Assignment 4

Information Retrieval

- Mayank Chauhan, MT18008

Methodology

Preprocessing

- Collect the documents.
- Split the data randomly maintaining the ratio for each class.
- Apply tokenization - remove punctuation, accent marks and other non alphanumeric characters like (, >, $, \, \\n, /t, _ etc.
- Apply Normalization and Lemmatization.
- Now, we have the final list of tokens for our index.

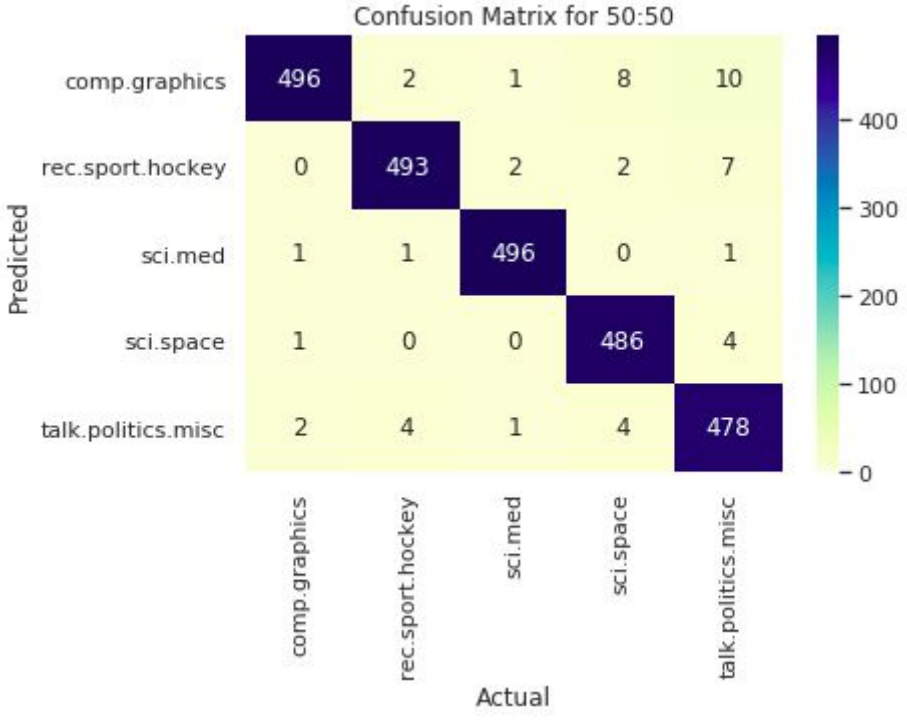
Building Inverted Index

- Create two separate Inverted indexes.
- For second question , the tf-idf is used for finding the top features.

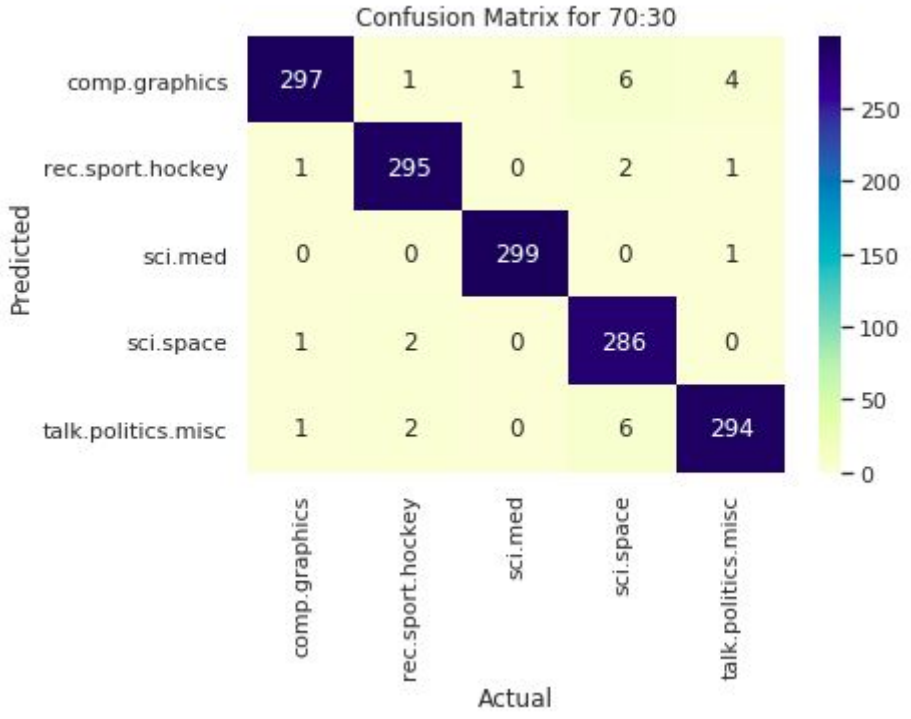
Classifier : Multinomial Naive Bayes

Analysis

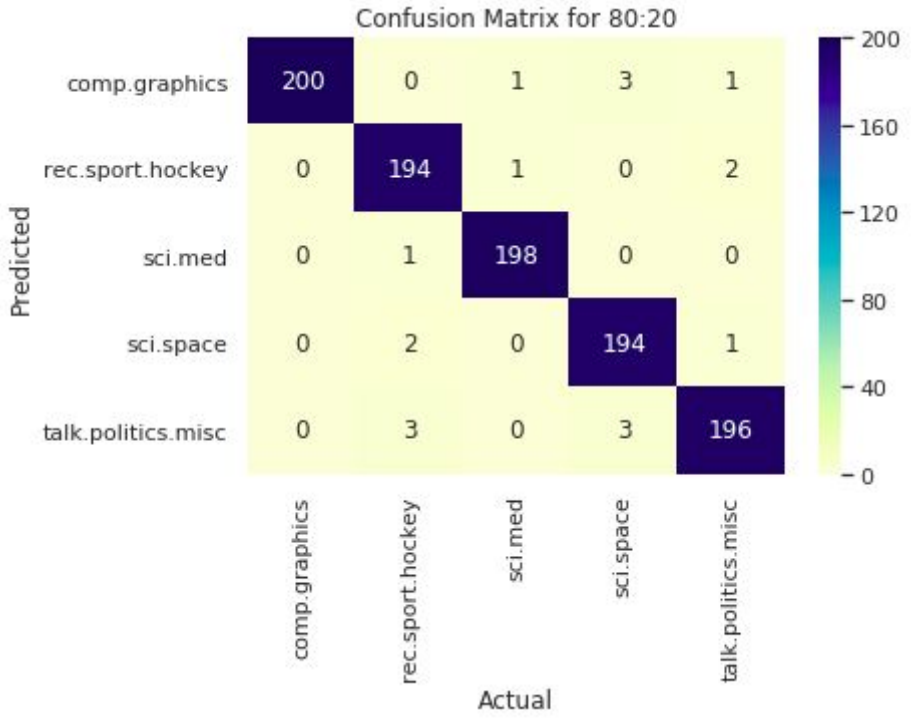
Accuracy by using 50% data for training is 97.96 %.



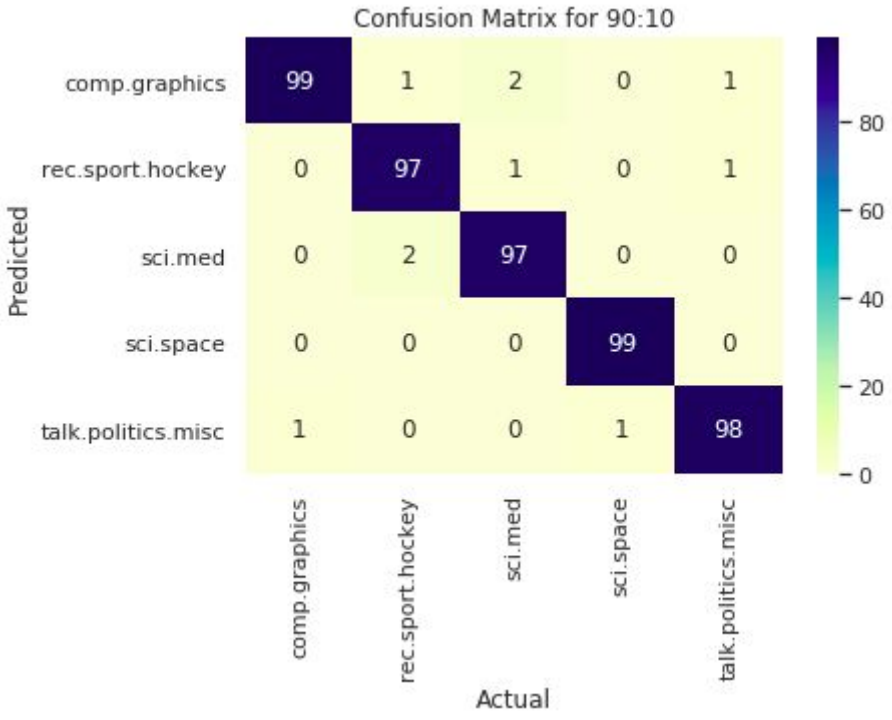
Accuracy by using 70% data for training is 98.06%.



Accuracy by using 80% data for training is 98.2 %.



Accuracy by using 90% data for training is 98.0 %.



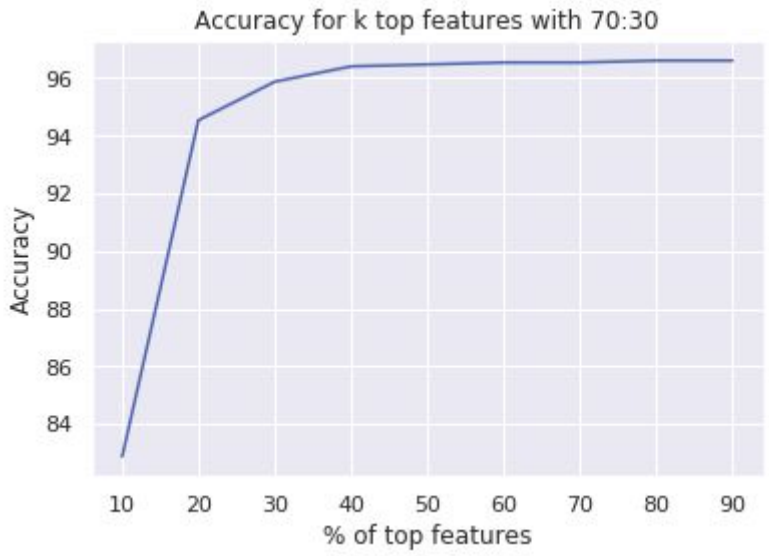
The accuracy on the test data increased when we increased training data from 50% to 70%. But after that, adding any new data in the training process is not helping us, in getting higher accuracy. The new features are not adding any diversity in the features set.

total features: 56777

Using Tf-idf for Feature Selection

% Feature Used	Accuracy
10	82.86%
20	94.53%
30	95.86%
40	96.39%
50	96.6%

Accuracy Vs % Features Used



We are getting 98.06% in the question 1, for 70:30 split, by feature selection the accuracy is not increasing. We can get 96% accuracy by training on only half of the features.
Therefore, reducing the time complexity. Hence, the accuracy is not decreasing significantly, so the purpose of feature reduction is fulfilled.