CEG 7380 Information Retrieval

Apurvaa Kaamesh U00918186
Vidhya Lakshmi Sankaranarayanan U00924508

I. Implementation

1. **Parsedocs.py –** This file reads all the individual files and directories in the dataset and extracts the doc_id, class_name, and subject and body contents.

2. **Util.py –** This file performs two pre-processing operations like removing Stopwords and performing stemming.
   Stopword removal – Using Stopwords list for English language from NLTK package.
   Stemming **–** We have used Porter stemmer.

3. **Index.py –** We use the Inverted Index model here. Initially, we tokenize the subject and body contents extracted from dataset using Regular Expression tokenizer. In the Index file, for each term, we have The postings list, Term Frequency, class name, IDF and Positions of each term in the document.

4. **Feature_Extraction.py** – In this file, we generate the Class definition file, Feature definition file and Training data file for different feature values (TF, IDF and TFIDF). Class definition file stores the labels for the different classes available that would help us in classifying the data.
   Feature definition file stores the individual terms and assigns each term an ID. This helps in minimizing the storage required to store complete terms in training data file.
   Training data file is generated in LIBSVM format. This format works well with sparse datasets where all the features with value zero are eliminated.
   Each line in the file represents a document. We have the class label for that document alongwith list of feature id's and their respective values.

5. **Classification.py** - In this file, we classify the entire training data set using different algorithms and compare their accuracy. The Supervised algorithms used in the design are Multinomial Naïve Bayes(MNB), Multivariate Bernoulli(BNB), K-nearest neighbour(KNN) and Support Vector Machine(SVM).
   To calculate accuracy and standard deviation of each classifier, we use the 5-fold cross validation method and the Scoring metric used are F1, Precision and Recall. F1 metric synthesizes both Precision and Recall.

For better results, we use training dataset with term frequency features for MNB, training dataset with IDF features for BNB and training dataset with TFIDF features for KNN and SVM classifier.
Also, for SVM classifier, we used the linear kernel, for improving the accuracy.

6. **Feature Selection.py –** To reduce the computation and improve accuracy by classifying the data based on some selected features, we use two feature selection methods, Chi-square (Chi2) and Mutual Information (MI).
   Chi2 checks for the independence between term and class. If Chi2 is zero, that concludes that terms and classes are independent of each other. As Chi2 value increases, the correlation between terms and classes increases.
   Mutual Information: Mutual Information computes the extent of dependency between term and classes.

   After selecting the features, we classify the data once again and compare accuracy between the different supervised algorithms. The main aim here is to decide, how many features when selected give highest accuracy. To decide this, we check for different values for K features(100,500,1000,5000 and 10000).

7. **Clustering.py** – In this file we perform clustering on the training dataset using both Flat clustering algorithm K-Means and Heirarchial clustering algorithm Agglomerative (Bottom-up approach).
   KMeans Clustering - Kmeans Clustering performs grouping based on computing Euclidean distance between the features and the centroids. Features having minimal distance with the centroid belong to that cluster.
   We compute Silhouette and Mutual Information score for both the algorithms and compare them.
   The main task is to decide on the number of clusters required for classification. To decide this we compute the scores for different number of clusters in the range (2-25).

II Experimental Results

- We observe that after parsing all the files in the dataset, there are about total number of 2000 documents. We have been provided with 6 Class labels as follows:

  Class 1: 'comp.graphics', 'comp.os.ms-windows.misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware','comp.windows.x'
  Class 2: rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey'
  Class 3: 'sci.crypt', 'sci.electronics', 'sci.med', 'sci.space
  Class 4: 'misc.forsale'
  Class 5: 'talk.politics.misc', 'talk.politics.guns', 'talk.politics.mideast'
  Class 6: 'talk.religion.misc', 'alt.atheism', 'soc.religion.christian'

- We build the Inverted Index after pre-processing the tokens. The total number of tokens in index is 31591. So, the feature definition file has 31591 feature id's.
- For comparison purposes, we generated three training files based on TF, IDF and TFIDF feature values respectively. Each training file has 2000 lines, one for each document in the dataset.
- Our classification results for the supervised classification algorithms without feature selection are as follows:

Classifier	**F1 Macro**	**Precision Macro**	**Recall Macro**
Multinomial Naïve	0.69 +/- 0.06	0.71 +/- 0.15	0.70 +/- 0.04
Bernoulli Naïve	0.45 +/- 0.10	0.57 +/- 0.06	0.45 +/- 0.12
K Nearest Neighbor	0.15 +/- 0.04	0.57 +/- 0.13	0.21 +/- 0.02
Support Vector Classifier	0.71 +/- 0.08	0.75 +/- 0.09	0.70 +/- 0.08

- After applying Feature selection method Chi2 and Mutual information, the accuracy of all these algorithms are as follows:

K value	**Multinomial**	**Bernoulli**	**K Neighbor**	**SVC**
100	0.57	0.68	0.57	0.59
500	0.73	0.77	0.58	0.67
1000	0.78	0.75	0.53	0.68
5000	0.86	0.68	0.35	0.72
10000	0.85	0.61	0.25	0.72

Chi2 feature selection

K value	**Multinomial**	**Bernoulli**	**K Neighbor**	**SVC**
100	0.62	0.69	0.55	0.60
500	0.76	0.76	0.50	0.60
1000	0.79	0.78	0.46	0.56
5000	0.84	0.66	0.29	0.27
10000	0.83	0.61	0.24	0.15

Mutual Information

- The clustering algorithms have the below effect on Silhouette and Mutual information score:

  for value 2
  Kmeans clustering
  silhouette score is 0.9915205792283464
  Hierarchical clustering
  silhouette score is 0.9915205792283464
  for value 3
  Kmeans clustering

silhouette score is 0.9756057289315788
Hierarchical clustering
silhouette score is 0.9756057289315788
for value 4
Kmeans clustering
silhouette score is 0.9718726732573182
Hierarchical clustering
silhouette score is 0.9718726732573182
for value 5
Kmeans clustering
silhouette score is 0.9693940745884608
Hierarchical clustering
silhouette score is 0.9693940745884608
for value 6
Kmeans clustering
silhouette score is 0.9627739557724379
Hierarchical clustering
silhouette score is 0.9627739557724379
for value 7
Kmeans clustering
silhouette score is 0.9609384404228345
Hierarchical clustering
silhouette score is 0.9609384404228345
for value 8
Kmeans clustering
silhouette score is 0.9205951411613863
Hierarchical clustering
silhouette score is 0.9205951411613863
for value 9
Kmeans clustering
silhouette score is 0.9100938660027817
Hierarchical clustering
silhouette score is 0.9210591479127231
for value 10
Kmeans clustering
silhouette score is 0.9192294087020125
Hierarchical clustering
silhouette score is 0.9192294087020125
for value 11
Kmeans clustering
silhouette score is 0.9045454012974676
Hierarchical clustering
silhouette score is 0.9064383642326613

for value 12
Kmeans clustering
silhouette score is 0.8231116179987295
Hierarchical clustering
silhouette score is 0.9048961338639009
for value 13
Kmeans clustering
silhouette score is 0.8223401312254216
Hierarchical clustering
silhouette score is 0.8790266915556142
for value 14
Kmeans clustering
silhouette score is 0.8221696838927873
Hierarchical clustering
silhouette score is 0.8772386330629344
for value 15
Kmeans clustering
silhouette score is 0.8246076244023314
Hierarchical clustering
silhouette score is 0.8770939817077404
for value 16
Kmeans clustering
silhouette score is 0.8530938244720357
Hierarchical clustering
silhouette score is 0.8721187461296671
for value 17
Kmeans clustering
silhouette score is 0.8532830106609999
Hierarchical clustering
silhouette score is 0.8228265200087777
for value 18
Kmeans clustering
silhouette score is 0.8189586230428018
Hierarchical clustering
silhouette score is 0.8233144781912763
for value 19
Kmeans clustering
silhouette score is 0.7530804856004623
Hierarchical clustering
silhouette score is 0.8237351026062272
for value 20
Kmeans clustering
silhouette score is 0.6556047661880333

Hierarchical clustering
silhouette score is 0.8246364315982346
for value 21
Kmeans clustering
silhouette score is 0.743786035200312
Hierarchical clustering
silhouette score is 0.8233154610758311
for value 22
Kmeans clustering
silhouette score is 0.7753721772922851
Hierarchical clustering
silhouette score is 0.823664277881277
for value 23
Kmeans clustering
silhouette score is 0.7472821211416304
Hierarchical clustering
silhouette score is 0.8221428695927464
for value 24
Kmeans clustering
silhouette score is 0.7756384428223368
Hierarchical clustering
silhouette score is 0.8224686589089358
for value 25
Kmeans clustering
silhouette score is 0.5778689892984681
Hierarchical clustering
silhouette score is 0.822788543344536
for value 2
Kmeans clustering
mutual information score is 0.00808917707502087
Hierarchical clustering
mutual information score is 0.00808917707502087
for value 3
Kmeans clustering
mutual information score is 0.01144619158823107
Hierarchical clustering
mutual information score is 0.01144619158823107
for value 4
Kmeans clustering
mutual information score is 0.014026476208641775
Hierarchical clustering
mutual information score is 0.014026476208641775
for value 5

Kmeans clustering
mutual information score is 0.017679415297953006
Hierarchical clustering
mutual information score is 0.017679415297953006
for value 6
Kmeans clustering
mutual information score is 0.01944146994694971
Hierarchical clustering
mutual information score is 0.01944146994694971
for value 7
Kmeans clustering
mutual information score is 0.021576396241209543
Hierarchical clustering
mutual information score is 0.021576396241209543
for value 8
Kmeans clustering
mutual information score is 0.02468059316190584
Hierarchical clustering
mutual information score is 0.02468059316190584
for value 9
Kmeans clustering
mutual information score is 0.026394482251794155
Hierarchical clustering
mutual information score is 0.026394482251794155
for value 10
Kmeans clustering
mutual information score is 0.028008480382836626
Hierarchical clustering
mutual information score is 0.028008480382836626
for value 11
Kmeans clustering
mutual information score is 0.02953431572574835
Hierarchical clustering
mutual information score is 0.03120229918816747
for value 12
Kmeans clustering
mutual information score is 0.03276069989149221
Hierarchical clustering
mutual information score is 0.03098909299528291
for value 13
Kmeans clustering
mutual information score is 0.032378221034406275
Hierarchical clustering

mutual information score is 0.030832709966213467
for value 14
Kmeans clustering
mutual information score is 0.03588148120388798
Hierarchical clustering
mutual information score is 0.033505932192950984
for value 15
Kmeans clustering
mutual information score is 0.03541689706145983
Hierarchical clustering
mutual information score is 0.0333233827955115
for value 16
Kmeans clustering
mutual information score is 0.04167735872894354
Hierarchical clustering
mutual information score is 0.03523169724578386
for value 17
Kmeans clustering
mutual information score is 0.03614282127895607
Hierarchical clustering
mutual information score is 0.03783499410950035
for value 18
Kmeans clustering
mutual information score is 0.042211398104366
Hierarchical clustering
mutual information score is 0.04089210303653776
for value 19
Kmeans clustering
mutual information score is 0.04010706145102954
Hierarchical clustering
mutual information score is 0.04195838665664576
for value 20
Kmeans clustering
mutual information score is 0.045353367087101394
Hierarchical clustering
mutual information score is 0.04418078015107005
for value 21
Kmeans clustering
mutual information score is 0.04611080491240086
Hierarchical clustering
mutual information score is 0.044836382368680525
for value 22
Kmeans clustering

mutual information score is 0.048049187287213274
Hierarchical clustering
mutual information score is 0.0458103697241725
for value 23
Kmeans clustering
mutual information score is 0.04628137468065605
Hierarchical clustering
mutual information score is 0.049357946909057905
for value 24
Kmeans clustering
mutual information score is 0.047341572238489116
Hierarchical clustering
mutual information score is 0.05055999528639307
for value 25
Kmeans clustering
mutual information score is 0.05306622575472888
Hierarchical clustering
mutual information score is 0.05174212150267318

III Discussion

- We observe that Naïve Bayes algorithm perform better in this training dataset. Also, with feature selection 1000, the computed result is better.
- The Metric "F1" has better results.
- Linear kernel for SVM classifier works best.
- Also, Chi2 based feature selection gives better features.
- K-Means clustering algorithm gives good score than agglomerative algorithm.

Feature Selection

Chi score for all features: 0.7885950291238968

Better Chi score for selected features of 1000: 0.7976541092391952

Silhouette score

Sihouette score stabilizes when number of clusters increases beyond 15 for KMeans clustering. Whereas for agglomerative clustering, Silhouette score drops drastically for increase in cluster beyond 15.

Mutual Information Score

Mutual Information score increases gradually for increase in number of clusters. Mutual information is similar for both Kmeans and agglomerative clustering.