Trabalho 1

Regressar a: Semana 7 (15/11...

O objetivo deste trabalho é desenvolver um programa em Python 3 que apresente dados estatísticos de número de ocorrência de palavras para um dado texto proveniente de um ficheiro. O texto pode ser um livro e, a partir deste programa, poderia ser analisado de acordo com a ocorrência de palavras (por exemplo, gerar nuvem de palavras, calcular a similaridade de textos, classificar a autoria de uma parte de texto, ...). Para testar, utilize livros no formato txt. Alguns exemplares podem ser baixados gratuitamente do Projeto Gutenberg: https://www.gutenberg.org/ebooks/.

Os requisitos deste trabalho são:

1) O programa deve obter texto de um ficheiro;

- O nome do ficheiro deve ser dado pelo usuário.

2) O programa deve separar as palavras de modo a remover elementos de pontuação (pontos, vírgulas, exclamações, interrogações, travessões, ...);

- A separação das palavras não deve ser case-sensitive (exemplo: 'casa' e 'Casa' devem ser tratadas como iguais).

- Não há necessidade de separar radicais ou remover prefixos ou sufixos das palavras (exemplo: 'casa' e 'casas' pode ser consideradas como palavras diferentes).

- Algumas facilidades podem ser assumidas, desde que devidamente indicadas por comentário no código ou alguma outra documentação do trabalho. Por exemplo, o(a) aluno(a) pode manter o texto de identificação do projeto no início e fim do ficheiro, mesmo que este texto não faça parte do texto do autor (assume-se, a princípio, que isto não trará graves problemas para a contagem de palavras de modo geral).

3) Como saída do programa, deve ser gerado um novo ficheiro contendo cada palavra e seu respectivo número de ocorrências (ou percentual de ocorrência).

4) Uma outra saída esperada do programa é a apresentação (print no console) das 20 palavras mais utilizadas no texto.

5) As partes importantes do código devem ser comentadas (descrever o que está sendo feito).

6) O programa deve ser organizado em módulos (funções e procedimentos).

Observações:

Haverá um decréscimo na nota (20%) para cada dia ou fração de atraso.

Alguns exemplares de livros em Português:

https://www.gutenberg.org/cache/epub/34719/pg34719.txt

https://www.gutenberg.org/files/62624/62624-0.txt

https://www.gutenberg.org/cache/epub/24401/pg24401.txt

https://www.gutenberg.org/cache/epub/54829/pg54829.txt

Alguns exemplares de livros em Inglês:

https://www.gutenberg.org/files/11/11-0.txt

https://www.gutenberg.org/files/1661/1661-0.txt

https://www.gutenberg.org/cache/epub/4078/pg4078.txt