Document Classification and Clustering

Team Name : Artificially Intelligent
Team Members :

RollNo	Name
201405610	VIKNESHWAR E
201405635	Krishna Chouhan
201405584	Tushar Kadu

Mentor : Shailesh Hiralal Jain

Abstract :

The primary focus of this paper is topic detection, i.e. assigning the documents in a collection C to “topics”.In its simplest form the goal of topic tracking is to assign a new document to one of the topics detected in C, or to decide that it is about a new topic not represented in C.

In Statistics terminology, topic detection is a clustering problem: we want to partition C into groups such that documents in each group are similar to each other, and dissimilar from documents in other groups.

In its simplest form, topic tracking is a classification problem. We have a collection C of documents, each labeled with a topic, and we want to assign a label to a new document. The unusual aspect of the problem is that our answer could be “none”, in which case the document is taken to represent a new topic.

Clustering and classification methods play a central role in the reduction of both the number of operations needed for document classification, and the retrieval time. Also, they can be designed to make accurate decisions on whether or not a document represents a new topic.

Dataset :

We have used 20newsgroups dataset for this project which can be found here.
http://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

Implementation :

1) Document Classification :

The documents are preprocessed using Python NLTK . During preprocessing , stopwords are removed from the documents and words are converted into lowercase and stemmed. Now documents are represented in form of vectors using Bag of Words approach . Then these words are converted into Tfidf matrix to represent the similarity of words between documents. Hence Tfidf values for each word in all the documents are computed. These steps are same for both training documents and test documents. The training corpus along with their class labels are passed to the classification algorithm. Once the classifier is trained, the testing corpus is applied to the classifier to predict the class labels. 5 Fold cross validation is used to test the classifier on the training corpus and then the classifier is

applied on Test Corpus finally. Classification Algorithms that are used here are as follows :

1) Multinomial Naive Bayes

2) Support Vector Machines

3) Random Forest Classifier

4) Maximum Entropy Classifier

5)K Nearest Neighbors

6)Decision Tree Classifier.

2) Document Clustering :

Documents are converted into vectors using Bag of Words approach as described above and Tfidf matrix is computed. These vectors are converted into lower dimensional space using LSA and then the corpus is passed to clustering algorithm (Kmeans).

The number p of terms occuring in a document collection can easily be in the thousands and may be larger than the number n of documents. Representing each document by a p-dimensional vector of (tranformed and weighted) term frequencies has at least two disadvantages. First, it is costly. Storing a document vector requires space proportional to the number of terms occuring in the document. Finding the distance between two document vectors requires work that is proportional to the number of terms occuring in the two documents. This assumes that sparse matrix techniques are used. Representing documents by vectors of dimensionality lower than the average number of terms in a document results in savings of space as well as time.

Once the clustering is done, validation measures such as Homogeneity,Completeness,V Measure,Adjusted Rand Index and Silhouette Coefficient are determined.

Results :

1) Document Classification :

Accuracy Table :

Classification Algorithm	Cross Validation	Final Testing
Multinomial Naive Bayes	91.27	83.32
Support Vector Machines	91.5	82.79
Maximum Entropy Classifier	89.81	82.34
Random Forest Classifier	86.14	77.82
K Nearest Neighbors	80.78	74.18
Decision Tree Classifier	64.76	56.49

As observed, Multinomial Naive Bayes and Support Vector Machines are observed to be the best classification algorithms for this dataset. Maximum Entropy Classifier gives 80+% accuracy as well.

2) Document Clustering :

K Means algorithm is applied on this dataset and the following measures are determined.
Homogeneity: 0.425
Completeness: 0.434
V-measure: 0.429
Adjusted Rand-Index: 0.283
Silhouette Coefficient: 0.223 .

Homogeneity , Completeness and V-measure are bounded between 0 and 1. 0 indicates bad clustering and 1 indicates perfect. Adjusted Rand Index and Silhouette Coefficient are bounded between -1 and 1. -1 indicates incorrect clustering and 1 indicates perfect clustering. Positive values indicate that the clustering is good. The drawbacks of the above measure is that the labels have to be known to determine them.