CSCI 5408 ASSIGNMENT 3

By: Bhargav Dalal (B00785773) & Hardik Galiawala (B00777450)

JUNE 23, 2018
DALHOUSIE UNIVERSITY

Table of Contents

1. Objective:

The main purpose of this assignment is to perform streaming of data from social media platform (Twitter) by using the big data analytical tool (Apache Spark). To apply machine learning techniques (classification algorithms) to the streamed data.

2. Task Description:

The task was to understand the process of streaming the data (tweets) from Twitter which is a social media platform. These tweets are analyzed and classified if they have a positive, negative or neutral sentiment. This prediction is done based on the model, which we trained using spark machine learning library.

We have used test dataset [1] to train our model using spark pipeline. The reason we used to test data set for training is that the training set has only two classes whereas we need to perform three-class classification as stated in the objective of this report.

The trained model is used to predict the streamed tweets in batches and we terminate the process after analyzing 2000 tweets due to cloud service limitations.

We have installed Python 3.6 and Apache Spark 2.3.1 on the server instance on Microsoft Azure. Once the environment is set up, we installed libraries such as “tweepy” and “pyspark”. We have extensively used regex to perform various string operations on the tweets.

3. Twitter Tweet Extraction:

We have worked on python to extract the data. “tweepy” is the package used to communicate with Twitter API. First, we install the tweepy package in python. Further, from the twitter, noted the require consumer & access key as well as consumer & access secret to establish and authorize the connection.

```
# Establish connection

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth)
```

Fig 3.1 – Establishing Connection with Twitter

This process is done during the streaming of tweets where we have created a socket which will act as a server to the spark application (tweepy_stream.py) where the tweets are fed. Once the connection is established, the tweets are extracted by cleaning/filtering the tweets using regex.

4. Sentimental Analysis Algorithm:

The sentimental analysis is done using pre-labeled dataset (data_final.csv) which we have mentioned in the "Task Description". The first column of the CSV file has three categories: Positive, Neutral and Negative. The respective scores for each category of words are 4, 2, and 0. The second column is tweet id, third has tweet date, fourth is the username and the last column is tweet text.

As we understand, this is a classification problem it is a good idea to use a machine learning algorithm which is based on classification. Thus, we use RandomForestClassifier() algorithm to predict the above-mentioned class labels. RandomForest algorithm is based on decision making based on trees. Also, it works well with less training data if we tune its hyperparameter (numTrees).

While training the model, we split the dataset in 7:3 ratio as train and test set respectively. The test accuracy is around 64% which is bad. But our focus was to understand the working of cloud technology, deploy the trained model, and perform prediction on the data stream.

5. Labelling Training Data:

We downloaded labeled data which was mentioned in the above section of the report.

6. Feature Selection:

We understand that feature selection is an important part of any machine learning process. We have included all the tweets with meaningful words as our features. We have filtered certain types of words which are mentioned in the code.

```
# stop words
add_stopwords = ["http","https","amp","rt","t","c","the"]
stopwordsRemover = StopWordsRemover(inputCol="words", outputCol="filtered").setStopWords(add_stopwords)

# bag of words count
```

Fig 6.1 – Removing useless words

We have cleaned all the tweets after extracting them for predicting their sentiment. We have removed retweets, emojis, and extra spaces.

```
def clean_str(self, string):
        string = re.sub(r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*(),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+',
        string = re.sub(r"\n|\t", " ", string)
        string = re.sub(r"(.)\1{2,}", r"\1\1", string)
        string = re.sub(r"(..)\1{2,}", r"\1\1", string)
        string = re.sub(r"(...)\1{2,}", r"\1\1", string)
        string = re.sub(r"(....)\1{2,}", r"\1\1", string)
        string = re.sub(r'RT ', '', string)
        return string
```

Fig 6.2 – Cleaning tweets

7. Output:

We are storing the results of our analysis of the tweets in a CSV file (output.csv). This file contains two columns, the first one is the tweet text and the other is the result which is in the closed range [0 – 2] with Positive, Neutral and Negative respectively.

A	B	C
@coviramirez: Kudos to Toni Gonzaga and her new interna	1	
My 600 Lb life.	1	
Puta?	1	
I liked a @YouTube video Tropic Thunder (5/10) Movie CL	1	
@dianaaziz_: I need a museum date!! Gallery date!! Zoo d	1	
@Evettexo: Have y'all seen the movie acrimony? What y'a	1	
@tonystark_1993: ??? ???????????, ????? ???????????? ??	1	
Princess diaries	1	
How to lose a guy in 10 days This is a movie about a journa	0	
@Benris84: TFW you go to the movies and sat next to som	1	
@jessphoenix2018 @blumspew That was the first movie f	0	
@VijayUAEfans: Genuine Review? Which movie deserves I	1	
???????? ?????? #?????? #??????? #????????? #???????	1	
@BabuSenthil: #2POINT0ModeOn The basement of the m	0	
the parking lot at the movie theater was packed but they'r	0	
@Itsmadiha_: #HappyBirthdayHamzaAliAbbasi @iamham	1	
free sex video dump black live porn watch online adult mo	1	
@OkkadiFanIkada: Topic diversion ?? Intakanna peekeder	1	
Public agent brunette teen and extra small gets fucked And	2	
?????? ? ???????? ?????????//?????????MM?????????????	1	
@MrIanMalcolm: From the first Indian movie to have live	1	
@OhMyDisney: Your choice of 5 @DisneyPixar movies wil	0	
@eliesaab: Laura Harrier would be the perfect corpse brid	1	
@bnhaincorrect: yaoyorozu: getting a girlfriend looks easi	1	
@yusraaXO: Whose lousy son is going to take me to the m	0	
@menongautham @steaphenCVF Bro Release the movie s	1	
@GodzillasHeart Tru ?. Movie made me fall in love with W	1	

Fig 7.1 – Output

8. Code Submission:

GitHub Repo link: [Assignment3Repo]

Bibliography

[1] "sentiment140," [Online]. Available: http://help.sentiment140.com/for-students/.