NLP Project Report
Round-1

Team Name
KAFKAESQUE

TEAM MEMBERS

1. SOHAM BAIJAL 19UCS098
2. SHIVAM SHUKLA 19UCS040
3. KESHAV MUNDRA 19UCS074

GitHub Project Code Link

https://github.com/sbaijal/NLP_Project_Round1.git

Data Description

Two books downloaded from http://www.gutenberg.org

Book 1 (T1) - **Crime and Punishment**
Book 2 (T2) - **War and Peace**

Data Preparation

The data is imported in raw form and prepared by removing the not required headers and chapter names and licence documents at the bottom of the book.

Data Preprocessing Steps

After obtaining only the book text we tokenize to remove '/n' , '/r' and store the words in a list.
We then apply two preprocessing steps to convert our raw data into a meaningful list of tokens to extract information.
Steps of Preprocessing

1) Converting all the tokens in the list to lowercase
2) Applying lemmatization

Note : Substantial changes were seen in output after applying the above two mentioned steps

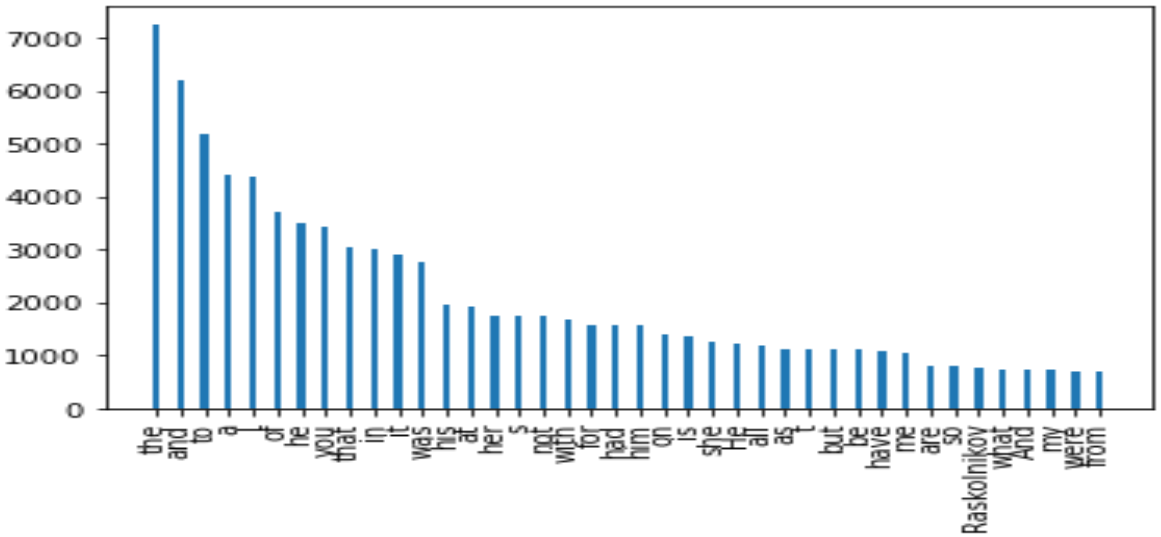
Problem statement

- Import the text, let's call it as T1 and T2
- Analyze the frequency distribution of tokens in T1 and T2 separately
- Create a Word Cloud of T1 and T2 using the token that you have got

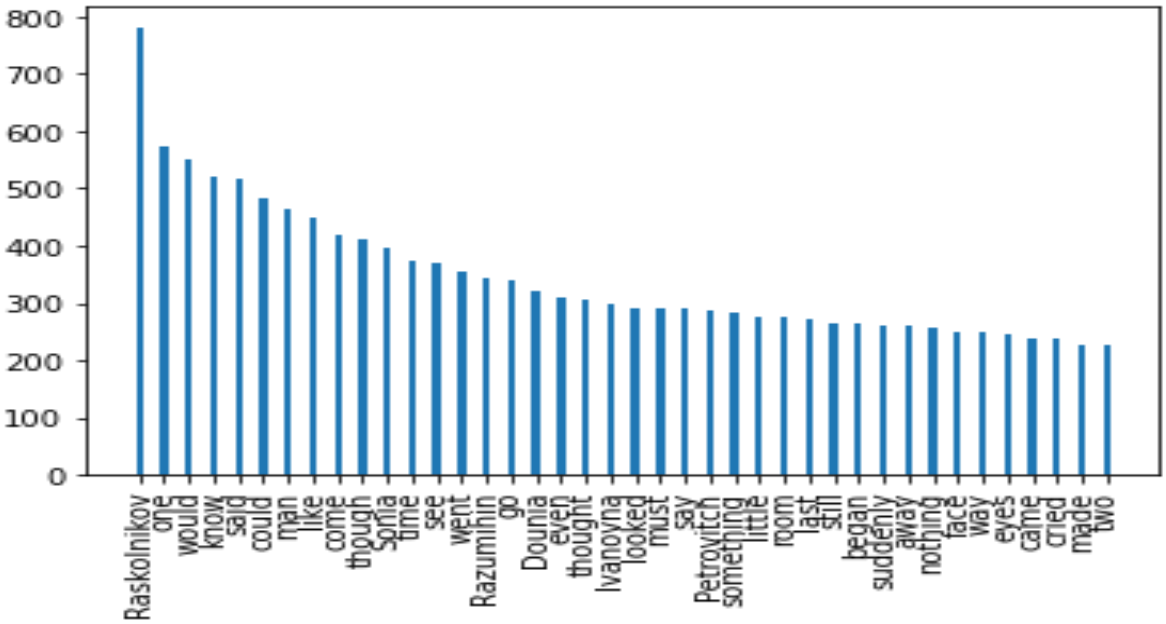
- Remove the stop words from T1 and T2 and then again create a word cloud - what's the difference it gives when you compare with word cloud before the removal of stop words?
- Evaluate the relationship between the word length and frequency for both T1 and T2 — what's your result?
- Do PoS Tagging for both T1 and T2 using anyone of the four tag sets studied in the class and get the distribution of various tags

Plots

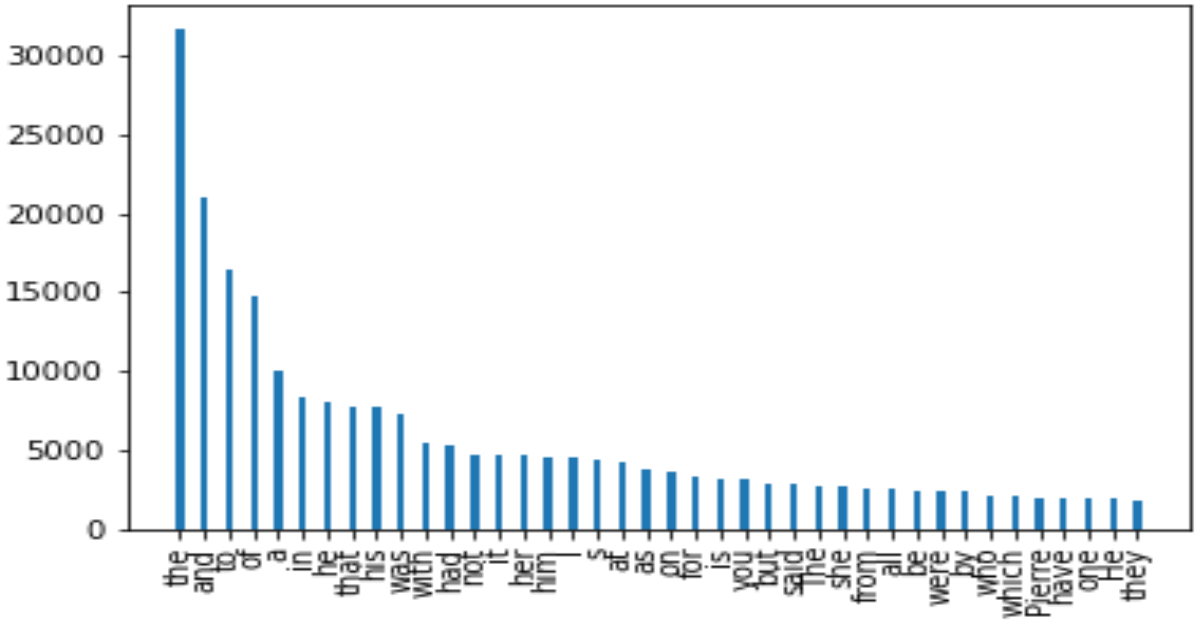
1) T1 Text frequency distribution of top 40 most frequent words including Stop Words



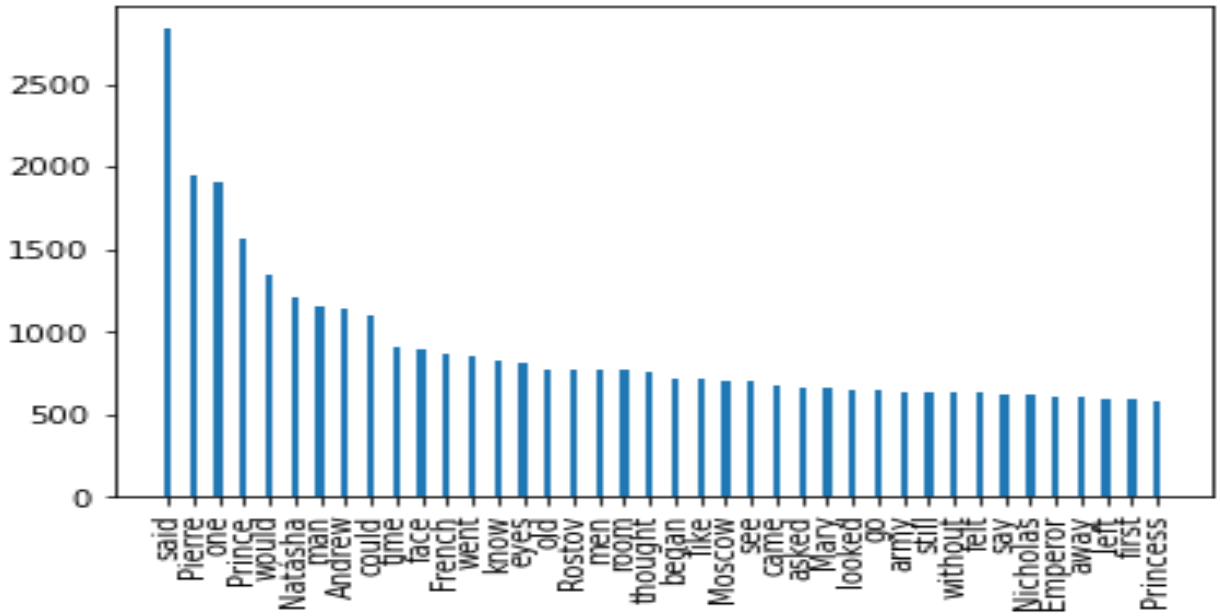
2) T1 Text frequency distribution of top 40 most frequent words excluding Stop Words



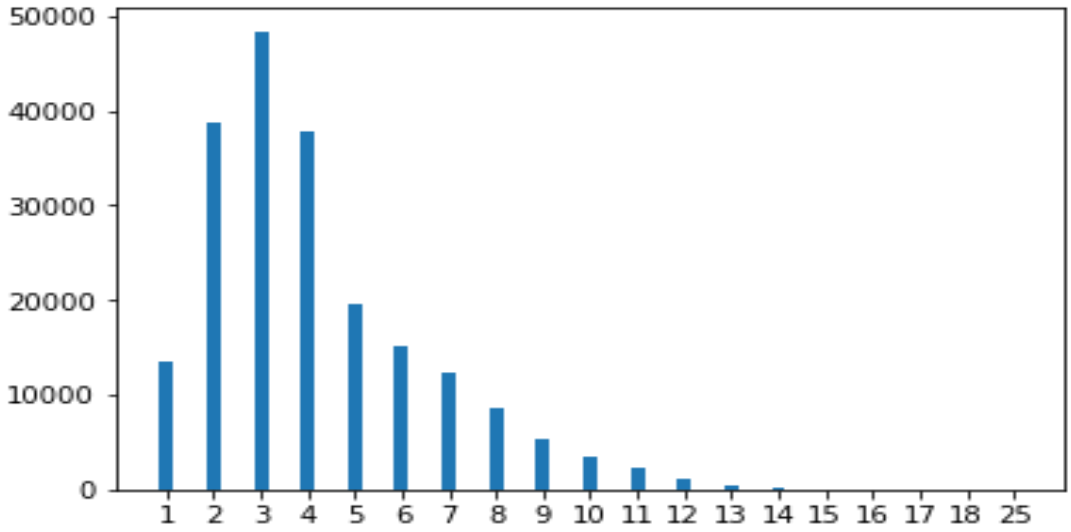
3) T2 Text frequency distribution of top 40 most frequent words including Stop Words



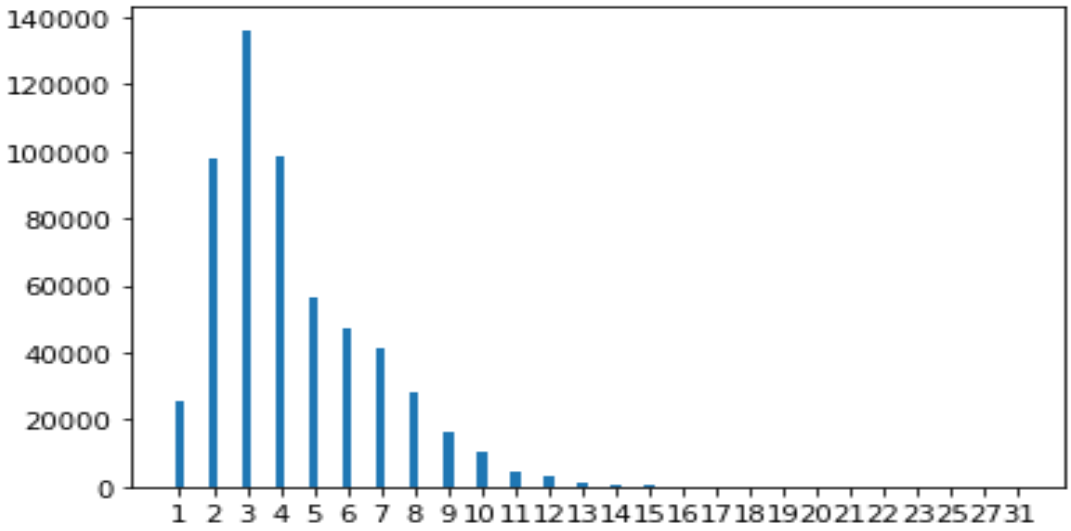
4) T2 Text frequency distribution of top 40 most frequent words excluding Stop Words



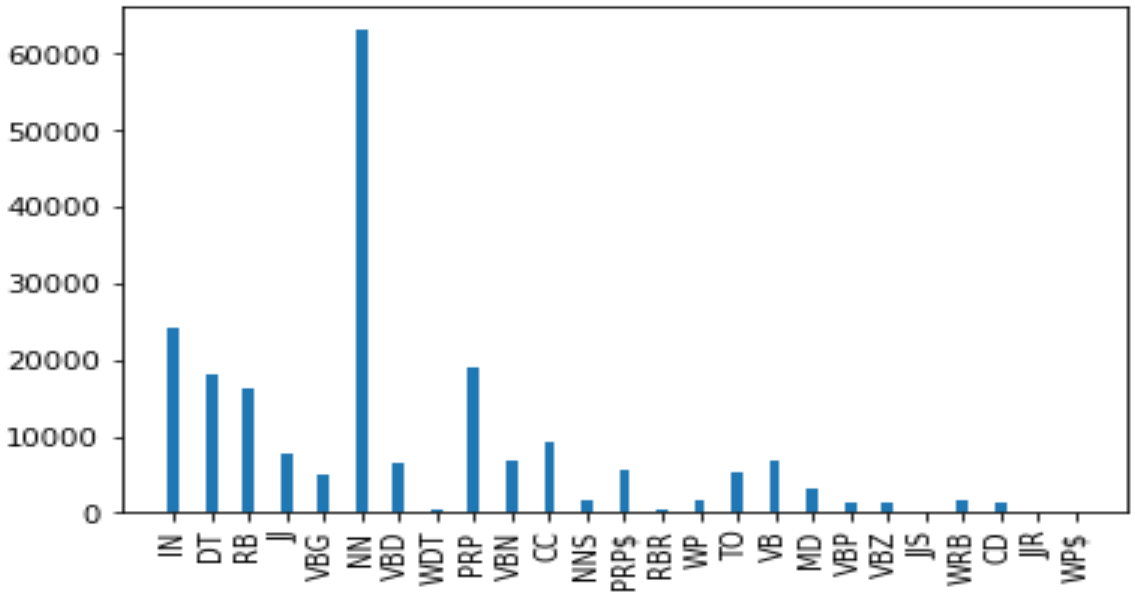
5) T1 Word Length Frequency Distribution Graph



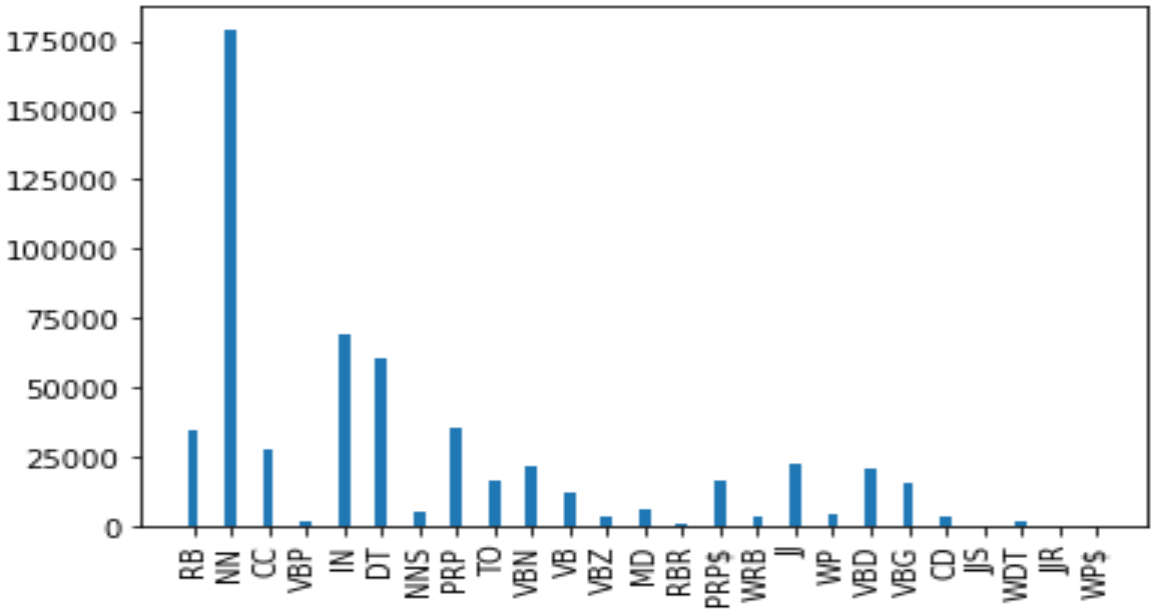
6) T2 Word Length Frequency Distribution Graph



7) T1 POS TAGGING frequency bar plot



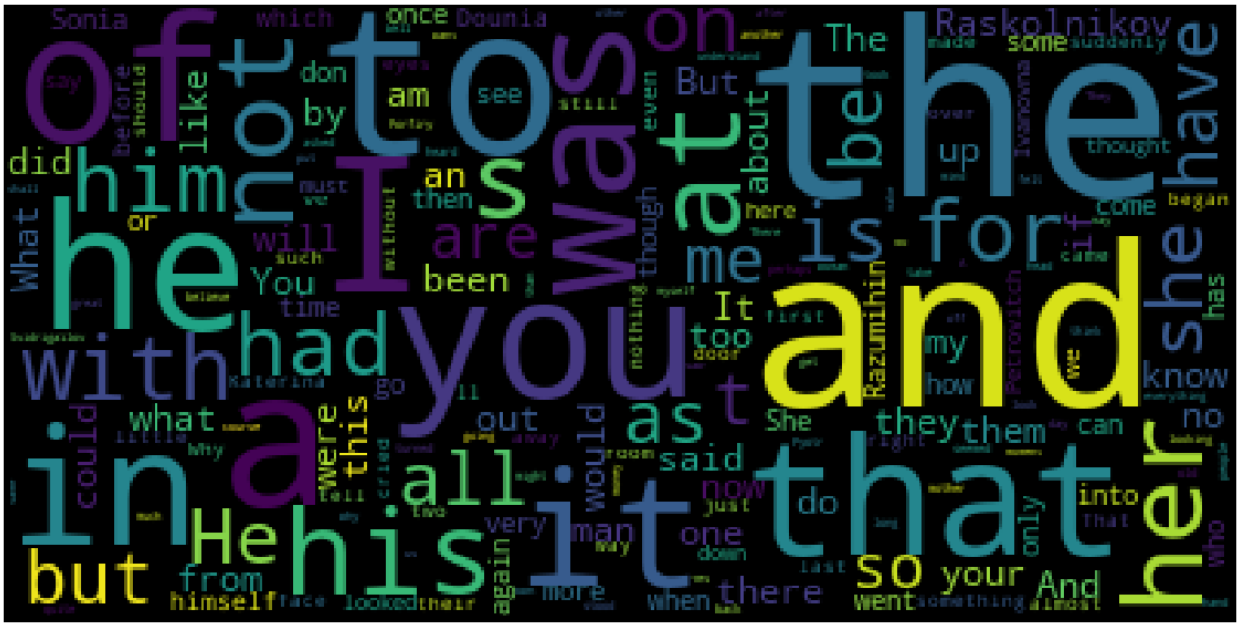
8) T2 POS TAGGING frequency bar plot



Figures

T1 Text Word Clouds

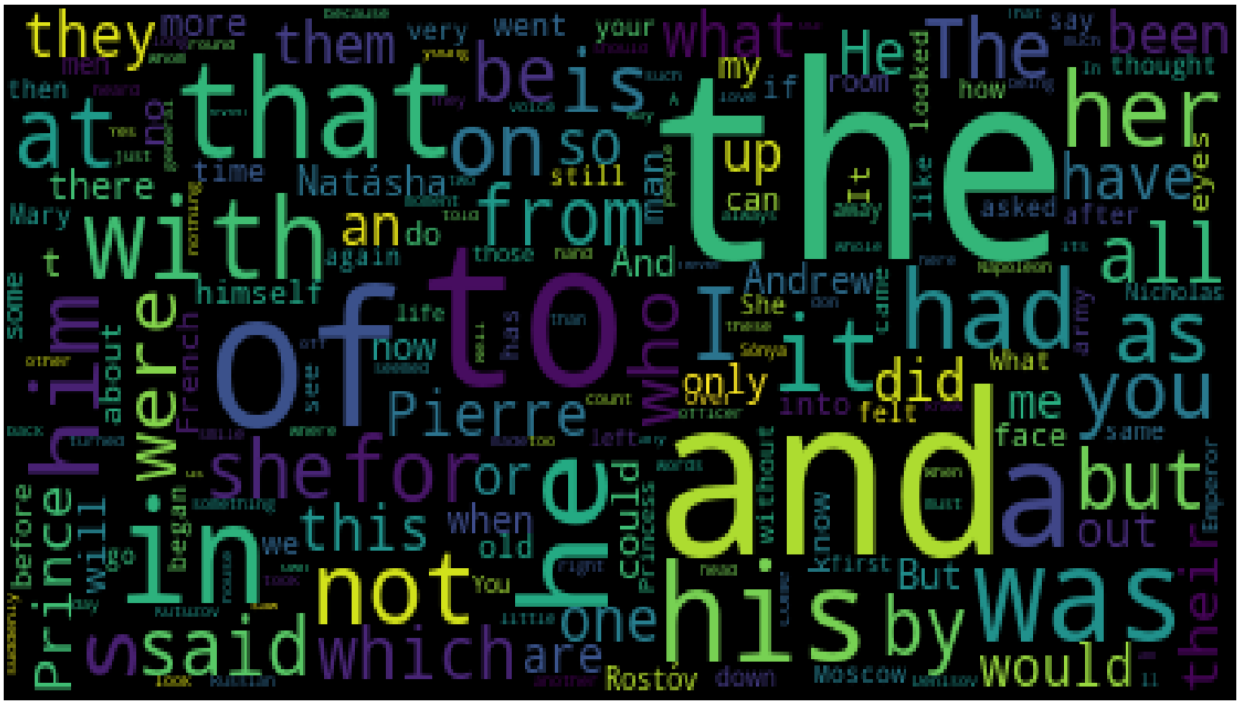
1) Word Cloud with Stop Words included



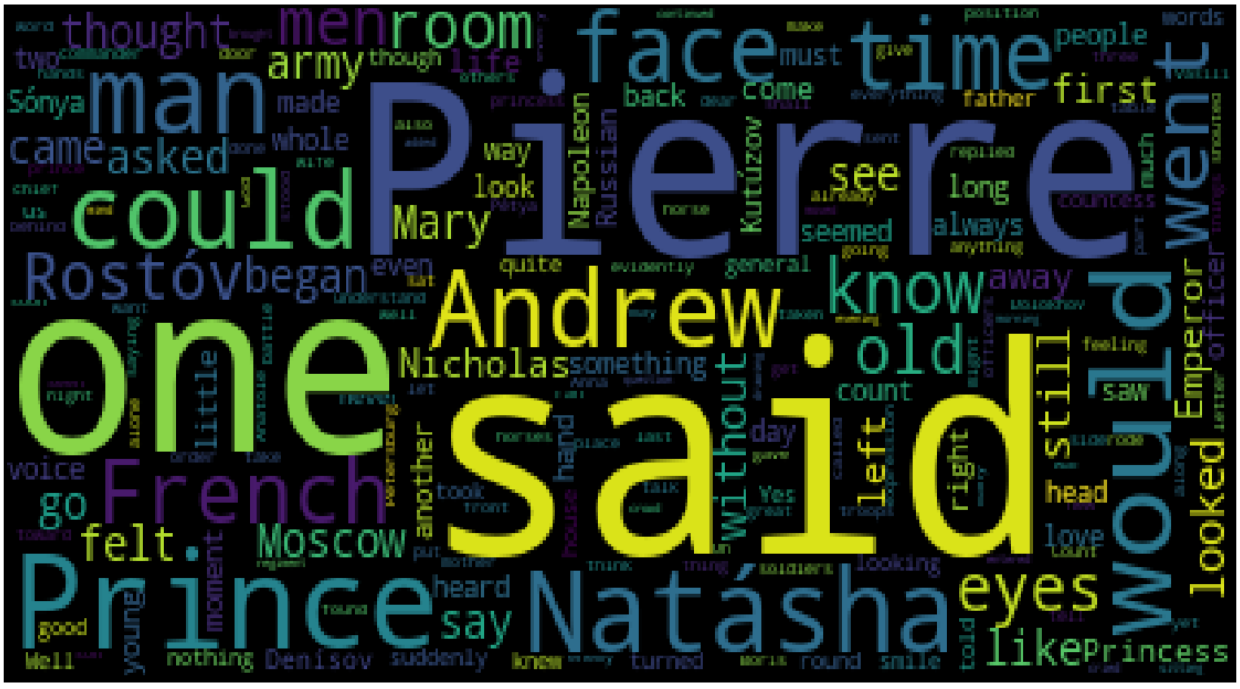
2) Word Cloud excluding Stop Words

T2 Text Word Clouds

1) Word Cloud with Stop Words included



2) Word Cloud excluding Stop Words



Tables

1) T1 Pos Tagging Output Table

TAG	FREQUENCY
IN	24209
DT	18034
RB	16130
JJ	7602
VBG	5026
NN	63091
VBD	6540
WDT	349
PRP	18990
VBN	6765
CC	9178
NNS	1760
PRP$	5732
RBR	563
WP	1702
TO	5239
VB	6822
MD	3163
VBP	1390
VBZ	1468
JJS	279
WRB	1617
CD	1223
JJR	96
WP$	20

2) T2 Pos Tagging Output Table

TAG	FREQUENCY
RB	34684
NN	178612
CC	27777
VBP	1802
IN	69348
DT	60750
NNS	5806
PRP	35342
TO	16627
VBN	21999
VB	12631
VBZ	3441
MD	6278
RBR	1222
PRP$	16727
WRB	3638
JJ	23070
WP	4911
VBD	21345
VBG	15650
CD	3429
JJS	795
WDT	2085
JJR	354
WP$	111

Output with your Inferences

1) When preprocessing steps such as lowercasing and lemmatization were applied the results obtained were more consistent than that observed before applying preprocessing.
2) Removal of stop words showed prominent character names and proper nouns, more consistent with the observations of a reader, going through the text corpus.
3) The frequency distribution bar plot showed very similar kurtosis and skewness(**right/positive skewness**) with the mode of the plot being 3 which signifies that even though they are different text corpus the core distribution of the words remains similar.
4) The frequency distribution of both text corpus is also congruent with the word frequency of the English Vocabulary in the sense that words with large length are less frequent that the words with smaller length. So the word frequency is inversely proportional to word length for length greater than 3.