Assignment 1

Introduction to NLP and Text mining

In this assignment, you will be extracting some information from the Herman Melville novel Moby Dick. Later you will build a spelling corrector. The objective of this assignment is to get you familiar with text mining using natural language processing tools. This assignment is an individual effort.

1- How many tokens and unique tokens in the text (words and punctuation symbols)? (10 points)
2- Apply lemmatizations on the verbs in the text, recalculate the number of tokens and unique tokens. (10 points)
3- What percentage of tokens is ' HISTORY' or 'history'? (10 points)
4- What are the 10 most frequently occurring (unique) tokens in the text? What is their frequency? (10 points)

(40 points)
Spelling recommender/checker
For the spelling recommender/checker, your code should provide recommendations for the user input using the edit distance and print the word that is the closest to the user input.

You should expect the user input to be either a misspelled word ex: validrate or an actual word- one word at a time. The user input consists of alphabetical letters only that can include both uppercase and lowercase letters.

Hint: you can use NLTK corpus (words) as a dictionary- you are also free to use any other dictionaries.

```
import nltk
nltk.download('words')
from nltk.corpus import words
correct_spelling = words.words()
```

(20 points)
Deliverability report and readme.txt file

(5 points) Bonus
Adding comments to your code

Deliverability: Your code should be submitted via Webcourses along with a report that includes your answers and your explanation of how you solved the given question. A readme file that includes information about how to run your code is required.
Notes: It is recommended to use python and NLTK library as it has many built-in functions. However, you are free to use any other programming language that you feel comfortable with.