Ex1

Pré-processamento

Pré-processamento é uma etapa importante que envolve a limpeza e transformação dos textos brutos dos documentos para facilitar a análise. Tendo como objetivo filtrar as informações, removendo palavras não significativas e transformando o texto em uma representação mais adequada para a análise.

StopWords

Stopwords são palavras que são comuns e que geralmente não contribuem para o significado do texto como, por exemplo: ‘as’, ‘e’ , ‘os’, ‘de’ e etc.

Stemming e Lemmatization

As duas são técnicas para reduzir palavras para a sua forma base/raiz. A diferença é que *stemming* apenas remove prefixos e sufixos, já *lemmatization* envolve regras gramaticais. Por exemplo, ‘processando’ em *stemming* ficaria ‘process’ já em *lemmatization* ficaria ‘processar’.

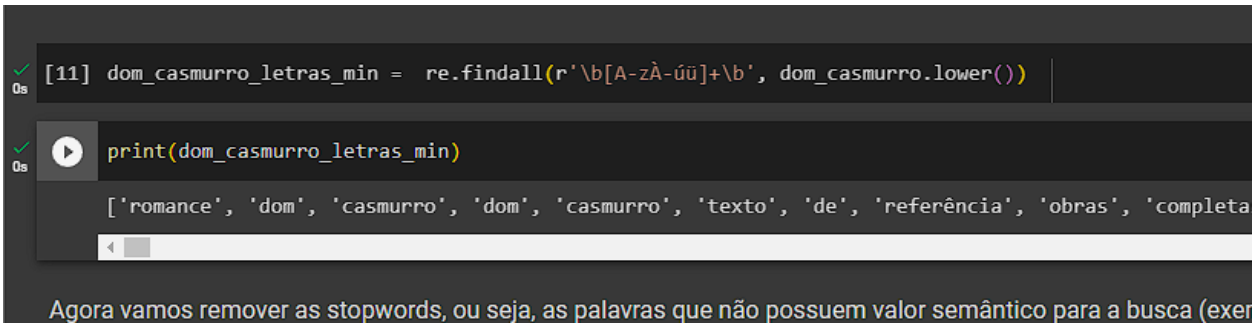
Explicação passo a passo:

```
[6]  nltk_id = 'machado'

[7]  nltk.download(nltk_id)
```

Primeiramente, é baixado da biblioteca NLTK recursos relacionados a Machado de Assis.

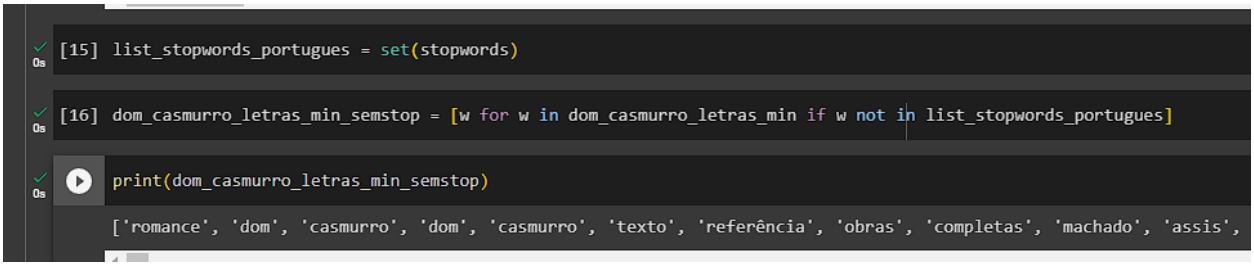
```
miscelanea/mams08.txt: A Paixão de Jesus (1904)
miscelanea/mams09.txt: Gonçalves Dias (1906)
miscelanea/mams10.txt: A Estátua de José de Alencar (1906)

[9]  dom_casmurro = nltk.corpus.machado.raw('romance/marm08.txt')

     print(dom_casmurro)
os dois amigos da universidade me levantaram um tumulo com esta inscr
tirada do profeta Ezequiel, em grego: 'Tu eras perfeito nos teus
caminhos'. Mandaram-me ambos os textos, grego e latino, o desenho da se
```

Corpus acessa a coleção ‘machado’, carrega o texto de ‘marm08.txt’ presente no diretório ‘romance’ e armazena em ‘dom_casmurro’

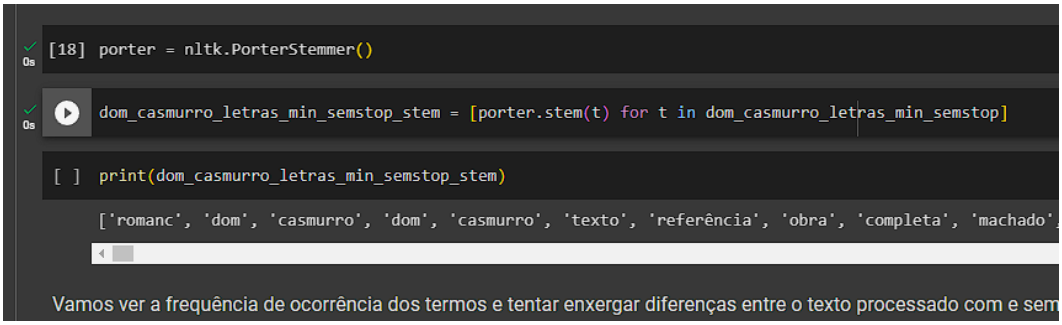


É então extraído todas as palavras em letras minúsculas de ‘dom_casmurro’ e armazenadas em ‘dom_casmurro_letras_min’

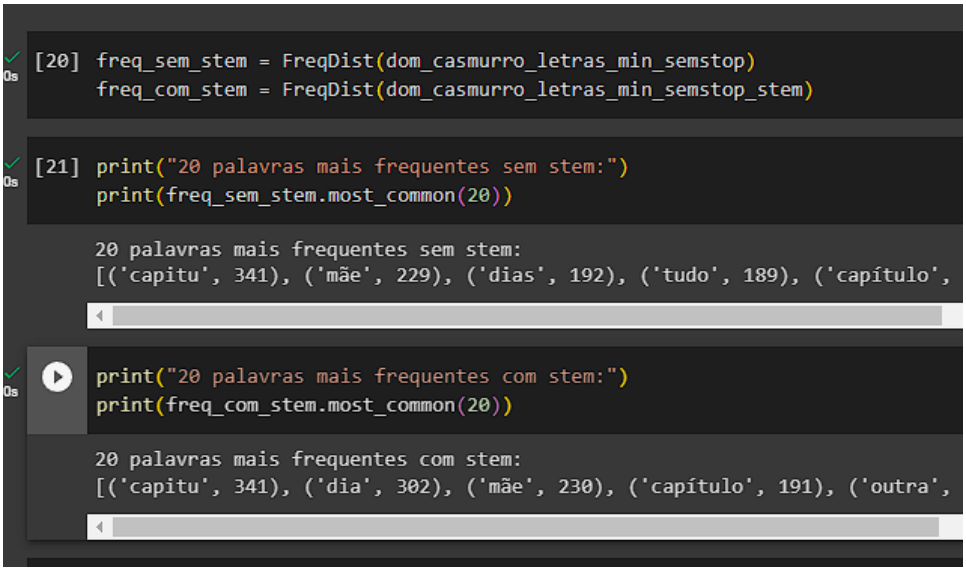
A coleção de stopwords é baixada do pacote NLTK e logo é armazenado as *stopwords* da língua portuguesa para ‘stopwords’



É criado uma lista cujo recebe todas as palavras da lista ‘dom_casmurro_letras_min’ exceto aquelas presentes na lista de *stopwords*



É realizado o *stemming* usado o algoritmo de Porter para normalizar as palavras (reduzi-las a uma forma raiz, removendo prefixos e sufixos). Estas são colocadas em 'dom-casmurro_letras_min_semstop_steam'



Usa-se *FreqDist* da NLTK para calcular a frequência de cada palavra e então printa-se as 20 palavras mais frequentes

Por fim, é exibido os gráficos de frequência das palavras sem e com *stemming*