ANALYSIS ON TWITTER DATASET

DATA7201 – Project report

Tejeshvini Ashre
4560340

Structured Abstract

This is a comprehensive report of exploration, analysis and findings on the twitter API dataset. The tools used for this project are PySpark for querying and analysing and pandas for visualizing the results. The key analysis included finding out the most tweeted hashtags for a particular period of time, analysis of behaviour of the users with highest tweets including that hashtag. This also includes text analysis and making a wordcloud of the text in the tweets of users most tweeting about it. It was found that an online voting campaign was on during 2014 to aid in selecting a winner for the best music artist for 2014. We found that the words in the wordcloud formed were mostly popular music artists and bands that were globally famous. However, it couldn’t be cross checked with the actual winner as there were many winners in different category of awards.

Table of Contents

Introduction

Big data analytics

Big data Analytics is the process of exploring and analysing huge and varied datasets which are usually referred to as Big data. This process helps organizations discover correlations, market trends, customer behaviour, hidden patterns etc to make informed business decisions. Therefore, big data analytics help data scientists, statisticians and various analytics professionals to analyse growing volumes of structured transaction data and unstructured data like web server logs, social media data, website traffic data etc.

Unstructured or semi-structured data types don’t usually fit well in traditional data warehouses which are based on relational databases designed for structured data sets. Data warehouses may not be able to handle. Sometimes big data that is updated frequently like real time stock trading data, weather data etc. Data warehouses may not be able to handle processing demands posed by such datasets.(Rouse, 2010)

Distributed systems

A distributed system is a collection of independently operating computing elements that coordinate and communicate in order to appear to its users as a single coherent system.(Steen & Tanenbaum, 2016)

These machines have a shared state, operate concurrently and can fail independently without affecting the whole system’s uptime.There are four major reasons for building distributed systems: resource sharing, computation speedup, reliability, and communication. In this section, we briefly discuss each of them.

Resource sharing : A number of different sites with different capabilities are connected together such that a user at one site can access resources available at a different site

Computation speed up : When a particular computation operation can be divided into sub computation operations, which can be run parallelly, a distributed system allows the sub computation operation to be distributed among different sites. Therefore, these computations can be executed simultaneously and therefore aid computation speed up. Additionally, if one site is overloaded with jobs, some of the jobs can be moved to other sites with lighter loads. This is called load sharing.

Reliability : As a distributed system consists of independently operating computers, failure of one of them should not affect the operation of other systems. Any failure should be detected by the system and appropriate steps are taken to recover from the failure. The site is no longer used by the system and its functions are transferred to another site till the failed site repaired completely.

Practical examples of distributed computing systems : Telecommunication networks, Aircraft control systems, Sensor networks. (Motivation)

Dataset Analytics

The main focus of this analysis is to analyse twitter data for a particular period of time. In this particular case, we answered questions like, what was the most tweeted hashtag for a particular period of time? Who was tweeting it? What URLs were a part of this hashtag? Was there any campaign going on in relation to the most tweeted hashtag? What words apart from the hashtag are most being used by the tweeters tweeting the most tweeted hashtag.

We used the twitter dataset which is a collection of tweets. This data is available through the free twitter API and accounts to 1% of data spanning 6 months – July to December 2014. This data has a complex schema with various attributes and sub-attributes like user, entities, tweet text, retweeted, language of the tweet, geo location, hashtags etc. Below id the overall schema of this dataset.

The schema for the entire dataset is extensive. Hence, we subset the main data frame with only the attributes that we require for analysis.

The tool used were pyspark for querying, aggregating and analysing the dataset, and python was used for visualizing the results.

Firstly, we select a time period of 10 days to analyse the data – 11th December to 19th December 2014. We then drop the unwanted attributes and retain only the attributes that help us in the analysis – user, entities, language, retweeted, favorited, text, delete, retweeted_status. These attributes further have sub attributes that help us understanding various aspects of tweets and the users on twitter.

The entities attribute has a sub attribute ‘hashtag’ that include the hashtags tweeters use in their tweets. The most tweeted hashtags for this 10 day period were aggregated and following was the result

```
+--------------------+--------+
|                text|   count|
+--------------------+--------+
|                  []|35219534|
|          [MTVStars]|  558557|
|[ipad, ipadgames,...|   18887|
|     [FWEnVivoAwards]|   16641|
|[MTVSTARS, MTVSTARS]|   15416|
|          [ZDRSEDK4]|   14151|
|          [TuitUtil]|   13802|
|             [Quran]|   13341|
|    [GlobalArtistHMA]|   13077|
|[android, android...|   12767|
|[ÖzgürBasınSustur...|   10849|
|                [RT]|   10422|
|         [openfollow]|   10296|
|        [テイルズ, アスタリア]|   10292|
|[BrazilWantsOTRAT...|   10220|
|          [えどがわイケメン]|   10200|
|             [الهلال]|    9779|
|[شركات, مؤسسات, ا...|    9666|
|               [NMA]|    9623|
|        [CheatCodes]|    9375|
+--------------------+--------+
```

The most tweeted hashtag was found to be ‘MTVStars’. The difference between the first and second most tweeted hashtag was found to be high. ‘MTVStars’ has a hashtag count of 558557 whereas the second most tweeted hashtag [ipad, ipadgames,…] has a hashtag count of 18887. Variations of the hashtag ‘MTVStars’ is the fourth most tweeted hashtag. It was found that MTVStars of 2014 was a social media voting campaign of enormous scale.

The campaign involved users voting for their favourite artists amongst the biggest "stars" of 2014. Users could take part in this online voting campaign by tweeting #MTVstars followed by their favourite music artist or band. Then the music stars were selected based on various attributes like social media following, digital performance, chart performance and their relevancy throughout the year.

This projects aims to analyse the users who took part in this campaign and if it was successful in helping MTV choose the winner for that year. Therefore, our analysis of users, tweets etc is on this hashtag (#MTVStars).

The main data frame with the attributes above over the dates 11th to 19th December was filtered only by hashtags that contained ‘MTVStars’ in it named as MTVStars_df

The MTVStars_df was subset with only the user attribute to get the users that were most tweeting about MTV Stars. The urls linked to the tweets were aggregated. The urls sub attribute from the user attribute was used to analyse this. The top 20 urls with users tweeting about MTV Stars were as follows.

```
+--------------------+------+
|                 url| count|
+--------------------+------+
|                null|518927|
|http://stars.mtv.tv/|  9858|
| http://stars.mtv.tv|  8213|
|http://onedirecti...|  1026|
|http://www.mtv.co...|   906|
|https://twitter.c...|   712|
|https://twitter.c...|   520|
|http://youtube.co...|   473|
|http://www.onedir...|   368|
|http://www.youtub...|   356|
|https://twitter.c...|   345|
|https://soundclou...|   247|
| http://fb.com/kesha|   194|
| http://paramore.net|   176|
|http://www.justin...|   174|
|http://stars.mtv....|   172|
|http://twitter.co...|   165|
|https://twitter.c...|   153|
|http://stars.mtv....|   148|
|https://www.faceb...|   148|
+--------------------+------+
```

As we can see the top 2 urls are stars.mtv followed by one direction, youtube etc. All the top urls were major websites.

The top users tweeting about MTVStars were as follows:

```
+------------------+-----+
|              name|count|
+------------------+-----+
|#MTVStars Coldplay|18031|
|         MTVCpVote|13143|
|       NICKI MINAJ|10362|
|      ✝NICKI MINAJ✝| 9163|
|        Katy Perry| 7276|
|        Kesha Rose| 6977|
|             kesha| 6038|
|            Xyrien| 5641|
|        kesha rose| 3961|
|     Saved Account| 3941|
|             saved| 3712|
|       Nicki Minaj| 3509|
|     Justin Bieber| 2944|
|         Lady Gaga| 2471|
|             nicky| 2159|
|             MTVCp| 2025|
|   @itsBetancourth| 1986|
|      Lana Del Rey| 1959|
|             Ke$ha| 1733|
|           Beyonce| 1730|
+------------------+-----+
```

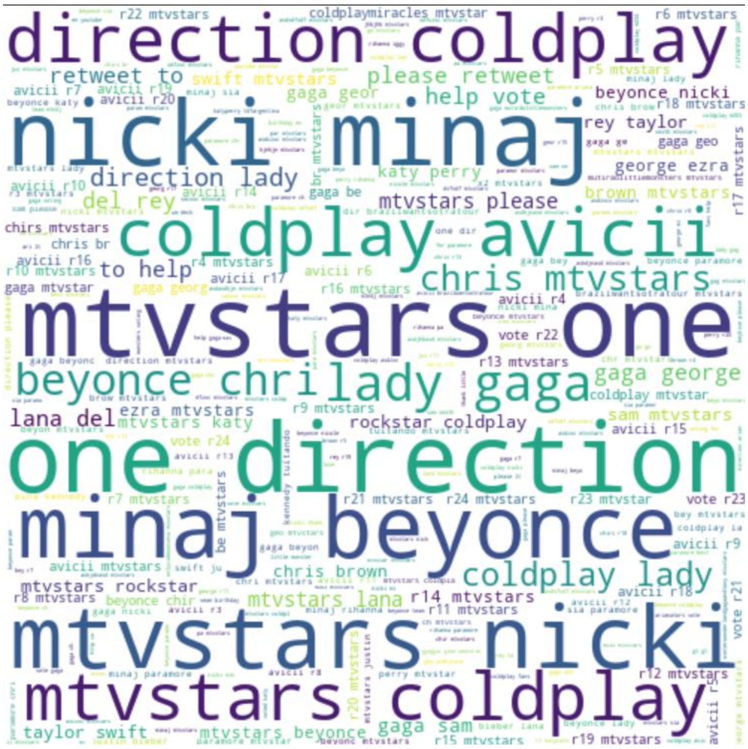
Many of the names in the above list seem to be top music artists – Nicki Minaj, Kesha, Justin Bieber, Beyonce etc. However, when the number verified accounts was calculated in the MTV Stars data frame, there were only 2 verified accounts.

```
+--------+------+
|verified| count|
+--------+------+
|   false|606979|
|    true|     2|
+--------+------+
```

These verified accounts were the official accounts of MTV London and MTV New Zealand.

```
location='London, UK', name='MTV Music'
location='New Zealand', name='MTV NZ'
```

We analysed what were the words most being used in the texts of the tweets by the top users tweeting about MTV Stars. This was done in order to check which music artist received the highest number of votes. A list with top 5 users was made and their statuses were collected and converted to a pandas dataframe. Their statuses were then tokenized and the stopwords were removed. Then a word cloud of this bag of words was made.



From the word cloud, we can see that mtvstars hashtag has the highest frequency. However other words are Nicki Minaj, One Direction, coldplay, Beyonce etc are also seen in a bigger size than rest of the words.

Since Nicki Minaj has the largest word size, we assume that the highest voted name was Nicki Minaj for the campaign MTV Stars.

Discussion and conclusion of the analysis

The approach used in this project was to focus on a specific period of time and look at the hashtags most tweeted about. We found that some hashtags like MTV Stars etc were tweeted a lot more than the other hashtags in the period 11th December 2014 to 19th December 2014. The difference between the first and the second most tweeted hashtag was also found to be huge(MTVStars - 558557 ; [ipad, ipadgames,…] – 18887). As there was an online campaign running during that period which required users to tweet #MTVStars followed by their favourite music star of the year 2014 in order for the star to win the award.

The user tweeting about the hashtag ‘MTV Stars’, associated urls, tweets containing this hashtag were analysed. It was found that the urls linked to these hashtags were from major websites like stars.mtv.com, youtube.com. We also aggregated the users most tweeting about this and most of the top 20 users has usernames of famous music artists and bands like Coldplay, Nicki Minaj, Kesha, Justin Beiber etc. However, when it was checked if any of these accounts were verified, it was found that none of the top twenty accounts were verified. The only verified accounts taking part in the online voting campaign were MTV London and MTV New Zealand. We also analysed the text or words in tweets containing this hashtag to see the music artist with highest number of votes in the top 5 users tweeting about MTV

Stars. From the world cloud it was found that Nicki Minaj had the highest number of words in the tweets containing MTVStars hashtag tweeted by the top 5 tweeters for the campaign.

The major takeaway from this analysis is that a lot of users had taken part in the MTV Stars online voting campaign. The top users tweeting about it were fan accounts of major music artists like Nicki Minaj, Justin Beiber etc. These were not the official accounts of the artists themselves but fan accounts. However, since we sampled for a very short duration of the entire campaign that ran from September to December, the insights from our results didn't apply in the bigger picture. Therefore, our sample was not representative of the user population that took part in the campaign.

Total wordcount : 1529 (Excluding headings)

Appendix

https://www.tutorialspoint.com/create-word-cloud-using-python

https://sparkbyexamples.com/pyspark/pyspark-explode-array-and-map-columns-to-rows/

https://shortyawards.com/7th/mtvstars-of-2014

Motivation, D. S. Retrieved from http://www.padakuu.com/article/185-distributed-system-motivation

Rouse, M. (2010). What is Big Data Analytics and why is it important? Retrieved from https://searchbusinessanalytics.techtarget.com/definition/big-data-analytic

Steen, M. v., & Tanenbaum, A. S. (2016). A brief introduction to distributed systems. *Computing*.

Code

```
from __future__ import print_function

import pyspark

from pyspark import SparkContext

from pyspark.sql import SQLContext, DataFrame

from pyspark.sql.functions import *

sc= pyspark.SparkContext("yarn")

sqlContext = SQLContext(sc)

df = sqlContext.read.json('/data/ProjectDataset/statuses.log.2014-12-
1*.gz')

df_sub = df.na.drop(subset=["user.id"]).select(["user","entities", "lang",
"retweeted", "favorited", "text","delete","retweeted_status"])

from pyspark.sql.functions import array_contains

entities_df = df_sub.select('entities.*')

hashtags = entities_df.groupby('hashtags.text').count()

print(hashtags.sort(desc('count')).show())
```

```
MTVStars_df =
df_sub.filter(array_contains(df_sub.entities.hashtags.text,'MTVStars'))

MTVStars_retweeted_status = MTVStars_df.select('retweeted_status.*')

retweet_count = MTVStars_retweeted_status.groupby('retweet_count').count()

print(retweet_count.sort(desc('count')).show())

retweet = MTVStars_retweeted_status.groupby('retweeted').count()

print(retweet.sort(desc('count')).show())

MTVstars_user = MTVStars_df.select('user.*')

MTVstars_urls = MTVstars_user.groupby('url').count()

print(MTVstars_urls.sort(desc('count')).show())

MTVstars_tweeters_name = MTVstars_user.groupby('name').count()

print(MTVstars_tweeters_name.sort(desc('count')).show())

MTVstars_ver = MTVstars_user.groupby('verified').count()

print(MTVStars_ver_only)

MTVStars_new=MTVStars_df.select("text","user.name")

MTVStars_top1 = MTVStars_new.where(col("name").isin({"#MTVStars
Coldplay","MTVCpVote","NICKI MINAJ","†NICKI MINAJ†","Katy Perry"}))

MTVStars_top1.toPandas().to_csv('MTVStars_top1.csv')

import numpy as np

import pandas as pd

from os import path

from PIL import Image

from wordcloud import WordCloud, ImageColorGenerator


import matplotlib.pyplot as plt

wordcloud_df = pd.read_csv("MTVStars_top1.csv")

text_strings = wordcloud_df["text"].to_string()

text_tokens = nltk.word_tokenize(text_strings)

from nltk.corpus import stopwords

stopword_list = stopwords.words('english')
```

```
wordcloud_words = ' '.join([token for token in text_tokens if token not in
stopword_list])

wordcloud_words = wordcloud_words.lower()

x=' '.join([token for token in text_tokens if token not in stopword_list])

def create_word_cloud(string):

    cloud = WordCloud(width = 500, height = 500,background_color = "white",
max_words = 500, stopwords = set(STOPWORDS))

    cloud.generate(string)

    cloud.to_file("wordCloud6.png")

create_word_cloud(str(wordcloud_words))
```