Processamento/Teoria de Linguagens e Compilação
LCC (3ºano) + MEFis (1ºano)

Trabalho Prático nº 1 (ER + Filtros de Texto)

Ano letivo 23/24

Objetivos e Organização

Este trabalho prático tem como principais **objetivos**:

- aumentar a capacidade de escrever *Expressões Regulares (ER)* para descrição de *padrões de frases* dentro de textos;
- desenvolver, a partir de ER, sistematicamente *Processadores de Linguagens Regulares*, ou *Filtros de Texto (FT)*, que filtrem ou transformem textos com base no conceito de regras de produção *Condição-Ação*;
- utilizar o módulo 're' — com as suas funções de search(), split(), sub()—do Python para implementar os FT pedidos.

Para o efeito, esta folha contém 5 enunciados, dos quais deverá escolher pelo menos um.
Neste TP, que se pretende que seja resolvido rapidamente, aprecia-se a imaginação/criatividade dos grupos ao incluir outros processamentos!

Deve entregar a sua solução **até dia 15 de outubro**. O ficheiro único com o código Pyhton que resolve o problema, **devidamente comentado**, deve ter o nome '`plc23TP1grNGr.py`', com '`NGr`' substituído pelo número do grupo, e será submetido através do Bb até à data afixada.
O programa desenvolvido será apresentado aos membros da equipa docente, totalmente pronto e a funcionar e será defendido por todos os elementos do grupo, em data a marcar.

1 Processador de Pessoas listadas nos Róis de Confessados

Construa um programa Python para processar o texto contido no ficheiro '`processos.txt`', que transcreve o livro de *Róis de Confessados* existente no *Arquivo Distrital de Braga*, com o intuito de calcular frequências de alguns elementos, ou fazer outras operações, conforme solicitado abaixo.
Para isso comece por criar uma cópia do ficheiro eliminando os registos (as linhas) repetidos. Depois responda às seguintes alíneas:

a) Calcular a frequência de Processos por ano (primeiro elemento da data);

b) Calcular a frequência de Nomes Próprios (o primeiro em cada nome) e Apelidos (o último em cada nome) por séculos, analisando o nome do *Confessado*, do seu pai e da sua mãe;

c) Calcular a frequência de processos que são Recomendados por, pelo menos, um *Tio* (referido no campo Observações quando este está presente);

d) Identificar todos os Pais que tenham mais do que 1 Filho *Confessado*;

e) Imprimir o primeiro registo num formato JSON que julgue adequado.

Crie uma página HTML (ficheiro '`index.html`') para apresentar os resultados do seu processador.

2 Processador de Registos de Exames Médicos Desportivos

Neste exercício pretende-se trabalhar com um dataset gerado no âmbito do registo de exames médicos desportivos.
Construa, então, um programa Python para processar o ficheiro de texto `"emd.csv"` e produzir o solicitado nas alíneas seguintes:

a) Calcular as Idades extremas dos registos no dataset;

b) Calcular a distribuição por Género no total;

c) Calcular a distribuição por Modalidade em cada ano e no total, devendo apresentar as Modalidades por ordem alfabética;

d) Calcular a percentagem de Aptos e não aptos por ano;

e) Ajudar a normalizar as colunas do Nome, visto que o ficheiro original está inconsistente: quando o género é feminino cumpre-se o estabelecido no cabeçalho (primeiro o nome próprio e depois o apelido), mas quando o género é masculino essa ordem estabelecida no cabeçalho está trocada. Para isso, deve escrever num ficheiro de saída, em formato JSON, os pares de nomes masculinos que *julga* estar trocados (para que *à posteriori* o utilizador possa manualmente corrigir, se assim entender).

Crie uma página HTML (ficheiro '`index.html`') para apresentar os resultados do seu processador, no que respeita às primeiras 4 alíneas.

3 Processador de registos de Doenças Cardíacas

Neste exercício pretende-se trabalhar com um dataset gerado no âmbito do registo de doenças cardíacas.
Construa, então, um programa Python para processar o dataset `"myheart.csv"` e produzir o solicitado nas alíneas seguintes:

a) Calcular a percentagem da Doença no total da amostra e por Género (considere como total só os que estão doentes);

b) Calcular a distribuição da Doença por Escalões Etários. Considere os seguintes escalões: [30-34], [35-39], [40-44], ...;

c) Calcular a distribuição da Doença por níveis de colesterol. Considere um nível igual a um intervalo de 10 unidades, comece no limite inferior e crie os níveis necessários até abranger o limite superior;

d) Determinar se há alguma correlação significativa entre a Tensão ou o Batimento e a ocorrência de doença;

e) Criar gráficos para as distribuições, explorando o módulo `matplotlib`.

Crie uma página HTML (ficheiro '`index.html`') para apresentar os resultados do seu processador.

4 Processador de Cartas da Etiópia

Neste exercício pretende-se trabalhar com um dataset gerado no âmbito do registo de cartas do séc. XVI/XVII relativos à chegada das naus portuguesas à Etiópia e India.
Construa, então, um programa Python para processar o dataset `"cartasetopia.csv"` e produzir o solicitado nas alíneas seguintes:

a) Calcular a frequência de Cartas por ano e mês;

b) Calcular a distribuição de Cartas por Local;

c) Calcular a frequência com que cada Interveniente aparece como destinatário, remetente ou mencionado em todas as cartas;

d) Criar um ficheiro de saída em formato `JSON` agrupando Título e Resumo de cada carta. Inclua depois uma mecanismo que permita procurar sobre esse objeto `JSON` as cartas onde ocorra um dado termo;

e) Construir um Grafo de Conhecimentos que mostra os nomes (apelidos) dos Intervenientes que estão conectados uns aos outros como destinatário, remetente ou mencionado em cada carta. Para visualizar o grafo, descarregue os triplos para um ficheiro `DOT` que possa depois ser aberto por um visualizador como o que pode ser encontrado em `http://www.webgraphviz.com/`.

Crie uma página HTML (ficheiro '`index.html`') para apresentar os resultados do seu processador.

5 Processador de um Arquivo Musical

Neste exercício pretende-se trabalhar com um vasto arquivo sobre canções populares portuguesas criada há alguns anos atrás com o intuito de registar informação diversa sobre o nosso cancioneiro tradicional.

Construa, então, um programa Python para processar o dataset `"arq-son.txt"` e produzir o solicitado nas alíneas seguintes:

a) Calcular a frequência de registos por Província e por Local (de preferência considerar apenas o Concelho quando referido após o lugar);

b) Calcular a percentagem de canções que têm pelo menos uma gravação `"mp3"`, indicando o título dessas canções;

c) Calcular a distribuição por instrumento musical;

d) Identificar todos os Musicos/cantores registados e calcular o número de vezes que são mencionados;

e) Construir um Grafo de Canções/Cantores que associa cada canção aos cantores/tocadores referidos no registo. Para visualizar o grafo, descarregue os triplos para uma ficheiro `DOT` que possa depois ser aberto por um visualizador como o que pode ser encontrado em `http://www.webgraphviz.com/`.