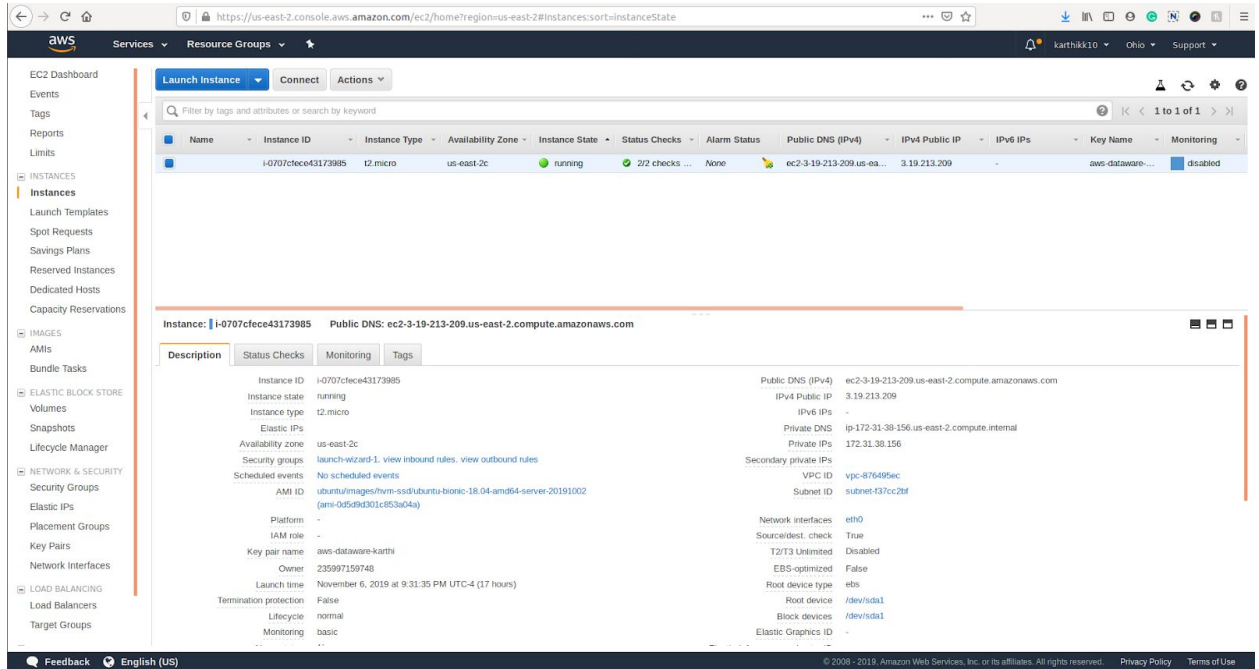


CSCI5408- ASSIGNMENT 2

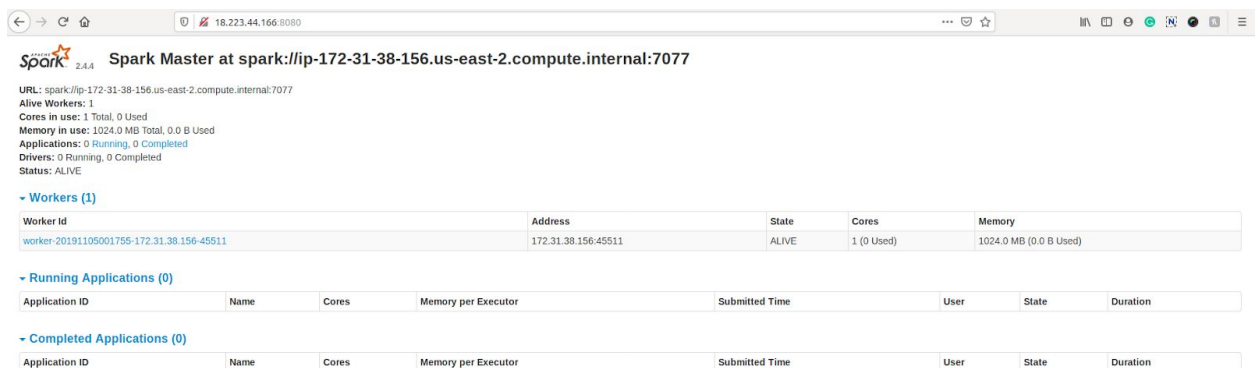
A. CLUSTER SETUP

I have used AWS (Amazon Web Services) to run a cloud instance of Ubuntu. In that cloud instance, I have installed MongoDB, Apache Spark, Python



For the installation and setup of Apache Spark, I have followed the steps provided in the lab slides [1]

After setting-up Spark, I have initialized a master and slave running on the same cloud instance.



As per my setup, Spark runs in Client mode and not in Cluster mode.

B. DATA EXTRACTION & TRANSFORMATION

Extraction:

(i) Twitter Data:

To extract Twitter data, I have used Tweepy library [2]
I used search API to extract data from Twitter
Data such as text, author, time were extracted.

(ii) News Article Data:

To extract News Article data, NewsAPIClient [3] library is used
The text description in NewsAPIClient does not provide the entire content and hence the URL from the response is taken and a call is made to that URL.
From the response, the content provided by the NewsAPIClient is matched and the complete data is fetched.
Data such as content, author, time and the source of the articles were extracted

Transformation:

- In both cases, special characters were removed.
- User mentions and hashtags were removed in case of Twitter data
- Emoticons and URLs were removed
- Data were converted to lowercase characters
- Lemmatization was done on the data and pronouns were removed since they were not useful to the context of this Assignment. This is done so that (schools will be converted to school) words are resolved to their dictionary form
- Data with apostrophe('s) was also removed

Loading:

The transformed data is then pushed to the cloud instance's MongoDB and the count of the data extracted is around 2312.

C. DATA PROCESSING:

Pre-processing:

The text data from MongoDB is extracted and is appended into a single string.

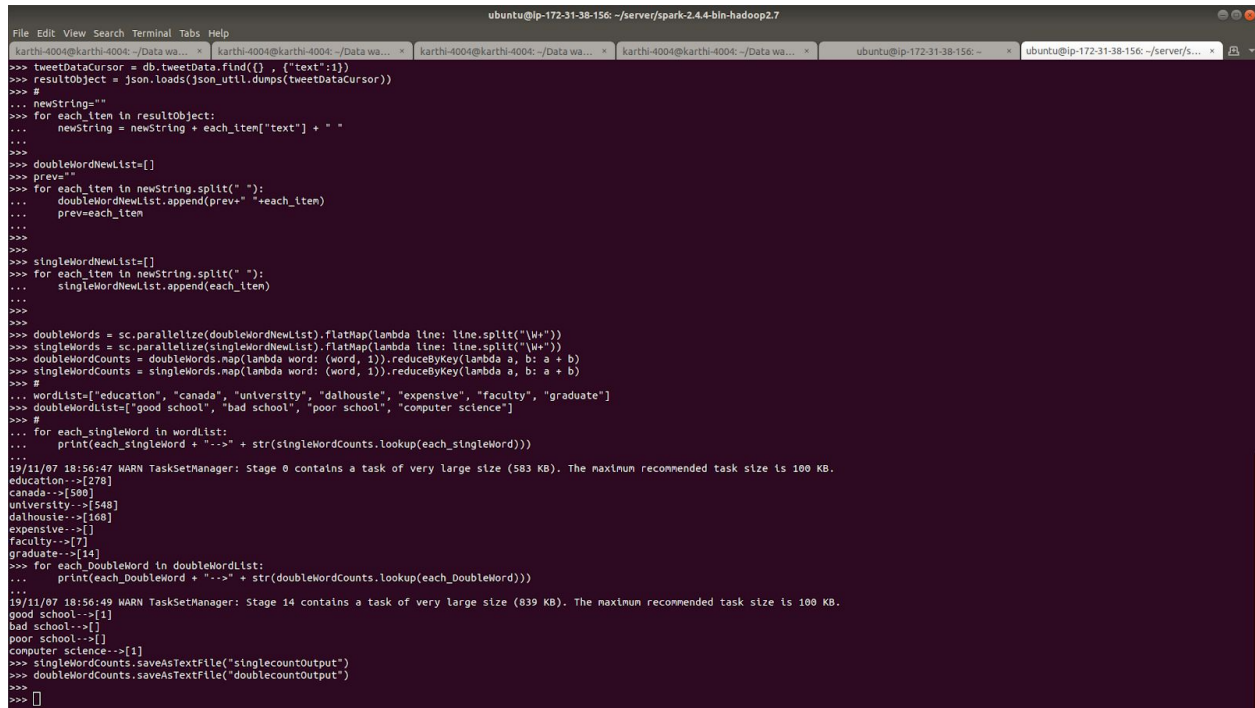
CSCI5408- ASSIGNMENT 2

For finding word count of single words, the string is converted to a list and is fed to the SparkContext for mapReduce operation.

Whereas for finding the word count of two words, the string is converted to two words and is fed to the SparkContext.

Processing:

Data processing is done using the pyspark shell and the output is attached as the screenshot.



```
File Edit View Search Terminal Tabs Help
ubuntu@ip-172-31-38-156: ~/server/spark-2.4.4-bin-hadoop2.7

>>> tweetDataCursor = db.tweetData.find({}, {'text':1})
>>> resultObject = json.loads(json_util.dumps(tweetDataCursor))
>>> #
... newString=""
>>> for each_item in resultObject:
...     newString = newString + each_item["text"] + " "
...
>>> doubleWordNewList=[]
>>> prev=""
>>> for each_item in newString.split(" "):
...     doubleWordNewList.append(prev+" "+each_item)
...     prev=each_item
>>>
>>> singleWordNewList=[]
>>> for each_item in newString.split(" "):
...     singleWordNewList.append(each_item)
>>>
>>> doubleWords = sc.parallelize(doubleWordNewList).flatMap(lambda line: line.split("W+"))
>>> singleWords = sc.parallelize(singleWordNewList).flatMap(lambda line: line.split("W+"))
>>> doubleWordCounts = doubleWords.map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
>>> singleWordCounts = singleWords.map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
>>> #
... wordList=["education", "canada", "university", "dalhousie", "expensive", "faculty", "graduate"]
>>> doubleWordList=["good school", "bad school", "poor school", "computer science"]
>>> #
... for each_singleWord in wordList:
...     print(each_singleWord + "->" + str(singleWordCounts.lookup(each_singleWord)))
...
19/11/07 18:56:47 WARN TaskSetManager: Stage 0 contains a task of very large size (583 KB). The maximum recommended task size is 100 KB.
education->[278]
canada->[598]
university->[548]
dalhousie->[168]
expensive->[1]
faculty->[7]
graduate->[14]
>>> for each_DoubleWord in doubleWordList:
...     print(each_DoubleWord + "->" + str(doubleWordCounts.lookup(each_DoubleWord)))
...
19/11/07 18:56:49 WARN TaskSetManager: Stage 14 contains a task of very large size (839 KB). The maximum recommended task size is 100 KB.
good school->[1]
bad school->[1]
poor school->[1]
computer science->[1]
>>> singleWordCounts.saveAsTextFile("singlecountOutput")
>>> doubleWordCounts.saveAsTextFile("doublecountOutput")
>>>
```

Also attached the text files for reference.

CSCI5408- ASSIGNMENT 2

REFERENCES:

- [1]R. Gupta, H. Pamnani and M. Bhanderi, "Apache Spark", Dalhousie University, 2019.
- [2]"Tweepy Documentation — tweepy 3.8.0 documentation", *Docs.tweepy.org*, 2019. [Online]. Available: <http://docs.tweepy.org/en/latest/>. [Accessed: 01- Nov- 2019].
- [3]"Python client library - News API", *Newsapi.org*, 2019. [Online]. Available: <https://newsapi.org/docs/client-libraries/python>. [Accessed: 01- Nov- 2019].