Lei (Raymond) Chi

Text Categorization

This project implements the Naïve Bayes method for text categorization. The program begins by reading the file and its corresponding category. The **tokenize_file()** function then processes the articles by removing punctuations using **string.punctuation**, tokenizing with **nltk.tokenize**, eliminating stopwords with **nltk.corpus**, and performing stemming using **PorterStemmer**. After tokenization, the program calculates the prior and likelihood values, which are then used in the Naïve Bayes equation to determine the category with the highest probability.

Two types of smoothing methods were explored in this project: Laplacian smoothing and Jelinek-Mercer (JM) smoothing. JM smoothing performed best with corpus 1 and 3, achieving accuracy above 90%, but reduced the accuracy of corpus 2 to only 80%. As a result, Laplacian smoothing was chosen, with a constant alpha of 0.058 (tuned through multiple testings). The results below were obtained using this method. For corpus 2 and 3, the data was split in a 55/45 ratio.

To run the code:

1. Place **CHI_naive_bayes.py** in the **/TC_provided** directory.
2. Run the code.
3. When prompted, enter the name of the file containing the list of labeled training documents (e.g., **./corpus1_train.labels**).
4. Enter the name of the file containing the list of unlabeled test documents (e.g., **./corpus1_test.list**).
5. The program will generate a file named e.g., **predicted_corpus1_test.labels** in your directory (the **corpus#** is determined by the first word of the file name you input).
6. To check accuracy, use the command: **perl analyze.pl predicted_corpus1_test.labels corpus1_test.labels**.

Corpus 1 performance:

```
Found 5 categories: Oth Dis Pol Str Cri
Processing prediction file...

394 CORRECT, 49 INCORRECT, RATIO = 0.889390519187359.

CONTINGENCY TABLE:
        Oth     Dis     Pol     Str     Cri     PREC
Oth     13      0       2       0       0       0.87
Dis     3       88      0       1       0       0.96
Pol     5       0       123     4       1       0.92
Str     3       1       18      128     7       0.82
Cri     1       0       1       2       42      0.91
RECALL  0.52    0.99    0.85    0.95    0.84

F_1(Oth) = 0.65
F_1(Dis) = 0.972375690607735
F_1(Pol) = 0.888086642599278
F_1(Str) = 0.876712328767123
F_1(Cri) = 0.875
```

Corpus 2 performance:

```
Found 2 categories: O I
Processing prediction file...

341 CORRECT, 62 INCORRECT, RATIO = 0.846153846153846.

CONTINGENCY TABLE:
        O       I       PREC
O       258     38      0.87
I       24      83      0.78
RECALL  0.91    0.69

F_1(O) = 0.892733564013841
F_1(I) = 0.728070175438596
```

Corpus 3 performance:

```
Found 6 categories: Wor USN Sci Spo Fin Ent
Processing prediction file...

387 CORRECT, 43 INCORRECT, RATIO = 0.9.

CONTINGENCY TABLE:
        Wor     USN     Sci     Spo     Fin     Ent     PREC
Wor     151     10      5       0       0       1       0.90
USN     2       106     3       2       4       3       0.88
Sci     1       0       40      0       1       0       0.95
Spo     0       0       0       33      0       0       1.00
Fin     1       0       3       1       44      2       0.86
Ent     1       0       3       0       0       13      0.76
RECALL  0.97    0.91    0.74    0.92    0.90    0.68

F_1(Wor) = 0.934984520123839
F_1(USN) = 0.898305084745763
F_1(Sci) = 0.833333333333333
F_1(Spo) = 0.956521739130435
F_1(Fin) = 0.88
F_1(Ent) = 0.722222222222222
```