Identification of the class of English text among several classes

Introduction :

Generally , the project is a classification software system that works on unprepared dataset as a training dataset . The first task that the software should do is preparing the training dataset which is the first argument for the system . The training dataset should be collections of English texts such as articles ,or news . These texts should be classified into two or more classes (i.e there is a known class for each text file) . After the training phase based on a chosen dataset , the system should be able to classify any text file into one of the dataset classes using statistical learning techniques .

Objectives :

Mainly , the project is aimed to classify any English text into a class based on the training dataset that has been chosen before .

Theory :

The idea behind the project can be simplified by the definition of supervised learning in machine learning field . The aim of the project is to use classification methods such as Bayes classifier or Decision tree to predict the class for the text input .

Preparing the dataset phase should be done through certain steps , these steps should lead to a dataset that contains a vector of words for each class . These words should be filtered from punctuations , numbers , stop words , and any other symbol that doesn’t belong to English words . After these filtering steps , all words should pass through stemming process to get the root form for each word . To apply this process , the Stemmer (The algorithm that will be used to stemming process) should be efficient enough to produce appropriate results . After stemming process , the dataset should be ready for classification phase .

Steps :

As it has been said above , the system could be logically splitted to two main phases and they are Preparing data phase and classification phase consecutively . Each phase should have a plan for stand-alone construction purpose and another plan for the interaction with the other phase for system integration purpose . All of these steps will be discussed at this section .

Firstly , preparing data phase :

The processes that should be done for the dataset are :

1. Transform each text file for a vector of words based on spaces between the words .

2. Apply each vector of words that has been extracted from each text file to stop words filtering .
3. Then apply it to numerical and punctuation symbols filtering .
4. Then apply the result vector to stemming process .

These steps require efficient punctuation list , efficient stop words list , efficient Stemmer . We have chosen each of these requirements carefully , and we have tested two stemming algorithms to get the best one . Firstly we have applied Porter algorithm , it wasn’t the best so we have applied Stanford CoreNlp for only stemming purpose as an external system . Stanford CoreNlp is one of the most integrated for natural language processing . After testing it we have discovered its efficiency over Porter algorithm .

After these steps of planning , we implemented this phase as a module in Java . Hence , the dataset is required to measure the efficiency of our system modules . The dataset should satisfy the conditions that we have discussed it above . After searching we found an appropriate and certificated dataset .

Secondly , classification phase :

the classification phase is what we are studying right now . Its planning and implementation will take place at the next semester .

Future work :

The system can be widely used in modern technology applications . Generally if there is an available and suitable dataset for books and papers will be enough to classify millions of other unclassified books and papers could be useful for commercial, educational purposes . In particular we for developing this system to act as search engine indexing module that make an index for pages on the Internet with their keywords , in our case the indexing module will play a different role . The indexing module will index the page and could be a section of the page by its class . Which is very useful for searching by the meaning not by the keywords .

References :

1. BBC articles and news dataset . http://mlg.ucd.ie/datasets/bbc.html
2. Porter stemming algorithm . http://people.scs.carleton.ca/~armyunis/projects/KAPI/porter.pdf https://tartarus.org/martin/PorterStemmer/
3. Stanford CoreNLP . https://stanfordnlp.github.io/CoreNLP/
4. The elements of statistical learning , 2nd Edition . hastie , tibshirani , friedman .
5. Introduction to machine learning , 2nd Edition . Ethem Aplaydm .
6. Data Mining: Concepts and Techniques, 3rd Edition. Jiawei Han, Micheline Kamber, Jian Pei .