Functional Specification

Table of Contents

- 1. Introduction
  - 1.1. Overview
  - 1.2. Purpose
  - 1.3. Scope
  - 1.4 Business Context
  - 1.5. Glossary
- 2. General Description
  - 2.1. Product/System Functions
    - 2.1.1. Website Structure
    - 2.1.2. Types of Statistics
    - 2.1.3. Dmitrijs Milajev's Data Set
    - 2.1.4. Collecting Data and the Database
  - 2.2. User Characteristics and Objectives
  - 2.3. Operational Scenarios
    - 2.3.1. Navigate to General Statistics
    - 2.3.2. Navigate to Statistics Sorted by Stance and Country
    - 2.3.3. Search Statistics for Tweets Containing a String
    - 2.3.4. Send an Email
  - 2.4. Constraints
    - 2.4.1. Time Constraints
    - 2.4.2. Accuracy Control Constraints
    - 2.4.3. Tweet Availability Constraints
- 3. Functional Requirements
  - 3.1. User Selects an Area
  - 3.2. Search a Keyword
  - 3.3. Sentiment Analysis Algorithm
  - 3.4. Interacting with the Database
  - 3.5. Web Interface
  - 3.6. Statistical Generation
- 4. System Architecture
  - 4.1. Website
  - 4.2. Web Server
  - 4.3. MySQL Server and Database
- 5. High-Level Design
  - 5.1. Data Flow Diagram

    - 5.2. Context Diagram
- 6. Preliminary Schedule
    - 6.1. Pert Chart
        - 6.1.1. Design and Implementation of Sentimental Analysis
        - 6.1.2. Implement and Automate PHP, Python and R Scripts
        - 6.1.3. Documentation and Video Walkthrough
- 7. Appendices

1. Introduction

1.1. Overview

Since its launch in 2006, Twitter has quickly become one of the world's most popular microblogging sites with an average of over 317 million active users.

As a result, it grants us a glimpse into the thoughts of users in real-time.

1.2. Purpose

The goal of our project is to highlight the relationship between the network location of a posted tweet and its context.

1.3. Scope

Using the current affairs topic of the 2016 Brexit Referendum as a control, we will analyse harvested tweets via their associated hashtags (e.g. #brexit, #leave, #remain, #Yes2EU, etc).

We plan to gauge positive and negative public opinion within each region of the UK, namely that of Northern Ireland, Scotland, England and Wales.

The statistics of pro and anti-Brexit tweets will be compared with that of the official result to see if opinions have changed dramatically in each region.

International opinion will also be interpreted via the data harvested from hashtags such as #brexit, #Brexit, etc.

The statistical data will be plotted visually via its network location from each of the four regions and by country for international data visualisation.

1.4. Business Context

The potential business context with regards to this project is that of a political one as campaigners can gauge online support.

These estimates of online chatter can be used as a basis for engagement with potential voters and supporters.

The project has vast potential to be used as a research tool for those studying Political Science or for those with a vested interest in politics.

1.5. Glossary

API

'Application Programming Interface', a set of subroutines, tools and protocols which act as the fundamental building blocks for developing an application.

Cron

A time-based scheduler within Unix-based operating systems such as Linux, Ubuntu, etc.

Cron can be used to schedule commands or scripts to be run at specific times or in intervals.

Django

A Python-based framework aimed at developing web-based applications.

Framework

A platform containing a set of pre-built tools and protocols in order to allow users to create and develop applications quickly and easily.

JSON

'Javascript Object Notation', allows for information to be stored in a collection which can be easily accessed and read by humans.

mySQL

An open-source SQL database management system which utilises SQL in order to insert, update, delete, retrieve and create database entries.

Poultry

A tweet archiving manager which allows for collection of tweets to be grouped by date.

Works in conjunction with twarc.

Python

An object-oriented programming language used to create various programs and applications with a user interface.

R

A programming language for statistical computing and graphics.

R includes a large range of functions for the computation and visualization of data both linearly and nonlinearly.

SQL

'Structured Query Language', a programming language designed to be used with a database management system in order to read and retrieve information.

twarc

A command-line Python library that allows for Twitter JSON data to be archived.

2. General Description

2.1. Product/System Functions

2.1.1. Website Structure

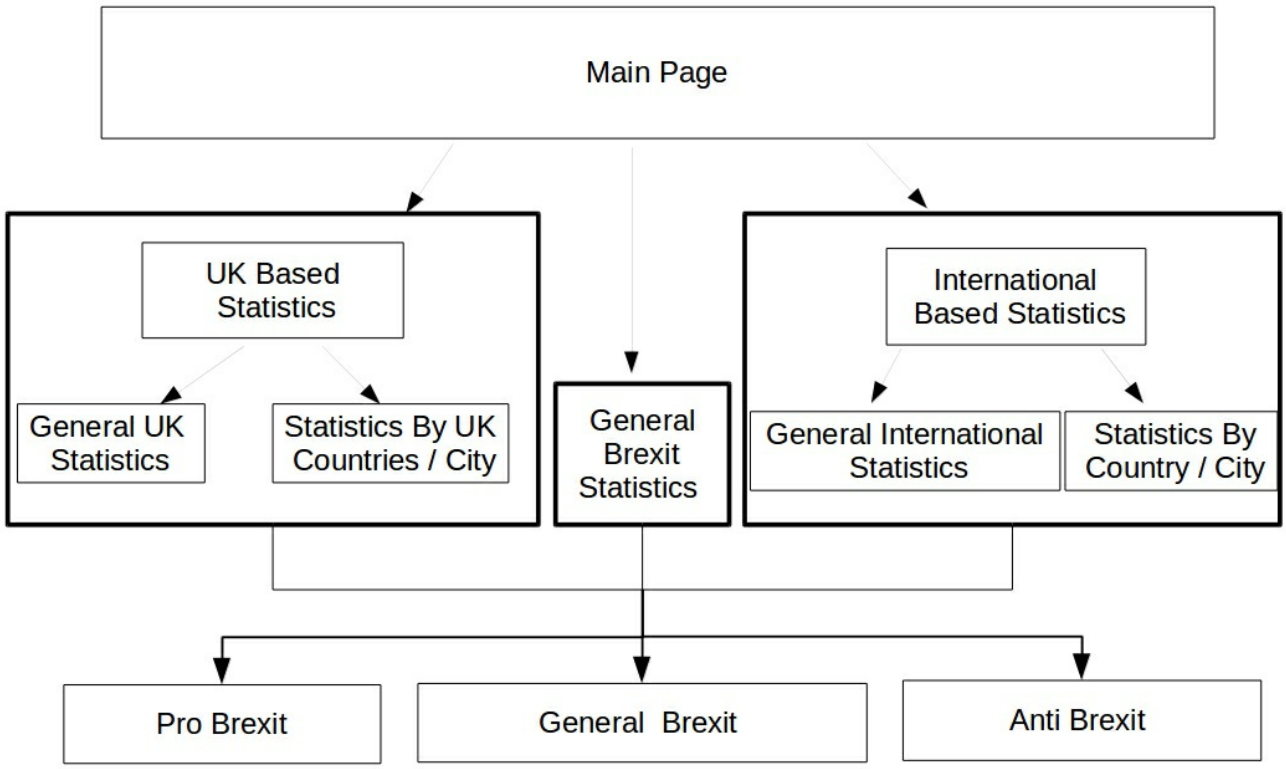
Our product will be presented to our users via a website.

The website's core content will be split up into three main statistical sections:

- Statistics on tweets located within United Kingdom
- Statistics on tweets irrespective of region
- Statistics on tweets from various non-UK countries.

These sections will also be split up by stance - i.e. Pro-Brexit (Leave), Anti-Brexit (Remain) and general Brexit statistics (neutral and/or undecided).
We will use Django to aid us in creating a visual, intuitive website which will visualise the data and statistics we have gathered.

The website will guide users through its structure. allowing them to access their desired information they want to find.



2.1.2. Types of Statistics

We will be showing different variations of statistics with significant value to Brexit - e.g. spikes in number of tweets per day with regards to Twitter discussion of Brexit and Brexit-related events. Major spikes often indicate events relevant to Brexit and how they trigger the public to voice their opinions.

We will show statistics based on sentiment and stance - e.g. showcasing the total tally of Brexit-centered tweets are positive versus the total tally of negative Brexit-centered tweets.

Methods of visually portraying said data will include that of bar charts, pie charts, box plots and many more. This will ensure that users can immediately understand the generated data.

We will use R to perform various statistical analysis on our data. A PHP script will pull data from the database and pipe it to an R script where the required statistical analysis will be performed.

2.1.3. Dmitrijs Milajev's Data Set

Dmitrijs Milajev has published a dataset of Brexit Tweet IDs and a Tweet Collection Manager software named Poultry. The data set contains the IDs of over twenty million tweets regarding Brexit collected between May and June 2016. Poultry allows us to split the data set by date, allowing us to better manage the set locally.

This dataset will be used in a multitude of ways. Firstly, we will test our sentimental analysis algorithms and perform statistical analysis on this particular dataset.

Secondly, this dataset is of the utmost importance as the Twitter API only permits the mining of tweets of no later than a week from the current time.

As this dataset is dated from the time frame of Brexit Referendum - i.e. May - June 2016- it will allow us to compare the sentiment towards Brexit now against that of the sentiment towards Brexit at the time of the referendum.

The dataset will be utilised with twarc by means of downloading the tweets in a JSON format using the tweet IDs as specifiers.

2.1.4. Collecting Data and the Database

We will collect our own data set of tweets through twarc - a Python library designed for the mining of tweets.

We will gather a collection of hashtags that reference Brexit such as #Brexit, #leave, #remain, etc. and use it alongside string filtering methods in order to determine which tweets will be added to our dataset of tweets.

We will categorize these data sets by date and run our sentimental analysis algorithms on them in order to determine the stance.

The same approach will be taken with Milajev's online dataset. Each tweet will then be inserted into our MySQL database. The sentimental analysis algorithms we will use have not been fully determined- however we will take a 'Bag-of-Words' approach combined with a lexicon scoring system. Naive Bayes sentimental analysis may also be used but more research must be done to finalise this decision.

2.2. User Characteristics and Objectives

Our user community will consist of several types of users ranging from but not limited to:

- Political parties
- Political campaign staff
- Journalists
- Web surfers who have an interest in politics
- Web surfers interested in big data analysis
- Students studying political science

As our system is primarily a web based system, all potential users should be able to use the website with relative ease and will not require a vast amount of prior technological experience. Journalists may use our website to gather interesting statistics and derive their own stories from the information and statistics we provide. Political party staff looking to provide their politicians with useful information on public opinion regarding Brexit may look to our website for statistics in order to gauge public opinion.

Brexit is a topic of popular discussion now and for the near future as the departure of the UK from the EU is unprecedented with a ripple effect that could potentially last decades as the relationship between the UK and Europe will be fundamentally re-written. People with an interest in this topic may seek information and wish to research public opinion towards Britain leaving the European Union.

We will not have a lot of time to catch the web surfer's attention so we will need to make our home page catch their attention quickly. We will portray some of the more interesting results of our analysis on the home page in a minimalist design so as to prevent cluttering the homepage and overloading the user with excessive information. A guide to finding

the categories of information the user may want will also be on the front page to aid navigation.

Political staff members to various parties will be some of the more frequent users of our website and get the most out of it. They will use the portrayal of our statistics as basis for a sample set opinion poll, namely that they will look for statistics which show that public opinion is in support of their party or political stance in order to strengthen their political position. Journalists writing about the topic of Brexit can reference statistics found whilst using our website.

Users can look for information within specific areas and with regards to specific political stances and sentiments. Due to the fact that there is a considerable amount of variable combinations with which users will narrow their search, the website structure must be clear, concise and easily navigable - one of our major aims. However, users such as casual web surfers may not have a specific topic in mind and simply wish to view the portrayal of the visual data in order to satisfy their curiosity. Thus, the website is obliged to cater for this particular type of user with a generalised section with statistics and information of various stances and regions.

2.3. Operational Scenarios

2.3.1 Navigate to general statistics

A user wishes to see statistics and information on general Brexit tweets regardless of stance and region - i.e positive, negative and/or neutral sentimental tweets. The user hovers the mouse over the navigation panel and the panel drops down with a menu. The user clicks ‘General Statistics’. The user is brought to the ‘General Statistics’ page.

2.3.2 Navigate to statistics sorted by stance and country

A user wishes to view statistics and information about Brexit tweets that reference a specific stance from a specific country, namely that of Northern Ireland, Scotland, England, Wales or outside of the UK.

The user hovers the mouse over the navigation panel and the panel drops down. The user then hovers the mouse over United Kingdom Statistics and hovers the mouse over the country they wish to select. At this point, three options appear: ‘Pro-Brexit’; ‘Anti-Brexit’ or ‘General’ and the user clicks on their desired choice.

If the country of their desired choosing is not within the realm of the UK,
the user hovers the mouse over the navigation panel and then clicks on ‘International Brexit Statistics’. Now, they are brought to a page with a list of countries with provided statistics. From here, they can choose their desired country and the statistics will be categorized by stance within the page if applicable.

2.3.3 Search statistics for tweets containing a string

A user wishes to search for tweets containing a particular string. They use the navigation panel to select ‘Search a Keyword’ and are taken to the ‘Search a Keyword’ page.

The user fills out a form which takes in the following information:

- String to be searched
- Type of data desired
  - E.g. Percentage of total Brexit tweet
  - E.g. Total number of tweets containing the string
- Country from which tweets originated (optional)
- Sentiment of tweet (optional)
- Political stance (optional)

- Method of representation - e.g. bar chart, pie chart, etc.

The user submits the form and the representation of data appears.

2.3.4 User sends an E-mail

User navigates to the contact page of the website. User fills in a contact form with their name, e-mail, and query. The user clicks ‘Send’ and an e-mail is send to our e-mail account. This can be useful for the user if they want to suggest an additional function to the website, report any issues with the website or ask us any queries.

2.4 Constraints

2.4.1 Time Constraints

The project must be completed by the 10th of March 2017 and we have been provided with a 6 week window in which we must develop the project and all necessary documentation.

2.4.2 Accuracy Control Constraints

It is impossible to guarantee the sentiment result of each tweet is accurate due to the fact that we will be unable to check the result from sentimental analysis of every single tweet we download. To counter this, we will test our sentimental analysis algorithms against tweets of sentiment we know, record the accuracy in order to use it as an estimate. We will also need to account for a margin of error when we generate the statistics.

2.4.3 Tweet Availability Constraints

In accordance with Twitter’s API, tweets that were posted one week prior to the current date are available for download/data mining. This ultimately means that we will be unable to gather tweets tweeted between the time we have our working system and the time in which Milajev's data set was collected.

3. Functional Requirements

3.1. User Selects an Area

Description

The user selects a single area from a choice of two drop down menus - one listing the regions of the UK - Northern Ireland, Scotland, England, Wales and an international one which provides a list of locations from which to choose that are outside the United Kingdom-e.g. Germany, France, Ireland, etc. This selection acts as a filter in order to allow users to view a sample set of statistics regarding Brexit in a particular region or country.

Criticality

This functional requirement is crucial to the fact that it acts as a filter for the system to narrow down the search results. Without it, the system will not be able to determine the statistics of the sentiment of the tweets and will be unable to produce a final result. This is also the user’s first point of contact with our application and as such it must be easily accessible and easy to use.

Technical Issues

The main technical issues that we will face will be ensuring that the dataset of harvested tweets has been correctly inserted into the database schema in order to ensure that users will be able to make queries and that the database itself can retrieve the data related to them.

Language will also be a an issue for us as our system will designed solely in English, therefore we need to take into account that we will only be able to process international tweets written in English which will be resolved by setting the "lang" parameter to "en".

We will also have to correctly code the location variables as provided in the Twitter API, such as "geocode" which takes in parameters for longitude, latitude and radius of search.

Dependencies with other requirements

Dependent on the user successfully starting the session with our application and on the dataset of tweets having been inserted correctly into the database.

3.2. Search Keyword

Description

This feature require that users input a string and will allow them to view statistics for Brexit-related tweets which contain this aforementioned string or keyword.

The system will then return a chart of the percentage of positive or negative tweets containing this keyword which will allow the user to ascertain whether or not the keyword has positive or negative connotations with regards to Brexit.

If their chosen keyword is not contained within any of the tweets, the system will return a prompt which will display "Your keyword was not found" and ask the user to try again with a new keyword.

Criticality

This function, while not critical, acts as an interesting feature for our system. Users can query the sentiment of certain keywords such as "Theresa May" or "Scottish Independence" with regards to Brexit and whether these keywords are associated with pro or anti-Brexit sentiment. As a result, users can have a customised view of the resulting data statistics.

Technical Issues

The technical issues here will be similar if not the same as 3.1. If a user wishes to search for a hashtag, the search will look for tweets containing that particular hashtag.

Dependencies with other requirements

See 3.1.

3.3. Sentiment Analysis Algorithm

Description

Once a user has successfully entered their search keyword, the system must run this algorithm on each individual tweet in order to determine whether it expresses positive or negative sentiment.

Criticality

This algorithm is critical in order to determine the sentiment of each tweet and tally the count of positive and negative opinion of the Brexit referendum and to plot this data visually.

Technical Issues

We will have a number of issues regarding this functional requirement. Part of the Sentiment Analysis Algorithm will include a lexicon based approach where words are rated between zero and one for positivity and negativity. This score will be taken into account when deciding the sentiment of a tweet.

Bag-of-Words as a standalone algorithm will not take negative grammar rules into account. For example, the sentences, “Brexit is a good idea” versus the sentence “Brexit is not a good idea” will both result in the algorithm declaring that both sentences are of equal positive sentiment because they both contain word ‘good’ which has positive connotations.

We would need to incorporate an n-gram set of negative grammar rules such as negative prepositions, conjunctions, blind negation words (a set of words which indicate a lack of quality - i.e. “Brexit needs proper planning”)(Palanisamy, Yadav, Elchuri, 2013), etc., alongside that of the Bag-of-Words algorithm in order to increase the accuracy of the data. We have not finalised our decision to use this particular algorithm and are still researching other sentiment analysis algorithms such as Naive Bayes algorithm.

3.4. Interacting with the Database

Description

Once a user has specified their search term(s), the system must establish a connection to the database containing the tweets. The system must ensure that the database is able to return any and/or all data relating to their query.

Criticality

This function is important as users must be able to access the correct information within the database via their queries. The database must have the correct columns with regards to hashtags, location, positive sentiment or negative sentiment and tweet date.

Technical Issues

As the aforementioned above is primarily back-end work, we need to ensure that all information has been inputted correctly via relations and that once a query is made, the correct information relating to said query will be returned.

Dependencies on other requirements

The user must enter a valid query – i.e. a query that consists either of the name of a country or region within UK or a valid string with characters ranging from ‘a-z’ or ‘A-Z’ and ignore any special characters such as ‘!?%$^’, etc.

3.5. Web Interface

Description

The website will the the primary source of user engagement with our system. All transactions with regards to the back-end database and contact forms will be done via PHP. The web-interface of our system will also allow users to select a

customised view of specific statistics which will be visually displayed.

#####Criticality

This is a crucial functional requirement as our system is heavily dependent upon visual elements as we are returning a graphic statistical representation. We must present the data to end-users in a logical and concise manner as well as ensuring that interface is easy to navigate and accessible in order to accommodate the requirements of our users. All aspects of the interface including that of menus, navigation buttons, links, etc must be precise and allow users to accomplish their tasks quickly and effectively.

Technical Issues

In accordance with the feedback that we were given in our project proposal presentation, our user interface must be validated against the heuristics of Human Computer Interaction (HCI) – i.e. Norman's Seven Principles of Design, Nielsen's Ten Usability Heuristics, etc. This is to ensure that the interface is a strong user-friendly model based on the fundamental HCI principle of usability with an efficient and minimalistic design.

3.6. Statistical Generation

Description

The statistics generated by our system will be required in order to create the visual representation of the data as specified by the user. These formulas will be composed of standard deviation in order to determine how the variance within each area, population probability so as to make an educated estimate about the opinion of the various populations within each region, standard mean and so on.

We will also need to account for a margin of error which can be resolved if we calculate the numbers with a basis of a confidence level in the range of 90 – 95% of probability.

The data read from the database entries will be counted and then supplied to a R script to act as the parameters for the various statistical formulas we will use.

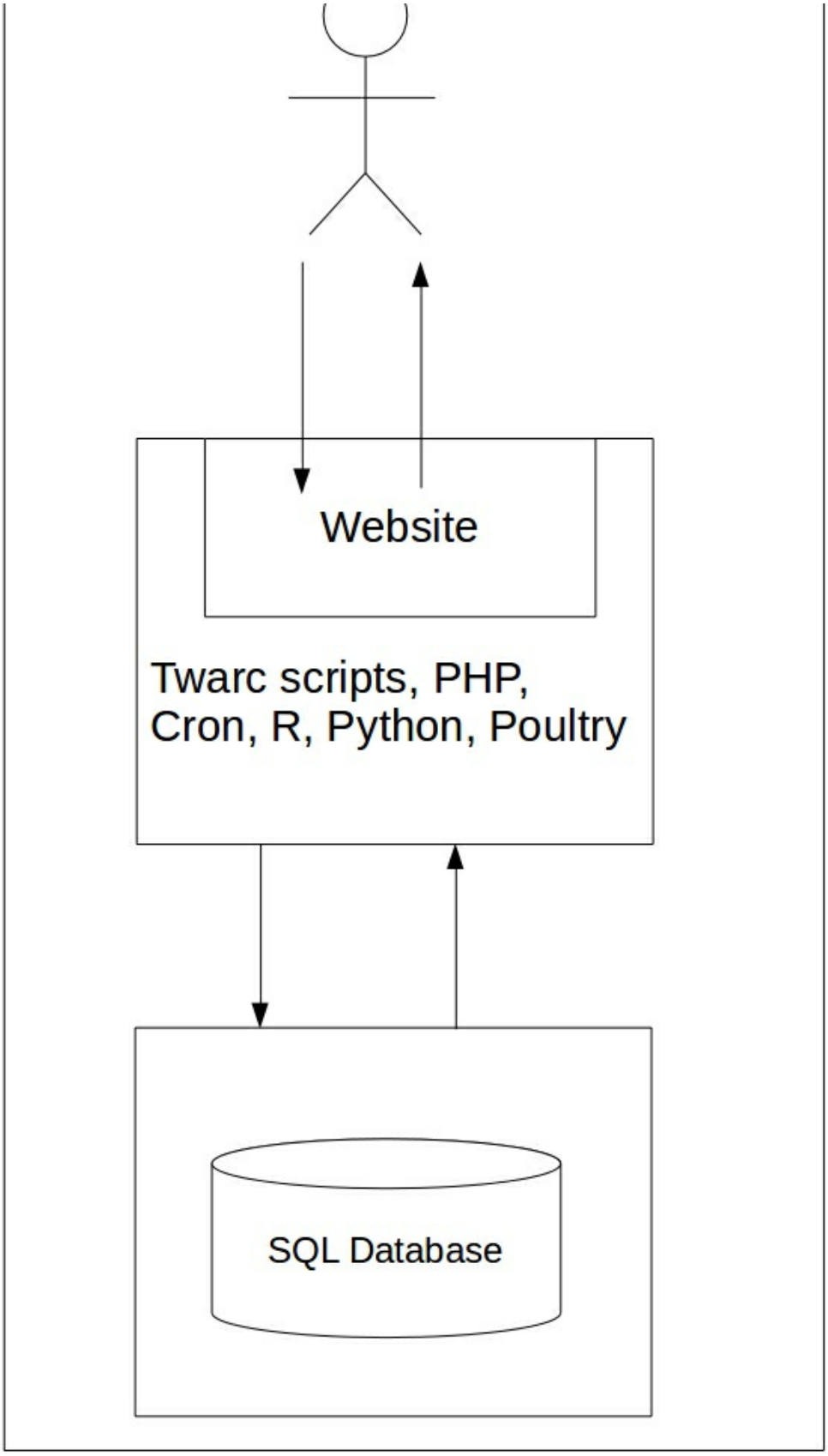
Criticality

This functional requirement is important for without it we will have incorrect data returned to our users and will provide false information.

Technical Issues

The technical issues surrounding this functional requirement are ensuring that all data gathered from the database queries such as total counts of positive and negative tweets, tweets per location searched, etc are correctly piped to the R script as parameters for the statistical formulas. We also must ensure the correct statistical formulas with respect to opinion polling of a sample set are run.

4. System Architecture



4.1 Website

The website will be the point of contact with the users. We will use a mixture of HTML, CSS, Javascript and PHP to create our website. The Django framework will be used to aid us in creation of a dynamic and accessible website.

4.2 Web Server

The web server is where our computation and analysis will take place. We will use Cron to run scripts and commands at scheduled times. Twarc will be used to download tweets, we may use poultry to filter the tweets. PHP will be used to interact with the MySQL Database. We will need Python to perform sentiment analysis, and R for the computation of statistics.

4.3 MySQL Server & Database

The MySQL server is where we will keep our database of tweets. PHP will insert, select and create tables and elements from tables. The database will hold information such as tweet content, geolocation, sentiment and stance amongst other data.

5. High-Level Design

In this section, we have provided a Data Flow Diagram and a Context Diagram as a per SSADM design approach in order to present the data process of our system.

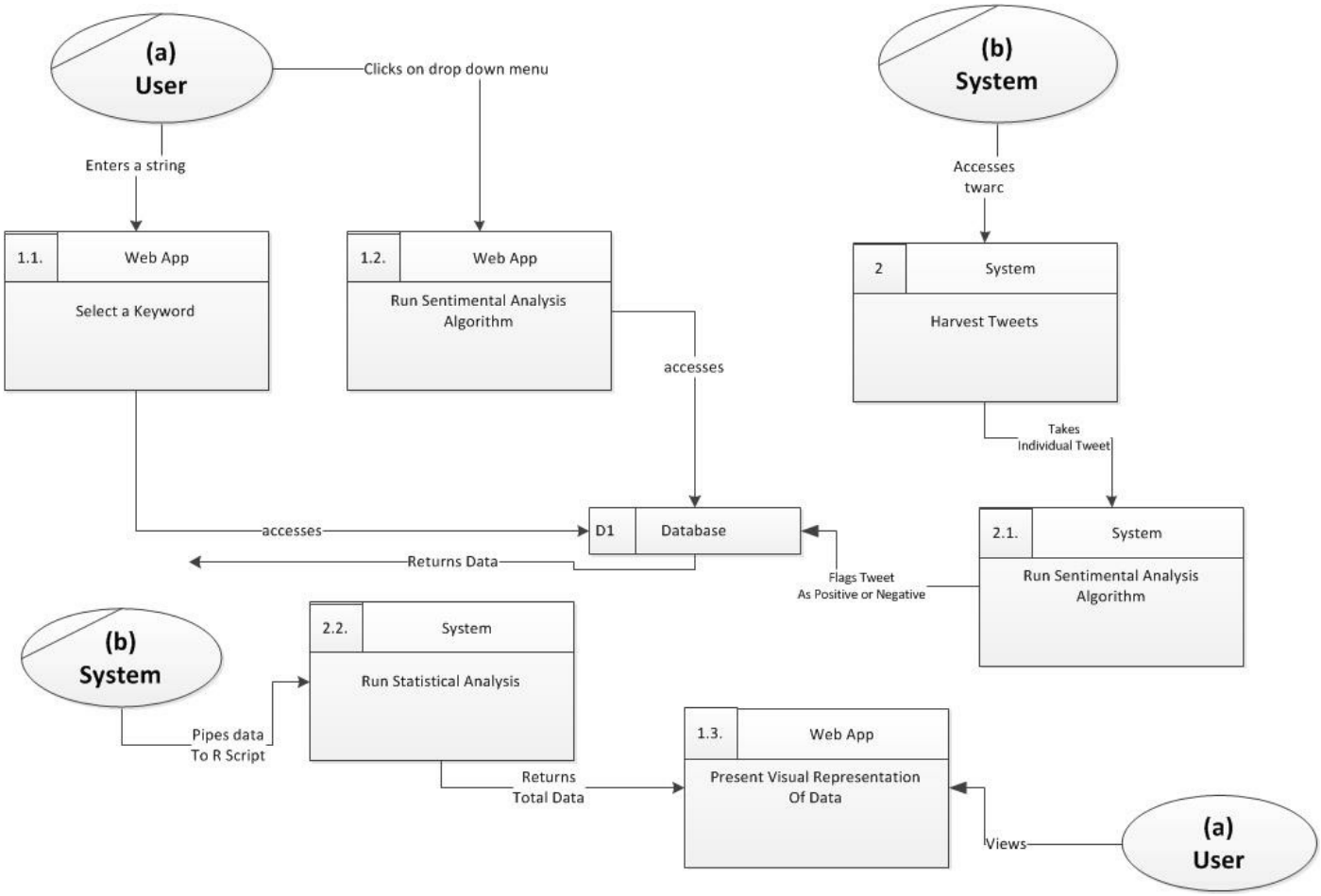
Data Flow Diagram

The Data Flow Diagram or DFD allows us to showcase the flow of data within the system. We have provided the datastore and external entities in order to display the data flow.

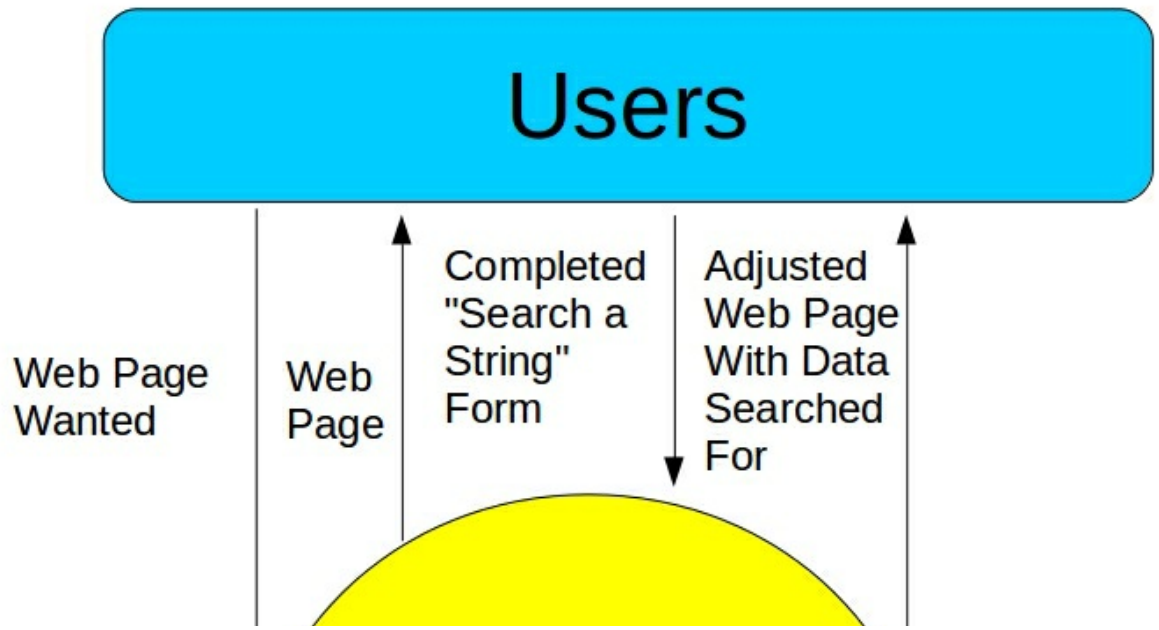
Context Diagram

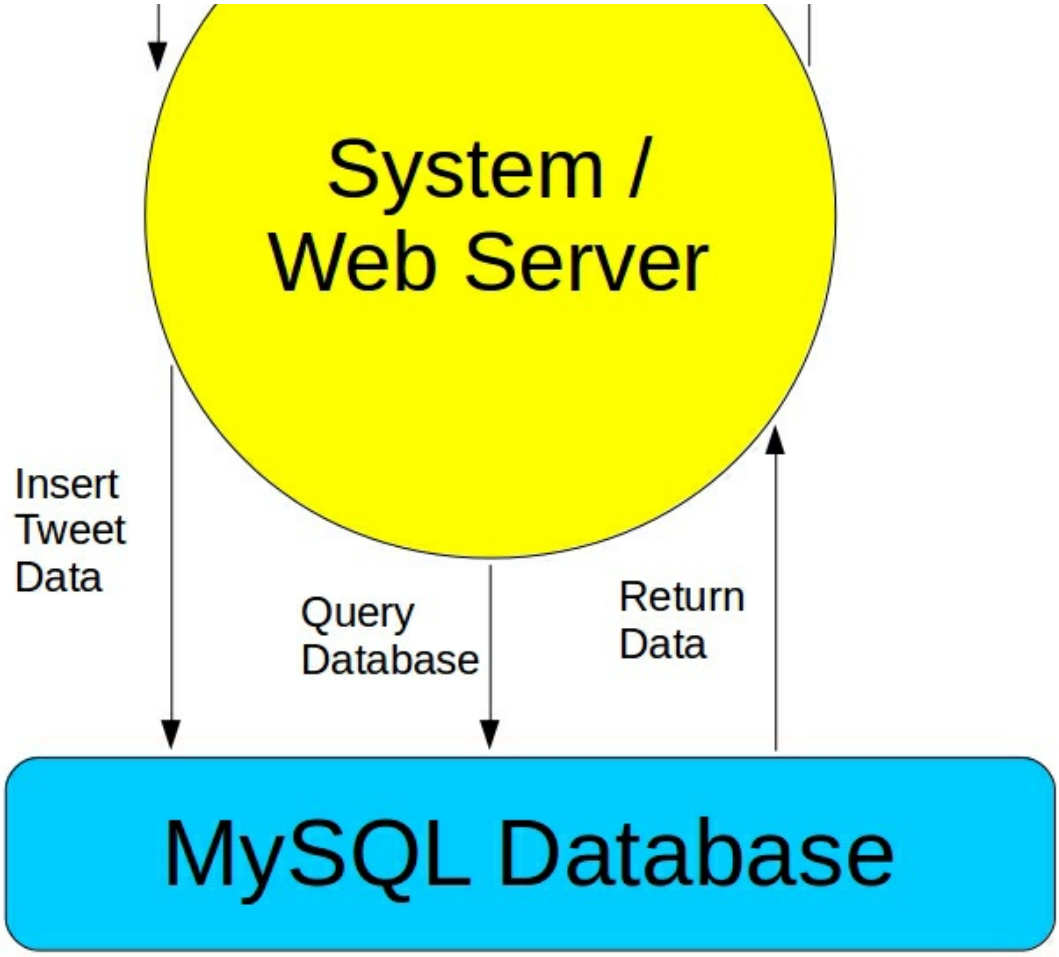
The Context Diagram shows the flow of information between our system, the database and the user.

5.1. Data Flow Diagram

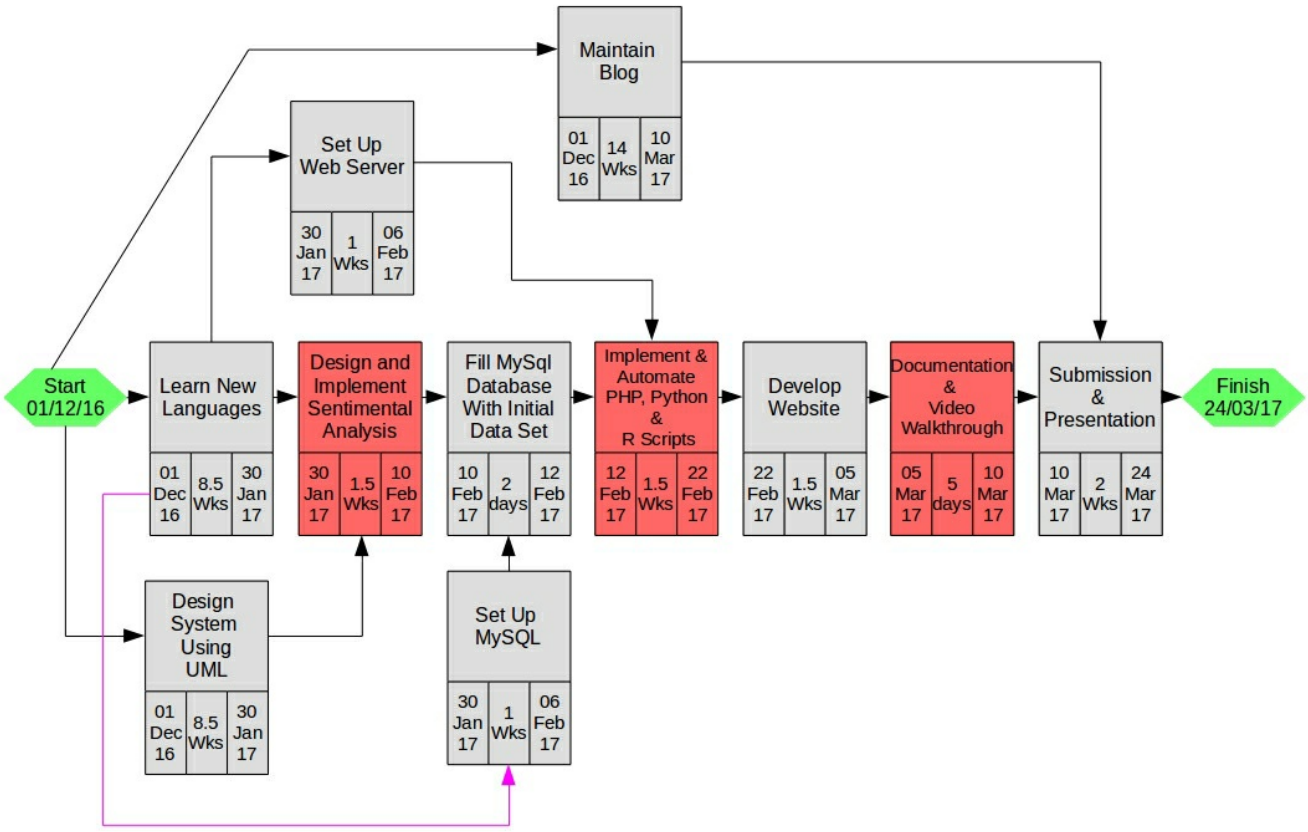


5.2 Context Diagram





6. Preliminary Schedule



6.1 Pert Chart

This Pert Chart shows the preliminary schedule for the project. Any box marked in red is broken down further below. The project tasks will mainly be split up in an Agile manner between us.

6.1.1 Design and Implementation Sentimental Analysis

This phase will involve designing the sentimental analysis algorithms, it will also include the development of scripts implementing said algorithms and finally automating the scripts on the server using Cron.

6.1.2 Implement and Automate PHP, Python and R scripts

This phase will include the following:

- Gather hashtags and strings by which to filter tweet downloads by
- Create scripts to download Brexit tweets following the aforementioned filters
- Develop PHP scripts to communicate with the MySQL database
- Develop statistical analysis scripts
- Automate developed scripts using Cron

6.1.2 Documentation and Video Walkthrough

This will include the Video Walkthrough, the Technical Specification and the User Manual.

7. Appendices

- Dmitrijs Milajevs provides
  - a collection of Brexit Tweet IDs
  - Poultry a Tweet Collection Manager
  - Twarc, a tweet download aid.
  - Available at http://www.eecs.qmul.ac.uk/~dm303/brexit/