**NLP project Round1 report**

Submitted by

Members of Team Language Revolution

Jatin Dahiya 19ucs033

Rishabh Sahu 19ucs129

Naman Dhanotia 19ucs128

# **Link to GitHub code repository**

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

# **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 pre-processing on the text
* Generating frequency distributions of the words
* Creating word clouds from the text before and after removing stop words
* Evaluating relationship between word length and frequency
* Parts of Speech tagging for the words in the text

**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) |

# **Python Libraries/Modules used**

Matplotlib : for drawing plots

Python re library (regular expressions library): For regular expressions

NumPy :for parameters of axes while plotting graphs

Nltk: :Used for tokenizing, removing stop words

Math :For calculating floor and ceil function values while plotting values

WordCloud :For creating word cloud

Collections :For getting the frequency mappings of the POS tags

# **Inferences after examining raw data**

This raw text contains copyright related information, chapter headings, random blank lines and unprocessed text which cannot be directly processed.

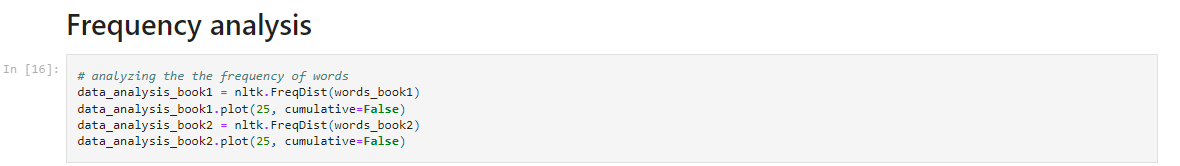
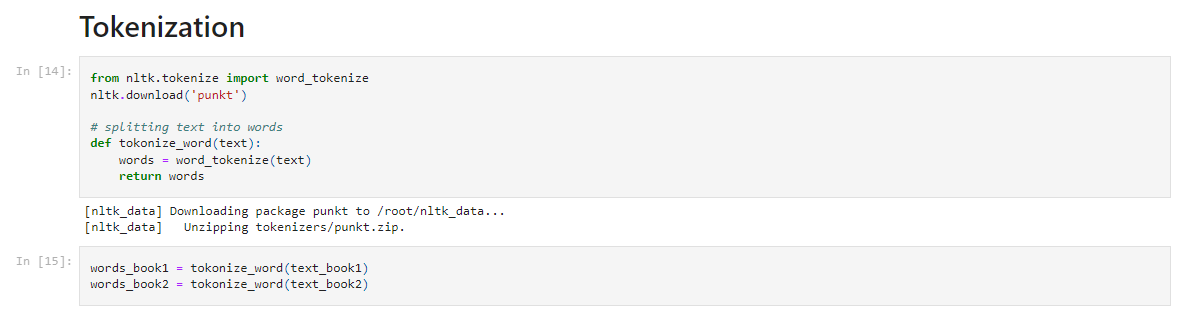
# **Data Pre-processing and Preparation steps**

We performed the following data pre-processing 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









Removing stop words



# **Illustrations (Word clouds and word wise 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

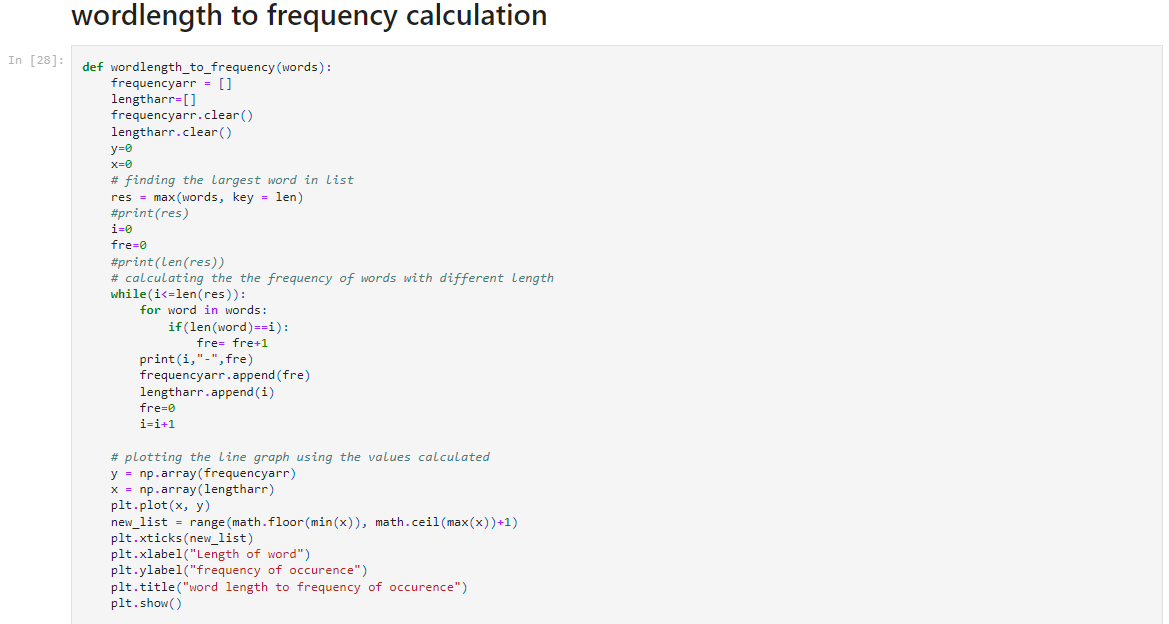
|  |  |
| --- | --- |
| Before | After |
| Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |
| A picture containing text  Description automatically generated |  |

# **Inference from word Clouds**

* The word clouds before and after removing stop words are quite different due to the high frequency of many of these stop words. One of the reasons may be that stop words can be used in a variety of contexts whereas nouns and verbs are more restricted to the situations to which they relate to.
* After removing stop words, we are able to find the set of words which provide us meaning and context about the document.

**Word length – frequency**

Here we are calculating the word length and their frequency of occurrence.



# **Illustration: Word length – frequency plots**

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

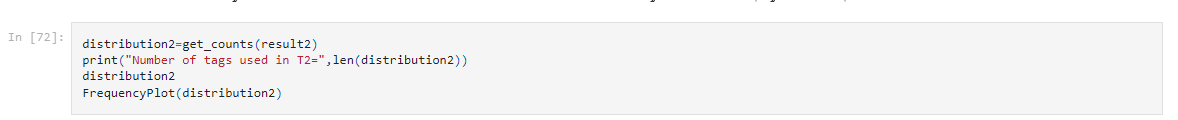
# **Inferences from word length- frequency plot**

* For both the books Words having length between 3 to 5 are the most frequently occurring words in these books. After those words with larger lengths (up to 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.

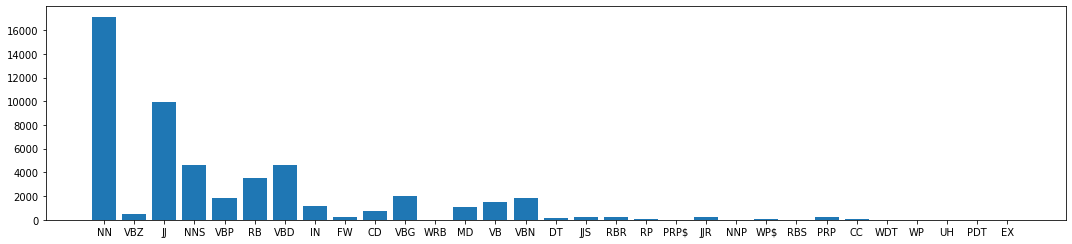
# **POS tagging**

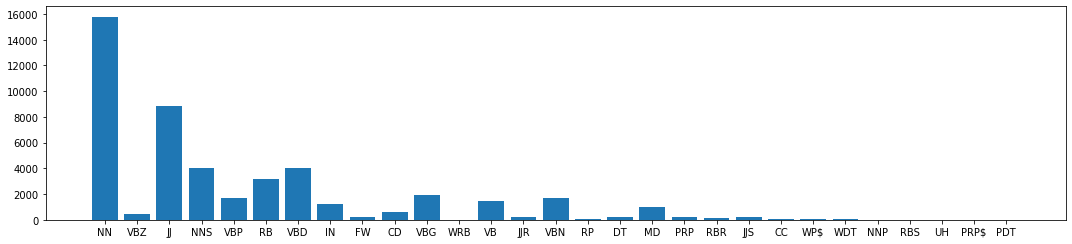
Here we are finding the tag associated with each word that was pre-processed.





# **Illustrations POS tagging**

Book T1****

Book T2 

# **Inferences POS\_tagging**

We applied pos\_tagging on the two books using pos\_tag function. The pos\_tag(words) function uses the Penn treebank as the default tag set as per official documentation.

In T1 the most frequently occurring POS Tag is ‘NN’ with count 17139 followed by ‘JJ’ having count 9972.

In T2 the most frequently occurring POS Tag is ‘NN’ with count 15804 followed by ‘JJ ‘ having count 8871.

**Conclusions**

In this Round 1 of our project, we performed the tasks of word pre-processing, word tokenisation, Word Cloud generation, POS tagging and also deduced many inferences from them about the books while also learning in the process.