DALHOUSIE
UNIVERSITY
Faculty of Computer Science

CSCI 5408 Data Management, Warehousing and Analytics

ASSIGNMENT 3

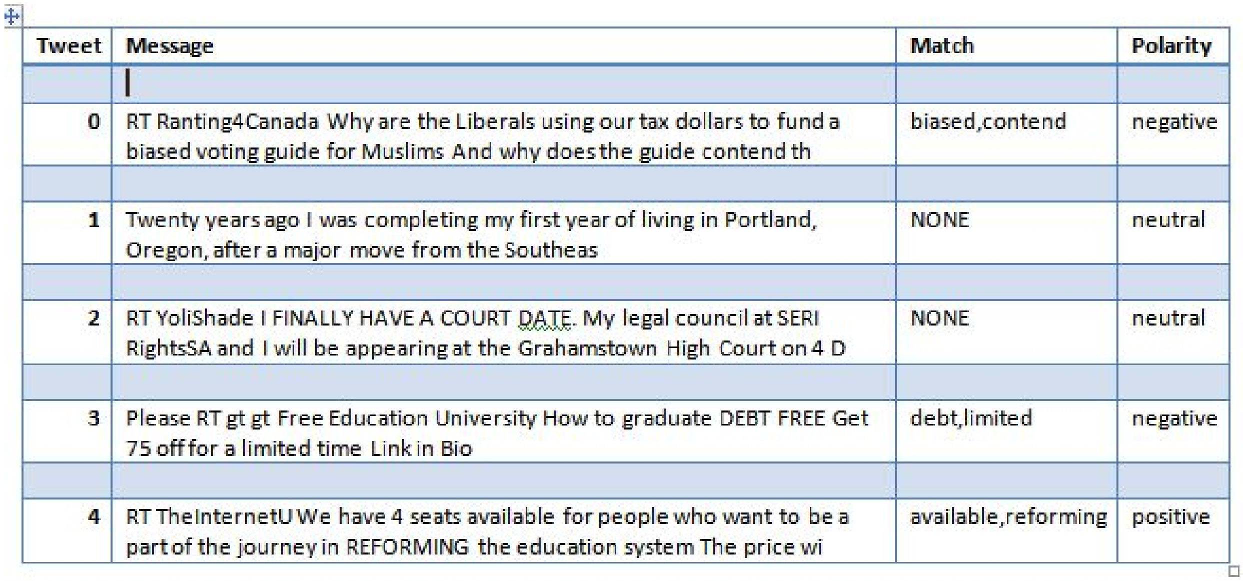
MADE BY: Zeel Shah(B00829477)

A. SENTIMENT ANALYSIS

This task is performed by the tweets which were extracted in Assignment 2. I have already cleaned the tweets with the help of Regex and python script to remove URL and/or any special characters.

Now, I have downloaded the text files of positive and negative words separately, which are attached as positive-words.txt and negative-words.txt. Bag of words[3] for each tweet is created and tweet text is now compared with the words in both files, if any match occurs from positive words file then it will be appended in list of positive words[5] and similarly for negative words[6].

For each tweet, match will be found from positive and negative words both. Count is done for positive and negative words such as positive_count and negative_count. To find the polarity, if positive_count is greater and negative_count, then polarity will be positive. If negative_count is greater than positive_count, then polarity will be negative. If both are equal then, polarity will be neutral. All these are stored in the format given such as: Tweet, Message, Match, Polarity.



Tweet	Message	Match	Polarity
0	RT Ranting4Canada Why are the Liberals using our tax dollars to fund a biased voting guide for Muslims And why does the guide contend th	biased,contend	negative
1	Twenty years ago I was completing my first year of living in Portland, Oregon, after a major move from the Southeas	NONE	neutral
2	RT YoliShade I FINALLY HAVE A COURT DATE. My legal council at SERI RightsSA and I will be appearing at the Grahamstown High Court on 4 D	NONE	neutral
3	Please RT gt gt Free Education University How to graduate DEBT FREE Get 75 off for a limited time Link in Bio	debt,limited	negative
4	RT TheInternetU We have 4 seats available for people who want to be a part of the journey in REFORMING the education system The price wi	available,reforming	positive

Fig 1. Tweets Match and Polarity

It is stored in the csv file 'Output_polarity1.csv' which is attached.

Now, I found the count of each match of positive word found from all the tweets, stored it as frequency of that particular word and stored it in the csv file 'Output_file_new.csv'.

Similarly, the frequency of each negative word is stored in the csv file. This is used to visualize word cloud in Tableau[1].

Frequently occurring words in the positive and negative tweets collected in a word cloud using Tableau is shown as below.

Also, I have visualized word cloud[4] individually for positive and negative words.

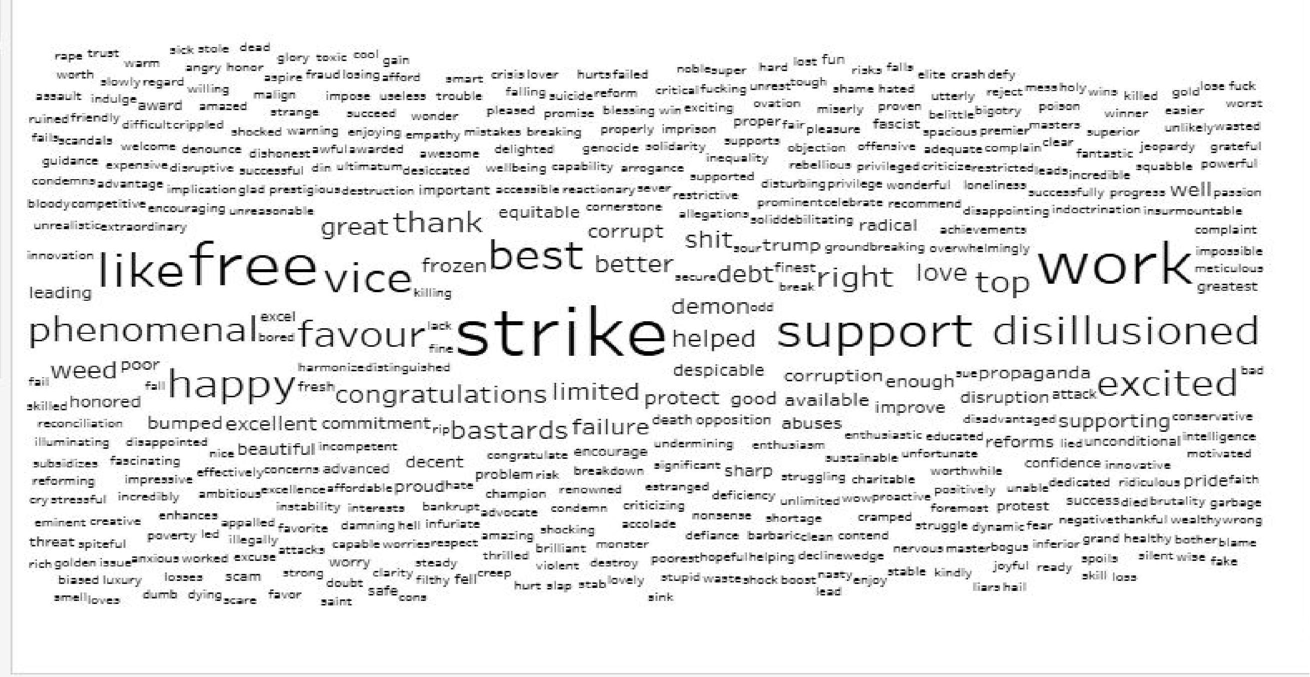


Fig 2. Word Cloud of positive and negative words

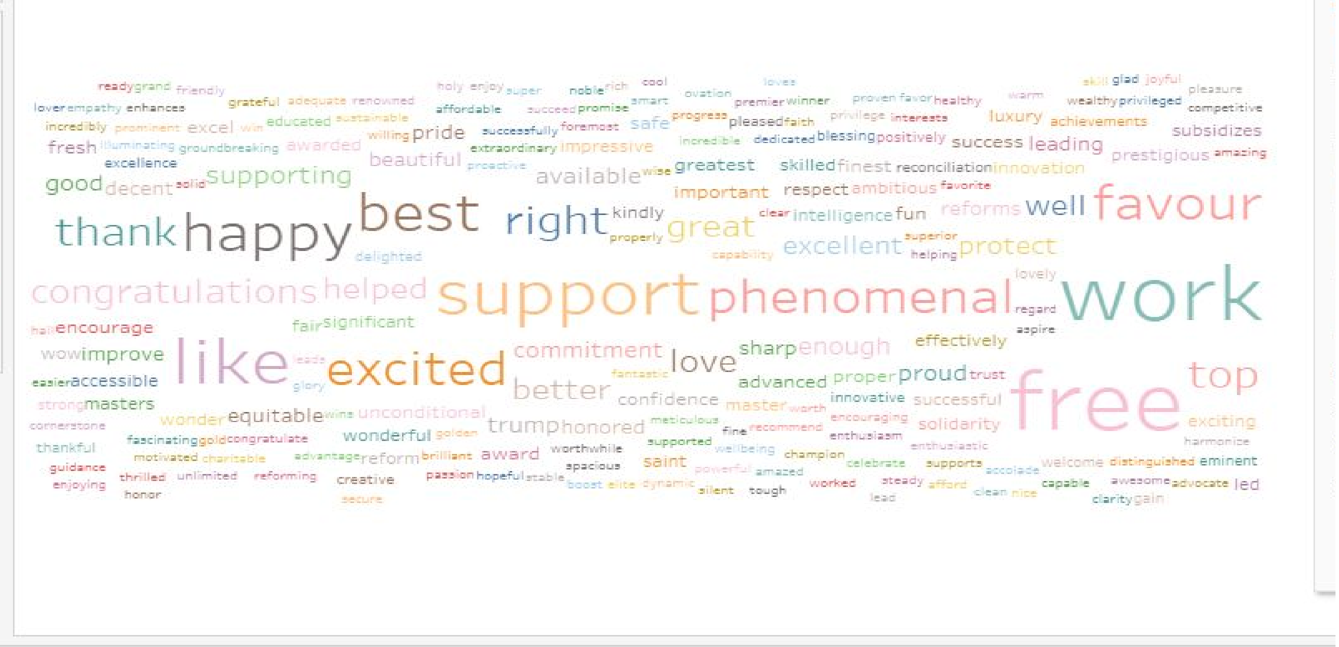


Fig 3. Word Cloud of positive words

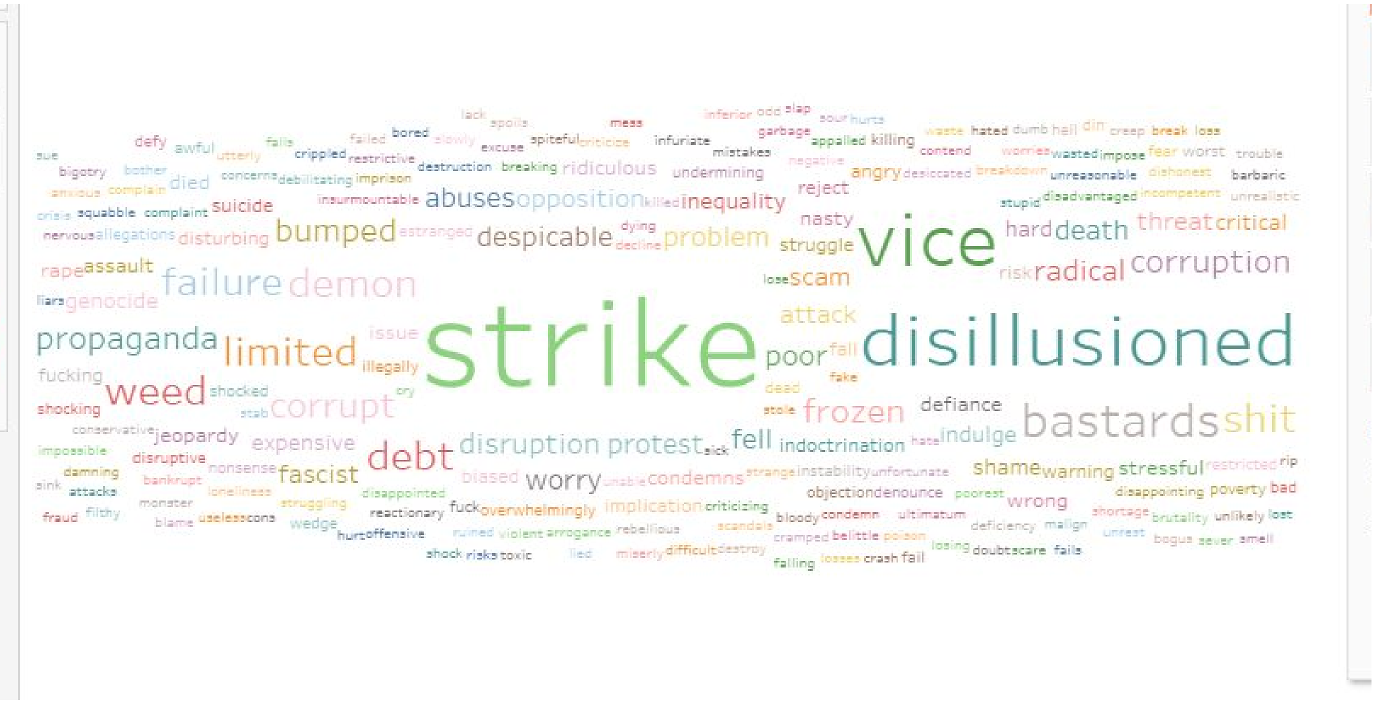


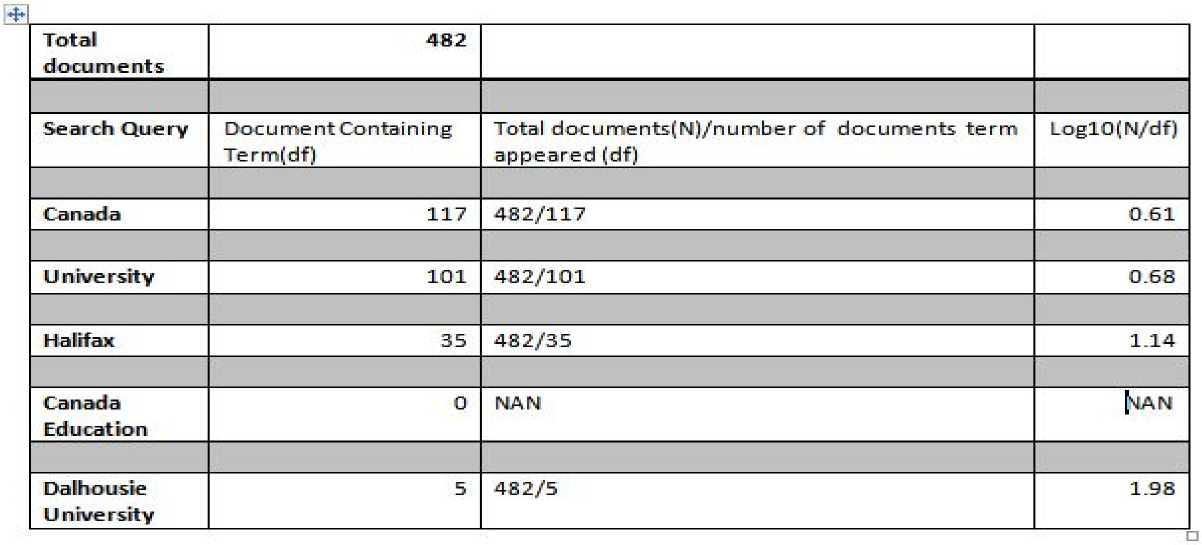
Fig 4. Word Cloud of negative words

B. SEMANTIC ANALYSIS

I have used the script from the assignment 2 for news article data which is already cleaned using regex and python script. For each article of news API, a separate text file is created for each one of them which contains Title, Content and Description and the total files generated were 482.

Occurence of words “Canada”, “University”, “Dalhousie University”, “Halifax”, “Canada Education” is found from all the documents and Document containing term(df) , Total Documents(N)/ number of documents term appeared (df) and Log10(N/df) is calculated.

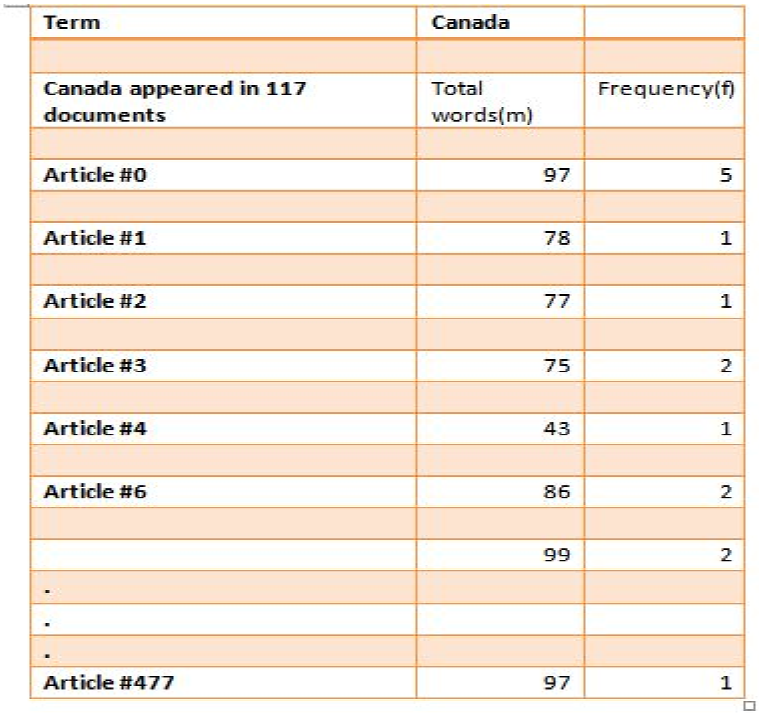
It is shown in the table given below.



Total documents	482		
Search Query	Document Containing Term(df)	Total documents(N)/number of documents term appeared (df)	Log10(N/df)
Canada	117	482/117	0.61
University	101	482/101	0.68
Halifax	35	482/35	1.14
Canada Education	0	NAN	NAN
Dalhousie University	5	482/5	1.98

Fig 5. TF-IDF for Search Keywords

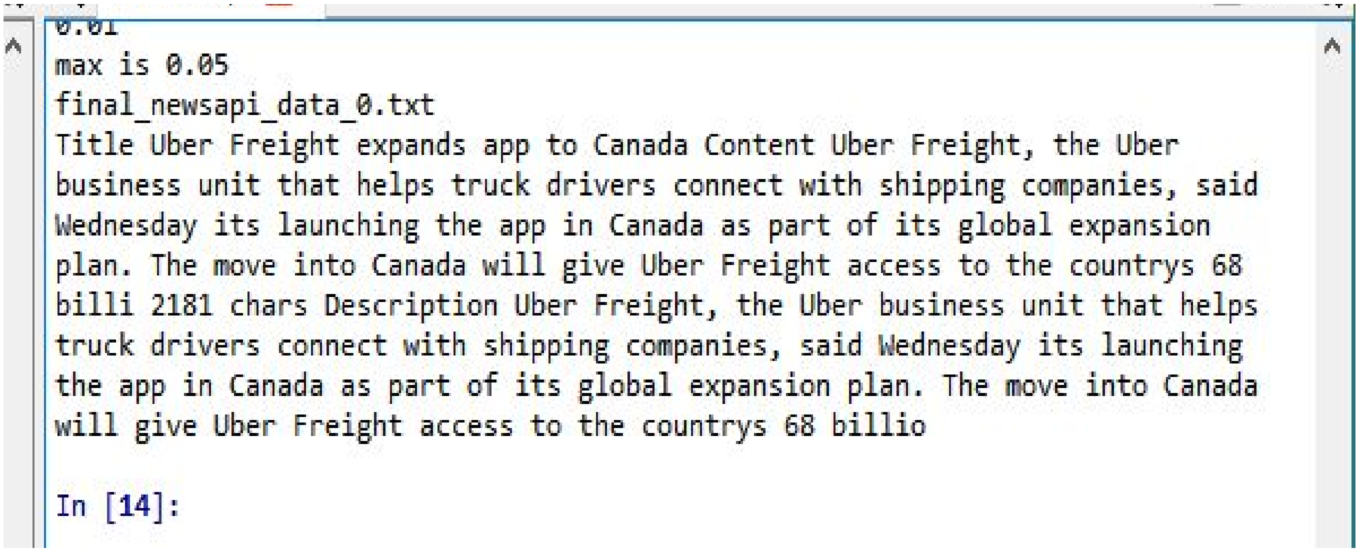
Next, “Canada” was found in total 117 documents and its frequency count is also made which is shown in the table below.



Term	Canada	
Canada appeared in 117 documents	Total words(m)	Frequency(f)
Article #0	97	5
Article #1	78	1
Article #2	77	1
Article #3	75	2
Article #4	43	1
Article #6	86	2
	99	2
.		
.		
.		
Article #477	97	1

Fig 6. Frequency Count of Canada for particular document

Now, by calculating the highest value of (f/m), I have printed the file name and news article and it is shown as below:



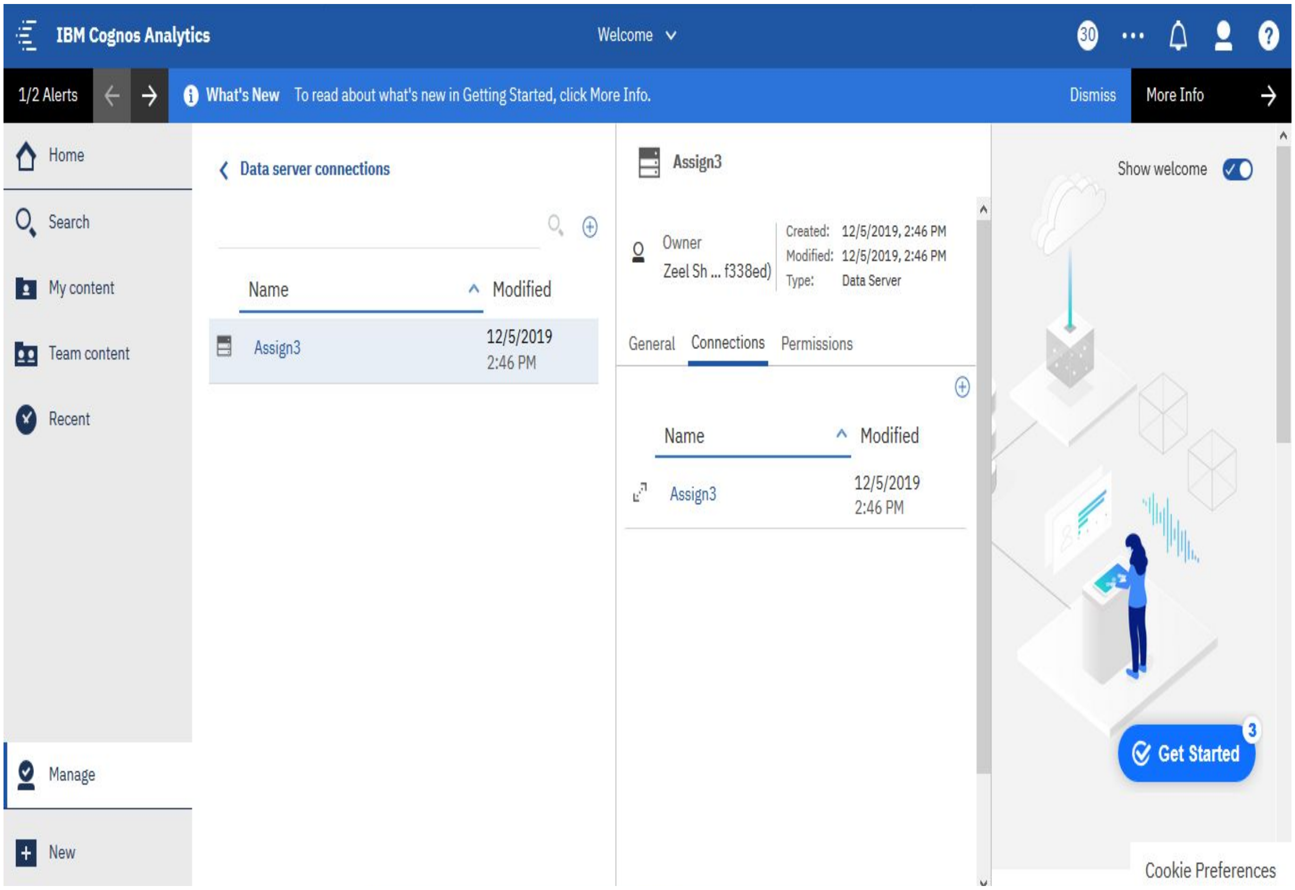
```
0.01
max is 0.05
final_newsapi_data_0.txt
Title Uber Freight expands app to Canada Content Uber Freight, the Uber
business unit that helps truck drivers connect with shipping companies, said
Wednesday its launching the app in Canada as part of its global expansion
plan. The move into Canada will give Uber Freight access to the countrys 68
billi 2181 chars Description Uber Freight, the Uber business unit that helps
truck drivers connect with shipping companies, said Wednesday its launching
the app in Canada as part of its global expansion plan. The move into Canada
will give Uber Freight access to the countrys 68 billio

In [14]:
```

Fig 7. News Article for highest value of Canada Count

C. BUSINESS INTELLIGENCE

For this task, Cognos BI setup was done during Lab session under the guidance of TA's. The connection was done properly with database and the dashboard of Cognos BI[2] is shown as below:



To read about what's new in Getting Started, click More Info.

Name	Modified
Assign3	12/5/2019 2:46 PM

Assign3

Owner
Zeel Sh ... f338ed)

Created: 12/5/2019, 2:46 PM
Modified: 12/5/2019, 2:46 PM
Type: Data Server

Name	Modified
Assign3	12/5/2019 2:46 PM

Fig 8. Dashboard of Cognos BI.

With the data scraped from assignment 1 of Dalhousie University, I have tried to deduce facts and dimension tables. I have found that 'Student' as fact table and dalcard, program, programdetails, faculty and employee dimension tables. Moreover, health table is another fact table and it has services as its dimension table.

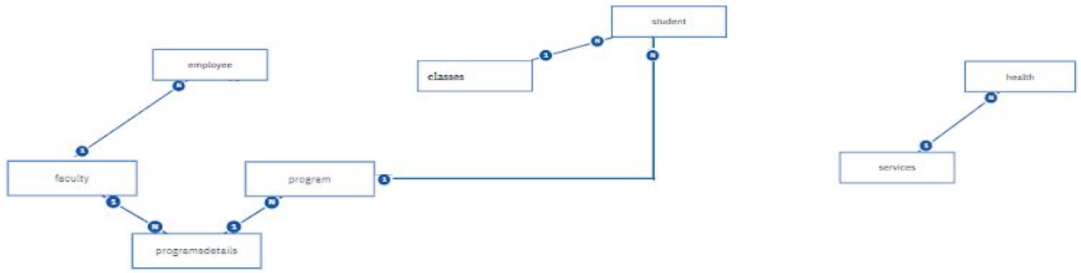


Fig 9. Snowflake schema for data of Dalhousie University.

The dimensions of the tables used are:

1. Student: Student_ID, First_name, Last_name, Course_ID
2. Classes: Student_ID, Class_Number
3. Program: Program_code, Program_description, level
4. Programdetails: Program_ID, Program_name,Program_type, Faculty_Name
5. Faculty: Faculty_ID, Faculty_name
6. Employee: Employee_Id, First_name, Last_name, Position, Faculty_Name
7. Health: Service_Id, Medical_Services, Location
8. Services: Service_Id, Services_name

Program table is divided into programdetails and it is linked to faculty table and this makes hierarchical schema.

Now, report is created using Cognos BI with the help of bar graph which represents the data that how many programs are present in each faculty and it is shown as below:

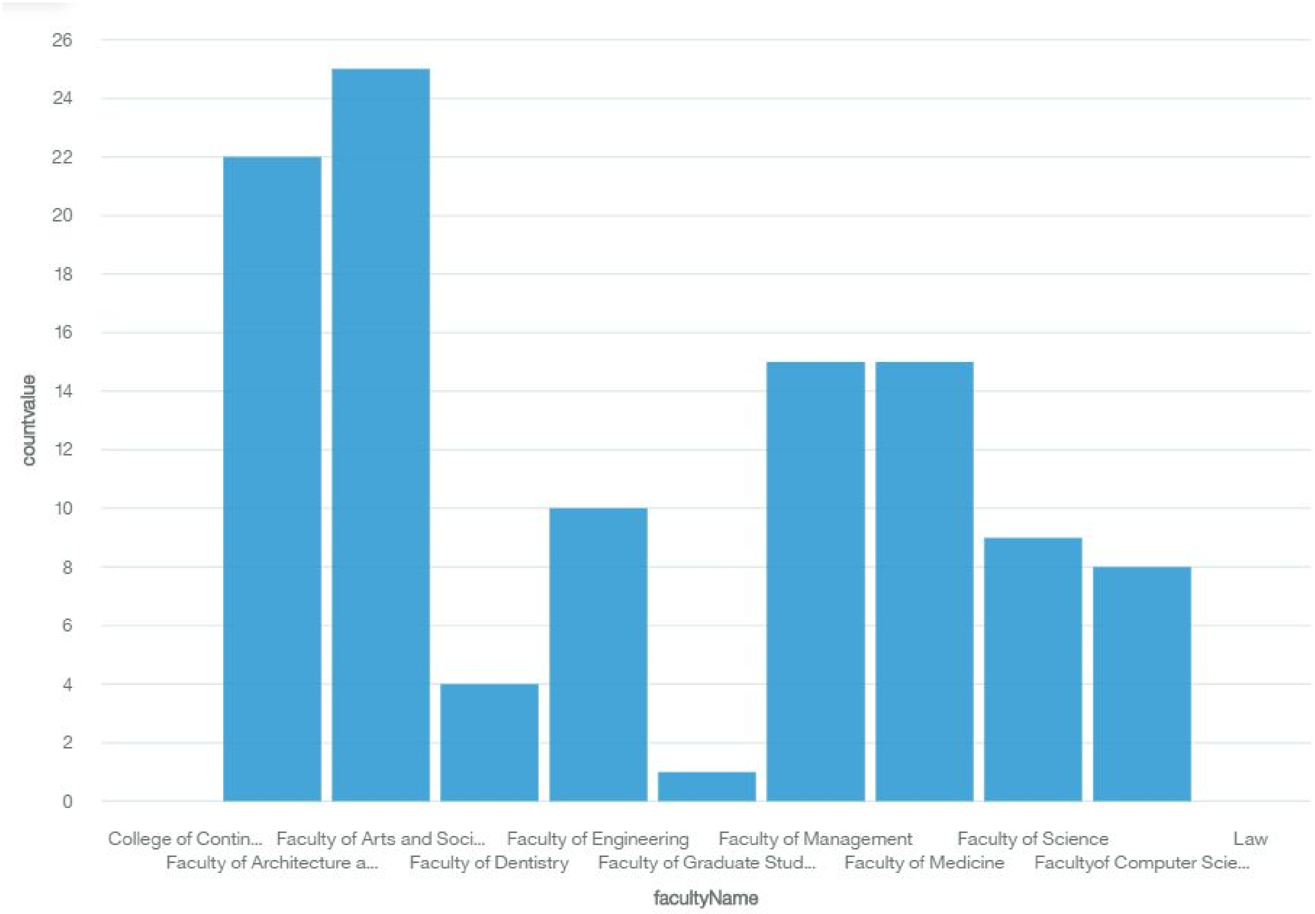
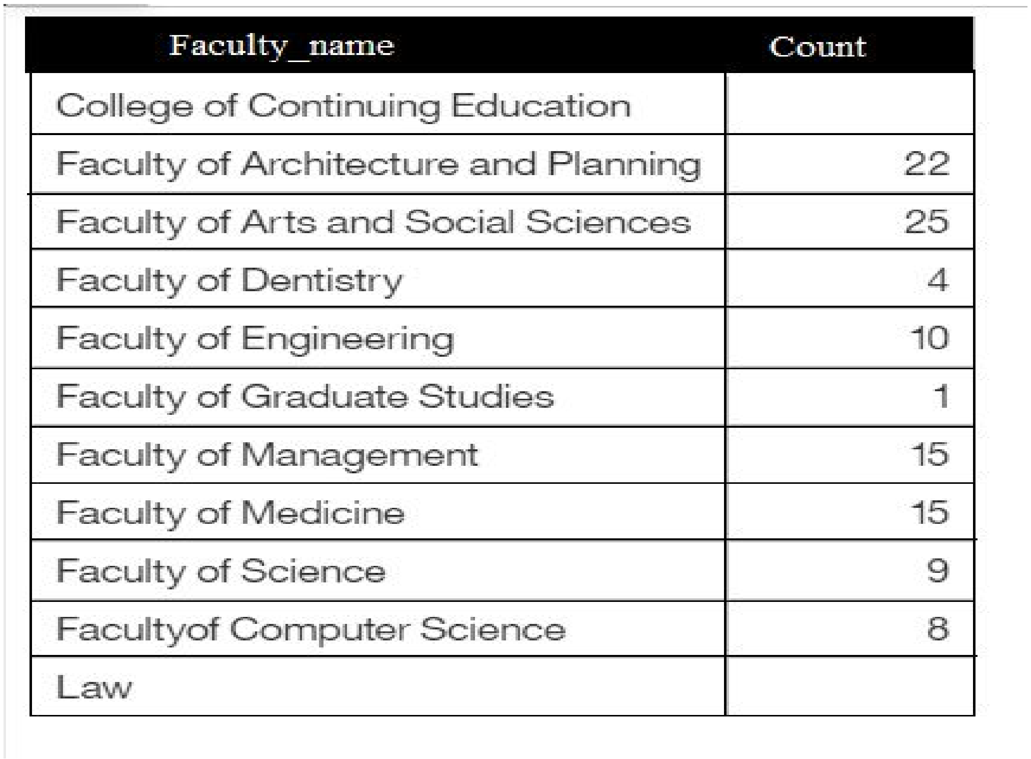


Fig 10. Bar Graph

Next part is answering the given questions using BI tools:

Question 1: Does Computer Science offer the highest number of programs?
Answer 1: No, according to the bar graph above, computer science has only 8 programs. The highest number of programs are 25 in the Faculty of Arts and Social Sciences.

Question 2: How many courses are there in each department or faculty?
Answer 2:



Faculty_name	Count
College of Continuing Education	
Faculty of Architecture and Planning	22
Faculty of Arts and Social Sciences	25
Faculty of Dentistry	4
Faculty of Engineering	10
Faculty of Graduate Studies	1
Faculty of Management	15
Faculty of Medicine	15
Faculty of Science	9
Facultyof Computer Science	8
Law	

Fig 11. Number of programs for each faculty

D. REFERENCES

1.“Tableau Data Visualization(Tutorial 7)”: Dal.brightspace.com. (2019). [online] Available at: https://dal.brightspace.com/d2l/le/content/100142/viewContent/1490370/View [Accessed 25 Nov. 2019].

2.“Cognos BI (Tutorial 6)”.Dal.brightspace.com. (2019). [online] Available at: https://dal.brightspace.com/d2l/le/content/100142/viewContent/1490370/View [Accessed 25 Nov. 2019].

3. Brownlee, J. (2019). *A Gentle Introduction to the Bag-of-Words Model*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/gentle-introduction-bag-words-model/ [Accessed 21 Nov. 2019].

4.Medium. (2019). *Word Clouds in Tableau: Quick & Easy.*. [online] Available at: https://towardsdatascience.com/word-clouds-in-tableau-quick-easy-e71519cf507a [Accessed 6 Dec. 2019].

5. M. Kulakowski, www.github.com 2012. [Online]. Available: https://gist.github.com/mkulakowski2/4289437 [Accessed: 27-Nov- 2019]

6. M. Kulakowski, www.github.com 2012. [Online]. Available: https://gist.github.com/mkulakowski2/4289441 [Accessed: 27-Nov- 2019]