CENTRO UNIVERSITÁRIO FEI

PEL 218 – Processamento de Linguagem Natural

Atividade 1

JONATHAN KENJI KINOSHITA

Matrícula: 120102-9

São Bernardo do Campo

2020

Tarefa

Escolher qualquer corpus (conjunto de documentos) ou livro de até 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)

Foi utilizado o arquivo “Iracema.txt”, disponibilizado no Moodle, para a realização desta tarefa

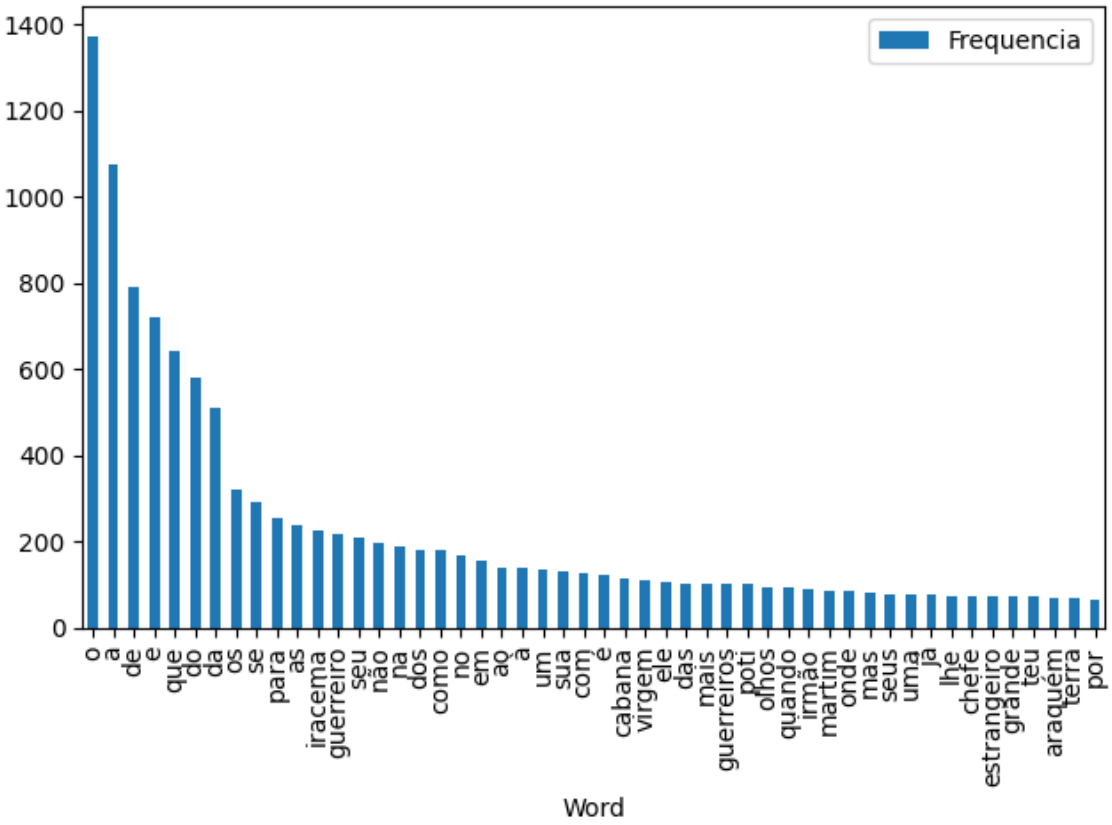
Código disponibilizado no GitHub

Quantidade de palavras distintas

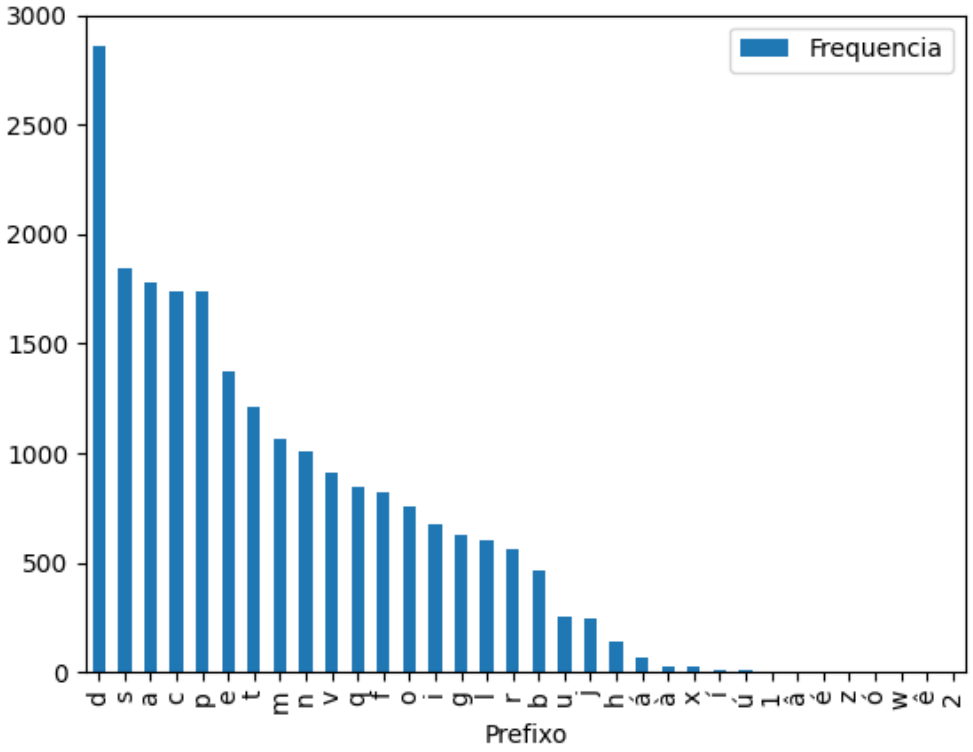
```
import re
from collections import Counter
import pandas as pd
import matplotlib.pyplot as plt
pd.set_option("display.max_rows", None, "display.max_columns", None)

words = re.findall(r'\w+', open('Iracema.txt').read().lower())
histWords = Counter(words).most_common()
qntWords = len(histWords)
print('Quantidade de Palavras Distintas: ',qntWords)
```

```
Quantidade de Palavras Distintas:  4631
```

Histograma das palavras

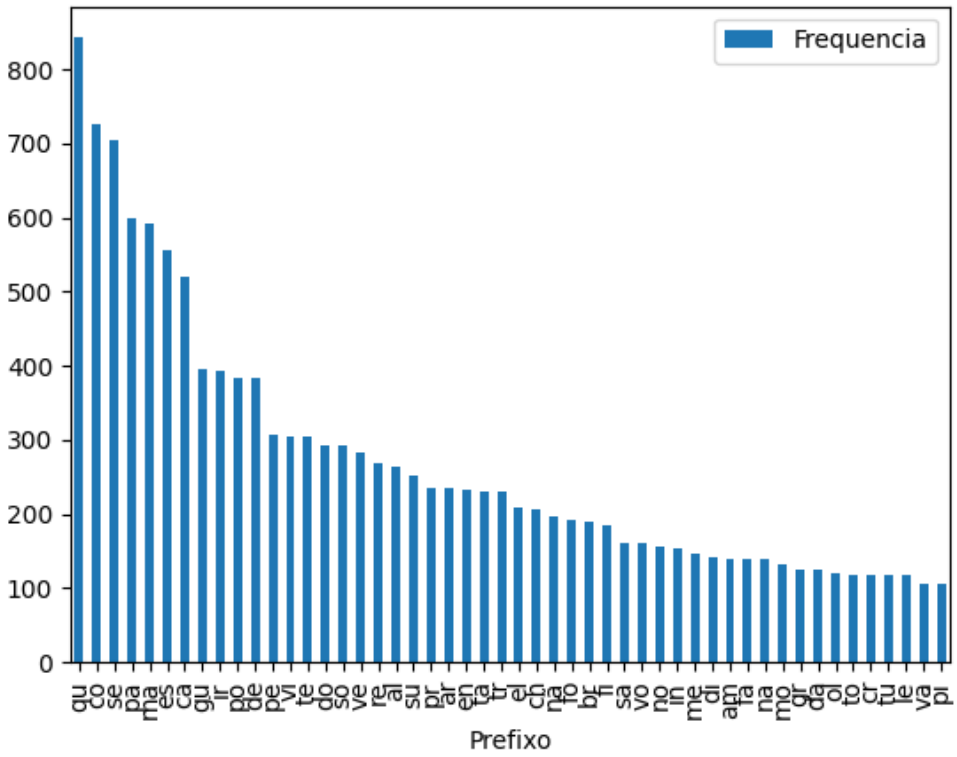


Histograma de prefixos de tamanho 1

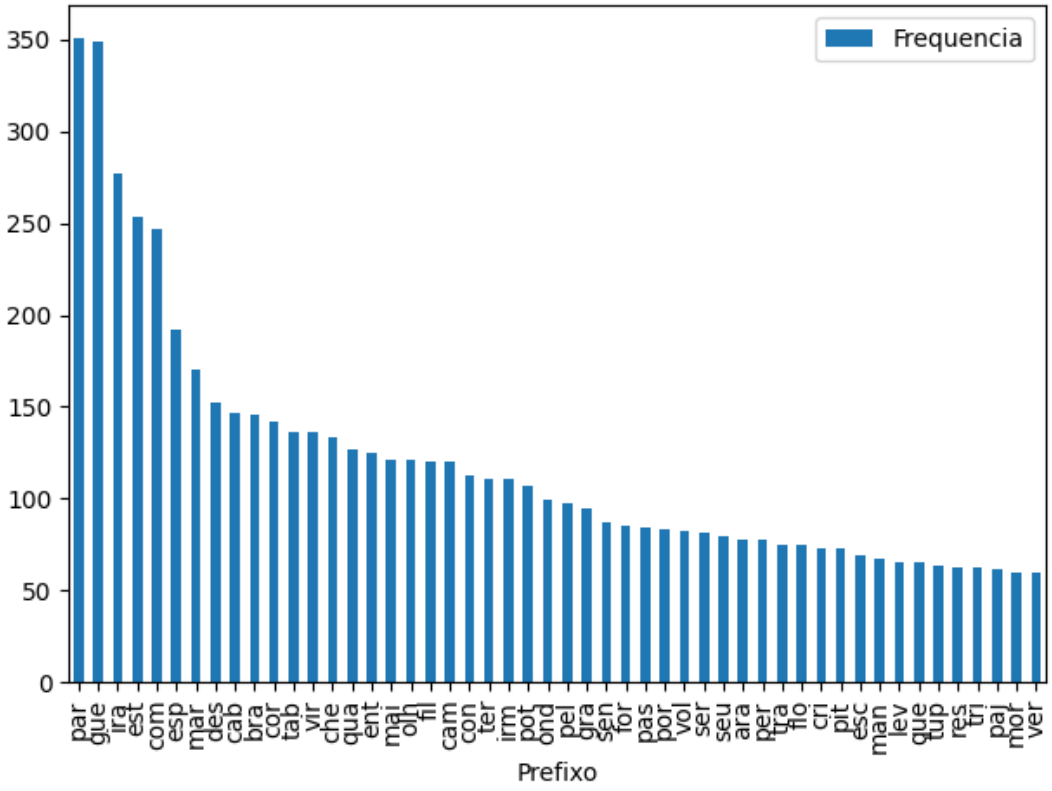


Histograma de prefixo de tamanho 2

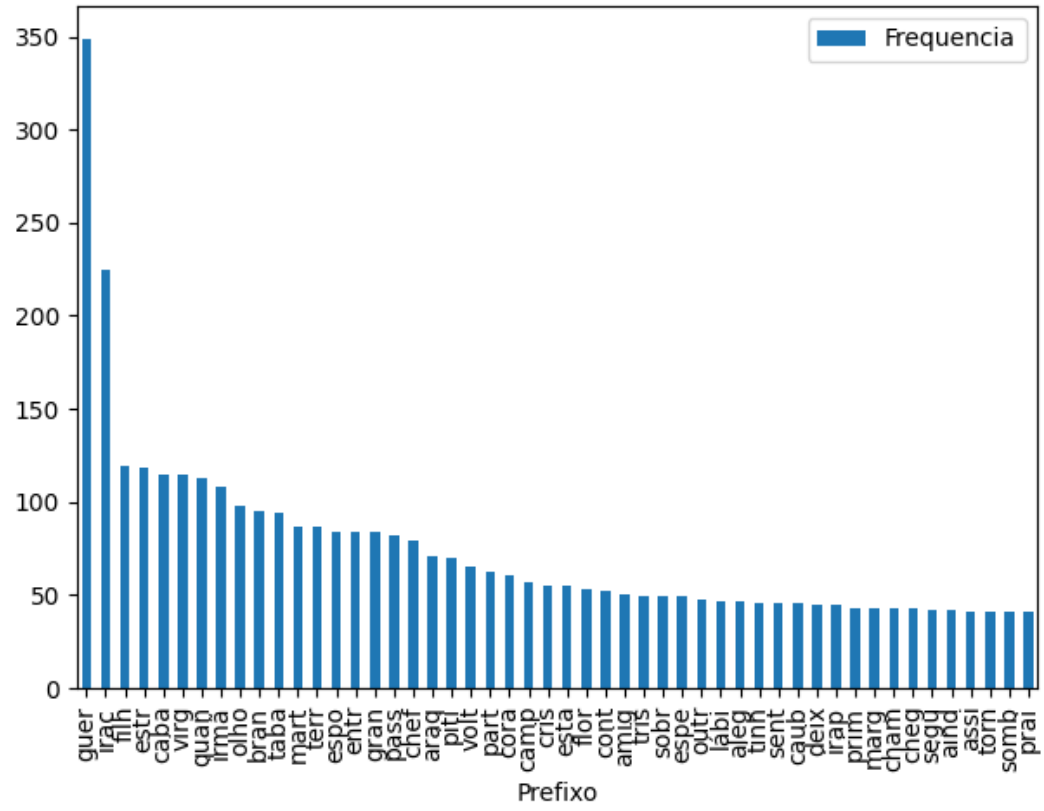
(limitado à 50 amostras)



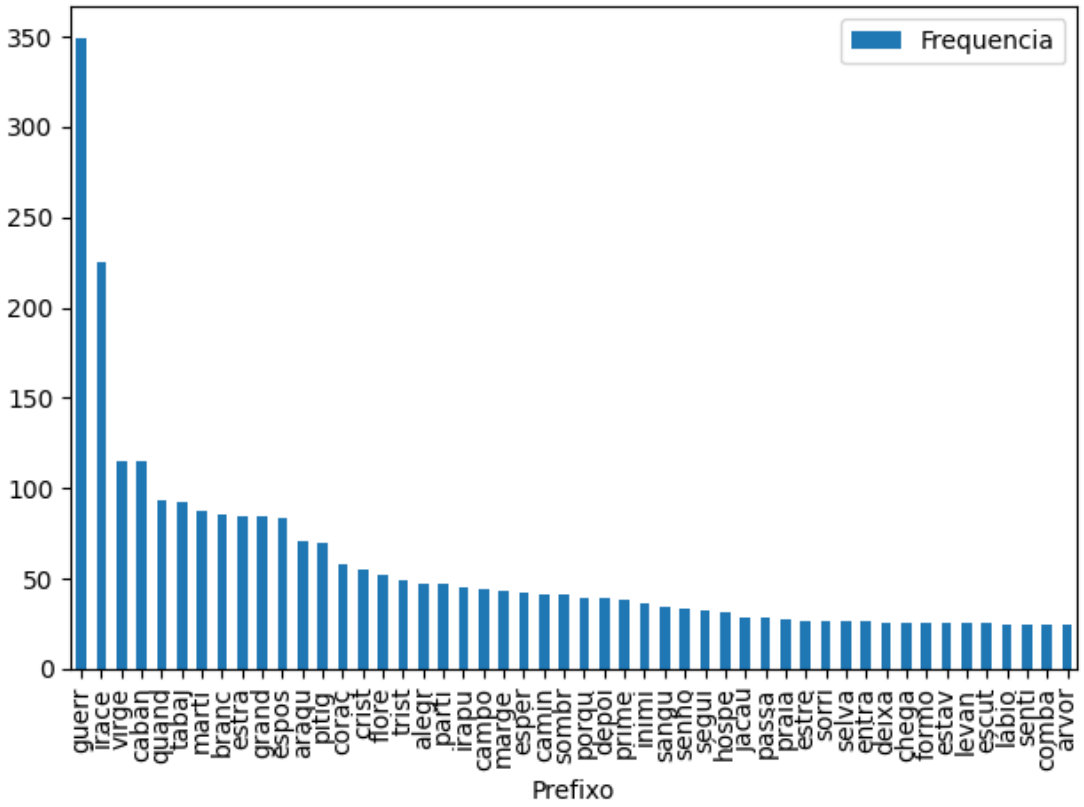
Histograma de prefixo de tamanho 3 (limitado à 50 amostras)



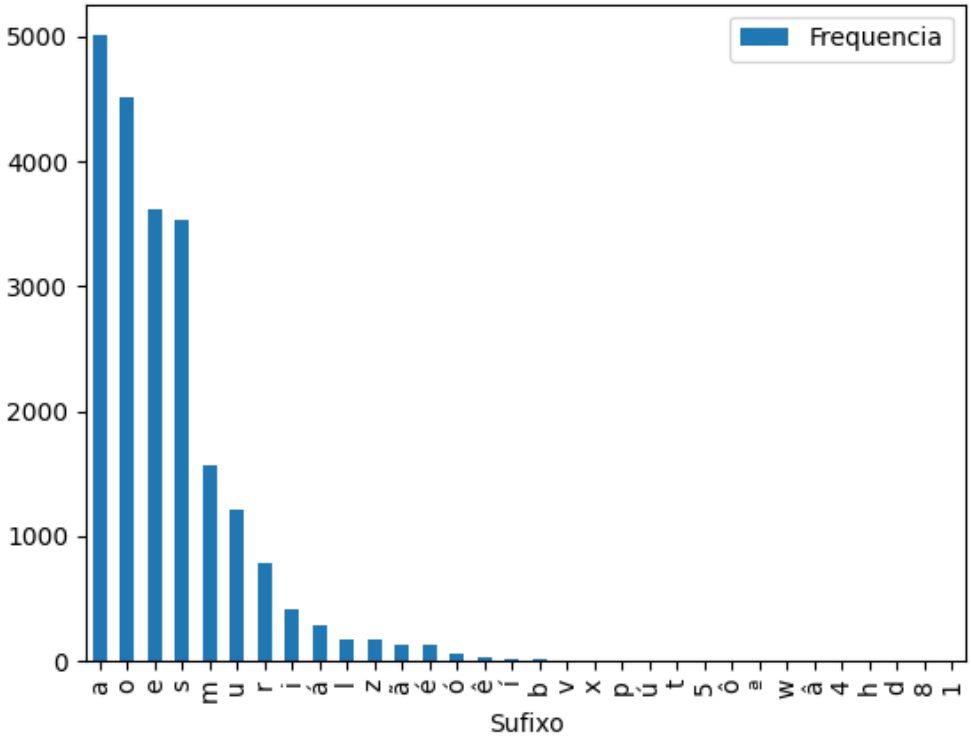
Histograma de prefixo de tamanho 4 (limitado à 50 amostras)



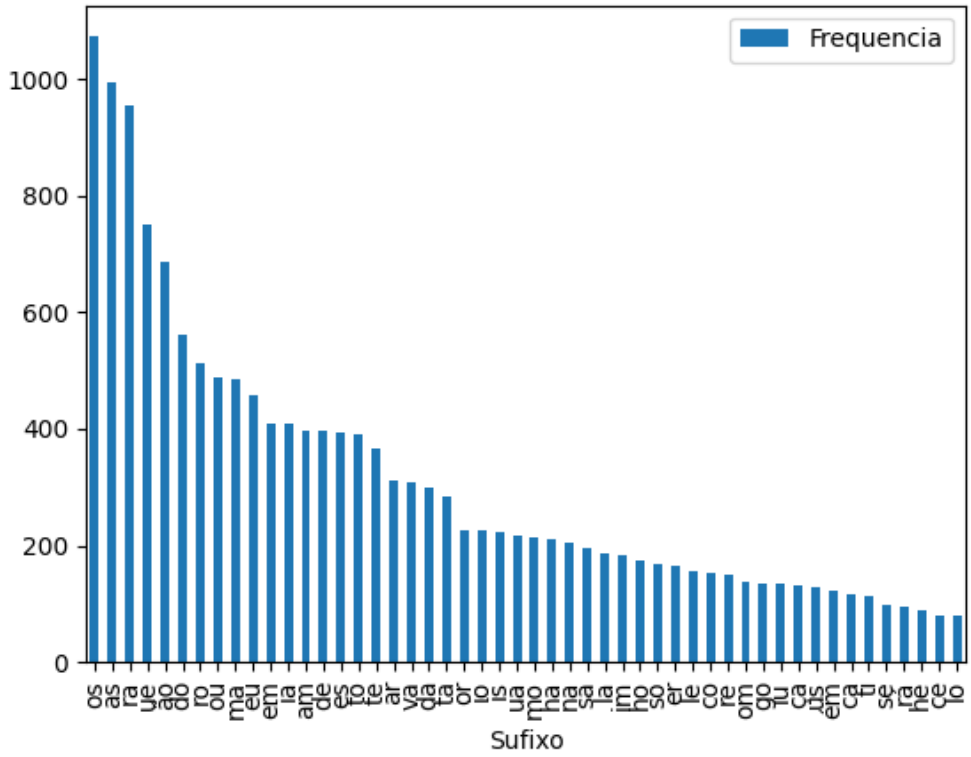
Histograma de prefixo de tamanho 5 (limitado à 50 amostras)



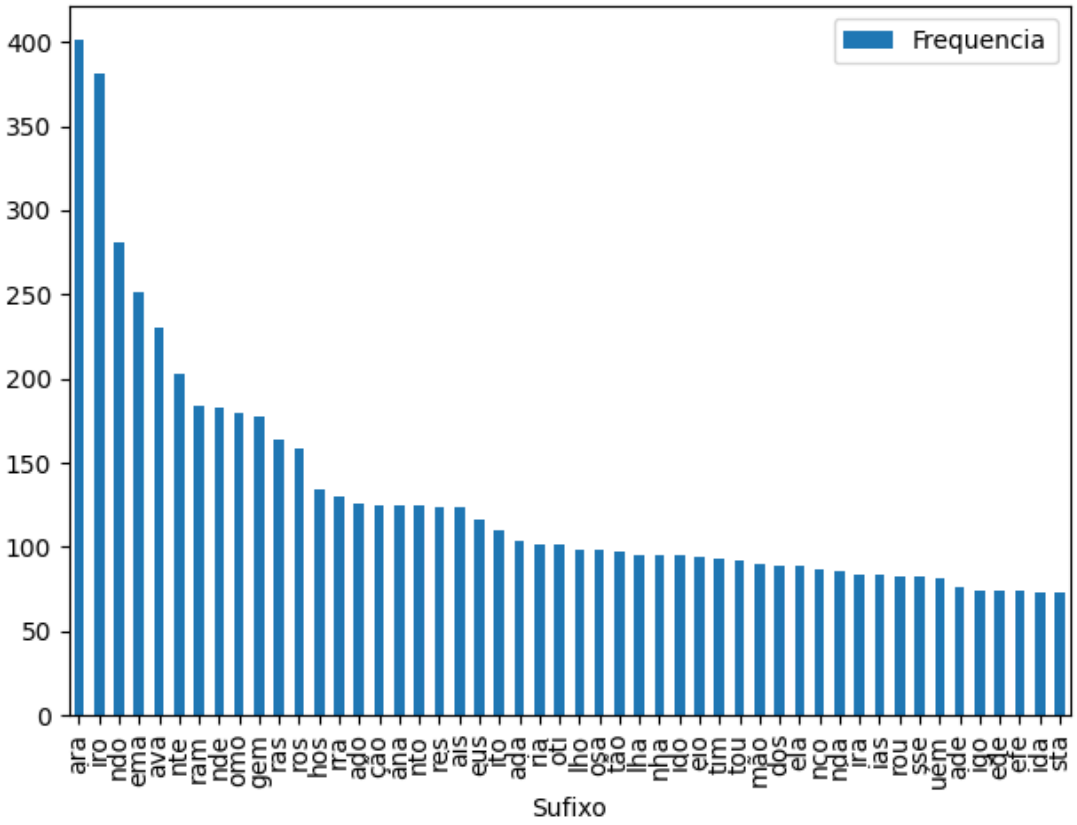
Histograma de sufixo de tamanho 1



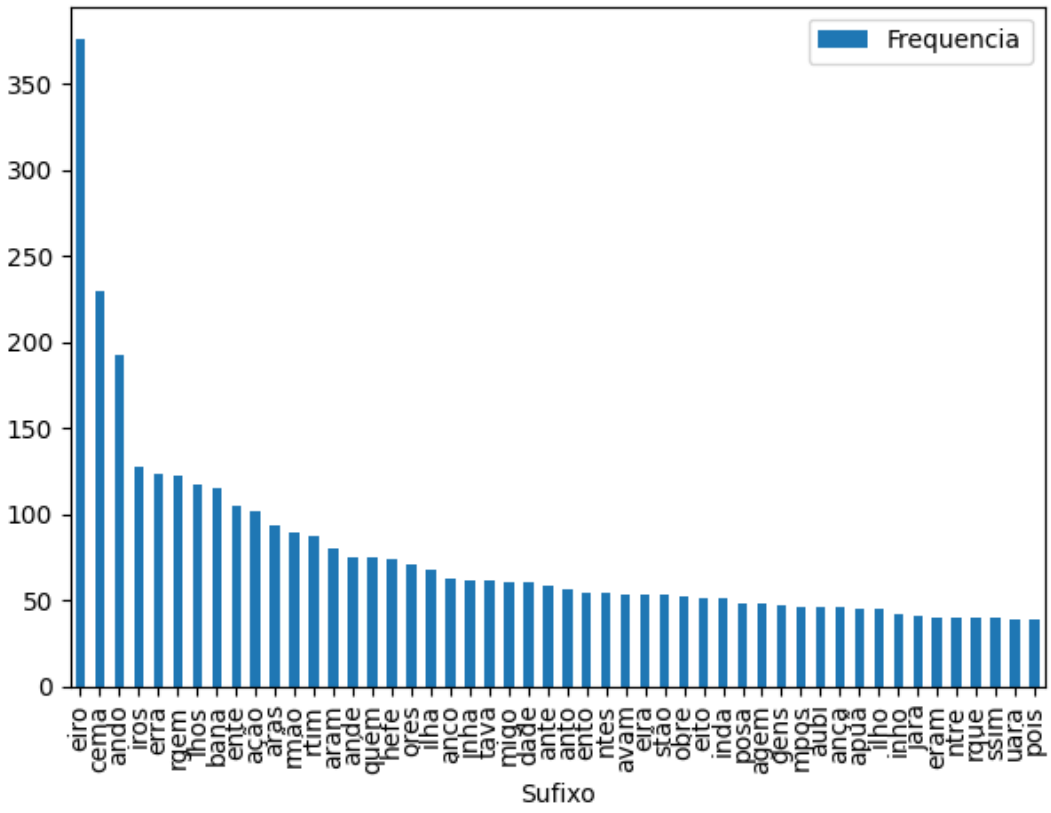
Histograma de sufixo de tamanho 2 (limitado à 50 amostras)



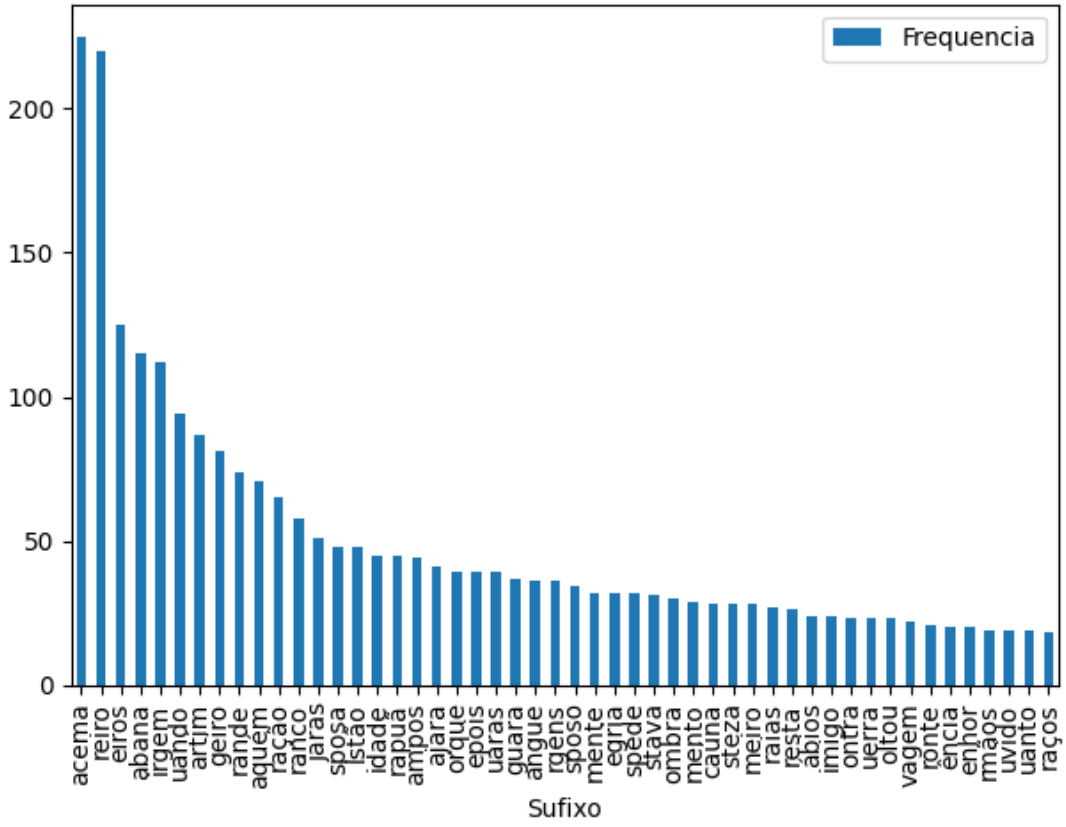
Histograma de sufixo de tamanho 3 (limitado à 50 amostras)



Histograma de sufixo de tamanho 4 (limitado à 50 amostras)



Histograma de sufixo de tamanho 5 (limitado à 50 amostras)



Nos histogramas de sufixos e prefixos foram retiradas todas as palavras que possuíam a mesma dimensão que o tamanho proposto