Natural Language Processing

This project is developed in order to build a basic insight on NLP using Spam Message Filter Dataset.

Following steps are involved in order to build a model regarding NLP:

1. Reading Dataset and performing basic operations:
   First we read dataset using pandas library. After that we will check for presence of null values and noise in label column.

2. Pre-processing Dataset:
   Next step required is the pre-processing of dataset:
   1. Remove punctuation: Punctuations aren't helpful in developing any kind of information regarding dataset, so we remove them.
   2. Tokenization: Words present in the text is then changed into tokens that is list of words.
   3. Remove stop words: Stop words are those words which act as connectives e.g. for, is, to, me etc. Such words are of no use in order to have information regarding the text and may only act as noises.
   4. Stemming: Stemming is the process of converting word to its root. It removes suffices like 'ing', 'ly', 's' e.g. Entitling -> Entitled -> Entitle.
   5. Lemmatizing: Lemmatizing derives the canonical form ('lemma') of a word. i.e. the root form. It is better than stemming as it uses a dictionary-based approach i.e. a morphological analysis to the root word.eg: Entitling, Entitled->Entitle.

3. Saving cleaned dataset:
   Dataset is saved after its cleaning so it can be used further accordingly.

4. Vectorizing Data:
   (1) Apply CountVectorizer.
   (2) Apply N-Grams.
   (3) Apply TF-IDF.

5. Feature Engineering:
   (1) Length of text message.
   (2) % of punctuation in text message.

6. Building ML Classifier:
   (1) Exploring Parameters using Grid Search.
   (2) Random Forest Classifier.

7. Final Evaluation of Model:
   (1) Accuracy Score of model is calculated.
   (2) Confusion matrix is used to evaluate the model.