**Document Classification for Presidential Election Speech Recognition**

For this project we are sing NLTK package of python to create a classification program. This program reads 2 files each for Hillary and Trump. These file contain the different speeches for individual candidates and use this data as training set. The each speech is tokenized and prune using NLTK. The filtered list of words is the tagged with respective candidate.

This list of tagged words is used to create the feature set using condition of a particular word “Contains” in the Top 2000 repeating (with most frequency) of complete word list.

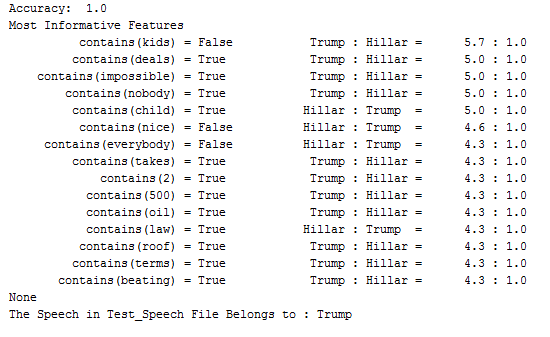
We are using the Naïve Bayes Classification to create the classifier.

As an output we are showing output of The Accuracy of the Classifier and the most informative Features and the output of the Test Speech classification.

Note: We have included the different test links for both candidate which you can copy to test\_speech file.

**Output:**

**Trump Test Speech:**



**Hillary Test Speech:**

