Identification of the class of English text among several classes

Abstract :

The aim of the project is to classify an unknown English text into one of several classes. A collection of files containing English text will be gathered from the Internet . Each file will be classified into one of a group of classes according to the content of the file. Learning techniques will be used to differentiate between and classify each file to the corresponding class. Success rates will be calculated.

Introduction :

The project is a research based one in which we study and understand the steps to be implemented to classify raw data of an English text file into a class depending on the training dataset that should be selected as a first step . The dataset to be used is divided into two groups namely a training one (80%) and another testing one (20%) . The Dataset should be prepared for supervised learning techniques (i.e all files are classified by the provider of the dataset into classes according to the content of the each file) . As for the training dataset, a software system will be designed and developed to load the content of each file within each class and then it will do certain steps to preparing data (Text processing) . These steps includes removing all words that don't add a meaning to the context of the file , removing all symbols and punctuation marks , and stemming . After these steps , we should get a vector of words to represent each class separately and a number to represent the frequency of occurrence for each word . We will repeat the same steps for the testing group and then we will build a classifier based on the training dataset that should be able to classify the unlabeled testing dataset . The percentage of success rate in deciding the right class to which each file in the testing dataset belongs will be measured to see the efficiency of the whole process .

Objectives :

At the end of the project we should be able to measure the percentage of success rate in deciding the right class for the testing dataset to evaluate this approach . During the experiments we will try to improve the steps to get the highest possible accuracy .

Theory :

Generally , we should divide the whole process into phases to study and improve each of which separately to get the required result . The process should be divided to five phases , and they are : data selection , preparing the dataset , building a classifier model , testing and validation , and accuracy measurement . The dataset that will be used should be large enough to give a consistent vector of words that will represent the class at the end . The testing group should be a part from the selected dataset and its label should be known to measure the accuracy of the classification . Preparing the dataset should be done through removing Stopwords , removing punctuation marks and symbols , getting the root form for

each remaining word through stemming process using stemming algorithm (Stemmer) , calculating the frequency of occurrence for each word to use the frequencies of words in the classification . The classifier that will be used should be chosen carefully after studying the advantages of each candidate classification method . Testing / Validation should be the criteria for improving and adjustment the process steps .

Steps :

As it has been stated in the theory section , the process should be divided into five phases .

Firstly , Dataset Selection :

The selected dataset should have four characteristics and they are : certificated , scientifically collected , clean , highly recommended . While we are searching about the dataset that will be used we found that BBC news and articles dataset satisfy all conditions.

Secondly , preparing data phase :

The processes that should be done for the dataset are :

1. Transforming each text file for a vector of words based on spaces between the words .
2. Applying each vector of words that has been extracted from each text file to stopwords filtering .
3. Applying the remaining words to symbols and punctuation marks filtering .
4. Applying the remaining words to stemming process .

These steps require efficient punctuation list , efficient stop words list , efficient Stemmer . We have chosen each of these requirements carefully , also testing more than one Stemmer is recommended to choose the best one . There are two candidate Stemmers and they are : Porter algorithm and Stanford CoreNlP software . Porter Algorithm is the most common English Stemmer and it's simple . Statistically the results of Porter algorithm is good but not the best . Stanford CoreNLP software is one of the most powerful integrated systems for NLP and uses complex techniques for stemming .

After Stemming process , a software tool should be developed to count the number of occurrences for each word , and then the words with highest frequency will be chosen to represent the file within each class . The minimum frequency will be chosen to increase the accuracy of the classification .

Thirdly , building a classifier model :

There are two possible choices and they are : Naive bayes classifier and Decision tree classifier . Both should be tested to choose the best according to the accuracy measurement .

Fourthly , Testing and validation Then Accuracy measurement :

Testing and validation will be done by trying to classify files that will be already labeled but we will do this step to determine the the accuracy measurement using the testing group (20% of The selected dataset) and then we will calculate the percentage of success rate in deciding the right class and in each experiment we will try to get the highest accuracy by increasing the minimum frequency for the words that will represent each class , or removing mutual words that will represent more than one class , and testing more than one Stemmer to get the best one and also The classifier could be changed , or any other method could help to increase the accuracy measurement

Future work :

The process 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 are expecting that this approach could be used to build a system that will act as search engine indexing module that make an index for pages on the Internet with their keywords , in this 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 .