Assignment 2

Report

Report is based on data extraction from different sources (Twitter and NewsAPI), Transformation on data, and Processing on Apache Spark to find the frequency of words provided.

Cloud Setup:

1. Creation of AWS instance.
2. Installed Apache Spark and Java.
3. Set global variables for Spark, Java, Python.
4. Run master node and slave node(if needed); Commands are as follows:
    a. *sudo ./spark-2.4.4-bin-hadoop2.7/sbin/start-master.sh* (for master node)
    b. *sudo ./spark-2.4.4-bin-hadoop2.7/sbin/start-slave.sh spark://ip-3.14.7.141:7077* (for slave node where 3.14.7.141 being my ip address)

Data extraction process:

- Maintained a modularized programming approach for easy readability and better understandability.
    - Tweets extraction
        - twitter_extract.py
            - Extracted 1000 tweet data for each of the keyword specified, hence total 5000 tweets extracted.
            - Each tweet data includes tweet text, retweet count, user location, coordinates.
            - Saved the result to a json file "*main_data_twitter.json*".
            - Fetched data from "*main_data_twitter.json*" and save it to MongoDB database "twitter".
            - Also saving all the twitter data to a text file "twitter_texts.txt".
    - News extraction
        - Newsapi_extract.py
            - Extracted all the news based on the keywords provided.
            - Saved to a json file "*main_data_news.json*"
            - Fetched from the json file and saved to MongoDB database "news".
            - Saving the news data to a text file "news_texts.txt" for Pyspark processing.

Data cleaning:

- Before saving the data to "*main_data_twitter.json*" and "main_data_news.json", I cleaned it based on different factors.
    - Removed all special characters as mostly they were attached with word itself (which could give improper results).
    - Put space between all the words, as many times 2 tweets used joined together.
    - Set all the data to lowercase for ease in processing it.
    - Removed hyperlinks as it was not needed.

- Data could be cleaned based on many other factors, which is very crucial for sentiment analysis, but it was unnecessary for counting frequency of words; hence did not clean further.

Data file format:

- During this assignment, I have used JSON as a main language to store and transfer data.
- Finally created a text file for processing in Apache Spark.

Spark Processing:

- spark-engine_news.py and spark-engine_twitter.py
    - used map-reduce approach to find the frequency of the words.
    - Created 2 RDD, one for storing words and other for storing phrases of the data as I had to find the frequency of phrases too.
    - For counting frequency of phrases such as 'good school' or 'bad school'; I have used python's in-build function zip() to create RDD of each consecutive 2 words.
        - For an instance, string = "this is my assignment". Hence zip() is used to make RDD of (("this is"),("is my"),("my assignment")).
        - This is done to find the frequency of phrases.
- Though, spark-engine_news.py and spark-engine_twitter.py has almost same code in it, and could be made as a single file serving both purposes, but kept different for future processing. As both are from different sources, hence there may be a need of different manipulation of data.

Result:

- For twitter data, I got following result:
    - education: 635
    - canada: 1144
    - university: 927
    - dalhousie: 846
    - expensive: 4
    - good: 37
    - bad: 6
    - good school: 2
    - bad school: 0
    - faculty: 9
    - computer science: 1
    - graduate: 28

- For news data, I got following result:
    - education: 3
    - canada: 67
    - university: 29
    - dalhousie: 1
    - expensive: 0
    - good: 4

- bad: 8
- good school: 0
- bad school: 0
- faculty: 2
- computer science: 0
- graduate: 0

Conclusion:

Although, the word ‘bad’ has appeared in the data which doesn’t convince enough that it is used for ‘Dalhousie University’ as data contains information of ‘Canada’, ’Halifax’ and many other keywords. But there hasn’t been a single occurrence of ‘bad school’, hence can be concluded that review of ‘Dalhousie University’ is not bad. Still it cannot be confirmed that all reviews are good, because this is just count of frequency of words, and nothing is done related to finding sentiments of data. Hence still many anomalies and ambiguities may exist in the result and it is unreliable to predict review of the university.

References:

[1]"BigData with PySpark", *Nyu-cds.github.io*, 2019. [Online]. Available: https://nyu-cds.github.io/python-bigdata/. [Accessed: 2- Oct- 2019]

[2]"Welcome to Spark Python API Docs! — PySpark 2.4.4 documentation", *Spark.apache.org*, 2019. [Online]. Available: https://spark.apache.org/docs/latest/api/python/index.html. [Accessed: 06- Nov- 2019]