Technical Report

Mapping Top Irish Brands outside Ireland
- sentiment analysis -

Oana Cozma

x13118897

oana.cozma@student.ncirl.ie

BSc (Hons) in Computing

Stream: Software Development

May 2017

Declaration Cover Sheet for Project Submission

SECTION 1 *Student to complete*

Name: **OANA COZMA**
Student ID: **X13118897**
Supervisor: **MANUEL TOVA-IZQUIERDO**

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarized overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature:____

Date:____________

NB. If it is suspected that your assignment contains the work of others falsely represented as your own, it will be referred to the College's Disciplinary Committee. Should the Committee be satisfied that plagiarism has occurred this is likely to lead to your failing the module and possibly to your being suspended or expelled from college.

Complete the sections above and attach it to the front of one of the copies of your assignment

Table of Contents

Executive Summary

Social media monitoring brings brands and businesses insights into how their product is being perceived by the customers. By analysing what drives the users to mention their brand name and what types of emotions they are expressing, businesses can better understand their customer's needs and therefore have the possibility to improve their products.

For this project, the monitoring tool is provided by Twitter. The choice of this platform has been made taken into consideration the costs and also the quality of the data that is analysed. Twitter is one of the most known social media platforms, with more than 300M monthly active users that drive 1B unique visits monthly to sites with embedded Tweets, according to the statistics presented by the company ("Company | About").

"Mapping Top Irish Brands outside Ireland - sentiment analysis" research paper will collect, analyse and present the findings for top international Irish brands. The following sections of the technical report will present the project in more detail.

1 Introduction

There are a number of well-known top Irish brands that have a great deal of exposure on the international markets. Brands such as Guinness, Kerrygold, Tayto, Barry's Tea and many others are exporting their products worldwide and benefit from an increasingly online presence due to their customer's reviews and comments.

Tweets are 140 characters messages and they are incorporated in a microblogging service. Often registered users are using facial expressions (eg: :) :-D and others) to express emotions and moods. Other features include the "@" symbol and the hashtag, both used in order to connect the tweet with another user or with a trending topic.

In order to complete a specific research for each brand in particular, this project is going to focus on top international Irish brands. For this research to be unbiased and objective, the mechanism behind rating the international Irish brands will be based on public figures regarding exports from Irish Economy and Public Finances (National Treasury Management Agency). The methodology behind choosing those brands will be further described in detail, in the dedicated section.

1.1 Background

Nowadays an increasing number of agents are continuously performing this type of actions in order to drive more market share and therefore increase their revenue.

By analysing what drives the users to mention their brand name and what types of emotions they are expressing, businesses can better understand their customer's needs and therefore have the possibility to improve their products. Mapping complaints that drive changes, as well as monitoring the online reputation of a business are necessary business strategies.

Researching how different international Irish brands are being perceived outside Ireland can generate a powerful impact on the marketing strategies that are implemented. Comparing brands from the same industry (eg: food and beverage) can generate insightful cause-effect conclusions for any business in the sector.

1.2 Aims

"Mapping Top Irish Brands outside Ireland - sentiment analysis" project aims to collect, analyse and represent data for 10 international businesses in regards to how the brands are perceived worldwide.

The sentiment analysis performed on tweets, for each top Irish brand in particular, will lead to a worldwide map showing the level of engagement, online presence and type of emotions (different levels positive / negative) for each country.

In this moment in time, there are no other similar studies available for public use. What differentiates this project from existing sentiment analysis tools that offer reports for brands is the fact that it is focus on the market outside Ireland, offering cross-language and cross-cultural observations.

The finding may be of use to further international marketing studies and may support individual brand's strategic management decision making process.

1.3 Technologies

In order to monitor and process Tweets in real-time, this project will use The Streaming APIs resources. A proper implementation of a streaming client needs to be in place in order to gather data.

Connecting to the streaming API requires a continuous HTTP connection as the data flow will be on-going. After the server opens the streaming connection and Twitter accepts the connection, the Tweets are streamed as they occur. The application will receive streamed tweets, perform any processing required and store the results in a data store. The connection closes after the process is terminated, at request or due to technical fault.

The algorithm used for parsing the data is an important part of the sentiment analysis process. Online there are numerous information about this type of analysis and basic algorithms for parsing and interpreting data will form the basis in this step. Additionally, specific parameters will be added for each brand and a scaling mechanism of the emotions (negative/positive) will be defined.

1.3.1 Twitter API key

In order to access data from Twitter, there are a number of steps to consider. We require obtaining API key, API secret, Access token and Access token secret. This step is completed by transforming a normal account into a developer one and creating a new Twitter application.

1.3.2 Twitter Streaming API

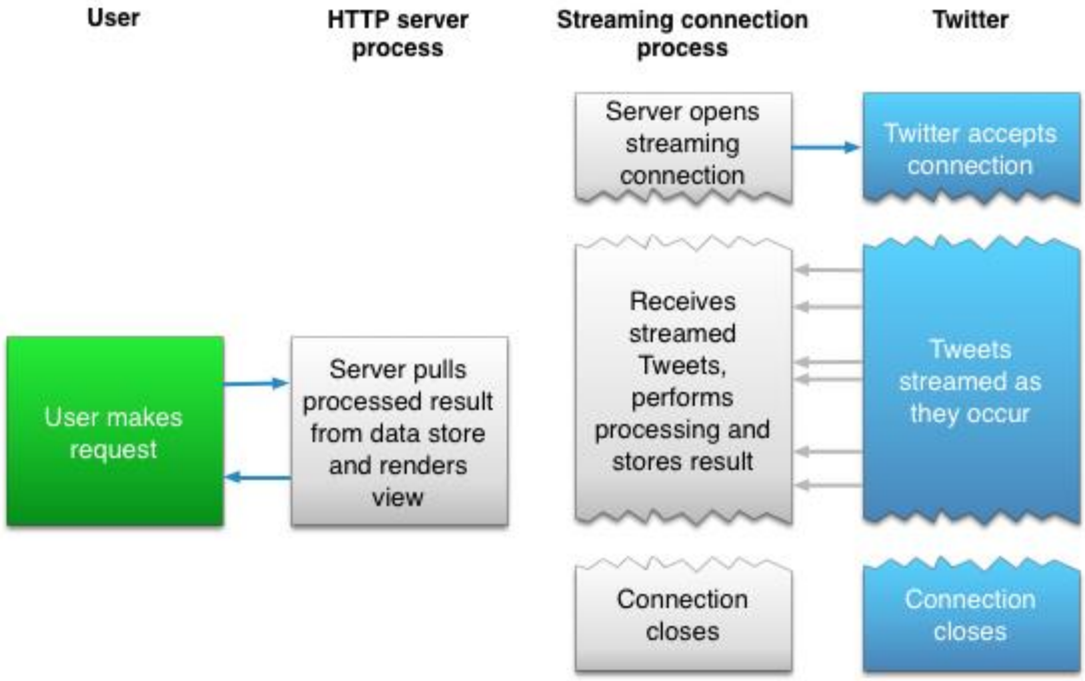
Twitter offers to all Developer Accounts the possibility to collect raw data from their users (tweets, user public information and other information regarding followers - connections between users).

At the moment, there are two options to access Twitter’s global stream of data, Streaming API and REST API. After considering both implementations and their differences, for the purpose of this research, we will be using the Streaming API tool.

Reasons behind this choice are as follows:

- REST API conducts singular searches for a period of maximum 2 weeks’ past tweets
- Streaming API allows the tweets data to be collected on a longer period of time (does not limit it to two weeks)
- REST API responds to one request that is after that formatted and displayed to the user
- Data is being preprocessed before it is being stored when using the Streaming API tool
- The streaming process, although it is more complex, offers the benefit of a real-time stream of data

Unlike the REST API, the streaming connection is run in a separate process, as outlined in the following design diagram (“Twitter Developer Documentation | Streaming APIs”).


User
HTTP server process
Streaming connection process
Twitter
Server opens streaming connection
Twitter accepts connection
Receives streamed Tweets, performs processing and stores result
Tweets streamed as they occur
User makes request
Server pulls processed result from data store and renders view
Connection closes
Connection closes

Streaming API offers developers three different streaming endpoints depending on the specific requirements. The Public stream is suitable for searches run by topics and data mining.

By implementing the POST statuses/filter both GET and POST requests are supported. In this way the application will return public statuses that match one or more filter predicates. There is a limit to consider in regards to the rate (up to 400 track keywords) for the default access level. The response formats are in JSON.

Parameters for the request that will be used are:

- Track
    - This parameter will contain the list of keywords on which the search is being performed
- Filter_level
    - none / low / medium
- Language
    - Specifying the language will only return Tweets detected to be in the specific language
- Follow

    - This parameter will be edited in order to not collect the tweets generated by the specific official brand accounts (eg: Guiness_UK)
- Locations
    - Option available only for the geo-located Tweets
- Stall warnings
    - Setting it TRUE will warn the developer if the streaming is in danger of being disconnected

Example of the API structure that contains the following parameters:

- Keyword: Ireland
- Location: France
- Language: french
- Including characters: “:)”, “:(“, “?” and retweets

https://twitter.com/search?f=tweets&q=ireland%20lang%3Afr%20near%3A%22France%22%20within%3A15mi%20since%3A2016-12-01%20until%3A2016-12-11%20%3A%29%20%3A%28%20%3F%20include%3Aretweets&src=typd

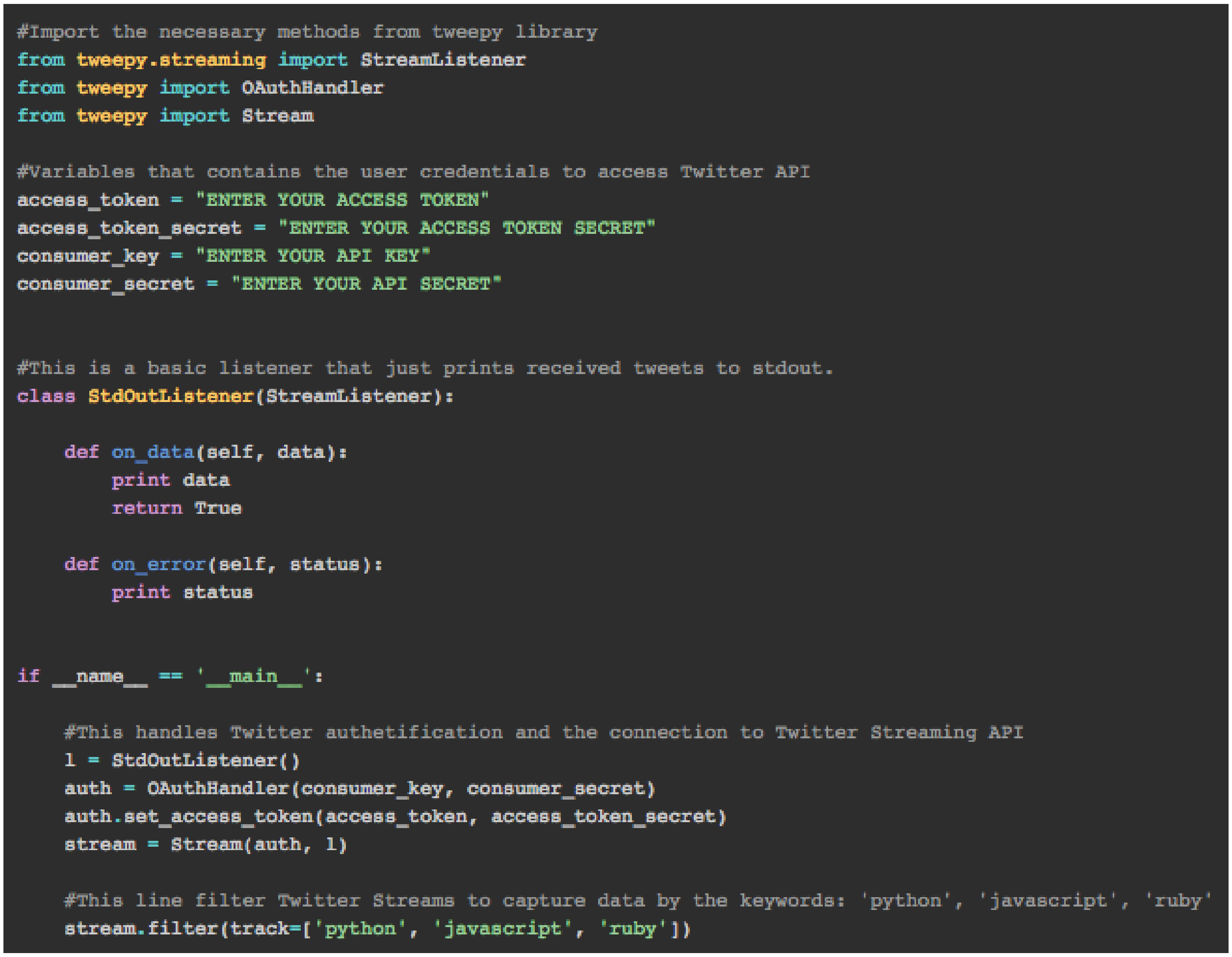
1.3.3 Python

In order to connect to Streaming API and to download data, we will use python specific library called “Tweepy”.

The command to execute the streaming and save the output in a file looks like this:

```
python twitter_streaming.py > twitter_data.txt
```

Following there is a demonstration of the implementation of *twitter_streaming.py* run by Adil Moujahid in the tutorial “An Introduction to Text Mining using Twitter Streaming API and Python”:



```
#Import the necessary methods from tweepy library
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

#Variables that contains the user credentials to access Twitter API
access_token = "ENTER YOUR ACCESS TOKEN"
access_token_secret = "ENTER YOUR ACCESS TOKEN SECRET"
consumer_key = "ENTER YOUR API KEY"
consumer_secret = "ENTER YOUR API SECRET"


#This is a basic listener that just prints received tweets to stdout.
class StdOutListener(StreamListener):

    def on_data(self, data):
        print data
        return True

    def on_error(self, status):
        print status


if __name__ == '__main__':

    #This handles Twitter authetification and the connection to Twitter Streaming API
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)

    #This line filter Twitter Streams to capture data by the keywords: 'python', 'javascript', 'ruby'
    stream.filter(track=['python', 'javascript', 'ruby'])
```

This will continuously gather data until the process is being stopped.

Another great application of python in this particular project refers to the specific libraries available for parsing the data (json), for manipulating the data (pandas), for creating charts (matplotlib) and for regular expressions (such as re).

This project also makes use of TextBlob python library which offers translation and text analysis. A series of NLP tasks are being offered by this library with just one method call.

1.3.4 JSON data format

All data is being returned by Twitter using a JavaScript Object Notation.

Following an example of a tweet in this format:

```
{
        "coordinates": null,
        "created_at": "Thu Oct 21 16:02:46 +0000 2010",
        "favorited": false,
        "truncated": false,
        "id_str": "28039652140",
        "entities": {
          "urls": [
            {
              "expanded_url": null,
              "url": "http://gnip.com/success_stories",
              "indices": [
                69,
                100
              ]
            }
          ],
          "hashtags": [
          ],
          "user_mentions": [
            {
              "name": "Gnip, Inc.",
              "id_str": "16958875",
              "id": 16958875,
              "indices": [
                25,
                30
              ],
              "screen_name": "gnip"
            }
          ]
        },
        "in_reply_to_user_id_str": null,
        "text": "what we've been up to at @gnip -- delivering data to happy customers
        http://gnip.com/success_stories",
        "contributors": null,
        "id": 28039652140,
        "retweet_count": null,
        "in_reply_to_status_id_str": null,
        "geo": null,
        "retweeted": false,
        "in_reply_to_user_id": null,
        "user": {
          "profile_sidebar_border_color": "C0DEED",
          "name": "Gnip, Inc.",
          "profile_sidebar_fill_color": "DDEEF6",
          "profile_background_tile": false,
  "profile_image_url":
"http://a3.twimg.com/profile_images/62803643/icon_normal.png",
  "location": "Boulder, CO",
  "created_at": "Fri Oct 24 23:22:09 +0000 2008",
  "id_str": "16958875",
  "follow_request_sent": false,
  "profile_link_color": "0084B4",
  "favourites_count": 1,
  "url": "http://blog.gnip.com",
  "contributors_enabled": false,
  "utc_offset": -25200,
  "id": 16958875,
  "profile_use_background_image": true,
  "listed_count": 23,
  "protected": false,
  "lang": "en",
  "profile_text_color": "333333",
  "followers_count": 260,
  "time_zone": "Mountain Time (US & Canada)",
  "verified": false,
  "geo_enabled": true,
  "profile_background_color": "C0DEED",
  "notifications": false,
  "description": "Gnip makes it really easy for you to collect social data for your
business.",
  "friends_count": 71,
  "profile_background_image_url":
"http://s.twimg.com/a/1287010001/images/themes/theme1/bg.png",
  "statuses_count": 302,
  "screen_name": "gnip",
  "following": false,
  "show_all_inline_media": false
},
"in_reply_to_screen_name": null,
"source": "web",
"place": null,
"in_reply_to_status_id": null
}
```

1.3.5 Wixx

Wixx is an open-source software tool that helps building and deploying websites in a fast and simple to use manner. Since this is our last step of this research, it is required only to cover the displaying of our findings. A pre-defined model was chosen and deployed in the cloud infrastructure wixx provides.

The website model incorporates design and features (such as menu, about section, contact section, subscribe) and reutilizes free available modules. For this assignment, the focus was placed more on the usability and addressing all the functional requirements in displaying/visualizing data findings.

1.3.6 Cloud services

For the connection to Twitter Streaming API to take place I made use of c9.io cloud web services. Decided to use this cloud space taken into consideration the storage needs and the availability.

Workspace size: Free version

Memory: 512 MB

Disk: 2 GB

This can easily be resized in care resource limits are being reached.

Another point which made the choice of cloud9 is the fact that they offer continuity even after logging out. The workspaces keep on running in the open terminal sessions. It is easy to access and edit from anywhere. C9 does offer the possibility for SSH Workspaces by providing access to user's own RAM, disk space and CPU.

1.3.7 Plotly

For the creation of graphs this project is collaborative using Plotly services. Modern data science focuses more on the quality and exciting ways of visualizing data. This helps making data research less boring and more engaging, raising more awareness amongst people outside the research fields and increasing the likelihood of the knowledge to be transmitted further.

Plot.ly is an online chart editor used to create interactive data visualization by uploading a .csv file (Comma separated values). Data in this form is easily interpreted and .csv files are being largely used when dealing with analyzing big data sets.

2 Data analytics

Following the Data sub-section in the Introduction of this paper, a more detailed approach into explaining the methodology applied to handle data is needed. This section will describe the data (input, process, output) and the steps involved when analyzing data from Twitter.

2.1 Background review

Researching up-to-date studies involving sentiment analysis on Twitter was a ground field for this paper. As software techniques evolve, getting information about the latest available methods in reaching this type of goal is mandatory.

The trending topic has evolved during the last years and a lot of thought and work has been put into understanding social media emotions. Understanding the opinion generated by a consumer in this case is done by analyzing the emotions expressed by a user in regards to a topic of choice.

For the purpose of this assignment, a lexicon approach method is being used when analyzing the sentiment of each tweet. Details regarding the implementation to follow in the next sections.

2.2 Methodology

2.2.1 Top Irish Brands

In this section, we will describe in detail the methodology used to generate the full list of brands that this project is focused on analyzing.

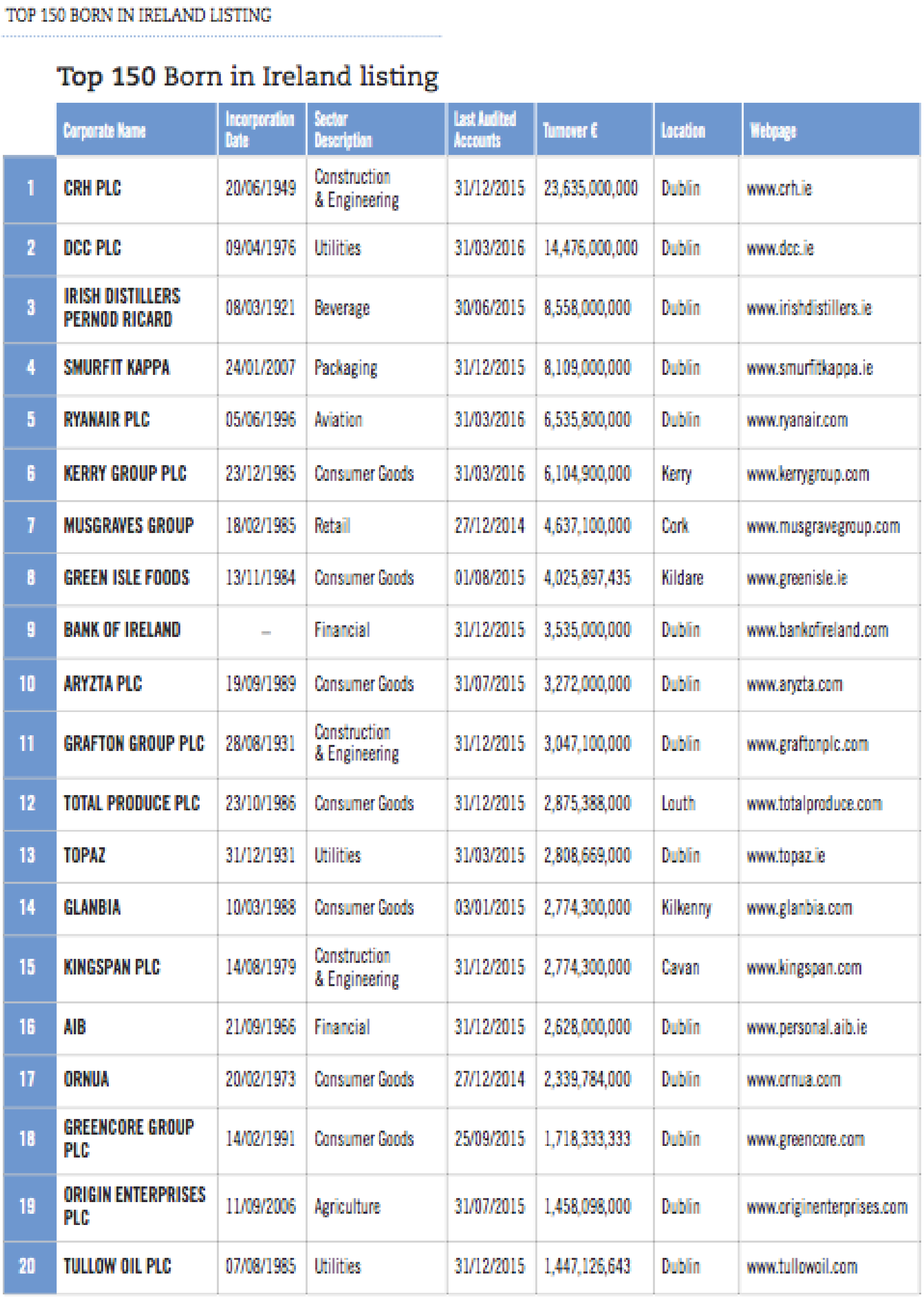
The first step was to research the comprehensive list of Irish Exporters and extract the Top 20 companies from official reports. In Ireland, Irish Exporters Association (referred to as IEA) represents a connecting force for all Irish exporters. They provide support to existing international exporters by liaising, offering consular services, training and sponsorship to members of the association.

Each year, IEA is publishing analysis and annual reports of the Irish exporters by turnover. As Simon McKeever, Chief Executive - Irish Exporters Association, mentioned in the Introduction of the 2016 publication, “Ireland’s exporting sector is one of Ireland’s greatest strengths and has been the key force behind the Irish economic, driving both growth and job creation.”, the public welcomes the launch of a new publication, *Top 150 Born in Ireland: An analysis of the leading indigenous Irish Exporters by turnover*.

As stated in the publication by Christine Cullen, Managing Director The methodology applied to compile the “Top 150 Born in Ireland 2016” companies list is as follows: “

- largest companies ordered by turnover in EUR. Currency conversions were performed with rates as at the end of August 2016
- Only companies that have filed accounts in 2014 or later
- Only non-dissolved companies with normal status
- Unlimited companies are included only if their audited accounts are available
- Duplicate companies have been removed is they shared accounts
- Charitable organizations have been excluded
- Inversion deals have been excluded
- Companies which started off as Irish owned, but which were taken over by larger non-Irish organizations have been excluded
- For groups, we have attempted to identify the most well-known brand, rather than the ultimate parent, the turnover figure displayed is that of the main trading entity within the Group.”

Section one of the paper is displaying the first 20 companies from the Top 150 Born in Ireland listing. They are, as follows:



TOP 150 BORN IN IRELAND LISTING

Top 150 Born in Ireland listing

	Corporate Name	Incorporation Date	Sector Description	Last Audited Accounts	Turnover €	Location	Webpage
1	CRH PLC	20/06/1949	Construction & Engineering	31/12/2015	23,635,000,000	Dublin	www.crh.ie
2	DCC PLC	09/04/1976	Utilities	31/03/2016	14,476,000,000	Dublin	www.dcc.ie
3	IRISH DISTILLERS PERNOD RICARD	08/03/1921	Beverage	30/06/2015	8,558,000,000	Dublin	www.irishdistillers.ie
4	SMURFIT KAPPA	24/01/2007	Packaging	31/12/2015	8,109,000,000	Dublin	www.smurfitkappa.ie
5	RYANAIR PLC	05/06/1996	Aviation	31/03/2016	6,535,800,000	Dublin	www.ryanair.com
6	KERRY GROUP PLC	23/12/1985	Consumer Goods	31/03/2016	6,104,900,000	Kerry	www.kerrygroup.com
7	MUSGRAVES GROUP	18/02/1985	Retail	27/12/2014	4,637,100,000	Cork	www.musgravegroup.com
8	GREEN ISLE FOODS	13/11/1984	Consumer Goods	01/08/2015	4,025,897,435	Kildare	www.greenisle.ie
9	BANK OF IRELAND	–	Financial	31/12/2015	3,535,000,000	Dublin	www.bankofireland.com
10	ARYZTA PLC	19/09/1989	Consumer Goods	31/07/2015	3,272,000,000	Dublin	www.aryzta.com
11	GRAFTON GROUP PLC	28/08/1931	Construction & Engineering	31/12/2015	3,047,100,000	Dublin	www.graftonplc.com
12	TOTAL PRODUCE PLC	23/10/1986	Consumer Goods	31/12/2015	2,875,388,000	Louth	www.totalproduce.com
13	TOPAZ	31/12/1931	Utilities	31/03/2015	2,808,669,000	Dublin	www.topaz.ie
14	GLANBIA	10/03/1988	Consumer Goods	03/01/2015	2,774,300,000	Kilkenny	www.glanbia.com
15	KINGSPAN PLC	14/08/1979	Construction & Engineering	31/12/2015	2,774,300,000	Cavan	www.kingspan.com
16	AIB	21/09/1966	Financial	31/12/2015	2,628,000,000	Dublin	www.personal.aib.ie
17	ORNUA	20/02/1973	Consumer Goods	27/12/2014	2,339,784,000	Dublin	www.ornua.com
18	GREENCORE GROUP PLC	14/02/1991	Consumer Goods	25/09/2015	1,718,333,333	Dublin	www.greencore.com
19	ORIGIN ENTERPRISES PLC	11/09/2006	Agriculture	31/07/2015	1,458,098,000	Dublin	www.originenterprises.com
20	TULLOW OIL PLC	07/08/1985	Utilities	31/12/2015	1,447,126,643	Dublin	www.tullowoil.com

As stated in the introductory section of the project, this study will only focus on beverage and consumer goods brands. From extracting only the companies in those two sectors and researching the brands behind the corporate names, we compiled the following detailed list:

	Corporate Name	**Sector**	**Brands**
1	Irish Distillers Pernod Ricard	Beverage	Jameson, Midleton, redbreast, dunphy's, huzzar, zaconey, absolut vodka, ballantines, Havana club, Malibu, Martell, glenlivet, beefeater gin, PERRIER-JOUËT, wyborowa, glenlivet, pernod, olmeca
2	Kerry Group PLC	Consumer Goods	LowLow, Cheestrings, Dairygold, Charleville, Denny, Richmond, Wall's, Matterssons, Fire&Smoke, yollies – yougurt lolly – KerryGoup
3	Green Isle Foods	Consumer Goods	Pizza: GOODfella's, SanMarco, frozen fish: DonegalCatch, frozen food: GreenIsle, Galway pie & pastry,
4	Aryzta PLC	Consumer Goods	Aryzta bakeries
5	Total produce PLC	Consumer Goods	TOP TotalProduce, TOPFruitFun, TOP Organic, TOP LoveLocal, TOP fresh'n ready, TOP Fruit for the senses, TOP Berries for the senses, TOP veg for the senses, Chef's Cut, Oppy, exczelent, perfekt, peniani, ideal, TotalProduceAgri, Coplaca, Kiwiberico, islaBonita, agroorigen bio, platano de Canarias, Ogra, total flowers, uniplumo, SunBurst, Progressive Produce, Pacific Gold,

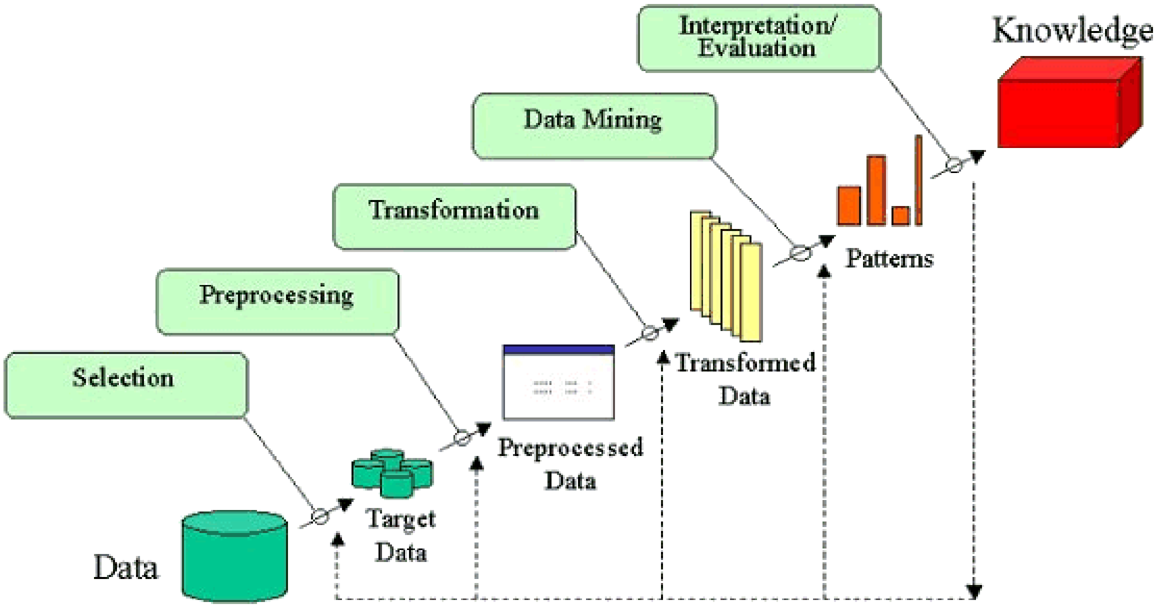
			Americas asparagus, Ole Pacifica, Hollywood Fries, Nature Bounty Organic, Dulce Sweet Onions, MICRO Baker
6	Glanbia	Consumer Goods	4 segments: Glanbia Performance Nutrition GPN (Optimum Nutrition (ON), BSN supplements, Isopure, thinkThin, Nutramino, ABB trusource), Glanbia Nutritionals GN, Dairy Ireland (Avonmore), Joint Ventures & Associates
7	Ornua	Consumer Goods	Kerrygold, Dubliner, Beo, pilgrims choice, Shannon gold
8	Greencore Group PLC	Consumer Goods	Bisto, Bistro to go, Heinz, Lettieri's, little dish, munch, Pandora pickles, sushi san, Sutherland deli, weight watchers chilled

Majority of the companies listed has a number of different business lines which combine a series of brand names. Some brand names are owned by the corporate from creation and others have bought the brand when fusion occurred. We will be treating all brands under the corporate name as equal. At this stage, all brands under each group are being considered in the research. At further stages of the analysis, some brand's name could not be processed (eg: Richmond - details in the Data Analytics section) and therefore the final list reduced.

The list above represents the basic introductory step in the data analytics process. Detailed information regarding the methodology used for gathering and mining the data are presented in the section next section of the paper, Data Analytics.

2.2.1 KDD (Knowledge Discovery in Databases)

The overall methodology applied to the process of extracting knowledge from data used broadly by researchers when dealing with large databases is called Knowledge Discovery in Databases (KDD for short).



Source: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

The graph above lists the steps of the KDD process. The research for this project is conducted respecting and outlining of this methodology and numerous references will be made to this process when it comes to analyzing data.

1st step: data cleaning: involves removing noise/outliers, managing strategies for missing fields

2nd step: data integration: getting more than one set of data in lined

3rd step: data selection: reduction and projection of the representing data in a goal-focused manner

4rd step: data transformation: data mining task: can be classification, regression, clustering and others; choosing the methods to be used for discovering the patterns in data

5th step: data mining task: can be classification, regression, clustering and others; applying methods to extract patterns

6th step: pattern evaluation: identifying the patterns representing knowledge based on measures

7th step: knowledge representation: visualization techniques are used to present the findings

Getting familiar with the data environment represents the first step. Researching Tweeter and it's data helped decide on the methods applied in order to collect data and for storing it.

We described in detail how the first approach in deciding what topics this research will focus on. With that in mind, gathering the raw data and storing it was done using python tweeter library. To process each tweet, storing the input text of each one of those data snippets is first step.

2.2.2 Sentiment analysis

As mentioned in the previous sections, sentiment analysis is the term used to describe the process of understanding and extracting feelings from data.

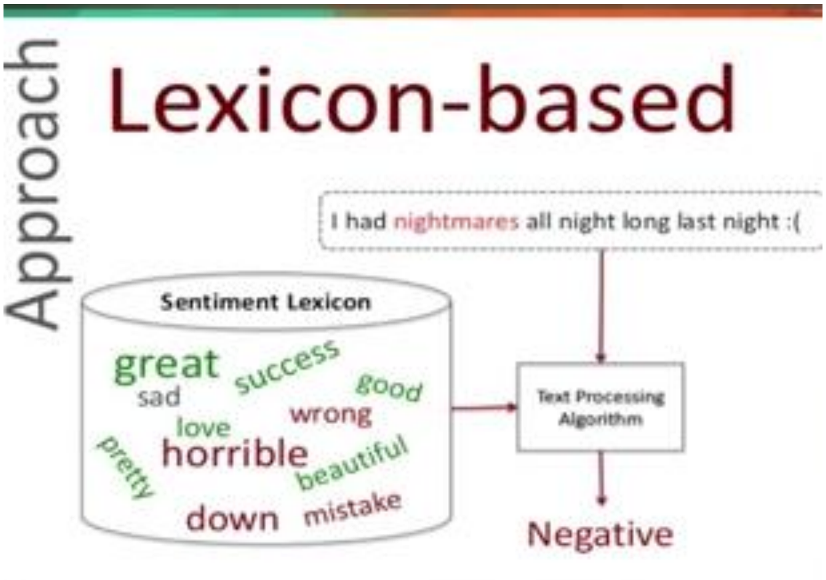
There are two main methods largely used when performing sentiment analysis on a given text:

* Lexicon approach – using a dictionary type of approach in determining the score: positive/negative/neutral of one's opinion; most cost and time efficient tool in performing opinion mining

* Machine learning – implementing a self-thought system that will address issues like sarcasm, intensions, irony and other human mind-like interactions that are hard to be programmed. This method involves manually scoring tweets at the beginning of the process as positive or negative so that the machine is writing its own rules and conditionals regarding sentiment of words associations and not stand-alone words unlike the lexicon approach.

After the tokenization, which implies splitting the data (tweet) into small tokens. This process is breaking each line of text into words and counting the number of times each word is presented in tweet. We are basically creating a “bag of words” for each tweet.

Lexicon-based approach is used as next step in the process to attribute a value for each “word” in the “bag of words”. In this method, a sentiment value is being generated from an already defined dictionary of words.



After getting the sentiment value of each word, the process then totalizes the tweet’s sentiment value between -1 and 1 (from negative to positive).

2.3 Implementation

Steps involved in the implementation of the data analytics research are outlined below.

2.3.1 Setting up the environment

As mentioned in the introduction, c9 is the chosen cloud environment for the development of this application. C9 offers pre-installed workspaces – python included.

Python 3.4.3 (default, Nov 17 2016, 01:08:31) is used in this project.

library for accessing the tweeter api:

```
sudo pip3 install tweepy
```

Tweepy is the library used by python developers to connect to Twitter API by just calling pre-defined methods.

Below is a code snipped user for connecting to the Streaming API:

```
#Import the necessary methods from tweepy library
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
(…)
#Variables that contains the user credentials to access Twitter API
access_token = "…."
access_token_secret = "…."
consumer_key = "…."
consumer_secret = "…."


class StdOutListener(StreamListener):
   def on_data(self, data):
      decoded = json.loads(data)
        (…)
      return True
   def on_error(self, status):
      print(status)
if __name__ == '__main__':
   #This handles Twitter authetification and the connection to Twitter Streaming
API
   l = StdOutListener()
   auth = OAuthHandler(consumer_key, consumer_secret)
   auth.set_access_token(access_token, access_token_secret)
   stream = Stream(auth, l)
```

2.3.2 KDD steps

Following the connection, gathering data is being done using keywords representing brands. Here is a code example for collecting data under the Irish Distillers corporate name:

```
   #This line filter Twitter Streams to capture data by the keyword: 'python'
   stream.filter(track=['Jameson', 'Midleton whiskeys', 'redbreast', 'dunphys',
'huzzar', 'zaconey', 'absolut vodka', 'ballantines', 'Havana club', 'Martell', 'glenlivet',
'beefeater gin', (…) ])
```

In the period of time 15 March – 25 April 2017, after a series of attempts followed by evaluation and updates to the code, the raw data gathered was around 50MB of size.

The cleaning of the data is being done in several steps, one of them outlined below:

```
def clean_tweet(tweet):
    new_tweet=tweet.replace('\n', ' ').replace('\r', '').lower()
    return new_tweet
n_tweet=clean_tweet(text)
```

The next step involved the selection of the data to apply the mining algorithm for. Json formats can be decoded in python using this approach:

```
tw_id = decoded['id']
tw_user = decoded['user']['screen_name']
text = decoded['text']
location=decoded['user']['location']
user_lang=decoded['user']['lang']
geo=decoded['geo']
lang=decoded['lang']
```

In order to determine the brand name for each tweet, after the initial filtering of the stream, we are forced to run a number of if statements, like the one bellow:

```
h=['bisto', 'bistro to go', 'heinz', 'lettieris','pandora pickles', 'sushi san', 'sutherland deli', 'weight watchers chilled']
if any(x in n_tweet for x in h):
    brand = 'Greencore'
```

When running sentiment analysis, we will be importing and using TextBlob library, as mentioned in the introduction. Considering the fact that not all tweets are written in English, calling the sentiment.polarity method to a non-English text returns the value of “zero”.

By addressing this further, the code snipped bellow shows how tweets in a different language are being first translated to English and after that the sentiment.polarity is being calculated.

```
from textblob import TextBlob
    analysis=TextBlob(text)
    if lang=='en':
      sentiment = analysis.sentiment.polarity
    else:
      try:
        n_text = analysis.translate(to='en')
        sentiment=n_text.sentiment.polarity
      except:
        sentiment = analysis.sentiment.polarity
```

The example of a tweet in this form is being stored in a file in the cloud:

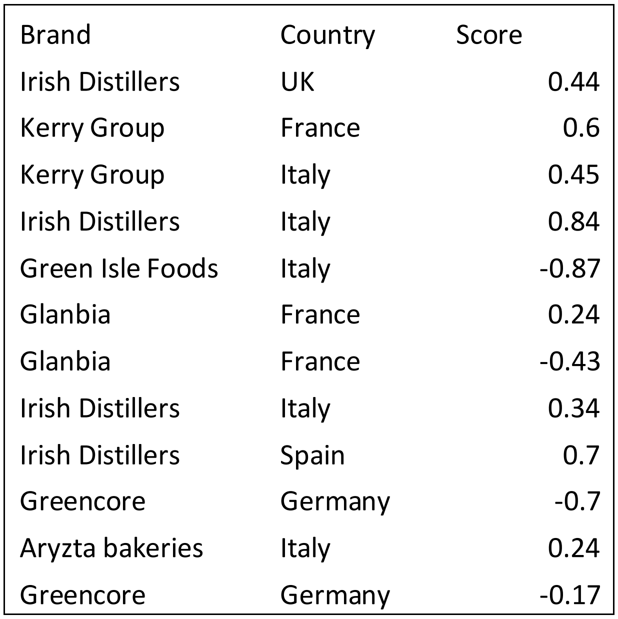
```
    saveMe = tw_user+ ':::' + n_tweet +':::' + str(sentiment) + ':::' + brand+ ':::' + location +':::'+user_lang+':::'+lang+'\n'
    output = open('data/data_tw_stream.csv','a')
    output.write(saveMe)
    output.close()
```

The data is then transformed in this type of record:

```
BooLinx:::rt a friend who would like a #goodfellas tee now available at
https://t.co/ksosy4o9h1 #londonltdclub https://t.co/rvpsnorufo:::0.4:::Green Isle
Foods:::Null:::en:::en
```

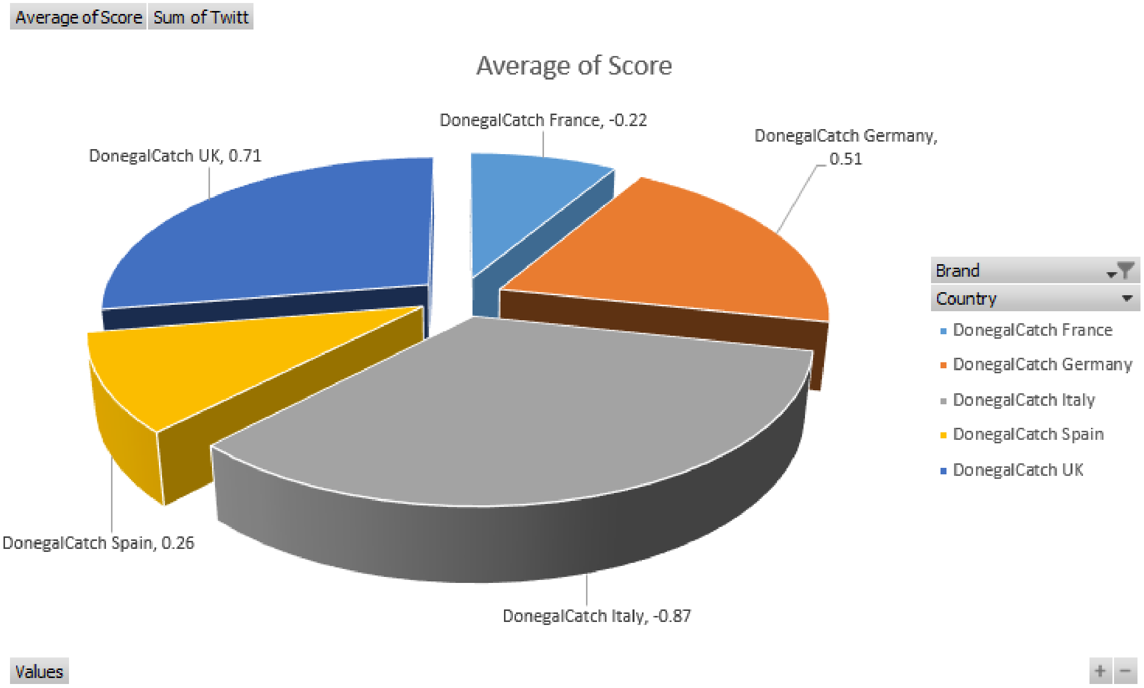
In other words, from the tweet example presented in the Introduction section, we are extracting and storing the user name, the text of the tweet, the sentiment value of the text, the brand's name, the location and two other language indicators (to help us in the process of locating the tweet we will be using both user_lang and tweet_lang parameter).

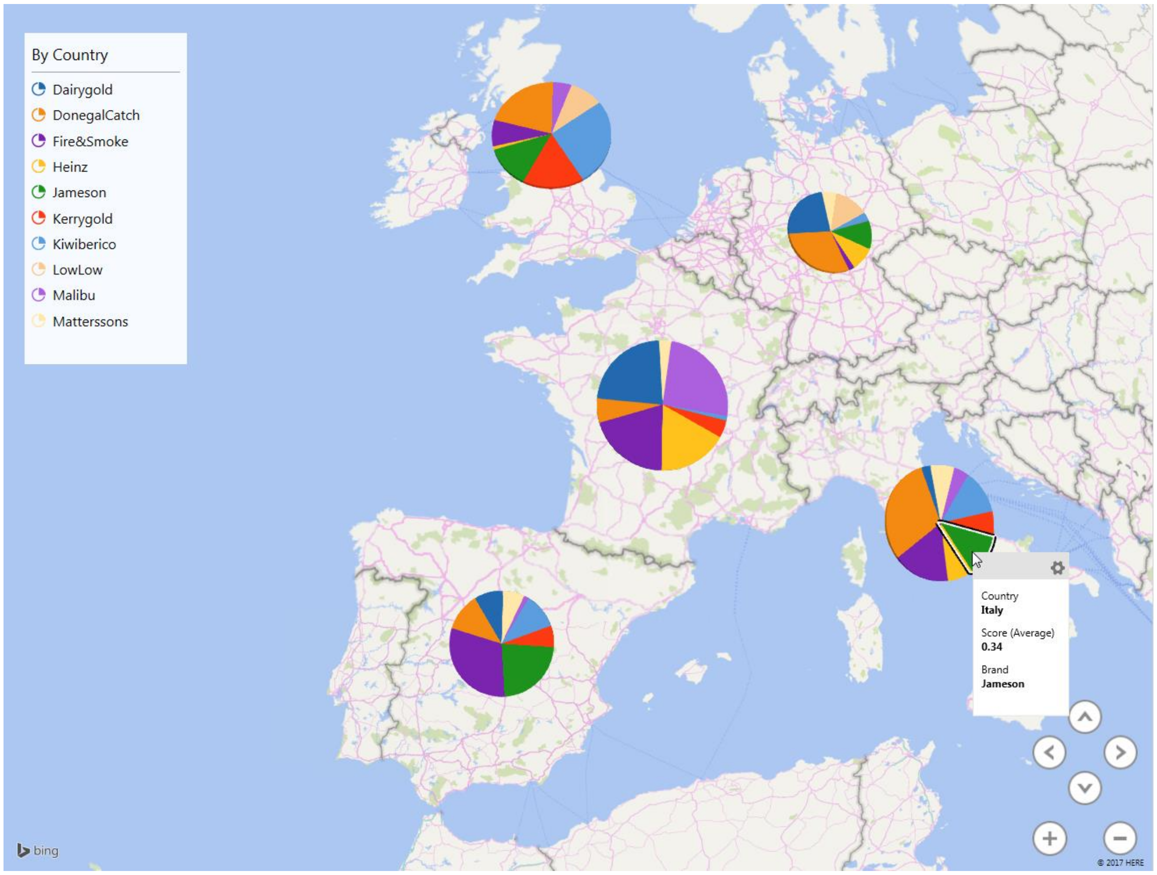
The data reduction step involves extracting only the next headers from the records described:



Brand	Country	Score
Irish Distillers	UK	0.44
Kerry Group	France	0.6
Kerry Group	Italy	0.45
Irish Distillers	Italy	0.84
Green Isle Foods	Italy	-0.87
Glanbia	France	0.24
Glanbia	France	-0.43
Irish Distillers	Italy	0.34
Irish Distillers	Spain	0.7
Greencore	Germany	-0.7
Aryzta bakeries	Italy	0.24
Greencore	Germany	-0.17

In the data visualization step, data takes one of the following forms:


Average of Score
Sum of Twitt
Average of Score
DonegalCatch France, -0.22
DonegalCatch Germany, 0.51
DonegalCatch UK, 0.71
DonegalCatch Spain, 0.26
DonegalCatch Italy, -0.87
Brand
Country
DonegalCatch France
DonegalCatch Germany
DonegalCatch Italy
DonegalCatch Spain
DonegalCatch UK
Values
+
−


By Country
Dairygold
DonegalCatch
Fire&Smoke
Heinz
Jameson
Kerrygold
Kiwiberico
LowLow
Malibu
Matterssons
Country
Italy
Score (Average)
0.34
Brand
Jameson
bing
© 2017 HERE

Different tools are being used in this step in order to make data visualization more easy to interpret by the final-user. 3D Map excel and Plotly are the open source software that is enforced to display the results of this research paper.

2.3.3 Evaluation

In the processing of the data several challenges were faced, some were addressed and resolved others have yet to be considered.

* Spam tweets

* Sarcasm – sentiment analysis fails

* trending topics (eg: Jameson Taillon cancer operation)

* twitter does not allow multiple streaming connections or repeated connection calls (successful or not)

* location of tweets (a lot of missing values; user manual input)

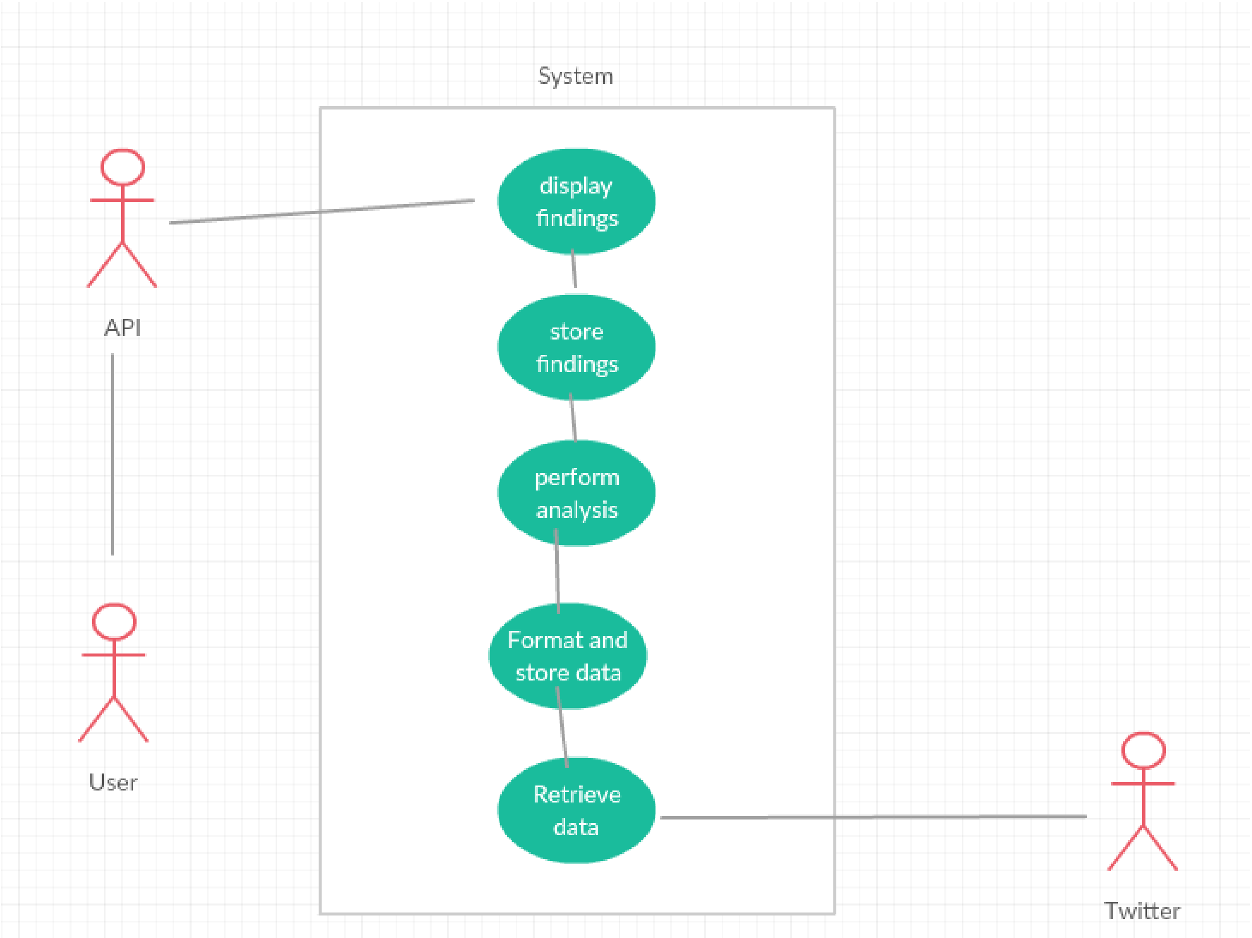
3 System

3.1 Requirements

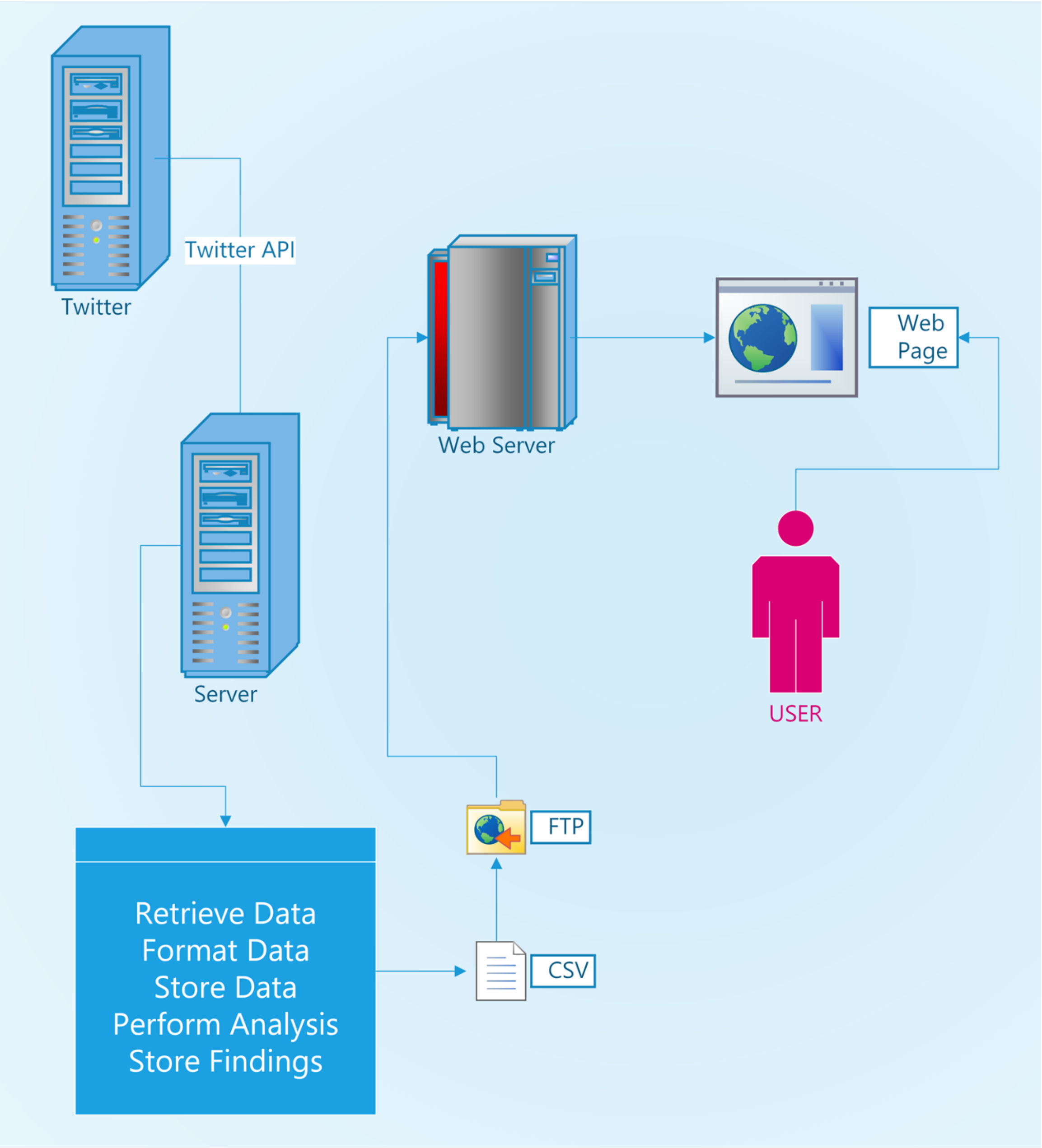
3.1.1 Functional requirements

Use Case Diagram

The Use Case Diagram provides an overview of all functional requirements.



As described in the above sections, the user interacts with the system through a web page that gets the information from the back-end application.



Requirement 1 - User can view the findings

Description & Priority

The main goal of the project is to display in a visual with high impact the findings of the analysis on each brand in particular. The user has a great interest in this requirement.

Use Case

Map view of one’s brand analysis findings

Scope

The scope of this use case is to display to the user the findings of the study

Description

This use case describes the process in which a user is interacting with the system in order to it to display a world’s map containing the countries on which sentiment analysis was being performed.

Use Case Diagram

Diagram should highlight actors and uses cases

Flow Description

Precondition

The system is in initialization mode

Activation

This use case starts when an <Actor> clicks on the drop down box and selects a certain brand they want to view

Main flow

- The system identifies the request
- The <Actor> selects a brand
- The system displays the findings for that particular brand
- The <Actor> interacts with the map

Alternate flow

- A1 : <title of A1>
- The system identifies the request
- The <Actor> does not select any particular brand
- The use case continues at position 3 of the main flow

Exceptional flow

- E1 : <title of E1>
- The system displays the findings of a particular brand
- The <Actor> selects a different brand and sends the request
- The use case continues at position 4 of the main flow

Termination

The system presents the next brand.

Post condition

The system goes into a wait state.

Requirement 2 - user can view country statistics

Description & Priority

Beside the findings, users need to be able to view statics over each country.

Use Case

Viewing the statistics on each country

Scope

The scope of this use case is to display information regarding the number of tweets presented in the study and the location of them.

Description

This use case describes the user interaction with the map.

Use Case Diagram

Flow Description

Precondition

The system is in initialization mode.

Activation

This use case starts when an <Actor> clicks on a highlighted country from the map displayed.

Main flow

- The system identifies the request - it is already displaying a map
- The <Actor> selects a country
- The system displays a pop up containing the details - statistics
- The <Actor> can exit that view

Alternate flow

- A1 : <title of A1>
- The system is already displaying a map
- The <Actor> does not select any country
- The use case continues at position 3 of the main flow

Exceptional flow

- E1 : <title of E1>
- The system is displaying statistics for a country
- The <Actor> does not exit the window
- The use case continues at position 4 of the main flow

Termination

The system presents the next brand.

Post condition

The system goes into a wait state.

3.1.2 User requirements

From a client's perspective, the final product must display in a simple and easy to read manner the findings of the project next to information regarding design (algorithm, weighing scale).

The possibility of manipulating data from the website, and displaying in real time the new results is also part of the user requirements.

3.1.3 Non-Functional Requirements

The particular non-functional attributes required by the system are as follows.

Performance/Response time requirement

Each interaction the use makes with the system needs to be able to respond in a performant manner. Editing the key terms in the research and manipulating data is of high importance to be made fast and efficient, as a background process, and displayed in a matter of seconds.

Availability requirement

In the process of gathering the data, the streaming needs to be connected with the system on an ongoing basis in order to support this process. Availability in this scenario is highly important as this requires the system to be online and connected to be able to receive the amount of data requested.

Robustness requirement

Due to the constant connection and interaction with the database, robustness is an important non functional requirement. When the user will manipulate the data mining process by designing new algorithms, the calls the user makes using the API need to correspond and comply to the system.

3.2 Design and Architecture

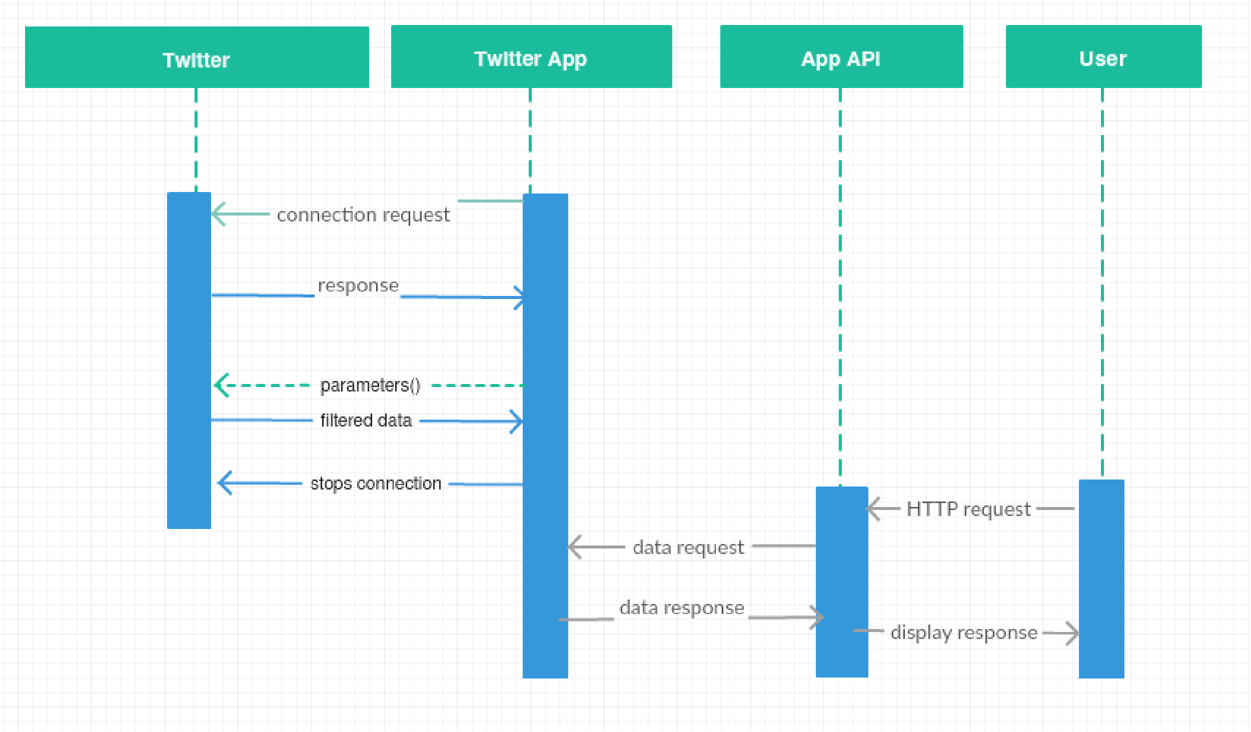
The analysis performed on each selected brand will follow the same rules for all, but will differentiate in terms of emotions-keywords. This process will be defined according to the pre-analysis (manually) of messages on each brand.

Beside analysing each selected brand separately and displaying the results individually, this project will also involve gathering information regarding the location of the tweet (country wise). This information being among the most important one in this research, the plan is to incorporate more than one method of finding the location of that tweet (language detection tools, profile bio of the user). This is especially important in order to geo-locate tweets that do not have a geo-location tag. This feature is optional and it is considered that few Twitter users are sharing their location coordinates by default.

Data will be stored in a local data storage and handled using python.

Data will be translated using Google translate tool before being stored in the local storage.

Sequence diagram:

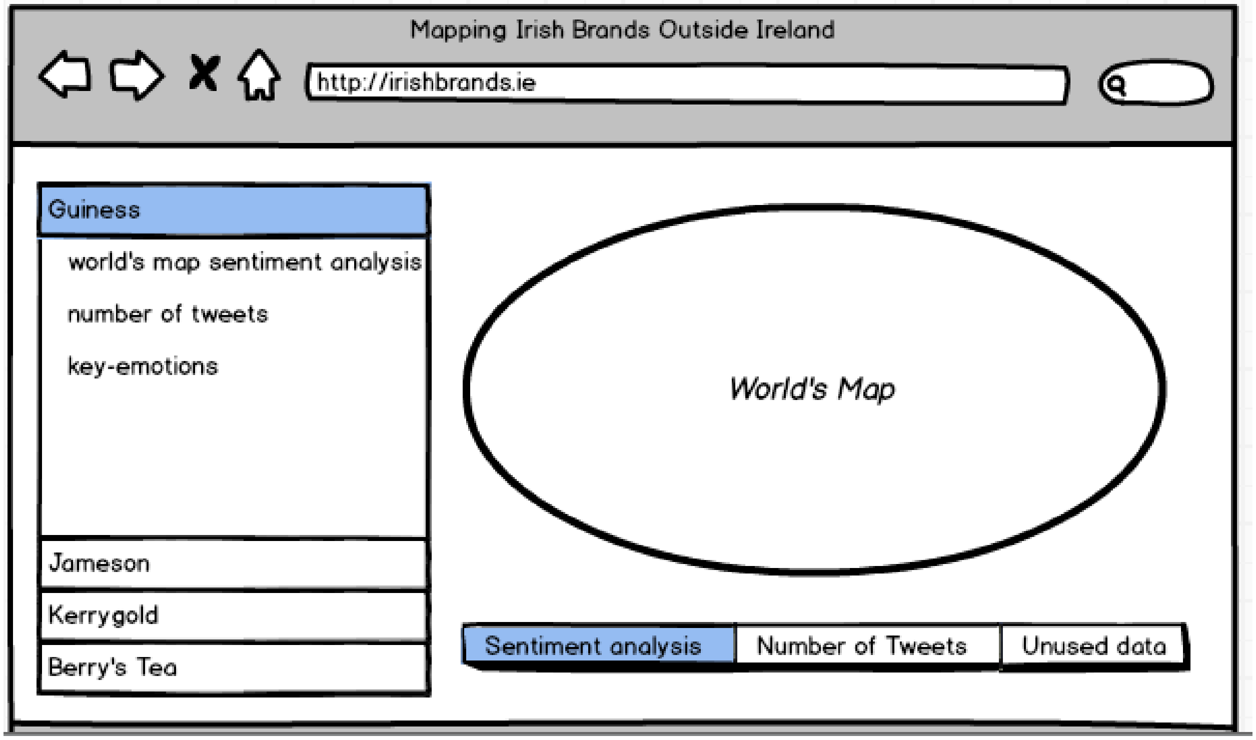


The above sequence diagram shows the relations between Twitter, Twitter Application, Application API and User.

3.3 Graphical User Interface (GUI) Layout

The GUI represents an API designed to display in an easy to interpret manner the findings of this research.

A prototype will be in place to display furthermore the bellow mockup:



As pictured in the mockup plan, the API will contain a list of the Irish brands being analysed and three different views for each:

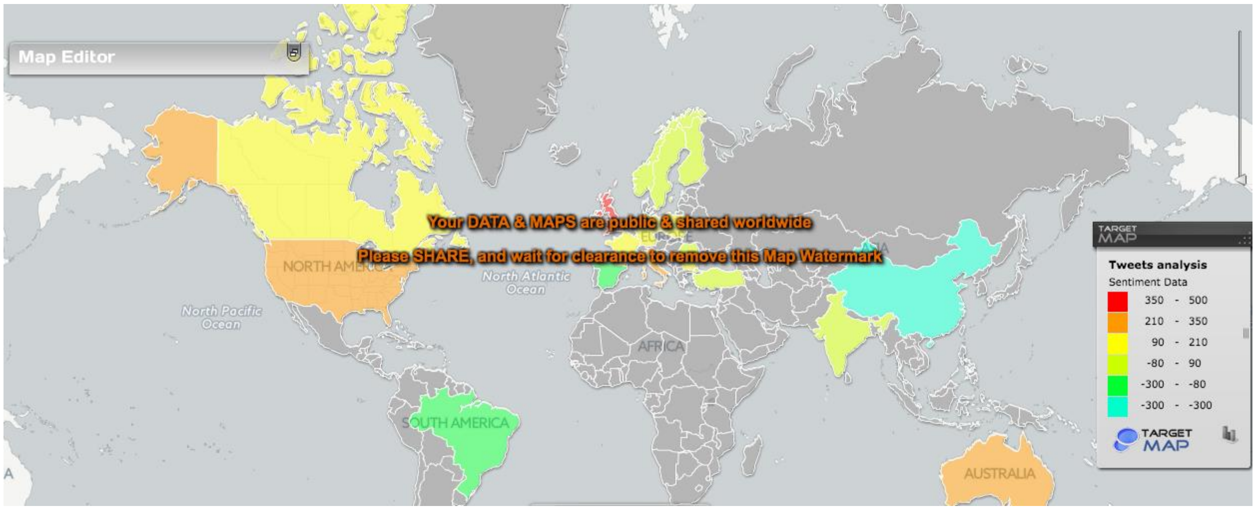
- World's map view
- Number of tweets view
- Key-emotions list

Displaying on each country, colored differently, depending on the intensity of the sentiment (positive, negative, neutral) will also display information regarding the number of tweets being analyzed.

The information about the number tweets without geolocation parameter or that were not included in the research will be available for display.

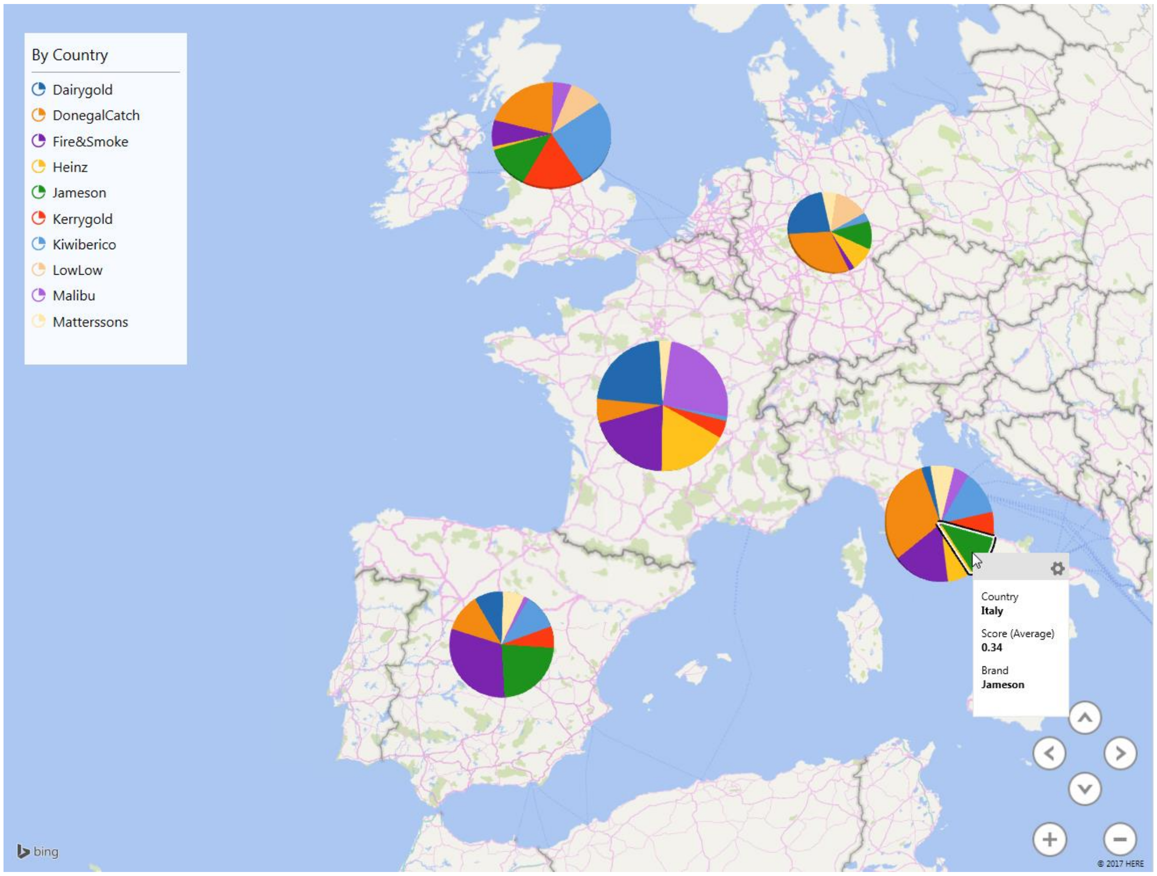
Each tweet will be weighted accordingly and the score will be stored in a local data storage. The keywords specific to each brand and their weighting scale will be available.

The initial plan was that the world's map views will be supported using TargetMapp.com - The data can be inserted using an excel form. Following this an example of how a map could look like. The map will also have to be interactive and responsive.



targetmap.com

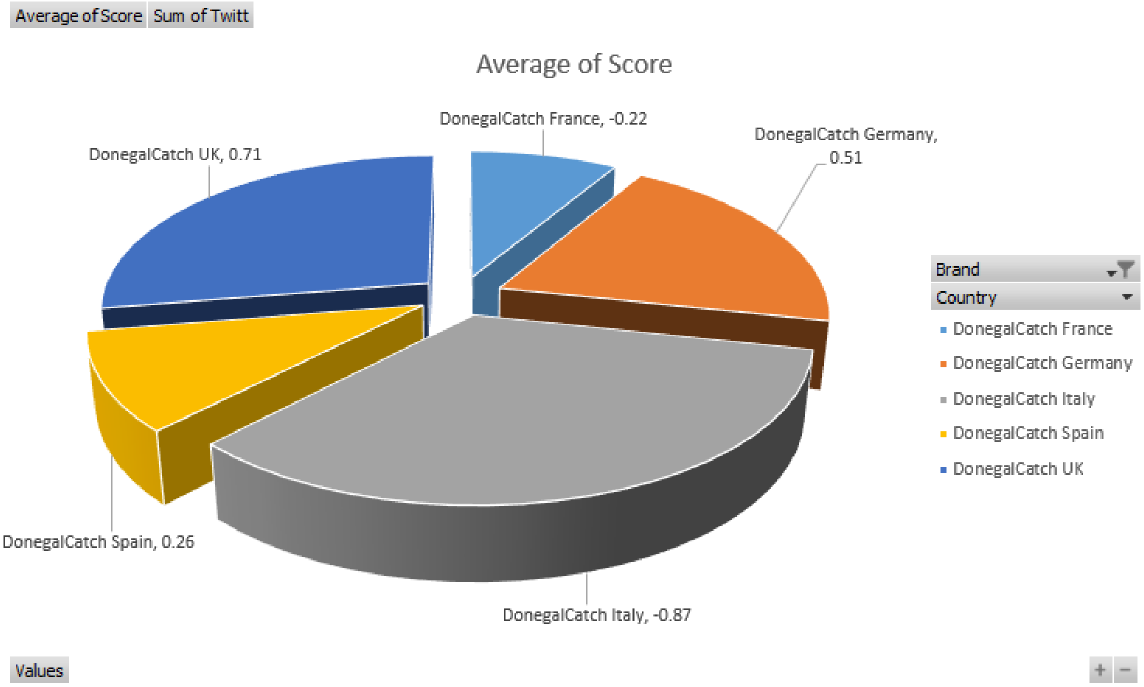
A more up to date approach incorporates graphics outside TargetMap.com. Here are some examples:

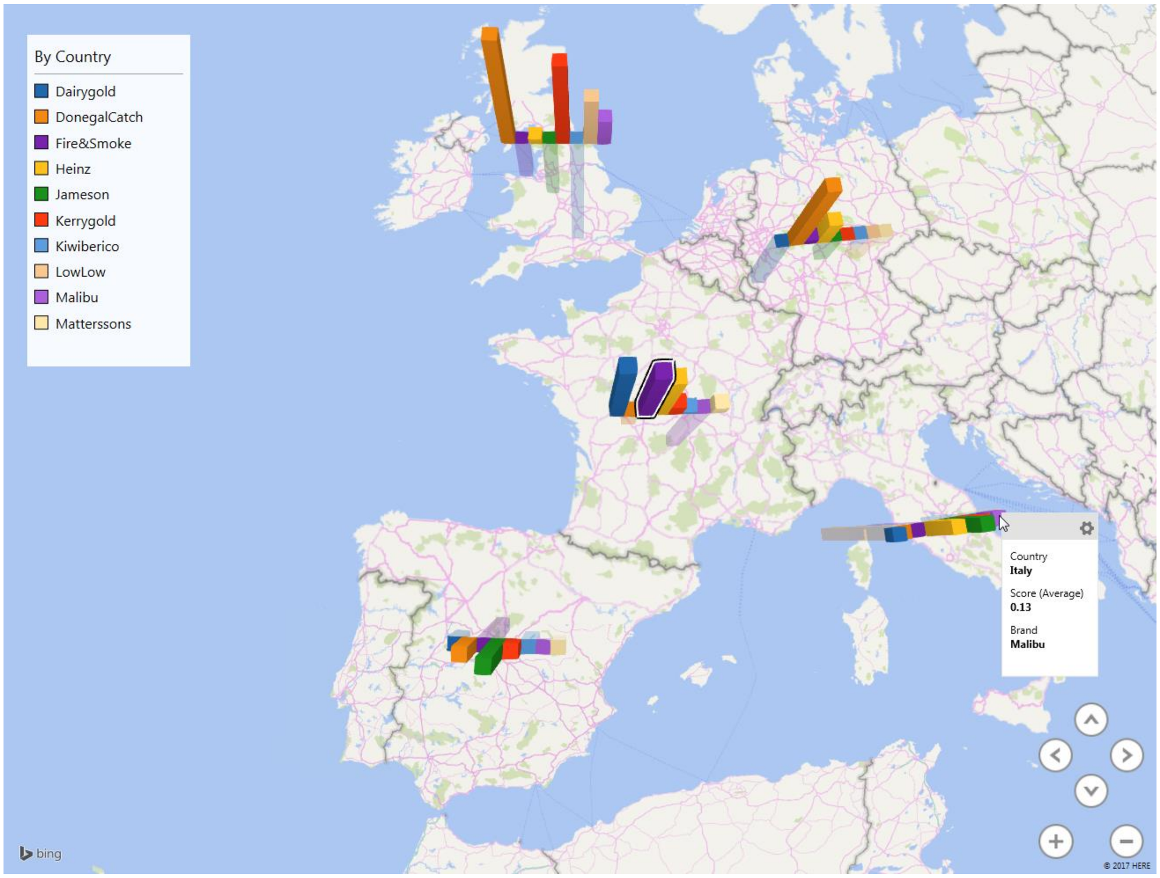


Country
Italy

Score (Average)
0.34

Brand
Jameson





Country
Italy

Score (Average)
0.13

Brand
Malibu

The GUI also displays information about the project and the possibility to subscribe to news from the author. The usability design features used are respecting the modern attributes such as: easy flowing, less structure, no extra information, keep in simple, one page, one design.

The website is being created and deployed using wixx services and can be accessed at the following link:

https://oanacozma.wixsite.com/irishbrands

4 Testing

Testing in software throughout the circles of the development of the project has a great importance and effect on the quality and completion of the work.

Professional programmers automating their test using something like Unit testing. This way all individuals involved in the development or maintenance of the product have a way of knowing that those units (pieces of code) are delivering or acting as expected.

4.1 Unit testing

In python, there is a library commonly used with this scope, called unittesting.

First, we need to perform the imports:

```
import unittest
```

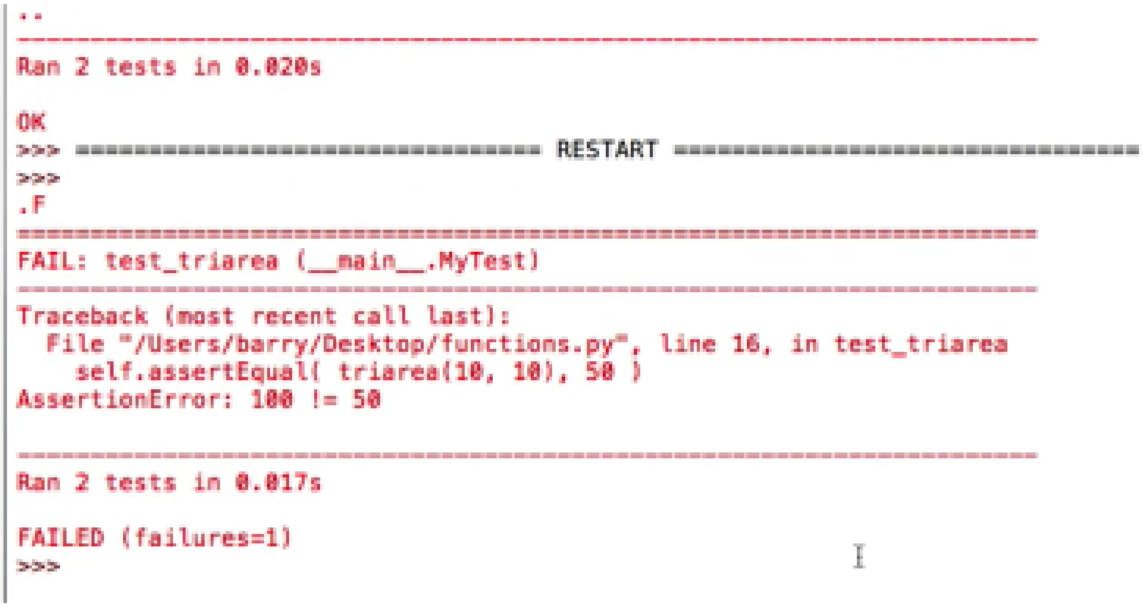
Secondly, the creation of the class that handles unit testing methods. Here we are testing the clean_tweets method on the main class by trying different possibilities and testing that the expected outcome is being given.

```
class MyTest(unittest, TestCase):

        def test_cleaning_noise(self):

                tweet_to_check = “this is “+”\n”+” messy..+”\r”

                clean_tweet = “this is  messy..”

                self.assertEqual( clean_tweets(tweet_to_check), clean_tweet)

                …

                self.assertEqual( clean_tweets(“\n\n\n\r “), “ “)
```

At the bottom, we are calling the testing method by using this line of code, inside the main tw_stream.py file:

```
unittest.main()
```

Depending on the data and the operations in place, there are two types of outcomes when running the unit testing methods. Here is a snapshot of possible outcomes:



```
..
----------------------------------------------------------------------
Ran 2 tests in 0.020s

OK
>>> ================================ RESTART ================================
>>>
.F
======================================================================
FAIL: test_triarea (__main__.MyTest)
----------------------------------------------------------------------
Traceback (most recent call last):
  File "/Users/barry/Desktop/functions.py", line 16, in test_triarea
    self.assertEqual( triarea(10, 10), 50 )
AssertionError: 100 != 50

----------------------------------------------------------------------
Ran 2 tests in 0.017s

FAILED (failures=1)
>>>
```

Details regarding the type of error, the place and the outcome are being given. This makes the bug fixing tasks a lot easier and the understanding of the error knowledgeable.

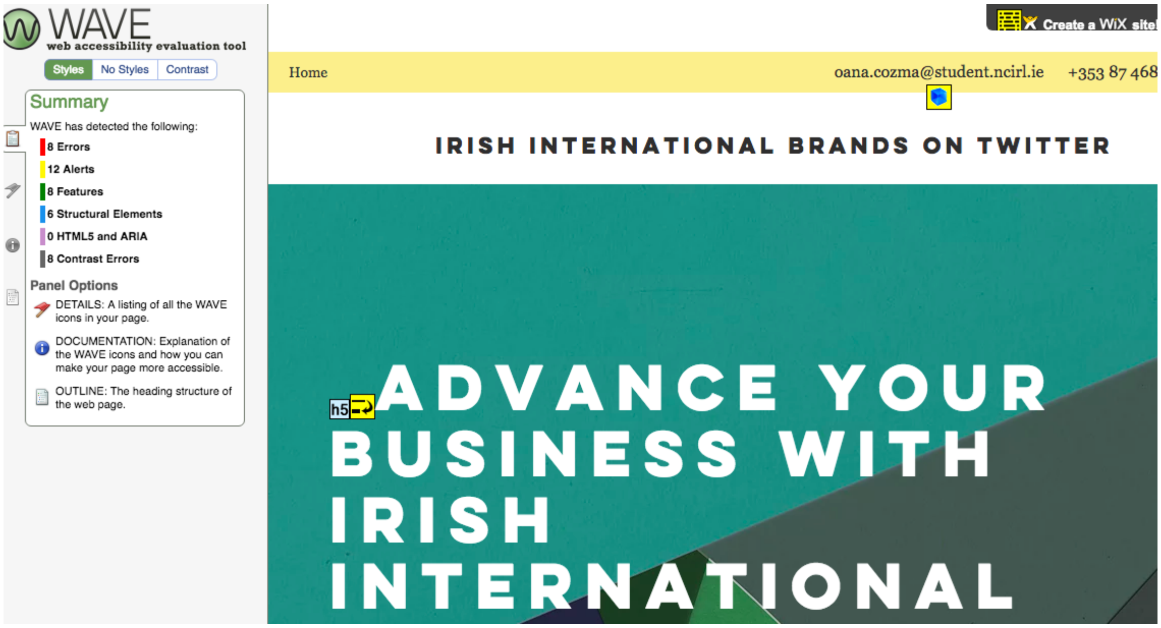
Another approach on testing that is available together with this module refers to the help given by commenting the output or giving instruction in regards to the return value of a function. When writing code, the comment written inside the test_ method appears when calling that function, making it easier to comprehend what is expected of that method to return, so that the input given when running the automated tests is being thought throughout.

4.2 Accessibility testing

Usability testing is being carried out on the final version of the web application, the one hosted on wix.com which deals with user interaction. Requirements are check to be met on a front-end approach by using the online free usability testing tool WAVE: http://wave.webaim.org/

This tool is checking the accessibility of the website by performing tests on errors, alerts, features, structural elements, HTML5 and ARIA and registering contract errors.

A copy of the WAVE report is being attached below:



WAVE
web accessibility evaluation tool
Styles
No Styles
Contrast
Create a WiX site!
Home
oana.cozma@student.ncirl.ie
+353 87 468
IRISH INTERNATIONAL BRANDS ON TWITTER
h5
ADVANCE YOUR BUSINESS WITH IRISH INTERNATIONAL

Summary

WAVE has detected the following:

- 8 Errors
- 12 Alerts
- 8 Features
- 6 Structural Elements
- 0 HTML5 and ARIA
- 8 Contrast Errors

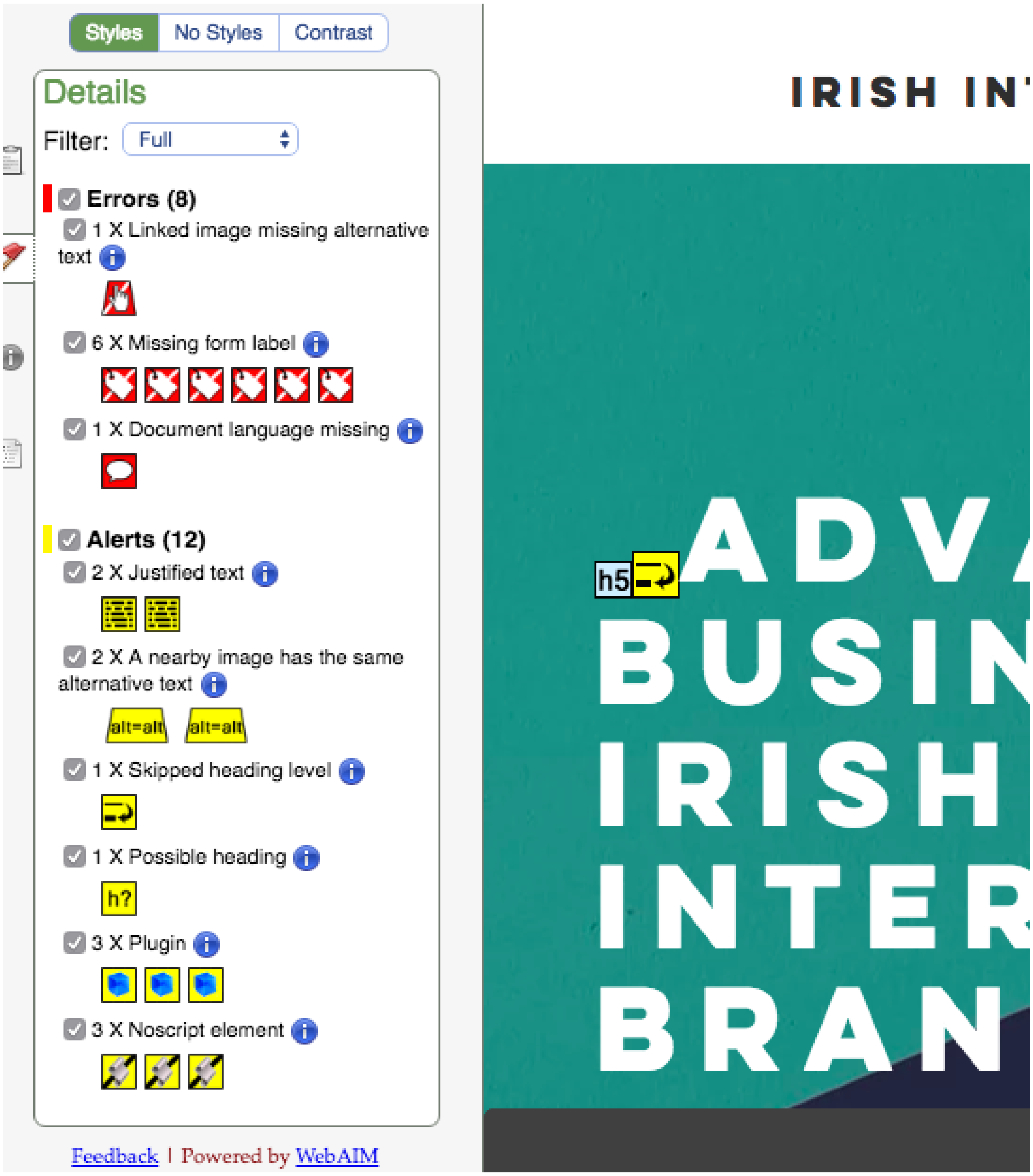
Panel Options

DETAILS: A listing of all the WAVE icons in your page.

DOCUMENTATION: Explanation of the WAVE icons and how you can make your page more accessible.

OUTLINE: The heading structure of the web page.

List of errors eventuated at the beginning of the development cicle:



Details

Filter: Full

☑ Errors (8)

☑ 1 X Linked image missing alternative text

☑ 6 X Missing form label

☑ 1 X Document language missing

☑ Alerts (12)

☑ 2 X Justified text

☑ 2 X A nearby image has the same alternative text

☑ 1 X Skipped heading level

☑ 1 X Possible heading

☑ 3 X Plugin

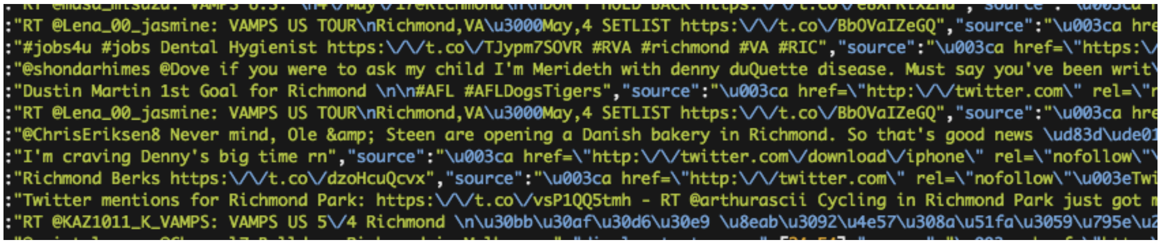
☑ 3 X Noscript element

After installing the Chrome extension, WAVE became available offline and made work more efficient. An excellent tool for staying on top of your requirements, errors and bugs for front end development.

5 Evaluation

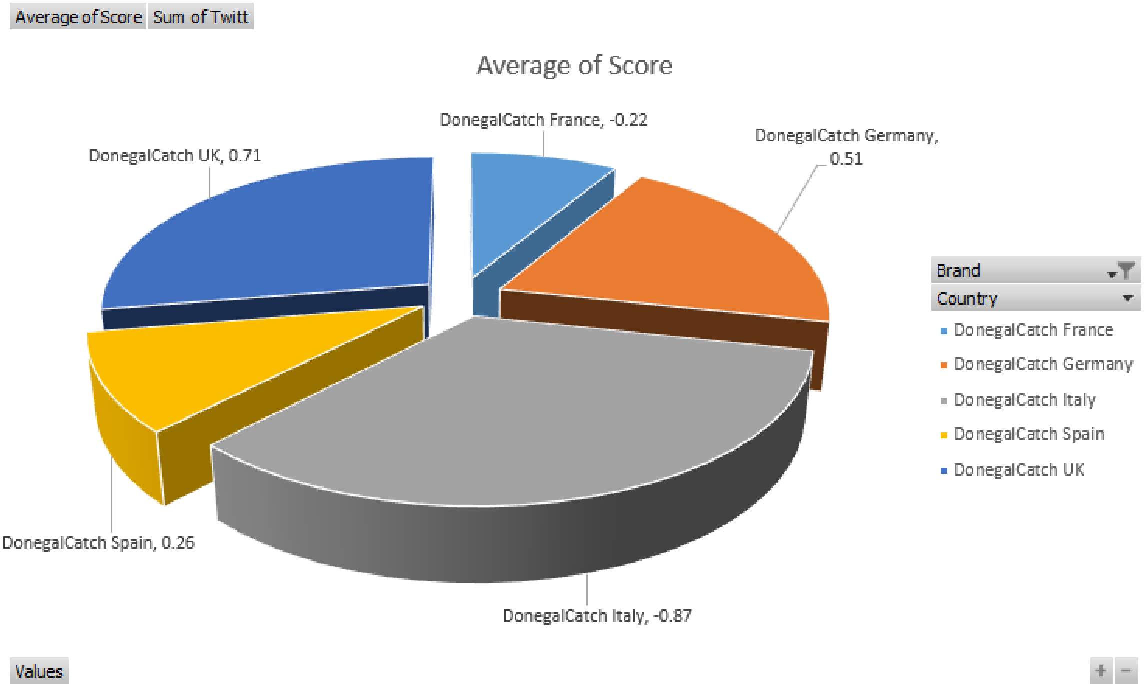
The data analysis applied the KDD principle throughout its steps in the implementation. Following the steps outlined assures a system that complies to the norms and takes into consideration all possible issues.

Going from this…



```
:"RT @Lena_00_jasmine: VAMPS US TOUR\nRichmond,VA\u3000May,4 SETLIST https:\/\/t.co\/BbOVaIZeGQ","source":"\u003ca hre
:"#jobs4u #jobs Dental Hygienist https:\/\/t.co\/TJypm7SOVR #RVA #richmond #VA #RIC","source":"\u003ca href=\"https:\/
:"@shondarhimes @Dove if you were to ask my child I'm Merideth with denny duQuette disease. Must say you've been writ\
:"Dustin Martin 1st Goal for Richmond \n\n#AFL #AFLDogsTigers","source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"r
:"RT @Lena_00_jasmine: VAMPS US TOUR\nRichmond,VA\u3000May,4 SETLIST https:\/\/t.co\/BbOVaIZeGQ","source":"\u003ca hre
:"@ChrisEriksen8 Never mind, Ole &amp; Steen are opening a Danish bakery in Richmond. So that's good news \ud83d\ude01
:"I'm craving Denny's big time rn","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\
:"Richmond Berks https:\/\/t.co\/dzoHcuQcvx","source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwi
:"Twitter mentions for Richmond Park: https:\/\/t.co\/vsP1QQ5tmh - RT @arthurascii Cycling in Richmond Park just got m
:"RT @KAZ1011_K_VAMPS: VAMPS US 5\/4 Richmond \n\u30bb\u30af\u30d6\u30e9 \u8eab\u3092\u4e57\u308a\u51fa\u3059\u795e\u2
```

to this …



The tweet itself represents a string of a set of key-value objects that are hard to read and understand much about it. It requires previous knowledge regarding the key’s and their significance, regarding the medium in which is presented and the relevance of social media events.

After attributing a value (from -1 to 1) representing the opinion/sentiment of a text in a tweet, this can have little significance or impact for a particular brand. What makes this type of research valuable is the amount of data that is being analyzed and the specific way in which tweets are being categorized. In this research paper tweets are being grouped by location, country code. This gives an indication as to how the brand is being perceived in that country by its customers.

6 Future development

The science of NLP (Neuro Linguistic Programming) in collaboration with software development and machine learning tools in data science makes the times we are living in even more exciting. Training the machine to interpret natural language and compile emotions is one field in which a lot more research has to be done.

Decrypting human emotions is something that we us humans have yet to learn more about. Teaching a machine to pick on attitude from lines of texts is ambitions and requires more research in order to be perfected.

At this point, the process of performing sentiment analysis lacks still from the quality and interpretation of language by the human brain.

Brands on social media are gaining field and importance in the last decade. Their online presence and focus on retaining and gaining customers online will more likely continue in the next years with an increase focus on researching and improving their company’s products or services.

Online data is nowadays the new way of making more out of science in terms of business and long-term impact, corporate wise. Budgets are being transformed to accommodate data research projects and efforts to interpret large amount of data is beginning to become a focus worldwide.

7 References

"Company | About". About.twitter.com. N.p., 2016. Web. 01 Dec. 2016.

"An Introduction To Text Mining Using Twitter Streaming API And Python // Adil Moujahid // Data Analytics And More". Adilmoujahid.com. N.p., 2016. Web. 11 Dec. 2016.

irishexporters.ie. (2017). TOP 150 Born in Ireland 2016. [online] Available at: http://www.irishexporters.ie/wp-content/uploads/2016/10/Top150_Born_in_Ireland_2016_Final.pdf [Accessed 10 Dec. 2016].

Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining,* AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

Docs.python.org. (2017). 26.4. unittest — Unit testing framework — Python 3.6.1 documentation. [online] Available at: https://docs.python.org/3/library/unittest.html#unittest.main [Accessed 10 March 2017]

Wave.webaim.org. (2017). WAVE Web Accessibility Tool. [online] Available at: http://wave.webaim.org/ [Accessed 10 Apr. 2017].

C9.io. (2017). Cloud9 IDE Support. [online] Available at: https://c9.io/support [Accessed 10 Jan. 2017].

Brownlee, J. (2017). What is Data Mining and KDD - Machine Learning Mastery. [online] Machine Learning Mastery. Available at: http://machinelearningmastery.com/what-is-data-mining-and-kdd/ [Accessed 01 Apr. 2017].

Textblob.readthedocs.io. (2017). API Reference — TextBlob 0.12.0 documentation. [online] Available at: http://textblob.readthedocs.io/en/dev/api_reference.html [Accessed 18 Mar. 2017].

8 Appendix

8.1 Project Proposal

Objectives

The main goal of this project is to perform an analysis on how top Irish Brands are perceived worldwide. In order to reach this goal, this research will utilize a specific social media platform, Twitter.

The sentiment analysis performed on tweets, for each top Irish brand in particular, will lead to a worldwide map showing the level of engagement, online presence and type of emotions (different levels for positive/negative) for each country.

In this moment in time, there are no other similar studies available for public use. What differentiates this project from existing sentiment analysis tools that offer reports for brands is the fact that it is focus on the market outside Ireland, offering cross-language, cross-cultural observations.

The finding may be of use to further international marketing studies and may support individual brand's strategic management decision making process.

Background

There are a number of well-known top Irish brands that have a great deal of exposure on the international markets. Brands such as Guinness, Kerrygold, Tayto, Barry's Tea and many others are exporting their products worldwide and benefit from an increasingly online presence due to their customer's reviews and comments.

In order to complete a specific research for each brand in particular, this project is going to focus on top 10 brands. The methodology behind choosing those brands will be further described in detail, in the next project deliverable. In order for this research to be non-bias and objective, the mechanism behind rating the international Irish brands will be based on public figures regarding exports from Irish Economy and Public Finances (National Treasury Management Agency).

After deciding on the list of top brands considered in this research, the work will focus on gathering as much data as possible. This will require a programmatic access to read and interpret streams of data from Twitter, based on a pre-defined algorithm. This stage requires a longer period of time (2-3 months) in order to gather data for a representative qualitative research.

The analysis performed on each selected brand will follow the same rules for all, but will differentiate in terms of emotions-keywords. This process will be defined according to the pre-analysis (manually) of messages on each brand.

Beside analysing each selected brand separately and displaying the results individually, this project will also involve gathering information regarding the location of the tweet (country wise). This information being amount the most important one in this research, the plan is to incorporate more than one method of finding the location of that tweet (language detection tools, profile bio of the user). This is especially important in order to geo-locate tweets that do not have a geo-location tag. This feature is optional and it is considered that few Twitter users are sharing their location coordinates by default.

Technical Approach

In order to monitor and process Tweets in real-time, this project will use The Streaming APIs resources. A proper implementation of a streaming client needs to be in place in order to gather data.

Connecting to the streaming API requires a continuous HTTP connection as the data flow will be on-going. After the server opens the streaming connection and Twitter accepts the connection, the Tweets are streamed as they occur. The application will receive streamed tweets, perform any processing required and store the results in a data store. The connection closes after the process is terminated, at request or due to technical fault.

The algorithm used for parsing the data is an important part of the sentiment analysis process. Online there are numerous information about this type of analysis and basic algorithms for parsing and interpreting data will form the basis in this step. Additionally, specific parameters will be added for each brand and a scaling mechanism of the emotions (negative/positive) will be defined.

Special resources required

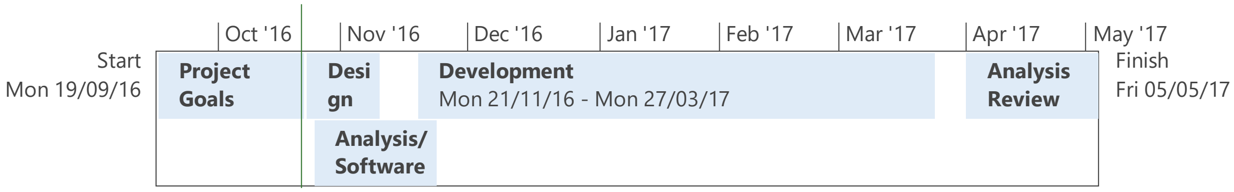
This research requires big amount of data to be stored in a data warehouse. Efforts will be made to utilize free available space in this regards.

Research of the options available needs to be considered and decisions in regards to the provider needs to follow. In the circumstances that the service of data storage will require fees, this information will be detailed in a future budget section of the project.

Another special resource required is considered to be the cross-language support that will derive from the data analyzed. This might involve the use of a translating tool. Research is needed and will follow with conclusions.

8.2 Project Plan

Using Microsoft Project, the timeline for the plan is as follows:



For the following project plan:

Task Name	Duration	Start	Finish
Software Project	**165 days**	**Mon 19/09/16**	**Fri 05/05/17**
Project Goals	**26 days**	**Mon 19/09/16**	**Mon 24/10/16**
Determine project idea	96 hrs	Mon 19/09/16	Tue 04/10/16
Deliver project pitch	1 day	Wed 05/10/16	Wed 05/10/16
Define project proposal	11 days	Thu 06/10/16	Thu 20/10/16
Initial research	6 days	Mon 17/10/16	Mon 24/10/16
Elaborate project plan	3 days	Thu 20/10/16	Mon 24/10/16
Design	**14 days**	**Tue 25/10/16**	**Fri 11/11/16**
Define methodology for choosing the brands	4 days	Tue 25/10/16	Fri 28/10/16
Research tweets location option	5 days	Mon 31/10/16	Fri 04/11/16
Define specific algorithm for each brand	5 days	Mon 07/11/16	Fri 11/11/16
Analysis/Software Requirements	**22 days**	**Thu 27/10/16**	**Fri 25/11/16**
Conduct needs analysis	3 days	Mon 14/11/16	Wed 16/11/16
Research translation tools	3 days	Thu 17/11/16	Mon 21/11/16
Define analysis algorithm	7 days	Thu 17/11/16	Fri 25/11/16
Research data storage options	5 days	Mon 21/11/16	Fri 25/11/16
Development	**90 days**	**Mon 21/11/16**	**Mon 27/03/17**
Develop application - coding	15 days	Mon 21/11/16	Fri 09/12/16
Testing	5 days	Mon 12/12/16	Fri 16/12/16
Streaming data	55 days	Mon 09/01/17	Fri 24/03/17
Development complete	0 days	Mon 27/03/17	Mon 27/03/17
Analysis Review	**25 days**	**Mon 03/04/17**	**Fri 05/05/17**
Document lessons learned	5 days	Mon 03/04/17	Fri 07/04/17

Design the maps	15 days	Mon 03/04/17	Fri 21/04/17
Conclusions	10 days	Mon 24/04/17	Fri 05/05/17

Technical Details

The application needs to be able to connect to the Streaming APIs. The streaming process that receives the tweets and performs additional operations (parsing/filtering/aggregation) before storing the result will be written in R Studio.

The motivation behind choosing R Studio as a tool to analyze data is that it incorporates a cross-platform environment for R (statistical language) and it provides also graphical workspace, tab-completion and full-featured text editor.

The graphics outlining the finding from the research will be using outsource software and will incorporate an easy to use/ easy to interpret (color emphasis on each country, depending on the rating of the emotions).

Evaluation

The system will be evaluated from a marketing perspective using comparison tools with existing social media sentiment analysis tools.

The project will be considered successful at the completion of the tasks outlined and by presenting all the data gathered in a user-friendly manner, easy to interpret data. The data output will be incorporate the finding of this research paper and also conclusions revealed.

8.3 Monthly Journals

Reflective Journal
Student name: Oana Cozma
Programme: BSc in Computing
Month: September

My Achievements

This month, I was able to pick and define the project idea. After considering other project ideas, I decided that I want to do a sentiment analysis on tweets on top Irish brands. What differentiates my research from what you can find online is that the emotions (positive, negative) linked to the brands will be mapped outside Ireland. To choose the brands that this research will showcase, I am considering several methods of classification for all existing international Irish brands.
My contributions to the projects included researching existing studies regarding sentiment analysis on Twitter, geolocation of tweets (country), ways of classifying top brands for this research and technologies to be considered.

My Reflection

After the Project Pitch, it felt good to have my idea approved by the judges. They made me aware of the fact that it may be possible that only 10% or less of the tweets have location services turned on by the user, this could be an impediment in gathering enough data in some cases, on a shorter period of time. I plan to access the streaming data provided by Twitter and this will allow me to do a qualitative analysis.
This research will also involve cross-language challenges and cross-cultural observations.

Intended Changes

Next month, I plan to do more research regarding streaming data from Twitter API and to come up with a classification of brands in order to start working on setting up specific keywords for the analysis.
I realised that I need to start streaming tweets sooner in order to gather enough data to complete a relevant search.

Supervisor Meetings

Date of Meeting: 5th October
Items discussed: Project Pitch
Action Items: Approved

Reflective Journal
Student name: Oana Cozma

Programme: BSc in Computing
Month: December

My Achievements

This month, I managed to get the first part of the Technical report out of the way. The presentation coming up and a prototype for the website is being put in place.

My Reflection is that is hard to estimate how will I manage the data with the little information I have regarding streaming tweets and sentiment analysis.

Next month, I will be focusing more on the back end part of the project as this looks to be the most challenging aspect of the project.

Supervisor Meetings

Date of Meeting: December
Items discussed: Mid term presentation preparation
Action Items: Approved

Reflective Journal
Student name: Oana Cozma
Programme: BSc in Computing
Month: February

My Achievements

This month, I concluded the methodology behind chosen corporates and their brands and I did further research into twitter data and their API services.

My Reflection Is that data won't be broadly available and that my expectation regarding exceeding storage limits are over estimating the reality. In this new light, I worry about not having enough relevant data to conduct the research on.

Intended Changes: start the streaming API process asap so enough data will be gathered.

Supervisor Meetings

Date of Meeting: February
Items discussed: Streaming, data input
Action Items: Approved

Reflective Journal
Student name: Oana Cozma
Programme: BSc in Computing
Month: March

My Achievements

This month, I managed to get the connection available live, set up the technical environment and research the methods of doing sentiment analysis in python.

My Reflection Is that data won’t be broadly available and that my expectation regarding exceeding storage limits are over estimating the reality. In this new light, I worry about not having enough relevant data to conduct the research on.

Intended Changes: geo location tag is becoming redundant as there is mostly of a NULL value and represents the coordinates of the location, therefor harder to manipulate when dealing with large amount of data. Location tag is more likely to deliver that information and the focus will be on that from now on.

Supervisor Meetings

Date of Meeting: March
Items discussed: Streaming, data output
Action Items: Approved