TP 3:

Exercício 1:

Primeiramente é feita todas as importações de bibliotecas importantes que serão utilizadas no decorrer do programa, tais como: nltk (Natural Language Toolkit), matplotlib e entre outras.

Após isso é feito o download do catálogo de algumas obras de Machado de Assis através de uma função do nltk chamada ‘download’ o qual faz o download desse catálogo e os lista na tela. Logo em seguida, se é escolhida uma dessas obras (Dom Casmurro) e a mesma é exibida na tela.

Posto isso, é realizado todo o processo de remoção da pontuação e da acentuação de todos as palavras do texto em questão. Em seguida é executado um passo muito importante, que é a retirada das *stopwords,* que são palavras que não possuem valor semântico.

Por fim, é calculado a frequência do texto com a utilização de stemming e sem a utilização de stemming e é mostrado também graficamente a diferente entre as duas maneiras.

É importante citar que, a principal diferença entre stemming e lemmatization é que no processo de stemming é removido os útlimos caracteres da palavra, podendo levar à sentidos incorretos, sem analisar um contexto geral; já no lemmatization, também é feita a remoção de alguns caracteres do termo em questão, porém é analisado todo o contexto e sendo assim, é retornado a palavra reduzida que mais faz sentido.