Assignment 3 Report

CSCI 5408: Data Management & Warehousing

Submitted by:

Aakash Patel (B00807065)

A. Sentiment Analysis

- To perform sentiment analysis, I loaded my file with tweets of previous assignment (retrieved from Twitter using Search API).
- I performed required cleaning to remove URL and any special characters. I did consider Comma, Apostrophe and Space for not to be removed. They are not removed since they make sense for tweet interpretation and have major impact on twitter interpretability on removal.
- I processed each tweet in sequence and generated its bag of words.
- Bag of words represent count of each word's occurrence.
- Each word of the Bag of words is then looked for if it contains positive and negative words. Positive and Negative words to be compared with, are in lists prepared from positive and negative word list standard defined by the online source **[1].**
- If the tweet contains more positive words then Polarity is marked positive, more count of negative words then marked Negative and Neutral if both counts are equal. Results CSV File with Matches and Polarity is generated for every tweet.

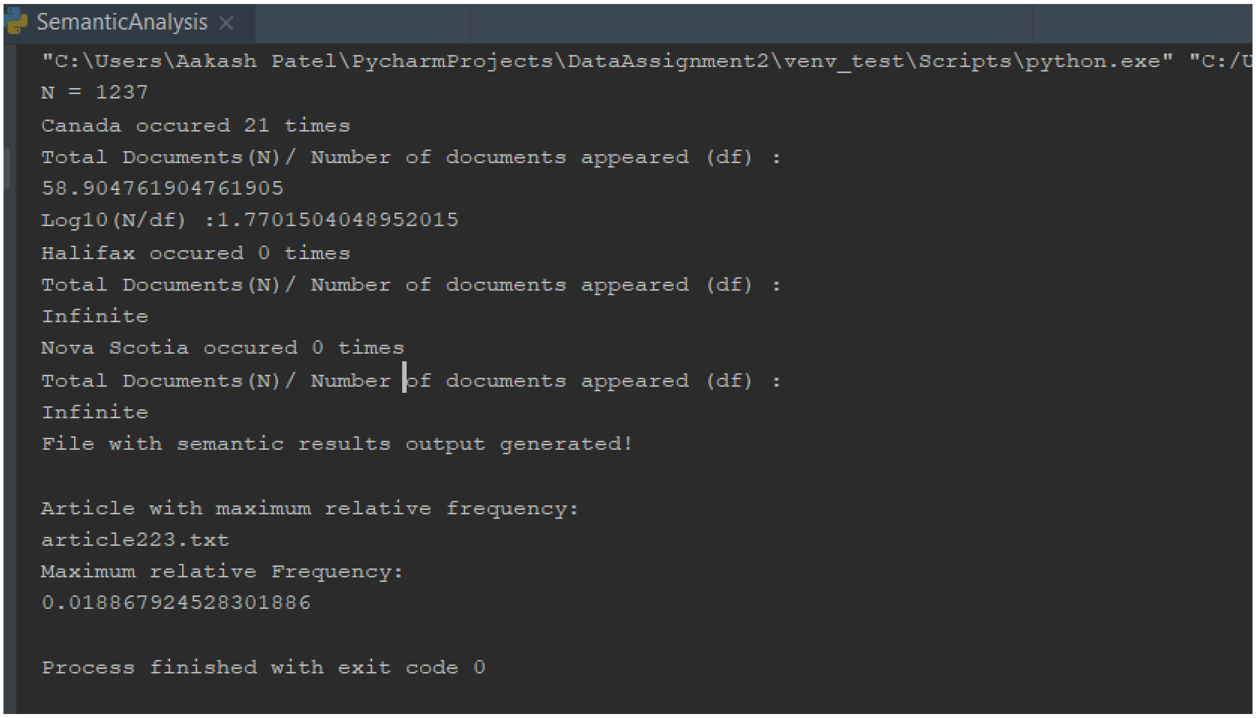
B. Semantic Analysis

1. To perform semantic analysis, I firstly determined N (i.e Number of total documents) programmatically and then iterated through all the required words to be assessed. For every word all articles created as part of second assignment, are evaluated for count.

2. The content of the article is cleaned for URLs and special symbols (if any).

3. Count is then used to evaluate two other measures provided its non-zero.

4. As part of the second half of the script all articles are re-iterated to find total number of occurrences for word "Canada"(f) within the article , total word count of the article and its relative frequency (f/m). All this information is stored in newly generated CSV file ("semantic_results.csv") and article with maximum relative frequency is displayed in console.



```
"C:\Users\Aakash Patel\PycharmProjects\DataAssignment2\venv_test\Scripts\python.exe" "C:/U
N = 1237
Canada occured 21 times
Total Documents(N)/ Number of documents appeared (df) :
58.904761904761905
Log10(N/df) :1.7701504048952015
Halifax occured 0 times
Total Documents(N)/ Number of documents appeared (df) :
Infinite
Nova Scotia occured 0 times
Total Documents(N)/ Number of documents appeared (df) :
Infinite
File with semantic results output generated!

Article with maximum relative frequency:
article223.txt
Maximum relative Frequency:
0.018867924528301886

Process finished with exit code 0
```

Figure 1: Semantic Analysis Script Output

C. Business Intelligence:

1. Firstly I did Cognos BI tool setup as guided by Lab session material. I created data modules using tables imported. Here I bifurcated data for Vehicle and School for better understanding and ease at further procedures.

Following facts were analysed:

- Number of vehicle sales
- School Board Expenses yearly

For Vehicle Sales data, the fact is number of vehicle sales sold whereas for School Board data, the school board yearly values happens to be the fact.

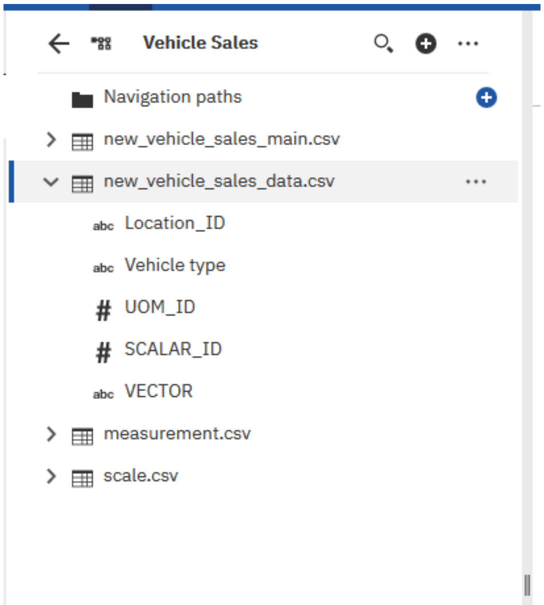
Below dimensions were used for Vehicular data:

- New Vehicle Sales Main – It defines main value for vehicle Sale fact
- Measurement – It defines measurement parameter taken into consideration.
- Scale – It defines Scale for the measurement.

Below dimensions were used for School data:

- School Board Expense Main – It defines main value for school expense fact.
- Measurement – It defines measurement parameter taken into consideration.
- Scale – It defines Scale for the measurement.

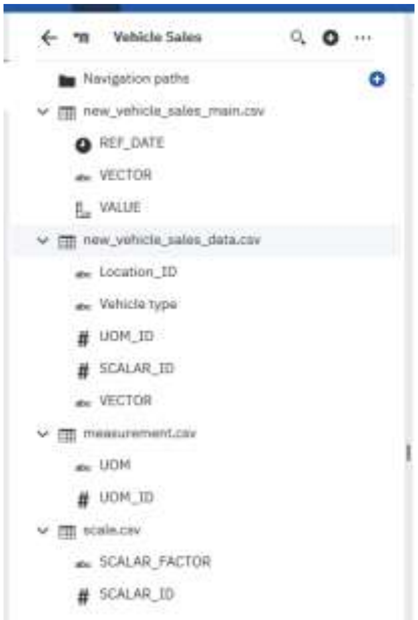
Following are the fact and dimension tables taken into consideration:



Vehicle Sales

- Navigation paths
- new_vehicle_sales_main.csv
- new_vehicle_sales_data.csv
  - abc Location_ID
  - abc Vehicle type
  - # UOM_ID
  - # SCALAR_ID
  - abc VECTOR
- measurement.csv
- scale.csv

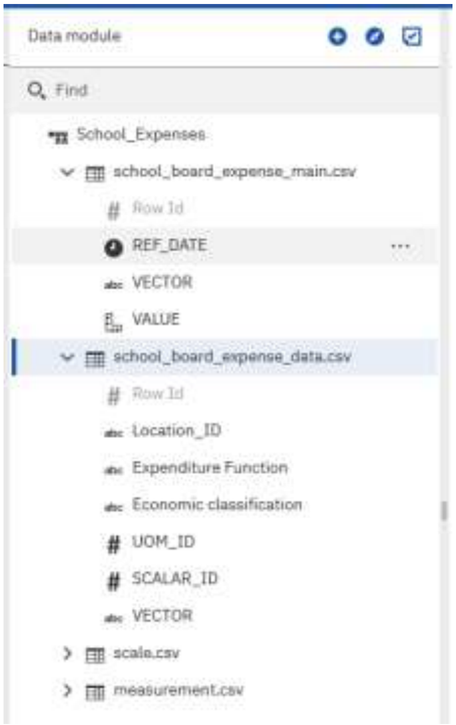
Figure 1: Vehicle Sales Data Fact Table



Vehicle Sales

- Navigation paths
- new_vehicle_sales_main.csv
  - REF_DATE
  - abc VECTOR
  - VALUE
- new_vehicle_sales_data.csv
  - abc Location_ID
  - abc Vehicle type
  - # UOM_ID
  - # SCALAR_ID
  - abc VECTOR
- measurement.csv
  - abc UOM
  - # UOM_ID
- scale.csv
  - abc SCALAR_FACTOR
  - # SCALAR_ID

Figure 2: Vehicle Sales Dimension Tables (with all attributes)

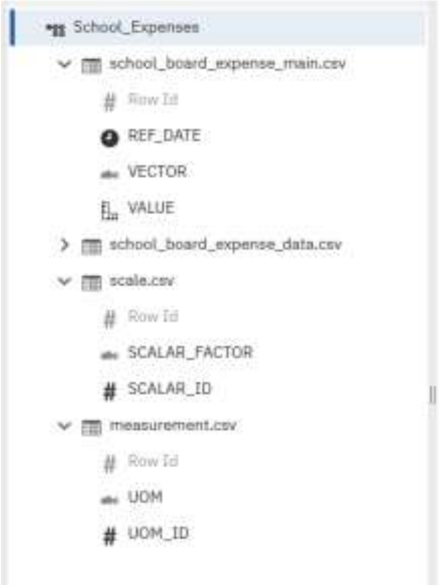


Data module

Find

- School_Expenses
  - school_board_expense_main.csv
    - Row Id
    - REF_DATE
    - VECTOR
    - VALUE
  - school_board_expense_data.csv
    - Row Id
    - Location_ID
    - Expenditure Function
    - Economic classification
    - UOM_ID
    - SCALAR_ID
    - VECTOR
  - scale.csv
  - measurement.csv

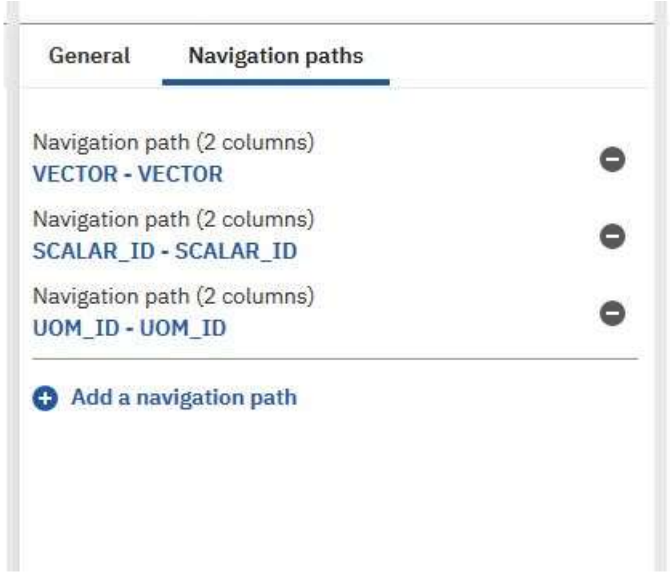
Figure 3: School Board Expense Data Fact Table



- School_Expenses
  - school_board_expense_main.csv
    - Row Id
    - REF_DATE
    - VECTOR
    - VALUE
  - school_board_expense_data.csv
  - scale.csv
    - Row Id
    - SCALAR_FACTOR
    - SCALAR_ID
  - measurement.csv
    - Row Id
    - UOM
    - UOM_ID

Figure 4: School Expenses Dimensions Table (with all attributes)

Navigation Paths hierarchies:

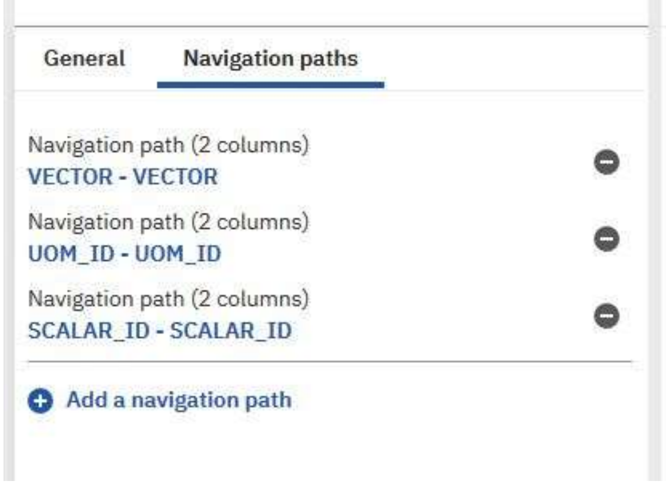


Navigation path (2 columns)
VECTOR - VECTOR

Navigation path (2 columns)
SCALAR_ID - SCALAR_ID

Navigation path (2 columns)
UOM_ID - UOM_ID

Figure 5: Navigation Path hierachy for School



Navigation path (2 columns)
VECTOR - VECTOR

Navigation path (2 columns)
UOM_ID - UOM_ID

Navigation path (2 columns)
SCALAR_ID - SCALAR_ID

Figure 6 Navigation Path Attribute hierarchy of Vehicular Data

- Data Warehouse design schemas:

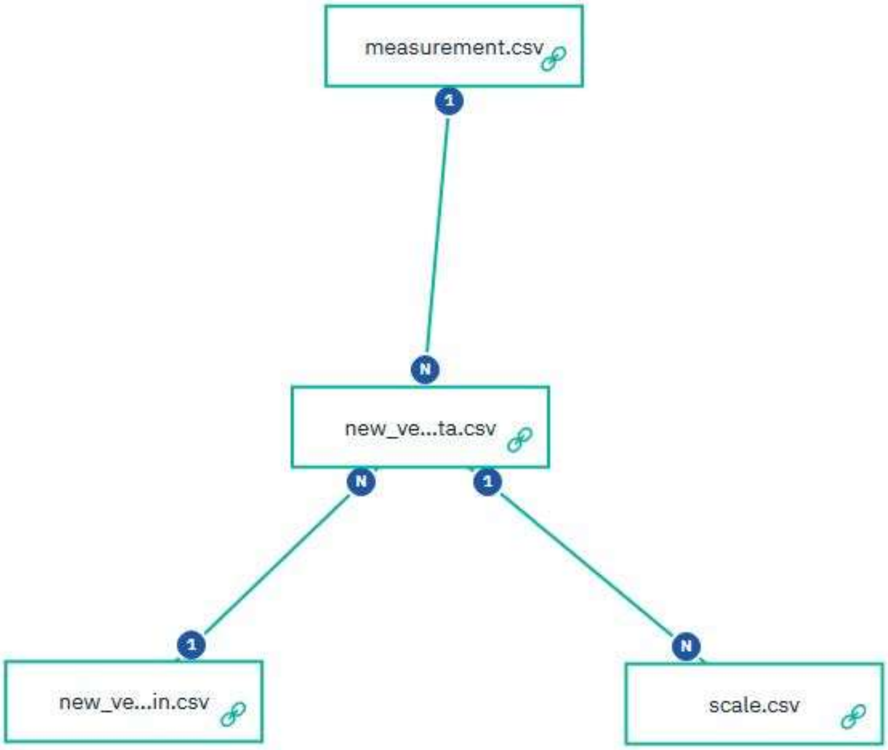


Figure 7: Vehicle Warehouse star schema

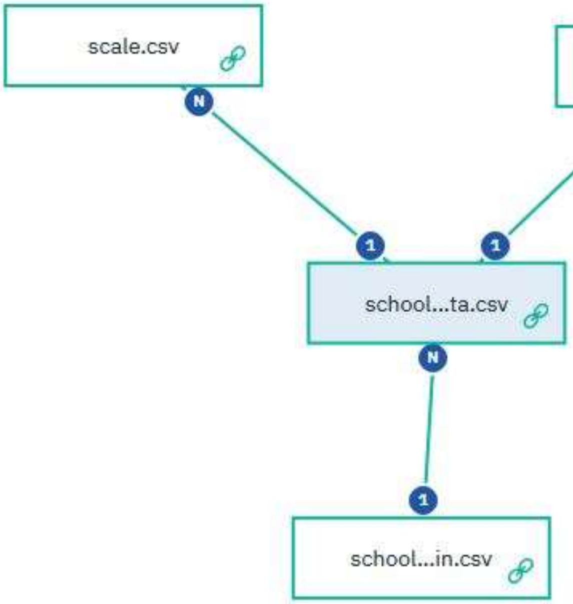


Figure 8: School Warehouse star schema

- Cognos BI reports:

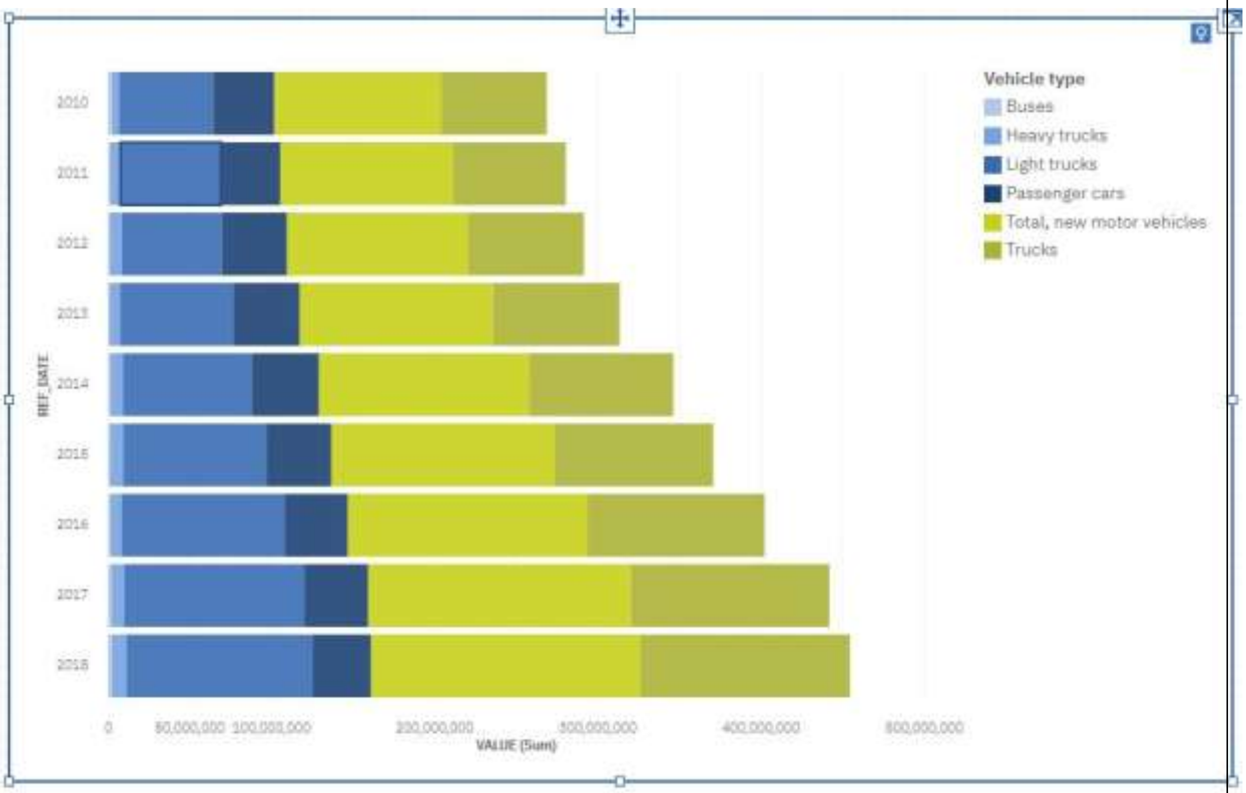


Figure 9: Vehicle Yearly Sales for various Vehicle types

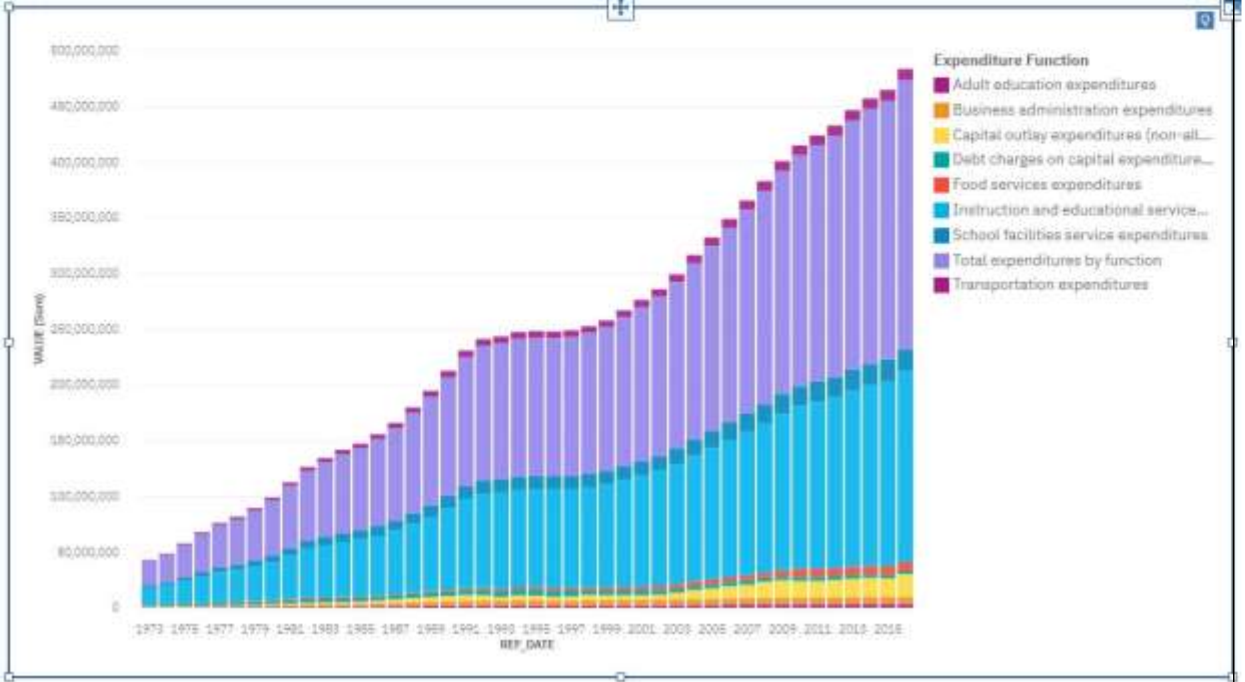


Figure 10: School Board Yearly Expenses for different expenditure functions

- Using the BI framework following questions can be answered:
  - Is there growth in motor vehicle sales?
    - Motor Vehicle sales have increased considerably from 2010 to 2018.
  - Is there growth for expenditure amount of school boards?
    - School Board expenses have increased overall from 1973 to 2016.

Note: Second question is in accordance to the data available.

References

[1] "shekhargulati/sentiment-analysis-python", GitHub, 2019. [Online]. Available: https://github.com/shekhargulati/sentiment-analysis-python/blob/master/opinion-lexicon-English/. [Accessed: 19- Jul- 2019].

[2] Us1.ca.analytics.ibm.com, 2019. [Online]. Available: https://us1.ca.analytics.ibm.com/bi/?perspective=ca-modeller&id=1105780077_392f3e37dfdd4b2d9517e68fe2f4542d_sessionTemp&ui_appbar=true&tid=1105780077_392f3e37dfdd4b2d9517e68fe2f4542d_sessionTemp. [Accessed: 22- Jul- 2019].

[3]H. Python et al., "How to remove any URL within a string in Python", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/11331982/how-to-remove-any-urlwithin- a-string-in-python?noredirect=1&lq=1. [Accessed: 22- Jul- 2019].

[4] H. Python et al., "How to remove any URL within a string in Python", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/11331982/how-to-remove-any-urlwithin- a-string-in-python?noredirect=1&lq=1. [Accessed: 22- Jul- 2019].