

## Sentiment Analysis:

To perform sentiment analysis on tweets, I have used the data collected in Assignment-2 [2] (I have uploaded the script that I used in Assignment and added a reference). The tweets extracted are based on the keywords provided ('canada', 'halifax', 'university', 'dalhousie university', 'canada education'). I used regular expressions to clean the data. The final data is free from special symbols and URLs. I have considered 3000+ tweets which are extracted into a csv file. I have collected positive and negative words to compare with words in the tweet. After creating bag of words for each tweet, I have compared the words with positive and negative words list, written the matched word, tweet and polarity into a csv file using python script (Please refer sentimentanalysis.py) and visualized the frequently occurred words using Tableau. [4]

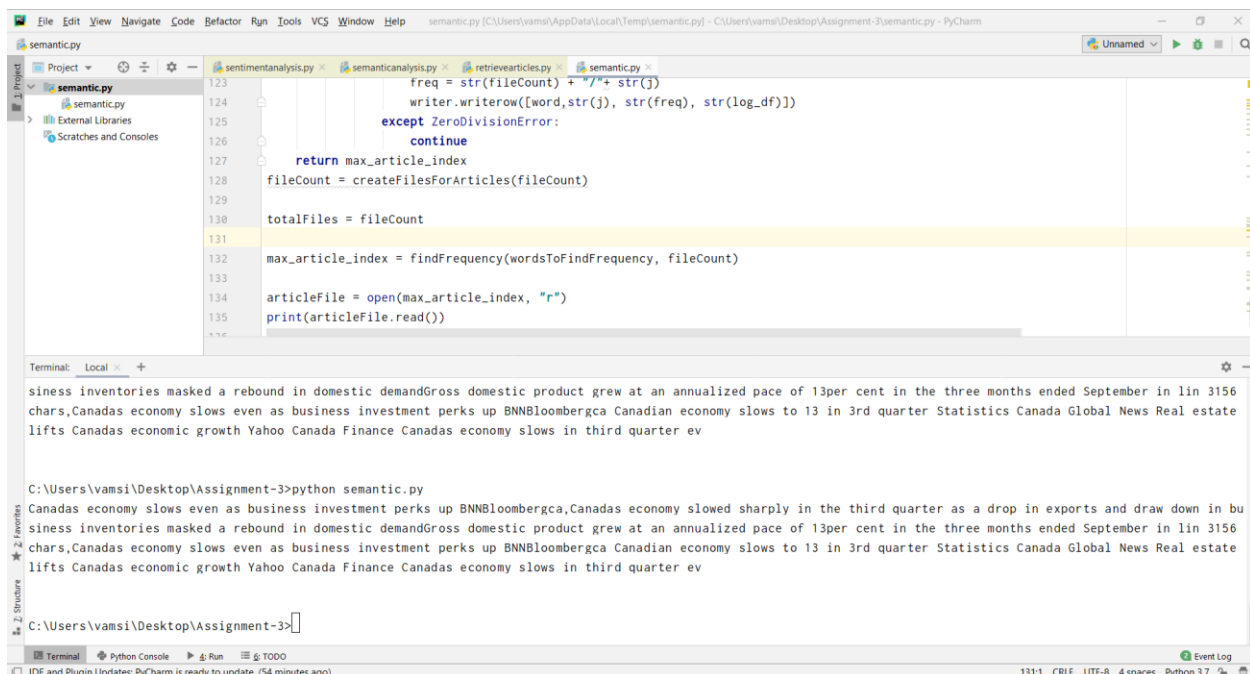


Figure 1: Tableau visualization of sentiment analysis data

## Semantic Analysis:

In this part of the assignment, I have written a script to extract news articles and extracted approximately 500 articles and cleaned them using regular expressions [2]. I have stored those articles in JSON format and converted it to CSV format. I have stored 'Title', 'Content' and 'Description' of the news articles extracted. Then I stored individual articles into individual text files (497 text files) using python script. I searched for keywords 'canada', 'halifax', 'university', 'dalhousie university', 'canada education' in the individual text files by splitting the words to calculate TF-IDF (Term Frequency – Inverse Document Frequency). I have calculated Total number of documents, documents having the keywords, number of times the word occurred in the document,  $\log_{10}(N/df)$  and stored them in CSV file in the given format (Please check TF-IDF.csv)

Later, I have found the document with maximum number of occurrences of 'Canada' and performed frequency count of the word in each document. I stored that data in CSV file (Please refer FreqCount.csv) in the given format. Finally, I have calculated the highest relative frequency by computing  $f/m$  (Please refer semanticanalysis.py).



The screenshot shows the PyCharm IDE with the `semantic.py` file open. The code in the editor includes logic for finding the maximum article index based on word frequency. The terminal window at the bottom displays the output of running `python semantic.py`, showing a list of news snippets and the file path where the results are stored.

```
123     freq = str(fileCount) + "/" + str(j)
124     writer.writerow([word, str(j), str(freq), str(log_df)])
125     except ZeroDivisionError:
126         continue
127     return max_article_index
128     fileCount = createFilesForArticles(fileCount)
129
130     totalFiles = fileCount
131
132     max_article_index = findFrequency(wordsToFindFrequency, fileCount)
133
134     articleFile = open(max_article_index, "r")
135     print(articleFile.read())
```

Terminal Output:

```
C:\Users\vamsi\Desktop\Assignment-3>python semantic.py
business inventories masked a rebound in domestic demandGross domestic product grew at an annualized pace of 13per cent in the three months ended September in lin 3156
chars,Canadas economy slows even as business investment perks up BNNBloombergca Canadian economy slows to 13 in 3rd quarter Statistics Canada Global News Real estate
lifts Canadas economic growth Yahoo Canada Finance Canadas economy slows in third quarter ev

C:\Users\vamsi\Desktop\Assignment-3>
```

Figure 2: Maximum relative frequency article for semantic analysis

## Business Intelligence:

I have considered the entities faculty, undergrad\_programs, grad\_programs, grad\_courses, library, adviser, subject etc., as dimensions. All the mentioned tables provide the source for the fact table. [1]

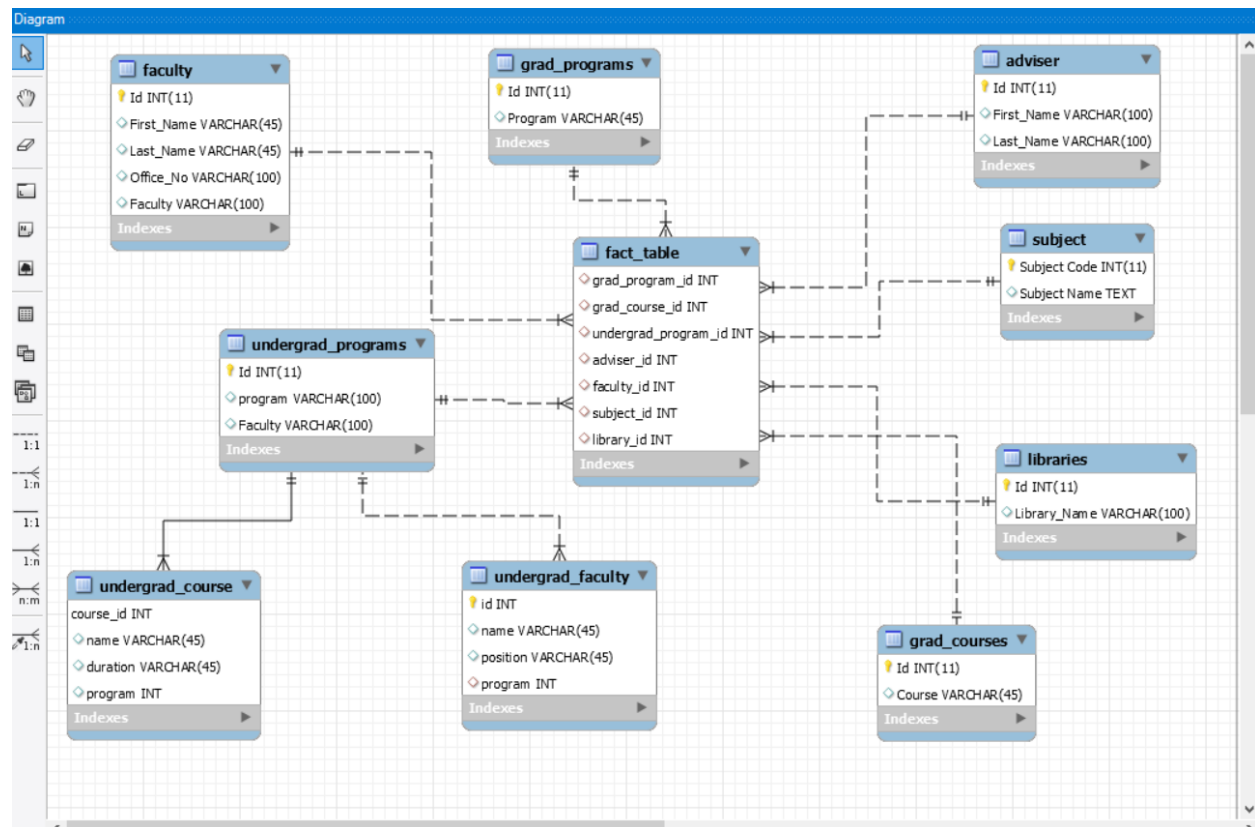


Figure 2: Snowflake scheme with fact table

The primary keys of the above tables act as foreign key in fact table and hence one-to-many relationships are established.

The dimension table **undergrad\_programs** has attributes 'id', 'program', 'faculty'. Dimension 'faculty' has attributes 'id', 'First\_Name', 'Last\_Name', 'Office\_No', 'Faculty'. Similarly, the other dimensions has attributes as mentioned in the data model presented above in figure 1.

**Attribute Hierarchies:** The dimension 'undergrad\_programs' can have attribute hierarchy in the form of entities 'undergrad\_courses' and 'undergrad\_faculty'. There can be many attribute hierarchies depending on the requirement of the dataset.

**Cognos BI:** The below is the screenshot of the snowflake schema in Cognos BI.

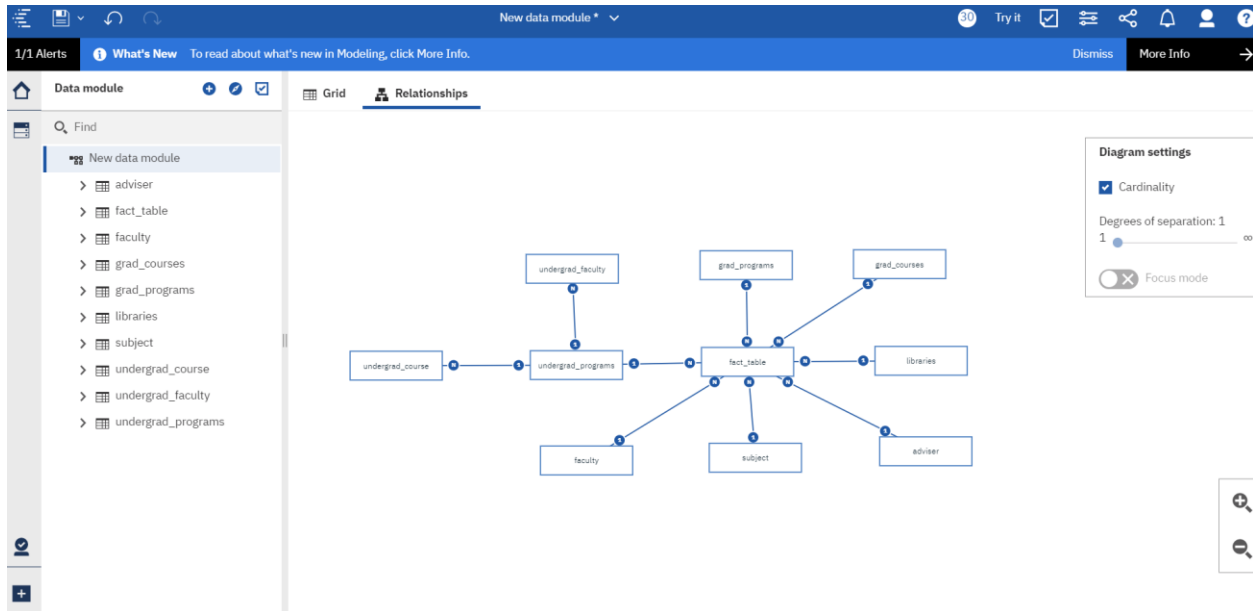


Figure 4: Snowflake schema in Cognos BI

I have created entities in AWS server and established a connection from Cognos BI and retrieved the tables and their data. By using the relationships established in AWS, tables can be visualized with the relationships as above.

18. a. From the data that I extracted form dal website, computer science doesn't offer highest number of course as we can see in figure 5, faculty of arts and sciences offers highest number of courses.

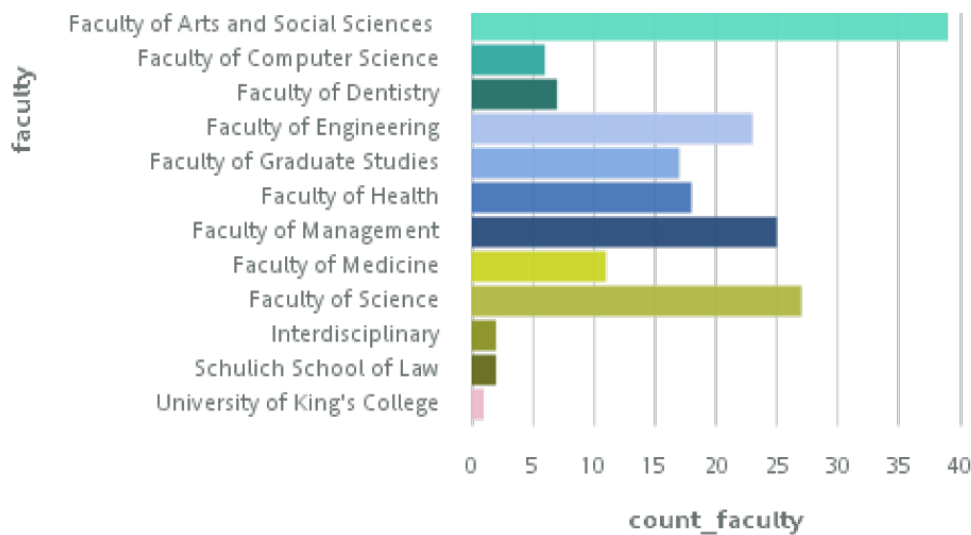


Figure 5: 18(a)

18. b. I have visualized the number of courses offered by each faculty in the below figure.

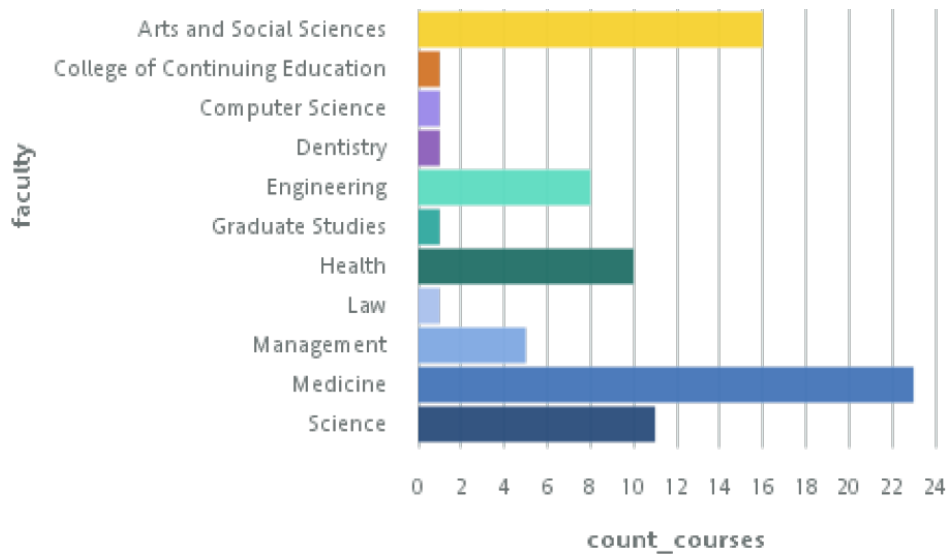


Figure 6: 18(b)

## References:

- [1] Vamsi Gamidi, Assignment-1, submitted to CSCI-5408
- [2] Vamsi Gamidi, Assignment-2, submitted to CSCI-5408
- [3] 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.
- [4] Free Tableau Videos [Online]  
Available: <https://www.tableau.com/learn/training>  
[Accessed on November 22, 2019]