DALHOUSIE UNIVERSITY

CSCI-5408 Data Management, Warehousing and Analytics

Assignment: 3

Name: Menni Prem Kumar (B00843422)

Date: December 5, 2019

Sentiment Analysis:

I have used CSCI_5408 assignment_2 twitter data extract script [1] to fetch tweets data. Using regular expressions, tweets data is cleaned by removing special characters and URL's. Result of the script is a csv file contains tweet text. I have downloaded positive and negative list of words from online source [2]. Using these words, Sentiment analysis was performed on each tweet based on these words.

Since, Tweet can be positive, negative and neutral. A bag dictionary is created, each tweet stored in the form of key value pairs (Keys and words, values as word count in the tweet). A tweet can have both negative and positive words, having a greater number of positive words in a tweet cannot guarantee that the tweet is positive and the same for negative words. Considering this method, I have created a csv file containing Tweet, word matched with positive and negative words list, tweet polarity considering that word.

I have considered tweets which are having neither of the positive and negative words as neutral and printed in the csv file. Please find the figure 1 which shows sample dictionary record created.

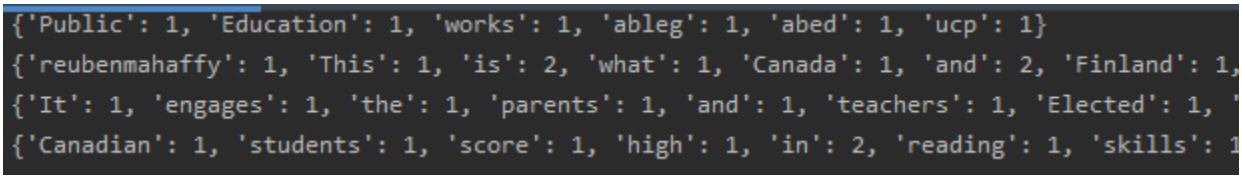

Figure: 1

After tagging each tweet with positive and negative words, tweet with no such positive and negative words as neutral. I have created an csv file with the details about tweet, word matched, polarity of tweet. Please find the sample result in figure 2.

	Tweet	Message	Match	Polarity
	1	Halifax police apologize to black community for pain caused by streetchecks	pain	negative
	2	Halifax restaurateur who pioneered citys slow food scene dies at 59	slow	negative
	2	Halifax restaurateur who pioneered citys slow food scene dies at 59	dies	negative
	3	My drip is talent saving is not boring so I might go Halifax a uni student	talent	positive
	3	My drip is talent saving is not boring so I might go Halifax a uni student	boring	negative
	4	My drip is talent saving is not boring so I might go Halifax	talent	positive
	4	My drip is talent saving is not boring so I might go Halifax	boring	negative
	5	ArcticNet2019 will exam research across Canadas Northern and Arctic Region focussing on current challenges they face Clima		neutral
0	6	Congratulations to the Bassett Wrestling team on their 5418 victory over GW Danville and their 5415 victory over H	victory	positive
1	7	Survived quick flight from Charlottetown to Halifax One more to go before heading to Montreal		neutral
2	8	Man Tasered by Halifax police after allegedly assaulting officer		neutral

Figure: 2

Using Tableau and csv file generated, I have performed frequency count and visualised the positive and negative words in word cloud. Please find the word cloud picture in figure 3.

Figure: 3

Semantic Analysis:

I have used CSCI_5408 assignment_2 news article data extract script [1] to fetch news data. Using regular expressions, special characters and URL's were removed from news article. Result of the script is to create each text file (news file) containing article- title, content, description for 500 news articles.

Computing TF-IDF (Term frequency – Inverse Document frequency):

Considering 500 text files generated from the extraction script as 500 input files, I have searched for the key words Canada, Halifax, Canada Education, Dalhousie University in each document and created an csv file containing data of key word, key word containing document count, total documents – number of documents key word appeared, log of document

frequency. I have printed only the key words which occurred in the articles. Please find figure 4 which shows TF-IDF for article data.

Note: Document frequency is (total documents / number of documents key word appeared)

Total Documents	500		
Search Query	Document containing term(df)	Total Documents(N)/ number of documents term appeared(df)	Login(N/df)
canada	62	500/62	2.087473713
university	56	500/56	2.189256408
dalhousie university	13	500/13	3.649658741
halifax	35	500/35	2.659260037

Figure: 4

Occurrence of Canada:

I have performed frequency count of Canada by reading each file as an input and created an csv file containing data of the file name which contains Canada in it and total words in the file, frequency of Canada in it. Please find figure 5 which shows sample lines of data.

	A	B	C	D
1	Term	Canada		
2	canada appeared in 50documents	Total words (m)	Frequency (f)	Relative frequency (Rf)
3	article/article8.txt	68	1	0.014705882
4	article/article13.txt	62	1	0.016129032
5	article/article15.txt	72	2	0.027777778
6	article/article62.txt	49	1	0.020408163
7	article/article84.txt	52	3	0.057692308
8	article/article101.txt	66	1	0.015151515
9	article/article104.txt	68	1	0.014705882
10	article/article108.txt	68	1	0.014705882
11	article/article114.txt	74	2	0.027027027
12	article/article115.txt	76	1	0.013157895

Figure: 5

Relative frequency:

Using total words and frequency of Canada in the file, relative frequency was calculated by dividing total words in the file with frequency count. The article with high relative frequency will be printed as an output. Please find the figure 6 which shows output includes article, file name, relative frequency.

```
C:\Users\premk\AppData\Local\Programs\Python\Python37-32\python.exe "C:/Users/premk/Desktop/First Term/Data Warehousing, Analy
File name:  article/article159.txt
Relative frequency:  0.10526315789473684
ContentNoneDisney ranks No 1 in App Store in US CanadaDisney ranks No 1 in App Store in US Canada

Process finished with exit code 0
```

Figure: 6

Business Intelligence:

I have considered departments, courses, programs, faculty as main facts to be analyzed. In departments, dimensions were Engineering - medicine are dimension table. In courses I have considered Undergraduate, graduate, professional are dimensions. In programs, I considered MACS, MCS etc., as dimensions. For faculty, name, mail id, office considered as dimensions. Attributes for each dimension are like name, description, date, code, contact info etc.,

Using the above dimensions and attributes, I have connected my AWS hosted database to Cognos BI and generated the schema for complete attributes. Please find the star schema below in figure 7.

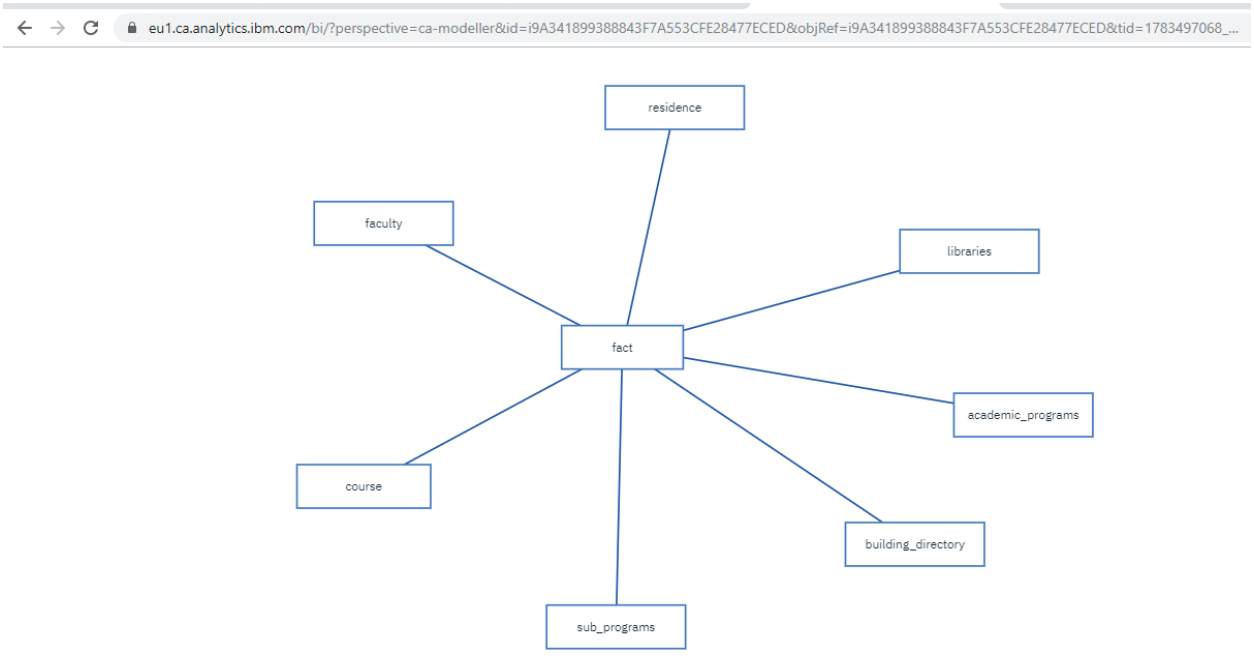

Figure: 7

Using Cognos BI framework (OLAP), I have calculated the number of programs provided by each department and Medicine is the highest to provide different types of programs. Computer

science do not provide highest number of programs. Please find the output of program count in figure 8.

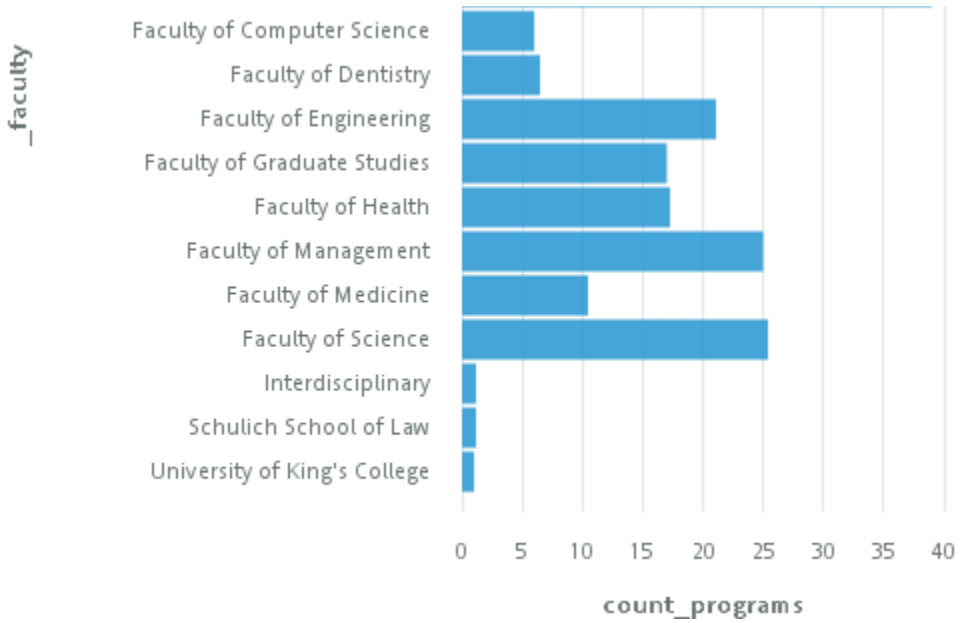

Figure: 8

Courses which are present in each department were provided in the figure 9.

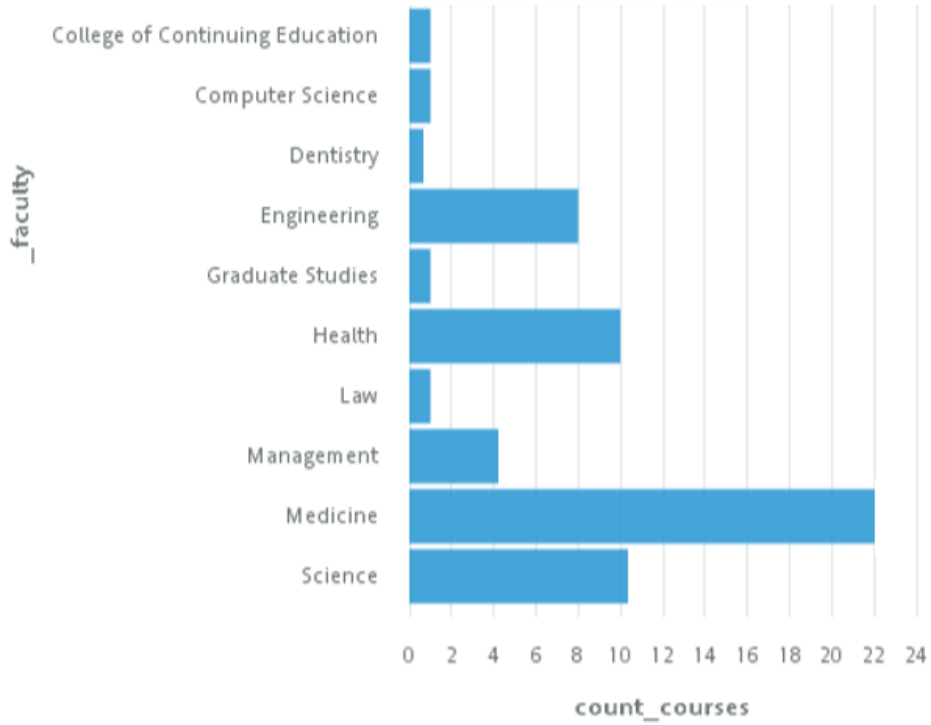

Figure: 9

References:

[1] Menni Prem Kumar. " Assignment -2" Submitted for CSCI 5408, Nov. 06, 2019. [Accessed: 02- Dec- 2019].

[2] Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

[3] "API reference index", Developer.twitter.com, 2019. [Online]. Available: https://developer.twitter.com/en/docs/api-reference-index.

[4] "Tweepy Documentation — tweepy 3.8.0 documentation", Tweepy.readthedocs.io, 2019. [Online]. Available: https://tweepy.readthedocs.io/en/latest/.

[5] "Documentation - News API", Newsapi.org, 2019. [Online]. Available: https://newsapi.org/docs.

[6] "IBM Cognos Analytics", IBM Cognos Analytics, 2019. [Online]. Available: https://www.ibm.com/ca-en/products/cognos-analytics. [Accessed: 02- Dec- 2019].

[7] "Data visualization", Tableau Software, 2019. [Online]. Available: https://www.tableau.com/learn/articles/data-visualization. [Accessed: 02- Dec- 2019].

[8] "Dalhousie University - Halifax, Nova Scotia, Canada", Dalhousie University, 2019. [Online]. Available: https://www.dal.ca/.

[9] Menni Prem Kumar. " Assignment -1" Submitted for CSCI 5408, Sep. 29, 2019. [Accessed: 02- Dec- 2019].