CS7611-COMPILER LABORATORY

NEWS CLASSIFICATION

Submitted by:

A.Varsha - 2017103606

S.Subashini -2017103600

Problem statement:

Classifying the news based on the headlines.

Languages used:

Lex , python.

Github link:

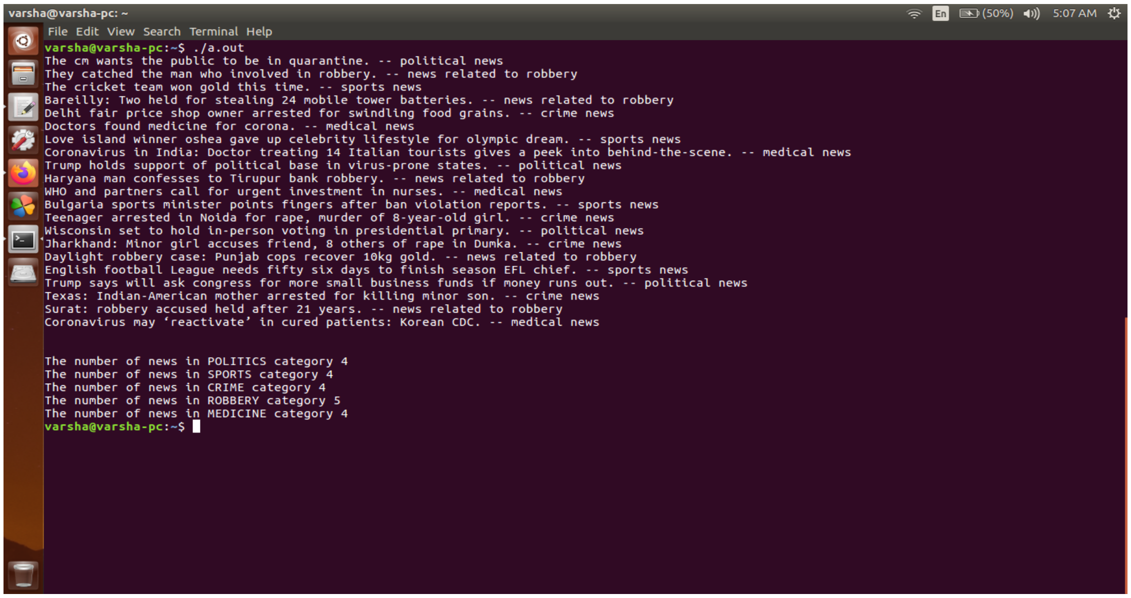
https://github.com/VarshaAnandavel/Varsha

Implementation details:

We have implemented in both python and lex. In lex implementation, the input is read from a file and the output will be the news and the category to which it belongs to. Ie, if the news is related to politics , ‘**political**’ will be displayed. We have also displayed the total number of news in each category. Using python, we have displayed the number of news in each category provided with a dataset. The precision and accuracy will also be displayed.

Output:

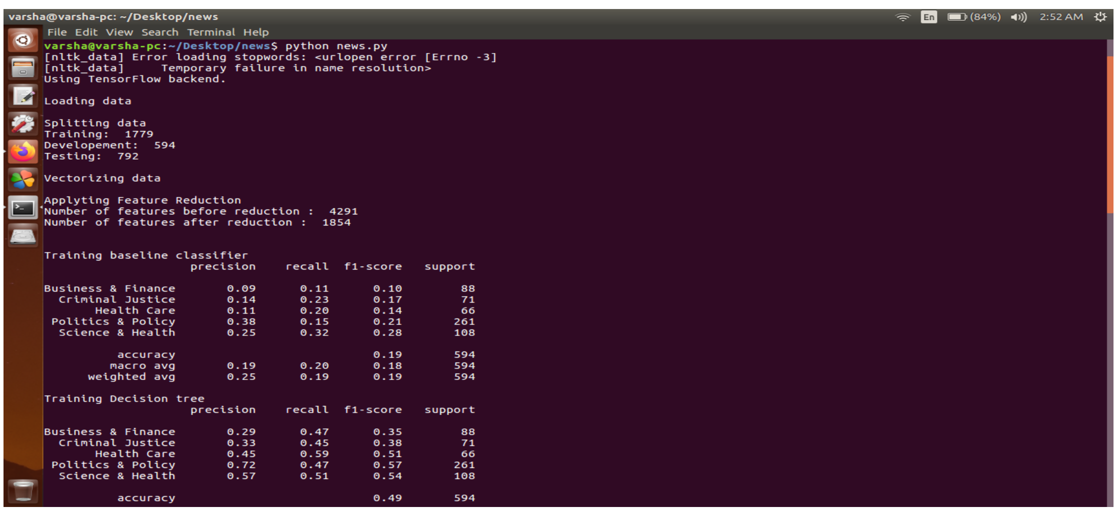
Output of lex code:



```
varsha@varsha-pc:~$ ./a.out
The cm wants the public to be in quarantine. -- political news
They catched the man who involved in robbery. -- news related to robbery
The cricket team won gold this time. -- sports news
Bareilly: Two held for stealing 24 mobile tower batteries. -- news related to robbery
Delhi fair price shop owner arrested for swindling food grains. -- crime news
Doctors found medicine for corona. -- medical news
Love island winner oshea gave up celebrity lifestyle for olympic dream. -- sports news
Coronavirus in India: Doctor treating 14 Italian tourists gives a peek into behind-the-scene. -- medical news
Trump holds support of political base in virus-prone states. -- political news
Haryana man confesses to Tirupur bank robbery. -- news related to robbery
WHO and partners call for urgent investment in nurses. -- medical news
Bulgaria sports minister points fingers after ban violation reports. -- sports news
Teenager arrested in Noida for rape, murder of 8-year-old girl. -- crime news
Wisconsin set to hold in-person voting in presidential primary. -- political news
Jharkhand: Minor girl accuses friend, 8 others of rape in Dumka. -- crime news
Daylight robbery case: Punjab cops recover 10kg gold. -- news related to robbery
English football League needs fifty six days to finish season EFL chief. -- sports news
Trump says will ask congress for more small business funds if money runs out. -- political news
Texas: Indian-American mother arrested for killing minor son. -- crime news
Surat: robbery accused held after 21 years. -- news related to robbery
Coronavirus may 'reactivate' in cured patients: Korean CDC. -- medical news


The number of news in POLITICS category 4
The number of news in SPORTS category 4
The number of news in CRIME category 4
The number of news in ROBBERY category 5
The number of news in MEDICINE category 4
varsha@varsha-pc:~$
```

Output of python code:



```
varsha@varsha-pc:~/Desktop/news$ python news.py
[nltk_data] Error loading stopwords: <urlopen error [Errno -3]
[nltk_data]     Temporary failure in name resolution>
Using TensorFlow backend.

Loading data

Splitting data
Training:  1779
Developement:  594
Testing:  792

Vectorizing data

Applyting Feature Reduction
Number of features before reduction :  4291
Number of features after reduction :  1854


Training baseline classifier
                    precision    recall  f1-score   support

Business & Finance       0.09      0.11      0.10        88
  Criminal Justice       0.14      0.23      0.17        71
       Health Care       0.11      0.20      0.14        66
 Politics & Policy       0.38      0.15      0.21       261
  Science & Health       0.25      0.32      0.28       108

          accuracy                           0.19       594
         macro avg       0.19      0.20      0.18       594
      weighted avg       0.25      0.19      0.19       594

Training Decision tree
                    precision    recall  f1-score   support

Business & Finance       0.29      0.47      0.35        88
  Criminal Justice       0.33      0.45      0.38        71
       Health Care       0.45      0.59      0.51        66
 Politics & Policy       0.72      0.47      0.57       261
  Science & Health       0.57      0.51      0.54       108

          accuracy                           0.49       594
```

```
varsha@varsha-pc: ~/Desktop/news
File  Edit  View  Search  Terminal  Help
            accuracy                           0.49       594
           macro avg       0.47      0.50      0.47       594
        weighted avg       0.55      0.49      0.50       594

Training Random Forest
                    precision    recall  f1-score   support

Business & Finance       0.37      0.49      0.42        88
  Criminal Justice       0.35      0.51      0.41        71
       Health Care       0.48      0.70      0.57        66
 Politics & Policy       0.77      0.49      0.60       261
  Science & Health       0.55      0.57      0.56       108

          accuracy                           0.53       594
         macro avg       0.50      0.55      0.51       594
      weighted avg       0.59      0.53      0.54       594

Training Multinomial Naive Bayesian
                    precision    recall  f1-score   support

Business & Finance       0.51      0.58      0.54        88
  Criminal Justice       0.51      0.52      0.51        71
       Health Care       0.49      0.64      0.56        66
 Politics & Policy       0.74      0.60      0.66       261
  Science & Health       0.60      0.69      0.64       108

          accuracy                           0.61       594
         macro avg       0.57      0.60      0.58       594
      weighted avg       0.62      0.61      0.61       594

Training Support Vector Classification
                    precision    recall  f1-score   support

Business & Finance       0.44      0.27      0.34        88
  Criminal Justice       0.41      0.10      0.16        71
       Health Care       0.56      0.29      0.38        66
 Politics & Policy       0.53      0.88      0.66       261
  Science & Health       0.61      0.32      0.42       108

          accuracy                           0.53       594
         macro avg       0.51      0.37      0.39       594
      weighted avg       0.52      0.53      0.48       594
```

```
varsha@varsha-pc: ~/Desktop/news
File  Edit  View  Search  Terminal  Help
  Science & Health       0.61      0.32      0.42       108

          accuracy                           0.53       594
         macro avg       0.51      0.37      0.39       594
      weighted avg       0.52      0.53      0.48       594

Training Multilayered Perceptron
                    precision    recall  f1-score   support

Business & Finance       0.52      0.45      0.48        88
  Criminal Justice       0.49      0.46      0.48        71
       Health Care       0.42      0.42      0.42        66
 Politics & Policy       0.67      0.68      0.67       261
  Science & Health       0.57      0.62      0.59       108

          accuracy                           0.58       594
         macro avg       0.53      0.53      0.53       594
      weighted avg       0.58      0.58      0.58       594



Predicting test data using Multinomial Naive Bayesian
                    precision    recall  f1-score   support

Business & Finance       0.39      0.51      0.44        97
  Criminal Justice       0.67      0.57      0.61        99
       Health Care       0.47      0.62      0.54        90
 Politics & Policy       0.73      0.62      0.67       333
  Science & Health       0.62      0.65      0.64       173

          accuracy                           0.61       792
         macro avg       0.58      0.59      0.58       792
      weighted avg       0.63      0.61      0.61       792



Incorrectly classified

Title:  how obamacare could cut your car insurance bill
True Category:  Health Care
Predicted Category:  Politics & Policy
varsha@varsha-pc:~/Desktop/news$
```