Assignment 5

Dibyendu Roy Chaudhuri
MT19034

Naive Bayes and KNN

Methodology:

1. Firstly, I have created a dictionary of terms that have a list of document ids whiches hold that term and frequency of that term in that document.
2. Secondly another dictionary of documents containing different terms it contains, their frequency.
3. I have also maintained another list where document names are stored and mapped with their document ids.
4. Then using the above dictionary, I have generated another list of five lists where each list contains, the terms each class contains.
5. Using the above data structure, I have calculated tf-Idf scores of terms where each class is assumed to be a document and MI scores.
6. Now I have chosen the best k features to calculate a common document space to represent each training and test documents. This document space is essential for Implementing KNN algorithm.
7. Values of k are varied from 1 to 5.
8. While implementing the naive Bayes algorithm, I have used add-one smoothing to find the posterior probability.

Preprocessing Step:

1. Tokenize is done on following delimiters-

("\s", "-", ".", "@", "t", "\n", ‘"’, ">", ",", "?", ":", "{", "(", "[", ")", "}", "]", "<", "_", "!", "/", "|", "\", "*", "=", "^","&", "%", "$", "!")

2. Convert whole text into lower case.
3. Convert num to text.
4. Lemmatization is used.
5. Pickle library is used to store the index file.
6. Alphanumerical words are removed.
7. Stop words are removed.

Assumption:

1. For Tf formula, binary formula is used.
2. KNN vector is built using 0-1 only.
3. Word length greater than 1 is taken as valid terms.
4. Stop words are removed.
5. I have assumed there is no alphanumeric word. So I just removed numeric digit during processing.
6. Add-one smoothing is used during calculating posterior probability.