**Week 1: Classification with KNN**

**Data Used: Reuters-21578 document**

**Text Classification implementing knn from scratch**

The problem consisted of sets of documents to be categorized according to its corresponding features.

APPROACH:

Most common approach is called as bag of words approach to solve this problem. In this problem , both datasets consisted sets of documents.

* + - * **Text Processing:**

Here, " PorterStemmer" from nltk package is used to carry out stemming process on the document . Also, " stopwords" from nltk package is used to eliminate the stop words.

* + - * **Document Representation :**

The training dataset consists of words hence to process the data tf-idf vectorizer is used/

In this tf calculates the frequency of a term in a document while idf measures how important that term is .

* vec\_tfidf = TfidfVectorizer(tokenizer=Stemmer ,sublinear\_tf = True, min\_df = 0.005, stop\_words = 'english' )

Here , stemmer function is called in the tfidf vectorizer which consists the Porter Stemmer and stopwards function. Making "sublinear\_tf "True , it helps the accuracy metric by not increasing significanceof a term only because it has maximum frequency.

While mindf ignores terms that have document frequency strictly less than the threshold and gives maximum accuracy.

* + - * **Implementing the Classifier**

1. Classifier consists of fit and predict method.
2. Fit()method accepts the training set and labels ,while outputs a learned model
3. Predict() method takes the unobserved test set and K as input and performs cosine similarity in order to get top k rows of the top ranked distance values. After that we select most frequent class among these rows and return the predicted index values that resemble the predicted labels
4. Here, cosine similarity is used as a distance metrics as it is gives maximum accuracy in document classifications.
5. Through cross-validation after splitting training set , k = 11 proves to be the best k-value to consider.