SEMANTIC ANALYSIS DOCUMENTATION

INTRODUCTION:

The Semantic analysis focuses on finding the highest relative frequency of the word “Canada” in the entire collection of 1578 reuter files /news articles. The frequecy of the word “Canada” is used to find the highest relative frequency by dividing the total number of words in the file by the number of occurences of “Canada” in the file.
hrf = (f/m).
hrf -> Highest relative frequency
f ->frequency of “Canada” in the file
m ->total number of words in the file

PROCESS:

1. I have written a javascript to read all the 1578 Reuter files which were generated as part of Assignment-2.
2. The script written to clean the reuter files will replace the tags with appropriate and meaningful words and the special characters are removed.
3. General expressions are used to remove the special characters.
4. The script counts the frequency of number of documents(df) containing the words “Canada”, “Halifax” and “Nova Scotia”.
5. This value is then used to find the log to the base10 value -> Log10(N/df)
6. N-> the total number of documents. ie. 1578

df-> number of documents containing the words “Canada”, “Halifax” and “Nova Scotia”.
7. Finally, the script calculates the document which has the highest relative frequency of the word “Canada” by calculating f/m.

Highest relative frequency = f/m
f-> frequency of “Canada” in each reuter document
m->total number of words in each reuter document

CONCLUSION :

The Reuter document #573 has the highest relative frequency for the word “Canada”.

Screenshot of Output from the console:

Console

```
<terminated> semantic_analysis [Java Application] C:\Program Files\jdk-11.0.1\bin\javaw.exe (Jul 21, 2019, 10:07:36 PM)
Log10 for Canada:1.7357090009744454
Log10 for Halifax:Infinity
Log10 for Nova Scotia:Infinity
Article which has the highest relative frequency:Article #573
```

semantic_analysis.java | sentiment_analysis.java

```
    public void compute_TF_IDF() {
        int count_Canada = 0;
        int count_Halifax = 0;
        int count_novascotia = 0;
        int canada_match = 0;
        int count = 0;
```