Processamento de Linguagens
MiEI (3ºano)

Trabalho Prático nº 1 – parte A (GAWK)

Ano lectivo 16/17

1 Objectivos e Organização

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

- aumentar a experiência de uso do ambiente Linux e de algumas ferramentas de apoio à programação;
- aumentar a capacidade de escrever *Expressões Regulares (ER)* para descrição de *padrões de frases*;
- desenvolver, a partir de ERs, sistemática e automaticamente *Processadores de Linguagens Regulares*, que filtrem ou transformem textos;
- utilizar o sistema de produção para *filtragem de texto* GAWK.

Para o efeito, esta folha contém 4 enunciados, dos quais deverá resolver um escolhido em função do número do grupo (NGr) usando a fórmula $exe = (NGr \ \% \ 4) \ + \ 1$.
Neste 1º TP que se pretende que seja resolvido rapidamente (1 semana), os resultados pedidos são simples e curtos. Aprecia-se a imaginação/criatividade dos grupos ao incluir outros processamentos!

Deve entregar a sua solução **até Domingo dia 12 de Março**. O ficheiro com o relatório e a solução deve ter o nome "`pl16TP1aGrNN-`– *em breve serão dadas indicações precisas sobre a forma de submissão.*

O programa desenvolvido será apresentado aos membros da equipa docente, totalmente pronto e a funcionar (acompanhado do respectivo relatório de desenvolvimento) e será defendido por todos os elementos do grupo, em data a marcar.
O **relatório** a elaborar, deve ser claro e, além do respectivo enunciado, da descrição do problema, das decisões que lideraram o desenho da solução e sua implementação (incluir a especificação GAWK), deverá conter exemplos de utilização (textos fontes diversos e respectivo resultado produzido). Como é de tradição, o relatório será escrito em LaTeX.

2 Enunciados

Para sistematizar o trabalho que se pede em cada uma das propostas seguintes, considere que deve, em qualquer um dos casos, realizar a seguinte lista de tarefas:

1. Especificar os padrões de frases que quer encontrar no texto-fonte, através de ERs.
2. Identificar as acções semânticas a realizar como reacção ao reconhecimento de cada um desses padrões.
3. Identificar as Estruturas de Dados globais que possa eventualmente precisar para armazenar temporariamente a informação que vai extraindo do texto-fonte ou que vai construindo à medida que o processamento avança.
4. Desenvolver um Filtro de Texto para fazer o reconhecimento dos padrões identificados e proceder à transformação pretendida, com recurso ao Sistema de Produção GAWK.

2.1 Processador de transações da Via Verde

A Via Verde envia a cada um dos seus utentes um extracto mensal no formato XML como se exemplifica no ficheiro anexo `viaverde.xml`.
Depois de analisar com cuidado o formato desse ficheiro anexo, pretende-se que desenvolva um Processador de Texto com o GAWK para ler um extrato mensal da Via Verde e:

a) calcular o número de 'entradas' em cada dia do mês.

b) escrever a lista de locais de 'saída'

c) calcular o total gasto no mês.

d) calcular o total gasto no mês apenas em 'parques'.

2.2 Album Fotográfico em HTML

No âmbito do núcleo português do Museu da Pessoa (`http://npmp.epl.di.uminho.pt/`), dispomos de uma vasta coleção de fotografias (em formato JPEG ou PNG) e de um ficheiro anotado em XML com meta-informação sobre cada uma. Esse ficheiro `legenda.xml`, em anexo, identifica o ficheiro da foto, o local e a data onde foi tirada, e, entre outras coisas mais, **quem** são as pessoas fotografadas.
Neste momento pretende-se criar uma espécie de album em HTML que tenha uma lista de pessoas fotografadas e para cada elemento da lista permita ver a respetiva foto, ou de imediato por baixo do título usando o comando IMG

```
<LI><b>NomeDaPessoa</b></LI>
    <center> <img  src="NomeDoFicheiro.jpg"/> </center>
```

ou usando uma ancora como hiper-link que associe o nome do fotografado ao ficheiro, do tipo

```
<A HREF="NomeDoFicheiro.jpg"/>NomeDaPessoa</A>
```

Assim, pretende-se que desenvolva um Processador de Texto com o GAWK para ler um ficheiro de meta-informação do tipo do legenda.xml e gerar um ficheiro HTML com o corpo do desejado album.
Além disso o seu processador deve listar todos os locais fotografados, sem repetições.

2.3 Autores Musicais

Além da coleção de entrevista e fotografias do npMP, o Professor José João Almeida tem uma diretoria (de nome `musica`, que é anexada em formato ZIP) com dezenas de ficheiros de extensão `'.lyr'` que contêm a letra de canções famosas precedidas de 2 ou mais campos de meta-informação (1 por linha) com o título da canção, o autor da letra (pode ser 1 ou mais pessoas), o cantor, etc. Uma linha em branco separa a meta-informação da letra. Podendo ainda ter em alguns casos um terceiro bloco (igualmente separado da letra por uma linha em branco) com a música escrita na notação midi.
Depois de analisar com cuidado o formato desse ficheiro anexo, pretende-se que desenvolva um Processador de Texto com o GAWK para ler todos os ficheiros `'.lyr'` da diretoria `musica` e:

a) calcular o total de *cantores* e a lista com seus nomes.

b) calcular o total de canções do mesmo *autor* (mesmo que em alguns casos sejam várias pessoas considere como único).

c) escrever o nome de cada *autor* seguido do título das suas canções; se mais do que uma, separadas por uma vírgula.

Nota: Para resolver adequadamente este exercício é fundamental saber que o GAWK pode ser invocado através do comando

```
>  gawk -f programa.gawk *
```

e neste caso o programa será aplicado a todos os ficheiros da diretoria atual como se estivesse a processar um só ficheiro. Assim sendo, o bloco BEGIN e o bloco END só são executados uma vez, e as variáveis são todas globais conservando os valores entre processamento de ficheiros.

2.4 Dicionauro

Ainda na coleção JJ, já citada, de arquivos variados existe uma outra diretoria designada `Dicionauro` (ZIP em anexo) que contém vários ficheiros de extensão `'.txt'` com entradas em Português de termos de um Thesaurus. Cada termo ocupa a 1ª linha de um bloco de texto, linha essa que é iniciada pela sigla **'PT'**. depois dessa virá um número variado de linhas com a tradução desse termo para línguas diferentes, a categoria gramatical, exemplos de uso, a área ou domínio do saber, as partes que compõem esse objeto, uma possível definição (linha iniciada com a sigla **'Def'**.
As entradas do Thesaurus dentro de cada ficheiro estão separadas entre si por uma linha em branco.
Depois de analisar com cuidado o formato desse ficheiro da diretoria anexa, pretende-se que desenvolva um Processador de Texto com o GAWK para ler todos os ficheiros `'.txt'` da diretoria `Dicionauro` e:

a) criar um índice em HTML com todos os termos de entrada no Thesaurus e, sempre que exista, a respetiva definição (na linha de baixo) e categoria gramatical.

b) escrever lista de todos os domínios diferentes referidos e o número de entradas em cada domínio.

Para resolver simples e adequadamente este exercício, tome em consideração a **Nota** do exercício anterior.