TWITTER DATA ANALYSIS ON TECHNOLOGIES

Principles of Big Data Management Project (CS5540)

Phase-1 Report

Team:

Tiyyagura Sindhusha(16280708)

Pradeepika Kolluru(16283597)

Thoshita Movva(16279838)

Main Objective:

The main objective of the project is to store, analyze and visualize the twitter's tweets(based on keyword tech/technologies,…).

Objective for Phase-1

1. Collect the tweets (around 100K tweets) from twitter using twitter streaming APIs.
2. Extracting Hashtags and URLs from the tweets that are collected in the 1st task
3. Running WordCount program using Apache Hadoop and Apache

Spark on extracted HashTags and URLs.

Technologies Used:

Python

Twitter Streaming APIs

Apache Hadoop

Apache Spark

Task -1: Collecting the tweets (around 100K tweets) from twitter using twitter streaming APIs:

1. Generating twitter streaming APIs

- Log in to the twitter developer site (https://developer.twitter.com/en/apps)
- Create an application and generate consumer and access API keys

2. Installing python3, pip for python3 and tweepy python module

- Install python3
  `sudo apt-get install python3`
- Verify python installation
  `python --version`
- Install Python package Index (PIP) package management system used to install and manage software packages

`sudo apt-get install python3-pip`

- All other python required modules can be installed using the following command:
  `pip install <module_name>`

3. Collecting twitter tweets by running a python program

- The python program collects the tweets into a text file by using Twitters streaming API(tweepy python module).
- In order to connect to the API, one must give valid credentials from the twitter developer account.
- The final output data (tweets in JSON format) is redirected to a output file (tweets_tech.json file)

Links:

- Python code: https://github.com/sindhusha-t/twitter-data-analysis/blob/master/Phase-1/source code files/twitter_streaming.py.py
- Output file: https://github.com/sindhusha-t/twitter-data-analysis/blob/master/Phase-1/twitter_data_files/tweets_tech.json

Output file generated: 107MB

Task -2: Extracting Hashtags and URLs from twitter tweets:

- The python program parses tweets file into JSON format and

extracts required Hashtags and urls and writes to two different files.

Links:

- Python code: https://github.com/sindhusha-t/twitter-data-analysis/blob/master/Phase-1/source code files/twitter_extracting_hastags_urls.py
- Output Files: https://github.com/sindhusha-t/twitter-data-analysis/blob/master/Phase-1/twitter_data_files/hashtags.txt https://github.com/sindhusha-t/twitter-data-analysis/blob/master/Phase-1/twitter_data_files/urls.txt

Task -3: Running WordCount program using Apache Hadoop and Apache Spark on extracted HashTags and URLs

Running in Hadoop

1. Compiling the java source code and creating class files
   `javac -classpath ${HADOOP_CLASSPATH} -d <CLASSES_FOLDER> <JAVA_SRC_FILE>`
2. Creating a jar file for the java class files
   `jar -cvf <JAR_FILE_NAME> -C <COMPILED_CLASSES_PATH>`
3. Storing the input file in HDFS
   `hadoop fs -put input.txt /hdfs_path/wordcountExample/input/`

4. Running the Hadoop program
   `hadoop jar <JAR_FILE> <CLASS_NAME> <HDFS_INPUT_DIR> <HDFS_OUTPUT_DIR>`
5. Getting the output file from HDFS
   `hadoop fs –get <HDFS_OUTPUT_DIR> <LOCAL_DIR>`

Links:

1. Java code: https://github.com/sindhusha-t/twitter-data-analysis/blob/master/Phase-1/source code files/WordCount.java
2. Output files: https://github.com/sindhusha-t/twitter-data-analysis/tree/master/Phase-1/hadoop_output

Running in Scala

1. run scala environment using command:
   `spark-shell -master yarn-client`
2. loading the scala program
   `:load WordCount.scala`
3. run the main class
   `WordCount.main(null)`

Links:

1. Scala code: https://github.com/sindhusha-t/twitter-data-analysis/blob/master/Phase-1/source code files/WordCount.scala
2. Output file: https://github.com/sindhusha-t/twitter-data-analysis/tree/master/Phase-1/Scala output

Hadoop and Scala LOG Files:

https://github.com/sindhusha-t/twitter-data-analysis/tree/master/Phase-1/logs

References:

[1] https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/

[2] https://docs.inboundnow.com/guide/create-twitter-application/

[3] https://github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition/blob/master/ipynb/Chapter 1 - Mining Twitter.ipynb