CSCI 5408

Data Management, Warehousing Analytics:

Report - Assignment 2

November 11th, 2019

Submitted by:

Menni Prem Kumar

Dalhousie ID : B00843422

Brightspace: pr775390

Cloud Set-up steps:

- Open ~/.profile file: sudo nano ~/.profile (add below three lines in profile document)
- export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
- export SPARK_HOME=~/server/spark-2.4.3-bin-hadoop2.7
- export PYSPARK_PYTHON=python3
- cd ~/server sudo ./spark-2.4.3-bin-hadoop2.7/sbin/start-master.sh (Starting Spark Master)
- sudo ./spark-2.4.0-bin-hadoop2.7/sbin/start-slave.sh spark://ip-172-31-21-62.us-east-2.compute.internal:7077 (Starting spark Slave)
- sudo ./spark-2.4.0-bin-hadoop2.7/bin/pyspark (Starting Spark shell)
- sudo ./spark-2.4.0-bin-hadoop2.7/sbin/stop-slave.sh spark://ip-172-31-21-62.us-east-2.compute.internal:7077 (Stopping spark Slave)

Data Extraction / Cleaning Process:

Twitter Data:

I have Created a twitter developer account. Authentication tokens [2] provided in it are used to create a access path to extract the twitter data. I have used tweepy cursor function to extract the meta data from the twitter using search operation. Since I have to collect tweet data for each key word, I have used for loop to parse the key values to tweepy cursor. Tweepy provided the meta data for the tweets and using key value pairs, I have extracted required meta data from the set. I have used CSV file to store the required meta data. While accessing retweets, attribute error occurred, and I have handled with catch block and extracted full text of the retweet using retweeted status attribute. I have used Regular expression to clean the data by removing the links and special characters from the data. I haven't replaced any empty values for the metadata extracted because I don't want to corrupt the data with wrong value. A sample record of data is provided in Table 1.

Sample data of tweet:

Created at	text	location	Retweet count
2019-11-02 14:22	RT tufc Match Squad Vs FC	From Torquay	3

Table: 1

News Article Data:

I have created an account in newsapi.org, key is assigned to my account. An URL is created using the instructions provided in the Newsapi.org [3] website using required key words. I have used python requests package to get the data from the URL [1]. I have created a query using the URL to fetch the complete data containing keywords like Dalhousie, Canada. Resulted data is dumped into json object and then converted to dictionary. Since it is dumped into dictionary, I have used article as a key to extract the article data from the dictionary. I have used Regular expression to remove all special characters from the data. Article details like title, Content were extracted from the Resultant article

meta data and stored it in CSV file. No empty values were replaced with any common values like Nan because, having Nan or NULL is same. A sample record of data is provided in Table 2.

Sample data of news:

Author	Title	Description	Content
Josh Ocampo	Should You	In your quest to find an especially	In your quest to find an especially great

Table: 2

Data Processing:

I have manually created text file using the CSV files created while extracting tweets, news data. I have used FileZilla [4] application to transfer files from my desktop to AWS cloud server. I have created a pyspark script which will execute in AWS cloud server. Spark is configured with the application name DWMA_Assignment2, SQL Context and Spark context were initialized. I have stored the complete file data in a variable using Spark context, used flat map to create RDD using the data stored in the variable. I have spitted the data using "," and converted data to lower case. I have used filter and count functionalities to count the occurrence of key word. Count and key word name were stored in output file.

Observation:

I could able to find that, Canada is the most common word in the tweets and news articles followed by education.

References:

Professional Internet Site:

[1] Documentation on "News API: Extracting News Headlines and Articles" [Online]

Available : https://python.gotrained.com/news-api/

[Accessed on November 1, 2019]

[2] Documentation on "Twitter data extraction" [Online]

Available : https://developer.twitter.com/en/docs/tweets/search/overview

[Accessed on October 31, 2019]

[3] documentation on "News API data extraction" [Online]

Available: https://newsapi.org/docs/endpoints/everything

[Accessed on November 1, 2019]

[4] FileZilla the free ftp solution [Application]

Available: https://filezilla-project.org/

[Accessed on November 1, 2019]