TWITTER DATA ANALYTICS

DATA7201

Reia Natu- 45634116

ABSTRACT

The dataset used for analysis is a collection of tweets available via the Twitter API; covering a 1% random sample of the entire Twitter stream over 6 months, specifically between July 2014 and December 2014. The 2 key areas in developing insights for this project are:

1) Building a Linear Regression Model:
   The data will be sub-sampled into 1000 observations and a linear regression model pipeline will be generated in order to predict the number of retweets associated to a particular tweet.
2) Analysing Donald Trump Tweets 2014:
   This will analyse the tweets of Donald Trump on days in September 2014. Key analysis involves identifying popular hashtags and words used in tweets and visualising them.

To generate these insights, the data will first be pre-processed and cleaned, and results will then be generated using Spark, PySpark and Python.

TABLE OF CONTENTS

INTRODUCTION

Big Data is used to describe a collection of data which is huge in size and that continues growing exponentially with time. To better understand Big Data, here are some of the characteristics it possesses:

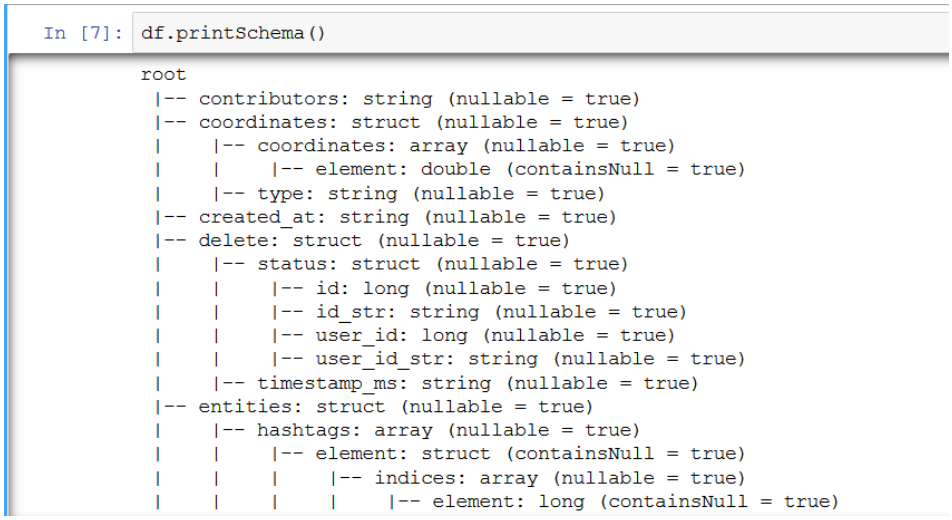
Properties	Description
Velocity	This refers to the speed of generating Big Data. For e.g.: Messages or videos that go viral in seconds on social media.
Variety	Data can be structured, unstructured or semi-structured, i.e. in the form of texts, audio, graphs, etc. and can be obtained from various sources like sensors, social media, etc. Further, it can possess different dimensions.
Veracity	This refers to reliability of data. It can be tricky to manage the quality and precision of Big Data. E.g.: if there are many worker nodes running simultaneously to perform a task, achieving reliability could be difficult.
Volume	This refers to a large amount of data transformed in seconds which traditional systems fail to handle. E.g.: the lower limit of Big Data is found to be approximately 1 TB. The number of likes for a particular post on social media could cross millions in a day.
Value	It refers to the cost of maintaining Big Data. Big data is valuable in terms of processing time with respect to generating insights, mining and visualising the data.

Traditional systems are not capable of handling exceptionally large amounts of various data types created and collected at a high speed. It becomes difficult for conventional data warehousing to store, analyse, manage and process Big Data. This raises the need to rely on Distributed Systems (DS). The DS system architecture distributes the storage over different nodes encompassing a cluster. Moreover, it is possible to scale-up the system architecture merely by adding more machines, whenever required. The velocity can be adjusted by balancing the load amongst the cluster nodes; failure of a machine would thus mean replacing it with another one to do the same job, while maintaining quality performance.

Many companies across the world employ DS systems. For instance, Netflix has around 160 million subscribers. The company keeps track of movie ratings, preferable genre, etc. for each of its subscribers. This accounts for a large volume of real-time data which is processed in milliseconds. Millions of these records cannot be processed or analysed in seconds by conventional database systems like sequential computing processes. Thus, DS systems like Hadoop and technologies like Spark form the backbone of Netflix's success because of their ability to maintain a trade-off between computational cost, time and quality of data interpretability.

DATASET ANALYTICS

The Twitter dataset's schema is essentially based on a tweet object. Below is a snapshot of a part of the schema including certain sub-entities.



```
In [7]: df.printSchema()

root
 |-- contributors: string (nullable = true)
 |-- coordinates: struct (nullable = true)
 |    |-- coordinates: array (nullable = true)
 |    |    |-- element: double (containsNull = true)
 |    |-- type: string (nullable = true)
 |-- created_at: string (nullable = true)
 |-- delete: struct (nullable = true)
 |    |-- status: struct (nullable = true)
 |    |    |-- id: long (nullable = true)
 |    |    |-- id_str: string (nullable = true)
 |    |    |-- user_id: long (nullable = true)
 |    |    |-- user_id_str: string (nullable = true)
 |    |-- timestamp_ms: string (nullable = true)
 |-- entities: struct (nullable = true)
 |    |-- hashtags: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
```

The tweet object consists of sub-entities like 'user', 'hashtags' , 'retweet count' and so on.

This project explains the following 2 objectives and further derives insights:

1) Generating a linear regression model to predict the retweets associated with a particular tweet:

   The parameters of the model include:

   - Lang : describing the language in which the tweet is
   - Followers_count : involves the number of people who follow a particular user
   - Friends_count : denotes the following count of a user
   - Favourite_count : describes the number of times a tweet is categorised into favourites
   - Retweet_count : is the count of the number of times a tweet has been retweeted
   - Statuses_count: count of the number of statuses uploaded by a user

Except language, which is a categorical variable, the remaining parameters are continuous variables.

Retweet_count is the target variable, others being predictors for the analysis.

The linear model is constructed on a sub-sample of 1000 observations wherein the input data is split into a 70% : 30% ratio for training and testing, and evaluated using RMSE and R-squared values as performance metrics. Predictions are then made on the test set to check how likely it is for a particular tweet to be retweeted.

2) Analysing Donald Trump Tweets of 2014:
   Key insights into this part include:
   - Checking for popular words in the tweets
   - Checking for popular hashtags used in the tweets
   - Visualising the earlier two insights as a word cloud

DATASET DESCRIPTION

The Twitter data available on the source, i.e. on the DATA7201 Hortonworks cluster consists of tweets from July 2014 to December 2014. It is stored as daily log files containing random tweets for a particular day. The cluster has 184 files where each file represents tweets for that day for an aggregate period of 6 months.

The linear regression model to predict the number of retweets a particular tweet gets, is built on tweets from the month of September 2014.

```
>>> df = sqlContext.read.json("/data/ProjectDataset/statuses.log.2014-09-**.gz")
```

Additionally, the analysis considers data of Donald Trump Tweets involves combining data of Trump tweets in the range of 1st to 28th September 2014.

The attributes used for the proposed analysis are as seen in the table below:

Attributes	Characteristics
User	Has many sub-attributes containing information about the username, profile image, count of followers, friends and statuses, etc.
Entities	This also has many attributes associated with hashtags in tweets, website URLs, etc.
Text	This describes the actual content of a tweet.
Lang	This denotes the language in which the tweet is written. Eg: For the current analysis, only English tweets are going to be considered; these are stored under lang 'en' standing for English.

PRE-PROCESSING STEPS

1) Analysis 1: Linear Regression Model
   The prediction of retweets for a particular tweet involves the following columns: 'retweet_count', 'user.favourites_count', 'lang', 'user.followers_count' , 'user.statuses_count' and 'user.friends_count' as printed in the schema below:

```
>>> analysis_df1.printSchema()
root
 |-- favourites_count: long (nullable = true)
 |-- followers_count: long (nullable = true)
 |-- friends_count: long (nullable = true)
 |-- statuses_count: long (nullable = true)
 |-- lang: string (nullable = true)
 |-- retweet_count: long (nullable = true)
```

STEPS:

- The initial step involves dropping missing values, if any, in the columns considered in the schema.
- My primary focus is on tweets in English so the dataframe will be filtered to get only those tweets that are in English as seen in the figure below.

```
>>> analysis_df1.show()
+----------------+---------------+-------------+--------------+----+-------------+
|favourites_count|followers_count|friends_count|statuses_count|lang|retweet_count|
+----------------+---------------+-------------+--------------+----+-------------+
|            2741|           1258|          131|         25177|  en|            0|
|             593|            819|         1489|          3969|  en|            0|
|             228|            105|          320|           230|  en|            0|
|            6416|            965|         1818|         21387|  en|            0|
|              17|             16|           28|            17|  en|            0|
|               0|            123|            1|        128213|  en|            0|
|               0|              2|            0|         34624|  en|            0|
|           13867|           3919|          468|         93393|  en|            0|
|            1021|             31|          471|          1962|  en|            0|
|            2138|            444|          396|          6015|  en|            0|
|             982|           1279|          898|         42681|  en|            0|
|             562|           1441|          423|         16455|  en|            0|
|              44|           1196|         1287|         52262|  en|            0|
|            3369|            147|          215|          1288|  en|            0|
|             243|            606|          977|          3085|  en|            0|
|              29|            358|          476|          4085|  en|            0|
|               0|            353|            1|         14534|  en|            0|
|              16|           3174|         3463|           650|  en|            0|
|            5980|            531|          715|         10340|  en|            0|
|             229|             78|          198|            29|  en|            0|
+----------------+---------------+-------------+--------------+----+-------------+
only showing top 20 rows
```

- A sample of 1000 is taken from this data of September 2014 and a linear regression model is built.
- For implementing the linear regression model in PySpark, the pyspark.ml module is used to build a model consisting of numeric data. The 'retweet_count' is the label column for the analysis, in other words, it can be considered as the target or dependent variable. The remaining columns from the schema are assembled as feature vectors by a vector assembler as seen below.

```
>>> v_df.show(3)
+--------------------+-------------+
|            features|retweet_count|
+--------------------+-------------+
|[2741.0,1258.0,13...|            0|
|[593.0,819.0,1489...|            0|
|[228.0,105.0,320....|            0|
+--------------------+-------------+
only showing top 3 rows
```

This data is then fed into a pipeline that performs transformation. After transforming the data, it is split to create training and test sets, where 70% is used for training the model and 30% for validation. Further, the prediction of a particular tweet being retweeted will be generated using the test set. To check the goodness of the model built, it is necessary to calculate the differences between predicted values and the actual values which is computed in terms of the RMSE and R Squared values.

2) Analysis 2: Trump Tweets Analysis

- The columns used for data on trump tweets involve columns bearing hashtags and text.
- The tweets for this analysis are filtered and grouped by language as English and filtered for tweets in September 2014 to generate the required data as seen below.

Out[44]:

	handle	text	time	lang	retweet_count
1	realDonaldTrump	Hillary Clinton didn't go to Louisiana, and no...	2014-09-01T00:40:07	en	20890.0
2	realDonaldTrump	Great trip to Mexico today - wonderful leader...	2014-09-01T00:43:00	en	10252.0
3	realDonaldTrump	Just arrived in Arizona! #ImWithYou\nhttps://t...	2014-09-01T00:57:54	en	3322.0
9	realDonaldTrump	I hear Churchill had a nice turn of phrase, bu...	2014-09-01T04:40:26	en	9105.0
10	realDonaldTrump	We are Watching A Leader Who for the First Tim...	2014-09-01T04:40:54	en	9401.0
11	realDonaldTrump	Mexico will pay for the wall - 100%!\n#MakeAme...	2014-09-01T04:58:19	en	10340.0
12	realDonaldTrump	There will be no amnesty!\n#MakeAmericaGreatAg...	2014-09-01T04:58:53	en	6504.0
13	realDonaldTrump	Hillary Clinton doesn't have the strength or t...	2014-09-01T05:06:52	en	7138.0
14	realDonaldTrump	Under a Trump administration, it's called #Ame...	2014-09-01T05:07:04	en	8741.0
15	realDonaldTrump	Mexico will pay for the wall!	2014-09-01T10:31:17	en	30488.0
16	realDonaldTrump	Thank you to @foxandfriends for the great revi...	2014-09-01T10:40:13	en	7082.0
17	realDonaldTrump	Poll numbers way up - making big progress!	2014-09-01T10:46:44	en	10162.0
20	realDonaldTrump	Thank you for having me this morning @American...	2014-09-01T14:22:26	en	6355.0
28	realDonaldTrump	I am promising you a new legacy for America. W...	2014-09-01T17:05:25	en	9741.0
34	realDonaldTrump	I will be interviewed by @ericbolling tonight ...	2014-09-01T22:53:52	en	3945.0
42	realDonaldTrump	Just heard that crazy and very dumb @morningmi...	2014-09-02T12:28:58	en	5944.0
43	realDonaldTrump	People will be very surprised by our ground ga...	2014-09-02T12:32:31	en	9397.0
44	realDonaldTrump	I visited our Trump Tower campaign headquarter...	2014-09-02T12:35:44	en	7267.0
49	realDonaldTrump	Great new poll Iowa - thank you!\n#MakeAmerica...	2014-09-02T16:55:57	en	10351.0
57	realDonaldTrump	Thank you Great Faith Ministries International...	2014-09-03T17:27:53	en	11097.0

- Further, Python is used to create visualisations of popular words and hashtags .

The analysis involving text from tweets (for popular words) primarily focuses on text analytics. Hence, before analysing this text, there are certain string manipulation steps undertaken to clean the tweet in order to get the data fit for use*. They involve:

1. Removing hashtags and account names from the tweet text.
2. Converting the tweet to a lower case.
3. Removing hyperlinks, punctuations, special characters, digits, underscores and stop words like ‘is’, ‘are’, ‘the’.

The snapshot below shows the clean tweet after these steps are incorporated.

	handle	text	time	lang
0	realDonaldTrump	hillary clinton go louisiana go mexico drive stamina make america great	2014-09-01T00:40:07	en
1	realDonaldTrump	great trip mexico today wonderful leadership high quality people look forward next meeting	2014-09-01T00:43:00	en
2	realDonaldTrump	arrived arizona	2014-09-01T00:57:54	en
3	realDonaldTrump	mexico pay wall	2014-09-01T04:58:19	en

*Note: Hashtags are analysed using the hashtag column corresponding to tweets

RESULTS

Analysis 1: Linear Regression

To begin with, the sample size considered was only 1000 from amongst twitter data of September 2014. The training sample that was chosen was from this sample where as seen in the figure below, there were zero retweets. Thus, no trends could be identified for this tenure.

```
>>> analysis_df1retweetresult.show()
+----------------+---+---------------+-------------+--------------+----+-------------+
|favourites_count| id|followers_count|friends_count|statuses_count|lang|retweet_count|
+----------------+---+---------------+-------------+--------------+----+-------------+
+----------------+---+---------------+-------------+--------------+----+-------------+
```

Further, the summary of prediction of retweets on the test set showed a zero coefficient and intercept value explaining no trend observed or predicted because of the negligible retweet count.

```
+---------+
|residuals|
+---------+
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
|      0.0|
+---------+
only showing top 20 rows
```

```
>>> print("Coefficients: " + str(lr_model.coefficients))
Coefficients: [0.0,0.0,0.0,0.0]
>>> print("Intercept: " + str(lr_model.intercept))
Intercept: 0.0
```

```
>>> print("R Squared (R2) on test data = %g" % lr_evaluator.evaluate(lr_predictions))
R Squared (R2) on test data = nan
>>> print("RMSE: %f" % trainingSummary.rootMeanSquaredError)
RMSE: 0.000000
>>> print("r2: %f" % trainingSummary.r2)
r2: nan
```

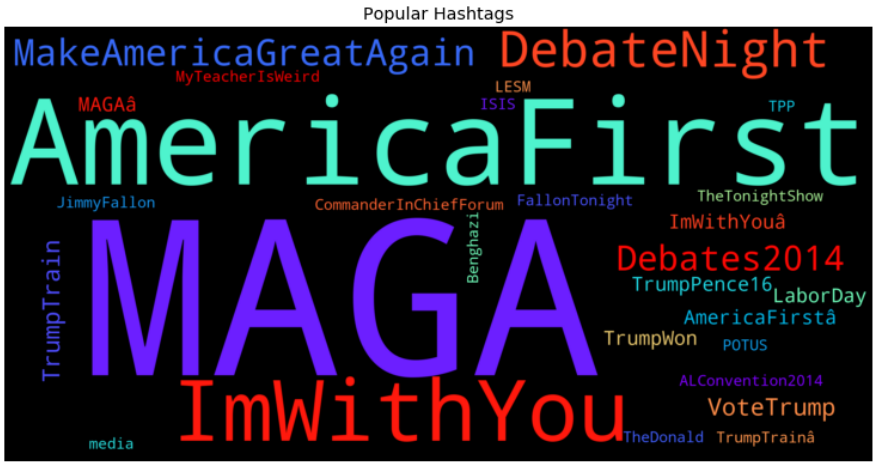
Analysis 2: Trump Tweet Analysis

1) Popular Word Analysis:

The word cloud below shows the frequency of words used in Trump tweets. It is evident that 'thank' is one of the most popular words.

2) Popular Hashtag Analysis:

The word cloud below shows the hashtags as per frequency, used in Trump tweets. It is evident that ‘MAGA’, ‘AmericaFirst’ are the most popular hashtags.



DISCUSSION

- From the outcome of the linear regression model, it seems that though the variables considered were continuous, it was not quite a good fit. The root cause could be the small sample size used as training data. Increasing the size of the training data could perhaps enhance model performance. The model built for this analysis aims to choose the best parameters from amongst a large set of parameters for training, and to test for more parameters or faster processing it is viable to work with a smaller set of data. The current pipeline can benefit say influencers to forecast their interactions on social media. Additionally, it is possible to check for retweets based on other languages too, in order to analyse the trend, behaviour and impact of tweets across multiple cultures.
- The Trump Tweet analysis helps to identify words with maximum reach and hashtags used commonly. It can be useful for campaigning due to the insights gained on the effective reach of hashtags and use of specific words.
- The word cloud is a user-friendly and easily interpretable visualisation of data which illustrates popular words and hashtags used in tweets. This also helps in identifying trends in tweets for a given timeframe.
- The algorithms on Big Data can be run even on small data. Big Data analytics can become an optimal process in the true sense when for instance, a smaller sample of data becomes a representative of the population (Big Data). This implies that it is better to build a model on a sample and understand key trends and then run queries on a bigger volume to derive insights, thereby making the best use of time and other resources.