# 1st Progress Report

Final Year Project (2017/2018)

## Sentiment-Aware POI search

**Bachelor of Science in Computer Science**

**Hong Kong Baptist University**

Cheung Siu Sin 16214684

Ma Yu Kan 16228073

# Table of Contents

# 1 Abstract

This proposal will discuss the problem, proposed solution, the role of the user, language used, hardware and software requirement and finally the distribution of the workload.

This project, Sentiment-Aware POI search, will separate into two parts, first of all, the part of searching the location required by the user based on the category of the buildings. This part is similar to the Google Map's searching part which also providing results based on user's location and the keywords given by the user.

So, what makes this project special? Besides of just searching the location just like the Google Map, we tend to offer a result which gives a more accurate and optimized result by making the searching become sentiment awareness and we will use Twitter as the source of the comments, which we will cover it up in the later passage.

# 2 Problem

## How does the algorithm work?

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect the analysis, emotion analysis, review mining, etc. However, they are now all under the umbrella of sentiment analysis or opinion mining. While in industry, the term sentiment analysis is more commonly used, but in academia, both sentiment analysis and opinion mining are frequently employed.

Our project will adopt the technique of Sentiment Analysis to analyze, which is the hardest part of this project that what is the way of presenting the words or sentences from user input. In general, people are easy to understand the mean of the opinion or sentence of others, but it is difficult for the computer to know about the sentiment of users' opinion or neutral language. We cannot teach the computer to understand the neutral language, but we can 'teach' the computer that how to process the neutral language of a textual item. This is a big challenge for us because there does not have an open source for our reference, we need to study more about Sentiment Analysis from others' articles beyond the classroom, and we need to study that what are the evaluations we can adopt, and how to make these evaluations as a program and it works.

# Categorization of Disorganized Data

And another challenge, it is about the categorization of a disorganized dataset of Twitter. As all know, the post does not have a specified subject in Twitter, which is about the user's feeling or opinion. Thus, we need to extract the entities from the dataset (Entity, opinion holder and time extraction is the classic problem of named entity recognition by Bing Liu), such as political entities, sports entities, etc. and to clean the data from the irrelevant news as well as advertisements. After extraction, we also need to categorize the extracted entities. In natural language text, people often write the same entity in different ways. For example, Motorola may be written as Mot, Moto, and Motorola. We need to recognize that they all refer to the same entity.

This challenge is so interesting for us because it is a mathematical question about cluster analysis. We have learnt the knowledge about K-means, which aims to partition **N** observations into **K** clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. But we need to apply the theory that we have learnt in school into our system.

# 3 Proposed Solution

## Scope of the solution

At this part, we will talk about how the system supposes to be used by the user after talking about the algorithm and the technical stuff.

First of all, the user will type in the keywords or select the category of buildings they want, the location will acquire automatically after granting the permission from the user and will store in the session for future use.

After receiving the location, the keywords or category from the user, the server will receive the data as a JSON strings and the data will pass through to the algorithm, as the algorithm will calculate based on the data it has and will send back the result to the user as the Top-10 result based on the user requested area.

Then, the calculated data will send back to the web application and the application will show it out as a list and will display on the map when the user selects the places.

After introducing how the system working abstractly, we will dig deeper at what exactly the server working inside.

# Step of sentiment analysis

After the study about sentiment analysis, we recognized there are 5 steps to achieve the sentiment analysis:

1. To predict a tweet whether it is positive, neutral or negative.
2. To predict a tweet of known to be about a given topic whether it can be expressed a positive, neutral or negative sentiment towards of the topic.
3. To evaluate a tweet of known to be about a given topic that the sentiment conveys towards to the topic on a 5-point scale ranging from Highly Positive (2.0 points) to Highly Negative (-2.0 points).
4. To classify a set of a tweet which is in Positive and Negative classes.
5. To estimate the scale of a set of a tweet of known to be about a given topic which across 5 classes, ranging from to Highly Negative.

# 4 Role of users

The system can be used by the travellers which want to get local comments on the attractions. For example, they may want to want how the restaurant is before booking the seat. So by using this application, the user will able to check out the best rating attractions on their planning of the trip.

Besides, it can be used by the travellers, it also can be used by the locals, even the may not familiar to all the places where they stay in, so by using this application, they can search for the comments of these places.

However, as the tweets database also based in the United States, so this application only can be used in the United States. But we assume that this application can be used in the country which speaks English and if the database of that country is added.

# 5 Implementation requirements

This part will talk about the requirement of the implementation of our project, it will separate into two parts, the first part will talk about the choices on the programming language and the second part will talk about what is the hardware and software requirement on our server and the client.

## Choices on the programming languages

There are many types of languages can be chosen on our project, such as C/C++, Python, JavaScript, Java etc. But we prefer Java language to develop the server side. It is because we have a good experience in development of Java language and its architecture.

For the client side, JavaScript will be used, besides, difference framework, for example, express.js, AngularJS and Knockout, will also be used as the construction of the client. In order to create a mobile-friendly application, bootstrap will also be used to build a responsive, mobile-friendly application

Apart from the programming language, the choices of the hardware and software are also important on the reliability of the project and there is the requirement of our project.

# The requirement of the server

CPU: 4 Core CPU
RAM: 8 GB RAM
Hard Disk: 500G Disk Space
Operation System: Ubuntu Server 16.04.3 LTS
With Network Connection

# The requirement of the client

## Supporting Web Browser

Apple Safari 10 or above
Google Chrome 56 or above
Mozilla Firefox 55 or above

## Support Smart Device

Apple iPhone with iOS 9 or above
Google Android 6 or above

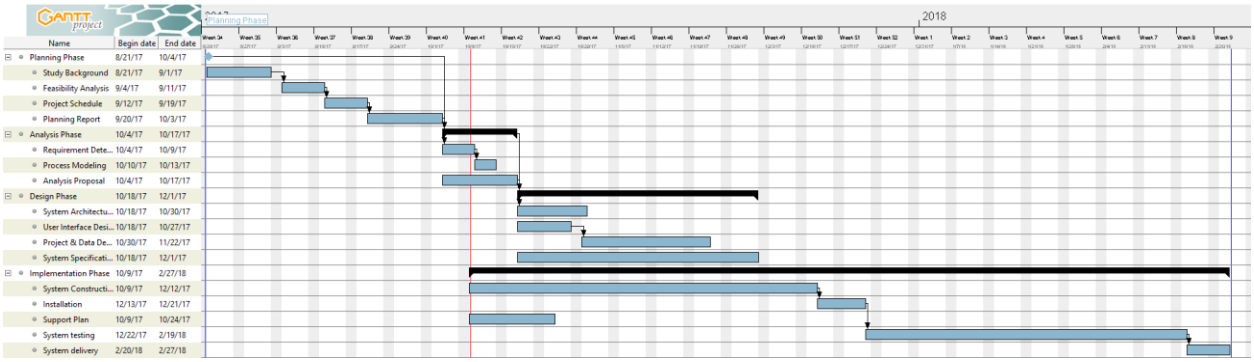
# 6 Detailed Project Plan

## Task List

| **Task ID** | **Planning** | **Duration (days)** | **Dependency** |
|---|---|---|---|
| A1 | Study Background | 10 | |
| A2 | Feasibility Analysis | 6 | (A1) Study Background |
| A3 | Project Schedule | 6 | (A2) Feasibility Analysis |
| A4 | Planning Report | 10 | |
| **Task ID** | **Analysis** | **Duration (days)** | **Dependency** |
| B1 | Requirement Determination | 4 | |
| B2 | Process Modeling | 4 | (B1)Requirement Determination |
| B3 | Analysis Proposal | 10 | |
| **Task ID** | **Design** | **Duration (days)** | **Dependency** |
| C1 | System Architecture Design | 9 | |
| C2 | User Interface Design | 8 | (C1)System Architecture Design |
| C3 | Project & Data Design | 18 | (C2)UI Design |
| C4 | System Specification | 33 | |

| **Task ID** | **Implementation** | **Duration (days)** | **Dependency** |
|---|---|---|---|
| D1 | System Construction | 47 | |
| D2 | Installation | 7 | (D1)System Construction |
| D3 | Support Plan | 12 | |
| D4 | System testing | 42 | (D2) Installation |
| D5 | System delivery | 6 | (D4) System testing |

# Gantt chart




Full-Size Photo: https://goo.gl/KQBMZR

# 7 Distribution of the workload

Cheung Siu Sin will responsible for the client side and the back-end side, for example, the user interface of the web application and the server structure.

Ma Yu Kan will responsible for the algorithm side of this application, for example, the design of the algorithm, the implementation of the algorithm and the optimization of the algorithm.