Ronan McDaid – 40042962

Final Year Project – Problem Specification

Title: Web-based visualisation platform for Twitter data

Supervisor: Dr Anna Jurek

# Problem Description

Below is the project description from the “Project List 2016-17” document on QOL.

“Social media is one of the most significant information exchange technologies of the 21st century. Social media channels, such as Twitter and Facebook, provide very convenient and efficient ways of communicating and sharing information publicly. Making sense of the high-volume data generated through social media has become very popular in both industrial and academic environments. The purpose of this project is to develop an interactive platform which allows to stream, analyse and visualise Twitter data in real time.

The functionality of the platform can for example include:

• Topic analysis

• Sentiment analysis

• Conversation analysis”

# Solution

As a solution to the challenges listed above I have decided to create an interactive dashboard which will present the user with various visualisations of data such as graphs, tables and charts. The user will first enter a search term which will trigger the system to stream Tweets containing the term into a database and remove any information provided with each Tweet by the Twitter API that will not be used for analysis. The data obtained by the search will be analysed by various algorithms. The results will feed into visualisations, displaying information to do with popularity of users and their Tweets, location that Tweets originate from and the time and date that the Tweets were sent. I would also like to include some form of sentiment analysis if even at a basic level to allow the user to assess whether the Tweets are positive or negative. I would also like it to be possible for the user to save or print the results of their search for future reference. The final project should be simple for the user to operate and navigate and the dashboard should be adjustable so the user can focus on a specific feature and set of data.

# Overview

Social media platforms such as twitter have become one of the most popular ways of communicating and sharing information across the globe. Twitter has an estimated 310 million monthly active users and 500 million tweets being sent on a daily basis. The huge volume of data being sent presents opportunities to analyse and spot trends for both industrial and academic purposes. I have been tasked to create an interactive application which allows the user to stream, analyse and visualise Twitter data.

The analysis of this data has many real-world applications including in business, politics and entertainment to name just a few. Businesses may look at Twitter data relating to their own products or to assess demand for a service they are considering providing, they could also see how their business is viewed by the public using sentiment analysis. Political parties could also use sentiment analysis to judge their popularity and they could also view which topics are popular in their area in order to shape policy that addresses local concerns and win them voters. For entertainment purposes artists can chose to alter set-lists for concerts depending on which songs have been popular in the area, for example The Rolling Stones often have their followers on Twitter request the songs to be played in the encore of concerts they will be attending.

These are just a few examples of how the vast quantity of data available from Twitter can be filtered and analysed to suit the various needs of groups or individuals on a large or small scale.

I see this project as having four major stages as follows:

## Learning to use Twitter API

Learning Twitter API will be the first task for me in order to complete a successful project, having limited experience using Twitter and no previous experience using Twitter API, I will need to spend time familiarising myself with the features of both. This will be done by spending initial time researching and studying tutorials about the Twitter API and it’s various uses and throughout my project as I come across features and learn how to use them.

As I progress through my project and learn about the application of Twitter API I expect to come across features that will present me with opportunities to perform analysis that I had not previously considered. I am sure this will be challenging but I hope to learn as I work and make best of these opportunities to improve my system and provide more useful functions to the user.

## Obtaining and storing data

Before I perform analysis on the data I will need to capture it from Twitter by either streaming live data or searching Tweets that have been posted in the past or possibly combining the two. This will be something that I will need to examine the benefits and complications of each case. For example, by streaming live data I will be able to provide analysis of the most recent data available but it will also require waiting until there is significant data to perform meaningful analysis. This wait may be relatively short for searches on popular topics with high volumes of tweets but may prove impractical for more niche areas. Using historical tweets would solve this issue but the timeframe of data you want to gather may vary depending on your search. Throughout the course of my project I will attempt to work with both historical and streamed data to see what is most suitable for providing analysis.

Twitter’s API passes data in JSON format with many pieces of information related to each Tweet­ such as username, content, geolocation, time created, retweets or if the Tweet is a reply. These are just a few examples of the large selection of data made available and there are many pieces that I may not wish to use at all so I will need to trim the data down and store only the relevant parameters. Alternatively, I could store all the information provided and trim it down at the point of constructing queries but this would be less efficient than narrowing down the parameters chosen before storage.

My early research suggests that MongoDB is a popular platform for ingesting and storing Twitter data so I will examine it as a possibility for my project. It is however a resource that I have never used before so I will have to assess it’s usability and how easy it will be to implement into my project compared with other SQL based applications.

## Analysing the data

Once I have obtained the data from Twitter I will need to analyse what areas I would like to investigate and the types of analysis I wish to perform. For example, upon searching for a keyword of ‘Belfast’ it may be of interest to see where the Tweets have been sent from and possibly excluding Belfast or Northern Ireland results to see where in the world the topic of Belfast is most popular. Twitter allows you to see how many times a particular Tweet has been retweeted or favourited. It may be of interest to show the most popular tweets in relation to the search area by ranking them by either of these figures. Similar to this it could be possible to view which user has the most followers and therefore reaches most people with their Tweets. It would also be possible to show the popularity of the topic you are investigating over time by comparing timestamps and looking at any peaks in volume of Tweets.

Another possibility for analysing the data from Twitter is by examining the Tweets and using sentiment analysis to see if they are speaking about the topic in a positive or negative manner. This is very suitable for tweets as they are consistent in length (no more than 140 characters) which cannot be said for posts on other platforms such as Facebook. An example where this would be interesting is applying it to the 2016 US Election, by searching the names of either candidate and seeing how many Twitter impressions spoke about each and whether they were complimentary or critical. This may be a more difficult feature to implement but would certainly be of useful and provide interesting information to the user.

## How to visualise the data

After creating algorithms and methods for analysis on the Twitter data I will create visualisations to represent the results in a more readable and user-friendly format than just the raw figures. I aim to create dashboards showing a combination or graphs and tables so that the user can quickly digest the information and look for trends and points of interest in the results.

Some examples I wish to implement as mentioned previously would be tables showing the most popular Tweets or users related to the search, ranked by retweets/favourites or followers. I would also like to create time-series graphs showing the volume of Tweets sent over a period of time which would allow the user to assess the level of interest in their topic in this time. This could be done over weeks, days or even hours to show what time of day most Tweets were sent at.

Should I implement sentiment analysis it may also be useful to include pie charts which would provide an instant impression on attitudes towards the topic either positive, negative or neutral.

I feel that it is important to create a balance of displaying a good amount of information on screen without overwhelming the user so I hope to make the display customisable with the user able to adjust the size of certain areas of the dashboard and minimise or close those that they do not wish to use. A final desirable feature would be to save or print reports showing some of the information returned from the search.

# Goals for this project

* As this is my first project using Twitter API it will require me to learn and familiarise myself with the API, the data available and its applications
* I aim to create an efficient and accurate method of ingesting and storing the data from Twitter API
* When ingesting the data, I would like to trim it down to the information which will be relevant as there are over 100 pieces of information relating to each Tweet available and much of this will not be used
* Ability to stream live tweets or search historical tweets based on a keyword search (as previously mentioned I will experiment with both to see which is most suitable)
* Create algorithms and methods to filter, count or group data based on selected criteria
* Provide time-series graphs showing Tweets over a period of time
* Provide a feature to show Tweets by location
* I would like to create a functional and easy to use dashboard that allows the user to customise their search and view the data they have requested on easy to read graphs, tables or charts

## Added desirable goals

* A desirable feature of the project would be to include sentiment analysis to allow the user to see if tweets were positive or negative about the subject in which they are interested
* Another desirable feature would be the ability to show the most popular tweets or users related to a particular topic by ranking them in terms of retweets, favourites or followers
* It would also be an added feature if the user could print or save a report of the information returned allowing them to share their findings with others more easily

# Criteria for Success

My final project should provide a user-friendly and intuitive interface that is highly functional and aesthetically pleasing. On start up the user should be prompted to enter the subject they wish to search for. Once the search term has been set the system will either stream current tweets or search for historical tweets for a short period of time and select only the desired parameters which are pre-defined and insert them into the database. Once the data has been gathered the methods and algorithms for analysing the data will run and feed the information into the methods for visualisation. The visualisations will then be available for the user to view and adjust the size, maximise or minimise the various components to make the information they desire easily readable. Finally, the user has the option to print or save the information and visualisations displayed on the dashboard.

## Requirements

* Connect to Twitter API
* Trim data to exclude unwanted parameters
* User defined keyword search
* Download and store tweets and meta-data
* Algorithms to perform data analysis
* Visualisation of data
* Interactive dashboard to present visualisations
* Print or save a report of the information and visualisations from the dashboard

# Software

* Language: Java
* Twitter4j – A Java library for interacting with Twitter API
* MongoDB/MySQL\* will trial MongoDB
* Twitter API

# Time Planning

My initial estimates allow for problems I may encounter such as bugs and preparation for demonstrations in December. I have also set the target of having my coding and dissertation completed before May, meaning that if there are unforeseen complications I have left sufficient ‘slack’ to accommodate this.