Universidade do Minho

Mestrado Integrado em Engenharia Informática

4º ANO, 2º Semestre, 2019/2020

Scripting no Processamento de Linguagem Natural

Ana Pereira
A81712

Shahzod Yusupov
A82617

May 4, 2020

Índice

1 Introdução

Este trabalho prático inserido na Unidade Curricular de Scripting no Processamento de Linguagem Natural consistiu na exploração e aplicação da ferramenta **matplotlib** no contexto do processamento de linguagem natural.

Esta ferramenta trata-se de uma biblioteca *Python* que permite criar visualizações interativas de uma grande variadade de gráficos.

Neste relatório, iremos expor exemplos da aplicação desta biblioteca para representar alguns dados estatísticos provenientes do processamento de textos.

2 Gráficos

De modo a mostrar as capacidades desta ferramenta, decidimos apresentar através de vários tipos distintos de gráficos alguns dados provenientes do processamento de duas obras: *Os Lusíadas* e *Harry Potter* (versão inglesa).

Assim, ao correr o programa é possível escolher de entre 4 opções que podemos escolher para serem desenhadas as figuras dos gráficos, sendo possível guardar essas mesmas figuras.

- **Ocorrências de pronomes pessoais em Harry Potter**

  Para obter estes dados foi apenas necessário fazer uma contagem dos pronomes pessoais ingleses, dado a obra estar em Inglês. Tendo de seguida sido desenhado dois gráficos com estes valores.

  Assim, foram efetuados os seguintes gráficos: um gráfico circular (*pieplot*) e um gráfico de linha simples (*line plot*). Através do gráfico de linha é possível observar o número concreto de ocorrências enquanto que o circular dá-nos uma melhor ideia da proporção de cada pronome ao longo da obra.

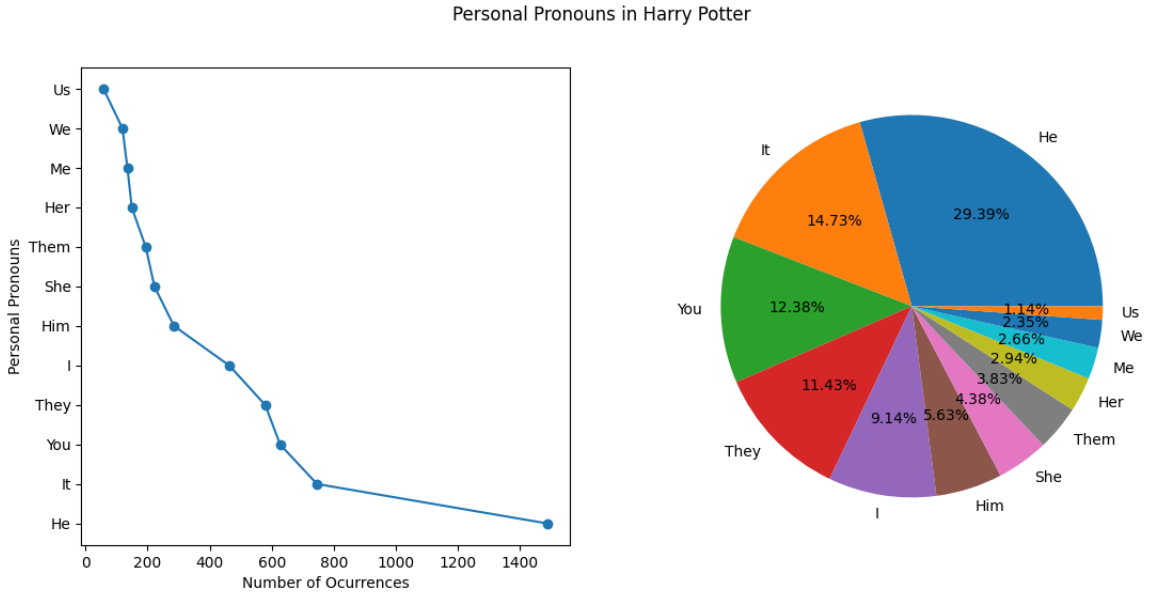

Figura 1: Ocorrências de Pronomes Pessoais em Harry Potter

- **Menções dos deuses romanos n'*Os Lusíadas***

  Como na opção anterior foram contadas as ocorrências dos deuses romanos principais presentes na obra. Esta informação foi apresentada em dois gráficos: um gráfico de barras (*barplot*) e um gráfico donut (*doughnut plot*).

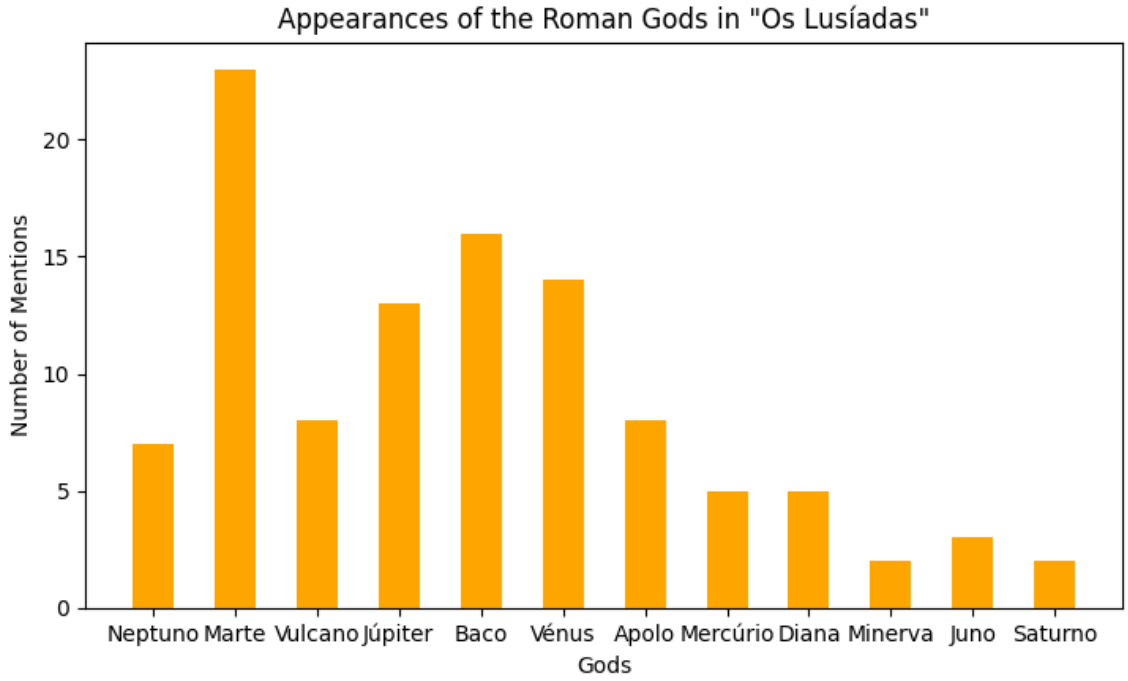

Figura 2: Menções dos Deuses Romanos n'Os Lusíadas - Gráfico de Barras

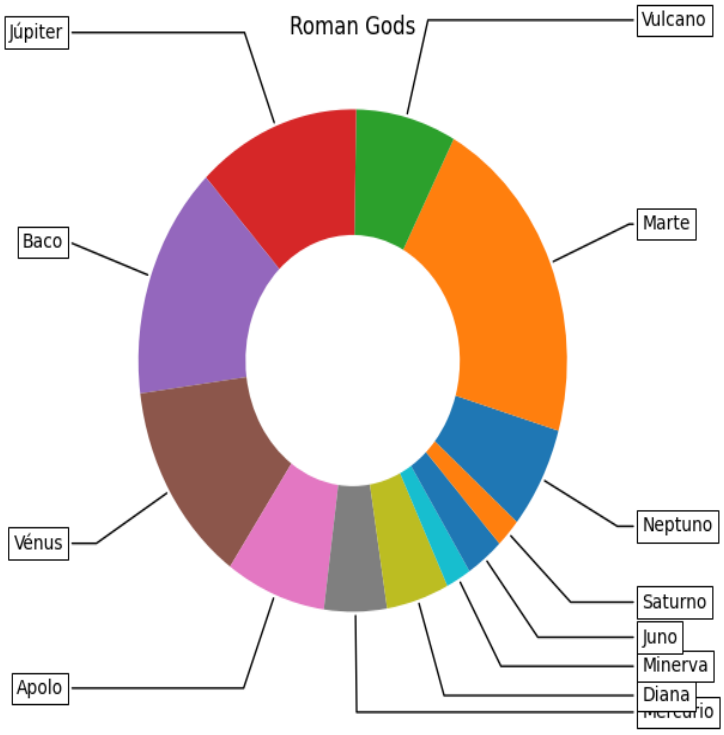

Figura 3: Menções dos Deuses Romanos n'Os Lusíadas - Gráfico Donut

- **Polaridades das palavras da obra *Os Lusíadas* por canto**

  Esta opção requer a identificação da polaridade das palavras presentes no texto. Para tal foi usada a ferramenta **linguakit** de modo a obter o ficheiro *Lusiadas.tagged* com a identificação de cada palavra. De seguida, através do ficheiro *sentilex2.txt*, que contém as polaridades de várias palavras portuguesas, obtivemos o número de palavras com polaridade positiva e negativa em cada canto da obra.

  Os resultados obtidos foram representados num gráfico de barras que nos indica a tendência encontrada em cada canto, isto é, cada representa a diferença de palavras positivas e negativas do canto em questão.

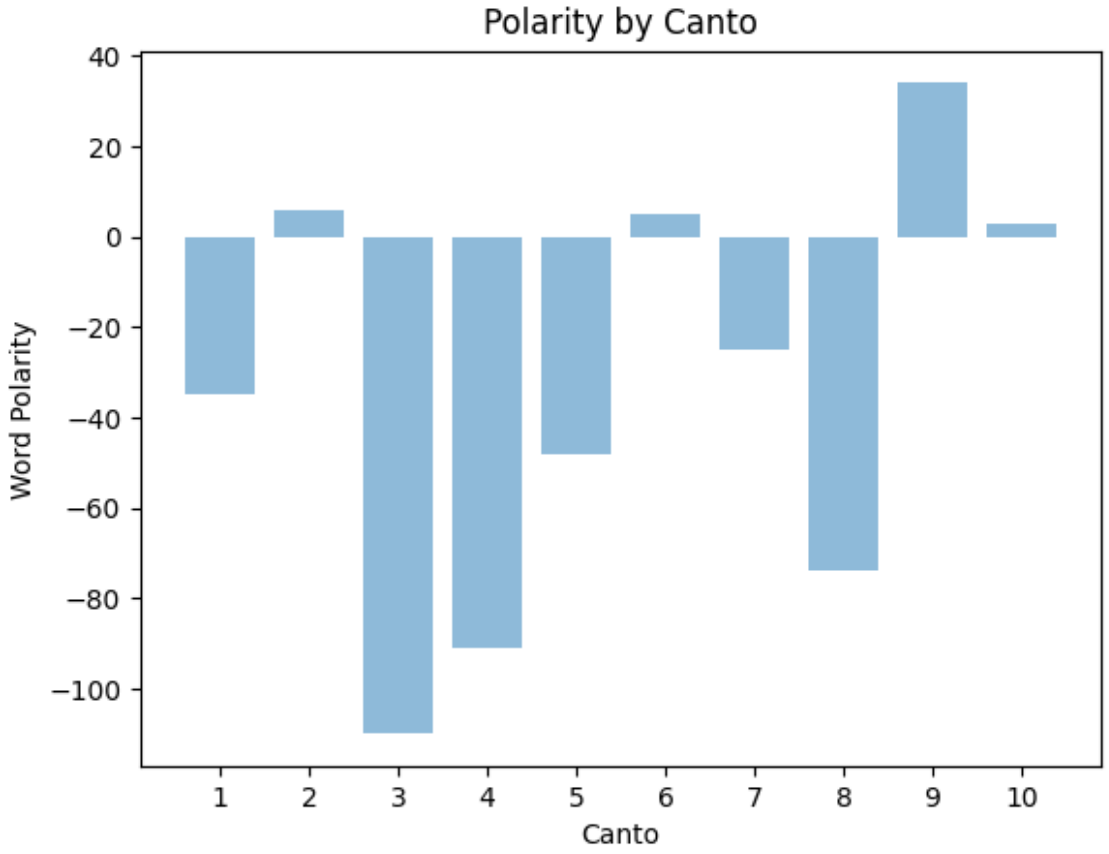

Figura 4: Polaridades das palavras d' Os Lusíadas por Canto

Podemos observar que, marioritariamente, as emoções expressas em cada canto são negativas, talvez devido às varias adversidades com que os portugueses se deparam ao longo da viagem, excetuando alguns casos como no Canto 9 que retrata a chegada dos portugueses à *Ilha dos Amores*, transmitindo na mariora das vezes emoções de carácter positivo.

- **Ocorrências de palavras iniciadas por "mar" a obra *Os Lusíadas***

  Visto esta epopeia relatar a viagem marítima dos portugueses até à China, o mar é algo bastante presente ao longo da obra. Assim, é interessante ver a distribuição de palavras iniciadas por **mar**, tal como *marítimo* ou *marinheiro*, no decorrer do texto.

  Após a contagem das ocorrências das palavras que começassem por *mar*, foram retiradas aquelas com ocorrências abaixo de um certo *threshold*, de modo a apresentar gráficos legíveis.

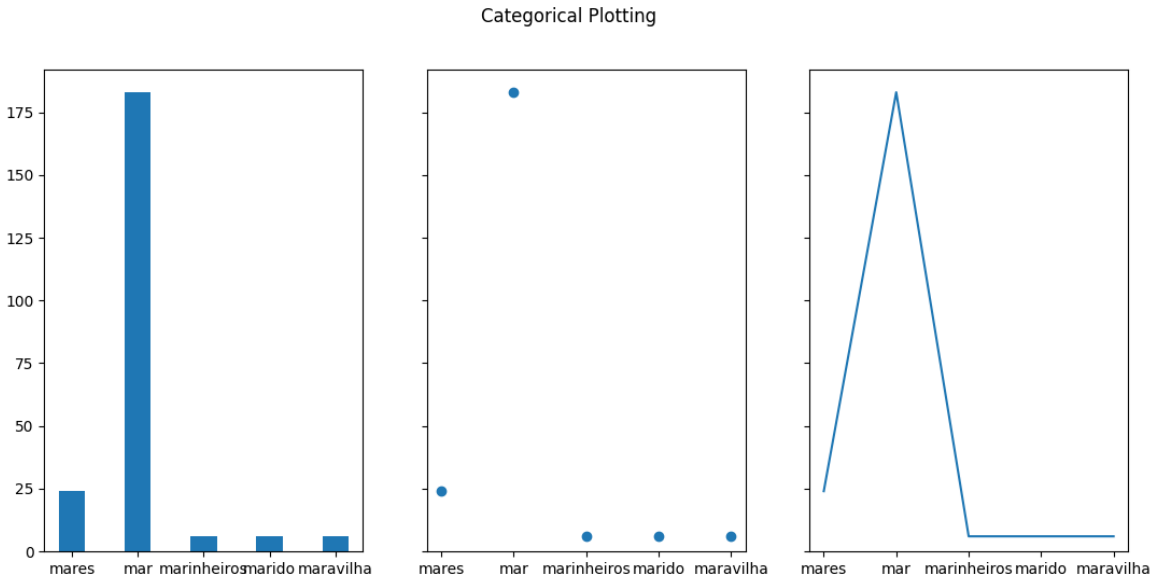

Figura 5: Ocorrências de palavras iniciadas por *mar*

A representação das frequências do número de ocorrências foi feita em **gráfico de barras**, de **pontos** e de **linhas**, respetivamente, em que podemos ver facilmente a variação de cada uma. Talvez alguns não sejam os mais indicados para representar este tipo de dados, por exemplo o gráfico de linhas, por não se tratarem de variáveis contínuas, porém a adição deste tem como finalidade apresentar a diversidade de tipos de gráficos que é possível construir com esta biblioteca.

3 Conclusão

Através deste trabalho prático foi possível aprender o funcionamento de uma nova ferramenta, bastante útil para a representação de dados, podendo ser aplicada nas mais variadas áreas. Neste caso foi exemplificado o seu uso desenhando alguns dos gráficos mais comuns que o **matplotlib** nos disponibiliza, com o intuito de representar dados provenientes da análise de textos.

Concluindo, sentimos que o trabalho prático foi efetuado com sucesso e os objetivos propostos para foram cumpridos.