Assignment 3
Text Classification of text documents
IR 2020

Deadline: 15th November

November 2, 2020

This assignment is on classification of text documents. It is highly recommended that you use python3 for this assignment as libraries like nltk will make many things easier (stop word removal and lemmatization). However, if you use any other language, you most probably have to design these modules yourselves which might not perform as good as nltk library in python. Also you can use packages from scipy for classification purposes.

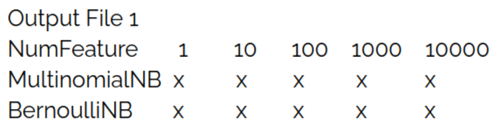
Assumptions: Remove stop words, punctuation marks, make everything to lowercase and perform lemmatization to generate tokens from the document (use nltk library in python). Assume positional and class conditional independence of the terms in document for Naive Bayes. For vector space classifications, assume each document is represented by its normalized tf-idf vector representation. The tf-idf vector construction follows the same procedure as stated in assignment 2.

Find your dataset here

1. Naive Bayes with feature selection
   (a) Select top x features using mutual information from both train data. Vary x in $\{1, 10, 100, 1000, 10000\}$.
   (b) Using each of the above x, train a multinomial Naive Bayes on the given train data, with add-one smoothing.
   (c) Using each of the above x, train a Bernoulli Naive Bayes on the given train data.
   (d) Print F_1 score for each of the classifier on the test data for each of the feature value.
2. Vector space classification - Linear: Use Rocchio classifier to classify documents in the test data and print the F_1 score. For Rocchio classifier, use the decision rule as follows. Assign d to class c iff $|\mu(c) - \nu(d)| < |\mu(\bar{c}) - \nu(d)| - b$. Vary b within the range $\{0, .01, .05, .1\}$ and print F_1 score for the b values.
3. Vector space classification - Non linear: Use kNN classifier to classify documents in the test data and report F_1 score. Vary k in $\{1, 10, 50\}$. For

similarity score use inner product of vector representation of two documents. Print F_1 scores on test data.

Instructions for submission Submit your codes named as Rollno_Taskno.py. Your code will be executed as
python3 your-code.py path-data-directory output-file
You should print F_1 scores in space separated manner in the output file as simple text. Sample output should look like



```
Output File 1
NumFeature    1     10    100   1000   10000
MultinomialNB x     x     x     x      x
BernoulliNB   x     x     x     x      x
```

Rest two files resembles similar structure.

Plagiarism Rules: If your code matches (more than 50%) with another student's code, all those students whose codes match will be awarded with zero marks without any evaluation. Therefore, it is your responsibility to ensure you neither copy anyone's code nor anyone is able to copy your code.

Code error: If your code doesn't run or gives error while running, you will be awarded with zero mark.