CSCI 5408 Data Management, Warehousing and Analytics

Report

B00841566 **Punarva vyas**

Cloud setup steps:

1. I have enabled ports:8080 and 8081 on aws machine instance.
2. Added the Oracle Java PPA to apt run using command: sudo add-apt-repository ppa:webupd8team/java and sudo apt-get update .
3. Installed openjdk using command: sudo apt-get –y install openjdk-8-jdk-headless
4. Installed Python using: sudo apt-get install python3 .
5. Installed spark:
   wget http://apache.forsale.plus/spark/spark-2.4.4/spark-2.4.4-bin- hadoop2.7.tgz .

   Unpacked the folder by curl -L http://apache.forsale.plus/spark/spark-2.4.4/spark-2.4.4-bin- hadoop2.7.tgz> spark-2.4.3-bin-hadoop2.7.tgz .

6. To set the path permanently I have used following: sudo nano ~/.profile .
7. Then added following to the file:
8. export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
9. export SPARK_HOME=~/server/spark-2.4.4-bin-hadoop2.7
10. export PYSPARK_PYTHON=python3
11. I started the master using following command: sudo ./spark-2.4.4-bin-hadoop2.7/sbin/start-master.sh .
12. To start the slave node I have used following command: sudo ./spark-2.4.4-bin-hadoop2.7/sbin/start-slave.sh spark://ip- 172-31-40-128
    I have included the screenshot of my running master slave machine in screenshots folder.

Data Extraction and cleaning process:

1. First created the developer account to acess the api.
2. Authenticated for using the api using the secret key provided and the acess token.
3. Created a list of keywords for which I need to extract the data.
4. To extract search results, cursor method was used in tweepy library,3500 items are retrived for each word.To extract the news get everything method was used with page_size as 100.
5. The retrieved tweettext and newscontent is cleaned using the regex.Removal of “http:”,special characters,emoticons and smileys is done.
6. The cleaned data is inserted into the csv file.That csv file is then inserted into the mongodb database.The inserted data can be found in screenshots folder.The output csv file is shown below.[1-3]

id	tweet_text	tweet_created	source	favorite_c	retweet_c
1.19E+18	RT IATSECANADA H	06-11-2019 13:01	Twitter for	0	6
1.19E+18	Ask yourself would	06-11-2019 13:01	Twitter for	0	0
1.19E+18	conanupdating AH	06-11-2019 13:01	Twitter for	0	0
1.19E+18	MOOMANiBE Hanc	06-11-2019 13:01	Twitter for	0	0
1.19E+18	RT CBCNews Two C	06-11-2019 13:01	Twitter for	0	12
1.19E+18		06-11-2019 13:01	Twitter for	0	5
1.19E+18	RT GlobalEdmontor	06-11-2019 13:01	Twitter for	0	14
1.19E+18	leilanifarha Ottawa	06-11-2019 13:01	Twitter W	0	0
1.19E+18	UBC innovation	06-11-2019 13:01	Hootsuite	0	0
1.19E+18	RT always vote Sask	06-11-2019 13:01	Twitter W	0	3
1.19E+18	RT Starbucks Turn c	06-11-2019 13:01	Twitter for	0	182
1.19E+18	4 The519 has also e	06-11-2019 13:01	Twitter W	0	0
1.19E+18	RT CU President Th	06-11-2019 13:01	Twitter W	0	4
1.19E+18	Gordon Laxer Best	06-11-2019 13:01	Twitter for	0	12
1.19E+18	Winnipeg foster pa	06-11-2019 13:01	Facebook	0	0
1.19E+18	1 Politician wants H	06-11-2019 13:01	Twitter W	0	0
1.19E+18		06-11-2019 13:01	Hootsuite	0	0
1.19E+18	Where in Canada a	06-11-2019 13:01	Twitter for	0	0
1.19E+18	RT Tsiehta gtlem Ar	06-11-2019 13:01	Twitter for	0	11
1.19E+18	RT IchbinUjjaini This	06-11-2019 13:01	Twitter for	0	7
1.19E+18	Justin Trudeau s att	06-11-2019 13:01	Twitter for	0	761
1.19E+18	RT BioNovaNS Janir	06-11-2019 13:01	Twitter for	0	1
1.19E+18	Here are	06-11-2019 13:01	Twitter W	0	1

Fig 1.

Processing of the csv file:

The tweets and the news content are extracted from the mongodb database and stored in the text file programmatically.Then the output file is given to spark which returns the count of all the keywords mentioned.The output of it is stored in another text file and output is shown below.[4-5]

```
Canada --> 1384
Halifax --> 904
university --> 419
Dalhousie --> 143
faculty --> 43
education --> 23
graduate --> 12
expensive --> 2
```

Fig 2.

References:

[1] "Get and Work With Twitter Data in Python Using Tweepy," *Earth Data Science - Earth Lab*, 05-Feb-2018. [Online]. Available: https://www.earthdatascience.org/courses/earth-analytics-python/using-apis-natural-language-processing-twitter/get-and-use-twitter-data-in-python/. [Accessed: 01-Nov-2019].

[2] "Python client library," *News API*. [Online]. Available: https://newsapi.org/docs/client-libraries/python. [Accessed: 02-Nov-2019].

[3] "API reference index - Twitter Developers," *Twitter*. [Online]. Available: https://developer.twitter.com/en/docs/api-reference-index . [Accessed: 02-Nov-2019].

[4] *Python MongoDB Find*. [Online]. Available: https://www.w3schools.com/python/python_mongodb_find.asp. [Accessed: 04-Nov-2019].

[5] *Writing CSV files in Python*. [Online]. Available: https://www.programiz.com/python-programming/working-csv-files . [Accessed: 04-Nov-2019].