CSCI 5408 Data Management, Warehousing, and Analytics

Assignment – 2

Banner ID: - B00843607 Darpan Patel

A. Cluster Setup: -

I have created Apache spark cluster in windows [1] (local machine).

Steps:

1. Install JDK 8, set JAVA_HOME = path to jdk file (ex. ‘C:\Program files\Java\jdk’) and PATH = “%JAVA_HOME%\bin”
2. Download “spark-2.4.4-bin-hadoop2.7.tgz” unpack and put the folder at “C:” drive (Or any location of your choice)
3. Set SPARK_HOME = “path where spark folder is located” (ex. ‘C:\spark’) and PATH = “%SPARK_HOME%\bin”
4. Check “pyspark” command in cmd.
5. Go inside “bin” folder in spark.
6. Run ‘spark-class org.apache.spark.deploy.master.Master’ (start Master and note IP and port address)
7. Run ‘spark-class org.apache.spark.deploy.worker.Worker spark://IP:PORT’

B. Twitter and News API Data Extraction and Cleaning: -

` For extracting tweets:

- Created developer account in Twitter
- Create application to get twitter credentials to access tweeter APIs.
- Using “api.search” [2] to get tweets [tw.Cursor(api.search, q=query, tweet_mode="extended").items(numOfTweets)]

For extracting news:

- Create account to use NewsAPI, get API key.
- I have used api.get_everything()[3] to get articles.
  Example

```
api.get_everything(q='University',
                                      from_param='2019-10-05',
                                      to='2019-11-02',
                                       language='en',
                                      sort_by='relevancy',
                                       page_size = 100)
```

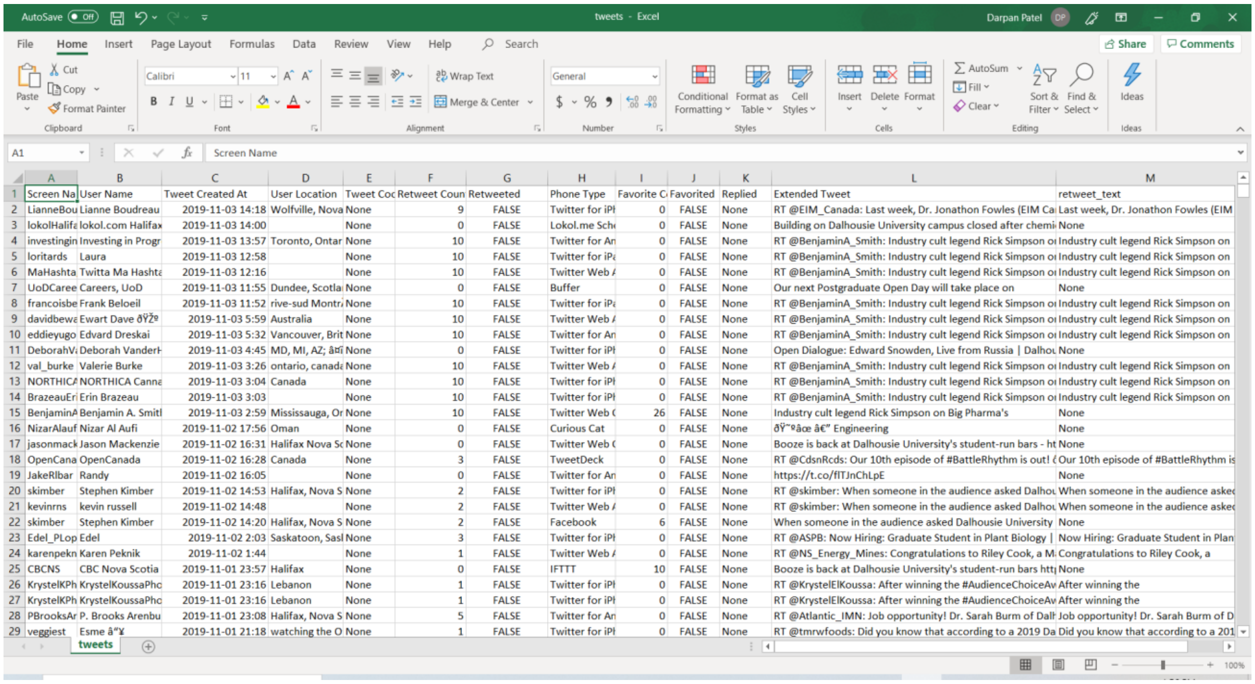
I have extracted total **3572** tweets and **500** news articles.

To clean data I have removed special characters, links, emoticons and converted everything to lower case using regex. After cleaning check for empty values in tweets and news, if present remove that row.

C. Import data to MongoDb: -

- mongoimport --db data --collection tweets --type csv --file C:\Users\darpa\Desktop\Tweepy_tutorial\final_codes\news.csv --headerline
- mongoimport --db data --collection news --type csv --file C:\Users\darpa\Desktop\Tweepy_tutorial\final_codes\tweets.csv --headerline

D. Data file in .CSV: -



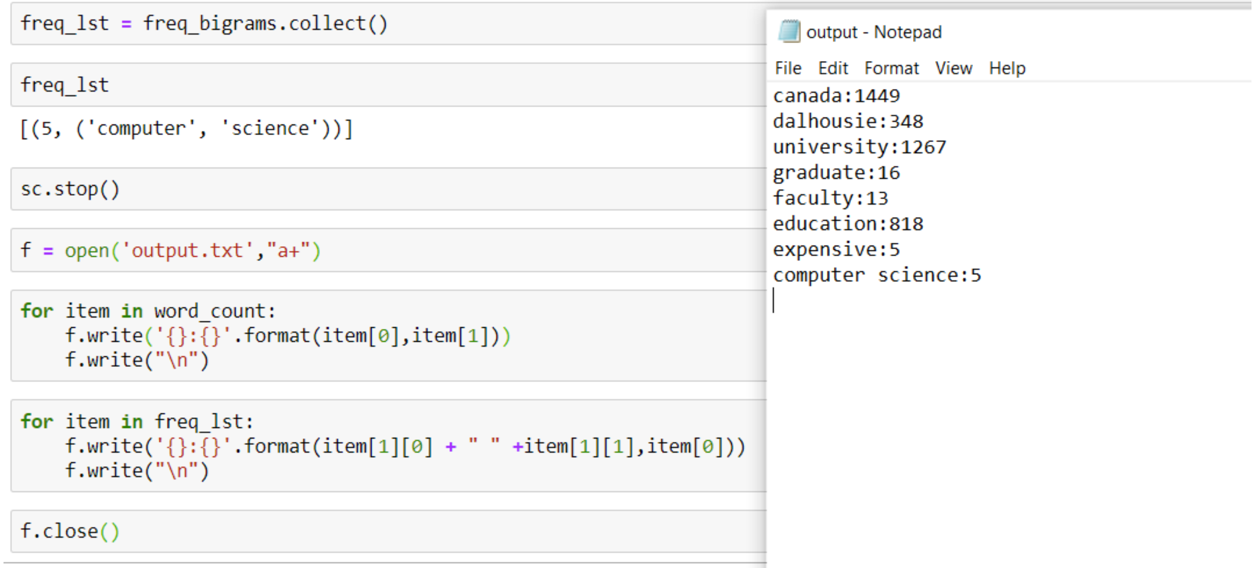
A1 · Screen Name

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Screen Na	User Name	Tweet Created At	User Location	Tweet Coc	Retweet Coun	Retweeted	Phone Type	Favorite C	Favorited	Replied	Extended Tweet	retweet_text
2	LianneBou	Lianne Boudreau	2019-11-03 14:18	Wolfville, Nova	None	9	FALSE	Twitter for iPh	0	FALSE	None	RT @EIM_Canada: Last week, Dr. Jonathon Fowles (EIM Ca	Last week, Dr. Jonathon Fowles (EIM
3	lokolHalifa	lokol.com Halifax	2019-11-03 14:00		None	0	FALSE	Lokol.me Sch	0	FALSE	None	Building on Dalhousie University campus closed after chemi	None
4	investingin	Investing in Progr	2019-11-03 13:57	Toronto, Ontar	None	10	FALSE	Twitter for An	0	FALSE	None	RT @BenjaminA_Smith: Industry cult legend Rick Simpson o	Industry cult legend Rick Simpson on
5	loritards	Laura	2019-11-03 12:58		None	10	FALSE	Twitter for iPa	0	FALSE	None	RT @BenjaminA_Smith: Industry cult legend Rick Simpson o	Industry cult legend Rick Simpson on
6	MaHashta	Twitta Ma Hashta	2019-11-03 12:16		None	10	FALSE	Twitter Web A	0	FALSE	None	RT @BenjaminA_Smith: Industry cult legend Rick Simpson o	Industry cult legend Rick Simpson on
7	UoDCaree	Careers, UoD	2019-11-03 11:55	Dundee, Scotla	None	0	FALSE	Buffer	0	FALSE	None	Our next Postgraduate Open Day will take place on	None
8	francoisbe	Frank Beloeil	2019-11-03 11:52	rive-sud Montr	None	10	FALSE	Twitter for iPa	0	FALSE	None	RT @BenjaminA_Smith: Industry cult legend Rick Simpson o	Industry cult legend Rick Simpson on
9	davidbewa	Ewart Dave ðŸŽº	2019-11-03 5:59	Australia	None	10	FALSE	Twitter Web A	0	FALSE	None	RT @BenjaminA_Smith: Industry cult legend Rick Simpson o	Industry cult legend Rick Simpson on
10	eddieyugo	Edvard Dreskai	2019-11-03 5:32	Vancouver, Brit	None	10	FALSE	Twitter for An	0	FALSE	None	RT @BenjaminA_Smith: Industry cult legend Rick Simpson o	Industry cult legend Rick Simpson on
11	DeborahVa	Deborah VanderH	2019-11-03 4:45	MD, MI, AZ; â¤ï	None	0	FALSE	Twitter for iPh	0	FALSE	None	Open Dialogue: Edward Snowden, Live from Russia | Dalhou	None
12	val_burke	Valerie Burke	2019-11-03 3:26	ontario, canada	None	10	FALSE	Twitter Web A	0	FALSE	None	RT @BenjaminA_Smith: Industry cult legend Rick Simpson o	Industry cult legend Rick Simpson on
13	NORTHICA	NORTHICA Canna	2019-11-03 3:04	Canada	None	10	FALSE	Twitter for iPh	0	FALSE	None	RT @BenjaminA_Smith: Industry cult legend Rick Simpson o	Industry cult legend Rick Simpson on
14	BrazeauEri	Erin Brazeau	2019-11-03 3:03		None	10	FALSE	Twitter for iPh	0	FALSE	None	RT @BenjaminA_Smith: Industry cult legend Rick Simpson o	Industry cult legend Rick Simpson on
15	BenjaminA	Benjamin A. Smitl	2019-11-03 2:59	Mississauga, Or	None	10	FALSE	Twitter Web C	26	FALSE	None	Industry cult legend Rick Simpson on Big Pharma's	None
16	NizarAlauf	Nizar Al Aufi	2019-11-02 17:56	Oman	None	0	FALSE	Curious Cat	0	FALSE	None	ðŸ˜ºâœ â€” Engineering	None
17	jasonmack	Jason Mackenzie	2019-11-02 16:31	Halifax Nova Sc	None	0	FALSE	Twitter Web C	0	FALSE	None	Booze is back at Dalhousie University's student-run bars - ht	None
18	OpenCana	OpenCanada	2019-11-02 16:28	Canada	None	3	FALSE	TweetDeck	0	FALSE	None	RT @CdsnRcds: Our 10th episode of #BattleRhythm is out! ð	Our 10th episode of #BattleRhythm is
19	JakeRlbar	Randy	2019-11-02 16:05		None	0	FALSE	Twitter for An	0	FALSE	None	https://t.co/flTJnChLpE	None
20	skimber	Stephen Kimber	2019-11-02 14:53	Halifax, Nova S	None	2	FALSE	Twitter for iPh	0	FALSE	None	RT @skimber: When someone in the audience asked Dalhou	When someone in the audience asked
21	kevinrns	kevin russell	2019-11-02 14:48		None	2	FALSE	Twitter Web A	0	FALSE	None	RT @skimber: When someone in the audience asked Dalhou	When someone in the audience asked
22	skimber	Stephen Kimber	2019-11-02 14:20	Halifax, Nova S	None	2	FALSE	Facebook	6	FALSE	None	When someone in the audience asked Dalhousie University	None
23	Edel_PLop	Edel	2019-11-02 2:03	Saskatoon, Sasl	None	3	FALSE	Twitter for iPh	0	FALSE	None	RT @ASPB: Now Hiring: Graduate Student in Plant Biology |	Now Hiring: Graduate Student in Plan
24	karenpekn	Karen Peknik	2019-11-02 1:44		None	1	FALSE	Twitter Web A	0	FALSE	None	RT @NS_Energy_Mines: Congratulations to Riley Cook, a M	Congratulations to Riley Cook, a
25	CBCNS	CBC Nova Scotia	2019-11-01 23:57	Halifax	None	0	FALSE	IFTTT	10	FALSE	None	Booze is back at Dalhousie University's student-run bars htt	None
26	KrystelKPh	KrystelKoussaPho	2019-11-01 23:16	Lebanon	None	1	FALSE	Twitter for iPh	0	FALSE	None	RT @KrystelElKoussa: After winning the #AudienceChoiceAv	After winning the
27	KrystelKPh	KrystelKoussaPho	2019-11-01 23:16	Lebanon	None	1	FALSE	Twitter for iPh	0	FALSE	None	RT @KrystelElKoussa: After winning the #AudienceChoiceAv	After winning the
28	PBrooksAr	P. Brooks Arenbu	2019-11-01 23:08	Halifax, Nova S	None	5	FALSE	Twitter for An	0	FALSE	None	RT @Atlantic_IMN: Job opportunity! Dr. Sarah Burm of Dalh	Job opportunity! Dr. Sarah Burm of D
29	veggiest	Esme â"¥	2019-11-01 21:18	watching the O	None	1	FALSE	Twitter for iPh	0	FALSE	None	RT @tmrwfoods: Did you know that according to a 2019 Da	Did you know that according to a 201

Figure 1: Tweets in csv format

E. Data Processing: -

To process data, first I have merged ‘tweets.csv’ and ‘news.csv’ with axis = 0 in dataframe. Now, using flatMap and reduceByKey in spark I have calcuted frequency for 1-Grams (Canada, Dalhousie, university, graduate, faculty, education, expensive) and Bi-Grams (good school, bad schools, good schools, bad school, computer science). Whole program was run on cluster with 1 slave node (14.7 Gb RAM assigned)



```
freq_lst = freq_bigrams.collect()
```

```
freq_lst
```

```
[(5, ('computer', 'science'))]
```

```
sc.stop()
```

```
f = open('output.txt',"a+")
```

```
for item in word_count:
    f.write('{}:{}'.format(item[0],item[1]))
    f.write("\n")
```

```
for item in freq_lst:
    f.write('{}:{}'.format(item[1][0] + " " +item[1][1],item[0]))
    f.write("\n")
```

```
f.close()
```

```
canada:1449
dalhousie:348
university:1267
graduate:16
faculty:13
education:818
expensive:5
computer science:5
```

Figure 2 - Frequency Count Output

References

[1] “Installing Apache Spark(PySpark)” *Meduim*. [Online]. Available: https://medium.com/@loldja/installing-apache-spark-pyspark-the-missing-quick-start-guide-for-windows-ad81702ba62d.[Accessed: 27-Oct-2019].

[2] “tweepy.api – Tweeter API wrapper” Available: http://docs.tweepy.org/en/latest/api.html [Accessed: 01-Oct-2019].

[3] “Python client library,” *News API*. [Online]. Available: https://newsapi.org/docs/client-libraries/python. [Accessed: 01-Oct-2019].

[4] *PySpark – Word Count Example*. [Online]. Available: https://pythonexamples.org/pyspark-word-count-example/. [Accessed: 03-Nov-2019].