C O V E N T R Y
U N I V E R S I T Y

Faculty of Engineering and Computing

Department of Computing

Data Science And Computational Intelligence

M08CDE Individual Project

# Intelligent maps that self-customise information presentation based on what they know about its user

Author: Imran Kader Chowdhury

SID: 6694003

Supervisor: Dr. Faiyaz Doctor

Submitted in partial fulfilment of the requirements for the Degree of Master of Data Science and Computational Intelligence

Academic Year: 2015/16

## Declaration of Originality

This project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet, etc) has been acknowledged within the main report to an entry in the References list.

I agree that an electronic copy of this report may be stored and used for the purposes of plagiarism prevention and detection.

I understand that cheating and plagiarism constitute a breach of University Regulations and will be dealt with accordingly.

## Copyright



Signed: Imran Kader Chowdhury Date: 15/08/2016

**Office Stamp**

## Abstract

Digital maps have become a crucial tool for our everyday life. People use maps every day for navigation and to discover new places. However, users are being overwhelmed by the amount of information presented on the digital map. When the user explores the route, the important business locations are prioritised and displayed first. However, a digital map can be made personalised and prioritise places that are important to specifically to that person at that precise moment, if the map can learn and adapt from his historical geospatial data. In this research, machine learning prediction models such as Decision Tree and Random Forest was applied to predict locations from a person's previous data such as current geolocation along with time and day, activities, feelings and emotions associated with the place. These data were collected from a mobile application from six volunteers over a seven-day period. The geographical data of London and West Midland was extracted from OpenStreetMap database. OpenStreetMap is an open sourced database that contains the geographic data of the entire planet. Moreover, a user based recommendation model has been built using Apache Mahout machine learning library to recommend geolocations to a person. The recommendation engine has been constructed using collaborative filtering algorithm and applied to the prediction result. The model has been tested using cross-validation, precision and recall. Finally, a contextualised map for the Android platform has been built using open source projects ('OpenStreetMap' and 'Mapsforge'). The model was deployed on Apache Hadoop to see that the system can be made scalable. To store data Hbase database has been used which runs on top of HDFS (Hadoop Distributed File System). At the end of the research, a survey was conducted among the volunteers to get valuable feedback and evaluate the application from user's point of view.

# Table of Contents

## *Additional Materials on the Accompanying CD*

**Resources:**

1. Android application Source Code
2. OpenStreetMap Dataset (London and West Midland Extract)
3. Processed Data
4. Data Collected from User
5. Data Models
6. Source Code of Collaborative Filtering
7. Source Code of Random Forest
8. Android Package Archive File

**Documents:**

1. Project Brief
2. Project Report
3. Project Plan
4. Project Milestone
5. User Manual
6. Presentation
7. Consent Form
8. Survey Questions
9. Survey Report
10. Er Diagram
11. Sequence Diagram
12. Use Case Diagram
13. Meetings with Supervisor
14. Ethics Document
15. Ethics Approval Certificate
16. Participants Information Sheet

## Acknowledgements

I would like to show my gratitude to my supervisor Dr. Faiyaz Doctor for sharing his wisdom and guiding me overcome complex issues during the research. I would also like to acknowledge the OpenStreetMap team and thousands of volunteers from all over the world contributing to develop a detailed geospatial dataset of the planet. I would like to thank my friends for supporting me during the research. I am extremely grateful to my brothers, sisters and my parents for helping me in many ways during the research. I would like to thank the people participated in various parts of this research, for their time and consideration. I would like to show my gratitude to my tutors Dr. Faiyaz Doctor, Abdulrahman Altahhan, Vasile Palade for sharing their knowledge in machine learning which helped me a lot accomplishing the task.

# 1 Introduction

The map is a canvas to display what we know about the world around us. People use digital maps for many different purposes such as for navigation, searching for places, for cycling, route planning and much more. Digital map has become an essential tool for our everyday life. However, the map is inundated with information which makes it difficult for the user to choose the right one in right place and time. There are lots of labels on the map which makes the location searching even more challenging. When the user zooms out the map, the names get overlapped. To alleviate this issue, the labels that are less important are kept hidden from the view. Again when the user zooms in, more important labels are shown first. Traditional map rendering engine prioritises famous and well-known business locations over other places. However, a map can be more intelligent and highlight locations that are relevant to a particular user. If the engine learns about the daily activity of a person, such as where he/she goes, what type of activity he/she is involved with in that place and how does he/she feels about that place, the map can show and highlight locations that are more appropriate and relevant to the person in that context. In this research, a machine learning model using decision tree and the random forest applied on the data collected from six volunteers over a week period to analyse and predict where the person is likely to go next and what type of activities he/she will be doing. From the prediction, we can make a more contextualised and self-customizable map that displays information that the person is likely to be interested in at that moment. For example, a person goes to a particular type of places in the morning and a different kind of place in the evening. The map learns from his daily behaviour and activities and adapts the route accordingly in such a way that it will display only the type of places where he/she would like to visit at the particular time of day. The engine doesn't only analyse the pattern of daily and weekly activities but also considers other aspects such as feelings and emotion of a person associated with the place. Moreover, a recommendation engine using collaborative filtering has been applied to the data to recommend new places to the person which he/she might be interested. The purpose of this research is to investigate different aspects of contextualization, to build a personalised and user specific map. The primary objective of the project is divided into three sections such as collecting and processing data, investigate different machine learning models to analyse, predict and recommend locations to the user and finally to build a contextualised map application to reflect the results. These objective leads to the following research questions:

**Research Questions:**

1. To what extent this recommendation model can recommend locations to a person from similar user's preferences.

2. To what extent the machine learning model can predict where a person will go by analyzing their personalized data such as feelings, emotion and activities associated with time, day and geolocation.

## 1.1 Background to the Project

The first map service launched by Google in 2005 was much like a paper map. It came in the form of a web application with very few features such as dragging and zooming. People used the maps only to search and explore places. Three years later in 2008 the map was first introduced in GPS-enabled smartphones, which was a breakthrough event in the history of the development of the map services. Since the system was able to know user's current geo-location, applications became more data-driven and location aware, resulting the users to have a much better experience. Today Google map provides services such as top-down satellite view, panoramic street view, real-time traffic conditions, route planning for the different mode of transportations such as walking, car, bicycle or even public transports along with real-time navigation. Maps are now becoming an imperative tool for the users and people are using these features in their everyday life. Although the user experience was improved to a great extent, there is still much more a map can do for a user if it can learn and adapt to his daily activities.

The next phase of the evolution is to make the map more user specific. A unique map can be generated for every single user based on time of day and the place they are currently in. The map should be more contextualised and data-driven. Since from the geolocation data we know exactly where the buildings are as well as how they look like and also things that are in those buildings and we can also infer some information from the users such as how they feel about a place, whether they want to go there alone or with friends, what kind of activities they do in those locations, do they go there in the morning every weekdays or on the weekends only, do they go there by foot or by a car, which route they chose to go there and much more. From this data, the map can be made more personalised and display information in the map that is appropriate for that person in that context.

## 1.2 Project Objectives

The objective of the project is to investigate different methods of collecting data from users, analyze the data and recommend appropriate places to the user and finally built a contextualized map application to show the recommended places. The objectives can be listed below.

1. To collect data such as current-geolocation, feeling, activities, emotion from the user.

2. To build a recommendation engine to analyze and recommend places to people.

3. To build a map application to show the recommended places.

### 1.2.1 Data Collection

Two types of data have been collected for this research. One is the OpenStreetMap geodata, and the other is the user data such as geo-location, feelings, activity, emotion and rating obtained from six volunteers over a period of a week. OpenstreetMap is licensed under open content license and is free to use for any research purposes. OpenStreetMap data contains the data of the whole planet, however, for this research only geodata of London and West Midlands has been extracted and used. On the other hand, user data was collected through a mobile app. For this research at early stage data was simulated to test the functionality of the system, however later data was collected from six volunteers over a period of a week. Also at the end of the research, a survey was conducted to get some review from the users.

### 1.2.2 Analysis and Model

The OpenStreetMap data was extracted and processed for analysis. The data was obtained using XML parser in Java programming language. The locations were classified, and the data model was refurbished. Two machine learning classification model, decision tree and the random forest has been applied on the data collected from the user to predict the class. The user data contains six features such as day, time, friends, activities, feelings and emotion with a corresponding class id. A class is a combination of type and sub-type of a place, for example, a type could be restaurants and sub-type could be fish_and_chips. There are total 1341 classes in the dataset. Finally, to recommend places, collaborative filtering algorithm has been applied. Collaborative filtering is a

user-based recommendation approach where the algorithm suggests items to the user from the preference of another user if they both have some common interests.

### 1.2.3 Results and Evaluation

In this project, Random Forest is used for prediction and Collaborative Filtering has been used for user-based recommendation. Random Forest was used to predict locations and was evaluated with k-fold cross validation evaluation technique. The model produced 65.238% accuracy (approx.). The collaborative filtering algorithm was assessed using average absolute difference recommender method. The mean absolute error calculated was 1.79

## 1.3 Overview of This Report

This report is started by addressing the problem statement and introducing the research proposal and different aspects of the research. Various research papers have been reviewed in the second chapter to investigate existing solutions or proposed solutions in various parts of the project. The literature survey topic has been covered the following research areas: OpenStreetMap evaluation and data accuracy survey, sophisticated methods of collecting data from the users for sentiment analysis, machine learning prediction models such random forest and decision tree, collaborative filtering recommendation system and some related work on geospatial analysis. The methodology and software development process is described in the third chapter. In chapter four, the functional and non-functional requirements of the system have been discussed. Different aspects of the project have been analysed in chapter five, such as a description of the data sources, data modelling, different machine learning models (prediction and recommendation) and infrastructure. Chapter six is the design section where the system architecture design, sequence diagram, use case diagram and er diagram has been described. In chapter seven, different stages of implementation have been presented with appropriate screenshots of the system. In chapter eight the machine learning models have been tested using k-fold cross validation, precision and recall and the accuracy of the system is analysed. In chapter nine, project management system is explained and in chapter ten a critical analysis of the overall report has been described. Finally, in chapter eleven the report is concluded by presenting the findings of the research and possible future work. At the end of the report, the list of resources and literature used for this project has been listed down and finally different part of the final implemented project is presented in the appendix section.

# 2 Literature Review

This project contains several research areas. Firstly, the research was focused on quality and accuracy of OpenStreetMap dataset and geospatial information analysis. This part of the study reviewed about the accuracy and methodologies of OpenStreetMap data and how this data can be used for analysis. The second part of the research is about user data collection and sentiment analysis. The user data such as current-geolocation, activity, feeling and emotions were collected through a mobile app in this project. However, some research paper has been reviewed to explore different methods to automate and collect user data from wearable devices and sensors. Thirdly the research was focused on studying different machine learning prediction models such as decision tree and random forest, to learn how these prediction models can be applied to the collected dataset and how well they perform on predicting the outcome. Finally, a comparative study of various recommendation systems has been carried out to investigate the best approach. Moreover, some work similar to this research has been studied for critical analysis.

The study begins with the hunt for open source geospatial data sources which leads to the OpenStreetMap project, widely used for spatial analysis all around the world. Spatial data analysis and cartography have been a very popular field of research over the last decade. Access to geospatial data was controlled by the government and commercial organisation at the early age, and the end users had very limited access to these data often by paying high charges for the license. Although the data was restricted to the end users, this geo-data business model was a success. However, the evolution of the internet of things and social media over the last decade has made significant development in this field. As the global-positioning system came into existence, portable and cheap satellite navigation systems became available for everyone, which confronted the geo-data business model. In 2004, the OpenStreetMap project was first introduced under open content license. The aim of this project was to maintain a geo-information database of the world and make the data free and available to everyone so that the users no longer requires to pay for the data and license. OMS contains rich geographical data such as streets, buildings, rivers, lands and such. Thousands of volunteers are working all over the world and adding new data to OSM planet database every day. OSM provides free software for the contributors to edit and update the data into the database in real-time as-well-as it provides tools for importing the data in GPS enabled devices. OSM also provides various communication media for the OSM community members and contributors through online forums, emails and meetings (Mooney and Cocoran 2013). By the year 2014, OSM had 1.90 million registered end-users and contributors. During the last couple of years, scientific research disciplines such as geoinformation science, cartography, geography, ecology

and computer science have been involved with the project both in academic and commercial research. Various research has been conducted to validate the OSM data, and it has been concluded that the geo-data of OSM are complete and accurate in some parts of the planet in compared to other proprietary geo-datasets (Zipf and Nis 2012; Zipf and Zielstra 2010; Helbich et al. 2012). Evaluating the accuracy and quality of OSM data has become a field of research and various software tools has been developed by the OSM community to validate the dataset and analyse the quality. (M Haklay et al. 2010; Jokar Arsanjani et al. 2013). There are several other approaches proposed in different researches to improve the data through various algorithms (Amelunxen et al. 2010). Many academic types of research have been published to investigate the evolution of the OSM data across the planet (Mooney et al. 2012; Zipf and Neis 2012).

Next part of the research was about user data collection. Initially, a model was made from user's geo-location data, but later parameters such as human emotion, feelings, activities associated to geo-location were incorporated to contextualise the recommendation model. In this research, a mobile application has been developed to capture this data from the users. However, some studies in the field of sentiment analysis have been explored to investigate some other automated approach of data collection. The researcher has tried different methods to automate capturing human emotions and predicting them. Some research proposed the analysis of user engagement in social media such as Twitter while other research proposed obtaining data from various wearable devices and to analyse them to predict someone's feelings and emotions and activities. Both methods have been explored in this research.

Sentiment analysis from social media is a part of the Natural Language Processing(NLP) research field. NLP is the process of extracting meaning and information from text. There are many social media and blogs such as Twitter, Reddit where people express their thoughts and feelings. There is a broad range of research field for data mining from social media and process and analyse this stream of data in real-time to predict someone's feelings. Some recent studies of sentiment analysis led by Bermingham and Smeaton, in 2010. They examined people's tweets that contain emoticons. They classified positive emotions if the tweet contains ':)',':-)' and negative if the tweet contains ':(',':-(' and such. They built machine learning model using Naïve Bayes, Support Vector Machine and Max-Ent. Another research led by Pak and Paroubek in 2010 proposed different classification techniques. They also crawled tweets with emoticons along with some online newspapers such as 'New York Times' and 'Washington Post' associated with the Twitter account. Both approaches built on top of bigram and unigram model. An n-gram model is a language model that uses probabilistic methods to predict the next word of a sequence of words that are in the form of

Markov model. The training and testing data were collected from random samples from the stream of tweets. Barbosa and Fang directed another research in 2010 with a different method. They extracted the features from the tweets such as punctuation, link, hashtags, retweets, exclamation and punctuation combining with the polarity of each part of speech(POS). However, it was not mentioned in the paper how the data was collected. Gamon performed research on sentiment analysis in 2004. His research area was extensively focused on extracting and analysing linguistic features such as parts-of-speech from the text. There are also some acronyms used by people in social communication media such as 'lol','ROFL','bff','gr8' and such which was studied by Apoorv Agarwal and Boyi Xie et al. in 2010 to classify sentiments using feature based and tree kernel models.

Social media is one data source that can be used to analyse sentiment. Another data source can be the data coming from wearable sensors. A lot of research has been published to build automated systems to recognise human emotions. Keshi Dei, Harriet J. Fell and Joel MacAuslan studied how to identify human emotion from speeches. Frank Dillard, Thomas Polzin and Alex Waibel also performed similar research in 2010. They used artificial neural networks to recognise emotions from speech. Some researchers tried to identify emotion from facial expression. Isidoros Perikos, Epaminondas Ziakopoulos and Ioannis Hatzilygeroudis performed research on the identification of emotion from facial expression in 2014 using the neural network. Another automated emotion detection technique involves physiological wearable devices, which captures pulses, respiration, oxygen level, blood pressure, perspiration and chemicals present in sweat. Some researchers used Electroencephalogram(EEG) and Electrocardiography(ECG) (KH Kim, SW Bang et al. 2004) to recognise emotional state such as anger, joy, sadness, fear, hate, love, disgust, grief. During the research, some music was played, and video clips were shown to the volunteers to stimulate the emotion (J Kim, E Andre 2008). They conducted a series of 20 experiments on the 20 volunteers in a period of 20 consecutive days, and each test lasts for 20 minutes. The researchers were able to successfully classify six states of emotion ('Sadness', 'Dislike', 'Joy', 'Stress', 'Normal', 'No-Idea') with a stunning 81% accuracy. Human feelings and emotion have been captured through a mobile application in this research. Since the primary objective of this study is not to analyse sentiment rather predict human behaviour and recommend places accordingly, for this reason, more sophisticated methods of sentiment analysis was not applied and investigated further. However, some machine learning model such as random forest and the decision tree was implemented on the data collected from six volunteers over a period of a week to predict geolocations from activities, feelings and emotion of a person.

The next part of the research is to classify the geo-data and predict geo-locations for a given set of features such as activity, feeling and emotion. Various classification and prediction model has been investigated for this purpose. Classification is an attractive field of research in machine learning. A research paper was published by Jehad Ali, Rehanullah Khan, Nasir Ahmed and Imran Maqsood in 2012 titled 'Random Forest and Decision Trees', has compared the classification results of this two models. The experiment was applied on twenty datasets from UCI Machine Learning Repository. The classification results obtained from the Random Forest model was then compared with the result of Decision Tree. The parameters were correctly and incorrectly classified instances, accuracy, precision and recall. For small datasets, the difference is marginal but for large datasets. However, Random Forest yields much better results. For example, the parentage of accurately classified instances was increased from 69.24% to 96.12% for Random Forest when the total number of cases of the breast cancer dataset was increased from 288 to 698. Random Forest belongs to a more sophisticated machine learning class called ensemble learning. Ensemble learning is a method of combining and using multiple machines learning algorithms to achieve better prediction result. Random Forest utilises the Decision Tree model and builds many trees during the time of training and calculates the mean prediction of each tree. The trees are generated by taking a random sample from the dataset.

Another important research domain is location aware recommender system (LCARS). There are a number of research work published on this topic.

Research led by a team of five (Hongzhi Yin, Bin Cui, Yizhou Sun, Zhiping Hu, Ling Chen) in 2014 has applied the model based approach to predict the interests of the users. Their model was designed to recommend venues such as shopping malls and restaurants or events like exhibition by considering user interests. However, they have applied their model to museum dataset only and used collaborative filtering as recommendation model on user ratings. They have also hybridised their model to fit the non-Gaussian rating (Diggle et al., 1998 and Yu et al., 2006). By limiting the number of venues and event for the user (Horozov 2006), collaborative filtering on spatial items (Ricci and Shapira et al., 2011) yields better results. There are other approaches proposed in different papers for the spatial recommendation. The task gets even more compelling when the model is trained by observing local commuting of a user and recommend places when the user goes to an entirely new venue and has no previous activity record (Levandoski et al. 2012). The research has analysed the data from Foursquare and has concluded that 45% users usually travels within an area of 10 miles' radius and 75% users travels within a zone of 50 miles' radius. Another study shows that 0.45% of the activity was recorded from the users when they are away from their home city (Scellato et al. 2011). The research concluded that for the case when the user is in a new

town and has no activity record at that area, the item-based collaborative filtering fails to recommend places to the user.

From this literature survey we have learned that predicting human activities, feeling and emotion from sensors achieved 81% accuracy by a research (J Kim, E Andre et al. 2008). However, in this research the activity, feeling and emotion was collected through a mobile app. The data is acquired from user inputs and therefore the data is 100% accurate. Moreover, researchers have faced difficulties recommending geo location to user in new area. My approach has addressed this issue by classifying all the location in the dataset. My prediction model is not predicting where a person go rather it is predicting the type of places a person might be interested in.

# 3 Methodology

A lot of research has been done on recommendation system since automated recommendations have now become essential in every sector. There are different types of recommendation approach available such as item based, user based and hybrid recommendation systems. In this project, both item based and user based recommendation has been applied on geospatial data. The geospatial data was taken from OpenStreetMap which is a crowd-sourced dataset and was made open-sourced under open database license (ODBL). Although it contains the geolocation data of the entire planet, for this research only the data of London and West Midlands has been extracted and used. Moreover, some personalised data was collected from the users to analyse and give recommendations. The data was collected from six volunteers and it includes their current geo-location, rating, emotions, feelings, activities they are currently doing and also if they are alone or with friends. The user data has been collected through a mobile application and was stored into HBase distributed data storage system. Later a data model was prepared from this data source to apply various machine learning techniques such as item-based recommendation, collaborative filtering and decision tree/random forest. The analysis was performed in Hadoop distributed file system using MapReduce paradigm and the data was stored in HBase data storage. A restful interface was made to deliver the data to the client application in JavaScript Object Notation(JSON) format. The client application then interprets this received data and presents the information on a map. The model was tested and evaluated using standard machine learning testing techniques like k-fold cross-validation.

## Development Process:

The software development process for this project was a combination of prototyping and incremental approach. The software prototyping model is a method were many different versions of the software is to be developed as prototypes. Each prototype is being presented to the client and their feedbacks and requirements are being analysed. The software is then further developed to adopt these changes and a new version is released. This process is repeated several times until a stable release version has been achieved.

There are three parts of the project, one is the application for collecting data from the user, the other one is the map application to present the recommendations and finally the recommendation engine that analyses the data collected from the users and send recommendations to the map application. For this project, the first version of the software that was developed was an item-based recommendation system. The geolocation data from OpenStreetMap was processed and categorised. A mobile application for collecting data was built to collect the geolocation information from the users. An algorithm was developed to find geo locations that are of similar to the user's interests and finally the data was put on the map. In the second version of the software, the recommendation engine was further improved by adding user-based recommendations using the collaborative filtering algorithm, making it a hybrid recommendation system. A web application was made to make the data collection process faster. The user-interface of the map was improved by adding colour markers such as red, orange and green where red represents not recommended, orange represents recommended, and green represents highly recommended. In the final version of the product, the mobile app for collecting data was further improved by adding features like capturing a person's geolocation along with his feeling, emotion, activities, whether he/she is with friends and such. Also, a rating system was introduced so that a person can express whether he/she likes or dislikes a place, for example, a restaurant. These features were also incorporated into the recommendation engine making it a hybrid contextualised recommendation engine. The recommendation engine would further analyse the recommended data model and would apply decision tree/random forest algorithm to analyse the features like feelings and emotions and such and provide more contextualised recommendation.

The prototyping development method is not a standalone process but rather it contains another method for each development phase. Although three phase of development was described above, each step was developed under incremental development approach. At each stage, the features and

objectives were broken down into smaller segments so that they can be changed easily later on. The incremental model is a process which is a way of performing several smaller waterfall models during the development of each phase. Once a part is completed then the next increment of the development proceeds. The main steps of waterfall model are requirement analysis, design, implementation, testing, integration, deployment, maintenance. The reason a waterfall methodology could not be applied here is that the requirement analysis, design and implementation were changed throughout the process the changes were adopted accordingly, and the final software was evolved through different development phases as described above. The research conducted for this project was also quantitative and not qualitative, since the actual data from users were collected for this research such as personalised logged data, geolocation data and surveys and the results was concluded by analysing these data. The surveys that were conducted for the study has three parts such as participant's information collection, usability and the last part is about the evaluation of the final product.

The first section of the survey is to collect participant's information such as name, email, address, phone number. Participants name and email are mandatory fields to identify a user uniquely, and rest are optional. The second part of the survey is about usability where points like how the user uses the existing systems and what do they use them for was focused. The final part was about user feedback and user experience to evaluate the final application that was developed.

# 4 Requirements

In this section, the functional and non-functional requirements have been discussed. The functional requirements investigate what tasks different parts of the system are responsible for. And the non-functional requirement describes things that are required in order to deploy the system. The application consists of two components such as data collection component and recommendation component. The data collection component is a mobile application that asks the user some questions about his current situation. The questions are divided into five sections such as where he/she is now, what he/she is doing, how he/she is feeling, how would he/she rate the place and what is his current emotional state. The collected data is sent to the recommendation engine that analyzes and recommends geolocation to the user.

**Functional Requirements:**

- The data collection system should be able to collect the data from the user.
- The data collection system should be able to send data to the recommend engine through a web-service.
- The recommendation engine should be able to recommend places to the user.
- The map application should be able to plot the recommended places on the map.

**Non Function Requirements:**

- The data collection application and the map application was developed on android platform. A minimum android API version level 10 (Gingerbread) is required to build and deploy both the applications
- The application can be deployed on android smart phone and it requires internet connectivity wi-fi/mobile data and location services to be turned on.
- The recommendation engine has two parts such as prediction model and recommendation model. For prediction, python development environment is required with machine learning library sci-kit learn.
- The system can be deployed on a Linux based system with python and java development environment.
- For the recommendation model apache mahout machine learning library is required.
- The system requires Hadoop distributed framework configured on a Linux server.
- HBase distributed database system in required for storing data.
- The system should be available at all time, the system should be fault tolerant, scalable and reliable.

# 5 Analysis

This section of the report describes a high-level view of the system and investigates different part of the system such as primary data sources, dataset descriptions, data collection method and different data models. Moreover, the applied method has been compared with other methods. Finally, user feedback survey report has been presented and analysed.

Overview of the system:

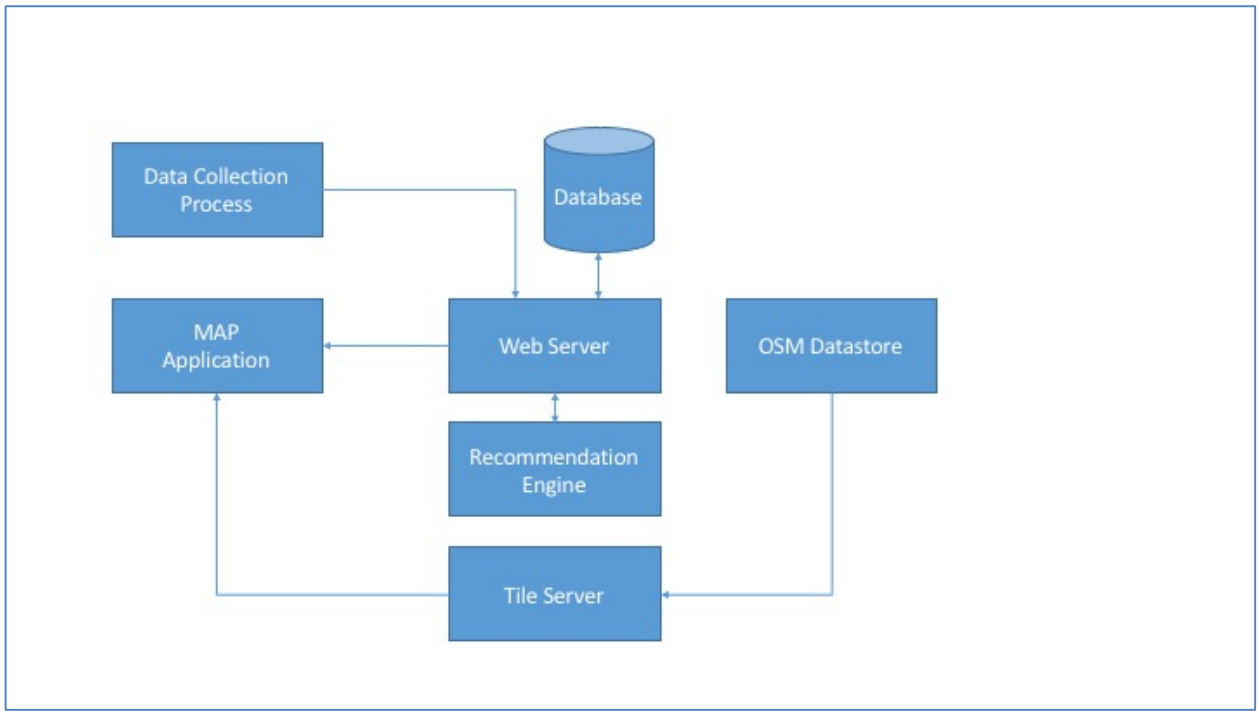



Figure 1: High-Level View of the System

The process starts with the data collection application. Every time user wants to see the map he/she will be asked some questions such as where he/she is now, what he/she is doing, how is he/she feeling and such. The system must know the context first, in order to make a contextualised map that can recommend places based on context. The data is then stored in the HBase NoSQL data store hosted on top of Hadoop Distributed File System.

The recommendation engine analyses the data and sends the result back to the map application.

There are two primary data sources here. One is the geo-location data from OpenStreetMap and the other one is User Generated data. OpenStreetMap contains geodata of the entire planet, however, for this research only the data of London and West-Midland has been extracted. Below OpenStreetMap data is described:

**OpenStreetMap Dataset Description:**

OpenStreetMap is a fast growing crowd-sourced geographical information archive. The OSM data was collected from volunteers all around the world, this type of information is known as 'volunteered geographical information' or VGI in short. The dataset is updated and maintained on a regular basis. Openstreetmap has successfully captured around 29% area of England over the last four years. The data is made available for academic research and also appropriate for commercial use.

The OSM contains the geolocation data of the entire planet. The data is stored in a Planet.osm file in the form of structured XML which is about 660GB in size. A java command line application known as osmosis needs to be used in order to analyse and extract information from this file. However, they also provide regional extracts which are of smaller size and generally contains geo information of a city or small area. For this research, I have used the extract of London data and west midlands data. A typical structure of an OSM file can be described below.

```
<?xml version="1.0" encoding="UTF-8"?>
<osm version="0.6" generator="CGImap 0.0.2">
 <bounds minlat="54.0889580" minlon="12.2487570" maxlat="54.0913900"
maxlon="12.2524800"/>
 <node id="298884269" lat="54.0901746" lon="12.2482632" user="SvenHRO" uid="46882"
visible="true" version="1" changeset="676636" timestamp="2008-09-21T21:37:45Z">
  <tag k="traffic_sign" v="city_limit"/>
 </node>
 <way id="26659127" user="Masch" uid="55988" visible="true" version="5"
changeset="4142606" timestamp="2010-03-16T11:47:08Z">
  <nd ref="292403538"/>
  <tag k="highway" v="unclassified"/>
 </way>
 <relation id="56688" user="kmvar" uid="56190" visible="true" version="28"
changeset="6947637" timestamp="2011-01-12T14:23:49Z">
  <member type="node" ref="294942404" role=""/>
 <relation/>
</osm>
```

Above is a snapshot of the data model of OpenStreetMap. The OSM data model is consists of three basic elements such as node, ways and relations. Each of these elements contains one or more tags. A tag describes an element. Each tag contains a key and a value. The key defines the type of an item and value contains the details of that item. An element may contain more than one tag. For example, key='amenity' and value='restaurant'.

The primary element node is used to represent a single point on the surface of the earth. Each node contains a unique id and coordinates (latitude and longitude). A node can represent a single point feature or configuration of a way. A node usually doesn't have any tags when they are used to represent ways. A node can also be used as a member of a relation where a relation represents the function of that node in a set of relevant elements in the dataset. Another important element is the way which is a collection of nodes ranging in between 2 to 2000, which is used to represent a polyline. The way is an important data structure that is used to represent linear attributes such as roads, highways, railways or river. The way is also used to define the boundaries of an area, for example, a forest or a building. The element of the set if called relation which relates all the elements together to create a new meaning of them. For example, a route can contain a collection of ways to define a highway or a bicycle route. A multi-polygon relation might contain several ways and nodes to describe the inner and outer boundary of an area. This is the basic structure of the original dataset which was later pre-processed and made a new structure for the convenience of the computation.

Processed data:

```
<node id="993861491" lat="51.5160255" lon="-0.3019331" n="Tesco Express"
t="shop" st="supermarket" class="253"/>
<node id="1000463949" lat="51.5099809" lon="-0.2865237" n="Empire"
t="shop" st="supermarket" class="253"/>
<node id="1003093177" lat="51.5775031" lon="-0.3715401" n="Cycle King"
t="shop" st="bicycle" class="264"/>
```

From the original dataset, the features were extracted and the nodes were categorised. Each node now contains a unique id, latitude and longitude, name of the place, type of the place, the subtype of the place and class or category id. A class is a combination of a type and subtype. There are total 63545 locations divided between 1341 categories. A class/category is defined by a class id. Example of some class is given below:

| Type | Subtype | Class |
|---|---|---|
| Shop | Supermarket | 253 |
| Restaurant | Indian | 5 |
| Office | Estate Agent | 914 |

**Data Collection:**

The data is collected from a mobile application. The application asks user different questions about the context. There are six sections in the process of data collection such as the list of nearby places, activities, feelings, rating, emotion, whether he/she is alone or with friends and also the time and

| Day | Time | Activities | Feeling | With Friends | Emotion | Rating |
|---|---|---|---|---|---|---|
| Monday | Morning | Work/Housework | Optimism | YES | Liking | Poor |
| Tuesday | Afternoon | Preparing Meals/Eating | Love | NO | Joy | Average |
| Wednesday | Evening | Taking Medication | Submission | | Surprise | Good |
| Thursday | | Religious Observances | Aware | | Anger | Excellent |
| Friday | | Shopping | Disapproval | | Sadness | Extraordinary |
| Saturday | | Use of Telephone | Remorse | | Fear | |
| Sunday | | Community Mobility | Contempt | | | |
| | | | Aggressiveness | | | |

day in taken from the system to understand the daily and weekly patterns of his behaviours. The list of places comes from google place API and dynamic. The content of each section is as below:

XML representation a sample data is presented below:

```
<user>
    <user_id>1</user_id>
    <place_id>2699776232</place_id>
    <name>Burger King</name>
    <lat>52.4443114</lat>
    <lon>-1.495705</lon>
    <class>813</class>
    <time>MORNING</time>
    <day>MON</day>
    <activity>PREPARING MEALS/EATING</activity>
    <emotion>SURPRISE</emotion>
    <feeling>LOVE</feeling>
    <rating>1</rating>
    <with_friends>0</with_friends>
    <timestamp>2016-06-26 09:39:49.205</timestamp>
</user>
```

This data is then processed so that it can be trained using machine learning models. There are seven features and one target which makes eight labels in total. The above data can be represented in the following way:

| | Feature | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|
| Label | userid | day | time | friends | activity | feeling | emotion | class |
| Value | 1 | MON | MORNING | 0 | EATING | LOVE | SURPRISE | 813 |

From this input data, test data is generated for the decision tree to predict the class. Here, from the data we can see that the context is, the user in eating in a place which belongs to the class 813(813 is the class id for fast food) on Monday morning and he/she is loving the place, not with friends and he/she is also surprised. Since we know this information now, there is no reason to show anymore 'fast food' shops on the map, rather we would like to know where the person will likely to go next. From his previous data, we know what type of activities he/she gets to do and what his feelings and emotions are usually like on Monday morning and afternoon. From this information and previous data, we prepare test cases and apply a decision tree to predict the class. Below is a sample of the test data.

| | Feature | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|
| Label | userid (X1) | day (X2) | time (X3) | friends (X4) | activity (X5) | feeling (X6) | emotion (X7) | class (Y) |
| Value | 1 | MON | MORNING | 0 | EATING | LOVE | SURPRISE | ? |
| | 1 | MON | MORNING | 0 | HOUSEWORK | AWARE | LIKING | ? |
| | 1 | MON | MORNING | 0 | SHOPPING | AWARE | JOY | ? |
| | 1 | MON | NOON | 0 | WORK | OPTIMISM | SADNESS | ? |
| | 1 | MON | NOON | 0 | EATING | LOVE | SURPRISE | ? |

**Decision Tree:**

A decision tree is a predictive modelling approach widely used in machine learning and statistics for classification and regression. The tree is produced by observing labelled features in order to come to a conclusion about the target. It starts by breaking down the dataset into smaller subsets recursively until the tree is constructed. There is three type of nodes in the tree. The first node is called root node which corresponds to the best feature. All the other nodes that branch out from the root node, that have two or more branches are called the decision node. And the leafs of the tree are called the leaf nodes or target node which is the conclusion of that branch. The features from the table above(table number) can be represented as a tree as below.

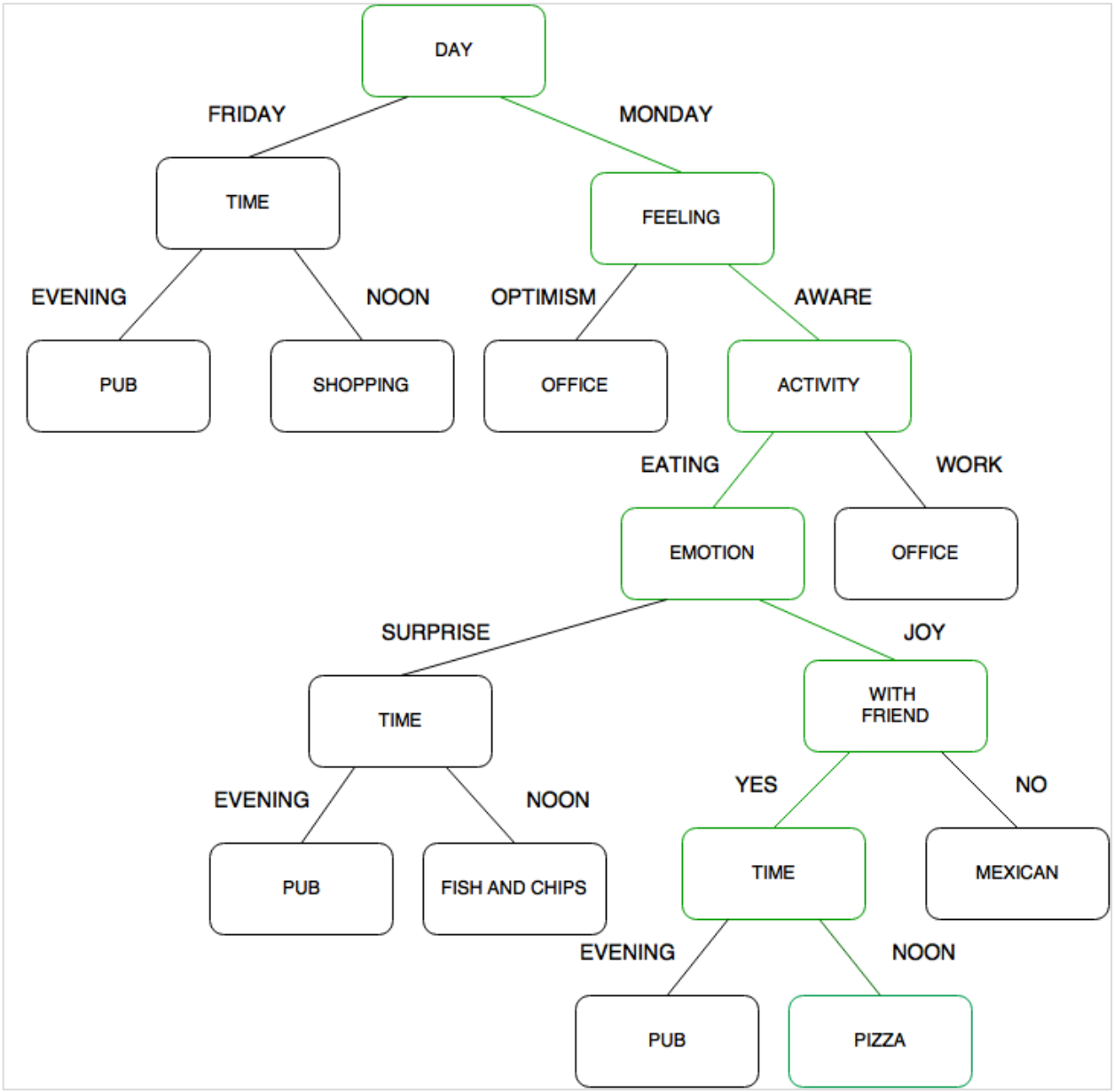



Figure 2: Decision Tree

This is a textual representation of a decision tree. The tree is generated during training and the features are encoded into numbers. From this tree we can see that if today is Monday and the person is aware and eat with friends, we can predict that the person will likely to go to a pizza shop in the afternoon. This is one case of many from the test cases and for each one of the case, a prediction is to be made. Different prediction cases might have a common prediction result.

The algorithm for constructing decision tree uses Entropy and Information Gain. The term entropy is used to indicate the impurity of the cases. A node needs to split into more branches until the entropy reaches zero. The entropy of a node can be calculated as:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

where P(i)s the probability of case i. In classification and regression tree algorithm (CART), Gini impurity is also calculated to measure the number of incorrectly labelled elements. The entropy of the leaf nodes is zero. However, from the impurity, we can measure the information gain at each level. Information gain is the difference in entropy of a node to that of its parent node. Information gain tells us the decreased amount of entropy at each node. Information gain can be formulated as:

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

For a multivariate problem like this one, the tree gets larger and more complex. To improve the prediction accuracy, the tree needs to be reduced in size. The process of selecting less effective sections and remove them from the tree is called pruning. Pruning helps to avoid over-fitting. Over-fitting occurs for complex models when the tree gets excessively large. Moreover, if the model is over trained that will cause over-fitting. Over-fitting causes poor prediction results as the models start to generalise the cases due to over-training. However, for large datasets, these problems still arise resulting the lower accuracy. To solve this problem and to get better results Random Forest has been applied to the dataset. Random Forest takes random samples from the dataset and creates a number of smaller trees instead of creating a big one. Random Forest belongs to a machine learning class called ensemble methods. Ensemble methods combine alternative machine learning models to produce a better result.

**Random Forest:**

The random forest proposed by Breiman in 2001 is an ensemble method. Ensemble method takes different weak learners together to produce one strong learner. In the case of the random forest, it takes decision tree classification model as the weak learner. It takes random samples from the dataset and produces a number of decision trees. From the figure below, during training, the dataset is split into T number of trees each with a sample of N random cases to create a group consisting of around 66% of the dataset. From each case, m number of features are selected in random which is used to perform binary split of the node. This process is repeated for all the nodes on all other trees. The accuracy and performance of random forest are better than decision trees in the case when the dataset is very large. Unlike decision tree, the model does not end up making one large tree which leads to over-fitting, rather a number of smaller trees which reduces the error.

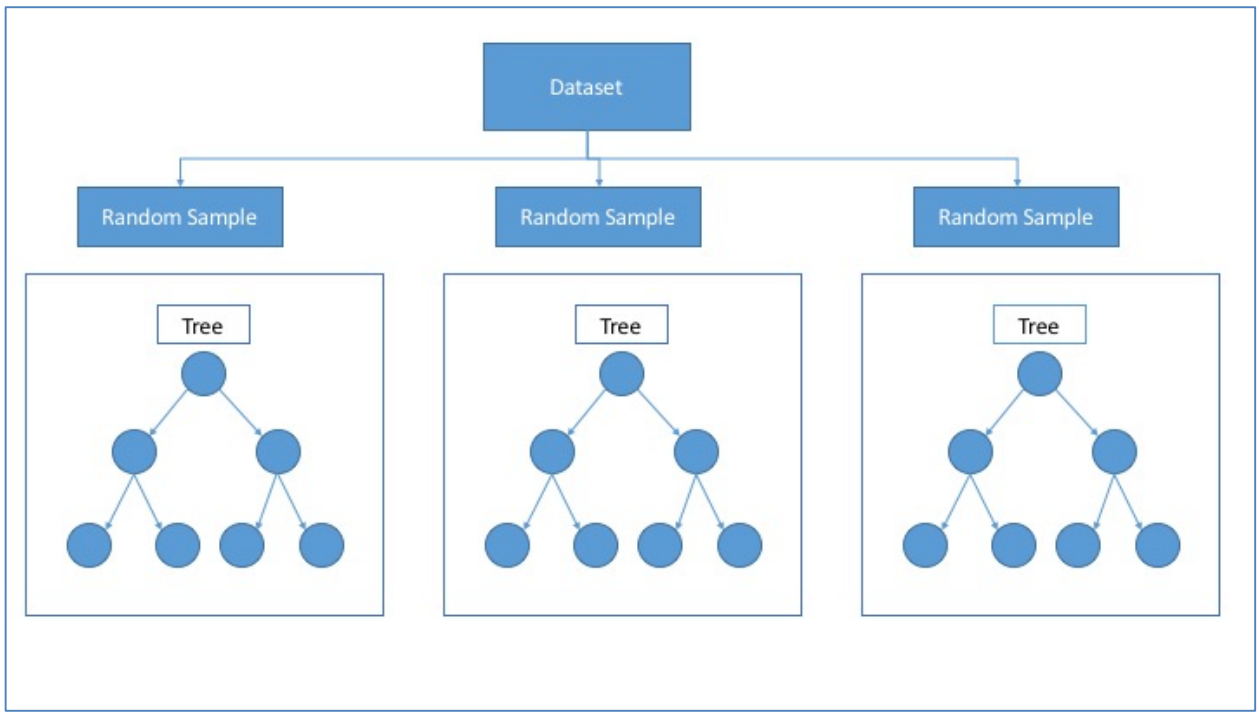



Figure 3: Random Forest

**Collaborative filtering:**

Collaborative filtering is a user based item recommendation model. It recommends an item to a user based on other user's preference if two of them has similar interests. The concept of collaborative filtering can be illustrated by the following table:

| User | Location 1 | Location 2 | Location 3 | Location 4 |
|---|---|---|---|---|
| 1 | Yes | No | Yes | Yes |
| 2 | - | Yes | No | No |
| 3 | Yes | Yes | No | |
| 4 | No | - | Yes | - |
| 5 | Yes | Yes | ? | No |

In the table the preference of 5 users for 4 locations has been plotted, whether a person likes or dislikes a location is represented by Yes and No. From this information, how likely it is for user 5 to like location 3. As we can see user 5, user 2 and user 3 likes location 2. And user 2 and user 5 dislikes location 4. Since user 5 has some common interest with user 2 and user 3 and none of them likes location 3 so it is likely that user 5 will not like location 3 as well. So location 3 will not be recommended. In this research, apache mahout has been used for collaborative filtering. The process is below:

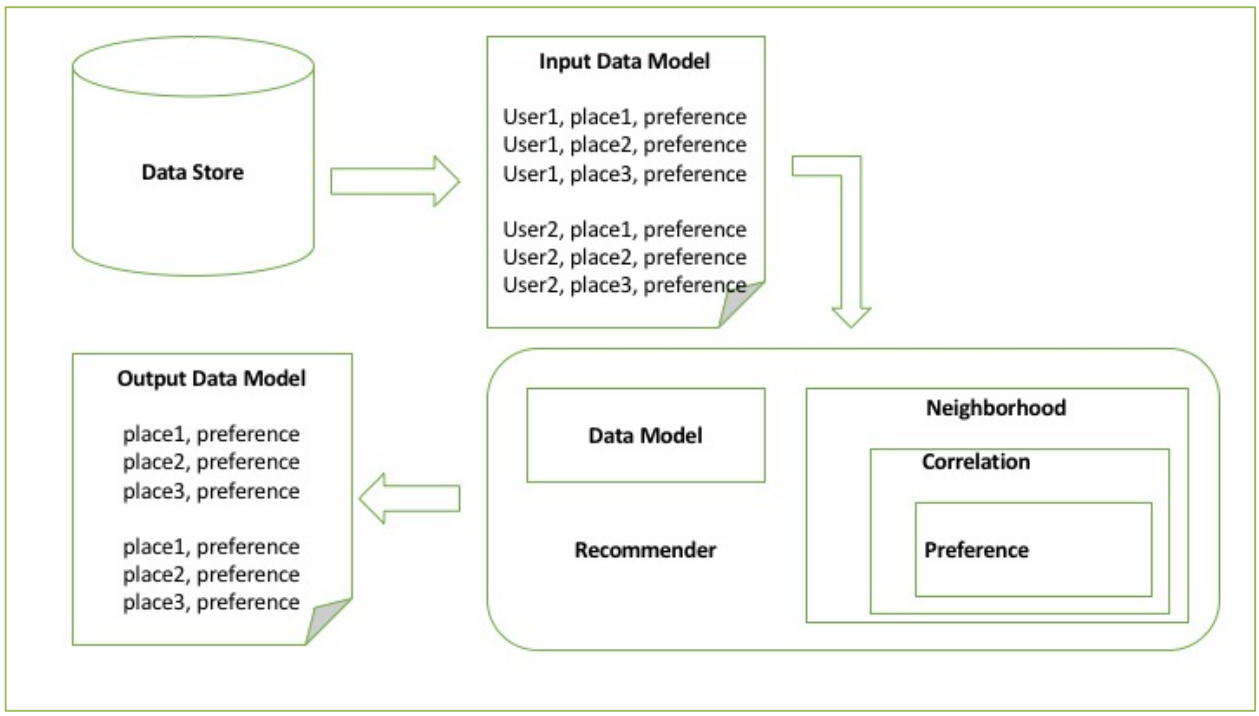



Figure 4: collaborative filtering

In order to apply collaborative filtering algorithm, the data has to be prepared in a particular way. The input data model was prepared by querying the database for the user and his friends' data. The was formatted into 'user id', 'place_id', 'preference'. The preference is numeric and usually 0,1 and greater than 1. All the places the user visited has a preference greater than 1 and the similar places to those places has preference 1. All the other places within 1km radius have been flagged by preference 0. Moreover, if the user gives a poor rating to a place that place will also be flagged as preference 0. The recommendation model is then applied to this data model. The model user similarity matrix with the users who have similar interests. It creates the user neighbourhood with N nearest users. The nearest is calculated from the user similarity matrix. The similarity can be formulated by:

$$\mathbf{simil}(x,y) = \frac{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})(r_{y,i} - \bar{r_y})}{\sqrt{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})^2 \sum_{i \in I_{xy}} (r_{y,i} - \bar{r_y})^2}}$$

Where $r_x$ and $r_y$ is the rating of item i by user x and user y. $r_{-x}$ and $r_{-y}$ is the average rating for user x and y. $I_{xy}$ is a set of item that was rated by both users. The model then calculates the preference by taking the weighted mean of all the rating of that user. Finally, the output data model is created in the form of 'place_id', 'preference'.

**Deployment Strategy:**

The system has been tested on a Hadoop distributed environment on a Linux machine with CentOS 6.0_x64 OS. To store data HBase distributed database storage system as been used which runs on top of Hadoop Distributed Fie System. Although the dataset used for testing was small but the system was made to be work in distributed environment as a proof of concept to demonstrate that the system can be scaled up and is capable of dealing with the data of entire planet and can handle millions of users in distributed computing environment.

However, there are other possible approaches that can be applied to store and analyze the data. One of the approaches is the use of Elastic Search storage system which can also store data in distributed fashion. Elastic Search was built on top of lucene text search engine. Lecene is a powerful search engine written in java, that uses hashing and indexing for faster data retrieval. Elastic Search would result a much better performance and scalability and can address the fault tolerance issue by replicating data across nodes. However, the reason for selecting Hbase for this project is that, Hadoop is the most popular distributed storage and distributed processing framework and Hbase is compatible with Hadoop ecosystem which allows random read/write operations on HDFS.

**Usability Analysis:**

A survey was conducted to understand how people use map applications and user's feedback on this application to evaluate the product. From the survey we get valuable insights about the product. We have learned that google maps is very popular among other existing map application and 83% of the population uses google map for searching for places and for navigation. We have also learned that 40% of the population uses public transport, 20% of the people uses cars and 40% of the population walk in their local area. 16.7% of the population always uses maps while commuting whereas 33.3% people never uses maps while walking or in public transport. However, 50% of the population uses maps occasionally and majority of this population uses car as a mode of transportation and generally uses the map for navigation. In the second part of the survey, people were asked some questions about the usability of this application. It is observed from the survey that 50% of the population is satisfied with the reliability, 66.6% of the population thinks the look and feel can be improved, 33.3% of population is concerned about privacy and security and 83.3% population thinks the application is easy to use. The volunteers were also asked to rate the product and the average rating turned out to be 3.3 in a scale of 5.

The insights acquired from the survey is presented below.

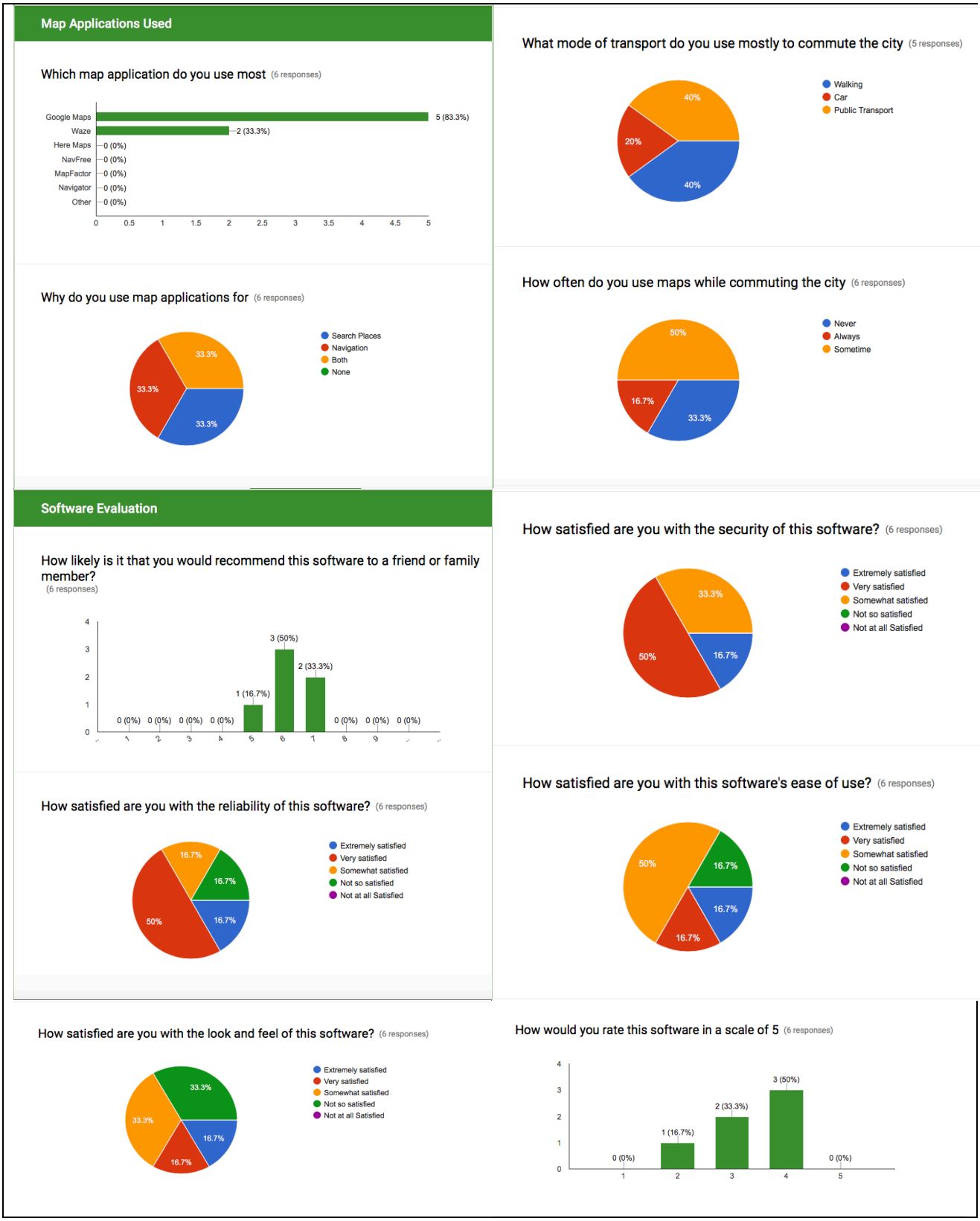



Figure 5: Software Evaluation Survey Report

# 6 Design

In this section we describe the system architecture, a high level functional sequence diagram, entity relationship diagram and the use case diagram.

System Architecture:

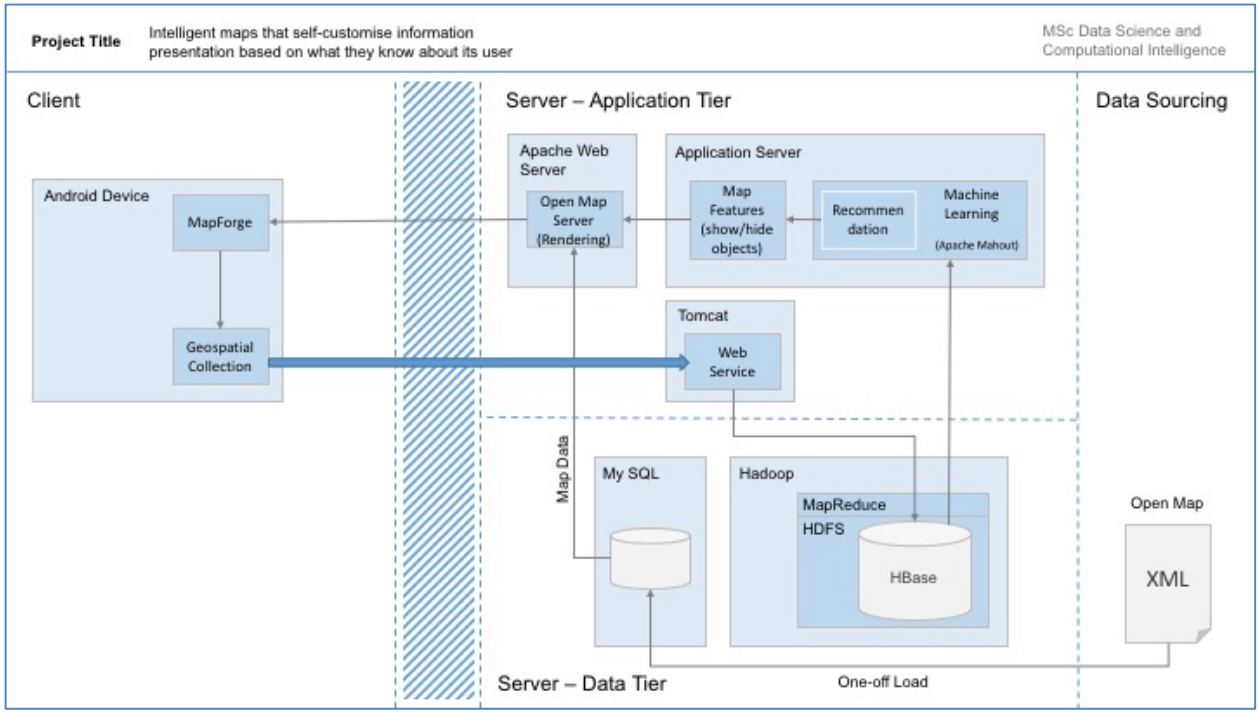



Figure 6: Architecture of the System

The client application consists of two subsystems, one is the map application and the other one is the data collection application. The map application is responsible for rendering the map which comes from the Open Map tile server. The data collection subsystem is a wizard application that collects various information from the user every time the user launches the map application. The systems ask the users a series of questions such as where he/she is now, what is he/she doing there, how does he/she feel about the place, whether he/she is with friends, his emotional status and how would he/she rate the place which indicates whether the person likes or dislikes the place. The purpose of this section is to learn about the context and to send this information to the recommendation model so that the recommended engine can contextualise the recommendations. The data is sent to the web service that stores the data in the HBase database. HBase is a distributed data storage system that runs on top of Hadoop Distributed File System (HDFS). This makes the system scalable and extendable. The application server layer, however, is responsible for querying data from the database, preparing data models, execution of the recommendation engine and finally sends the data back to the map through apache web server.

The functional sequence of the recommendation model can be described below:

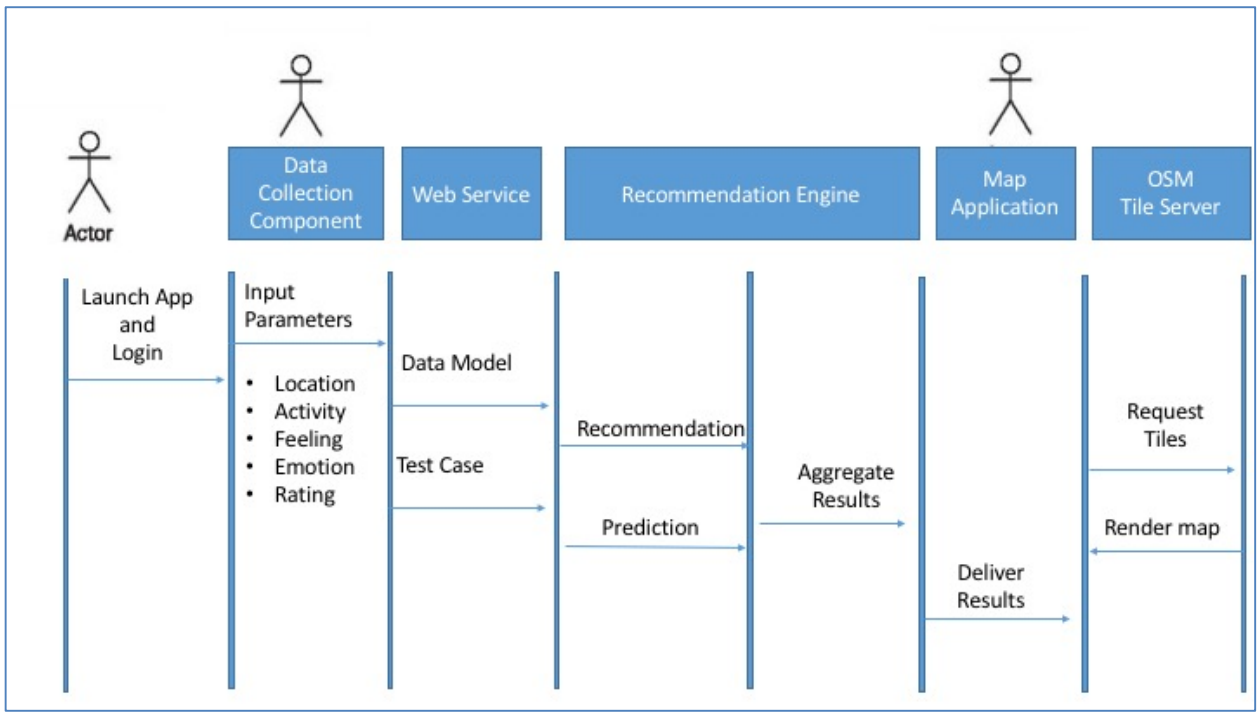



Figure 7: Functional Sequence Diagram

The client sends the input parameters to the web service which gets the data from the data storage and prepares the test cases and data models for Random Forest and Collaborative Filtering The input parameter are user id, day, time, place, emotion, feeling, rating. A sample data model for random forest is below:

| userid | day | time | friends | activity | feeling | emotion | class |
|---|---|---|---|---|---|---|---|
| 1 | SAT | MORNING | 0 | EATING | AWARE | SURPRISE | 812 |

For the test data model, the parameters are same, only the class is unknown which the prediction model will predict. And the data model for collaborative filtering is as below:

| User id | Place id | Preference |
|---|---|---|
| 1 | 469779385 | 2.0 |
| 1 | 566313869 | 1.0 |
| 2 | 2881796422 | 1.0 |

The prediction and recommendation results are then aggregated and is sent back to the client in Java-script Object Notation (JSON)Format. A sample JSON data is provided below:

```
[
    {
        “place_id”:” 2699776232”,
        “distance”:”0.1km”,
        “lon”:” -1.495705”,
        “name”:” Burger King”,
        “sub_type”:””
        “user_id”:”1”
        “type”:”fast_food”
        “lat”:” 52.4443114”
        “rank”:”1.7425”
    }
]
```

The client processes this data and displays the information on the map.

The entity relationship diagram of the database is described below:

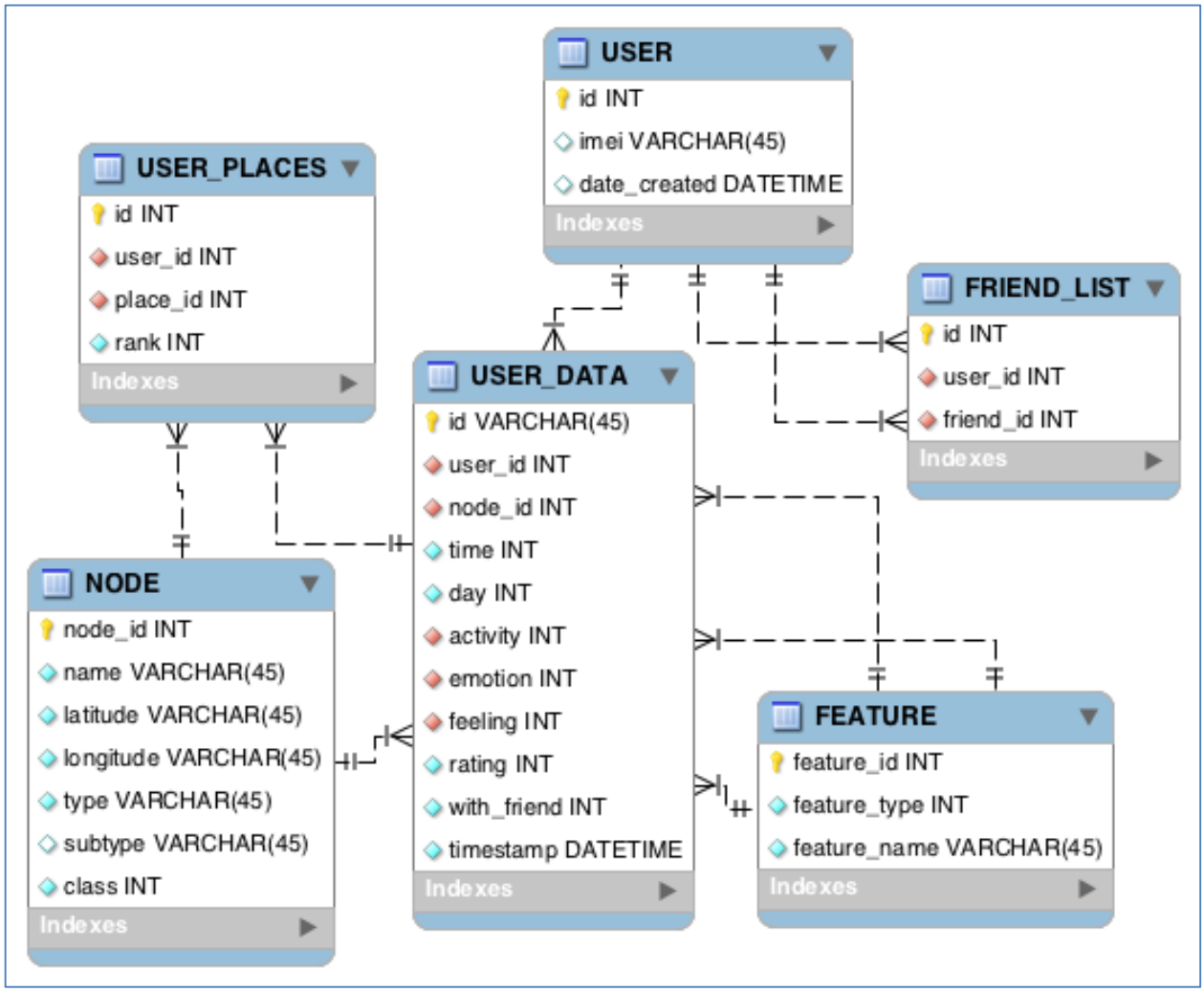



Figure 8: Entity Relationship Diagram

The database consists of six tables. Each user is identified by a unique id, in this case his IMEI number. The IMEI of the user is stored in the 'user' table which relates to the 'friend list' table. When a person is adding friends, his own id is mapped to his friend's id in the 'friend list' table. The 'User data' table contains the data sent from the data collection subsystem. The 'Node' table contains all the places of London and West-Midland. The recommendation engine recommends places from the 'node' table and the recommendations are stored in the 'user places' table. The feature table contains activity, feeling and emotion labels, identified by feature id which is the foreign key of user data table. The features are separated by feature type. For activity the feature type is 1 for emotion 2 and 3 for feeling.

Use case diagram:

The data is collected through a mobile app designed to learn the user's current environment Below a use case diagram for the data collection process is described.

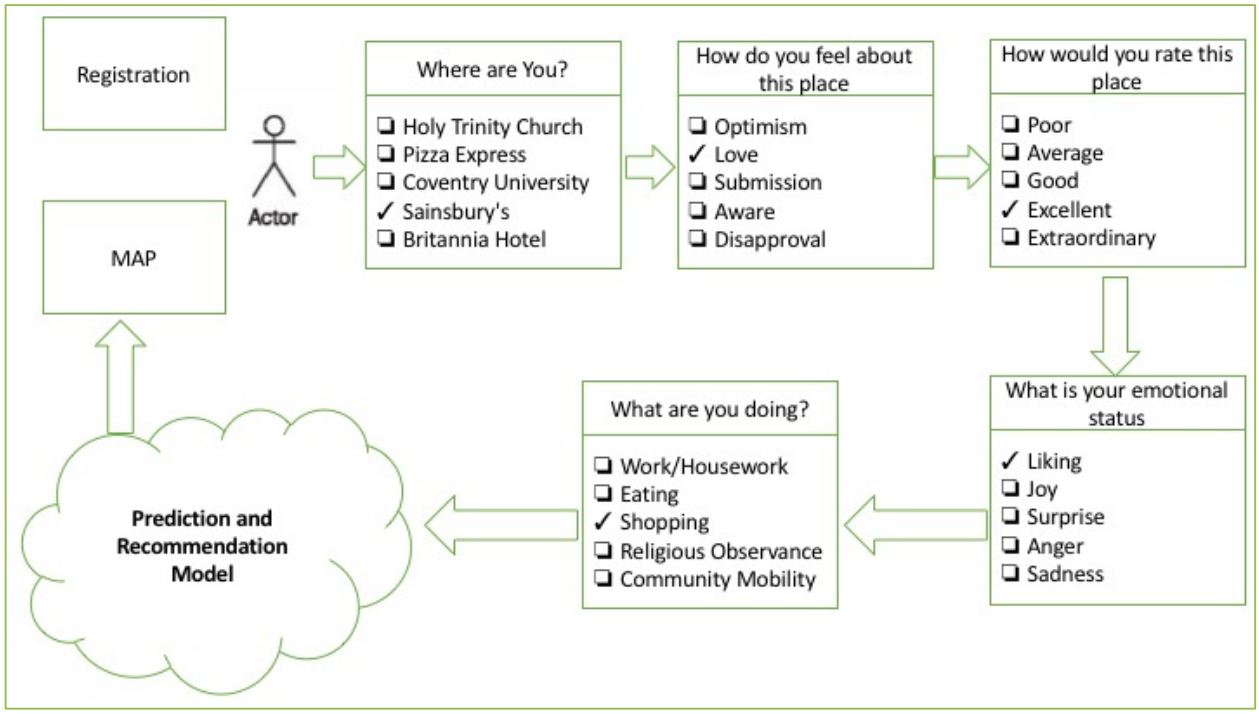



Figure 9: use-case diagram

The user functionality is divided into two components, the data collection application and the recommendation system. The registration process is automated and it takes only the IMEI number to uniquely identify the user. The user is then subjected to answer some question through a wizard application. In the above figure, a sample selection has been shown. The sections are about collecting the user's geo-location, feeling, emotion, rating, activity. Though each section has around five items, the user can add more item if the present items don't describe his current context. The data is sent to the server through the web service where recommendation engine recommends

places that are appropriate to that context and the result is sent back to the user. The second subsystem is the recommendation engine that translates the response from the server and put colourful markers on the map to highlight the recommended places on the map. The markers represent the priority of that place.

# 7 Implementation

The project is divided into two major components. The data collection component and the recommendation component. The recommendation process can be described by the diagram below:

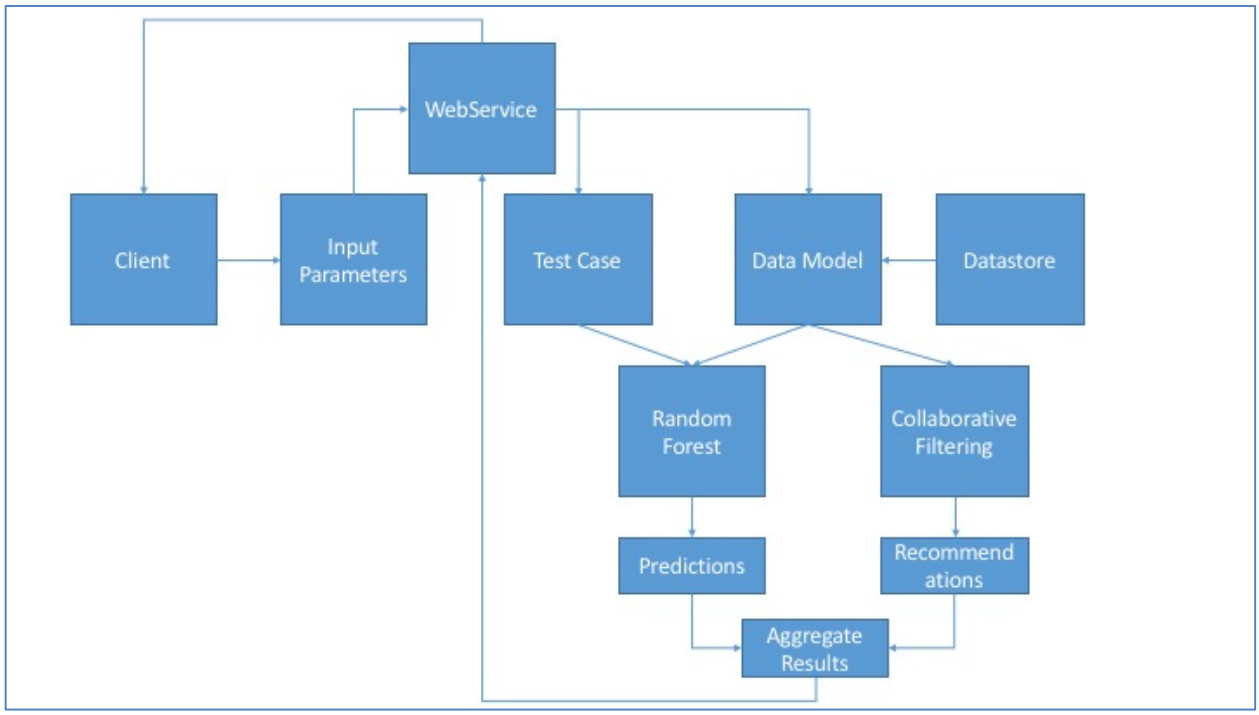



Figure 10: Functional flowchart of the system

The data is collected through an android application. When the user first launches the app, he/she has to go through a series of multiple choice questions, such as his/her current geo-location, activity, feeling, emotion, rating of that place and whether he/she is with friends. The questions are required to understand his/her current context. The application sends these inputs to the web service which prepares test cases and data models for random forest and collaborative filtering algorithm. The random forest is a predictive model used here, implemented using python and sci-kit learn machine learning library. The random forest is trained with previous user data and is tested with the test data created from the inputs, to get some predictions. The model predicts a class which is a combination of type and subtype i.e. if type is restaurant and sub-type is Chinese then all the Chinese restaurants are represented by a class id. The prediction tells us what type of

place the person will likely to go in near future. For example, if the current context is, the person is at work on Monday morning and he is sad, from this information the prediction model will predict what type of places the person will be interested in the afternoon.

Moreover, a data model for collaborative filtering is prepared to recommend new places. The data model is consisting of three components, user id, place id and preference. The preferences were set from the user's visits to a geolocation. If the person has visited a place before the preference would be greater then 1, depending on the number of time the person visited that place. All the places that are similar to those places within 1km radius have a preference 1 and all the other places within 1km radius have a preference 0. The preference being more than 1 means that the place is highly recommended for that person, preference 1 mean that the place is recommended and has normal priority and preference 0 means the place is not recommended to the person. The process can be described using the flowchart below:

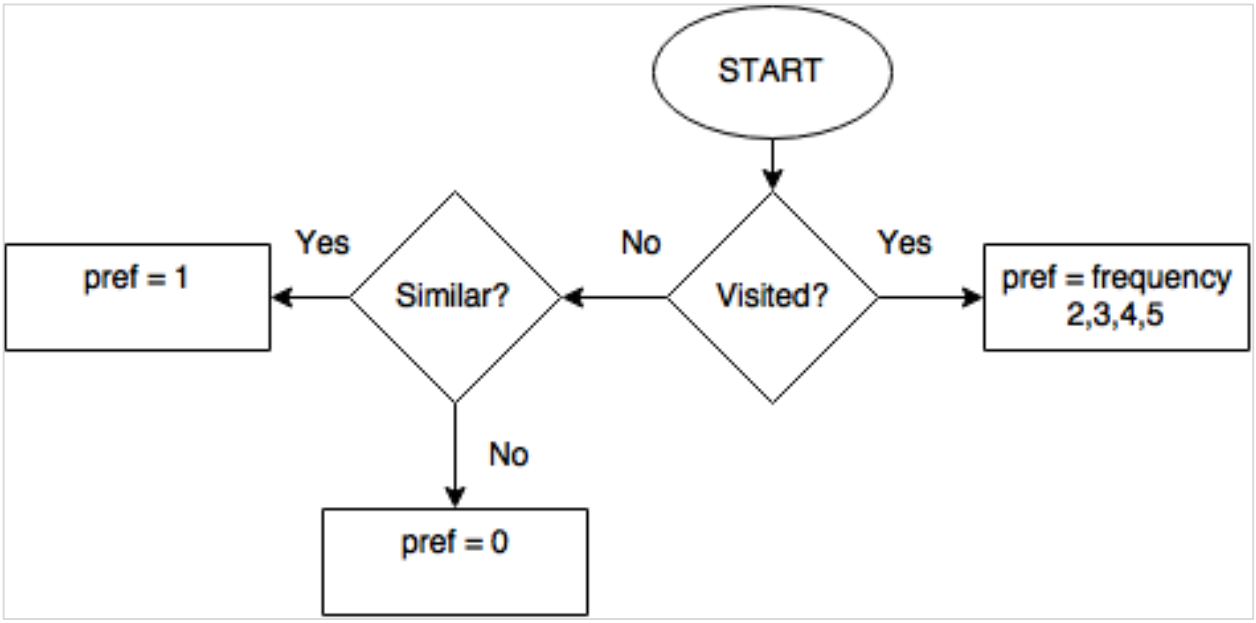



Figure 11: Flowchart for determining preference from visits

We apply collaborative filtering which recommends new places that the person might be interested in. Both of the results from prediction model and recommendation model are aggregated and sent back to the user. The map application receives this result and puts coloured markers on the map to highlight the places.


Description:

The 'MapsForge' supports different modes of map rendering methods. In this research I am using DownloadLayerViewer class which loads the map from OpenStreetMap server.

The map sends request to the application server which responds back the recommendations in JSON. The map then parses the JSON structured data and puts colorful markers to highlight the recommended locations. Green and Orange color represents highly recommended and recommended respectively.

Figure 12: Screenshot of the map

However, the map is rendering from third party server. To display map, open-source map framework, 'MapsForge' is used in this project. The map sends a latitude and longitude to the OpenStreetMap tile server. The Tile server prepares small tile images with embedded labels and sends the tiles to the application. When the user zooms in or out on the map, new tile images are created and rendered to the application.

Various tools and resources used in this project is discussed below:

**Platform**: The data collection component and the map application was developed in android platform. Both application supports a minimum version Android 2.3.3 (API level 10). The application service was deployed on a Linux server (Cent OS 6.0_x64), with Apache web server 2.4 and Apache Hadoop 2.7.2 configured. Also to store data HBase database storage system was configured in the server.

**Languages**: Both java and python was used for different components. Both the data collection app and the map application was written in java programming language and was build using maven. Java runtime environment version 1.7.0_79 is required to build the project. The prediction model was implemented in Python programming language using sci-kit learn machine learning library. Again the HTTP web service API and the recommendation model was written in Java. Also a

python library called HappyBase was used to access the Hase data store through Thrift API. The recommendation engine was written in java using Apache Mahout machine learning library.

**Tools**: To build the Java projects (Data collection and map application), Eclipse (Luna version 4.4.2) IDE (Integrated Development Environment) has been used. The project was built using maven android deployed on android smartphone (tested on android simulator WVGA4.2, API level 17, CPU architecture ARM (armeabi-v7a)). The source was built using AAPT (Android Application Packaging Tool). The source code was scanned using Lint for probable bugs. Once built, it will produce an APK file which needs to be deployed in the phone or simulator using ADB (Android Debugging Bridge). There are five dependencies that needs to be added in the dependency list. The dependencies are:

- Android app-compact v7 (android user interface style compatibility library)
- Android-support-v4 (requires for backward compatibility of latest user interface elements)
- Robotium-solo-4.0 (android user interface testing tool)
- Kxml-2.2.3 (XML parsing utility for java)
- Androidsvg-1.2.0 (bitmap processing utility for android)

For python projects Anaconda Navigator, a graphical user interface for conda package manager has been used with spyder IDE, a scientific python development environment. The source codes are compatible with python 2.7.

**Methodology**: The software development methodology for this project has been chosen to apply software prototyping. The software evolved with three versions. In the first version of the software, an algorithm was built to find places similar to the user visited places. In the second version of the software, collaborative filtering was applied to find recommendations and recommend new places. Later the location data was classified and each location was represented by a corresponding class. In the third and final version, human emotion, feeling, activity was collected through a mobile application and machine learning model like decision tree and random forest was applied to predict the type of place (class) the person will go in near future.

**Status**: A full end to end prototype of the app is ready to be deployed.

**Deployment strategy:** A Linux virtual private server was borrowed momentarily to test the deployment environment.

# 8 Testing

Classification and Regression Tree (CART) algorithms are used for predictive modeling. In this research both decision tree and random forest algorithms has been investigated. The machine learning models i.e. Random Forest and Collaborative Filtering has been tested and evaluated using machine learning model evaluation method K-Fold Cross Validation. However, for evaluating the performance of machine learning models a method called K-fold cross validation technique is applied.

Cross validation is a technique used for model validation by partitioning the dataset into test set and training set and measuring the accuracy by comparing the predicted values with the actual values. In this research the dataset acquired from the users was fairly small. The data was collected from six volunteers in Coventry over a period of a week. The dataset contains 126 instances and seven features. The features are id, day, time, feeling, emotion, activity and if he/she is with friends. The dataset was split into 33% testing subset and 66% training subset and K-fold cross validation was applied to measure accuracy for different values of k (k=10, k=20, k=30). The dataset is partitioned into K number of segments of equal size and each segment is validated. This process is repeated and the root mean squared error and standard deviation are calculated at each iteration. From the average of the error, the accuracy is calculated. Since the cross validation takes random sample from the dataset so the accuracy is not constant. The best achieved accuracy is approximately 65%. Below the cross validation results are plotted in the table:

| Cross validation | K=10 | K=20 | K=30 |
|---|---|---|---|
| Random Forest | 40.256% | 53.452% | 64.667% |
| Decision Tree | 11.11% | 48.463% | 55.631% |

The model was further evaluated by constructing confusion matrix. A confusion matrix or error matrix is a tabular representation of the true positive, true negative, false positive and false negative, used to analyze the data in more detail. The structure of a confusion matrix can be described from the table below:

| Confusion Matrix Construction | Predicted NO | Predicted YES |
|---|---|---|
| Actual NO | TRUE Negative | False Positive |
| Actual YES | False Negative | TRUE Positive |

Here the true negatives and true positives are accurately classified instances and false negative and false positive are the error or misclassified instances. The accuracy can be calculated by accuracy = (True Positive + True Negative)/ Total Number of Instances. The confusion matrix constructed from the actual dataset is presented below:

| | | Prediction | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | class | 2 | 245 | 251 | 253 | 297 | 397 | 411 | 785 | 809 | 834 | 913 | 916 | 1156 | 1175 | support | Precision | Recall | Accuracy |
| Actual | 2 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0.4 | 0.67 | 66.666667 |
| | 245 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0.5 | 1 | 100 |
| | 251 | 0 | 2 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0.67 | 0.67 | 66.666667 |
| | 253 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0.75 | 0.75 | 75 |
| | 297 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| | 397 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.5 | 1 | 100 |
| | 411 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 |
| | 785 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 1 | 100 |
| | 809 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 5 | 1 | 0.2 | 0.2 |
| | 834 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| | 913 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0.2 | 0.33 | 33.333333 |
| | 916 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 1 | 100 |
| | 1156 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 3 | 0.75 | 1 | 100 |
| | 1175 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | 1 | 0.75 | 75 |
| | | | | | | | | | | | | | | | | total: 42 | average: 0.66 | average: 0.62 | Overall Accuracy = 61.9047619 |

Figure 13: Confusion matrix

**Analysis of the confusion matrix:**

From the matrix we can see there are 14 classes and 42 instances in the test subset. Support tells us the frequency of a class in the dataset. The diagonal elements (green colored) represents the true positive and true negative values. Here, class 2 was classified correctly 2 out of 3 times. The precision tells us the ratio of the occurrences of the instances that are relevant and recall tells us the ratio of the correctly classified instances. The columns are the predicted values and the rows are the actual values. Here class 2 has appeared 5 times in the during prediction. The precision for class 2 is precision = correctly classified occurrences/ total occurrences = 2/5 = 0.4

And recall = correctly classified occurrences / support = 2/3 = 0.67

And the accuracy was calculated by,

accuracy = total correctly predicted instances/ total number of instances = 61.9 %

Moreover, this model was compared with other machine learning models such as K-nearest neighbor, decision tree algorithm (CART), and Support Vector Machine. The performance comparison graph is below:

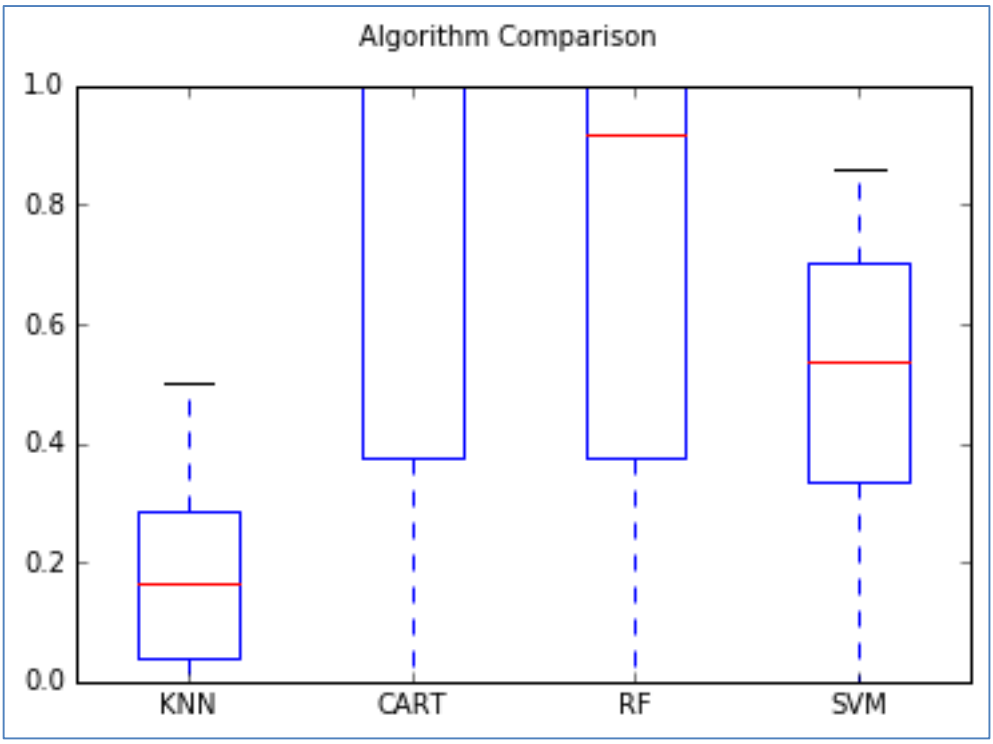



Figure 14: Comparison

The accuracy for different models:

| Model | Accuracy | Standard Deviation |
|---|---|---|
| KNN | 0.185714 | 0.152753 |
| CART | 0.683333 | 0.411299 |
| RF | 0.666667 | 0.401386 |
| SVM | 0.480952 | 0.294854 |

From the comparison we can see that both K-nearest neighbor and Support Vector machine has below 50% accuracy. However, Decision Tree and Random Forest has 68.3% and 66.6% accuracy respectively. From this comparison we can conclude that tree algorithms are suitable for this type of problem. Although decision tree produced better results than random forest, we cannot decide which one is better because a small dataset was used here. However, by surveying related literatures (Maqsood I. et al. 2012) we have learned that for large dataset, decision tree gets too big and complex resulting over-fitting and poor accuracy. However, Random Forest addresses this issue by taking random sample from the dataset and producing a number of smaller trees.

**Collaborative Filtering Evaluation:**

In this research collaborative filtering algorithm has been used to implement user based recommendation system. Collaborative filtering algorithm predicts rating of an item from measuring the user similarity. The similarity can be measured from Pearson Correlation Similarity

$$sim(u,v) = \frac{\sum_{i \in I_{uv}} (y_{u,i} - \hat{y}_u)(y_{v,i} - \hat{y}_v)}{\sqrt{\sum_{i \in I_{uv}} (y_{u,i} - \hat{y}_u)^2 \sum_{i \in I_{uv}} (y_{v,i} - \hat{y}_v)^2}}$$

Where sim (u,v) is the similarity between user u and v. Iu,v represents list of items both users have rated, Yu,i represents the rating of item i rated by user u and Yv,i represents the rating of item i rated by user v. And finally y^ represents the average rating of that user. From this equation we calculate the similarity for every pair of users and produce a similarity matrix for every user. From the similarity measure we can predict the rating a user would give to an item.

$$y^*(u,i) = \hat{y}_u + \frac{\sum_{j \in I_{y_{*j} \neq 0}} sim(v_j, u)(y_{v_j,i} - \hat{y}_{v_j})}{\sum_{j \in I_{y_{*j} \neq 0}} |sim(v_j, u)|}$$

Here Y*(u,i) represents predicted rating for item i and user u. To evaluate collaborative filtering algorithm, we split the dataset into 90% training and 10% testing set. We apply the above equation to predict the rating and calculate error by using average absolute difference recommender evaluator. The testing process can be illustrated below:

| User ID | Place ID | Actual Rating | Predicted Rating | Absolute Error |
|---|---|---|---|---|
| 1 | 234 | 1 | 1.2 | 0.2 |
| 1 | 756 | 3 | 2.9 | 0.1 |
| 1 | 8788 | 2 | 2.2 | 0.2 |
| | | | average absolute error | 0.167 |

Absolute error is calculated by taking the absolute difference between predicted and actual rating.

Average Absolute Error = SUM( |predicted rating – actual rating| )/ Number Of Instances

Absolute error for all the items in the test set is calculated and their average is the average absolute different which tells us the average error. The absolute error achieved from the recommender applied in this project is 0.1797573220296521

# 9 Project Management

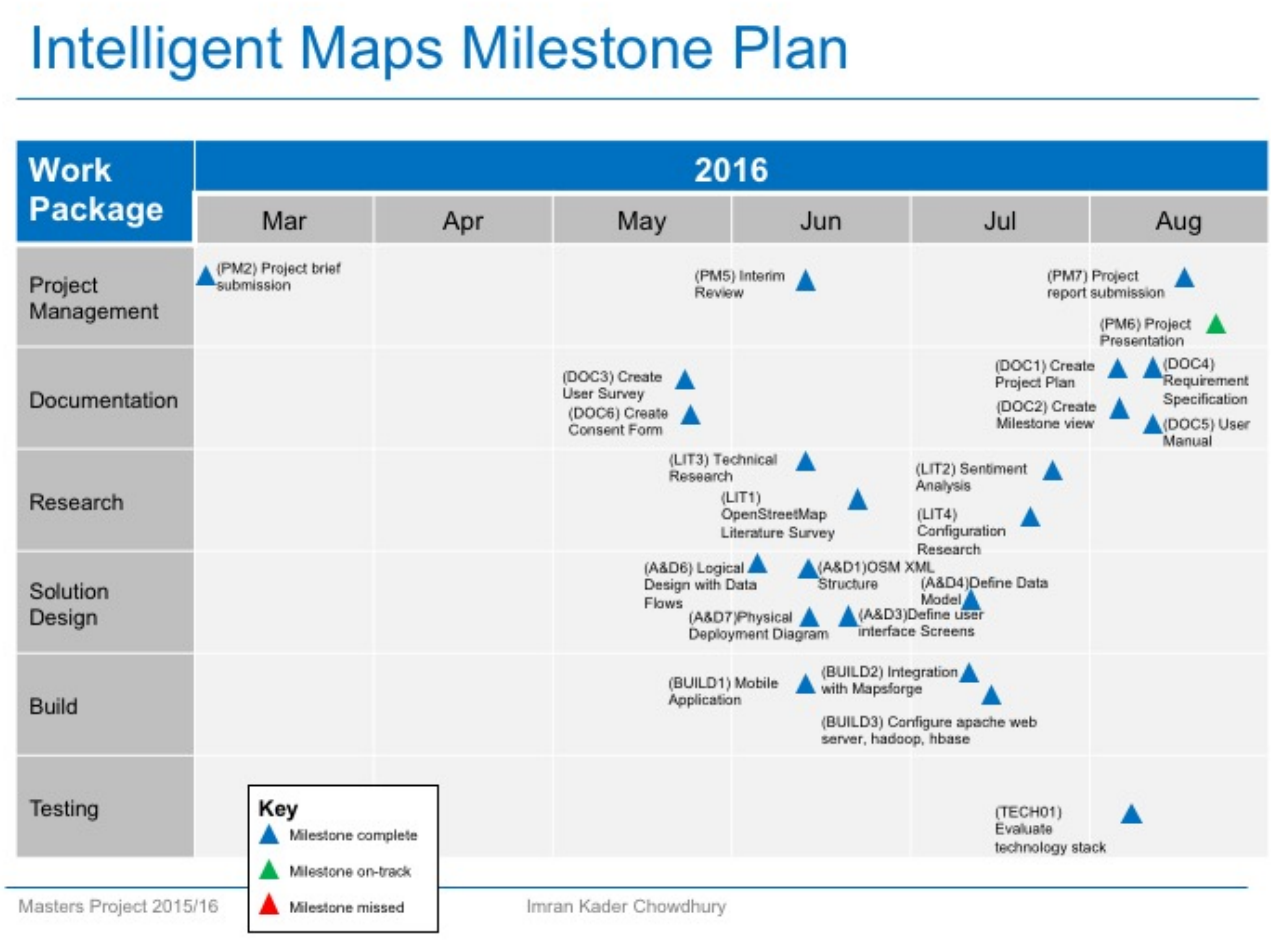



Figure 15: Project Milestone

## 9.1 Project Schedule

ject Plan: intelligent maps that self-customise information presentation based on what they know about its user

| Phase | Task ID | Task | Description | Effort | Start Date | Planned Date | Confirmed Date | Dependency | Status | Responsible |
|---|---|---|---|---|---|---|---|---|---|---|
| ect Oversight & Control | | | | | | | | | | |
| Project Management | Key Milestones | | | | | | | | | |
| | PM1 | Allocation of Supervisor | Identification of projects of interest and associated supervisory staff | 1.5d | 04-Jan | 01-Feb | 01-Feb | N/A | Complete | IKC |
| | PM2 | Project Brief submission | Creation and submission of project brief template | 2.0d | 01-Feb | 29-Feb | 29-Feb | PM1 | Complete | IKC |
| | PM3 | Ongoing project tracking | | 4.0d | 10-Jun | 28-Jul | 28-Jul | DOC1, DOC2 | Ongoing | IKC |
| | PM4 | Fornightly review with Supervisor | Ongoing review sessions with Supervisor | 2.0d | 10-Jun | 28-Jul | 28-Jul | PM1 | Complete | IKC, FD |
| | PM5 | Interim Review | | 1.0d | 13-Jun | 13-Jun | 13-Jun | PM1 | Complete | IKC, FD |
| | PM6 | Project Presentation | | 1.0d | 15-Aug | 15-Aug | 15-Aug | N/A | On Track | IKC |
| | PM7 | Project Report Submission | | | | | | | On Track | IKC |
| Documentation | Key Milestones | | | | | | | | | |
| | DOC1 | Create Project Plan | Creation of this project plan template | 1.0d | 07-Aug | 09-Aug | 09-Aug | N/A | Complete | IKC |
| | DOC2 | Create Gantt Chart/Milestone view | Creation of a plan on a page showing major milestones | 1.0d | 08-Aug | 10-Aug | 10-Aug | N/A | Complete | IKC |
| | DOC3 | Create user survey | Creation of survey using google form | 1.0d | 26-May | 26-May | 26-May | N/A | Complete | |
| | DOC4 | Requirements Specification Document final version | This is a sub-section of the main report | 1.0d | 10-Aug | 10-Aug | 10-Aug | N/A | | |
| | DOC5 | User Manual final version | This is a sub-section of the main report | 1.0d | 09-Aug | 09-Aug | 09-Aug | N/A | | |
| | DOC6 | Create applicatin use consent form | | 1.0d | 28-May | 28-May | 28-May | N/A | Complete | IKC |
| | DOC8 | Project Report - publication version | | | | | | | | |
| ject Implementation | | | | | | | | | | |
| Research | Implementation Methodology Selection | | | | | | | | | |
| | IM1 | Identify research topics | Identify those items that will require literature review | 2.0d | 10-Jun | 14-Jun | 14-Jun | N/A | Complete | IKC |
| | IM2 | Identify pilot user base | | 1.0d | 12-Jun | 12-Jun | 12-Jun | N/A | Complete | IKC |
| | IM3 | Evaluate and Select method for gathering user requirements | | 1.0d | 15-Jun | 15-Jun | 15-Jun | N/A | Complete | IKC |
| | IM4 | Define approach for technology selection | Development and deployment strategy | 1.0d | 18-Jun | 18-Jun | 18-Jun | N/A | Complete | IKC |
| | IM5 | Evaluate and Select build approach | E.g. evaluation of waterfall, agile or prototyping approaches that will be reactive to new learning | 1.0d | 20-Jun | 20-Jun | 20-Jun | N/A | Complete | IKC |
| | IM6 | Implementation approach defined | | | | | | | Complete | IKC |
| | Literature Review | | | | | | | | | |
| | LIT1 | Open Street Map feature analysis | Data Structure and validity | 10.0d | 20-Jun | 10-Jul | 10-Jul | N/A | Complete | IKC |
| | LIT2 | Sentiment analysis | Predict sentiment from social media and sensor data | 10.0d | 20-Jun | 10-Jul | 10-Jul | N/A | Complete | IKC |
| | LIT3 | Technical research | Machine learning prediction and classification modelling, location aware recommendation system, collaborative filtering | 10.0d | 20-Jun | 10-Jul | 10-Jul | N/A | Complete | IKC |
| | LIT4 | Configuration research | Research into installation and configuration of Hadoop, Hbase, HDFS, Thrift, Python, Map Forge | 10.0d | 20-Jun | 10-Jul | 10-Jul | N/A | Complete | IKC |
| | LIT5 | Core research completed | | | | | | | Complete | IKC |
| | Requirements Analysis | | | | | | | | | |
| | REQ1 | Conduct end user survey | | 1.0d | 01-Aug | 01-Aug | 01-Aug | DOC3 | | |
| | REQ2 | Use case analysis | | 1.0d | 04-Aug | 05-Aug | 05-Aug | N/A | | |
| | REQ3 | Incorporate end user survey into own requirements capture | | 1.0d | 06-Aug | 06-Aug | 06-Aug | N/A | | |
| | REQ4 | Requirements analysis finalised | | | | | | | | |
| | Technology Selection | | | | | | | | | |
| | TECH01 | Evaluate appropriate technology stack and | | 1.0d | 07-Aug | 07-Aug | 07-Aug | N/A | Complete | IKC |
| | TECH02 | Finalise implementation technologies | | | | | | | Complete | IKC |
| Solution Design | Analysis | | | | | | | | | |
| | A&D1 | Analayse Open Street Map XML data structures to identify salient attributes | Data analysis and Data pre processing | 4.0d | 01-Jun | 10-Jun | 10-Jun | N/A | Complete | IKC |
| | A&D2 | Define JSON format structure to persist location preference data | | 1.0d | 20-Jul | 20-Jul | 20-Jul | N/A | Complete | IKC |
| | A&D3 | Define user interface screens | | 2.0d | 10-Jul | 14-Jul | 14-Jul | N/A | Complete | IKC |
| | A&D4 | Define data models for predictions and recommendation engine | | 2.0d | 16-Jul | 20-Jul | 20-Jul | N/A | Complete | IKC |
| | A&D5 | Key analysis items influencing design | | | | | | | Complete | IKC |
| | Design | | | | | | | | | |
| | A&D6 | Create application logical design with data | | 1.0d | 06-Jun | 08-Jun | 08-Jun | N/A | Complete | IKC |
| | A&D7 | Create physical deployment diagram | | 1.0d | 06-Jun | 08-Jun | 08-Jun | N/A | Complete | IKC |
| | A&D8 | Major design components defined | | | | | | | Complete | IKC |
| Build | Install, Configure, Write | | | | | | | | | |
| | BUILD1 | Write mobile application (based on Android | | 15.0d | 10-Jun | 26-Jul | 26-Jul | N/A | Complete | IKC |
| | BUILD2 | Integrate mobile application with MapForge | | 5.0d | 20-Jul | 28-Jul | 28-Jul | N/A | Complete | IKC |
| | BUILD3 | Install and configure Apache web server | | 1.0d | 14-Jun | 18-Jun | 18-Jun | N/A | Complete | IKC |
| | BUILD4 | Install and configure Hadoop | | 1.0d | 14-Jun | 18-Jun | 18-Jun | N/A | Complete | IKC |
| | BUILD5 | Install and configure Hbase | | 1.0d | 14-Jun | 18-Jun | 18-Jun | N/A | Complete | IKC |
| | BUILD6 | Write predictive analysis component | | 6.0d | 20-Jul | 28-Jul | 28-Jul | N/A | Complete | IKC |
| | BUILD7 | Write recommendation component | | 6.0d | 10-Jul | 18-Jul | 18-Jul | N/A | Complete | IKC |
| | BUILD8 | Core components of application built | | | | | | | Complete | IKC |

EY

- Completed activity
- On track for completion
- Delivery likely to slip
- Due date missed
- Task not started

- Major milestone achieved

Roles

| | |
|---|---|
| IKC | Imran Kader Chowdhury |
| FD | Dr Fayez Doctor |

## 9.2 Risk Management

During the planning and development phase, several issue raised. Some issues were mitigated and some did cost few days to find an alternate solution. The risk factors can be divided into four sectors such as risks regarding development strategy, deployment strategy, data collection strategy and ethics and privacy conservation strategy. Each of this risk areas are discussed below:

**Development Strategy:** The map application could be achieved either by using Google map API or by using open source solutions. Although the project was inspired by Google maps, Google maps was not used for having restrictive access to its core features and flexibility of user interface customization, rather more time was devoted on research and development of 'MapForge', an open source map solution for OpenStreetMap. Choosing the open source solution helped mitigate the risk of using google map which could have been more complex and could have delayed the schedule.

Another risk factor was about selecting server side technology stack. OpenStreetMap has an open source server side application known as OSM Tile Server that is responsible for rendering the map. The tile server produces small segments ('tile images') of the map and sends these small images to the client application layer, the client side application puts these tile images together to construct the map. The initial proposed plan was to integrate the recommendation engine layer with OSM server. However, considering the amount of time and resources required for the integration, we decided to use OSM server as a third party service provider and keep the recommendation engine as an independent application layer.

**Deployment Strategy:** Choosing the right deployment platform was an important part of the research. We proposed to use the EMR(Elastic Map Reduce) service from AWS(Amazon Web Services). However, it took some days to familiarize with AWS and configure the web services. Later, due to high-scale pricing of the service we decided not used AWS and borrowed a virtual private server for a month instead. This risk factor did cost ten days.

**Data Collection Strategy:** Initially an automated data collection process was proposed through a background service of the mobile application. The service would send user's geo-location along with sensor data from gyroscope and accelerometer, to the web service at a specific interval. It was proposed that from the mobile phone sensor data we will predict someone's activity, feelings and

emotions. However, by surveying some literature in this field it was discovered that it is impossible to predict someone's activity, feelings and emotions from the data of gyroscope and accelerometer of the phone. However, from these sensor, we can predict whether a person is walking or sitting, standing etc. But this information is not sufficient to predict the activity of the person.

It was also proposed to collect geo-location through a background service at a small interval and to predict where the person is currently in. However, from a continuous stream of latitudes and longitudes, it would be very difficult to filter location and find the important ones. It is also impossible to know whether the person really went to the place or he is just passing by it. This approach also has serious ethical and privacy issue since we are collecting data in background.

To mitigate these risks, new data collection method was proposed later. To collect data, a mobile application was developed where the user will be asked about his current geo-location, his activity, feeling and emotion. The ethical issue has been mitigated since there is no longer any hidden background service and the quality and accuracy of the data was improved since we no longer need to predict the activity, feeling and emotion of a person instead these input are coming straight from the user.

**Ethics and Privacy conservation strategy:** There were some ethical risks in this project since we are collecting personal data such as geo-location, activity, feeling and emotion from the user. However, since there is no signup process in the app, user's name, email, phone number etc are not taken. Each user is identified by a unique number, in real life scenario that would be the IMEI number. During the research, to avoid the ethical and privacy risks the volunteers were asked to sign a consent form (attached in appendix F).

## 9.3 Quality Management

Software quality was evaluated by conducting some surveys and by reviewing literatures related to evaluating accuracy and quality of the Open Source projects used during the development. Moreover, the recommendation and prediction model was evaluated using k-fold cross validation technique. Different aspects of quality management are discussed below:

**User Survey:** A software evaluation survey was conducted among a group of six volunteers. It has been observed from the survey, 66.6% of the population is satisfied with the reliability of the software, 33.3% of the population is satisfied with the look and feel of the software. The volunteers were asked to put a rating on the app which resulted the average rating 3.3 in a scale of 5. From

the survey analysis we can conclude that there are enough room to improve the overall quality of the product.

**Open Source Projects**: This project was built on top of some open source projects. To render the map, we are using OSM Tile server which is an open source third party server. For the client side, we are using MapsForge which is an open source map application for android platform. Both projects are in active development and are maintained on a regular basis and many industrial projects and academic researchers are using it.

**OpenStreetMap Data**: OpenStreetMap data is collected from volunteers all over the world. Portable GPS devices with accuracy of +-7 meters are used to collect geodata. To evaluate the quality of OpenStreetMap data we have investigated some literatures. According to Open Data Ordnance Survey 2010, OpenSteetMap has successfully captured 30% area of England and and 4% of the data are digitized without attributes (Haklay M. et al. 2010). The OSM data was processed and the location was classified. There are total 63545 locations in the dataset categorized between 1341 classes.

**Software Evaluation:** K-fold cross validation technique was used to evaluate the Random Forest model. The accuracy of the model was calculated to be 65.238%. To evaluate the Collaborative Filtering recommender system, a method called average absolute difference recommender has been applied and the absolute average error was calculated to be 1.79

## 9.4 Social, Legal, Ethical and Professional Considerations

A part of the software requirement specification was to collect sensitive personal data from the user. A mobile application was developed to meet the requirement criteria. The app asks the user some questions regarding his/her current location, feeling, emotion and activities associated to a location. Although this type of data poses much risk toward someone's privacy, user's consent is taken for the app to reserve the right to use this content. When the app is first launched, the user is notified that the app is going to collect some data and the user has to accept the terms in order to proceed (Appendix C). The data however is collected from the user inputs which makes the data reliable and the user is aware of their actions. Moreover, the app is not taking any user identification information such as name, email or phone number, since there is no sign up process. Each user is identified by a unique number which in real life scenario is the IMEI (International Mobile Equipment Identity). For this research, the data was collected from some volunteers and their consent was taken at the beginning of the process (Appendix F).

# 10 Critical Appraisal

This project has provided me the opportunity to learn new tools and technology. I had the opportunity to learn and explore Java and Python development environment. I was introduced with android application development lifecycle. I learned how to build and deploy android project. During the development I came across different android user interface components such as List View, View Pager and Material Design Style. I have also acquired the knowledge of parsing XML and JSON data and to use REST API. This project gave me the opportunity to explore different machine learning techniques such as Decision Tree and Random Forest. I have also learned to test and evaluate these models. I was responsible for creating my own data set from real users and prepare data models. I learned to process large data files in distributed environment. I have learned how to apply machine learning on real life use cases using sci-kit machine learning library. I came across python libraries like pandas and numpy. I also had the opportunity to explore distributed computational frameworks such as Hadoop and Map-Reduce. I have acquired introductory knowledge about Amazon Web Services such as Amazon Elastic Cluster (EC2) and Amazon EMR (Elastic Map Reduce). I have gained introductory knowledge about HBase distributed Database management system. Most notably, I have gained expertise on building recommendation engines using Apache Mahout machine learning library. I have been familiarized with Linux based system.

This research has also improved my project management skills. I have practiced software development methodologies. I learned different stages of software development such as requirement analysis, solution design and development life cycle. I have learned to work under pressure to meet the deadlines. I have also learned to make surveys and analyze valuable insights from the survey. I also learned how to do literature research and apply the knowledge earned from the research in practice.

# 11 Conclusions

This report proposed a contextualized map application which recommends geolocations to a person by analyzing his previous data such as previous visited places, activity, feeling and emotion associated with those places. Data was collected from six volunteers over a period of a week which is insufficient to firmly come to any conclusion. However, the model was able to predict locations with 65.23% (approximately) accuracy. Although the functionality of the whole system has been tested and evaluated, it requires further evaluation in the future with a larger dataset.

## 11.1 Achievements

To collect data from the user, a mobile app has been developed which takes inputs from the user such as his/her geo location, activity, feeling and emotion. I have applied Random Forest machine learning model to predict where the person will go in near future. Also a recommendation engine using collaborative filtering model has been applied to recommend new places based on similar user's preferences. Both of these model was evaluated to measure to what extent the system can predict and recommend locations. This leads to the following research questions:

1. To what extent this recommendation model can recommend locations to a person from similar user's preferences.

   Answer: A recommendation model has been developed to find user similarity using Pearson Correlation Similarity and to find recommendations using Collaborative Filtering Algorithm. The recommender system was evaluated using average absolute error evaluator resulting 17% error. Based on this data, we can conclude, the model can 83% accurately recommend locations from user similarity.

2. To what extent the machine learning model can predict where a person will go by analyzing their personalized data such as feelings, emotion and activities associated with time, day and geolocation.

   Answer: CART algorithms were used in this research for predicting geolocations. However, Different machine learning models such as K-nearest neighbor, Decision Tree, Support Vector Machine and Random Forest has been compared. Although Decision tree algorithm had 68% accuracy it was concluded that Random Forest is suitable model with an average 65.23% accuracy.

## 11.2 Future Work

This project has the potential for further research and development. There are different parts of the project that can be made more sophisticated if more time and effort is invested. The future version of the project should have automated data collection process. Many different approaches can be useful for developing an automated system. Some wearable sensors can be used to capture the pulse rate, blood pressure, oxygen level, glucose level, chemicals in perspiration and by analyzing

this information, the person's emotion can be predicted. Also by capturing the photo of the individual and analyzing facial features and analyzing nodes and pitches from his/her vocal features, the person's emotion and feeling can be predicted using the artificial neural network. This area of the project needs to be further researched. Moreover, predicting human feeling and emotion from mobile phone sensor data (gyroscope and accelerometer) as proposed needs further investigation.

The app can be made a social app by adding features like social media integration, messaging, location sharing and navigation. Moreover, the user interface can be made interactive, and features like rating, review, and arranging meeting places will be a remarkable addition to the existing system. Some proposed attributes such as orientation and movement speed were not addressed in this research since these features are related to navigation. Orientation represents which direction the person is facing, and movement speed represents whether the person is walking or in a car. In the future, these features need to be incorporated into the system to provide a more robust and real-time recommendation.

The project can be enhanced to a great extent by adding the features mentioned above which will make the app a competitor to the existing commercial map applications. A tool can be developed to analyze the geodata of the entire planet to get valuable insights about the geographical information of the whole world. The map can be made more user-centric by targeting specific user groups such as a map for the cyclist or tourists.

# 12 Student Reflections

During the research and development of this project I came to learn different tools and technologies such as Java and python programming languages, android application development and deployment systems, making web services, applying machine learning models in real life user generated data and about recommendation systems. I have also learned big data platforms such as Apache Hadoop and machine learning libraries like sci-kit and Apache Mahout. I had the opportunity to apply data science knowledge such as SQL and NoSQL database management system, machine learning, configuring virtual server and REST API. The project requirements and specification was delivered in a very high intellectual fashion, and took a long time to determine the actual scope of work. Moreover, it took a significant amount of time to understand which of the proposed features are feasible and can be accomplished with the available time and resources. Inefficient time management was another issue I have faced during application development which delayed testing during the process. Too much time was devoted into research and development in contrast to the time spent on documentation. The infrastructure platform choice and configuring deployment environment also consumed significant amount of time. The application for collecting data from the user was delayed resulting smaller amount of dataset and poor data quality.

## Bibliography and References


1. Mooney, P., Corcoran, P. and Winstanley, A.C., 2010, November. Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 514-517). ACM.

2. Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design*, *37*(4), pp.682-703.

3. Neis, P., Zielstra, D. and Zipf, A., 2011. The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, *4*(1), pp.1-21.

4. Jokar Arsanjani, J., Helbich, M., Bakillah, M. and Loos, L., 2015. The emergence and evolution of OpenStreetMap: a cellular automata approach. *International Journal of Digital Earth*, *8*(1), pp.76-90.

5. Helbich, M., Amelunxen, C., Neis, P. and Zipf, A., 2012. Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata. *Proceedings of GI_Forum*, pp.24-33.

6. Bermingham, A. and Smeaton, A.F., 2010, October. Classifying sentiment in microblogs: is brevity an advantage?. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1833-1836). ACM.

7. Pak, A. and Paroubek, P., 2010, May. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc* (Vol. 10, pp. 1320-1326).

8. Barbosa, L. and Feng, J., 2010, August. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 36-44). Association for Computational Linguistics.

9. Gamon, M., 2004, August. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 841). Association for Computational Linguistics.

10. Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R., 2011, June. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (pp. 30-38). Association for Computational Linguistics.

11. Mohtarami, M., 2013. *From Semantic to Emotional Space in Sense Sentiment Analysis* (Doctoral dissertation, NATIONAL UNIVERSITY OF SINGAPORE).

12. Perikos, I. and Hatzilygeroudis, I., 2013, September. Recognizing emotion presence in natural language sentences. In *International Conference on Engineering Applications of Neural Networks* (pp. 30-39). Springer Berlin Heidelberg.

13. Binali, H., Wu, C. and Potdar, V., 2010, April. Computational approaches for emotion detection in text. In *4th IEEE International Conference on Digital Ecosystems and Technologies* (pp. 172-177). IEEE.

14. Ali, J., Khan, R., Ahmad, N. and Maqsood, I., 2012. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5).

15. Yin, H., Sun, Y., Cui, B., Hu, Z. and Chen, L., 2013, August. LCARS: a location-content-aware recommender system. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 221-229). ACM.

16. Horozov, T., Narasimhan, N. and Vasudevan, V., Horozov Tzvetan T, 2005. *Location-based recommendation system*. U.S. Patent Application 11/141,121.

17. Ricci, F., 2010. Mobile recommender systems. *Information Technology & Tourism*, *12*(3), pp.205-231.

18. Ricci, F., 2010. Mobile recommender systems. *Information Technology & Tourism*, *12*(3), pp.205-231.

19. Noulas, A., Scellato, S., Lathia, N. and Mascolo, C., 2012, December. Mining user mobility features for next place prediction in location-based services. In *2012 IEEE 12th International Conference on Data Mining* (pp. 1038-1043). IEEE.

# Appendix A – Project Specification

## Appendix A1: MASTER PROJECT BRIEF FORM 2015-17

**To be submitted by:**

1. Your details:

Full Name: Imran Kader Chowdhury
Student ID: 6694003
E-mail: chowdh62**@coventry.ac.uk**
Module Code (M08CDE/M29CDE/M95CS): M08CDE
Course of Study: Data Science and Computational Intelligence
Project Supervisor: Dr. Faiyaz Doctor

2. Project title (provisional) [Meaningful, relevant and concise]

Intelligent maps that self-customise information presentation based on what they know about its user

3. Outline (synopsis) of your project.[ What are the aim and objectives of the project?]

**Overview:**

This project was inspired from a talk in the GigaOM Roadmap 2013 conference in San Francisco where two senior designers working on Google Maps discussed about the present and future development of the product. When Google launched the map service, the user experience was much like exploring a paper map with very limited features such as dragging the map, zooming in and out. When the map was introduced in GPS enabled smartphones more location aware applications were coming up along the way and that brought a whole new level of user experience. Maps were no longer used just to explore places but also to navigate and trying not to get lost which has now become a basic feature and people are using it almost every day. Although the user experience has been improved a lot since then the map should be made more user specific, data-driven and contextualized.

**Aim:**

The aim of this project is to create a mobile map application that is more user specific and contextualized. Since we can gather information about the place user likes based on his/her location history and based on the places he clicks on the map, rated or reviewed, and based on the places his/her friends likes to visit, an algorithm can be built that can generate a map that are contextually more relevant to the user and choose to show the right information at the right time.

**Objectives:**

There are five key objectives of the project to achieve the aim described above:

1) **Data Collection**: The map data I intend to use is OpenStreetMap, an open source public dataset. To collect data from the user a mobile application has to be developed which will send a user's location to the server along with accelerometer and gyroscope sensor data. User engagement such as clicks, searches and time spent on each screen will be recorded to analyse and evaluate the product and surveys will be conducted through the application to get user feedback on the app experience. Also a user's social network may be accessed to collect ratings and reviews of the user and his/her friends on a place.

2) **Analysis:** User's latitude and longitude will be recorded with date-time along with other sensor data such accelerometer and gyroscope. Also how a user interacts with the map will be recorded and his/her own and friend's ratings and reviews on places in the social media will be stored. Various analytical operations will be conducted on the gathered dataset to find the places a user is interested in most, time spent there, his movement such as standing, walking from the sensor data and how often he visits a place. Based on this collective information's a ranking system will be built that will show the most relevant content on the map based on the context.

3) **Design:** This will be based on a cloud server where big data platforms such as Hadoop will be setup to store data and to process and analyse the data using MapReduce and mahout machine learning library will be used.

4) **Implementation:** The project includes a mobile application which will send required data to the cloud and OpenStreetMap which is an open source geolocation dataset. User's geolocation data will be stored in the Hadoop distributed file system which will be later analysed and processed through a recommendation engine to tailor the labels and the information will be sent to map server which will present the customized map to the user.

5) **Evaluation:** To evaluate the final product a survey has to be arranged in the application which will help to get user reviews and feedback. Also how a user interacted with the application will be recorded to assess how well the application kept the user engaged.

4. Intended user or group of users and their requirements. [ a) Who is the intended user or group of users? b) Why you think there is need for this project? c) What are the needs of the intended user that your product should satisfy?]

a) **Who is the intended user or group of users?**

The application can be targeted to a wide range of user groups such as students, tourists, cyclist, activists and potentially any user using a mobile map application. For the purpose of this study it will be University students.

b) **Why you think there is need for this project?**

The traditional mobile map applications presents all sort of data to the user regardless of whether the informations presented is relevant at that time. The current mobile map application serves the purpose of the model and user experience to discover and explore places around someone. By combining the user data with the location data a better map can be built that can learn about the user and adapt the user's needs based on a context and present informations that the user might be interested in. So a more user specific and personalized map will be generated for each user which will lead to a much better user experience.

c) **What are the needs of the intended user that your product should satisfy?**

The maps has become an essential tool, people use it to learn about places, discover new places and to navigate. With current technology and data a map can do more. By contextualizing the map it will present more relevant informations to the user at a specific place and time. The user doesn't have to search for anything since the location works as a query and it learns where the user will be going and what other place he might visit along the way.

5. Systems requirements and project deliverables. [ a) What are the characteristics/properties that the final product should possess? b) What are the process stages and the corresponding deliverables that will enable you to create the final product?]

a) **What are the characteristics/properties that the final product should possess?**

The product will be able to operate in the field. The data will be stored in the server and the system should be scalable. It should be able to store and manage large amounts of data and analyse the data to produce useful insights. It should be able to supply the user context map data to concurrent users in real time.

b) **What are the process stages and the corresponding deliverables that will enable you to create the final product?**

- The project will implement the software development life cycle: requirements analysis/gathering, use case development, architecture design, detailed design, POC development, detailed implementation, unit/system testing, UAT deployment, UAT testing, UAT results analysis and design/package modification, product finalisation
- The project may employ an agile development approach to develop certain parts of the architecture in response to changing requirements or design feasibility

6. Research [ a) How will you investigate/identify in detail the needs of the specified user in (3) b) How will you investigate the background of the project?]

a) **How will you investigate/identify in detail the needs of the specified user ?**

An online survey will be created and conducted across different type of user to investigate their needs. For this study a survey will be conducted among the students of Coventry University.

b) **How will you investigate the background of the project?**

To investigate the background, research will be conducted on existing map service providers to examine how they tackle the same problem set. Also technology journals and collecting information regarding what research has been done on building a ranking engine on spatial data and user behaviour analysis from sensor data and social media will be useful.

7. Evaluation. [ a) What makes a product successful? b) How will you demonstrate that your product fulfils the needs of the user in (3)? c) How will you evaluate the product?]

a) **What makes a product successful?**

The user defines whether a product is successful. How a user interacted with the application will be recorded to investigate how the application kept the user engaged

b) **How will you demonstrate that your product fulfils the needs of the user in (3)?**

Some survey and feedback should be conducted through the application to learn if the application successfully fulfils his/her needs.

c) **How will you evaluate the product?**

To evaluate the product a survey has to be conducted to collect user feedback and review on the product. Also some research has to be done to compare the product with the existing solutions.

8. Development skills. [ a) What information and resources do you need to complete the project successfully? b) Which of these do you need to acquire yourself? ]

a) **What information and resources do you need to complete the project successfully?**

- **Problem set and solution research:** online blogs, academic papers
- **Mobile app:** Java using tools such as IntelliJ
- **Data collection interface:** Web service implemented using Java and running on a Tomcat container
- **Data Storage:** Hadoop unstructured data store, Apache Spark to provide scalability
- **Data Analytics:** MapReduce implemented in Python with results exposed via web service interface

b) **Which of these do you need to acquire yourself?**

I need to learn how to store and analyze data stored in Hadoop distributed file system. I also need to do some research on recommendation engines and learn how to implement that using python.

9.Skill acquisition. [How do you intend to gain the skills, information and resources specified in (7)?]

To gain the skills I intend to talk to the experts in the university, my supervisor, internet resources such as blogs and open source communities, online tutorials and how-to guides for technology implementation and academic papers.

10. Estimate the number of hours you are planning to spend for each of the following tasks:

| | |
|---|---|
| Background research and learning new skills | 80 |
| Requirements gathering and investigation | 160 |
| Product design | 80 |
| Product development | 160 |
| Product evaluation | 40 |
| Final report preparation | 80 |
| Other (Please specify) | 00 |
| **Total number of hours** | 600 |

These categories should be changed to suit the type of project, namely product s/w development or investigation /

You **must** have completed the Ethics Online Procedure. Failure to comply will result in you not continuing with your project hence automatically fail your project. The grade for the Project Brief will only be awarded if the Ethics Online Procedure has been completed.

This form must be submitted electronically on your Moodle project web ***before*** 23:55 on the due date.

**I have completed my Ethics Application through the online CU Ethics system: NO ☐**

Signature ..... Date .....

COM PG Projects (M08CDE)

| Project Coordinator use only: |
|---|
| |

**Appendix A2: MARKING CRITERIA & FEEDBACK FOR MASTER PROJECT BRIEF**

| **Student Name:** | | **Student ID:** | | **Module Number:** | M08CDE | **Marks** |
|---|---|---|---|---|---|---|
| 1. Every project must have a clearly identified product (e.g. piece of software or proposed design). A list of deliverables should be provided, the delivery of which will constitute the objectives of the project. These should clearly identify the stages of the project and be sufficiently challenging. (item 5) | | | | | | ___/20 |
| 2. An appropriate requirements gathering and research methodology should be identified. (item 6) | | | | | | ___/15 |
| 3. Criteria for the evaluation of the product and the process should be clearly stated. (item 7) | | | | | | ___/15 |
| 4. A description of the skills, resources and knowledge already gained from previous study, as well as a list of any resources needed, modules to be studied and appropriate self-study should be provided. This should be fully comprehensive and realistic. (items 8 & 9) | | | | | | ___/10 |
| 5. The project should have a suitable title (item 2).<br>The synopsis, with aims and objectives, should be informative and relevant (Item 3). | | | | | | ___/20 |
| 6. Overall quality of the project brief, e.g. level of creativity, ability to encapsulate the project idea and degree of challenge. Accuracy and clarity of presentation. | | | | | | ___/15 |
| 7. A reasonable, sensible estimate of the time needed to complete all activities successfully should be given. (item 10) | | | | | | ___/5 |
| **I've approved and completed the Ethics procedure: NO** | | | | **Total Mark by supervisor** | | **___/100** |

**Assessor's Feedback**

| **Strengths of the project:** |
|---|
| |
| **Weaknesses of the project:** |
| |

COM PG Projects (M08CDE)

| **Areas in need of improvement:** |
|---|
| |

| **Assessor's name:** | | **Assessor's signature:** | | **Date:** |
|---|---|---|---|---|
| Moderator's Comments: | | | | Agreed Mark: /100 |

# Appendix B – Interim Progress Report and Meeting Records

| Date | Attendees | Agenda | Task Allocation | Due | Comment |
|---|---|---|---|---|---|
| 19/01/20016 13:00 | Dr. Faiyaz Doctor, Imran Kader Chowdhury | Discussed about the project proposal | Project supervisor allocation, Dr. Faiyaz, Imran Chowdhury | 29/02/2016 | N/A |
| 11/03/2016 14:00 | | Discussed about project features and concepts | N/A | 18/03/2016 | N/A |
| 18/03/2016 16:00 | | Discussed about different possible approaches | N/A | 21/04/2016 | N/A |
| 21/04/2016 14:30 | | Discussed about possible solutions | N/A | 11/05/2016 | N/A |
| 11/05/2016 16:00 | | Discussed about different implementation approaches | N/A | 31/05/2016 | N/A |
| 31/05/2016 15:00 | | Absent | | | |
| 10/06/2016 15:00 | | Discussed about project plan, data collection methods | N/A | 16/06/2016 | |
| 16/06/2016 15:00 | | Project update | Imran Kader Chowdhry | 28/06/2016 | |
| 28/06/2016 13:00 | | Project update | Imran Kader Chowdhury | 06/07/2016 | |
| 06/07/2016 16:00 | | Project update | Imran Kader Chowdhury | 18/07/2016 | |
| 18/07/2016 13:30 | | A user based recommendation system using collaborative filtering was implemented and showed to the supervisor | To contextualize the recommendations by adding features like activity, feeling and emotion, Imran Chowdhury | 27/07/2016 | Further investigation required about testing and evaluating the recommendation system |
| 27/07/2016 14:00 | | Discussed about project report structure | Write Project Report, Imran Chowdhury | 15/08/2015 | Final version of the project was shown to the supervisor. Data collection app was modified to collect activities, feeling and emotions. An implementation of the prediction model using Random Forest |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | was shown. The HTTP API was shown for server client communication. Data collection app was integrated with the map. The supervisor suggested to make the Data collection app automated and try to predict people's activities, feeling and emotion from mobile phone sensor (gyroscope and accelerometer) |

# Appendix C – User Manual


1

# Introduction

'My Social Map' is a location based recommendation system and map application. Each time the app is launched the user will be asked some questions to collect personal data such as geo-location, activity, feeling and emotion. Please choose the best suited answer from the options to get a better recommendation. Please sign in the consent form below and accept the terms and conditions of you agree to allow the app to keep record of these information.

To use the app you need to turn on internet connectivity wi-fi/mobile data and location service.



# Consent Form

**PERSONAL DATA PROTECTION ACT**
**CONSENT FORM**

1. In compliance with the Personal Data Protection Act ("PDPA"), I give my consent to collect and use my personal data (i.e. Name, contact numbers, mailing and email addresses) in order to perform research.

2. I also give consent to collect data such as geospatial data( latitude, longitude) through my mobile phone for the research.

3. I agree to take part in various surveys required for the research.

4. I agree for this application to collect my activities, feelings, emotion, rating associated to my geo-location.

5. By installing the application on my mobile phone I also agree to give access to my social networks (if requested) as long as I am assured that my personal data will be kept secured according to PDPA. Moreover I should be able to view my data upon request. I also give consent to make the data available online for further studies for others as well.

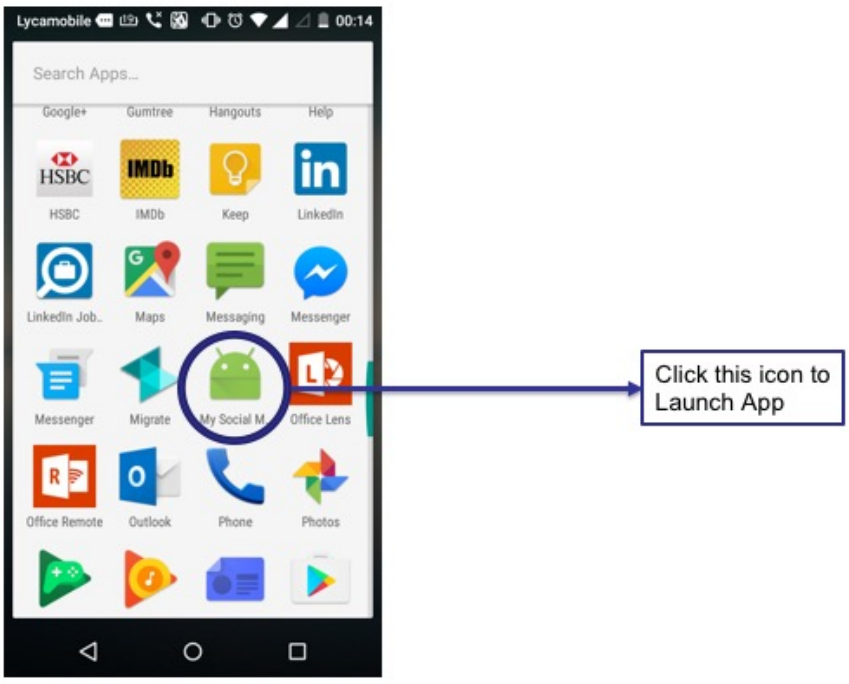
6. I hereby give my acknowledgement and consent to Imran Kader Chowdhury to use my personal data for his research.

7. I agree that my consent will remain in place until my withdrawal by officially notifying Imran Kader Chowdhury or his supervisor Dr. Faiyaz Doctor in writing at chowdh62@uni.coventry.ac.uk or aa9536@coventry.ac.uk

Name : _____
Contact No. : _____
Email : _____
Signature : _____
Date : _____

This app will collect personal data from the user. This consent form is required to be filled by every user to acknowledge that the user is authenticating the app to record these data

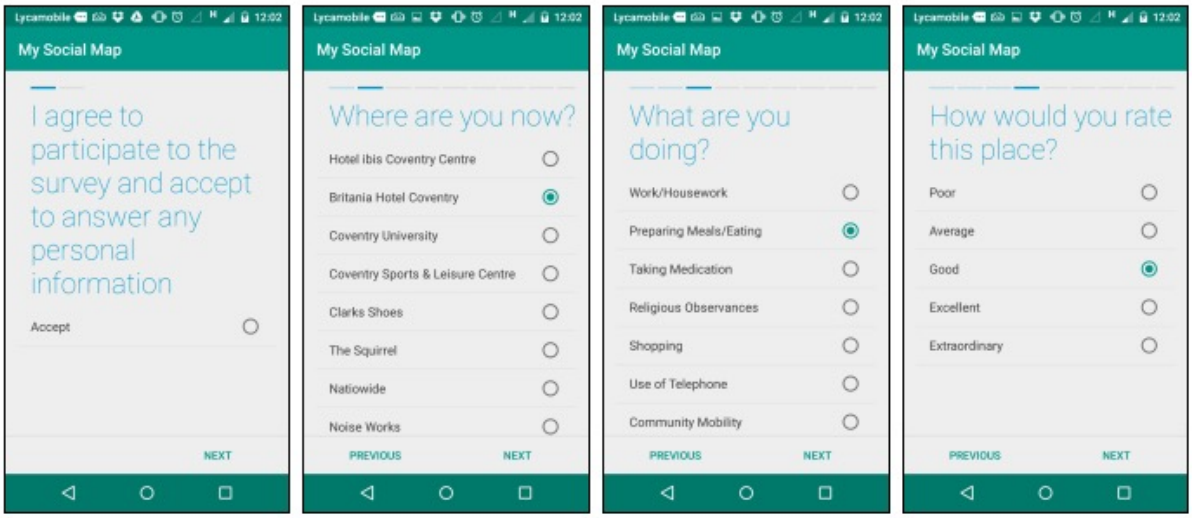
# App Icon




Masters Project 2015/16 Imran Kader Chowdhury 4

# App Launch Questions

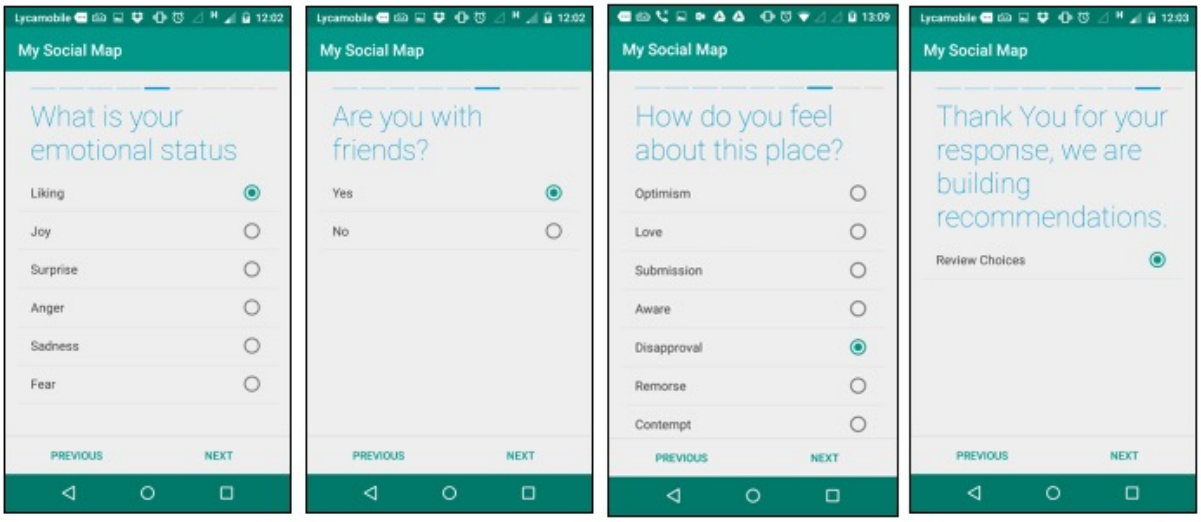
| Step 1: | Step 2: | Step 3: | Step 4: |
|---|---|---|---|
| The user will be asked a series of questions when the app is launched. Here the user agrees to answer all questions to proceed | The user needs to select the place they are currently in | The user needs to select the type of activity he is doing at that place | The user needs to rate the place |




Masters Project 2015/16 Imran Kader Chowdhury 5

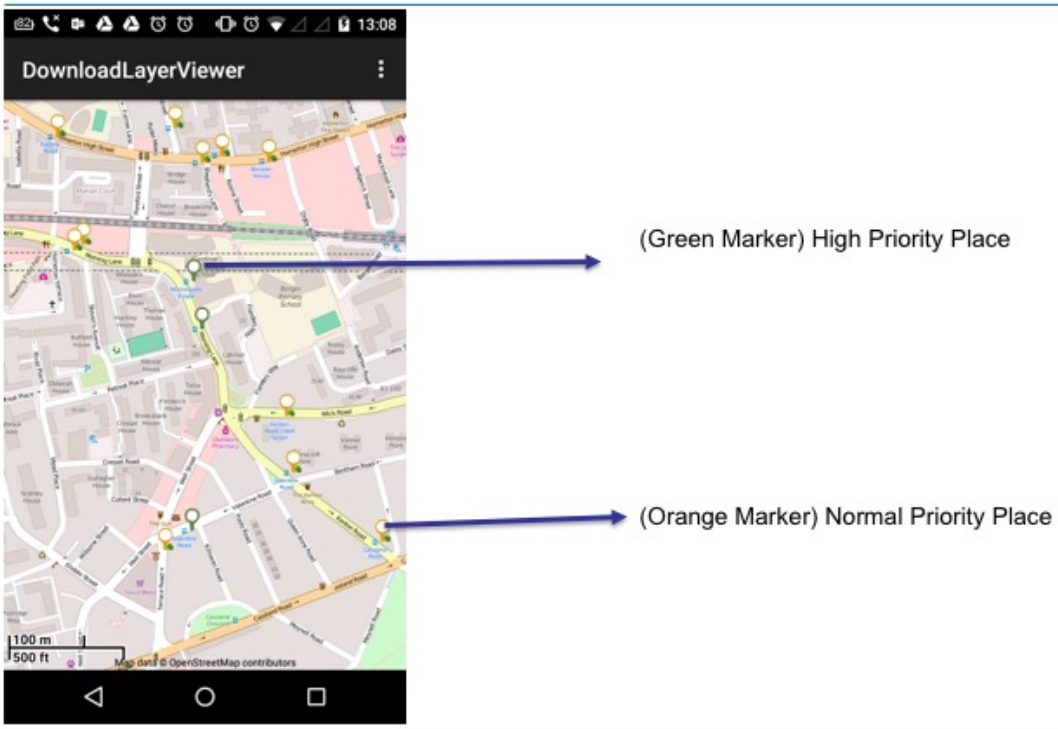
# App Launch Questions

| Step 5: | Step 6: | Step 7: | Step 8: |
|---|---|---|---|
| The user needs to select his current emotional state | The user needs to select whether he/she is with friends | The user needs to select how does he feel about the place | The user has reached at the end of the process and can review the answer and proceed to the map |




Masters Project 2015/16 Imran Kader Chowdhury 6

# View Recommendations




Masters Project 2015/16 Imran Kader Chowdhury 7

## Appendix D – Project Presentation

# Problem Statement

User wants to see places in the map that are important to him




# Opportunity


A map application that can predict and recommend locations

## Learning

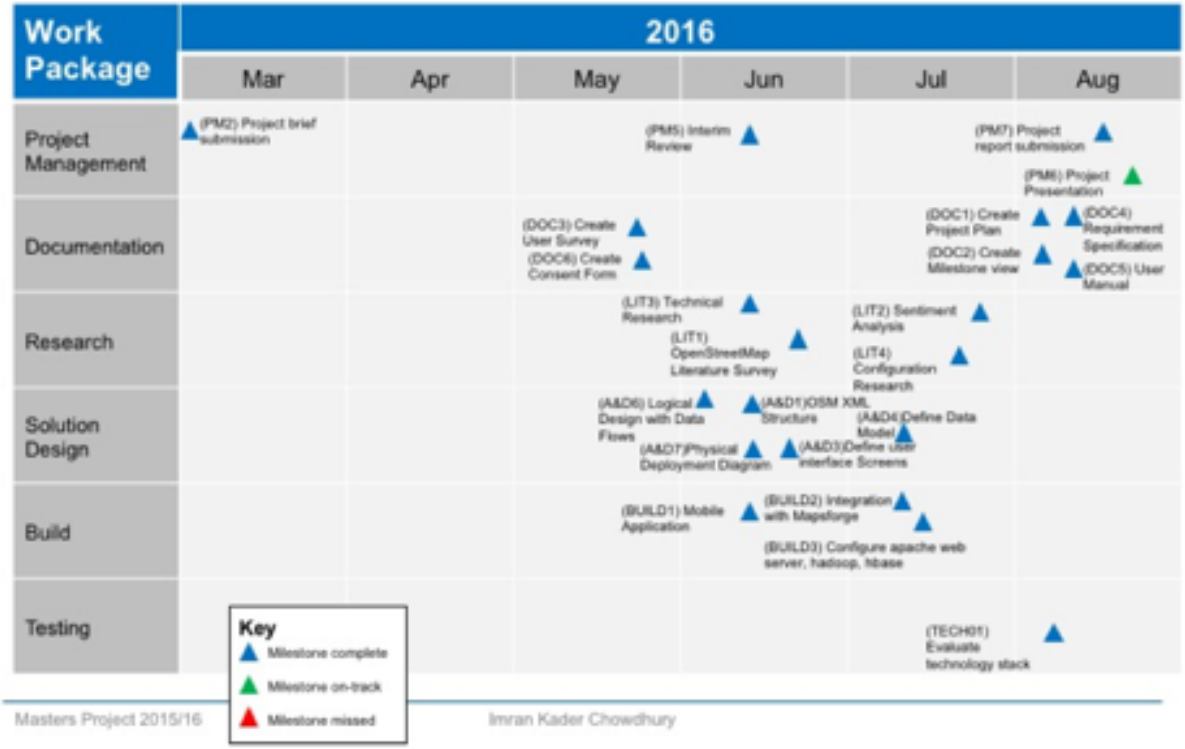
- Decision tree
- Random forest
- Collaborative filtering



## Ethics

- Enquire sensitive personal information
- User giving consent
- User identity hidden

# Intelligent Maps Milestone Plan




Masters Project 2015/16 Imran Kader Chowdhury

| Task ID | Task | Description | Effort | Start Date | Planned Date | Confirmed Date | Dependency | Status | Responsible |
|---|---|---|---|---|---|---|---|---|---|
| **Project Oversight & Control** | | | | | | | | | |
| **Key Milestones** | | | | | | | | | |
| PM1 | Allocation of Supervisor | Identification of projects of interest and associated supervisory staff | 1.5d | 04-Jan | 01-Feb | 01-Feb | N/A | Complete | IKC |
| PM2 | Project Brief submission | Creation and submission of project brief template | 2.0d | 01-Feb | 29-Feb | 29-Feb | PM1 | Complete | IKC |
| PM3 | Ongoing project tracking | | 4.0d | 10-Jun | 28-Jul | 28-Jul | DOC1, DOC2 | Ongoing | IKC |
| PM4 | Fortnightly review with Supervisor | Ongoing review sessions with Supervisor | 2.0d | 10-Jun | 28-Jul | 28-Jul | PM1 | Complete | IKC, FD |
| PM5 | Interim Review | | 1.0d | 13-Jun | 13-Jun | 13-Jun | PM1 | Complete | IKC, FD |
| PM6 | Project Presentation | | 1.0d | 15-Aug | 15-Aug | 15-Aug | N/A | On Track | IKC |
| PM7 | Project Report Submission | | | | | | | On Track | IKC |
| **Key Milestones** | | | | | | | | | |
| DOC1 | Create Project Plan | Creation of this project plan template | 1.0d | 07-Aug | 09-Aug | 09-Aug | N/A | Complete | IKC |
| DOC2 | Create Gantt Chart/Milestone view | Creation of a plan on a page showing major milestones | 1.0d | 08-Aug | 10-Aug | 10-Aug | N/A | Complete | IKC |
| DOC3 | Create user survey | Creation of survey using google form | 1.0d | 26-May | 26-May | 26-May | N/A | Complete | |
| DOC4 | Requirements Specification Document final version | This is a sub-section of the main report | 1.0d | 10-Aug | 10-Aug | 10-Aug | N/A | | |
| DOC5 | User Manual final version | This is a sub-section of the main report | 1.0d | 09-Aug | 09-Aug | 09-Aug | N/A | | |
| DOC6 | Create applicatin use consent form | | 1.0d | 28-May | 28-May | 28-May | N/A | Complete | IKC |
| DOC8 | Project Report - publication version | | | | | | | | |
| **Project Implementation** | | | | | | | | | |
| **Implementation Methodology Selection** | | | | | | | | | |
| IM1 | Identify research topics | Identify those items that will require literature review | 2.0d | 10-Jun | 14-Jun | 14-Jun | N/A | Complete | IKC |
| IM2 | Identify pilot user base | | 1.0d | 12-Jun | 12-Jun | 12-Jun | N/A | Complete | IKC |
| IM3 | Evaluate and Select method for gathering user requirements | | 1.0d | 15-Jun | 15-Jun | 15-Jun | N/A | Complete | IKC |
| IM4 | Define approach for technology selection | Development and deployment strategy | 1.0d | 18-Jun | 18-Jun | 18-Jun | N/A | Complete | IKC |
| IM5 | Evaluate and Select build approach | E.g. evaluation of waterfall, agile or prototyping approaches that will be reactive to new learning | 1.0d | 20-Jun | 20-Jun | 20-Jun | N/A | Complete | IKC |
| IM6 | Implementation approach defined | | | | | | | Complete | IKC |
| **Literature Review** | | | | | | | | | |
| LIT1 | Open Street Map feature analysis | Data Structure and validity | 10.0d | 20-Jun | 10-Jul | 10-Jul | N/A | Complete | IKC |
| LIT2 | Sentiment analysis | Predict sentiment from social media and sensor data | 10.0d | 20-Jun | 10-Jul | 10-Jul | N/A | Complete | IKC |
| LIT3 | Technical research | Machine learning prediction and classification modeling, location aware recommendation system, collaborative filtering | 10.0d | 20-Jun | 10-Jul | 10-Jul | N/A | Complete | IKC |
| LIT4 | Configuration research | Research into installation and configuration of Hadoop, Hbase, HDFS, Thrift, Python, Map Forge | 10.0d | 20-Jun | 10-Jul | 10-Jul | N/A | Complete | IKC |
| LIT5 | Core research completed | | | | | | | Complete | IKC |
| **Requirements Analysis** | | | | | | | | | |
| REQ1 | Conduct end user survey | | 1.0d | 01-Aug | 01-Aug | 01-Aug | DOC3 | | |
| REQ2 | Use case analysis | | 1.0d | 04-Aug | 05-Aug | 05-Aug | N/A | | |
| REQ3 | Incorporate end user survey into own requirements caoture | | 1.0d | 06-Aug | 06-Aug | 06-Aug | N/A | | |
| REQ4 | Requirements analysis finalised | | | | | | | | |
| **Technology Selection** | | | | | | | | | |
| TECH01 | Evaluate appropriate technology stack and | | 1.0d | 07-Aug | 07-Aug | 07-Aug | N/A | Complete | IKC |
| TECH02 | Finalise implementation technologies | | | | | | | Complete | IKC |
| **Analysis** | | | | | | | | | |
| A&D1 | Analayse Open Street Map XML data structures to identify salient attributes | Data analysis and Data pre processing | 4.0d | 01-Jun | 10-Jun | 10-Jun | N/A | Complete | IKC |

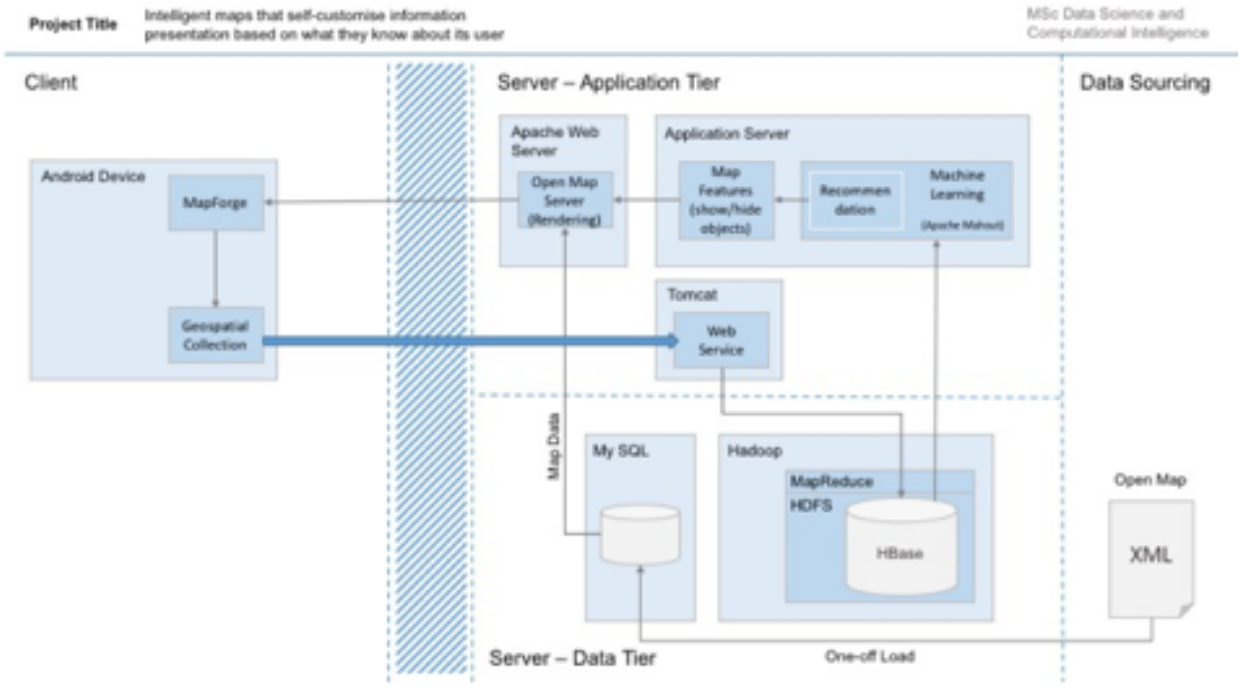
Masters

# Literature Research

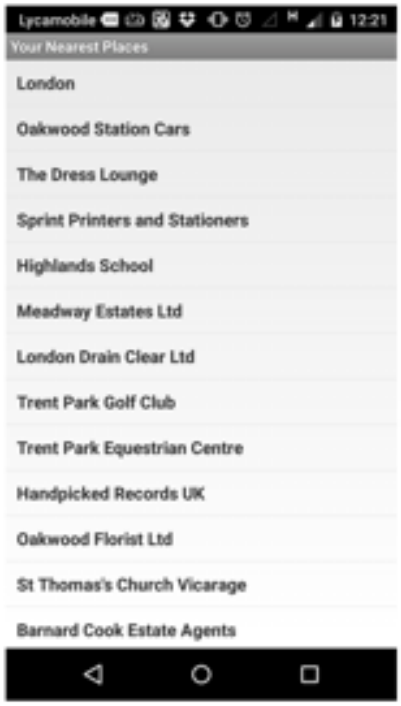
- OpenStreetMap
- Sentiment Analysis
- Machine Learning Models
- Recommendation Systems
- Geospatial Analysis

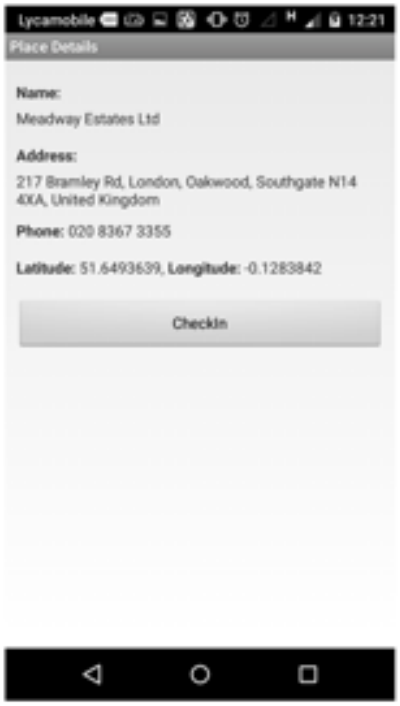


# Architecture Design



# Prototype Version 1

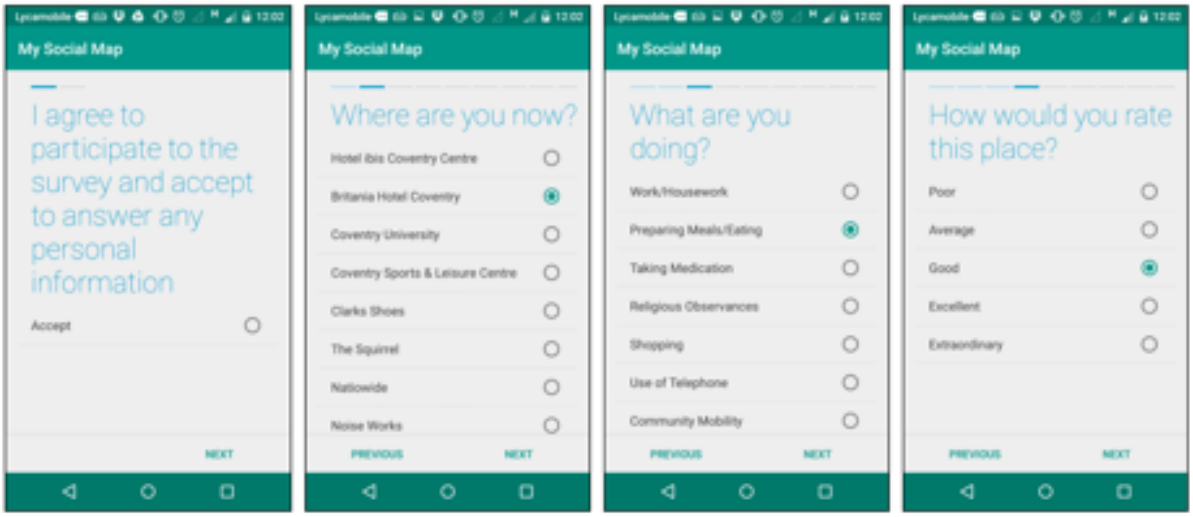



5

Masters Project 2015/16 Imran Kader Chowdhury

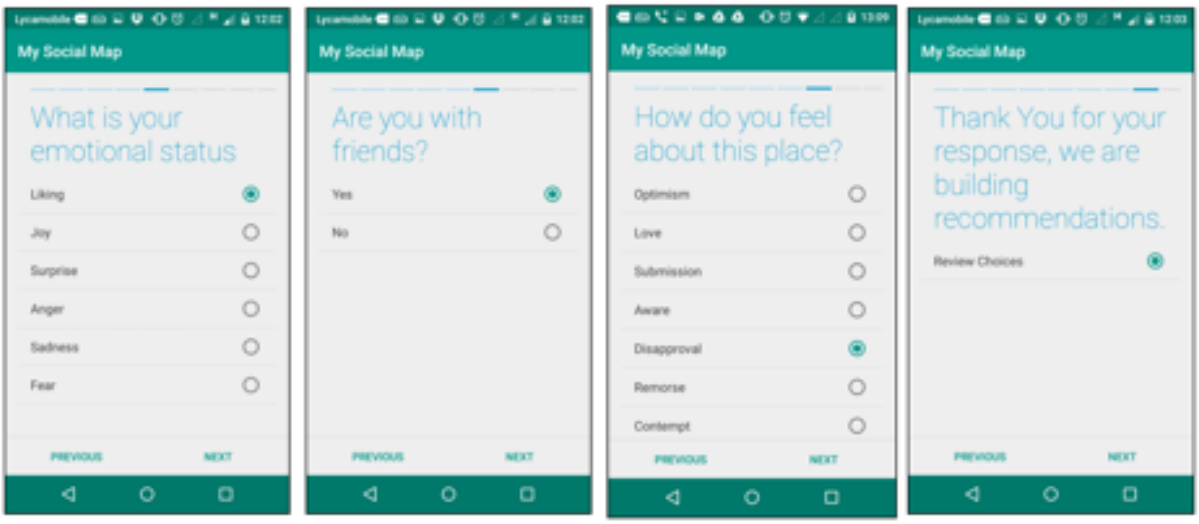
# Prototype Version 2




5

Masters Project 2015/16 Imran Kader Chowdhury

# Data Collection






# Data Model

| | Feature | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|
| Label | userid | day | time | friends | activity | feeling | emotion | class |
| Value | 1 | MON | MORNING | 0 | EATING | LOVE | SURPRISE | 813 |
| | 2 | TUE | NOON | 1 | WORK | AWARE | LIKING | 685 |
| | 3 | WED | MORNING | 2 | SHOPPING | REMORSE | JOY | 557 |
| | 4 | THU | EVENING | 3 | EATING | AWARE | SURPRISE | 429 |
| | 5 | FRI | MORNING | 4 | WORK | LOVE | SURPRISE | 301 |
| | 6 | SAT | NOON | 5 | SHOPPING | AWARE | LIKING | 173 |

# Test Data

## Input Data:

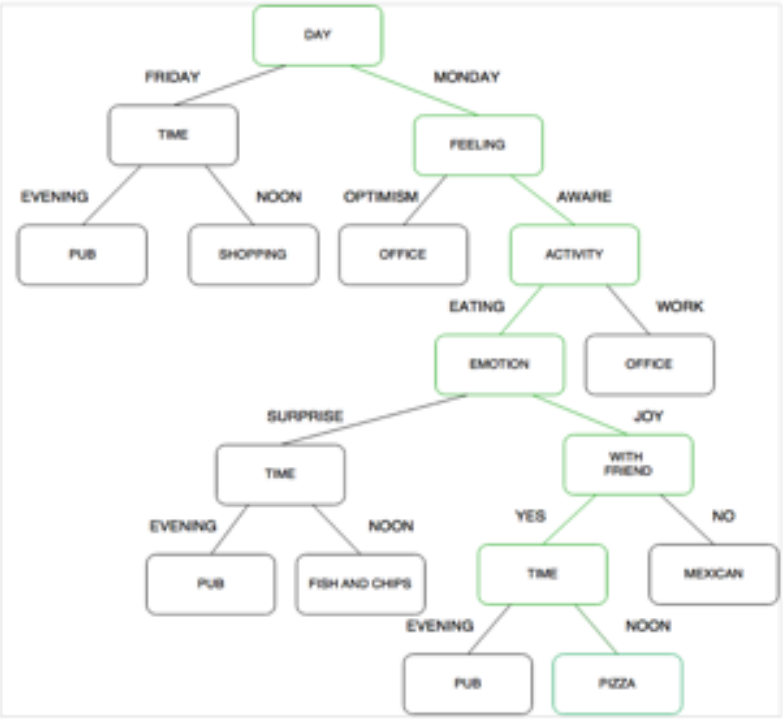
| | Feature | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|
| Label | userid | day | time | friends | activity | feeling | emotion | class |
| Value | 1 | MON | MORNING | 0 | EATING | LOVE | SURPRISE | 813 |

## Test Data:

| | Feature | | | | | | | Target |
|---|---|---|---|---|---|---|---|---|
| Label | userid (X1) | day (X2) | time (X3) | friends (X4) | activity (X5) | feeling (X6) | emotion (X7) | class (Y) |
| Value | 1 | MON | MORNING | 0 | EATING | LOVE | SURPRISE | ? |
| | 1 | MON | MORNING | 0 | HOUSEWORK | AWARE | LIKING | ? |
| | 1 | MON | MORNING | 0 | SHOPPING | AWARE | JOY | ? |
| | 1 | MON | NOON | 0 | WORK | OPTIMISM | SADNESS | ? |
| | 1 | MON | NOON | 0 | EATING | LOVE | SURPRISE | ? |

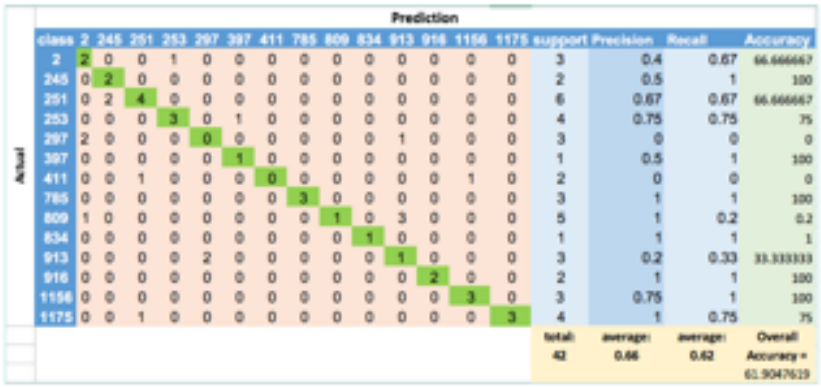
Masters Project 2015/16 Imran Kader Chowdhury

# Decision tree




Masters Project 2015/16 Imran Kader Chowdhury

# Confusion Matrix



| | | Prediction | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | class | 2 | 245 | 251 | 253 | 297 | 397 | 411 | 785 | 809 | 834 | 913 | 916 | 1156 | 1175 | support | Precision | Recall | Accuracy |
| Actual | 2 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0.4 | 0.67 | 66.666667 |
| | 245 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0.5 | 1 | 100 |
| | 251 | 0 | 2 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0.67 | 0.67 | 66.666667 |
| | 253 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0.75 | 0.75 | 75 |
| | 297 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| | 397 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.5 | 1 | 100 |
| | 411 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 |
| | 785 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 1 | 100 |
| | 809 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 5 | 1 | 0.2 | 0.2 |
| | 834 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| | 913 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0.2 | 0.33 | 33.333333 |
| | 916 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 1 | 100 |
| | 1156 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 3 | 0.75 | 1 | 100 |
| | 1175 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | 1 | 0.75 | 75 |
| | | | | | | | | | | | | | | | | total: 42 | average: 0.66 | average: 0.62 | Overall Accuracy = 61.9047619 |

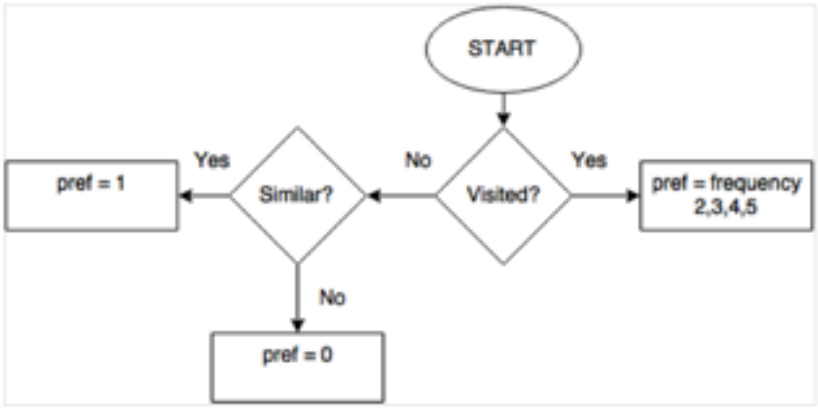


# Collaborative Filtering

| User | Location 1 | Location 2 | Location 3 | Location 4 |
|---|---|---|---|---|
| 1 | Yes | No | Yes | Yes |
| 2 | - | Yes | No | No |
| 3 | Yes | Yes | No | - |
| 4 | No | - | Yes | - |
| 5 | Yes | Yes | ? | No |

# Data Model

| User ID | Item ID | Preference |
|---|---|---|
| 1 | 234 | 1 |
| 1 | 756 | 3 |
| 1 | 8788 | 2 |
| 1 | 1150 | 0 |
| 1 | 35 | 0 |




Masters Project 2015/16 Imran Kader Chowdhury

# Evaluation

| User ID | Place ID | Actual Rating | Predicted Rating | Absolute Error |
|---|---|---|---|---|
| 1 | 234 | 1 | 1.2 | 0.2 |
| 1 | 756 | 3 | 2.9 | 0.1 |
| 1 | 8788 | 2 | 2.2 | 0.2 |
| | | | average absolute error | 0.167 |

For the actual Dataset, the error calculated was: 0.1797573220296521

Masters Project 2015/16 Imran Kader Chowdhury

# View Recommendations


(Green Marker) High Priority P

(Orange Marker) Normal Priori



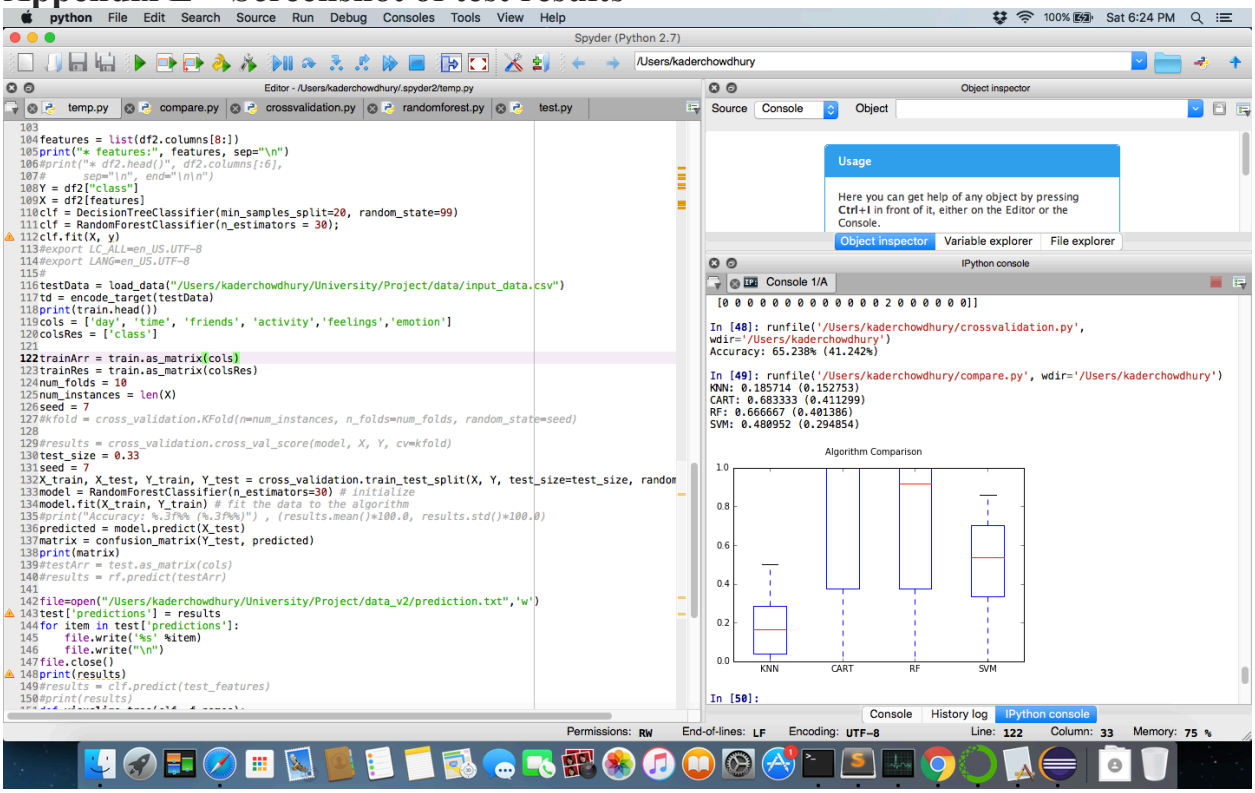
# Conclusion

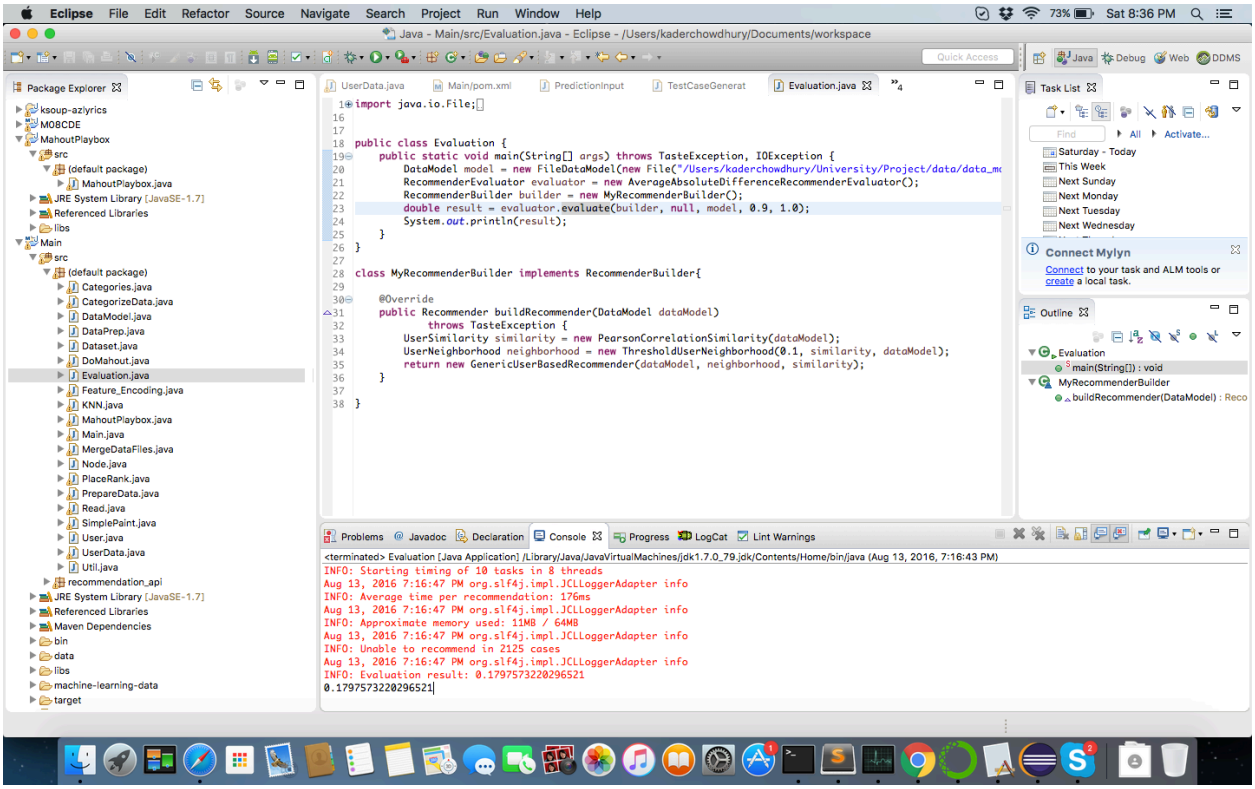
Potential Usage Scenario:

1. Cyclist map
2. Fresher's map
3. Tourists map

# Questions?

# Appendix E – Screenshot of test results



## Appendix F – Project Ethics

Following documents list has been attached to the report:
List of attached Documents:

1. 04 M08cde Survey - Google Forms
2. Certificate of Ethical Approval
3. Informed Consent Form
4. Participants Information Letter
5. Project Review

# 01. M08cde Survey - Google Forms



## Evaluating the usability of a contextualized map application

Participants Information

* Required

1. **Name** *

2. **Email** *

3. **Address** *

4. **Phone number**

5. **Gender** *
   *Mark only one oval.*
   - Male
   - Female

6. **Age Group** *
   *Mark only one oval.*
   - 18 and under
   - 19-24
   - 25-34
   - 35-44
   - 45 and over

## Map Applications Used

7. **Which map application do you use most** *
*Check all that apply.*

- [ ] Google Maps
- [ ] Waze
- [ ] Here Maps
- [ ] NavFree
- [ ] MapFactor
- [ ] Navigator
- [ ] Other

8. **Why do you use map applications for**
*Mark only one oval.*

- ◯ Search Places
- ◯ Navigation
- ◯ Both
- ◯ None

9. **What mode of transport do you use mostly to commute the city**
*Mark only one oval.*

- ◯ Walking
- ◯ Car
- ◯ Public Transport

10. **How often do you use maps while commuting the city**
*Mark only one oval.*

- ◯ Never
- ◯ Always
- ◯ Sometime

## Software Evaluation

11. **How likely is it that you would recommend this software to a friend or family member?** *
*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Not at all Likely | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Extremely Likely |

12. **How satisfied are you with the reliability of this software? ***
*Mark only one oval.*

- ◯ Extremely satisfied
- ◯ Very satisfied
- ◯ Somewhat satisfied
- ◯ Not so satisfied
- ◯ Not at all Satisfied

13. **How satisfied are you with the security of this software? ***
*Mark only one oval.*

- ◯ Extremely satisfied
- ◯ Very satisfied
- ◯ Somewhat satisfied
- ◯ Not so satisfied
- ◯ Not at all Satisfied

14. **How satisfied are you with this software's ease of use? ***
*Mark only one oval.*

- ◯ Extremely satisfied
- ◯ Very satisfied
- ◯ Somewhat satisfied
- ◯ Not so satisfied
- ◯ Not at all Satisfied

15. **How satisfied are you with the look and feel of this software? ***
*Mark only one oval.*

- ◯ Extremely satisfied
- ◯ Very satisfied
- ◯ Somewhat satisfied
- ◯ Not so satisfied
- ◯ Not at all Satisfied

16. **How would you rate this software in a scale of 5 ***
*Mark only one oval.*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ |

17. **Do you have any thoughts on how to improve this software?**

Powered by
Google Forms

## 02. Certificate of Ethical Approval


# Certificate of Ethical Approval

Applicant:

Imran Chowdhury

Project Title:

Intelligent maps that self-customise information presentation based on what they know about its user

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Medium Risk

Date of approval:

02 June 2016

Project Reference Number:

P42094

# 03. Informed Consent Form

**PERSONAL DATA PROTECTION ACT**
**CONSENT FORM**

1. In compliance with the Personal Data Protection Act ("PDPA"), I give my consent to collect and use my personal data (i.e. Name, contact numbers, mailing and email addresses) in order to perform research.

2. I also give consent to collect data such as geospatial data( latitude, longitude), sensor data (accelerometer, gyroscope) through my mobile phone for the research.

3. By installing the application on my mobile phone I also agree to give access to my social networks (if requested) as long as I am assured that my personal data will be kept secured according to PDPA. Moreover I should be able to view my data upon request. I also give consent to make the data available online for further studies for others as well.

4. I hereby give my acknowledgement and consent to Imran Kader Chowdhury to use my personal data for his research.

5. I agree that my consent will remain in place until my withdrawal by officially notifying Imran Kader Chowdhury or his supervisor Dr. Faiyaz Doctor in writing at chowdh62@uni.coventry.ac.uk or aa9536@coventry.ac.uk

Name : _____
Contact No. : _____
Email : _____
Signature : _____
Date : _____

**PERSONAL DATA PROTECTION ACT**
**CONSENT FORM**

1. In compliance with the Personal Data Protection Act ("PDPA"), I give my consent to collect and use my personal data (i.e. Name, contact numbers, mailing and email addresses) in order to perform research.

2. I also give consent to collect data such as geospatial data( latitude, longitude) through my mobile phone for the research.

3. I agree to take part in various surveys required for the research.

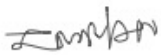
4. I agree for this application to collect my activities, feelings, emotion, rating associated to my geo-location.

5. By installing the application on my mobile phone I also agree to give access to my social networks (if requested) as long as I am assured that my personal data will be kept secured according to PDPA. Moreover I should be able to view my data upon request. I also give consent to make the data available online for further studies for others as well.

6. I hereby give my acknowledgement and consent to Imran Kader Chowdhury to use my personal data for his research.

7. I agree that my consent will remain in place until my withdrawal by officially notifying Imran Kader Chowdhury or his supervisor Dr. Faiyaz Doctor in writing at chowdh62@uni.coventry.ac.uk or aa9536@coventry.ac.uk

Name : Chowdhury Imran Kader
Contact No. :
Email : chowdh62@uni.coventry.ac.uk
Signature :
Date : 01/07/2016



**PERSONAL DATA PROTECTION ACT**
**CONSENT FORM**

1. In compliance with the Personal Data Protection Act ("PDPA"), I give my consent to collect and use my personal data (i.e. Name, contact numbers, mailing and email addresses) in order to perform research.

2. I also give consent to collect data such as geospatial data( latitude, longitude) through my mobile phone for the research.

3. I agree to take part in various surveys required for the research.

4. I agree for this application to collect my activities, feelings, emotion, rating associated to my geo-location.

5. By installing the application on my mobile phone I also agree to give access to my social networks (if requested) as long as I am assured that my personal data will be kept secured according to PDPA. Moreover I should be able to view my data upon request. I also give consent to make the data available online for further studies for others as well.

6. I hereby give my acknowledgement and consent to Imran Kader Chowdhury to use my personal data for his research.

7. I agree that my consent will remain in place until my withdrawal by officially notifying Imran Kader Chowdhury or his supervisor Dr. Faiyaz Doctor in writing at chowdh62@uni.coventry.ac.uk or aa9536@coventry.ac.uk

Name : Fleming Liam
Contact No. :
Email : flemin14@uni.coventry.ac.uk
Signature :
Date : 01/07/2016

**PERSONAL DATA PROTECTION ACT**
**CONSENT FORM**

1. In compliance with the Personal Data Protection Act ("PDPA"), I give my consent to collect and use my personal data (i.e. Name, contact numbers, mailing and email addresses) in order to perform research.

2. I also give consent to collect data such as geospatial data( latitude, longitude) through my mobile phone for the research.

3. I agree to take part in various surveys required for the research.

4. I agree for this application to collect my activities, feelings, emotion, rating associated to my geo-location.

5. By installing the application on my mobile phone I also agree to give access to my social networks (if requested) as long as I am assured that my personal data will be kept secured according to PDPA. Moreover I should be able to view my data upon request. I also give consent to make the data available online for further studies for others as well.

6. I hereby give my acknowledgement and consent to Imran Kader Chowdhury to use my personal data for his research.

7. I agree that my consent will remain in place until my withdrawal by officially notifying Imran Kader Chowdhury or his supervisor Dr. Faiyaz Doctor in writing at chowdh62@uni.coventry.ac.uk or aa9536@coventry.ac.uk

Name : Karydis Dimitrios
Contact No. :
Email : karydisd@uni.coventry.ac.uk
Signature :
Date : 01/07/2016

---

**PERSONAL DATA PROTECTION ACT**
**CONSENT FORM**

1. In compliance with the Personal Data Protection Act ("PDPA"), I give my consent to collect and use my personal data (i.e. Name, contact numbers, mailing and email addresses) in order to perform research.

2. I also give consent to collect data such as geospatial data( latitude, longitude) through my mobile phone for the research.

3. I agree to take part in various surveys required for the research.

4. I agree for this application to collect my activities, feelings, emotion, rating associated to my geo-location.

5. By installing the application on my mobile phone I also agree to give access to my social networks (if requested) as long as I am assured that my personal data will be kept secured according to PDPA. Moreover I should be able to view my data upon request. I also give consent to make the data available online for further studies for others as well.

6. I hereby give my acknowledgement and consent to Imran Kader Chowdhury to use my personal data for his research.

7. I agree that my consent will remain in place until my withdrawal by officially notifying Imran Kader Chowdhury or his supervisor Dr. Faiyaz Doctor in writing at chowdh62@uni.coventry.ac.uk or aa9536@coventry.ac.uk

Name : Sweiss Ramez Zaher F.
Contact No. :
Email : sweissr@uni.coventry.ac.uk
Signature :
Date : 01/07/2016

---

**PERSONAL DATA PROTECTION ACT**
**CONSENT FORM**

1. In compliance with the Personal Data Protection Act ("PDPA"), I give my consent to collect and use my personal data (i.e. Name, contact numbers, mailing and email addresses) in order to perform research.

2. I also give consent to collect data such as geospatial data( latitude, longitude) through my mobile phone for the research.

3. I agree to take part in various surveys required for the research.

4. I agree for this application to collect my activities, feelings, emotion, rating associated to my geo-location.

5. By installing the application on my mobile phone I also agree to give access to my social networks (if requested) as long as I am assured that my personal data will be kept secured according to PDPA. Moreover I should be able to view my data upon request. I also give consent to make the data available online for further studies for others as well.

6. I hereby give my acknowledgement and consent to Imran Kader Chowdhury to use my personal data for his research.

7. I agree that my consent will remain in place until my withdrawal by officially notifying Imran Kader Chowdhury or his supervisor Dr. Faiyaz Doctor in writing at chowdh62@uni.coventry.ac.uk or aa9536@coventry.ac.uk

Name : Syed Razack Syed Rafsan
Contact No. : _____
Email : syedrazs@uni.coventry.ac.uk
Signature : _____
Date : 01/07/2016

# 04. Participants Information Letter

| Participants Information Sheet | | | |
|---|---|---|---|
| **Last Name** | **First Names** | **Email** | **Phone** |
| Chowdhury | Imran Kader | chowdh62@uni.coventry.ac.uk | +447438976729 |
| Fleming | Liam | flemin14@uni.coventry.ac.uk | +447527323251 |
| Karydis | Dimitrios | karydisd@uni.coventry.ac.uk | +447464789651 |
| Syed Razack | Syed Rafsan | syedrazs@uni.coventry.ac.uk | +447341018377 |
| Thomas | Rhys | thoma186@uni.coventry.ac.uk | +447527223917 |
| Sweiss | Ramez Zaher F. | sweissr@uni.coventry.ac.uk | +447490046631 |

# 05. Ethics Review Feedback Form

Intelligent maps that self-customise information presentation based on what they know about its user P42094

**REGISTRY RESEARCH UNIT**

**ETHICS REVIEW FEEDBACK FORM**

(Review feedback should be completed within 10 working days)

**Name of applicant:** Imran Chowdhury .....

**Faculty/School/Department:** [Faculty of Engineering, Environment and Computing] CSE

**Research project title:** Intelligent maps that self-customise information presentation based on what they know about its user

Comments by the reviewer

| |
|---|
| **1. Evaluation of the ethics of the proposal:** |
| **2. Evaluation of the participant information sheet and consent form:** |
| **3. Recommendation:**<br>(Please indicate as appropriate and advise on any conditions. If there any conditions, the applicant will be required to resubmit his/her application and this will be sent to the same reviewer). |

| | |
|---|---|
| X | Approved - no conditions attached |
| | Approved with minor conditions (no need to re-submit) |
| | Conditional upon the following – please use additional sheets if necessary (please re-submit application) |
| | Rejected for the following reason(s) – please use other side if necessary |
| | Not required |

**Name of reviewer:** Anonymous .....

**Date:** 02/06/2016.....

Imran Chowdhury Page 1 of 1 02 June 2016