, the alphabetical line seems to have a higher cost associated, with an estimated cost of roughly 22s. Whereas, by population, most results seem to be found at roughly 20s. These results were completely as expected, as within the dictionary of places, if the prefix entered was ‘San’, ‘San Francisco’ was 14th in the dictionary alphabetically, whereas it as the 4th by population. Therefore, even after a prefix of only ‘San’, the destination could potentially be selected for population. However, alphabetically the user would have to type more letters before it made the top five suggestions, therefore, increasing the cost. This difference of 2s if expected to increase if a wider range of places were being held in the dictionary. The population, orange line, should stay roughly the same, whereas the alphabetical blue line, would increase with the more potential places. Figure 6.1, also shows this concept by running a place ‘Aberdeen’ and ‘Santa Cruz de Tenerife’, alphabetically sorted.

******

***Figure 6.1: ‘Aberdeen’ Vs ‘Santa Cruz de Tenerife, Alphabetically sorted.***

Figure 6.1, shows how ‘Aberdeen’ is being found at roughly half the cost of ‘Santa Cruz de Tenerife”. Again, this happens because of the sheer lack of places with the prefix ‘Abe’ and the vast about that carry ‘San’. Therefore, when user had typed ‘Abe’, if they had looked at the list of suggestions, it would be their guaranteed top choice.

Having the ability to change how the user would like to have the suggestions ranked played a vital role at keeping the cost to minimum. If the user knew they were searching a well-known, highly populated, destination (which most of the time they are) then the obvious choice is population. Moreover, most simulation tests resulted in the finding the minimum cost by being population overall, this can be seen in Table 6.1.

***Table 6.1: Alphabetically Ranked Vs Population Ranked, Overall Mean***

|  |  |  |
| --- | --- | --- |
| **No of places, simulated 100 times.** | **Mean Cost (Alphabetical Ranking)** | **Mean Cost (Population Ranking)** |
| **10** | ***19.73*** | ***17.55*** |
| **20** | ***18.09*** | ***17.89*** |
| **50** | ***18.90*** | ***17.43*** |
| **100** | ***19.28*** | ***17.74*** |

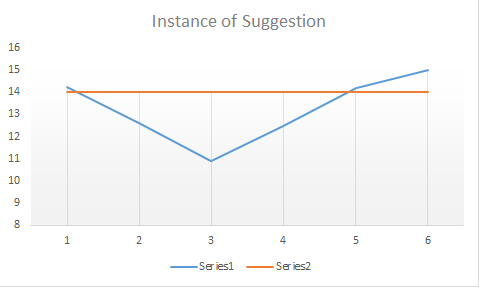
As can be seen in Table 6.1, ranking the suggestions by population gives a consistently lower results, when a random selection of places are simulated. Therefore, to keep cost to a minimum this is the most-efficient way of sorting and will be utilised throughout the rest of the project.

## 6.2. Instance of Suggestions Displayed

To date, it has been a user requirement to have a prefix of at least three letters before having the list of retrieved suggestions is displayed to the user. However, how can it be determined that this is the most cost-efficient option for the user? Table 6.2, shows five different simulations for the city ‘Jackson’, each of these simulations are executed 100 times to find the mean cost. Each simulation will be identical, except for the first simulation, the list of retrieved suggestions will be generated after a prefix of 1 letter. Then in the second simulation, the list of completions will appear after a prefix of 2, and so on. A visual representation for Table 6.2, can also be seen in Figure 6.2.

***Table 6.2: ‘Jackson’, total cost at instance of suggestions.***

|  |  |
| --- | --- |
| **Instance of Suggestion: Prefix Length** | **Mean Total Cost** |
| **1** | **14.22** |
| **2** | **12.62** |
| **3** | **10.89** |
| **4** | **12.48** |
| **5** | **14.18** |
| **6** | **15.01** |



***Figure 6.2. Suggestion Prefix Length***

In Figure 6.2, the average cost for each simulation is represented by the blue line whereas, the orange line represents the cost ‘Jackson’ would consume if they user only typed (7 letters at 2s per letter). It can be seen in Figure 6.2, that it not beneficial for the user to display the list of suggestions until prefix of at least two letters is inputted. After the 4th letter, the suggestions become irrelevant in this case, as the total cost taken is greater than it would be simply typing each individual letter. Therefore, rendering QAC unnecessary.

As can be seen in Table 6.2, displaying the suggestions after a prefix length of three letters gives the lowest total cost for ‘Jackson’. Therefore, to keep cost to a minimum this is the most-efficient way of displaying the completions and will therefore, be utilised throughout the rest of the project.

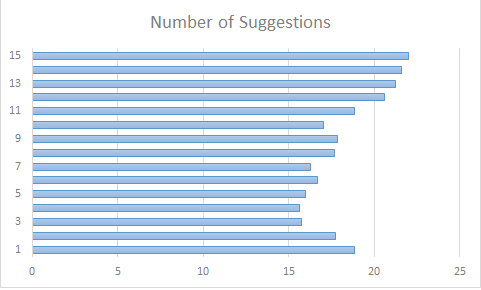
## 6.3. Amount of Suggestions Displayed

Whilst using any search bar that employs QAC functionality, it can be seen that the amount of suggested queries is minimised to between roughly four and eight. This section of the report discusses how these figures are calculated and at what amount of suggestions keeps the cost model to a minimum throughout this specific project.

To evaluate this theory, it would be bias to execute only one place into the simulation. As this specific destination may be the top populated place for their three letter prefix, causing this place to be the top ranked on a list of 1 suggestions, or 20 suggestions. Therefore, by generating 100 randomly selected places throughout the world and simulating them 100 times each to get an average cost for each place, will show a true reflection of the importance the number of suggested queries carries. This is shown in Table 6.2, with a visual representation shown in Figure 6.2.

***Table 6.2: Number of suggestions over 100 randomly selected places.***

|  |  |
| --- | --- |
| **Number of Suggestions** | **Mean Total Cost** |
| **1** | **18.82** |
| **2** | **17.72** |
| **3** | **15.73** |
| **4** | **15.62** |
| **5** | **15.99** |
| **6** | **16.65** |
| **7** | **16.27** |
| **8** | **17.66** |
| **9** | **17.87** |
| **10** | **17.03** |
| **11** | **18.83** |
| **12** | **20.59** |
| **13** | **21.26** |
| **14** | **21.62** |
| **15** | **22.01** |



***Figure 6.3: Number of Suggestions Over 100 Places***

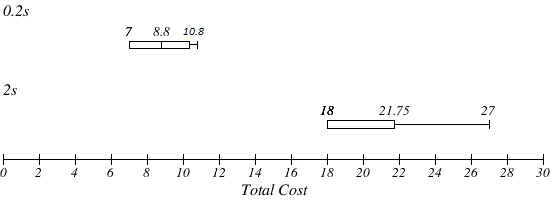
Table 6.3 and Figure 6.3, show clearly

The total cost for displaying only one suggestion was expected to much higher. However, with closer inspection this figures is completely accurate. This is because the cost of looking at the list without picking from the list of suggestions is the average of the amount suggested. So, in this case, 0.5s. This cost of looking without picking can be expected frequently when the amount of suggestion is smaller as the QAC wants to kick in and help, it just does not have the resources to do so. Also, with this cost being so miniscule, the amount was not noticed as much as if there was say, six suggestions, which would add 3.5s each time they looked without selecting. Where the higher cost for only displaying one suggestion comes from is when the user finds themselves in a state of looking without picking, they are forced back into typing another letter. Thereby, costing another 2s each time until their intended query becomes visible.

The cost for looking at the list of suggestions, with or without picking, will begin to cancel out the typing cost. As with more available suggestions, the time spent on typing letters will be redirected into looking at the list. This can be seen in Figure 6.3, where the amount of suggestions becomes higher than eleven, the total cost begins to increase to a maximum. Thereby, with only a select amount of letters typed, they user will have to scroll down the list where they may be successful and pick, which will cost 1s for every suggestion supplied until that point. Otherwise, it will cost an average of the amount of shown suggestions, where they will have to then type another letter in hope their query becomes visible in the next batch of suggestions.

The reason for choosing five suggestions to be presented to the user was determined by the fact Booking displayed this amount, and their process was to be replicated to give the user the same interaction. However, Figure 6.3, shows clearly that the most cost efficient way was displaying three or four suggestions, like Google, therefore, this could be integrated into the system to reduce costs.

## 6.4. Typing Cost



***Figure 6.4: Box Plot of Typing Cost Distribution***

## 6.5. Running Cost Model over Different Processors

Since we know that the total cost for the model is the sum of all places and that the need for sampling is crucial in a simulation this vast. This theorem is proved when the simulation was executed over two very different processors; an AMD A10 quad core processor @ 2Hz and a faster, more efficient i7-6700 CPU @ 4GHz. The results for running the cost model were recorded and can be seen in Table 6. .

***Table 6. : Running the cost model over different processors.  
  
a) Running 500 places, 500 times.***

|  |  |  |
| --- | --- | --- |
| Type of Processor | AMD A10 Quad Core @ 2GHz | i7-6700 CPU @ 4GHz |
| Mean (s) | 17.3 | 17.28 |
| Standard Deviation (s) | 0.28 | 0.34 |
| Total Cost (s) | 8,591.25 | 8,637.93 |
| Time Taken to Execute (s) | 7,529.54 | 1,513.44 |

***b) Running 1000 places, 1000 times.***

|  |  |  |
| --- | --- | --- |
| Type of Processor | AMD A10 Quad Core @ 2GHz | i7-6700 CPU @ 4GHz |
| Mean (s) | 17.75 | 17.48 |
| Standard Deviation (s) | 0.42 | 0.26 |
| Total Cost (s) | 17,754.25 | 17,988.73 |
| Time Taken to Execute (s) | 64,527.08 | 13,005.41 |

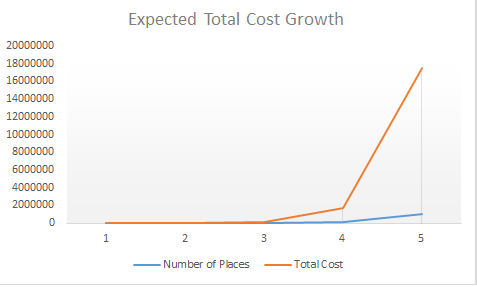
As can be seen in Table 6a, when the same 500 randomly generated places, each place being simulated 500 times, were executed on each processor they produced similar means, standard deviations and total costs. This was to be expected. The real shock was how dramatically the time to execute, on each of them, differed. These results stay consistent when the cost model executed another randomly generated 1000 places, with each place being simulated 1000 times, onto each processor. These results can be seen in Table 6b. It was expected that the i7-6700 processor would produce quicker results but operating just under 5 times quicker, truly shows how important the need for sampling is unless being executed in a data centre.

For instance, the cost model was developed and tested on the slower AMD A10 processor, therefore, running the 1000 places took just under eighteen hours to generate a total cost. This time-consuming task made it difficult to test the boundaries of what the highest number of places could be executed. Therefore, by sampling the data already gathered, the cost model was executed 100 of the previous 1000 places, which were each simulated 1000 times. The results are shown in Table 6.

***Table 6. : Running the cost model 1000 times for 100 places.***

|  |  |
| --- | --- |
| Type of Processor | AMD A10 Quad Core @ 2GHz |
| Mean (s) | 17.74 |
| Standard Deviation (s) | 0.87 |
| Total Cost (s) | 1,746.51 |
| Time Taken to Execute (s) | 638.8 |

As can be seen in Table 6, the mean and standard deviation for 100 places is similar to the mean and standard deviation for 1000 places, although the simulation only took ten minutes to execute. Therefore we can make an assumption that for running 10,000 places the mean and standard deviation should also stay similar. It is through this assumption that would can predict the total cost for 10,000 places being simulated 1000 time each to be roughly 175,000s. The expected total cost growth for up to 1 million places can be seen in Figure 6.



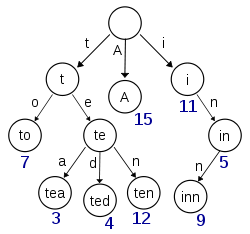
***Figure 6: Expected total cost growth for 1 million places***

If someone is a slower ‘typer’ but fast looker and vice versa

# Future Work

Although the current structure of the system will generate the cost model necessary places that would satisfy the user. There are certain features which could be implemented into the system to improve the finished product.

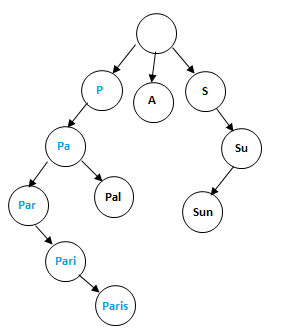
## 7.1. Trie Structure

 A search tree is an ordered tree data structure used to store dynamic sets or associative arrays, where the key is normally a String [6]. In computer science, a trie, which is also known as a prefix tree, is a form of these search trees. On the contrary to a binary search tree, it is the position in the tree that defines the key to which it associates, instead of the node in the tree storing the key. An example of a Trie structure can be found in Figure 7.1.

***Figure 7.1: Trie Structure [12].***

As can be seen in Figure 7.1, the successor of each node share a mutual prefix of the String associated with that node, where the root itself associates with an empty string. Values are not assigned to each node, they are instead associated with the resulting leaves. For example, in Figure 7.1 the keys are listed within the nodes and the values are presented below them. Therefore, each English word has a subjective integer value to which it associates. Using this trie structure for the implementation of the project, would exploit the trie's ability to quickly search for, insert, and delete entries. However, since storing words is the only necessary requirement, then a minimal deterministic acyclic finite state automaton (DAFSA) could be implemented instead, as this would use less space than a trie. The reason for this is that a DAFSA can condense branches that are the same from the trie which correspond to the same suffixes of different words being stored [6].

The main reason the implementation of a trie structure would be greatly beneficial for the simulation developed, is the amount of time it would take to run each city/place would dramatically decrease. Also allowing the user to reach their destination in a well-structured, orderly fashion. Currently, as stands, the program takes in a list of places/cites and stores it in an ArrayList. Once the first letter is typed, the program then runs through each String in that ArrayList that begins with the same prefix, whilst extracting them into another ArrayList of suggestions. For example, if the user types “PAR”, the program will remove each String beginning “PAR” and store it elsewhere to soon be display to the user. With each additional input from the user the list of suggestions will not be stored and the program will return to the beginning by clearing the ArrayList of suggestions. Therefore, when the prefix becomes “PARI” it will search the original ArrayList of places and extract all the places beginning with this new prefix. This backwards notion of clearing and starting again would continue until the full destination of “PARIS” was resulted. If the program was adjusted to implement a trie structure, the same search can be seen in Figure 7.1.



***Figure 7.1: Trie Structure for “PARIS”***

## Developing a Native Application

In today’s society, an obvious advancement for any project would be the ability for software to work on a smartphone or tablet. The ability to use the simulation as a mobile application, offers great flexibility for the user. The development of an app allows the user’s location to be determined, resulting in discovering the places that surround them if they find themselves in an unknown area. Therefore a study into what type of application to develop was utilised and how this location feature would operate. Adding a location service, would be a difficult but worthwhile addition for the end user. These location services would calculate the user’s current location, and constantly update the nearer a user got to their desired location.

### Android vs iOS

When developing a Native Applications, the two choices to be considered are: Android and iOS. Table 3 shows the comparison of these platform and any possible issues that might hinder the project.

***Table 3: Android vs iOS***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Issue | iOS | Android | Choosen platform | Reason |
| % of users worldwide | 17.7%[7] | 80.7%[7] | Android | Android offers a larger market. |
| Integration with mapping services | Supports many mapping services including Google and Bing [8] | Supports many mapping services including Google and Bing [8] | Undecided | Undecided as both offer the same mapping facilities. Therefore, both would be feasible. |

As Table 3 shows, since both Android and iOS have very similar mapping services, then the sensible decision would be to choose Android as they offer the largest market place of the application. Also, since the original cost model was developed in Java, this would be easily transferrable. If iOS was chosen, the cost model would have to be converted to a compiled programming language that was developed for Apple. For example, Swift, which could prove time consuming to learn a new programming language.

### 7.2.2. Location Services

The necessity for a location service is unnecessary for a company like Booking, as their popularity ranking will be constantly updated to produce the more popular places and the likelihood of the user wanting to visit the city closest to them is fairly unlikely. For example, if the location is “Glasgow” and the user types “GRE” into the search box, then they would expect for “Greece” and the “Greek Islands” to be suggested before “Greenock”, despite the distance between them. Therefore, the only time a location service would be necessary would be if the user was still located in “Glasgow” and they were searching the destination “Newcastle”. At this point the simulation would understand the likelihood that they meant “Newcastle, England” instead of “Newcastle, Australia”. This however, does not require an accurate location to be taken.

However, where there is great need for a location service application is when a user is searching for a place with a map service. Whilst in a precisely located destinations, a user may want to find the nearest cinema. At this point, ranking means nothing. The user would simply want to know the nearest cinemas location and how exactly to get there. In order to implement this feature into an app, it must be decides on what location service is most feasible.

There are three main technologies that are used to pinpoint a user’s locations: Global Navigation Satellite System, Cellular Positioning and Wi-Fi.

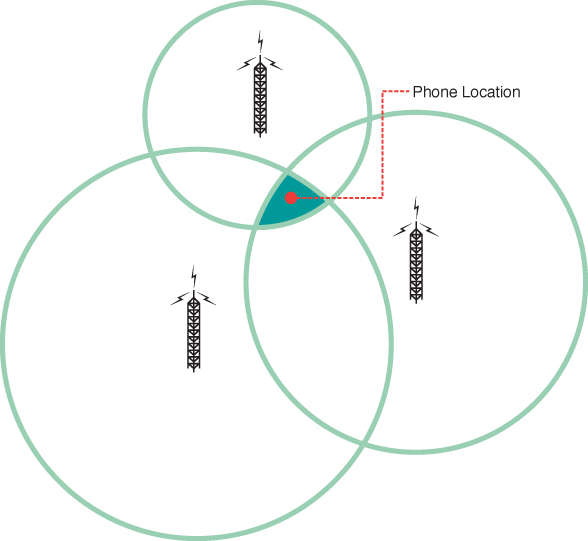
#### 7.2.2.1 Global Navigation Satellite System (GNSS)

The most well-known GNSS is the Global Positioning Service (GPS), which is funded by the United States, but there are several currently in operation. The Russian Federation have GLONASS, and Europe are soon to have Galileo (which they plan to have in operation by 2020). The main aim for a GNSS is to allow a user to determine the location they are at. The American GPS system is made up of twenty-four satellites, which continuously generate signals. These signals inform the user of the location of the satellite and they deliver precise information about when the signal was generated. The user’s location is predetermined by listening to the signals produced by four or more satellites, and determining their exact location. The result of the time taken to receive the signal and the distance from each of the satellites shared with location of the satellite is what determines the user’s locations [9].

The accuracy of GPS is approximately 3.5m [10], which although for the purpose of this project would suffice, a limitation would be the GPS receiver on a mobile device. A mobile phone’s GPS receiver can range from 2 - 10 meters, depending on the multipath conditions [11]. Multipath is when the signal takes multiple paths to reach the receiver, this could be due to the reflection on buildings. Therefore, since the application is likely to be used in the surroundings of an urban landscape, the GNSS location will be of lower accuracy than desired. With the signal being low outside, mixed with the background noise of an indoor environment, the signal could be as low as 10-100000 times weaker when indoors [12]. This signal to noise ratio, may cause severe problems when trying to gather the information from the four different satellite and without the data from all the satellites, calculating the location whilst indoors could prove very difficult.

#### 7.2.2.2 Cell Tower Based Location

Determining a user’s location using cell towers is similar to the technique used in GNSS. The mobile devices sends a signal to the nearest cluster of cell towers and it will determine how long to get a response. Then by taking the location of the cell towers, the phones location can be generated using the distances for each individual cell tower. The workings of this can be seen in Figure 7.2.2.2.



***Figure 7.2.2.2: An Example of Cell Towers***

It is known that the user must be in the circle as with the response time they are receiving, the farthest distance they could be would be a from the cell tower is a straight line (the radius of the circle). Although, there is no confirmation that signal is following a straight path, as similar to GPS, it may be being reflected off building. Taking these precautions, ensures the user is definitely at some point in the circle. Using more than one cell tower allows the user’s location to be accurately calculated, which can be seen in Figure 7.2.2.2, where the circles are overlapping one another.

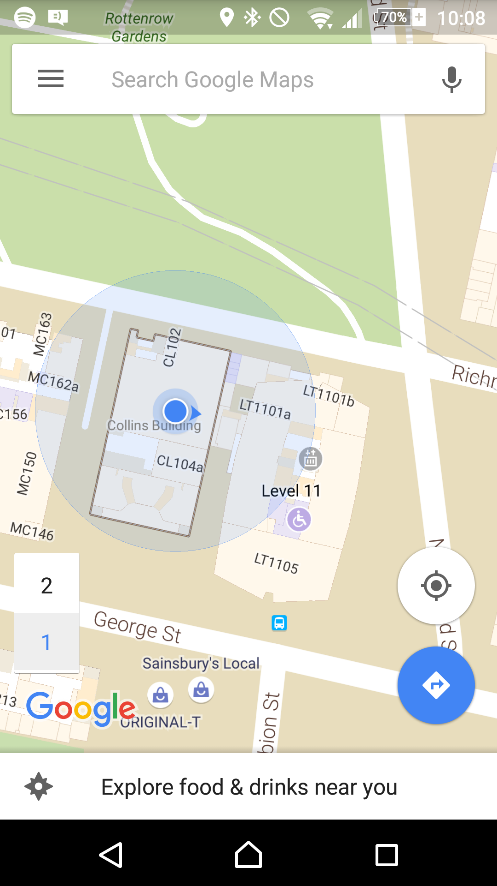
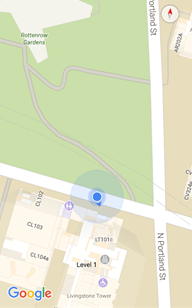
As long as the user maintained a mobile phone signal throughout, then the location could still be calculated whilst indoors. Additionally with the increase of Cell Towers in urbanised areas, Cell Tower location could be a viable method for the project. However, a major flaw is that cell tower locations are not made public knowledge, causing the use of Cell Tower Triangulation problematic [12].

#### 7.2.2.3 Wi-Fi Based Location

Similarly, the user’s location can be discover, through Wi-Fi based stations. The distance these stations are from the user can be easily determined thought criteria like signal strength. If numerous base stations are received and their location is identified, then the user’s location can be computer by triangulation. Therefore, this would be seem the most desired location service for the project. The only drawback for a user would be if there was no Wi-Fi connection made available, but nowadays, especially in urbanised areas, this is very unlikely.

#### 7.2.2.4 Android’s Already Employed Solution

Android as an operating system provides many built in features for finding the user’s current position and supplies this data to developers. This data is detected by using a combination of all three location services already mentioned [13]. To test how accurate this reading was, a location was detected outdoors, Figure 7.2.2.4a, and indoors, Figure 7.2.2.4b.



1. ***Outdoor Location Example. b) Indoor Location Example***

***Figure 7.2.2.4: Androids Location [15]***

Figure 7.2.2.4a, shows a small blue circle which is used to illustrate the user’s location, which accurately represent that actual position of the user at that time. How perfectly this signal is determined is due to how high the signal that is being transmitted by the three location services. Whereas, Figure 7.2.2.4b represents the users location within a building. This however, was not quite as accurate with the real location of the user being represented by the yellow circle.

Although there is an area of uncertainty as to where the user’s location is, this data is still practical for the project perspective. The likelihood of a user trying to navigate to a place within the building they are currently located in is slim. Relying greatly on the Wi-Fi, the idea to develop an application with location services is feasible.

# 8. Future Applications

## 8.1. Without Extension

## 8.2. With Extention

# Conclusions

## 8.1. Overall Reflection of the Project

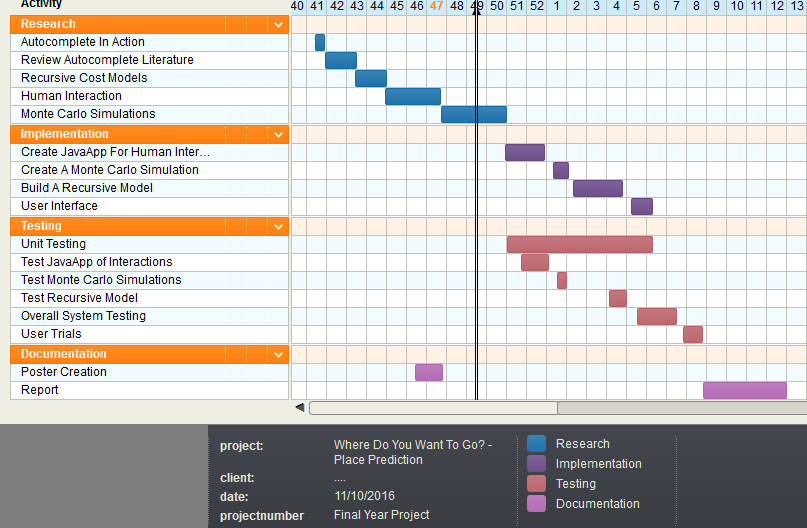
The development of the QAC cost model has been generally wholesome. The majority of different aspects were implemented without many significant problems. Of course as expected there was sections in building the state machine that were challenging to incorporate and to get working correctly. Having sufficient time to resolve any errors or complications that arose, allowed the cost model to be developed to a higher standard than having to be developed last minute. Regardless if the implementation had started at an earlier date, less of the report writing time would have been taken up, therefore would have been less time consuming in the weeks leading up to the deadline.

Any features that were not implemented within the current version of the cost model could be integrated at a later date. By implementing the cost model into Android devices in the form on an app, will vastly improve the scope of what the cost model will be capable of in terms of features and functionality.

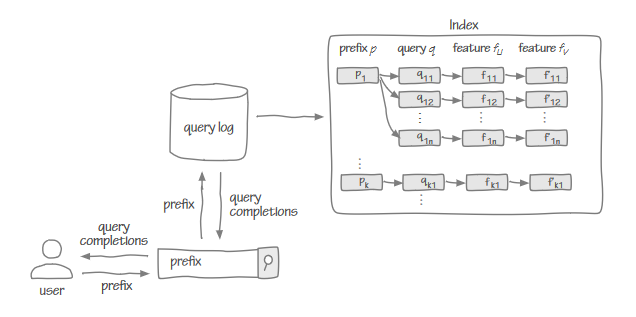
Such a relevant project into todays society.

# Appendices

## Appendix 1: Original Project Plan



Sampling, take 1000 cities. Fix them. Only going to take 1000s for the cities, take the average and that what it is to do. Make sure cities don’t change.



# References

[1] Fei Cai, Maarten de Rijke. [Online]  
“A Survey of Query Auto Complete in Information Retrieval”  
https://staff.fnwi.uva.nl/m.derijke/wp-content/papercite-data/pdf/cai-survey-2016.pdf

[2] Z. Bar-Yossef and N. Kraus.  
Context-sensitive query auto-completion.   
In WWW ’11, pages 107–116, 2011

[3] S. R. Kairam, M. R. Morris, J. Teevan, D. J. Liebling, and S. T. Dumais.   
Towards supporting search over trending events with social media.   
In ICWSM ’13, 2013.

[4] Yang Song, Dengyong Zhou, and Li-wei He.

Post-ranking query suggestion by diversifying search results Research and Development in Information Retrieval,

SIGIR 11, pages 815–824, New York, 2011b. ACM.

[5] Bret Nelson  
https://www.forbes.com/sites/brettnelson/2012/06/04/do-you-read-fast-enough-to-be-successful/#571fe7f6462e [Accessed March, 20017]

[6] Wikipedia – Trie Structure [Online]  
https://en.wikipedia.org/wiki/Trie, [Accessed March, 2017].

[7] Joe Rossignol [Online]  
https://www.macrumors.com/2016/02/18/ios-android-market-share-q4-15-gartner/ [Accessed March, 2017]

[8] Google [Online]  
 “Google Maps SDK for iOS and Android,”  
 https://developers.google.com/maps/documentation/ios-sdk/. [Accessed March 2017].

[9] E. K. a. C. Hegarty,   
“GPS Overview,”  
*Understanding GPS: Principles and Applications*, Artech House, 2005, pp. 3-4.

[10] K. M. Pesyna, Jr, R. W. Heath, Jr and T. E. Humphreys [Online].  
 “GPS World”  
http://gpsworld.com/accuracy-in-the-palm-of-your-hand/. [Accessed March 2017].

[11] George Deded [Online]  
Indoor GPS Positioning, Challenges and Oppurtunities.  
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.367.4023&rep=rep1&type=pdf

[12] Google [Online]  
“Location Strategies,”  
http://developer.android.com/topics/location/strategies.html. [Accessed March 2017].

[13] Greg Law  
University Route Planner.  
Images on Andriod taken, Chapter 2: Location Services, 2016, Pages 22-23.

1. www.booking.com [↑](#footnote-ref-1)
2. www.google.com [↑](#footnote-ref-2)
3. www.github.com [↑](#footnote-ref-3)
4. https://www.facebook.com [↑](#footnote-ref-4)
5. https://www.twitter.com [↑](#footnote-ref-5)
6. https://www.twitter.com [↑](#footnote-ref-6)