SENTIMENT ANALYSIS OF SOCIAL MEDIA WITH DATA VISUALISATION

BY

JOHN KELLY

SOFTWARE DEVELOPMENT

RESEARCH DOCUMENT

5TH APRIL 2017

TABLE OF CONTENTS

ABSTRACT

The purpose of this document is to outline the research conducted over the course of this project. Millions of users share their thoughts on microblogging sites about a multitude of issues. This is a data rich resource that contains real-time opinions on issues which this project will investigate natural language mining to find the sentiment. Included in this document is the thought process from which social media, technologies and cloud platforms to be used.

1. INTRODUCTION

The purpose of this document is to present a review of the research conducted as part of the fourth-year project. This document will be changing and adapting due to the iterations of the agile approach taken for this project. The aim for this project will be to create a generic sentiment analysis tool which will aid developers in creating different classification tools for any type opinion and platform using any type of classifier required to fulfil their goal. This project will be implemented as a sentiment analysis tool which can mine the public sentiment opinions towards a certain term or phrase.

This document contains several main sections. The first section, explores what sentiment analysis is. It will consider the different ways of classifying a section of text into different sentiments and the complexities that are faced with each type. The next section discusses the research into the different social media platforms and documents the choice of which would make an ideal candidate for this project. In the case of this project, the research that was carried out considered the number of people on each network, number of posts issued each minute, and the Application Programming Interface (API) that is required to retrieve the information to successfully carry out this project and any limitations that they come with. The next section presents the research into the different technologies that are appropriate for this type of project. This includes the frontend technologies as well as the backend technologies. Also, discussed here are the packages that will help with the development process. This includes the type of database that should be used and the reasons why one is better suited than the other types. Finally, the requirements will be researched and the minimum specifications will be decided upon.

As the number of people with access to the internet grows so does the number of users on popular microblogging and social media sites. This allows users to share opinions which is invaluable resource to market researchers. This project aims to help people gain an understanding and therefore have an improved image of what the public thinks of different products, topics or brands. Along with a greater understanding of their audience, more specific advertising could be implemented to the right audience.

2. SENTIMENT ANALYSIS

Sentiment Analysis is a very complex field of study. It examines the audience and how they react to brands, topics or any other subject. A fundamental use of a product like this would be for market researchers for companies to learn what products their target audience prefers and what aspects of it they like the most. It can also be used to keep up with current trends. This could be used for creating advertising, to echo the thoughts of the consumers on the chosen topic.

However, due to the intricacies of the English language, some of the linguistics aren't well understood. For example, "*I bought a phone today. It's really fast. Although, the camera isn't great.", and examine each section of it. "I bought a phone today",* this sentence is a complete neutral and expresses no sentiment or emotion towards anything to do with the phone. Now, take the next sentence, "*It's really fast*", this expresses a positive point of view towards the phone and how quick it is. Now taking the next sentence, "*Although, the camera isn't great*". This expresses a negative sentiment towards the phone due to the camera. Now examining it as a whole, it contains all three sentiments which are normally used in expressing how one feels about a topic, positive, neutral and negative. It cannot be said that it is positive or negative and because there is a sentiment expressed, it would be incorrect to assume that it is neutral. What would be the final thought on the tweet? Would it be fair to only give it one definitive answer? Is it possible to only provide one answer?

Another way of selecting the sentiment of a sentence is separating it into different sections. These sections will contain the keywords in the sentence which without these words, the sentence would be meaningless. They provide the actual meaning of it. For example, if we examine the sentence "*The tool is useful".* They're two keywords expressed here, "*tool"* and *"useful".* The use of the word "*useful*" expresses a positive attitude towards the tool. We can deduce that the main topic of this sentence is the tool and the sentiment towards the tool is positive. However, if we change the sentence slightly to, "*The tool is less than useful*", it completely changes the meaning of the sentence. This is called a sentiment shifter. (Donkor, 2013)

3. SOCIAL MEDIA

Choosing the right social media platform was a vital part in this project. It must be a widely-used platform to limit the number of posts from the same user. This will allow the application to have a more universal opinion on a chosen topic. The platform needs to support an Application Programming Interface (API) to allow programmatic access to the information.

3.1 FACEBOOK

The largest social media platform in the world is Facebook. With a current number of 1.59 Billion monthly active users. (Maina, 2016) Facebook is a key competitor to be used for a sentiment analysis tool for this project because of the number of active users it has. There's an average of 293,000 statuses and 510 comments are posted every 60 seconds. (Zephoria, 2016)

Facebook has one API called the Graph API which is restricted to returning public posts, pages, events, groups and check-ins. (Facebook, 2016b) This social network has several different privacy settings which many people take advantage of. In 2014, Facebook changed the default privacy setting to 'Friends Only' which means that all the status' people post won't be publically accessible unless they specifically select that option. (Magid, 2014) This means that this API won't be able to access posts from new users by default which limits the number of posts available. This API limits access to 200 calls per hour per user which will return a default limit of 25 results for each call. (Facebook, 2016a) According to a survey, 20% of Facebook users say that they have their profiles completely public. (Daily Mail, 2012)

3.2 TWITTER

Twitter is also a very popular social network platform as well. Although they don't provide a total monthly active user number. Researchers try to estimate the number by using different algorithms to give an idea of how many active users they have by basing it off

the number of tweets sent each day. Its estimated that Twitter has 320 Million monthly active users. (Maina, 2016) There's an average of 350,000 tweets sent every 60 seconds. (Internet Live Stats, 2016) Twitter also allows users to choose between two different privacy settings. Profiles are set to public by default which allows any one access the tweets by any user unless they specifically set their profile to the protected setting provided by Twitter. (Twitter, 2017c)

Twitter also limits the results to only public posts. However, in a survey, it was revealed that less than 10% of Twitter users have private profiles. (Moore, 2009) Twitter has two main API's that could be used in this project. The first is a Search API. This will allow the developer to create a query to search for certain terms and Twitter will return tweets from the past seven days to do with that term. This API limits access to 180 calls per fifteen minutes per user which will return a limit of 100 results for each call. However, if the user uses OAuth authentication, they will be able to make 450 request per fifteen minutes. (Twitter, 2016a)

The second API allows the developer access the current trends in a certain area. This is done by the trends/place API accepting a Where On Earth ID (WOEID) which is used to access the current trends in a country. The limit for this API is that it can only be called 75 times every 15 minutes, however, the trends are cached and refreshes every 5 minutes. (Twitter, 2016b)

Figure 1 shows the difference between how a normal REST API like the Search API would work compared to the Streaming API. The normal API is a once off connection which asks Twitter for information and Twitter will return it in a JSON format. Whereas, the Streaming API is a continuous connection made.
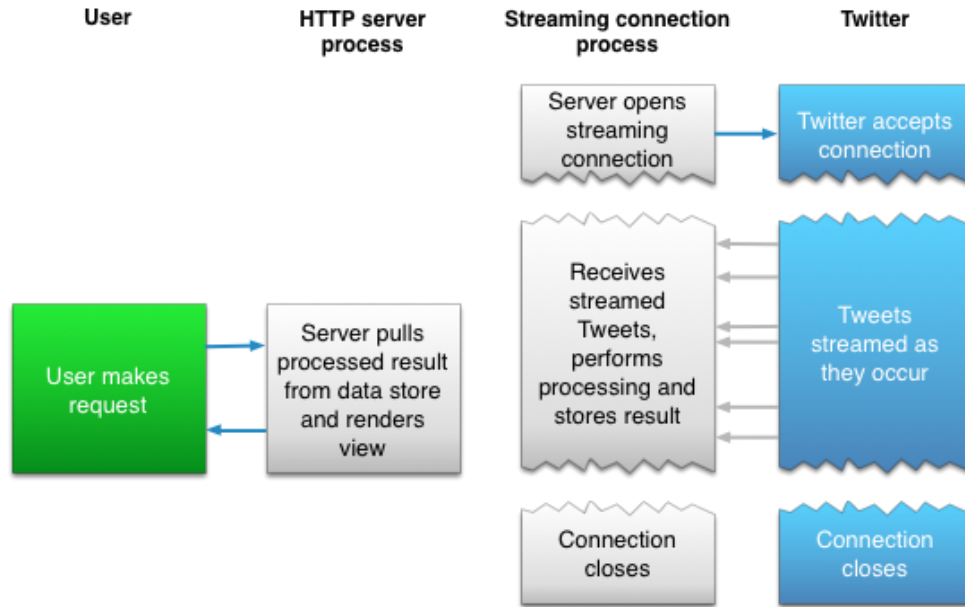
Figure 1. Shows the difference between Streaming and REST API's (Twitter, 2016c)

CONCLUSION

Twitter has a greater percentage of public profiles; it also provides more API calls per hour with a greater number of results returned. This enables this application to gather more information about topics and therefore, give a more accurate prediction on the sentiment of a searched term.

Although, Facebook has the great number of monthly active users, it's unfair to compare it to Twitter in that way because Twitter doesn't provide that actual number. Facebook's API is more limiting than the Twitter's API where this application is concerned. The number of posts per 60 seconds is also a big difference. This would suggest more topics of information being discussed on Twitter rather than Facebook.

Considering the difference between the two social networks, it has been decided that Twitter will be used for this application. The number of results will give this application a greater chance of providing an accurate sentiment towards the chosen term. Although, Twitter is chosen as a demonstration of this application working, the aim of this project to allow developers to choose any platform that they require.

4. FRONTEND TECHNOLOGIES

4.1 HYPERTEXT MARKUP LANGUAGE

Hypertext Markup Language is more commonly known as HTML. It's a standardised way of describing the structure of webpages viewed over the World Wide Web. It was standardised by the World Wide Web Consortium (W3C). Each document contains HTML elements which describes different sections of the page and how it's displayed. The latest version is HTML5. (W3C, 2016)

4.2 CASCADING STYLESHEETS

Cascading Style Sheet, more commonly known as CSS, is used to describe how the HTML elements should appear on the webpage. This includes the colour, width, height, etc. It's separate from HTML so it doesn't affect the actual structure of the webpage. (W3C, 2016)

Syntactically Awesome Stylesheets (Sass) is a style sheet language that compiles to CSS. It contains many features that CSS doesn't include. For example, variables, inheritance and mixins. (Sass, 2016) Each feature improves the productivity of the developer and aids in the readability of the syntax.

4.3 BOOTSTRAP

Bootstrap is a free front-end Hypertext Markup Language (HTML), JavaScript, and Cascading Stylesheet (CSS) framework. It's used to implement quick, easy, responsive solutions for web development. It efficiently scales to the device being used meaning, responsiveness is made easier and something the developer doesn't have to worry about. It divides the webpage up into twelve columns and unlimited rows. (Bootstrap, 2016)

4.4 ANGULARJS

AngularJS the JavaScript Framework. It's the second most popular technology in 2016. (Stack Overflow, 2016) It's created and maintained by Google and has great support in the community. It implements the Model View View-Model (MVVM) pattern which creates a division between the user interface (View) and their logic (Model). It does this by binding elements in the view and allows the model access it. This process is called Data Binding. It enhances the development process by allowing the application generate elements on the view and minimizes the code that needs to be written. (Conrad, 2012) Figure 2, shows the interaction between the different layers of the MVVM pattern.
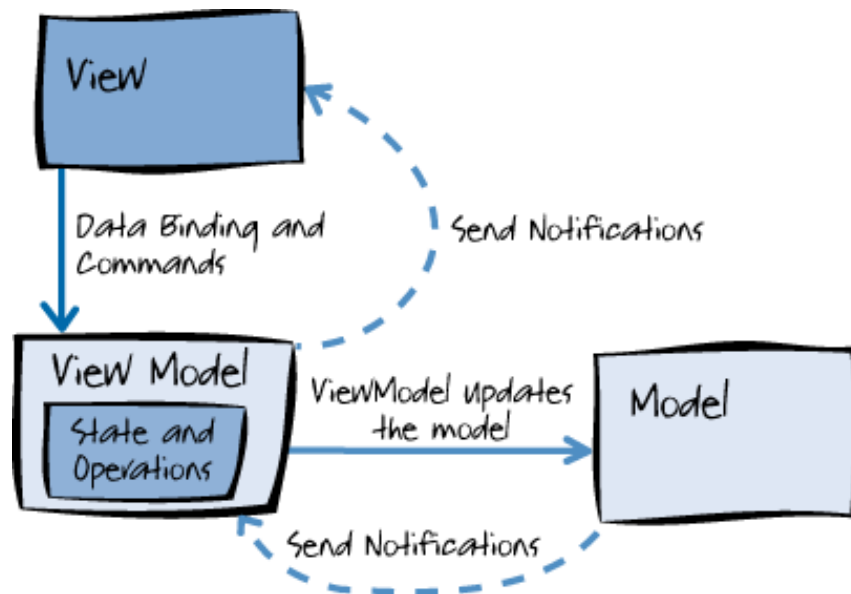
Figure 2. The Model View View-Model (Microsoft, 2016)

AngularJS has many components, however, one of the key components are Directives. This makes websites modular, which maximizes the reusability of code. It can be implemented through the HTML tags or as an attribute on a tag. It creates new HTML syntax and gives behaviours to that element. (Panda, 2014)

Controllers in an AngularJS application provide the behaviours of the application. This would be the Model in the MVVM pattern. It allows the application manipulate data to be

displayed on the webpage. Services provide a way the application can make API calls which retrieves data for the controller. (AngularJS, 2016a)

4.5 D3

D3 is a JavaScript library which makes full use of modern browsers to render different types of graphs. It stands for Data Driven Documents. It uses HTML, SVG and CSS to create the elements on the webpage. (D3, 2016) Each graph is created from the beginning, one element at a time which allows full customizability for the developer. This enhances the reusability of the library.

4.6 LODASH

Lodash is a JavaScript library which aids in the readability of code. It makes it easier to work with objects, strings, arrays and numbers. It eliminates the complexities from extracting of these tasks. (Lodash, 2017)

4.7 FLASK

Python has a micro framework called Flask which is used to develop websites. It uses Jinja as a templating language which allows Flask to have a very modular design which enables the developer to create webpages in parts and assemble the page on execution rather than have full HTML files that contains the entire structure for the webpage. Jinja works by extending other HTML documents which allows the developer to modularise the design. (Jinja, 2016)

The developer has created several applications in Flask and has experience with technologies like Jinja. While Jinja is a very powerful tool, it loses the readability aspect and the structure of the HTML compared to the previous technologies mentioned.

4.8 IONIC FRAMWORK

Ionic is a framework used to create high quality hybrid phone applications (apps). It allows users to develop apps using the web technologies. Ionic is based on the JavaScript framework AngularJS. It's the most popular cross platform mobile development technology. (Ionic, 2017a) As it's based on AngularJS, it will contain all the benefits mentioned above in the AngularJS section and due to that framework being very modular, it allows people to easily share the different modules that they create. Ionic provides a market place where developers can sell or share for free different resources for the applications. (Ionic, 2017b) This has the potential to speed up the development time substantially.

5. BACKEND TECHNOLOGIES

5.1 PYTHON

Python is a high-level language which focuses on readability of the code rather than the execution time. It's an open source technology and the community is well established. It has many inbuilt, as well as downloadable modules which can be used accessed via the PIP command in the command line. These tools are created to help speed up development time as well as develop different applications. Each of the modules are segments of code that have a specific purpose that was found that developers needed to repeat several times in different applications. Therefore, it was decided to create modules that the developer could import into the current application that they are working on without having to rewrite the code, instead they could call a function to do all the work.

5.1.1 TWEEPY

Tweepy is a module which creates a Python container that's used to access the Twitter API which is used to download tweets. They're two main types of API's for Twitter, the search API and the trends by place API. Tweepy creates an easy to use way of using the OAuth authentication and the AppAuthHandler feature to enable the max number of tweets to be downloaded. For example, for the search API, the number of request made to the Twitter server without the authentication is 180 requests whereas enabling the authentication will allow a max of 450 requests per fifteen minutes. (Twitter, 2016b) This enables a larger database to analyse the sentiment of and therefore giving a better understanding of the term being searched by the user.

This application will be using AppAuthHandler instead of OAuth authentication because it will sleep when it reaches the request limit for the exact time remaining till the next request can be made. This maximizes the number of tweet retrieved from Twitter.

5.1.2 NATURAL LANGUAGE TOOLKIT

A popular text analysis module in Python that will be necessary in this application is the Natural Language Toolkit also known as NLTK. This is an open source project which is community driven. It contains several text processing libraries for classification, tokenization, stemming and tagging. These libraries are used to parse and understand the meaning of language. NLTK is a fundamental module when aiming to mine text in a language.

The libraries used for the text analysis that will be considered for this project will be tokenization, stemming and tagging. Tokenization is used to split a string into substrings so they can be processed further. NLTK provides a Twitter specific tokenizer `nltk.tokenize.casual.TweetTokenizer`, this will split usernames as well as hashtags and emoticons which are popular in microblogging sites like Twitter. (NLTK, 2017a) Stemming is a technique used to find the base version of a word, e.g. processing will be shortened to process. (NLTK, 2015b) Tagging is a method used for dividing words into word classes. Once the sentence has been tokenized, the newly separated words can be tagged. This process produces a list of tuples with the word and the associated word class e.g. Noun, Verb or Adjective. This can be a very effective way when trying to find the sentiment of a piece of text. (NLTK, 2015)

There was an issue getting this module set up because all the libraries within it needing to be downloaded separately. This would normally work by running "`nltk.download()`" in terminal. However, one of the libraries would always timeout. This was because it was offline, so to get this to work; it needed to be downloaded manually and placed into the correct directory. This was done by:

- Clicking on the following link: http://dev.panlex.org/db/panlex_lite.zip
- After the download is completed, depending on your operating system, you'll need to place the file into one of the following directories:
    - `C:\nltk_data\corpora\`
    - `C:\Users\{{username}}\AppData\Roaming\nltk_data`
    - `/usr/share/nltk_data`
    - `/usr/local/share/nltk_data`

- o `/usr/lib/nltk_data`
- o `/usr/local/lib/nltk_data`
- o `~/nltk_data`

5.1.3 SCIKIT-LEARN

The English language has evolved over centuries of use and as the way people speak has changed so has the way people write has changed and both will continue to change, so to decide on the sentiment of a piece of text will require a classifier. Classifiers are machine learning algorithms which are a branch off Artificial Intelligence and will be used to determine the sentiment of a tweet. A classifier is a function used to map inputs vectors to an output class label. (Murphy, 2006) In this case, the input vectors will be the text from the tweet and the classifier will be used to map it to a positive or negative output class.

Classifiers are supervised learning algorithms, this means they all need to be trained. Each algorithm will be provided with a dataset of tweets which will have a sentiment associated with it. In machine learning terms, the sentiment associated with each tweet is known as a label and the tweet is known as the input. The input will have to be separated into features where in this case will be the individual words in the tweet. They will be implemented with the bag of words model which will contain the words that were extracted from the input during the training process. The bag of words technique is the widely used in text classification for its simplicity and good performance. (Baqapuri, 2012) This will be known as the feature vector and will aid in the classification of the sentiment. The classifiers will use the feature vector and the label to find trends within the data and will be able to use it to classify the new data. This process is illustrated in figure 3.
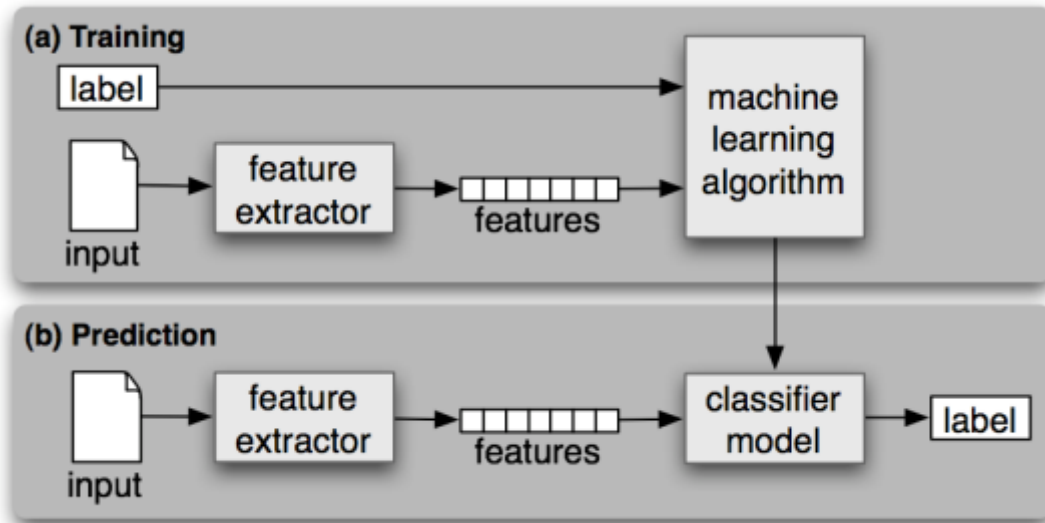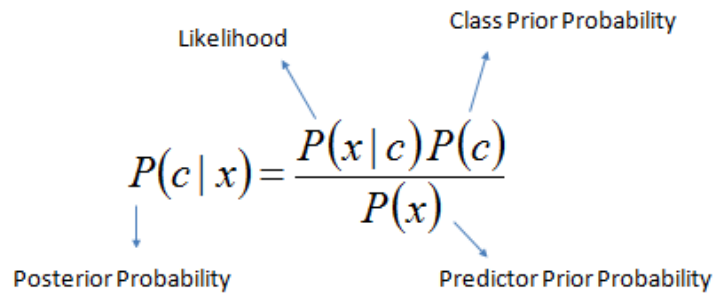
Figure 3. A visual representation of how machine learning works (Steven Bird, 2015)

They're many types of classifiers and different versions as well. This application will integrate three different classifiers in total, one Naïve Bayes, one Linear Model and one Support Vector Machine algorithm. These algorithms will be joined together by a voting algorithm which will take input from all the classifiers and find the majority vote and that will be chosen classification of the text.

Naïve Bayes is a popular choice for text classification. It's a probabilistic supervised learning algorithm which applies Bayes' theorem with the assumption of independence. (Sayad, 2017) This means that even if features have a dependency on each other, Naïve Bayes won't take that into account when classifying it. For example, a fruit is said to be an apple if the its round, red and three inches in diameter. Even if these features depend on each other to be an apple, this algorithm will view each feature as independent. When it comes to text classification, it receives an input, it will check the if features are in the bag of words, if it is, the algorithm will have a sentiment associated with it. Once all the words have a sentiment, the algorithm will then find the probability of the classification of the sentence. (Rish, 2001) This application will integrate two different versions of this algorithm.

$$P(c \mid x) = \frac{P(x \mid c) P(c)}{P(x)}$$

Likelihood

Class Prior Probability

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 4. Naïve Bayes algorithm (Sayad, 2017)

Multinomial Naïve Bayes algorithm is like Bernoulli Naïve Bayes algorithm because it also uses a binary feature vector which is used to show whether a word appears in the bag of word model. However, Multinomial expands on Bernoulli's by adding the number of times the word occurs in the sentence. (Scikit-learn, 2017) In the case of this application, if the word appears in the bag of words model a sentiment will be associated with it. That sentiment will be recorded and once every word in the sentence which is found in the bag of words model. The probability of all the words sentiments and the number of occurrences are calculated to find the final sentiment.

Linear Models are also a set of supervised learning algorithms. When these algorithms are trained with data that have two or more categories, it finds the best fit line for that data. The term Linear refers to the using a line and model is the equation that summarises the fitted line. (Trochim, 2006)

The Linear Model algorithm being used in this project is Logistic Regression. This is used to test the relationship between two or more variables. Once this algorithm is trained, it will create a best fit line for the training data. This line can be used to predict the classification on unseen data. Logistic Regression is a probabilistic regression algorithm, this means that it uses the best fit line method as well as probability to classify data. (Joglekar, 2015)

Menarche Data with Fitted Logistic Regression Line
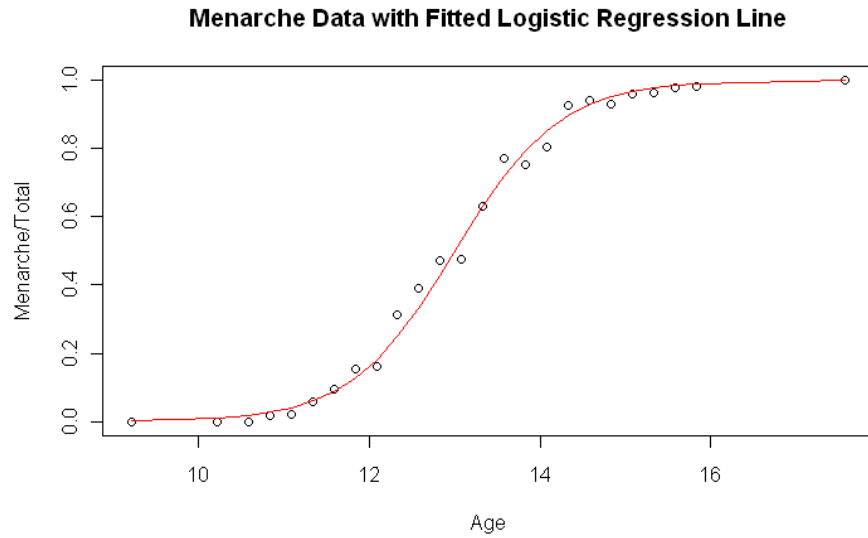
Figure 4. Logistic Regression example (King, 2016)

Support Vector Machine (SVM) will be the final algorithm. It works by plotting points onto a graph and separating the categories by a line/hyperplane depending on the number of categories. The line works by providing the maximum margin between the categories. (OpenCV, 2014) Figure 5 provides a visual representation of this method.
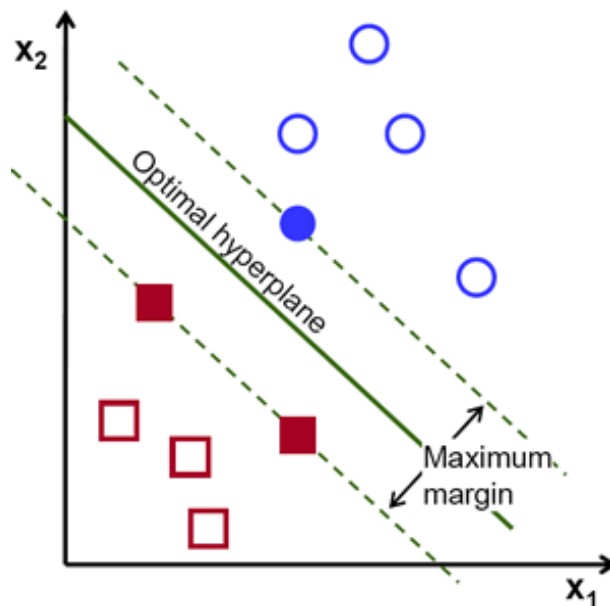
Figure 5. Support Vector Machine example (OpenCV, 2014)

5.1.4 PYTRENDS

This is a Python module that is used for retrieving the top 10 most search terms on Google. It's an unofficial API which means that if Google changes their backend, there is a chance of this API not working. It's not maintained by Google; however, it's an open sourced project so the community will be maintaining the project. This is used to retrieve the top searched terms in the United States of America and the United Kingdom. (Libraries, 2016)

6. DATA STORAGE

When it comes to data storage, there's two main types that are used almost exclusively in the development of applications, relational and non-relational databases. Both are useful for different reasons. Relational databases are structured with several tables which are used to keep track of the data whereas Non-Relational databases uses a document type storage method.

6.1 RELATIONAL DATABASE

Structured Query Language (SQL) is the most common technology used for relational databases. The database is structured with a column name as the table header and each row is a unique instance of the data in the table. (Rouse, 2006) This creates a tabular structure for the data to be stored in. Specific entries in the SQL databases can be accessed through a primary key which is a unique key which is provided to each individual entry in a table. There is also a foreign key with is defined in a second table which represents a primary key in another table. (Buckler, 2009b)

In Figure 4 there are two tables and each table has separate headers with several entries. In the employee table, the first column is name "id", this is used as the primary key for this table. Also, in the borrowed table, the first column is called "ref" this will be the primary key in this table. The second column in the borrowed table is called "employeeid" this will act as the foreign key in this table which allows the database keep track of which employee has borrowed what book.

employee:

id	firstname	lastname	birthdate
1	John	Smith	1976-01-02
2	Laura	Jones	1969-09-05
3	Jane	Green	1967-07-15

borrowed:

ref	employeeid	book
1	1	SitePoint Simply SQL
2	1	SitePoint Ultimate HTML Reference
3	1	SitePoint Ultimate CSS Reference
4	2	SitePoint Art and Science of JavaScript

**Figure 4.** An example of a Relational Database (Buckler, 2009b)

6.1.1 MYSQL

MySQL is a relational database management system (RDBMS) that is maintained by Oracle. It's an open source project which allows the community help with development of it. Due to its popularity, there are many tools out there that implements graphical user interfaces (GUI) versions to allow users to manipulate the data in a table form instead of command line. (Digital Ocean, 2014) It provides many key security features such as, when creating users, that user can have permissions to access certain functions within the table. This could include selecting entries, updating entries or inserting a new entry into the database. This solves many issues that could occur from unauthorized users from having access to delete the entire database for example. (MySQL, 2016)

MySQL scales vertically. This means to increase the performance of the server more RAM and CPU will be required which leads to added cost and redundancy as the server will need to be able to scale with demand. (Shalom, 2010)

6.2 NON-RELATIONAL DATABASE

Applications are now being developed using the Agile approach to software development. This means that projects adapt to changes. However, for the application to change the database will need to change alongside it. While relational databases are great for very structured data, they aren't very flexible and therefore scaling with the application leads to issues. For this reason, Non-relational database (NoSQL) are very flexible. NoSQL databases are used because they allow for rapidly changing structures and non-tabular data. They're also highly scalable with low latency and high performance as they scale horizontally. (Krisciunas, 2014) This means it scales by adding more machines rather than more CPU or RAM. This leads to better distribution of the data. (Shalom, 2010) Figure 4 expresses the difference between scaling horizontally and vertically.
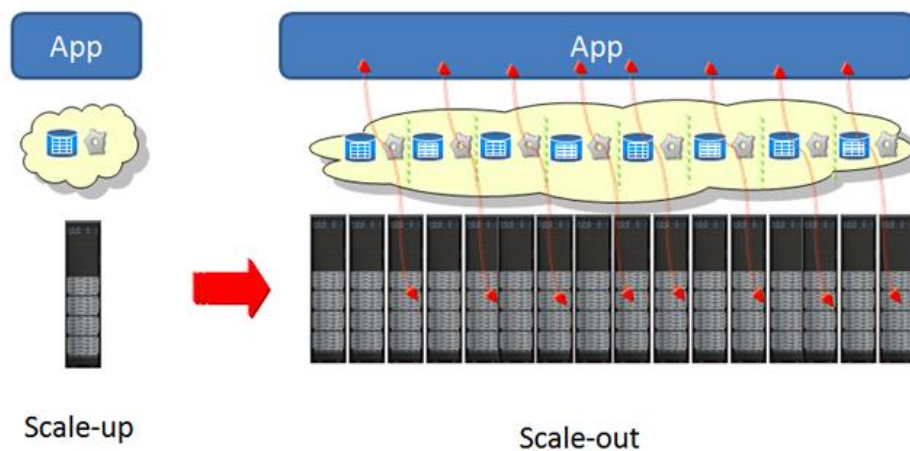
Figure 4. Difference between vertically scaling (scale-up) and horizontally scaling (scale-out)(Shalom, 2010)

There are many types of non-relational databases each of which has their own benefits for different projects.

- **Document Stores:** This is a semi-structured database which adheres to some of the rules of relational databases. However, it isn't strict on the structure of the database. It uses the JSON format to represent data and allows the developer to query by the properties of each object.

- **Graph Databases:** This data storage mechanism allows objects to be linked together forming a graph like structure. It allows the link to be named so the association is more readable. Each object can be linked to many other objects.

6.2.1 MONGODB

MongoDB is the most popular NoSQL database. It's a document storage database which stores its information in BSON which is based on JSON and stands for "Binary JSON". It's also an open source project which enables the community to commit changes to it. (Dayley, 2015) MongoDB uses a dynamic schema to store data which allows developers to add fields and delete fields when needed. It allows the storage of data structures like arrays and JSON objects which isn't available in MySQL. (MongoDB, 2017) MongoDB is used when dealing with JSON objects as it the document style storage enables it with ease.

7. TESTING

While creating an application, it's key to create tests for the features the developer completes as they create it. This is because the developer will be aware of how that element works because they just created it. Also, creating the tests, both functional and unit tests as the actual programming of the feature is completed allows the developers to ensure that, that feature is safe from future edits of the code. If another developer changes a piece of code that feature relies on, the tests will fail.

7.1 PROTRACTOR

This is a front-end testing framework for AngularJS. It is used to create both, functional and unit tests. Protractor maximizes the reusability of code with the use of Page Objects which are used to specify a single or group of elements within the HTML structure that can then be used to test the functionality of that object. It also makes use of mock data, which can be used in functional tests to check the accuracy of the algorithms used. (Protractor, 2016)

7.2 KARMA

Karma is a JavaScript tool used in the command line to start a web server which the application can run on. While the server is started, it runs unit tests on the code for the application. It enables the tests to run on different browsers to ensure maximum supportability. (AngularJS, 2017b)

7.3 JASMINE

Jasmine is the most popular tool for unit tests on AngularJS. It's a behaviour driven development framework for JavaScript. Jasmine has several features that allow the developer keep tests in a well-structured state. The 'describe' function can label a group

of tests to keep similar tests in a group. The '`it`' function is used to label a test for easy recognition if a failure occurs. (AngularJS, 2017b)

7.4 UNITTEST

Unittest is a standard unit test module provided by Python. It has many tools that enables testing Python code easier. An example of these tools are start-up and tear-down code for the developer to clean-up after the tests. Tests must begin with the prefix '`test_`' which sets the function as a test for the '`unittest.main()`' function to run it. (Python, 2017)

8. HOSTING

There's a number of possible hosting options for this project which will be researched here.

8.1 SELF-HOSTING

There's many reasons why self-hosting would be the best option for a project like this. The first one being that there's no technology limitations. This means the developer has complete control over the technologies used when choosing them without having to consider the cloud limitations. This would allow for a more varied choice of technologies as the cloud server won't be a factor in choosing them. Also, the ability to customise anything on the server like security settings will allow the developer to have complete access to the way the server behaves. (Court, 2017)

8.2 HEROKU

Heroku is a cloud platform which accepts all popular technologies like Python, Java and Ruby. It runs the application inside smart containers called Dynos. (Heroku, 2017a) These Dynos are lightweight Linux containers that runs a single users command. This command is normally how to run the application that is hosted on it. (Heroku, 2017b) Heroku has a free account, however, if it doesn't receive any traffic for 30 minutes, then the dyno will sleep until its needed again. (Heroku, 2017c)

9. PACKAGE MANAGER

Applications are made up of many different packages which are used to create the final product. These packages are also known as dependencies which enables the developer to speed up the development time of the application and allows them create quality applications because all the packages are features developed separately any project. This enables the feature to be developed and tested extensive and as most of these packages are open source it allows the developer to add more features to it. This allows the package to become more extensive and creates more time for the developer of the application to focus on the current project.

9.1 NODE

NPM is a package manager which is provided by Node, and allows you keep track, install and update your dependencies. It makes code more reusable as it installs the exact version of the dependencies. This avoids issues such as if the creator of the package that your application depends on removes the piece of code that you require or changes it and renders your application unusable. (NPM, 2016)

NPM will be used to install in this application to install all the required packages that are used in Node. The packages that it will be installing will be Gulp and AngularJS.

9.2 BOWER

Bower is a tool like Node which also keeps track of dependencies, however it is used to keep track of the frontend dependencies. It was created by Twitter in 2012. (Bower, 2016) Every website is made of different components and these components may have already been developed in part by other developers. The packages that Bower installs are small segments of code which has been created with a specific purpose. Each of these packages are the elements that can be assembled to help a developer develop the desired design they are looking for. (Bluescreen, 2014)

9.3 PIP

PIP is the preferred installer program for Python modules. It's included with the Python download by default. It allows modules to be downloaded singly, however if the module name and the version is included in a text file, it can be easily installed using the PIP command in terminal. This file has a standard name '`requirements.txt`'. Once the requirements are in the file using the command '`pip install -r requirements.txt`' will install all the contained modules. (Pypa, 2016)

10. VERSION CONTROL

As a project increases in size, having back up of the different versions become a vital part of the development process. This is important because if someone uploads a segment of code which renders another piece of code to not execute correctly, it allows the administrator to revert to a previous version that executes correctly. This becomes invaluable if several developers are working on a project which creates an entire set of new problems.

10.1 GIT AND GITHUB

Git is a distributed version control systems (DVCS). This means a developer can use Git to use a locally stored version control system on their own system as well as a central repository on the internet. It allows the users to add a message along with each change made, this is called a commit. Git keeps a track of these commits and enables the user to revert to time before the error had been made. Git allows the user to create different 'branch's' of the codebase which are completely independent of each other. These are created locally and allows the developer to work on different tasks for the project without messing up another branch. (Git, 2016)

Git is mainly a command-line tool which, for beginners, is difficult to learn so there are several alternatives. One of which is called SourceTree. This is a graphical user interface version (GUI) version of Git. This allows the developer to see what is way Git works and contains useful graphs which shows the different branch's and how they interact with the project's main branch. (SourceTree, 2016) However, due to the authors experience with Git, the command-line tool will be used for this project.

GitHub is a well-established place for developers to store their projects. When a developer creates a project, it is called 'creating a repository'. This repository can be public or private depending on the developer's preference. The developer can than create different branch's and upload his code to this repository. Once the developer has done so, it needs to be approved before being merged into the rest of the codebase. This is done by creating

a Pull Request. When there's several developers working on the same project it is essential that the code is reviewed and tested to ensure the highest quality. This is also helpful as GitHub also tracks who edited what line of code last and which enables a developer contact the one who wrote the code if there are any questions about it.

10.2 APACHE SUBVERSION

Apache subversion (SVN) is an open sourced version control system founded in 2000 by CallabNet. In a lot of ways SVN acts the same way as Git. It allows developers to work off the same repository without conflicting and overwriting code. It allows rollback if an error occurs and the project needs to go back to a previous commit. However, apart from the syntax difference, SVN works off a repository that's online. SVN doesn't adopt distributed source control so it doesn't allow any local version control system to be created. (Deveo, 2017)

11. CONTINUOUS INTEGRATION

When developing large scaled projects, it is difficult to remember everything and becomes increasingly easier to render another feature unusable. To solve this, people write tests for each feature they develop, so this keeps track of the functionality of that feature and will fail the tests if the new code breaks it. This becomes vital when you have more than one developer working on the piece of code. Although, it is possible to upload the code to the version control system even if the developer hasn't run the tests. Continuous Integration runs all the tests provided for the codebase whenever a Pull Request is created. This eliminates untested code from entering the codebase and improves the quality of the application.

11.1 JENKINS

Jenkins is a free Continuous Integration and Continuous Delivery tool. The user must host Jenkins on their own server which adds to the complexity of the setup of this tool. As well as that, the user is required to configure it, which means give it jobs to do. These jobs are a list of tasks that it must complete before the job can pass. As you can see from figure 5, once the developer commits their changes and opens a Pull Request, this will trigger Jenkins to start a job that the user has specified. In this case, it will run tests to ensure that the changes uploaded haven't rendered any other features to not work. As a project grows or when you have several developers working on a project this kind of tool is vital to the quality of the application.
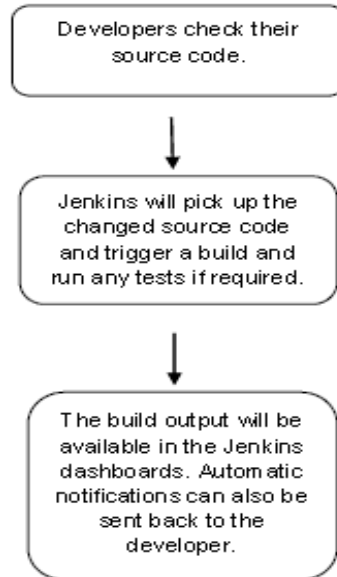
Figure 5. Shows the flow chart that demonstrates how Jenkins work (Tutorials Point, 2016)

11.2 TRAVIS CI

Travis CI is a tool to test and deploy code. This tool is free for open source projects, however, costs $129 per month for each account its linked to. This means if the user pays for this service, its available throughout all the projects on that account. Travis CI is also maintenance free as they host it themselves. (Waller, 2014) It's very simple to set up, all that's required is yml file in the root of the project and a pull request to be created. GitHub will trigger Travis CI and it will run the tasks set out in the configuration file. (Travis CI, 2017)

12. SIMILAR APPLICATION

As part of the research for this project, I had to research similar applications to this project.

12.1 TWITTER

Twitter provides a version of sentiment analysis. Its positioned in the advanced search option under the 'Other' heading. This heading allows the user select Positive, Negative and Questions to associate with the search term. However, all this feature does is add a smiley text based emoji, a sad face text based emoji and a question mark after the search term. This isn't a very accurate way of determining the sentiment for a chosen topic as it doesn't always filter the tweets properly. It also doesn't show any breakdown or try to determine the total sentiment of the topic. (Twitter, 2017d)

12.2 TWEET SENTIMENT VISUALIZATION

One of the more popular tools is called Tweet Sentiment Visualization. This tool, provides an input box which allows the user to search anything they would like to find sentiment analysis on. Once the user clicks 'Query' the application searches for tweets about the topics inputted. They use this by using the Twitter Search API. Once all the tweets are processed, it displays it on several different graphs. An example of one of these graphs is the sentiment graph, where it shows the sentiment on several different emotions such as happy, calm and nervous. (Ramaswamy, 2016)

12.3 HOOTSUITE

Many of sentiment analyses tools are paid tools. An example of this is Hootsuite which allows the user request a demo of the software. This tool is mainly advertised towards businesses and allows real-time access to tweets which is presumed to take advantage of the Streaming API that Twitter offers. (Hootsuite, 2016)

13. CONCLUSION

Microblogging sites have been identified as online word-of-mouth branding. (Jansen, 2009). For this reason, a tool that can examine the sentiment of publics word-of-mouth opinions is becoming more and more in demand.

The technologies for the font-end will be:

- AngularJS because the author is proficient in this technology and it maximizes the usability of the code with the use of features like Directives.
- D3 will be used to render the graphs needed to output the content.
- Lodash is chosen due to the enhancement for the readability of code.
- Sass will be used to compile the CSS as it maximizes the usability of code as well as providing useful features like variables.
- Bootstrap will also be used to maximize the responsiveness for the webpage and to aid in the general styling.
- The Ionic framework will be used to create the phone application.

The technologies for the back-end will be:

- Python will be the main technology as it maximizes readability and includes the key modules that are vital for this project. Such as Tweepy and NLTK which will prove to be a crucial component in the creation of this application.
- Travis CI will be used to ensure maximum quality and to establish no new components broke any previous feature.

This project was decided to use both the self-hosting and Heroku. The college server will be used to download and analyse the data for the website whereas Heroku will be implemented to demonstrate the phone application.

14. GLOSSARY

Term	Explanation
Classification	The category that the classifier categorises the text into.
Classifier	A machine learning algorithm used to categorise text into groups.
Classify	The method of deciding on the classification
Model	A trained classifier
Platform	A generic way of describing the source of the information e.g. Twitter and Facebook
Post	A generic way of describing the downloaded content from a source e.g. tweets from Twitter and statuses from Facebook
Term	A word or phrase the user wants to find the sentiment for.
Tweet	This is a post from a user on the Twitter social media platform

15. BIBLIOGRAPHY

AngularJS, 2016a. *AngularJS.* [Online]

Available at: https://angularjs.org/

[Accessed 27 November 2016].

AngularJS, 2017b. *Unit Testing.* [Online]

Available at: https://docs.angularjs.org/guide/unit-testing

[Accessed 23 March 2017].

Baqapuri, A. I., 2012. *Twitter Sentiment Analysis.* [Online]

Available at: https://arxiv.org/pdf/1509.04219.pdf

[Accessed 22 March 2017].

Bluescreen, 2014. *Bower vs. npm.* [Online]

Available at: http://ng-learn.org/2013/11/Bower-vs-npm/

[Accessed 22 October 2016].

Bootstrap, 2016. *Bootstrap.* [Online]

Available at: http://getbootstrap.com/

[Accessed 19 October 2016].

Bower, 2016. *About.* [Online]

Available at: https://bower.io/docs/about/

[Accessed 18 October 2016].

Buckler, C., 2009b. *How to Use MySQL Foreign Keys for Quicker Database Development.*
[Online]

Available at: https://www.sitepoint.com/mysql-foreign-keys-quicker-database-development/

[Accessed 5 November 2016].

Buckler, C., 2015a. *SQL vs NoSQL: The Differences.* [Online]

Available at: https://www.sitepoint.com/sql-vs-nosql-differences/

[Accessed 24 October 2016].

Conrad, A., 2012. *3 Reasons to Choose AngularJS for Your Next Project.* [Online]

Available at: https://code.tutsplus.com/tutorials/3-reasons-to-choose-angularjs-for-your-next-project--net-28457

[Accessed 30 November 2016].

Court, T., 2017. *Cloud Vs. Self Hosted – Which Digital Asset Management Solution Is Best For You.* [Online]
Available at: http://www.intelligencebank.com/uk/blog/cloud-vs-self-hosted-which-digital-asset-management-solution-best-you
[Accessed 2 April 2017].

D3, 2016. *Data Driven Documents.* [Online]
Available at: https://d3js.org/
[Accessed 25 October 2016].

Daily Mail, 2012. *Is private the new public? Study reveals Facebook users are defriending in greater numbers than ever.* [Online]
Available at: http://www.dailymail.co.uk/news/article-2106146/Is-private-new-public-Study-reveals-Facebook-users-defriending-greater-numbers-ever.html
[Accessed October 2016].

Dayley, B., 2015. *NoSQL with MongoDB in 24 Hours, Sams Teach Yourself.* 1st Edition ed. Indiana(Indianapolis): SAMS.

Deveo, 2017. *Which version control system should you choose?.* [Online]
Available at: https://deveo.com/git-vs-svn/
[Accessed 23 March 2017].

Digital Ocean, 2014. *SQLite vs MySQL vs PostgreSQL: A Comparison Of Relational Database Management Systems.* [Online]
Available at: https://www.digitalocean.com/community/tutorials/sqlite-vs-mysql-vs-postgresql-a-comparison-of-relational-database-management-systems
[Accessed 1 December 2016].

Donkor, B., 2013. *On Social Sentiment and Sentiment Analysis.* [Online]
Available at: http://brnrd.me/social-sentiment-sentiment-analysis/
[Accessed 22 October 2016].

Facebook, 2016a. *Rate Limiting on the Graph API.* [Online]
Available at: https://developers.facebook.com/docs/graph-api/advanced/rate-limiting
[Accessed 17 October 2016].

Facebook, 2016b. *Using the Graph API.* [Online]

Available at: https://developers.facebook.com/docs/graph-api/using-graph-api

[Accessed 16 October 2016].

Git, 2016. *About.* [Online]

Available at: https://git-scm.com/about/branching-and-merging

[Accessed 31 October 2016].

Heroku, 2017a. *Deploy and run apps on today's most innovative Platform as a Service.* [Online]

Available at: https://www.heroku.com/platform

[Accessed 2 April 2017].

Heroku, 2017b. *Dynos and the Dyno Manager.* [Online]

Available at: https://devcenter.heroku.com/articles/dynos

[Accessed 3 April 2017].

Heroku, 2017c. *Free Dyno Hours.* [Online]

Available at: https://devcenter.heroku.com/articles/free-dyno-hours

[Accessed 3 April 2017].

Hootsuite, 2016. *Overview.* [Online]

Available at: https://hootsuite.com/products/insights

[Accessed 26 October 2016].

Internet Live Stats, 2016. *Twitter Usage Statistics.* [Online]

Available at: http://www.internetlivestats.com/twitter-statistics/

[Accessed 23 October 2016].

Ionic, 2017a. *About us.* [Online]

Available at: http://ionic.io/about

[Accessed 26 March 2017].

Ionic, 2017b. *Discover & share ionic plugins for your app.* [Online]

Available at: http://market.ionic.io/

[Accessed 1 April 2017].

Jansen, B. J., 2009. *Micro-blogging as Online Word of Mouth Branding.* [Online]

Available at:

https://faculty.ist.psu.edu/jjansen/academic/pubs/jansen_Micro_blogging_as_Online_Word_of_

Mouth_Branding.pdf

[Accessed 22 March 2017].

Jinja, 2016. *Welcome.* [Online]
Available at: http://jinja.pocoo.org/
[Accessed 23 October 2016].

Joglekar, S., 2015. *Logistic Regression (for dummies).* [Online]
Available at: https://codesachin.wordpress.com/2015/08/16/logistic-regression-for-dummies/
[Accessed 25 March 2017].

King, W. B., 2016. *LOGISTIC REGRESSION.* [Online]
Available at: https://ww2.coastal.edu/kingw/statistics/R-tutorials/logistic.html
[Accessed 1 April 2017].

Krisciunas, A., 2014. [Online]
Available at: https://www.devbridge.com/articles/benefits-of-nosql/
[Accessed 5 November 2016].

Libraries, 2016. *GeneralMills/pytrends.* [Online]
Available at: https://libraries.io/github/GeneralMills/pytrends
[Accessed 23 March 2017].

Lodash, 2017. *Lodash.* [Online]
Available at: https://lodash.com
[Accessed 20 March 2017].

Magid, L., 2014. *Facebook Changes New User Default Privacy Setting To Friends Only -- Adds Privacy Checkup.* [Online]
Available at: https://www.forbes.com/sites/larrymagid/2014/05/22/facebook-changes-default-privacy-setting-for-new-users/#7df4c06659ac
[Accessed 20 March 2017].

Maina, A., 2016. *20 Popular Social Media Sites Right Now.* [Online]
Available at: https://smallbiztrends.com/2016/05/popular-social-media-sites.html
[Accessed 12 October 2016].

Microsoft, 2016. *The MVVM Pattern.* [Online]
Available at: https://msdn.microsoft.com/en-ie/library/hh848246.aspx
[Accessed 30 November 2016].

MongoDB, 2017. *MongoDB and MySQL Compared.* [Online]
Available at: https://www.mongodb.com/compare/mongodb-mysql
[Accessed 23 March 2017].

Moore, R. J., 2009. *Twitter Data Analysis: An Investor's Perspective.* [Online]
Available at: https://techcrunch.com/2009/10/05/twitter-data-analysis-an-investors-perspective-2/
[Accessed 16 October 2016].

Murphy, K. P., 2006. *Naive Bayes classifiers.* [Online]
Available at: http://www.ic.unicamp.br/~rocha/teaching/2011s1/mc906/aulas/naive-bayes.pdf
[Accessed 22 March 2017].

MySQL, 2016. *The MySQL Access Privilege System.* [Online]
Available at: https://dev.mysql.com/doc/refman/5.5/en/privilege-system.html
[Accessed 1 December 2016].

NLTK, 2015b. *Processing Raw Text.* [Online]
Available at: http://www.nltk.org/book/ch03.html
[Accessed 20 March 2017].

NLTK, 2015. *Categorizing and Tagging Words.* [Online]
Available at: http://www.nltk.org/book/ch05.html
[Accessed 20 March 2017].

NLTK, 2015. *nltk.classify package.* [Online]
Available at: http://www.nltk.org/api/nltk.classify.html
[Accessed 4 April 2017].

NLTK, 2017a. *nltk.tokenize package.* [Online]
Available at: http://www.nltk.org/api/nltk.tokenize.html
[Accessed 20 March 2017].

NPM, 2016. *What is npm?.* [Online]
Available at: https://docs.npmjs.com/getting-started/what-is-npm
[Accessed 22 October 2016].

OpenCV, 2014. *Introduction to Support Vector Machines.* [Online]
Available at:
http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html
[Accessed 26 March 2017].

Panda, S., 2014. *A Practical Guide to AngularJS Directives.* [Online]
Available at: https://www.sitepoint.com/practical-guide-angularjs-directives/
[Accessed 30 November 2016].

Protractor, 2016. *Protractor.* [Online]
Available at: http://www.protractortest.org/#/
[Accessed 23 October 2016].

Pypa, 2016. *User Guide.* [Online]
Available at: https://pip.pypa.io/en/stable/user_guide/
[Accessed 23 March 2017].

Python, 2017. *Unit testing framework.* [Online]
Available at: https://docs.python.org/2/library/unittest.html
[Accessed 23 March 2017].

Ramaswamy, H. &., 2016. *Visualizing Twitter Sentiment.* [Online]
Available at: https://www.csc.ncsu.edu/faculty/healey/tweet_viz/
[Accessed 26 October 2016].

Rish, I., 2001. *An empirical study of the naive Bayes classifier.* [Online]
Available at:
https://www.researchgate.net/profile/Irina_Rish/publication/228845263_An_Empirical_Study_of
_the_Naive_Bayes_Classifier/links/00b7d52dc3ccd8d692000000/An-Empirical-Study-of-the-
Naive-Bayes-Classifier.pdf
[Accessed 23 March 2017].

Rouse, M., 2006. *Relational Database.* [Online]
Available at: http://searchsqlserver.techtarget.com/definition/relational-database
[Accessed 2 November 2016].

Sass, 2016. *Sass Basics.* [Online]
Available at: http://sass-lang.com/guide
[Accessed 18 October 2016].

Sayad, D. S., 2017. *Naive Bayesian.* [Online]
Available at: http://www.saedsayad.com/naive_bayesian.htm
[Accessed 25 March 2017].

Scikit-learn, 2017. *sklearn.feature_extraction.text.CountVectorizer.* [Online]

Available at: http://scikit-

learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html#sklearn.

feature_extraction.text.CountVectorizer

[Accessed 4 April 2017].

Scikit-learn, 2017. *sklearn.naive_bayes.BernoulliNB.* [Online]

Available at: http://scikit-

learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html

[Accessed 24 March 2017].

Shalom, N., 2010. [Online]

Available at: http://ht.ly/cAhPe

[Accessed 23 March 2017].

SourceTree, 2016. *SourceTree.* [Online]

Available at: https://www.sourcetreeapp.com/

[Accessed 30 October 2016].

Stack Overflow, 2016. *Developer Survey Results 2016.* [Online]

Available at: http://stackoverflow.com/research/developer-survey-2016

[Accessed 17 October 2016].

Steven Bird, E. K. a. E. L., 2015. *Learning to Classify Text.* [Online]

Available at: http://www.nltk.org/book/ch06.html

[Accessed 3 December 2016].

ThinkNook, 2012. [Online]

Available at: http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-

22/

[Accessed 3 April 2017].

Travis CI, 2017. *Test and Deploy with Confidence.* [Online]

Available at: https://travis-ci.org/

[Accessed 23 March 2017].

Trochim, W. M., 2006. *General Linear Model.* [Online]

Available at: https://www.socialresearchmethods.net/kb/genlin.php

[Accessed 25 March 2017].

Tutorials Point, 2016. *Jenkins - Overview.* [Online]
Available at: https://www.tutorialspoint.com/jenkins/jenkins_overview.htm
[Accessed 25 October 2016].

Twitter, 2016a. *The Search API.* [Online]
Available at: https://dev.twitter.com/rest/public/search
[Accessed 17 October 2016].

Twitter, 2016b. *GET trends/place.* [Online]
Available at: https://dev.twitter.com/rest/reference/get/trends/place
[Accessed 1 November 2016].

Twitter, 2016c. *The Streaming APIs.* [Online]
Available at: https://dev.twitter.com/streaming/overview
[Accessed 24 October 2016].

Twitter, 2017c. *About public and protected Tweets.* [Online]
Available at: https://support.twitter.com/articles/14016
[Accessed 20 March 2017].

Twitter, 2017d. *Advanced Search.* [Online]
Available at: https://twitter.com/search-advanced?lang=en
[Accessed 23 March 2017].

Twitter, 2017. *FAQs about adding location to your Tweets.* [Online]
Available at: https://support.twitter.com/articles/78525
[Accessed 3 April 2017].

W3C, 2016. *HTML & CSS.* [Online]
Available at: https://www.w3.org/standards/webdesign/htmlcss
[Accessed 18 October 2016].

Waller, J., 2014. *Jenkins vs Travis.* [Online]
Available at: https://thoughts.wallproductions.com/2014/01/jenkins-vs-travis/
[Accessed 23 March 2017].

Zephoria, 2016. *The Top 20 Valuable Facebook Statistics – Updated September 2016.* [Online]
Available at: https://zephoria.com/top-15-valuable-facebook-statistics/
[Accessed 23 October 2016].