Medical Text Classification

Name
Varun Shah

Student ID
010823657

Rank
36

F1-Score
0.7340

Data Preprocessing
Performed following steps for data preprocessing:

1. Read data from “train.dat” and stored each document in the list lines. Same way read data from “test.dat” and stored each document in the list lines_test.

2. Classes and documents are split by “\t” and classes stored in classes list and documents stored in lines_train list

3. Using Hashing, transformed lines_train documents list to CSR Matrix and stored it to X_train. Also, applied Hashing technique to transform lines_test documents list to CSR Matrix and stored in Y_train

4. Classes stored into X_test as a target values to train a model

5. For hashing, three parameters were passed for it. Stopword with value ‘english’, norm with value ‘l2’ and lowercase with value true

Classifier Model Development
Performed following steps for Classifier Model Development:

1. Classification model developed with K Nearest Neighbors Classifier used with 120 n-neighbors (KNeighborsClassifier from sklearn.neighbors)

2. Classification model implemented using KNeighborsClassifier with 120 object to classify method

3. Model trained by calling fit method of KNeighborsClassifier and passing X_train and X_test as parameters

4. As the model trained, pred method of KNeighborsClassifier called to predict classes of y_train (test data set)

Methodology:

Fetched train data from file and split it into list of classes and list of documents. Classes represent target list and correspondent list represents that class. As per training data there are 5 classes.

Used hashing technique (HashingVectorizer of sklearn.feature_extraction.text) to term frequency count in each document and transformed the documents list to CSR Matrix setting following parameters:

- stop_words: removing English.
- norm: Used L2 Norm (Euclidian Distance) as there are multiple dimensions in data.
- lowercase: true. Setting all words to lowercase before indexing can provide correct frequency count in each document

K Nearest Neighbor Classification Model used for prediction. Following are the parameters set for training model and prediction.

- n-neighbors: 120 because ideally it is set to squareroot of number of documents
- weights: Uniform. In order to weigh all points in each neighbourhood equally
- algorithm: Auto. As it will decide the most appropriate algorithm (ball_tree, kd_tree, brute) based on the parameters passed to fit method

Classifier's fit method trains model taking following parameters:

- X_train CSR Matrix
- X_test target classes list.

After training the model, it is applied on Y_train CSR Matrix using predict method of classifier with following parameter:

- y_train: CSR Matrix

Using trained model with test data set, output stored in list and written in output file.