**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

**Link to github code repository**

[Rishabhsahu325/NLP\_Project\_Round1: NLP project (github.com)](https://github.com/Rishabhsahu325/NLP_Project_Round1)

**Python Libraries/Modules used**

Matplotlib

Python re library (regular expressions library )

Numpy

Nltk

Math

WordCloud

**Books chosen for applying the processing**

**T1:** The Adventures of Gil Blas of Santillane, Volume I (of 3)

**T2:** The Adventures of Gil Blas of Santillane, Volume II (of 3)

**Data Preprocessing and Preparation steps**

We performed the following data preprocessing steps

1. Removing chapter number and chapter Headings
2. Removing all punctuation marks
3. Changing all text to lowercase
4. Converting short forms like can’t to actual representations
5. Tokenising the text into a list of words
6. Removing chapter headings and unrelated data
7. Removing hyperlinks

**Illustration : Word length – frequency plots**

|  |  |
| --- | --- |
| T1 | T2 |
| Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |

**Inferences from word length- frequency plot**

For both the books 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. Very long words appear very rarely .Overall implying that most of the words lie in the length range of 3 to 5.

**Illustrations (Wordclouds and wordwise frequency plots)**

T1

|  |  |
| --- | --- |
| Before | After |
| Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |
|  | A picture containing text  Description automatically generated |

T2

|  |  |
| --- | --- |
| Before | After |
| Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |
| A picture containing text  Description automatically generated |  |

**Inference from word Clouds**

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.

**Sample results from pos\_tagging**

[('project', 'NN'), ('gutenberg', 'NN'), ('ebook', 'NN'), ('adventures', 'VBZ'),

('gil', 'JJ'), ('blas', 'NNS'), ('santillane', 'NN'), ('alain', 'VBP'), ('ren', 'JJ'), ('le', 'JJ'),

('sage', 'NN'), ('adventures', 'NNS'), ('gil', 'VBP'), ('blas', 'NNS'), ('santillane', 'RB'),

('translated', 'VBD'), ('french', 'JJ'), ('tobias', 'NN'), ('smollett', 'NN'), ('preceded', 'VBD'), ('biographical', 'JJ'), ('critical', 'JJ'), ('notice', 'NN'),

('le', 'NN'), ('sage', 'NN'), ('george', 'JJ'), ('saintsbury', 'NN'), ('twelve', 'VBP'),

('original', 'JJ'), ('etchings', 'NNS'), ('r', 'NN'), ('de', 'IN'), ('los', 'FW'),

('rios', 'NNS'), ('three', 'CD'), ('volumes', 'NNS'), ('vol', 'VBP'), ('ii', 'JJ'),

('london', 'NN'), ('j', 'NN'), ('c', 'VBP'), ('nimmo', 'RB'), ('bain', 'VBP'),

('14', 'CD'), ('king', 'VBG'), ('william', 'JJ'), ('street', 'NN'), ('strand', 'NN'),

('w', 'WRB'), ('c', 'JJ'), ('new', 'JJ'), ('york', 'NN'), ('scribner', 'NN'), ('welford', 'IN'),

('co', 'NN'), ('1881', 'CD'), ('contents', 'NNS'), ('vol', 'JJ'), ('ii', 'JJ'), ('book', 'NN'),

('fourth', 'JJ'), ('continued', 'VBD'), ('gil', 'JJ'), ('blas', 'NN'), ('leaves', 'VBZ'),

('place', 'NN'), ('goes', 'VBZ'), ('service', 'NN'), ('gonzales', 'NNS'), ('pacheco', 'VBP'),

('marchioness', 'JJ'), ('chaves', 'NNS'), ('character', 'VBP'), ('company', 'NN'),

('incident', 'NN'), ('parted', 'VBD'), ('gil', 'JJ'), ('blas', 'NNS'), ('marchioness', 'NN'),

('chaves', 'NNS'),

**Inferences POS\_tagging**

We are able to apply part of speech tagging to the words in these books using the penn treebank tagset. The pos\_tag(words) functions uses the penn treebank as the default tagset as per official documentation.