**NLP project Round1 report**

Submitted by Members of Team Language Revolution

Jatin Dahiya 19ucs033

Rishabh Sahu 19ucs129

Naman Dhanotia 19ucs128

**Problem Description**

To take two books from <http://gutenburg.org> in .txt format and perform the following Natural Language processing operations on them

* Apply Data preprocessing on the text
* Generating frequency distributions of the words
* Creating word clouds from the text before and after removing stopwords
* Evaluating relationship between word length and frequency
* Parts of Speech tagging for the words in the text

**Data Preprocessing steps**

We performed the following data preprocessing steps

Removing chapter number and chapter Headings

Removing all punctuation marks

**Data Preparation**

For preparing data to be analysed we performed the following steps

1. Tokenising the text into a list of words

**Problem Statement**

**Illustrations (Plots ,tables and figures and output)**

**Inferences**

The word clouds before and after removing stopwords are quite different due to the high frequency of many of these stopwords. One of the reasons may be that stopwords can be used in a variety of contexts whereas nouns and verbs are more restricted to the situations to which they relate to.

Words having length between 3 to 5 are the most frequently occurring words in these books . After that words with larger lengths (upto a certain length)are frequent followed by words of length 1 to2.

We are able to apply part of speech tagging to the words in these books using the penn treebank tagset.