2020

Twitter Analytics

DATA ANALYTICS AT SCALE

KIRTI KHADE || S4573313

Table of contents

1. Abstract

This section provides an abstract summary of the report.

In this project, twitter data will be analyzed using Hortonworks Data Platform (HDP), which is an Apache Hadoop distribution system. This system powers real-time analytics and helps to the accelerate decision-making process. The dataset that will be used for this project is twitter data for 6 months between the period July 2014 and December 2014. The project will detail two aspects of twitter data:

- ***Comparison of Tweets from the United Kingdom & Australian government***: This section will compare the tweets between the two government organizations in the 6 months.
- ***Comparing the tweets from the United Kingdom & Australian government on days with maximum tweets***: On the days when there are maximum number of tweets, an analysis will be performed to understand the significant events that took place on those days.

2. Introduction

This section briefly describes the general area of big data analytics along with highlighting the need for distributed system solutions with an example of why these solutions are needed.

The recent increase in data velocity, volume, variety, and veracity led to the development of big data analytics. Each human generates about 1.7 megabytes of data every second or 2.5 quintillion bytes of data every day. Technologies that can store and analyze such large amount of data need to be employed. In the past, data was stored at one location, but as the data got larger, the systems got slower. Today, data is distributed across different databases, which has enabled the storage and analysis of massive amounts of data. Technologies such as Hadoop, Pig, Spark, and Cloud computing not only systematically store data but also make sure that analyzing this data is efficient. This development has led to enormous progress in the field of machine learning and artificial intelligence. From identifying a person from a picture, to speech-to-text conversion, to language conversion, to image to text conversion - this technology has come a long way (Bernard, 2016, pp. 1-4).

Distributed system solutions and big data analytics are required for any company that collects and analyses a massive amount of data. For instance; **Twitter,** is a micro-blogging site, with about 330 million active users currently. This is the world's second most popular website with over 500 million posts on this site daily. The primary source of revenue for this website is promotional tweets. Companies pay twitter to appear in the feeds of the twitter users who might be interested in the products and services offered by the company. To identify these target users, companies have to analyze a vast amount of data using distributed system solution (Bernard, 2016, pp. 261-264).

3. Dataset Analytics

This section provides a brief description of the focus area, dataset, tool & techniques, and pre-processing. It also includes an overview of the results of the analysis.

This project will focus on analyzing the tweets from Australian(AU) and the United Kingdom(UK) government. This analysis is useful to political analysts, news agencies, capital investors, entrepreneurs, etc. This analysis can be used to understand the geo-political environment of a region and the affairs that the government is invested in and interested in.

3. 1 Dataset

From the twitter API which covers 1% random tweets of the entire twitter stream for 6 months, I will be analyzing the users having URLs as "gov.au" or "gov.uk" to get the users from the government organization from both the countries.

3.2. Tools and Technique

The tools and techniques used for this project are the following:

1. The analysis will be done using Hortonworks Data Platform (HDP) cluster which is an Apache Hadoop distribution system
2. To query this system pyspark was used, which is the python API to support Apache spark
3. Apache spark was used because it uses resilient distributed dataset(RDD) and intermediate data saving techniques for processing, which makes it more efficient and scalable as compared to Pig
4. Additionally, python was used for visualization of data and word cloud analysis
5. AWS s3 bucket was used to transfer data from cloud computer to local desktop

3.3 Pre-processing of data

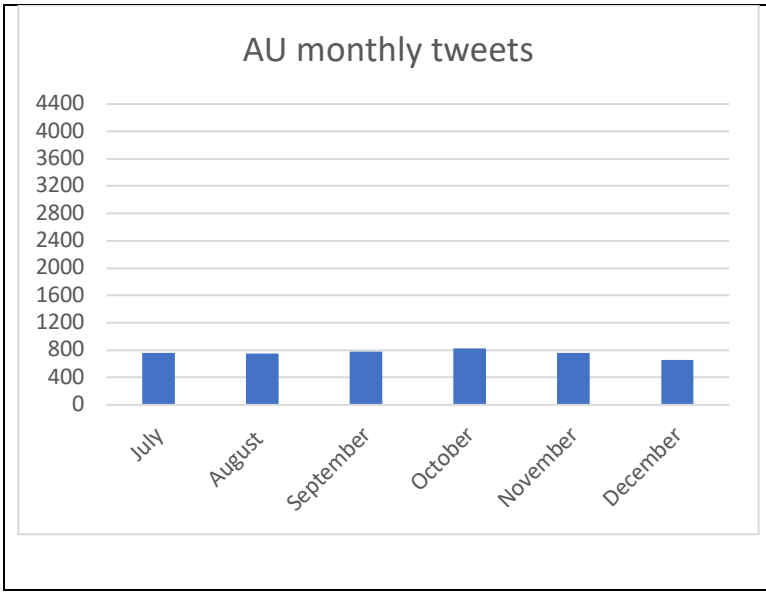
The following pre-processing step that was performed in the data:

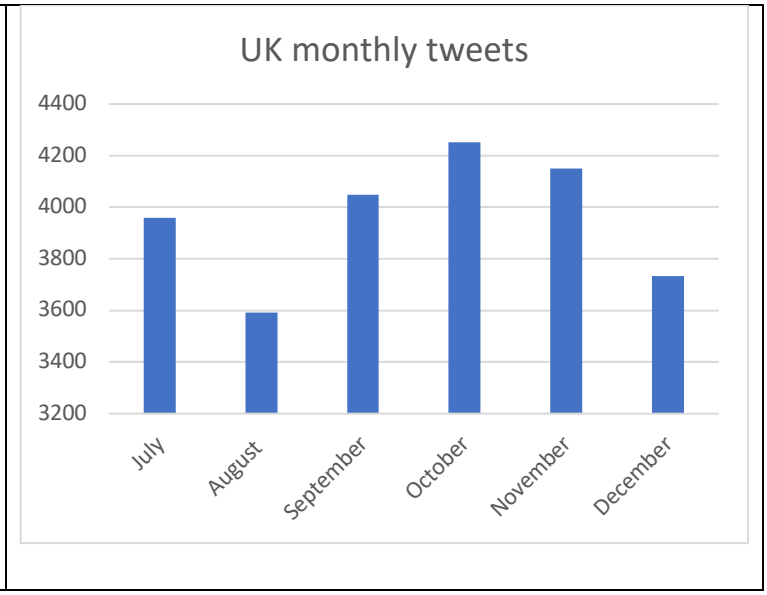
1. ***Selecting required columns***: For this analysis the following columns where chosen 'id', 'entities', 'timestamp ms', 'created at', 'user', 'text'
2. ***Selecting required users:*** Selecting users with url as 'gov.au' and 'gov.uk'
3. ***Timestamp ms/ Created at:*** Unix timestamp_ms into the human readable by removing the last three digits. For the month of July and October, no timestamp_ms column could be found. Hence, "created_at" was used. All the tweets with no timestamp/created_at were not counted.
4. ***URL:*** To compare the URL between the two countries, the URLs were manually edited to make sure that the URLs are understandable and are common in both the countries. For instance, all the cities and states of these countries are tagged under "cites/states"
5. ***Followers vs Status:*** For clearly understanding the relationship between the two attributes, a log transform has been applied to both these entities.

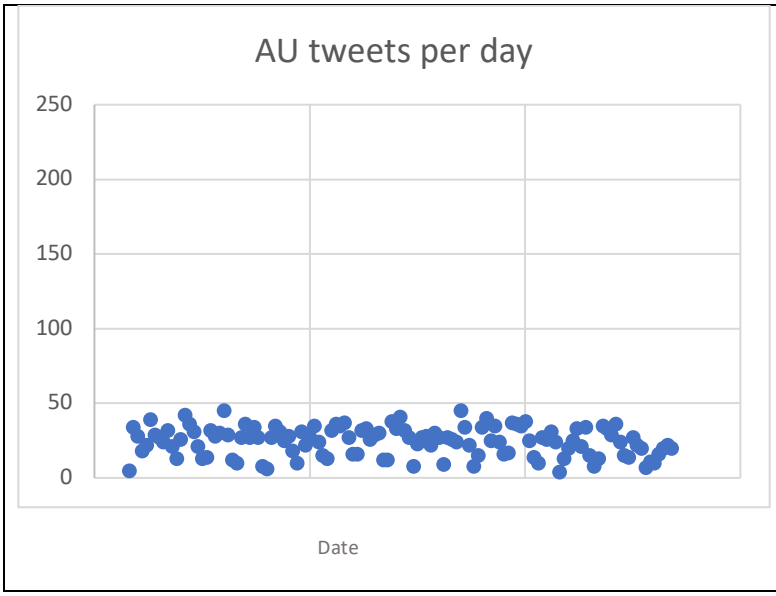
3.4 Six-month Analysis

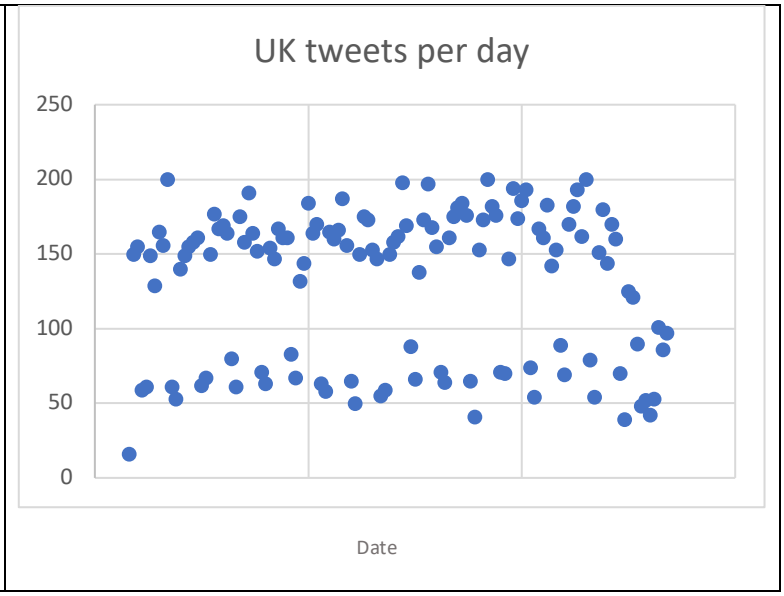
3.4.1 Trend of number of tweets

The number of tweets by UK government departments and organisations are about 6 times more as compared to AU. This is possibly because, in 2014, the number of government employee in the UK(7million) was 2 times as that of Australia(4 million). The population of UK is also 3 times that of AU.









3.4.2 Most used Hashtags

The most popular hashtags in Australian government organizations are:

1. ***#worldelephantday***: World elephant day is celebrated on 12th August every year (worldelephantday.org, 2012)
2. ***#economicdiplomacy***: The foreign minister of Australian (2013-2018) - Julie Bishop first discussed the economic diplomacy in September 2014, Sydney (exportfinance, 2014)

The most popular hashtags in UK government organisations are:

1. ***#elchiringuitodelmundial***: FIFA worldcup that took place from 12th June – 13th July, attracted this hashtag. This hashtag means "the bar of the world cup", used for tagging a bar where citizens can enjoy the WorldCup
2. ***#renfrew***: Renfrew is a town near Scotland, UK. In July 2014, there were reports of a massive blaze being starting deliberately (bbc, 2014)

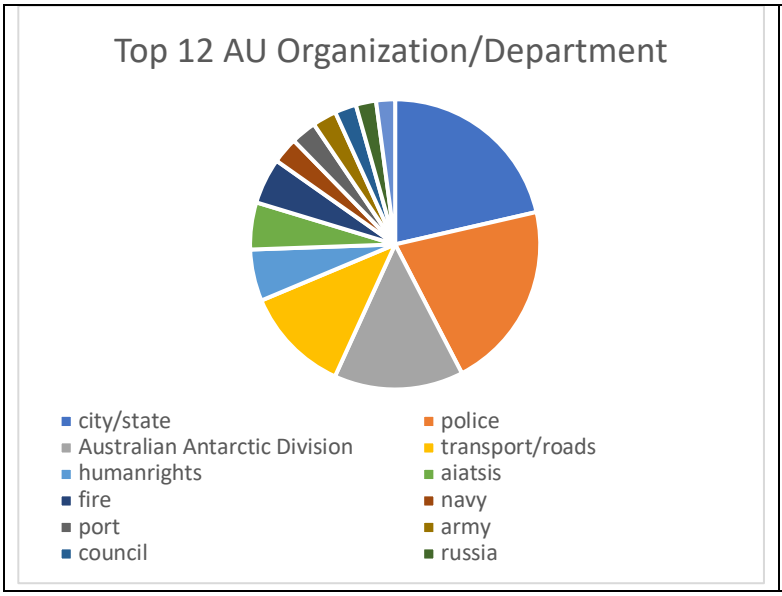
Other notable hashtags: ***Naidoc week*** (AU, 6-13 July) ***MH17*** *(AU,* 17 July 2014), **Ebola** (the UK, 29 December 2014) ***Royal*** (UK, always)

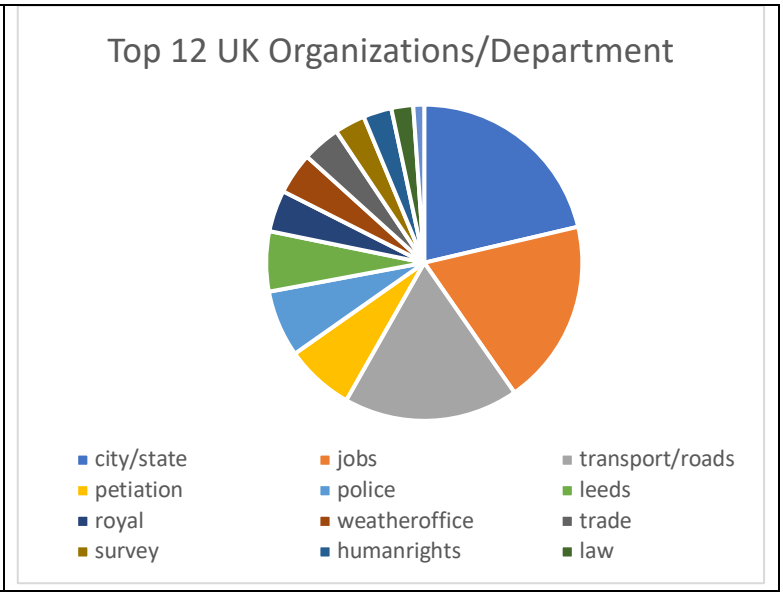
50 most used hashtags by AU government

50 most used hashtags by UK government

3.4.3 Which organization tweeted the most

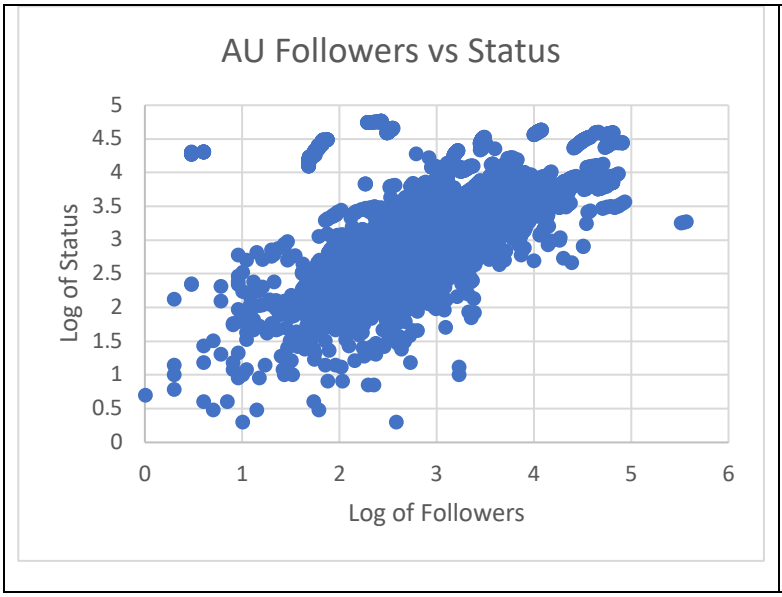
This field was calculated using the top 20 organizations that have tweeted the most. In Australia, local government from various cities/states tweeted the most, followed by police department and then the Australian Antarctic division. While in the UK, city/state local government, were followed by the department for employment.

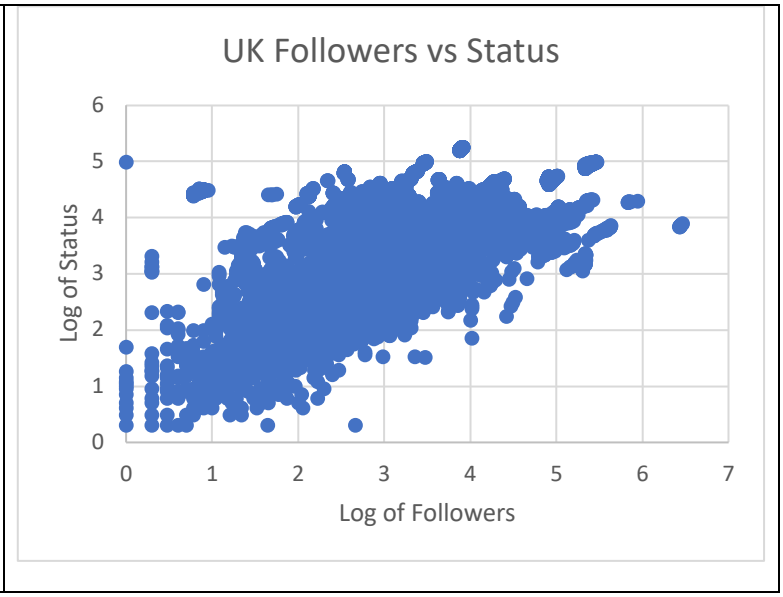




3.4.4 Followers vs. Status

In both countries, it can be seen that the number of followers increases with the increase in the number of status from the users. The number of followers and status shared by the organizations is higher in AU as compared to the UK. This is possibly due to differences in population(in 2014 UK has a population of 64.35 million and AU has 23.50).





3.4.5 Day on the most tweets

The days with the most tweets from Australia government organization (about 45 tweets) had the following events:

1. On ***2014-09-18***, the major event that took place on this day was the 9th meeting of the G20 heads was held in Brisbane. Which is why, there are prominent "g20", "Brisbane" and "Australia" hashtags (wikipedia, 2014)
2. On ***2014-09-25***, a Kookykidz market in Australia(Perth) attracted a crowd, and the movie "Frozen" was showcased. Which is why, hashtags like "kookykidz," "frozen," "kidzmarket", "movienight" (news.com.au, 2014)

The days with the most tweets from UK government organization (about 200 tweets) had the following events:

1. On ***2014-09-05***, the NATO summit took place in Wales, UK. This day also saw hashtags about Syria, Ukraine, and Iraq indicating the discussion on these countries (gov.uk, 2014)
2. On ***2014-11-19***, the Annual Survey of hours and earning declared the numbers for the year 2014. Hashtags like "job," "vacancies," "art jobs," "media jobs" were seen on this day (ons.gov.uk, 2014)
3. On ***2014-12-12***, no major event seems to have taken place. Hashtag about Christmas eve (#xmaseve) and shop local were majorly shared

4. Discussion and conclusions of the analysis

This section will summarise and discuss the main findings of the analysis and the main lessons learned.

4.1 Summary of the results

This analysis assumes that the 1% random samples of the twitter stream used for this project is representative of the entire dataset. From the comparison of the tweets from the government of the UK and Australia, the following conclusions can be drawn:

1. ***Frequency of tweets***: On a daily basis, about 25 tweets are posted by Australian government organizations/departments while this number is 130 for the UK. This difference is probably due to the difference in number of employees in UK and AU government.
2. ***Most used hashtags***: The most used hashtags showcase the significant events in both the countries and major changes. For instance, "**#*economicdiplomacy***" in Australia signifies a new model of economic development, and "***elchiringuitodelmundial***" in the UK signifies a major event i.e. FIFIA WorldCup, and the tendency of government to talk about the restaurants as compared to the world cup itself. Another example could be the fact that an event "**#kookykidz**" could generate so much interest from the government of Australia.
3. ***Most tweets by the organisation***: Although the kinds of organizations are same in almost every country, the difference in frequency of tweets showcase what is essential for which country at a particular time. For instance, the department of employment is much more active in the UK as compared to AU.
4. ***Followers vs. status***: Both the country's data has proved that the users with more the status count have more followers.

4.2 Learning for a data analyst

From this project, the following learning can be drawn:

1. With this analysis, one can understand the significant areas of the focus of the government, issues faced by government, citizens' behavior, and geopolitical environment. For instance, the UK government is interested in creating and providing more jobs. While, AU government has a keen interest in environmental issues as #worldelephantday was one of the major hashtags, as well as the Australian Antarctic division, was one of the most frequent tweeters.
2. In the past, this task may seem impossible due to the high amount of volume and variety of data. But today, with the help of distributed system solution I was able to perform this analysis in over 15 days.
3. With this massive amount of data in hand, it is vital to convert the data into insights, and insights into action, the distributed system solution provides an efficient way of analyzing data. This technology has enabled me to find the patterns for a specific domain in such a large data setting

Appendix

1. Wordcount

The total number of word count, excluding the abstract, headings, heading description, references and table of content is 1499

2. Codes

2.1 Pyspark

Pyspark was used to query the big data for the government(the United Kingdom and Australian), the following code was used with a change in filter for both the government.

```
from __future__ import print_function
import pyspark
from pyspark import SparkContext
from pyspark.sql import SparkSession
from pyspark.sql import SQLContext, DataFrame
from pyspark.sql.functions import desc, mean
import pandas as pd
from pyspark import Row
from pyspark.sql import functions as F
from pyspark.sql.functions import lower, col, explode
from pyspark.sql.functions import udf, col, lower, udf, regexp_replace
from pyspark.sql.functions import col, split
from pyspark.sql.types import StructType, IntegerType, StringType

df = sqlContext.read.json('/data/ProjectDataset/statuses.log.2014-*.gz')

# Filtering for the data needed
df = df.select('id' , 'entities' ,  'timestamp_ms' , 'user',  'text', 'retweet_count', ‘created_at’)

# Filtering for the valid users only
df_user_gov = df.filter(df.user.url.like('%gov.au%'))
type_gov = 'gov.au'
save_file = 'gov_au'



# Get the timestamp
df_timestamp = df_user_gov.select("timestamp_ms")
df_timestamp = df_timestamp.withColumn("date_short", F.substring(col("timestamp_ms"),1,10))
df_timestamp                                                                                                                    =
df_timestamp.withColumn('date_again',func.from_unixtime('date_short').cast(stypes.DateType()))
df_timestamp = df_timestamp.groupby('date_again').count()
df_timestamp_pd = df_timestamp.toPandas()
df_timestamp_pd.to_csv('/home/s4573313/trend_for_number_tweets_all' + save_file +  '.csv')



# Get the hashtags
df_hashtag = df_user_gov.select('id' , 'entities')
# Get the hastags
```

```
df_hashtag_explode = df_hashtag.withColumn('entities.hashtags.text', explode('entities.hashtags.text')).select('id' ,  'entities.hashtags.text')
# Explode w.r.t text in hashtags
df_hashtag_explode = df_hashtag_explode.withColumn('text', explode('text'))
# Convert everything to lower
df_hashtag_explode = df_hashtag_explode.select( 'id',lower(col('text')).alias('text'))
# Count of hashtags
df_hashtag_explode_ = df_hashtag_explode.groupby('text').count()
# hashtags with maximum tweets
df_hashtag_explode_ = df_hashtag_explode_.sort(desc('count')).head(20000)
# Convert to Pandas
df_hashtag_explode_pd = pd.DataFrame(df_hashtag_explode_)
df_hashtag_explode_pd.to_csv('/home/s4573313/hashtag_count_all' + save_file +  '.csv')


# Get the URLS
df_url_split = df_user_gov_uk.select('user.screen_name', pyspark.sql.functions.split(df_user_gov['user.url'], '/').getItem(2).alias('url'))
df_url_split = df_url_split.withColumn('url', F.regexp_replace('url','www.',''))
df_url_split = df_url_split.withColumn('url', F.regexp_replace('url', type_gov,''))
# Convert everything to lower
df_url_split = df_url_split.select('screen_name',lower(col('url')).alias('url'))
df_url_split_count = df_url_split.groupby('url').count().sort(desc('count')).head(20)
df_url_split_count_pd = pd.DataFrame(df_url_split_count)
df_url_split_count_pd.to_csv('/home/s4573313/url_count_'+ save_file + '.csv')


# Followers vs stat
df_user_stat = df_user_gov.select('id','user.screen_name' ,'retweet_count','user.location' , 'user.followers_count' , 'user.statuses_count').distinct()
df_follower_stat_sc = df_user_stat.groupby('followers_count' , 'statuses_count' , 'retweet_count').count()
df_follower_stat_sc_pd = df_follower_stat_sc.toPandas()
df_follower_stat_sc_pd.to_csv('/home/s4573313/followers_sc_count_'+ save_file + '.csv')

# Get the text
df_user_url_ = df_user_gov.select('text').distinct()
df_user_url_pd = df_user_url_.toPandas()
df_user_url_pd.to_csv('text.csv')
```

For the months of July and august, where the timestamp_ms column was not present, created_at was used:

```
df = sqlContext.read.json('/data/ProjectDataset/statuses.log.2014-07-*.gz')
split_col = split(df['created_at'], ' ')
df = df.withColumn(month, split_col.getItem(1))
df = df.withColumn('date', split_col.getItem(2))
df = df.withColumn('year', split_col.getItem(5))
# concatnate
df=df.withColumn("date_agaon",F.concat(col("year"), F.lit("/"), col("month"),F.lit("/"), col("date"),F.lit(" ")))

# Get the number of tweets per day for UK government
df_user_gov_uk = df.filter(df.user.url.like('%gov.uk%'))
df_uk = df_user_gov_uk.groupby('date_time_new').count()
```

```
df_uk_pd = df_uk.toPandas()
df_uk_pd.to_csv('/home/s4573313/uk_07.csv')


# Get the number of tweets per day for AU government
df_user_gov_au = df.filter(df.user.url.like('%gov.au%'))
df_au = df_user_gov_au.groupby('date_time_new').count()
df_au_pd = df_au.toPandas()
df_au_pd.to_csv('/home/s4573313/au_07.csv')
```

For the days with the maximum tweets, the following code was used to get the hashtags.

```
 df = sqlContext.read.json('/data/ProjectDataset/statuses.log.2014-*.gz')

# Filtering for the data needed
df = df.select('id' , 'entities' ,  'timestamp_ms' , 'user',  'text', 'retweet_count', 'created_at')

# Filtering for the valid users only
df_user_gov = df.filter(df.user.url.like('%gov.au%'))
type_gov = 'gov.au'
save_file = 'gov_au'

# Get the hashtags
df_hashtag = df_user_gov.select('id' , 'entities')
# Get the hastags
df_hashtag_explode = df_hashtag.withColumn('entities.hashtags.text',
explode('entities.hashtags.text')).select('id' ,  'entities.hashtags.text')
# Explode w.r.t text in hashtags
df_hashtag_explode = df_hashtag_explode.withColumn('text', explode('text'))
# Convert everything to lower
df_hashtag_explode = df_hashtag_explode.select( 'id',lower(col('text')).alias('text'))
# Count of hashtags
df_hashtag_explode_ = df_hashtag_explode.groupby('text').count()
# hashtags with maximum tweets
df_hashtag_explode_ = df_hashtag_explode_.sort(desc('count')).head(20000)
# Convert to Pandas
df_hashtag_explode_pd = pd.DataFrame(df_hashtag_explode_)
df_hashtag_explode_pd.to_csv('/home/s4573313/dfau01'  +  '.csv')
```

2.2 Python

The following code was used for visulisation and word cloud analysis from the data acquired from the pyspark

```
import pandas as pd
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt
import pandas as pd
%matplotlib inline

# Trend
df_trend = pd.read_csv('trend_for_number_tweets_allgov_au_a11 (1) 2.csv')
df_trend["Month"] = df_trend["date_again"].str.split("-", n = 2, expand = True)[1]
df_trend_monthly = pd.DataFrame(df_trend.groupby(['Month'])['count'].agg('sum')).reset_index()
fig = plt.figure()
ax = fig.add_axes([0,0,1,1])
months = ['August', 'September', 'October', 'November', 'December']
```

```
tweets = [441,4048,4252,4150,3733]
ax.bar(months,tweets)
plt.title("Tweet volume by each month", fontsize = 20)
plt.xlabel("Months")
plt.ylabel("Number of tweets")
plt.show()
plt.show()

# convert date to datetime format
df_trend['date_again'] = pd.to_datetime(df_trend['date_again'], errors='raise', dayfirst=False, yearfirst=False, utc=None, format=None, exact=True, unit=None, infer_datetime_format=False, origin='unix', cache=True)
# sort w.r.t datate time format
df_trend.sort_values(by=['date_again'], inplace=True)
# Replace na with zero
df_trend["count"] = df_trend["count"].fillna(0)
# change date to string again
df_trend["date_again"] = df_trend["date_again"].astype(str)
df_trend.head()
df_trend = df_trend[df_trend['date_again'] != 'NaT']
import matplotlib.pyplot as plt
%matplotlib inline
plt.scatter(df_trend['date_again'] , df_trend['count'] )
plt.title("Tweet volume at each date", fontsize = 20)
plt.ylabel("Count of tweets")
plt.show()


# Hashtags
df_hashtag = pd.read_csv('hashtag_count_allgov_au_a11.csv')
df_hashtag = df_hashtag.groupby('0').sum().reset_index()
df_hashtag = df_hashtag.sort_values(by = [1 , 0], axis=1,  ascending=False)
df_hashtag = df_hashtag.reset_index()
df_hashtag = df_hashtag.iloc[ :, 1:3 ]
df_hashtag = df_hashtag.reset_index()
df_hashtag.columns =  [ 'index' , 'a', 'b']
df_hashtag = df_hashtag.sort_values(by = 'b', axis=0,  ascending=False).head(50)

for i in range(df_hashtag.shape[0]):
   df_hashtag.iloc[ i,1] = (df_hashtag.iloc[ i,1] + " ") *  df_hashtag.iloc[ i,2]

comment_words = ''
# iterate through the csv file
for val in df_hashtag['a']:
   # typecaste each val to string
   val = str(val)
   # split the value
   tokens = val.split()
   # Converts each token into lowercase
   for i in range(len(tokens)):
      tokens[i] = tokens[i].lower()
   comment_words += " ".join(tokens)+" "
wordcloud = WordCloud(width = 800, height = 800,
            background_color ='white',
            min_font_size = 10).generate(comment_words)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
```

```
plt.tight_layout(pad = 0)

# URL
df_url = pd.read_csv('au_url.csv')
labels = df_url['departments']
sizes =  df_url['count']
fig1, ax1 = plt.subplots()
ax1.pie(sizes, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90)
ax1.axis('equal')
plt.show()
```

3. References

bbc. (2014, July 11). *bbc.com*. Retrieved from bbc.com: https://www.bbc.com/news/uk-scotland-glasgow-west-28258577

Bernard, M. (2016). *Big data in Practice* . West Sussex, United Kingdon: John Wiley and Sons Ltd.

exportfinance. (2014, September 26). *exportfinance.gov.au*. Retrieved from exportfinance.gov.au: https://www.exportfinance.gov.au/resources-news/news-events/latest-news/2014/september/australia-s-economic-diplomacy/

gov.uk. (2014, September 05). *.gov.uk*. Retrieved from .gov.uk: https://www.gov.uk/government/news/nato-summit-wales-2014-live

news.com.au. (2014, September 26). *news.com.au*. Retrieved from news.com.au: https://www.news.com.au/national/south-australia/more-than-1000-are-tipped-to-visit-port-adelaide-for-kooky-kidz-market-and-screening-of-disneys-frozen-at-harts-mill/news-story/1003a5aa1d991ef7861c6a6fca8a047f

ons.gov.uk. (2014, November 19). *ons.gov.uk*. Retrieved from ons.gov.uk: https://www.ons.gov.uk/surveys/informationforbusinesses/businesssurveys/annualsurveyofhoursandearningsashe

wikipedia. (2014, September 19). *wikipedia.org*. Retrieved from wikipedia.org: https://en.wikipedia.org/wiki/2014_G20_Brisbane_summit

worldelephantday.org. (2012, August 12). *https://worldelephantday.org/about*. Retrieved from https://worldelephantday.org/about: https://worldelephantday.org/about

https://www.datacamp.com/community/tutorials/wordcloud-python