CSCI 5408: Assignment 3

A. Sentiment Analysis

1. SentimentAnalysis.py script in the submission folder performs this task
2. Numpy and pandas libraries are imported.
3. Tweets collected (tweet_Cleaned.csv) from Assignment 2 are considered.
4. Such collected tweets are cleaned by storing into data frames and on each data frame series, URLs and emojis are removed
5. Such collected tweet texts are broken down into dictionary, where it contains the word and the word count
6. Each word in a list of bag of words is compared with the positive_words.txt and negative_words.txt.
7. Based on the count of match polarity is decided whether it is positive, negative or neutral if none matched.
8. The word with the highest frequency in each word is considered and corresponding tweet text, polarity, match(word with the highest frequency) is stored in sentiment.csv.
9. Visualization [1] is done in Tableau [2] with positive and negative tweets based on polarity. Below are the snapshots.

Positive words cloud:

Figure 1: Positive word cloud visualized in Tableau

Negative words cloud:

Figure 2:Negative word cloud visualized in Tableau

B. Semantic Analysis

1. SemanticAnalaysis.py and SemanticAnalysis2.py scripts in the submission folder perform this task.
2. Numpy, Pandas, and math libraries are imported.
3. News articles (news_cleaned.csv) collected from Assignment 2 are used in this task
4. Cleaning of each news article and storing each news article are done in SemanticAnalysis.py script. Please copy the news text files stored in news_files folder to src folder while running the code.
5. In SemanticAnalysis2.py script, term frequency of each keyword is found and inverse document frequency is calculated and stored in tf_idf.csv

Total number of documents	496		
Search query	Document Containing Term (df)	Total documents(N)/number of documents term appeared (df)	Log10(N/df)
Canada	151	3.28	0.52
university	126	3.94	0.6
dalhousie university	11	45.09	1.65
halifax	54	9.19	0.96
canada education	2	248	2.39

Figure 3: TF-IDF for the specific keywords

6. Frequency of Canada keyword found in the articles are calculated and stored in Canada_frequency.csv
7. Article 37 has the highest frequency of keyword Canada and is printed

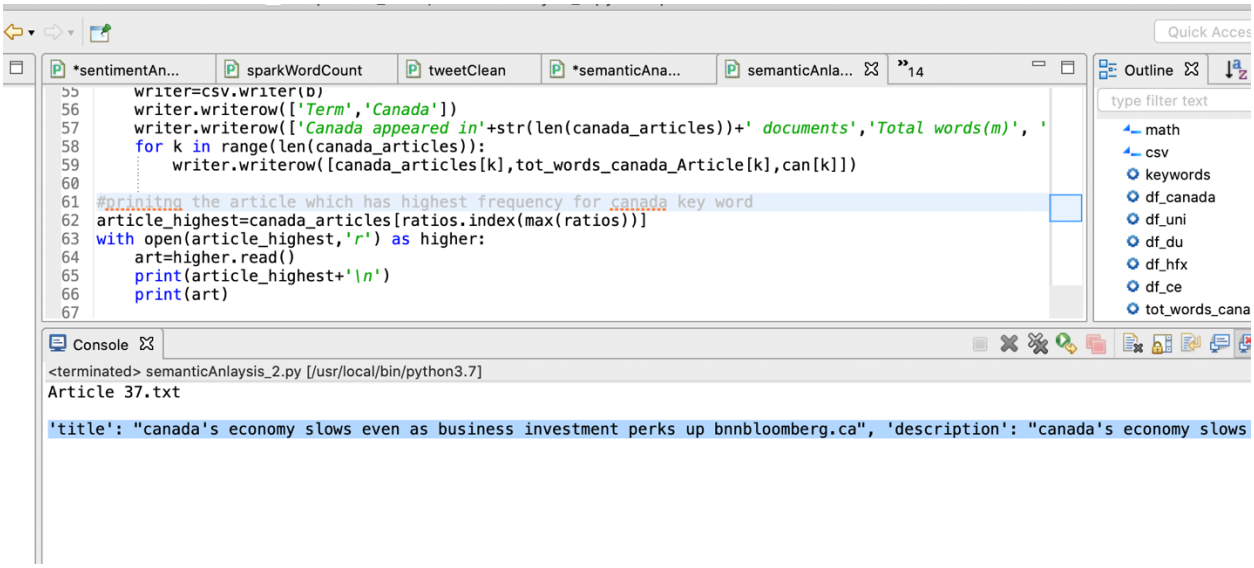

Figure 4: News article with the highest frequency for word Canada

C. Business Intelligence

1. Cognos BI is setup based on the instructions provided in the Tutorial 6[3].

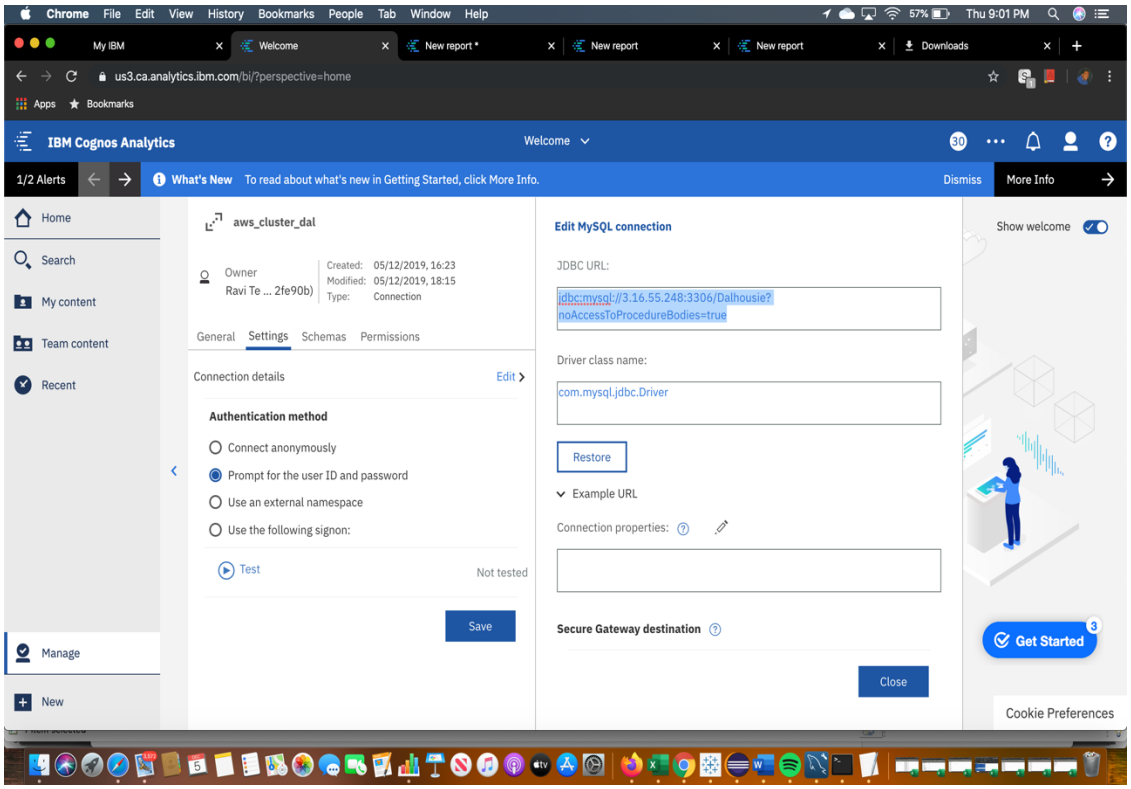

Figure 5: Setup of Cognos BI

2. Created a snowflake schema[4].
3. For the specific assignment, University_fact table is considered as Fact table, which has the following attributes: EmployeeId, DepartmentId, Number of courses, Number of programs
4. Dimension tables and their attribute hierarchies are as follows:
   Department- DepartmentId, DepartmentName
   Employee- EmployeeId, EmployeeLastName,EmployeeFirstName
5. Since snowflake schema is followed, Department dimension has 1:M cardinality with Program (ProgramId,ProgramName,ProgramLevel,DepartmentId) and Course(CourseId,CourseName,DepartmentId)
6. SQL script are added in the sql_script folder in the submission folder for reference

   Below is the snapshot of snowflake schema

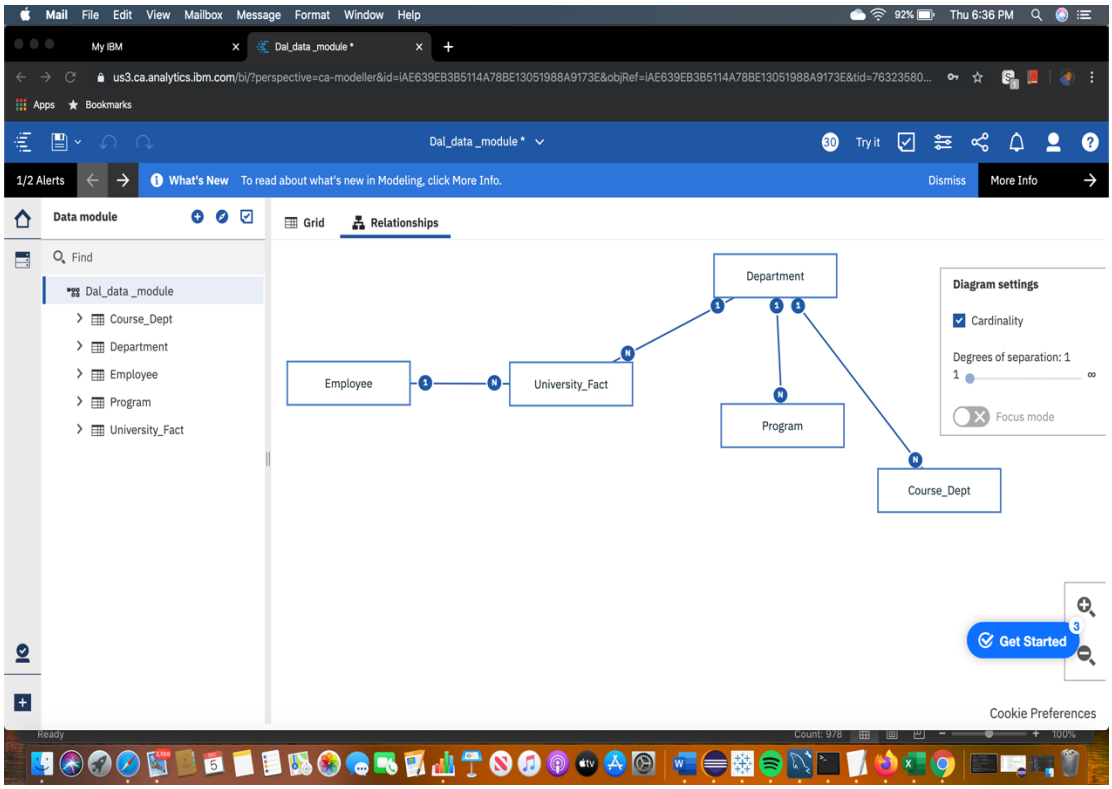

Figure 6: Snowflake schema

7. Using your BI framework, can you answer the following?
    a. Does Computer Science offer the highest number of programs?

        Answer: No

        As per the report generated[5] below, Engineering has more programs (44) while computer science offers 6 programs. Please click on the report to see details.

    b. How many courses are there in each department or faculty?

        Answer: yes

Report is generated in Cognos, by grouping by department Id and counting the Courses for each department. Please click on the below report to see details.

References:

[1] [Online]. Available: https://kb.tableau.com/articles/howto/creating-a-word-cloud. [Accessed 29 11 2019].

[2] [Online]. Available: https://dal.brightspace.com/d2l/le/content/100142/viewContent/1501687/View. [Accessed 29 11 2019].

[3] [Online]. Available: https://dal.brightspace.com/d2l/le/content/100142/viewContent/1490370/View. [Accessed 03 12 2019].

[4] [Online]. Available: https://dal.brightspace.com/d2l/le/content/100142/viewContent/1493737/View. [Accessed 04 12 2019].