ESCOLA SUPERIOR DE TECNOLOGIA IPCA

Trabalho Prático 1

Algoritmos e Estruturas de Dados 2

Licenciatura em Engenharia de Sistemas Informáticos

INSTRUÇÕES

1. A data de entrega deste trabalho prático é 29 de Março, 2020.
2. Só serão aceites trabalhos entregues através da atividade respectiva no Moodle.
3. Os trabalhos serão realizados em grupo de, no máximo, dois elementos.
4. O trabalho prático deverá ser acompanhado de um relatório em PDF.
5. De acordo com o artigo 16º do Regulamento de Inscrição, Aprovação e Passagem de Ano da escola, "*A prática ou a tentativa de prática de qualquer fraude acarreta a anulação da prova em que tenha lugar, mediante decisão do docente e constitui infração disciplinar grave, sem prejuízo da responsabilidade civil ou criminal que ao caso couber.*"

O Processamento de Linguagem Natural, uma sub-área da Inteligência Artificial, dedica-se ao estudo de métodos computacionais para o estudo das línguas humanas, ditas naturais. Um dos processos típicos na preparação de um texto para o seu processamento é a etiquetação morfossintática em que, a cada palavra, se associa uma categoria morfológica (*ex.* verbo, substantivo, adjetivo) e também de informação morfológica adicional (*ex.* número ou género nos substantivos, ou tempo pessoa e número, nos verbos).

Neste trabalho prático pretende-se criar uma ferramenta que seja capaz de analisar um **qualquer** ficheiro já etiquetado com informação morfossintática. Este ficheiro é composto por várias linhas, uma para cada palavra ou sinal de pontuação. Cada linha é composta por várias colunas, sendo que a primeira corresponde à palavra da frase original, a segunda ao seu lema (raiz da palavra) a a terceira à sua análise morfossintática. A última coluna inclui a certeza da ferramenta em relação à análise realizada. As colunas estão separadas por um caracter de espaço.

Considere-se o seguinte exemplo, da anotação da frase "Now there is a dreadful thought!":

```
Now now RB 0.998694
there there RB 1
is be VBZ 1
a a DT 0.998827
dreadful dreadful JJ 1
thought thought NN 0.0950141
! ! Fat 1
```

A aplicação a desenvolver deve:

1. Inicialmente, ler todas as linhas e guardá-las numa ou mais estruturas de dados dinâmicas. As linhas em branco, e as linhas correspondentes a sinais ortográficos podem ser ignorados (ou seja, só deverão ser consideradas as linhas referentes a palavras).

   A estrutura de dados a usar deverá ser escolhida de modo a ser possível responder às seguintes questões.

2. Construa uma tabela de frequências absolutas, relativas e acumuladas, referente à categoria gramatical usada (terceira coluna). Apresente a tabela ordenada por ordem crescente de frequência absoluta.

3. Construa uma tabela de frequências absolutas, relativas e acumuladas, referente ao tamanho das palavras existentes no texto.

4. Para, cada tipo de análise morfológica (terceira coluna), calcule com base na medida de certeza de etiquetação (quarta coluna) a sua média aritmética e desvio padrão.

5. Calcule as seguintes medidas de localização e dispersão, relativas ao tamanho das palavras: média aritmética, mediana, moda e desvio padrão.

6. Com base nas frequências das palavras (o seu número de ocorrências), calcule os quartis e implemente uma função que, dada uma palavra indicada pelo utilizador, permita obter, o quartil a que pertence.

7. Obtenha os dados necessários para a construção de um histograma das probabilidades (quarta coluna). Defina o número de classes e as respetivas frequências a considerar.

Serão valorizados os seguintes aspetos:

1. Estruturas de dados aninhadas (por exemplo, listas de listas);

2. Estruturas de dados eficientes;

3. Código modular e estruturado;

4. Existência de uma Makefile;

5. Descrição clara e objetiva dos algoritmos implementados no relatório;

6. Uso de LaTeX no relatório.