Information Retrieval (CS4051)
Programming Assignment No. 3
Spring 2022

Submission Date: May 25, 2022

Assignment Objective

This assignment focuses on feature selection for text classification or categorization. Feature selection is the process of selecting a subset of the features (in text documents it can be word/term, phrases, sequences, etc) occurring in the training set and using only this subset as features in text classification. Feature selection serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. Second, feature selection often increases classification accuracy by eliminating noise/redundant/irrelevant features. Generally, by feature selection you can improve the classifier performance easily. Text is especially rich in feature space and thus one of the challenge is to reduce the curse of the dimensionality. In this programming assignment you need to implement few feature selection schemes and later perform text classification using Naïve Bayesian approach as discussed during the lecture.

Datasets

The data set for this assignment is a different view of the famous WebKB dataset(http://www.cs.cmu.edu/~webkb/). In order to simplify your work, we transform it into a binary classification problem. You need to classify a given webpage into two categories "Course" and "Non Course". There are two folders in the zipped file one contains full-text for the course and non-course content. The other folder contains in-link for the both. You can also use this link information to improve your classification.

Feature Selection

1. Term Frequency/ Inverse Document Frequency (tf*idf)- as a first feature selection you need to select top 100 features based on tf*idf scoring scheme. It is you task to adjust tf and df frequencies accordingly to suits your selection.

2. Topic Terms co-occurrence based. You need to find top 50 most frequent noun (Topics base-set) from the entire collection You can use any language processing library to decide whether a term is noun or not (use any dictionary WordNet etc) {There are other innovative quick approaches as well- try it}. For the rest of the terms in the collection you need to find all occurrences of each other term with the topic base-set and add all co-occurrences term in each topic set. Hence you will be able to get some good topic set from all the collection. Using only these terms in all possible topic groups you can apply NB for the classification.

3. Lexical Chains - A lexical chain is a sequence of related words in writing, spanning short (adjacent words or sentences) or long distances (entire text). A chain is independent of the grammatical structure of the text and in effect it is a list of words that captures a portion of the cohesive structure of the text. You can get a lexical chain from the collection for each pair of words try to get it relation from WordNet. A common similarity like Wu & Palmer - It calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with

the depth of the LCS (Least Common Subsume) you can use any threshold value to decide the possible chains and their length. Collect all the possible lexical chains and use these chains to represent the document features. Using only these terms in all possible chains you can apply NB for the classification.

4. Mixed Features (Combining all above three) – Now try combing all features form above and apply NB for this space.

Files Provided with this Assignment:

1. Modified WebKB Dataset

Evaluation/ Grading Criteria

The grading will be done as per the scheme of implementations, feature selection and classification accuracy, precision, recall and F1- Measure.

Grading Criteria:
Preprocessing (2 marks)
Feature selection (you need to implement it from the scratch) (4 marks)
Naïve Bayesian classifier (2 marks)
Code Clarity/ Proper commenting (1 mark)
Bonus: GUI (1 mark for making the GUI 1 mark for Good Looking GUI)

<The End>