

Problem 2

Problem Statement:

In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use `.words()`, `.raw()`, `.sent()` for extracting counts)

The inaugural text data is downloaded and the methods – `raw()`, `words()` and `sents()` are applied to each of the president speeches' text file to get the number of characters, words and sentences in each text file.

Below is the required data in tabular format.

President Speech	No. of Characters	No. of Words	No. of Sentences
Franklin D. Roosevelt (1941)	7571	1536	68
John F. Kennedy (1961)	7618	1546	52
Richard Nixon (1973)	9991	2028	69

2.2) Remove all the stopwords from the three speeches.

The stopwords are imported from the nltk library. Stopwords include commonly used words like 'a', 'an', 'the', 'in'..etc which don't contribute much information when processing a text file. Hence, these words, along with punctuations are removed from the text files.

So, for all the presidents' speeches, the data now looks like this, after removing all the stopwords.

Roosevelt's speech

```
['nation',  
 'day',  
 'inaugur',  
 'sinc',  
 'peopl',  
 'renew',  
 'sens',  
 'dedic',  
 'unit',  
 'state',  
 'washington',
```

Kennedy's speech

```
['vice',
 'presid',
 'johnson',
 'mr.',
 'speaker',
 'mr.',
 'chief',
 'justic',
 'presid',
```

Nixon's speech

```
['mr.',  
'vice',  
'presid',  
'mr.',  
'speaker',  
'mr.',  
'chief',  
'justic',  
'senat',
```

2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)

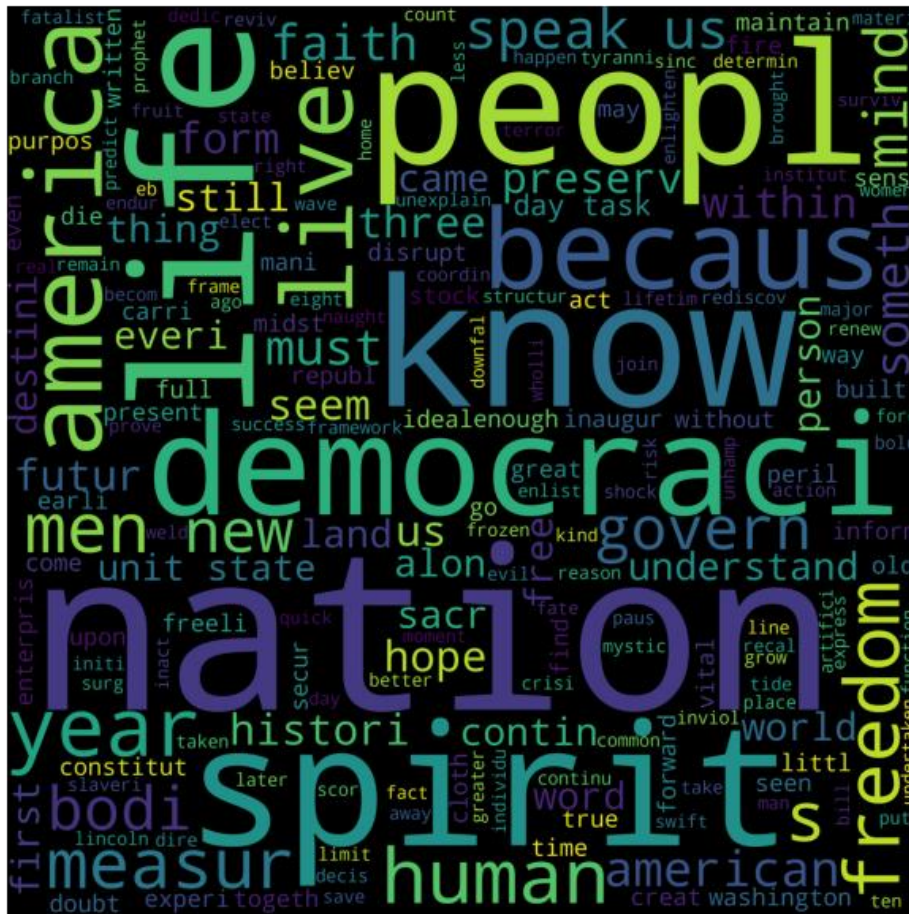
The top 3 frequently occurring words in each of the presidents' speeches is displayed below.

President Speech	Top 3 commonly occurring words
Franklin D. Roosevelt (1941)	'nation', 'know', 'peopl'
John F. Kennedy (1961)	'let', 'us', 'power'
Richard Nixon (1973)	'us', 'let', 'america'

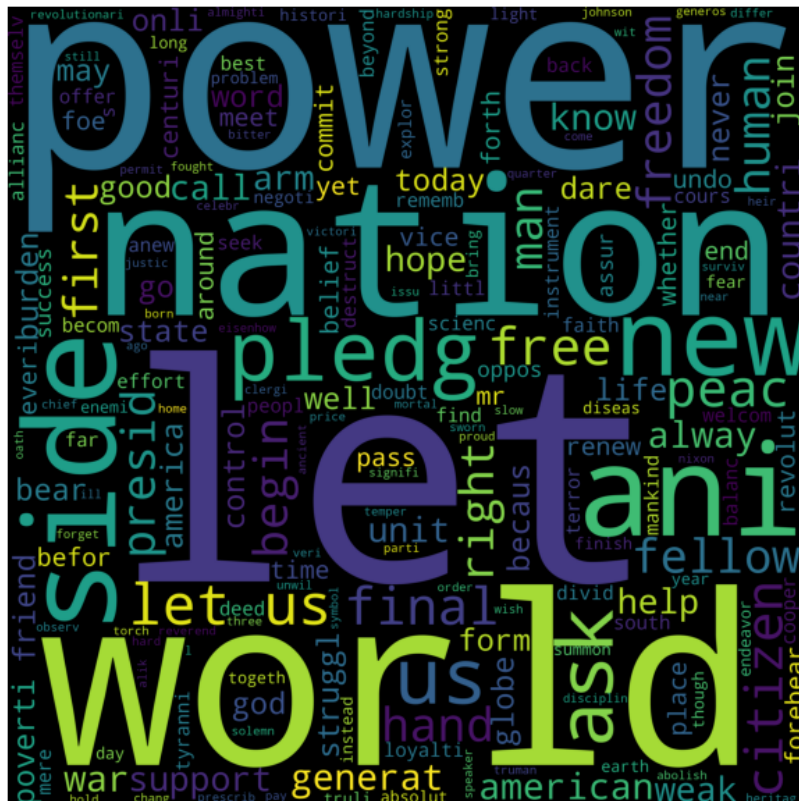
2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)

The word clouds for each of the presidents' speeches are illustrated below.

Word cloud for Franklin D. Roosevelt (1941) speech



Word cloud for John F. Kennedy (1961) speech



[illegible]