Objetivos

Pretende-se o desenvolvimento de um programa em linguagem Java que analise textos em língua inglesa. Os textos são lidos a partir de ficheiros de texto. O resultado deverá ser uma **listagem de palavras por ordem alfabética**, com indicação do **número de ocorrências de cada uma** e das respetivas **linhas em que ocorre**. Deverá ser também apresentado o **número total de palavras** do texto. O programa deverá ainda registar o **número de ocorrências de cada um dos caracteres alfabéticos**, registando também a **frequência relativa de cada carácter** (nº ocorrências / nº total de caracteres * 100%). O programa deverá permitir obter uma lista ordenada de forma decrescente dos n caracteres que ocorrem mais vezes no texto, sendo n um número inserido pelo utilizador.

Os textos podem ter caracteres de pontuação, espaços ou outros que deverão ser ignorados. Não deve ser feita distinção entre letras maiúsculas e letras minúsculas.

O programa deverá permitir a exportação dos resultados para um ficheiro em formato csv.

Para a contagem de palavras programa deverá usar uma **árvore binária** como estrutura de dados principal. As linhas onde cada palavra ocorre serão mantidas através de **listas ligadas**.

Os números de ocorrências dos caracteres encontrados no texto e as respetivas frequências relativas são guardados numa **fila com prioridade** implementada com uma árvore binária do tipo heap (*max binary heap*), representada implicitamente.

Não podem ser usadas coleções do Java.

Critérios de Avaliação

Será considerado na avaliação a qualidade dos algoritmos escolhidos e a forma de escrita do código-fonte (uso de comentários, organização do código, legibilidade, etc).

O trabalho pode ser realizado individualmente ou em grupo (máximo de 2 alunos) e será complementado com uma apresentação oral final (em Powerpoint ou equivalente, com peso de 20% na nota) para explicar o seu funcionamento, as estruturas de dados usadas e uma demonstração do programa.

A apresentação deverá incluir uma reflexão sobre os resultados obtidos com a contagem de caracteres que deverão ser comparados com as frequências relativas encontradas na língua inglesa.

(ver English letter frequency: https://en.wikipedia.org/wiki/Letter_frequency).

O trabalho será valorizado se tiver uma interface gráfica.

Deverá ser submetido o código fonte do trabalho (ficheiros .java, preferencialmente o projeto completo) e a apresentação em Powerpoint.

Exemplo:

Se o texto for:

```
There are two ways of constructing a software design:
One way is to make it so simple that there are obviously no deficiencies,
and the other way is to make it so complicated that there are no obvious
deficiencies. (*)
```

o programa deverá apresentar os resultados seguintes:

```
Word count:
The word "a" occurs one time in lines 1
...
The word “are” occurs three times in lines 1, 2, 3
The word “way” occurs two times in lines 2, 3
...
Number of words: 39

Letter frequency:
e: 23 (fr=13,3%)
o: 16 (fr=9,2%)

a: 13 (fr=7,5%)
...
```

(*) C. A. R. Hoare