**NLP Project Report- Round1**

Submitted by

Team ALP

Hemang Goyal-19ucs193

# **Link to GitHub code repository -**

<https://github.com/Hemang110/NLP_Project>

# **Problem Description -**

We were tasked to take two books from <http://gutenburg.org> in text format and perform Natural language processing techniques on the text acquired.

* Data pre-processing
* Analyze the frequency distribution of tokens in book1 and book2 separately
* Create a Word Cloud of book1 and book2 using the token that you have got
* Evaluating relationship between word length and frequency for book1 and book2
* Parts of Speech tagging for the words in the text

**Books used-**

T1: Animal Life in Field and Garden by Florence Constable Bicknell and Jean-Henri Fabre

Link -<https://www.gutenberg.org/ebooks/66755>

T2: David Vallory by Francis Lynde

Link- <https://www.gutenberg.org/ebooks/66754>

# **Python Libraries/Modules used**

Matplotlib : Creates a figure, Creates a plotting area in a figure

re library (regular expressions library): Provides support for regular expressions

nltk: : Used in tokenizing, removing stop words and pos tagging

Math : For calculating floor and ceil values

WordCloud : For creating word cloud

Collections : For getting the frequency mappings of the POS tags



# **Inference from Raw text-**

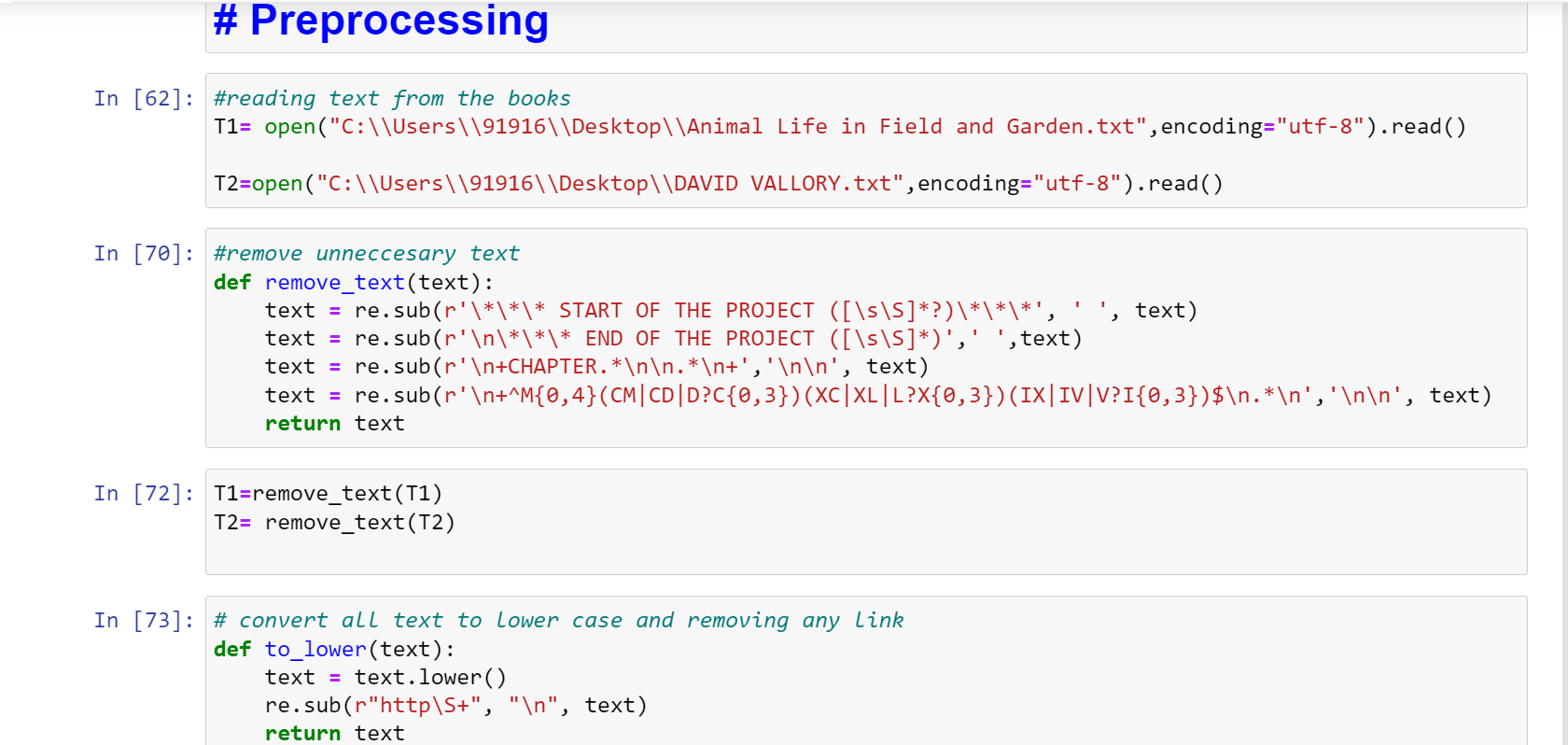
The Raw text contains Chapter number and headings, indexes, unprocessed data which is not fit for processing.

# **Data Pre-processing steps-**

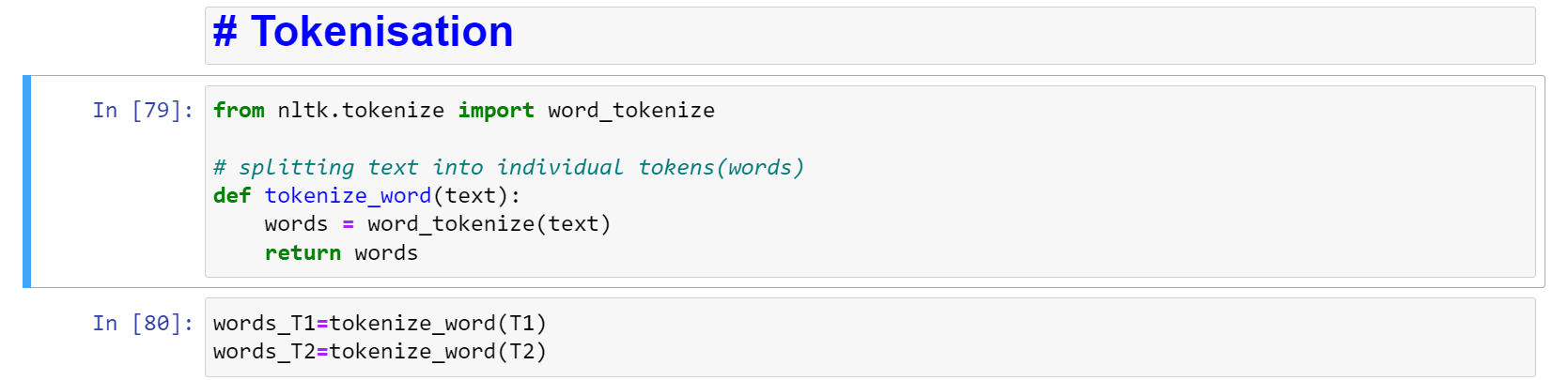
Performed following data pre-processing steps -

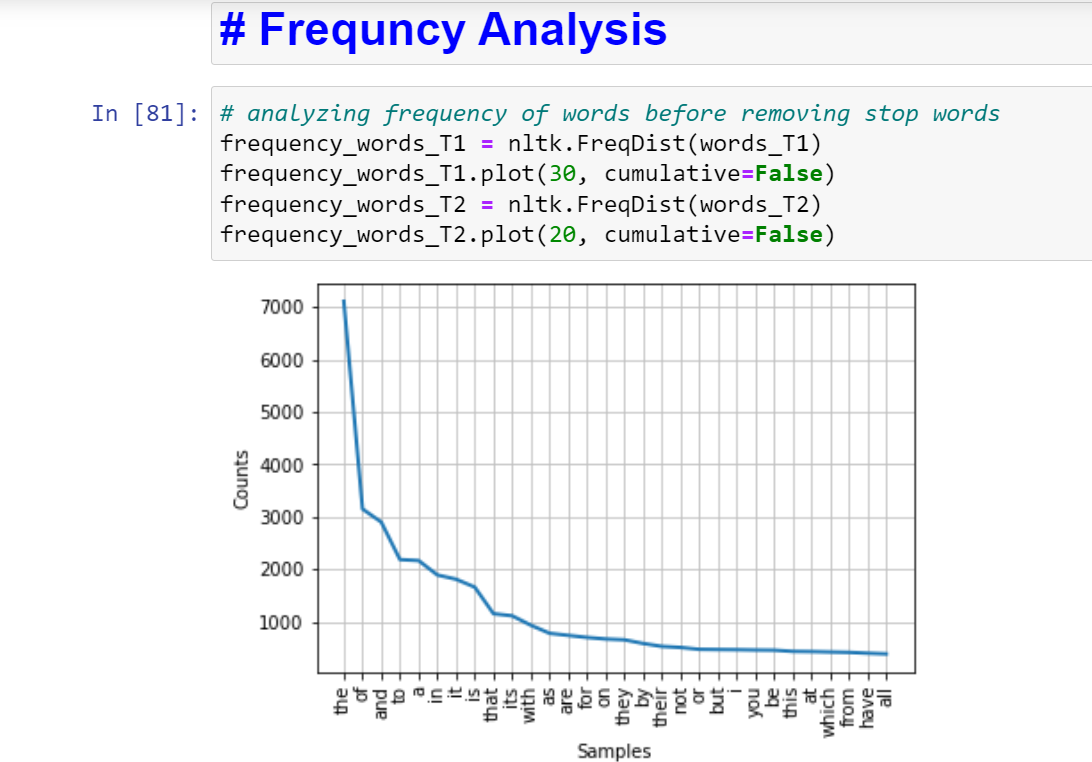
1. Used Regular expressions to remove chapter number and chapter Headings.
2. Removed all punctuation marks using regular expressions.
3. Changed text to lowercase.
4. Conserved the meaning of short forms like can’t, won’t, ‘m, ‘re etc. by converting short forms to actual representations.
5. Tokenized the text into a list of words.
6. Removed unnecessary data.
7. Removed hyperlinks

**Preprocessing snippets-**





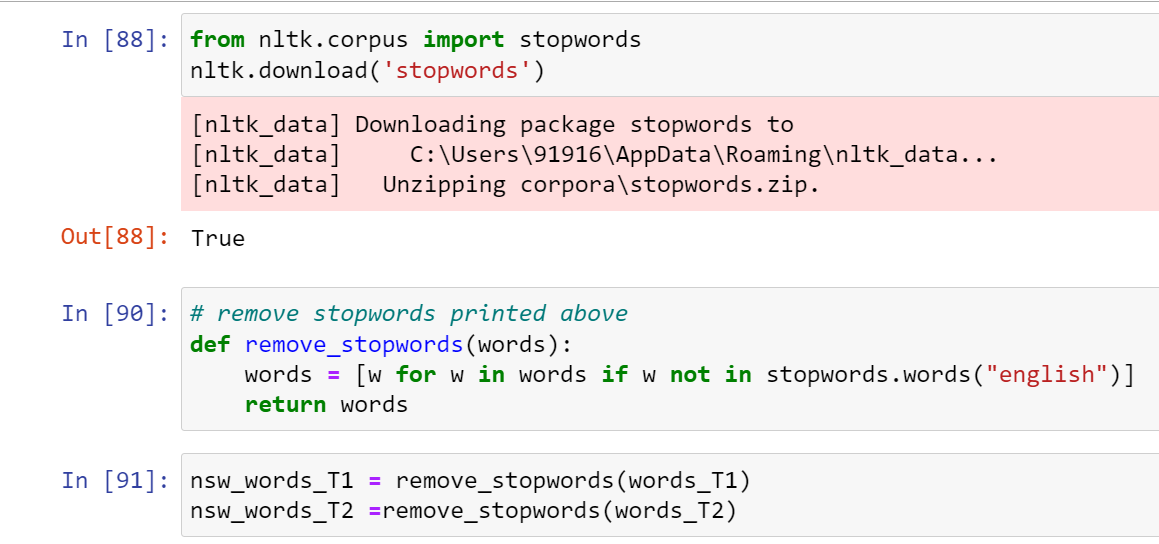
**Tokenization snippets-**

**Frequency Analysis Snippet- **

**Word Cloud Snippets-**

****

**Removing Stopword snippet-**

****

# **Illustrations -**

Plot of frequency analysis of T1 before and after stopword removal

| Before | After |
| --- | --- |
|  |  |

Plot of Word Cloud of T1 before and after stopword removal

| Before | After |
| --- | --- |
|  |  |

Plot of frequency analysis of T2 before and after stopword removal

| Before | After |
| --- | --- |
|  |  |

Plot of Word Cloud of T2 before and after stopword removal

| Before | After |
| --- | --- |
|  |  |

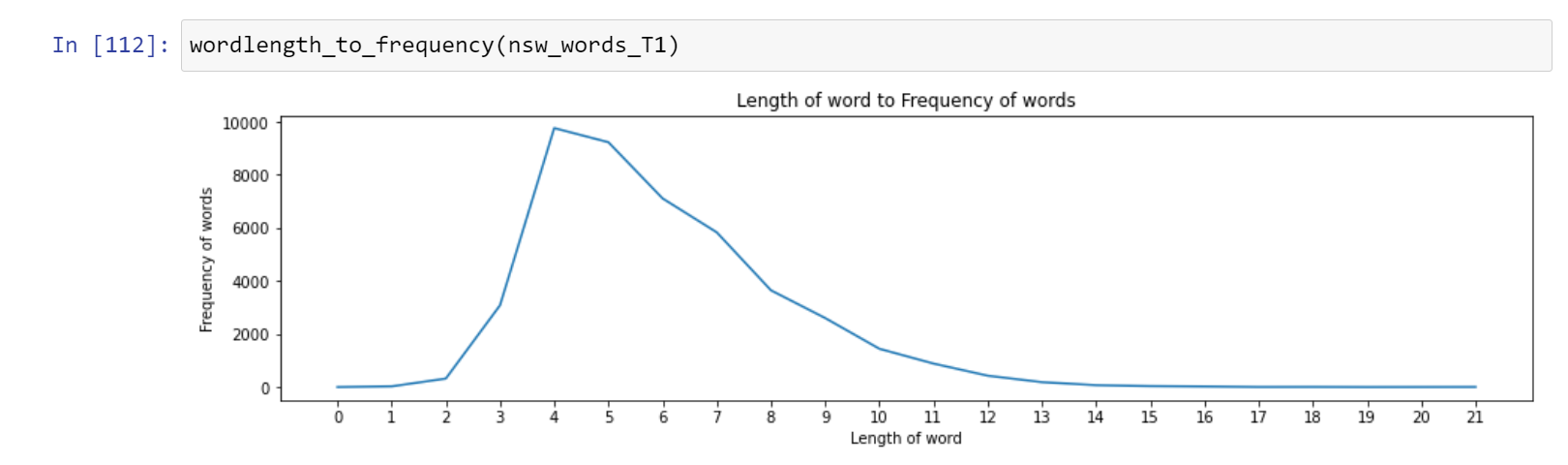
# **Inference from word Clouds and frequency plots**

* After comparing before and after images of both frequency plot and word cloud plot for T1 and T2, we can infer Stop words have a very high frequency of occurrence. the, to, and and were the most frequent and rightly so, as these determiners are very integral to writing grammatical sentences.
* After removing stop words, we are able to find the set of words which provide us meaning and context about the documents.
* Ex-For T1(Animal life) we can see words like insects,wings, beak which clearly provide us meaning and context about the documents.

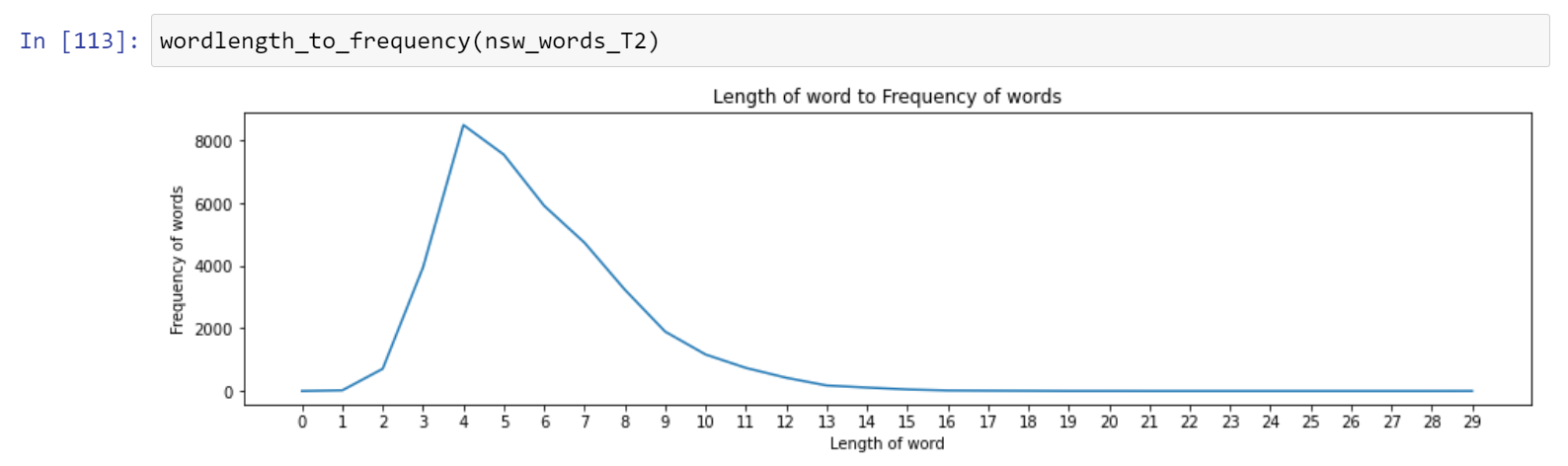
**Word length to frequency plots**

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

**For T1-**

****

**For T2-**

****

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

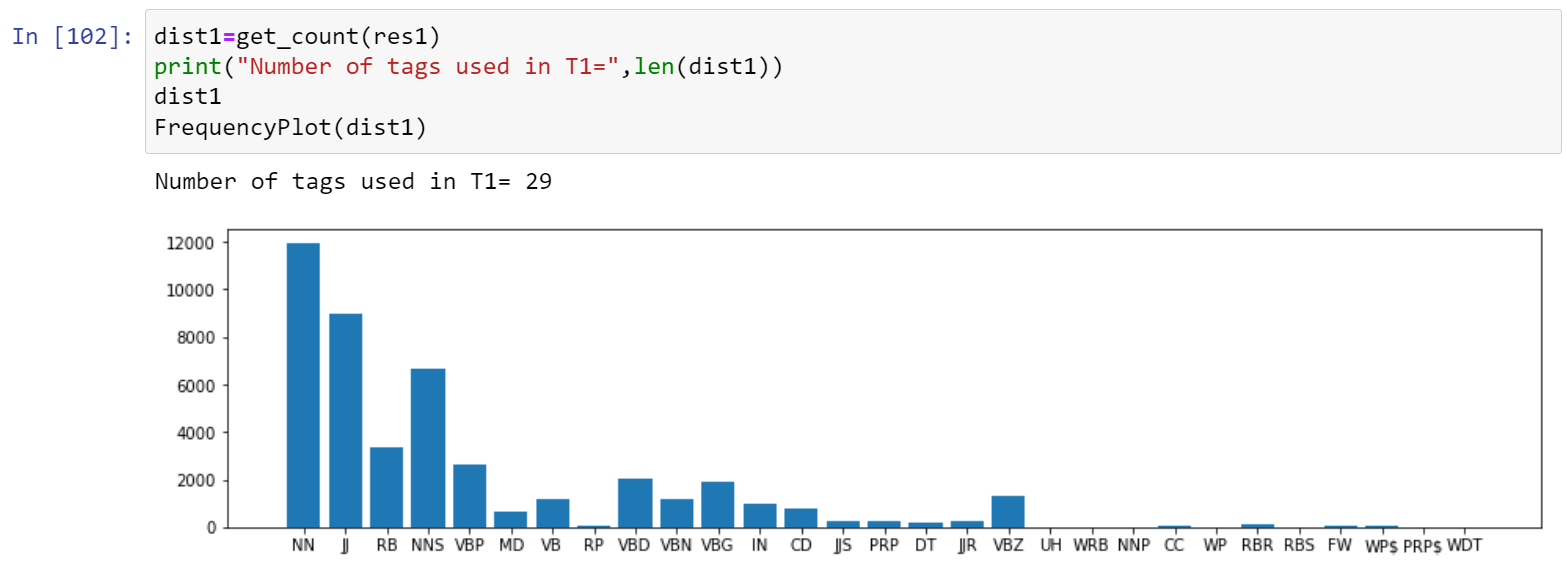
* For both the books T1 and T2 words with length between 4 to 6 have higher frequency of occurrence than other words. Very long and very small words appear very rarely. Optimal range of word length is between 4 to 6.

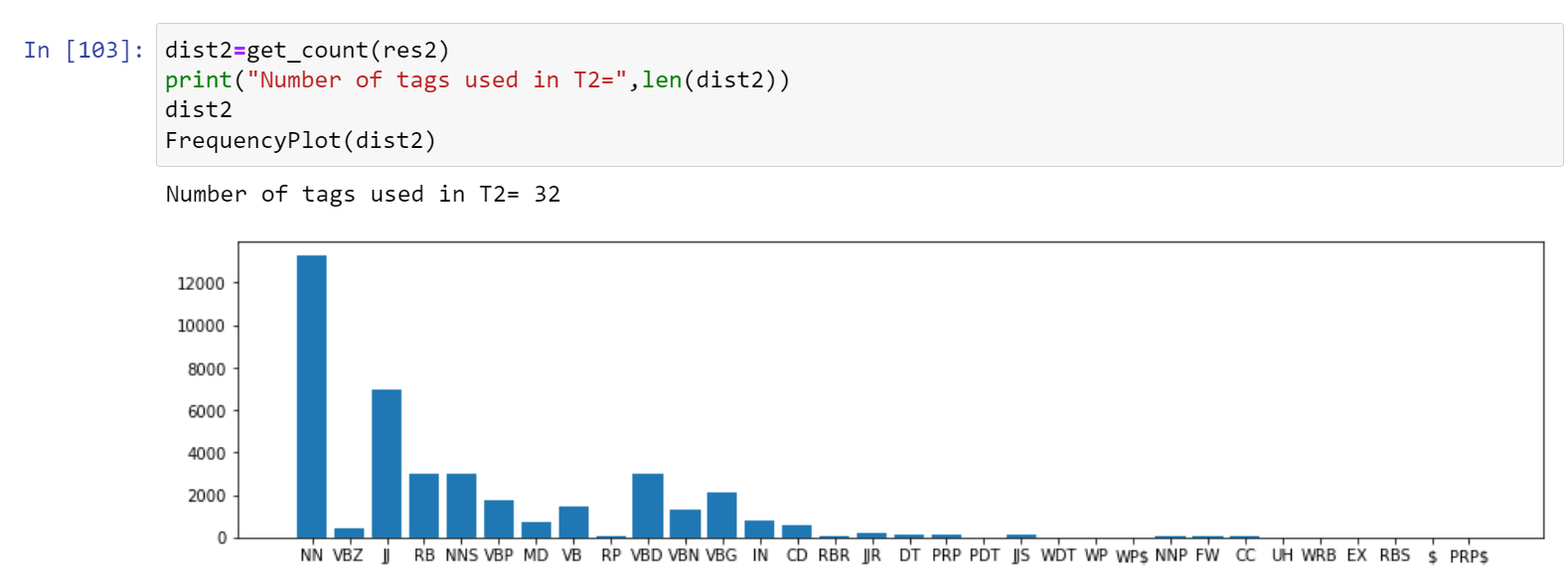
**POS tagging -**

Finding the tag associated with each word that was pre-processed.

# **Illustrations for POS tagging**

# 

For T1****

For T2 

# **Inferences POS\_tagging**

Applied pos\_tagging on T1 and T2.

Here are some inferences-

For T1, the most frequently occurring POS Tag is NN followed by JJ which is followed by NNS.

For T2, the most frequently occurring POS Tag is NN followed by JJ.

**Conclusion**

Performed some integral parts of nlp in form of tasks like word pre-processing, word tokenization, Word Cloud generation, POS tagging. Learned the requirement of preprocessing. Learned the importance of stopword removal. Learned about various pictorial representations of words, Learned the use of POS tagging.