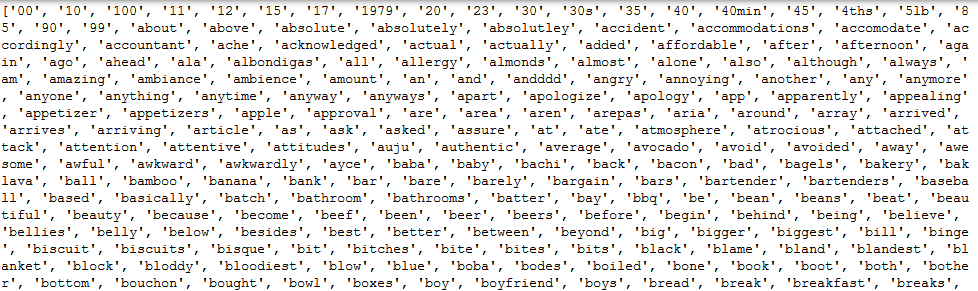
**OUTPUT SUMMARY**

The dataset used for text classification is *yelp\_labelled.txt* from *Sentiment Labelled Sentences Data Set.*

Following is the ouput summary:

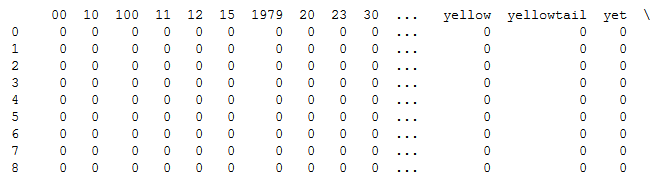
Step 1: Tokenization

Listing of the vocabulary from the training data. A portion of it is shown below

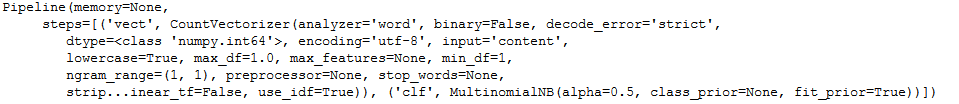


Step 2: Document-term matrix

Vocabulary and its occurrences per document is shown



Step 3: Fit the model using training data

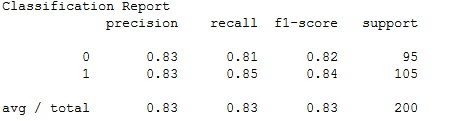


Step 4: Evaluate using test set

1. Accuracy



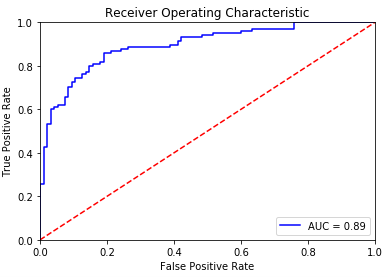
1. Classification Report



1. Confusion Matrix



1. ROC curve



**Observation:**

Best Parameter found is alpha=0.5 as it gives accuracy more than 80% most of the times. The default value of 1.0 gives accuracy below 80% at times.

i.e., on applying a Laplacian Smoothing of alpha=0.5, the model is found to perform better on the given dataset most of the times.

Area Under the curve = 0.89