ISCTE IUL
Instituto Universitário de Lisboa

Desenvolvimento de mini projeto em Python

ISCTE – IUL

Unidade Curricular – Fundamentos de Programação

Pós-Graduação em Ciência de Dados Aplicada

2019/2020

Histograma de Palavras

Identificação do Grupo de Trabalho:

Edgar Basto // Nº 93575 // edgar_basto@iscte-iul.pt

Bruno Araújo // Nº 80852 // bruno_miguel_araujo@iscte-iul.pt

Breve enquadramento:

De modo a apresentar uma visão alto nível do código desenvolvido podemos distinguir três blocos de código: a classe Texto(), a classe Palavra(), e a função main().

A classe Texto() tem como objetivo criar um objeto com o texto de uma notícia, que após processada gera um objeto da classe Palavra para cada palavra única na notícia. É nesta classe que está incluído o método que permite desenhar o histograma de palavras. No entanto, para obter este histograma é necessário descarregar e instalar a livraria *matplotlib.* Esta classe é composta por atributos e métodos a detalhar mais à frente.

A classe Palavra() cria um objeto com uma palavra e com uma dada frequência, isto é, quantas vezes ocorre ao longo de todo o texto a palavra.

E a função main() num formato de código mais procedimental faz a gestão do menu que irá receber inputs do utilizador.

Nas páginas seguintes explicar-se-á em detalhe os atributos e métodos destas classes.

Matplotlib:

Em suma matplotlib é uma libraria que permite gerar variados tipos gráficos através de uma interface orientada a objetos ou por um conjunto de funções. Documentação oficial em https://matplotlib.org/contents.html

Instalação da libraria através do pip (package manager do python) com o comando na CMD ou bash:

```
pip install matplotlib
```

No nosso programa foi importado o módulo pyplot que permite imprimir gráficos simples e interativos.

DOCUMENTAÇÃO

Classe Texto():

Para instanciar: Texto(str), onde *str* é uma string da notícia.

Atributos da Classe Texto():

i. original: onde é guardado o texto original que foi submetido
ii. lista: onde é guardada uma lista das palavras do texto (depois de limpo)
iii. objetosPalavras: lista de apontadores para os objetos Palavra()

Métodos da Classe Texto():

i. limparNoticia(str): Processa a string com a notícia, passando tudo para carateres minúsculos e separando as palavras. Verifica se cada palavra é alfabética, tem uma dimensão maior que 1 (para excluir "a", "e", "o") e incluí as palavras com hífen. Devolve uma lista com as palavras que passaram as condições mencionadas.
ii. criarPalavra(): Cria os objetos Palavra() com as palavras únicas e devolve uma lista com os apontadores para cada um dos objetos que criou. Exclui as Conjunções definidas. Passa para cada objeto Palavra, a própria palavra e a frequência com que esta aparece na notícia original.

    Lista de conjunções definidas para excluir:

    ['ao', 'um', 'mas', 'nem', 'já', 'ou', 'ora', 'que', 'quer', 'pois', 'por', 'de', 'da', 'do', 'se', 'para', 'as', 'os', 'até', 'em', 'no', 'na', 'nos', 'nas', 'às']

iii. devolveObjetos(): Devolve uma lista com os apontadores para os objetos Palavra.
iv. top(): Devolve uma lista de apontadores para os dez objetos com maior frequência.

    É feito um arranjo (sort) da lista com os apontares dos objetos Palavra por ordem decrescente, através de uma função lambda que verifica qual é o valor da frequência dentro de cada Palavra. Uma função lambda é uma pequena função anónima.

    Devolve uma fatia (slice) com apontadores para os objetos Palavra que têm a maior frequência.

v. hist(): Gera um histograma com as 10 palavras mais utilizadas e respetiva frequência. É utilizada a livraria *matplotlib* para gerar o gráfico. As labels são rodadas para ficar na vertical.

Classe Palavra():

Para instanciar: Palavra(str, int), onde *str* é a string da palavra, e *int* o valor da frequência.

Atributos da Classe Palavra():

i. palavra: onde é guardado o texto original que foi submetido.
ii. freq: onde é guardada uma lista das palavras do texto (depois de limpo).

Função main():

Função que simula um menu e gere as operações que o utilizador definir.

Opções disponíveis

1- Introduzir uma notícia.
2- Imprimir lista de todas as palavras únicas.
3- Obter Histograma com as 10 palavras mais utilizadas.
4- Sair do programa.

O programa corre até o utilizador escolher a opção 4. Se for introduzida uma opção diferente o programa solicita nova opção.