PEL218 - Exercício 01

Escolher qualquer corpus (conjunto de documentos) ou livro de ate 100 MB em português e extrair as seguintes informações:

- Quantidade de palavras distintas;
- Histograma das palavras;
- Histograma de prefixos de tamanho (1,2,3,4 e 5)
- Histograma de sufixos de tamanho (1,2,3,4 e 5)

Você poderá utilizar qualquer linguagem de programação (inclusive shell).

Relatório

Para esta atividade foi escolhido um conjunto de notícias do jornal A Folha de São Paulo contendo aproximadamente 30 mil artigos.

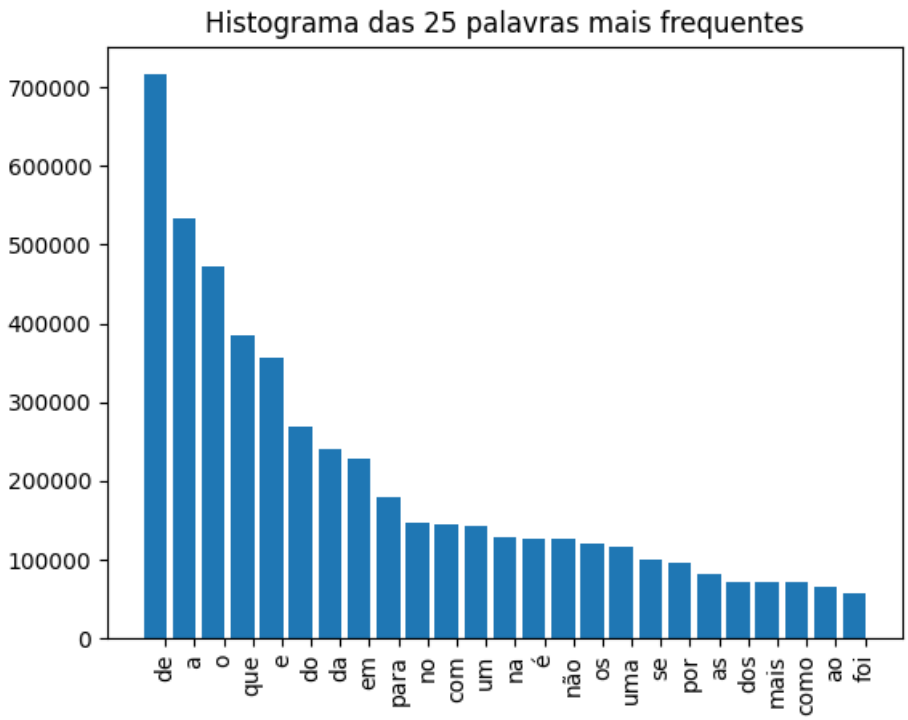
Para iniciar a análise foi necessário preparar o corpus. Esta preparação incluiu os seguintes passos:

1. Remover os caracteres especiais e pontuação;
2. Transformar todos os caracteres para minúscula;
3. Remover quebras de linha;
4. Remover espaços duplicados.

Após esta preparação o corpus continha 14.779.320 de palavras, porém muitas delas repetidas quando removemos as repetições pudemos identificar 184.462 palavras únicas.

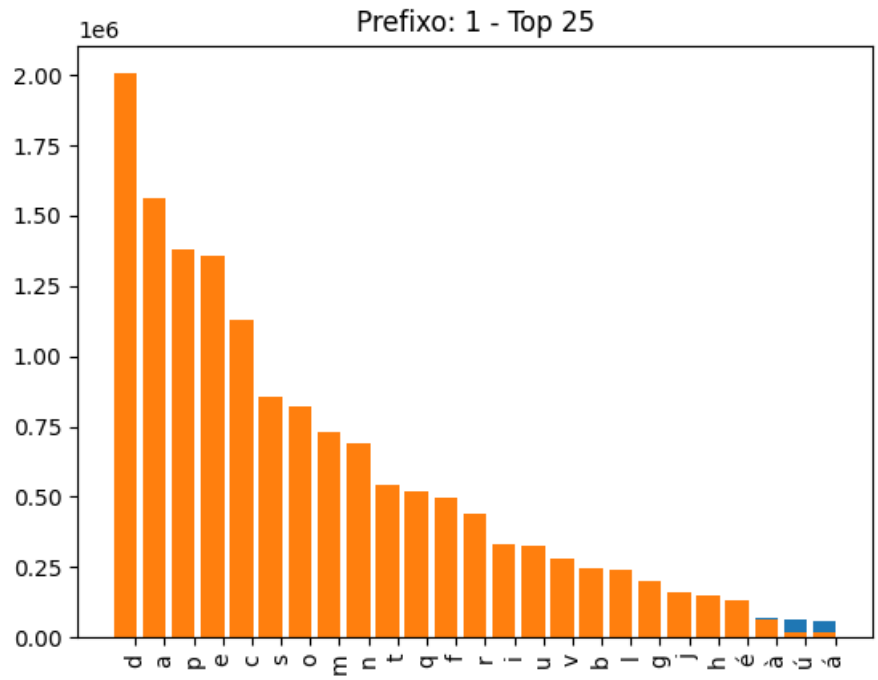
Frequência de palavras

Analisando o corpus pela frequência de palavras temos a lista abaixo com as 25 mais frequentes:

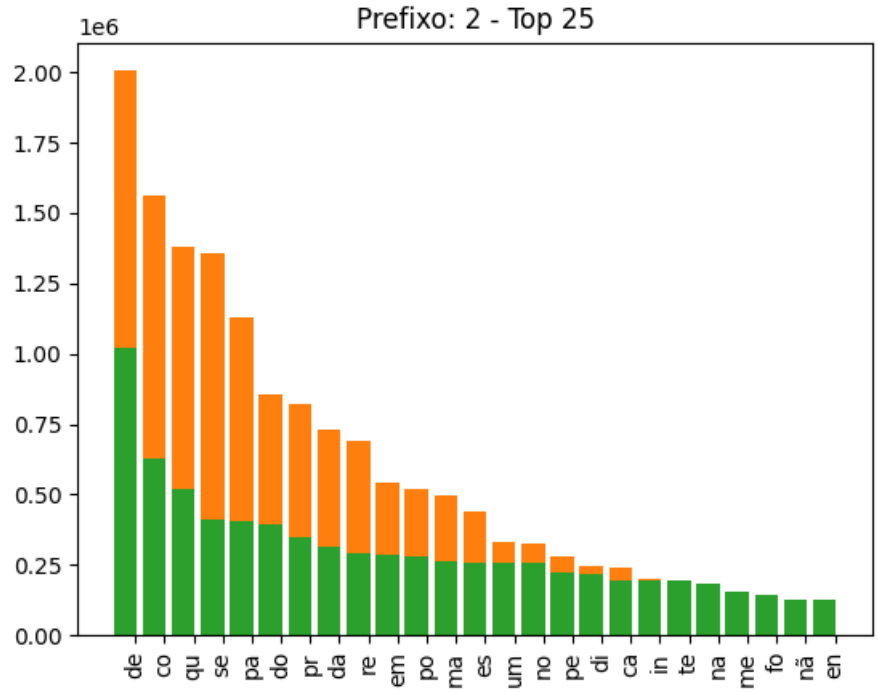


Prefixos

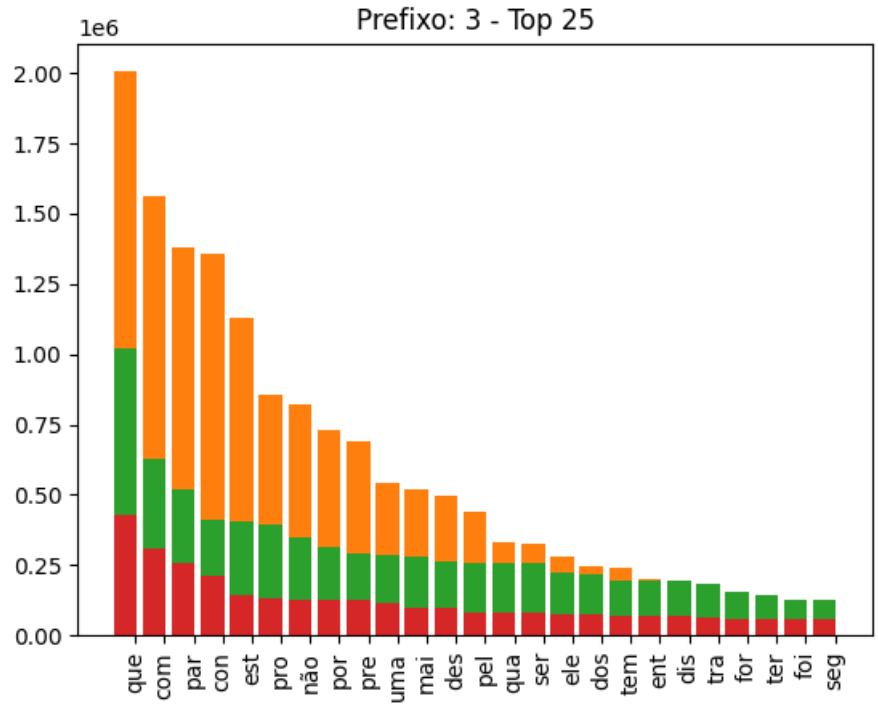
Prefixo de tamanho 1



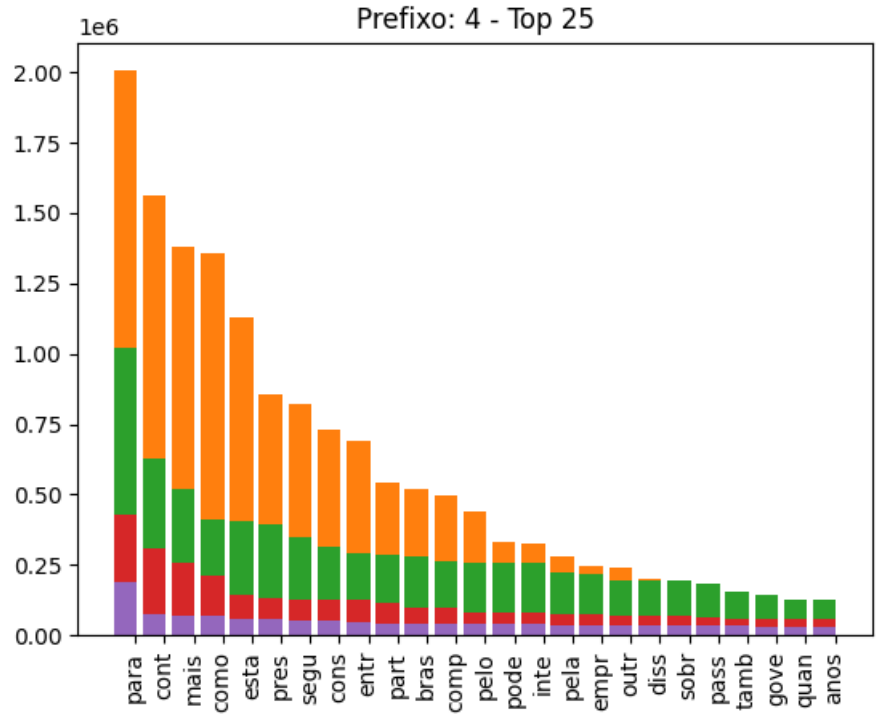
Prefixo de tamanho 2



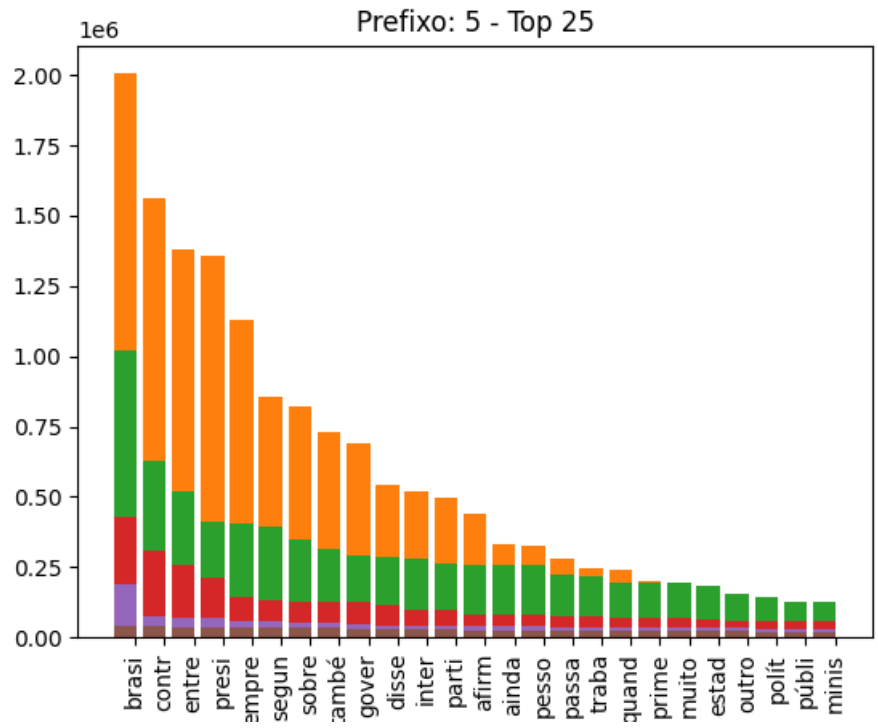
Prefixo de tamanho 3



Prefixo de tamanho 4

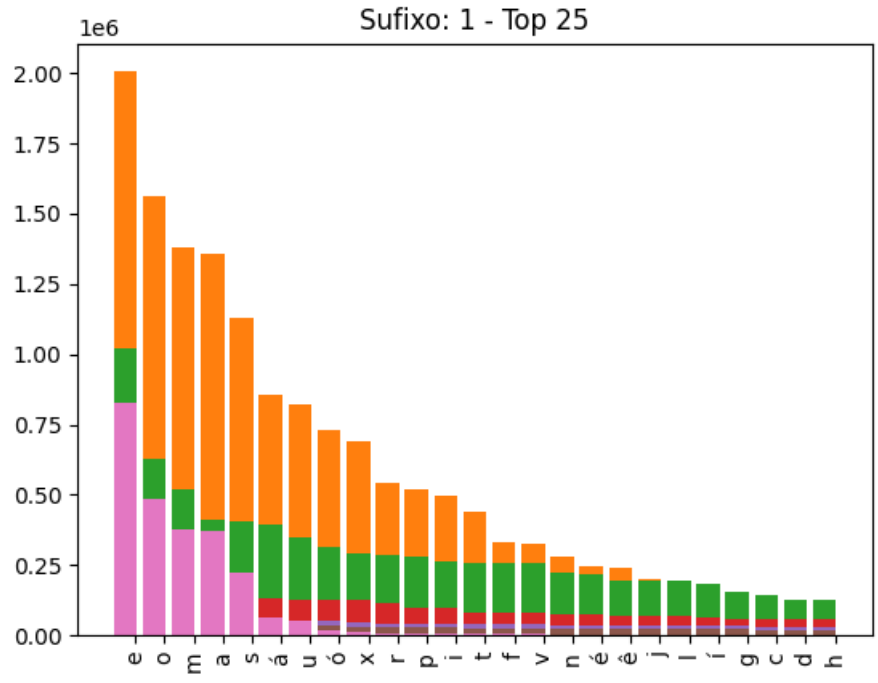


Prefixo de tamanho 5

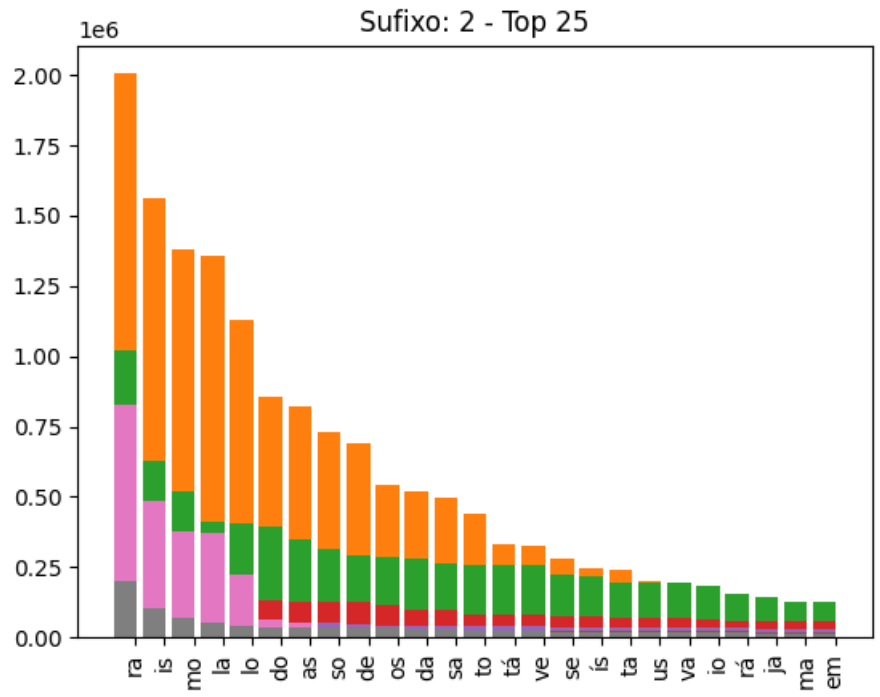


Sufixos

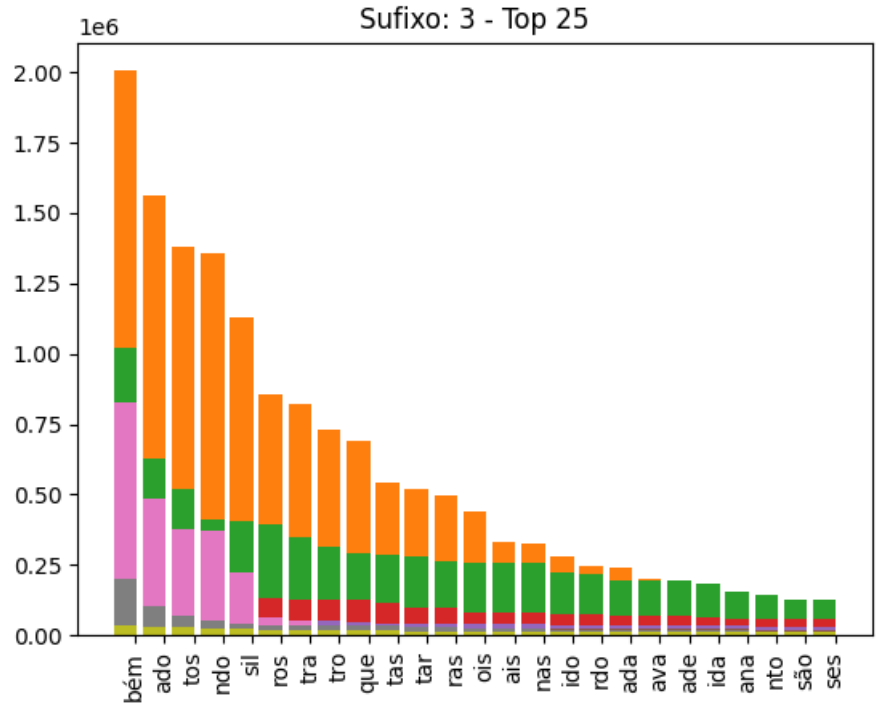
Sufixo de tamanho 1



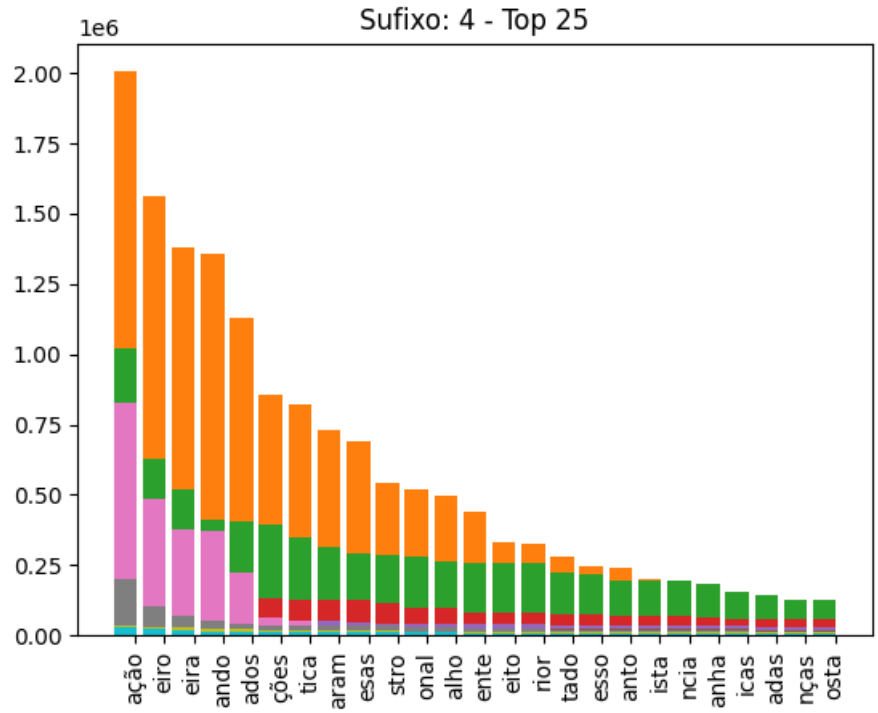
Sufixo de tamanho 2



Sufixo de tamanho 3



Sufixo de tamanho 4



Sufixo de tamanho 5