Assignment 4

CSCI 5408 - Data Warehousing, Management And Analytics

Submitted by: Nishant Amoli
Banner ID: B00835717

Submitted to: Dr. Saurabh Dey
Faculty of Computer Science
Dalhousie University

Abstract

Sentiment analysis and opinion mining are important parts of data science and information gathering [1]. Doing so makes it possible to virtualize the important information based on the historic data and learn from it. This can help in efficient decision making. There are many Business Intelligence tools that can be used in order to achieve this goal. One such BI tool is Tableau [2] which is used for this assignment to visualize the sentiments of the people using their twitter data.

Introduction

This assignment comprises the fundamentals of Data Analytics and Data Visualization. It is subdivided into various tasks that cover the concepts of cleaning and transformation of data, performing certain analytics, analysing sentiments and finally visualizing the key findings using Tableau.

A. Sentiment Analysis

Twitter Data Transformation

The objective of this task is to perform sentiment analysis of the tweets that were extracted as a task for Assignment 3 based on the keywords “Canada”, “University”, “Dalhousie University”, “Halifax” and “Canada Education”. As a part of Assignment 3, the tweets were also cleaned and transformed and then stored in a json file named ***cleaned_tweets.json***. The mentioned file is used to retrieve the cleaned twitter data in order to perform sentiment analysis and can be found in the project directory. Although the tweets were already cleaned, all the hashtags, URLs, special symbols, mentions, e.t.c were previously removed, a little data transformation and refining has been done using a python script named ***sentiment_analysis_script.py*** and can be found in the project directory.

Bag of words

The same script is used in order to create a bag of words for every tweet. The demonstration of the data structure containing the bag of words is as follows:

[
0: { tweet: “ Happy in Canada. ”, ‘bag_of_words’: { ‘Happy’:1, ‘in’:1, ‘Canada’:1 } }
1: { tweet: “A brief history of time. ”, ‘bag_of_words’: { ‘A’:1, ‘brief’:1, ‘history’:1, ‘of’:1, ‘time’:1 } }
2: { tweet: “the voice of the God. ”, ‘bag_of_words’: { the:2, ‘voice’:1, ‘of’:1, ‘God’:1 } }
.
.
.
]

The code can be found in the python script ***sentiment_analysis_script.py*** in the project directory.

Sentiment Analysis

In order to perform sentiment analysis, the positive [3] and negative [4] bag of words were used. Both the files can be found in the project directory in text format. The code can be found in the python script ***sentiment_analysis_script.py*** in the project directory. The tagged tweets are stored in a csv file named ***tagged-tweets.csv*** and can be found in the project directory.

	A	B	C	D
1	S.No	Message/Tweet	Match	Polarity
2	0	wifi should be available free throughout canada during this crisis this will help with public h	free available crisis	positive
3	1	canadian scientists make covid19 research breakthrough isolating virus only in canada the	breakthrough supportive innovation	positive
4	2	they also worked against 14 indigenous communities that wouldve benefited 40 years of	benefits worked	positive
5	3	dude we have the highest average education level in canada are you dense	dense	negative
6	4	our new education guide residential schools in canada aims to raise awareness of the histo	free reconciliation	positive
7	5	if this doesnt scare the shit out every american it should canada ill hold on to your beer	right shit scare	negative
8	6	they also worked against 14 indigenous communities that wouldve benefited 40 years of	benefits worked	positive
9	7	they also worked against 14 indigenous communities that wouldve benefited 40 years of	benefits worked	positive
10	8	its a busy time all over and were not slowing down here its national is hiring multiple sumr	work great	positive
11	9	the province is directing public sector employees who travel outside canada to stay home	support	positive
12	10	signing of peace and friendship treaty staged to educate people	peace	positive
13	11	saskatchewan education system prepping for covid19 cbc news via		neutral

Figure 1: The image demonstrating the csv file having the tagged tweets that is created using the python script.

Data Visualization

After performing sentiment analysis on the transformed twitter data, the tagged data have been visualized using Tableau. Below are the images demonstrating the report generated using Tableau. These images can also be found in the folder ***Visualised Data*** within the project folder.

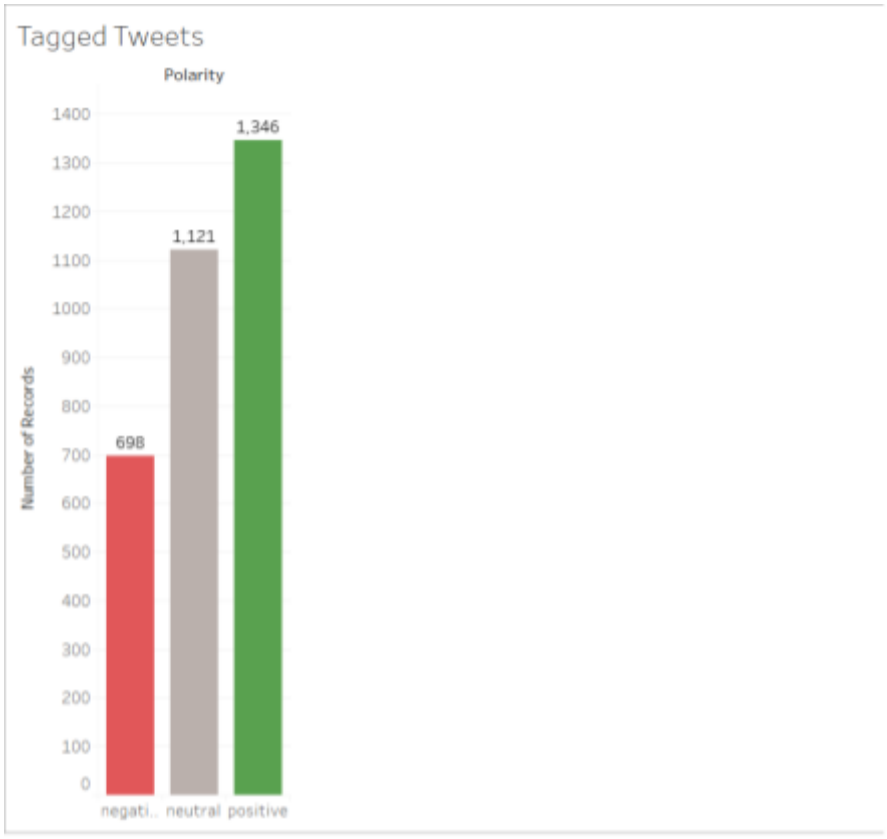

Figure 2: The graph generated using Tableau demonstrating the total number of tweets and their polarities.

Figure 3: A report generated using Tableau demonstrating the word cloud of most occuring positive and negative words.

Figure 4: Pack bubble representation of tweets generated using Tableau.

B. Semantic Analysis

News Articles Data Transformation

The objective of this task is to perform semantic analysis of the news articles that were extracted as a task for Assignment 3 based on the keywords “Canada”, “University”, “Dalhousie University”, “Halifax” and “Business”. As a part of Assignment 3, the news articles were also cleaned and transformed and then stored in a json file named ***news.json***. The mentioned file is used to retrieve the cleaned news data in order to compute document similarity using the TF-IDF method based on the keywords mentioned above and can be found in the project directory. Further transformation of news articles has been performed using a python script named ***semantic_analysis_script.py*** and the transformed data that contains ‘title’, ‘description’ and ‘content’ are stored in the json file named ***cleaned_news.json.*** The python script can be found in the project directory. This script also contains the code to compute TF-IDF. The computed result is stored in the csv file named ***TF-IDF.csv***.

	A	B	C	D
1	Total Documents	140		
2	Search Query	Document containing term (df)	Total Documents(N)/ number of documents term appeared (df)	Log10 (N/df)
3	canada	36	140/36	0.59
4	university	10	140/10	1.15
5	dalhousie university	4	140/4	1.54
6	halifax	19	140/19	0.87
7	business	17	140/17	0.92
8				

Figure 5: The image demonstrating TF-IDF.csv that is created using the python script.

Also, this script is used in order to find the documents having the highest frequency of the word, “Canada” and the computed result is stored in the csv file named ***canada-frequency-distribution.csv***.

	A	B	C
1	Term	Canada	
2	Canada appeared in 36 documents	Total Words (m)	Frequency (f)
3	Article #1	80	1
4	Article #2	111	1
5	Article #3	83	1
6	Article #4	99	1
7	Article #5	98	1
8	Article #6	98	1
9	Article #7	90	3
10	Article #8	89	4
11	Article #9	82	2
12	Article #10	108	1

Figure 6: The image demonstrating a part of canada-frequency-distribution.csv that is created using the python script.

At last, the news article with the highest relative frequency can be printed on the console screen along with the relative frequency using the same script, i.e. ***semantic_analysis_script.py***.

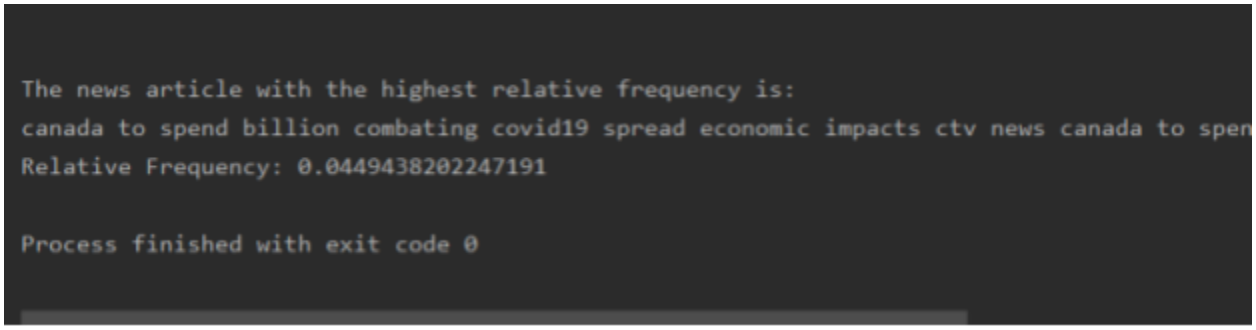

Figure 7: Image demonstrating the tweet with the highest relative frequency of the word "Canada" printed on the console screen after executing the python script.

Conclusion

In this assignment, concepts of data analytics, sentiment analysis and data visualization were covered, and along with that tools and methodologies including Tableau and TF-IDF. Data extraction, data transformation and data processing are important aspects of the field of Data Science and this assignment has provided an exposure to the importance of all of these.

References

[1] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] "Products," *Tableau Software*. [Online]. Available: https://www.tableau.com/products. [Accessed: 11-Apr-2020].

[3] "positive-words.txt," *Gist*. [Online]. Available: https://gist.github.com/mkulakowski2/4289437. [Accessed: 10-Apr-2020].

[4] "negative-words.txt," *Gist*. [Online]. Available: https://gist.github.com/mkulakowski2/4289441. [Accessed: 10-Apr-2020].

G. S. Dhillon, K. T. Mani, and M. Bhanderi, *Brightspace - Dalhousie University*. [Online]. Available: https://dal.brightspace.com/d2l/le/content/110354/viewContent/1624572/View. [Accessed: 11-Apr-2020].