1818
2018

DALHOUSIE UNIVERSITY

Assignment 3
Report-2

CSCI5408 – Data Warehousing & Analytics

Submitted by:

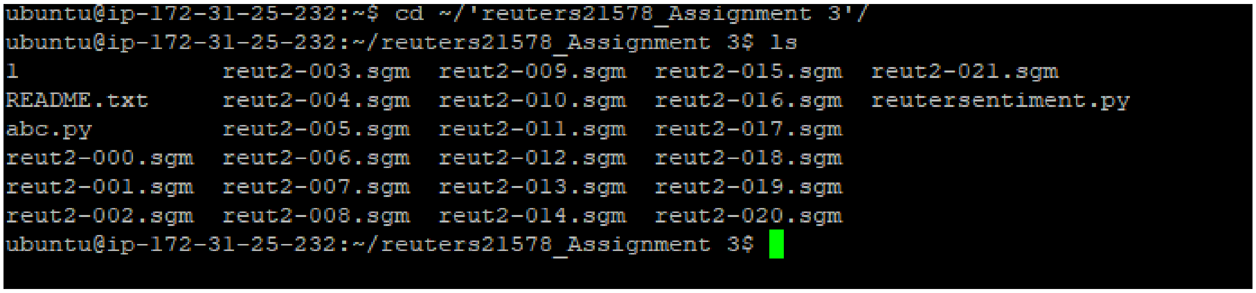
Shivam Gupta

B00810723

Shivam.Gupta@dal.ca

Data Upload

- I have used my cloud instance to gather tweets, and I worked on the same tweets which were fetched in assignment – 2.
- I have used WINSCP tool to upload all the Reuters data into my cloud instance.



```
ubuntu@ip-172-31-25-232:~$ cd ~/'reuters21578_Assignment 3'/
ubuntu@ip-172-31-25-232:~/reuters21578_Assignment 3$ ls
1               reut2-003.sgm  reut2-009.sgm  reut2-015.sgm  reut2-021.sgm
README.txt      reut2-004.sgm  reut2-010.sgm  reut2-016.sgm  reutersentiment.py
abc.py          reut2-005.sgm  reut2-011.sgm  reut2-017.sgm
reut2-000.sgm   reut2-006.sgm  reut2-012.sgm  reut2-018.sgm
reut2-001.sgm   reut2-007.sgm  reut2-013.sgm  reut2-019.sgm
reut2-002.sgm   reut2-008.sgm  reut2-014.sgm  reut2-020.sgm
ubuntu@ip-172-31-25-232:~/reuters21578_Assignment 3$
```

- I have done analysis of both data on my cloud instance.

Data Extraction, Transformation & Analysis for tweets:

- I have done everything using vim editor and python and written a script to perform sentiment analysis for Stream data and Search data.
- I have not considered any metadata and just the text part of tweets.
- Attaching the code in the submission.
- Files Attached:

positivewords.txt negativewords.txt

List of all the positive and negative words in a text file in cloud instance.

searchsentiment.py.txt Streamsentiment.py.txt

Python scripts for Search and stream tweets.

searchsentimentoutput.txt streamsentimentoutput.txt

Output for python files.

Data Extraction, Transformation & Analysis for Reuters:

- I have cleaned the documents in a python scripts and fetch all the data from the sgm files.
- Clean chunk of text is considered in the script, however new document is not created but only the text part of body is considered for all the articles.
- I have computed the TF and IDF and TFIDF in the python script and ‘Canada’ is searched in all the files.
- All the documents having ‘Canada’ are ranked according to TF-IDF value.
- At the end, all the sentences were fetched, where “Canada” is present according to the TF-IDF value.

reutersentiment.py.
txt

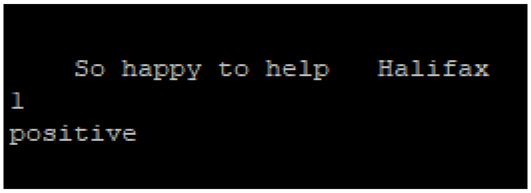
Stream tweets data:

- Positive tweets: 442
- Negative tweets: 3077
- Neutral tweets: 3300
- Total tweets: 6819

Search tweets data:

- Positive tweets: 801
- Negative tweets: 747
- Neutral tweets: 4052
- Total tweets: 5600

Sample tweet with polarity: here 1 is polarity.



```
    So happy to help   Halifax
1
positive
```

For Reuters data, I have provided the output file with all the extracted sentences.

References:

[1] Positive words, https://gist.github.com/mkulakowski2/4289437

[2] Negative words, https://gist.github.com/mkulakowski2/4289441

[3] Stop words, https://gist.github.com/sebleier/554280

[4] TF IDF calculation, https://www.elephate.com/blog/what-is-tf-idf/

[5] Assignment-2 twitter tweets

[6] www.stackoverflow.com**:** Basic error resolution

[7] WINSCP tool, https://winscp.net/eng/docs/start