IR ASSIGNMENT 5 README FILE

MT19021

Running the codes:

I have made 3 folders in the submission:

- **IPYNB FILES**: Contains the jupyter notebooks of all codes along with proper documentation and all the results obtained.
- **PYTON FILES**: Contains the python(.py) files of all codes.
- **PICKLE FILES**: Contains all the pickles created after feature extraction.

ALL THE PICKLES NEED TO BE PRESENT IN THE SAME FOLDER AS THE CODE FILES WHILE RUNNING THE CODE.

Also, if the feature extraction code needs to be run then the folder 20newsgroups also needs to be present at the same location as the code file.

Pre-processing steps: The pre-processing steps used for the dataset.

- All the text is converted to lower case
- All meta-data which is present in the first few lines of each document is removed from the text
- All email-ids are removed from the text
- Tokens which consist of letters as well as digits like play2 are removed from the text.
- All the punctuation marks are removed from the text
- All stop-words are removed from the text

Lemmatization is applied on the text of the dataset.

Methodology Used:

1. **Naïve Bayes algorithm:**
    - The training is done on the feature set obtained on doing the union of the top extracted features of all classes.
    - During training the prior probabilities of the class and the probability of a term being present in a class are used.
    - Now for each of the testing document the score for each class needs to be calculated.
    - To find the score only those terms that are present in the feature set are used.

    - The test doc is assigned the class for which the maximum score is obtained.

2. **KNN algorithm:** Each of the testing document needs to be assigned to a class. For this the following steps are followed:
    - Cosine similarity values are found between the testing doc vector and each of the training doc vectors.
    - The training docs are sorted in the descending order of the cosine similarity value.
    - The top k training documents are taken from the sorted list and they are the k-nearest neighbors.
    - Now the majority class among the classes of the k nearest neighbors is assigned as the class of the testing doc.