Processamento de Linguagens
MiEI (3ºano)

Trabalho Prático nº 2 (GAWK)

Ano lectivo 18/19

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 5 enunciados, dos quais deverá resolver um escolhido em função do número do grupo (NGr) usando a fórmula $exe = (NGr \ \% \ 5) \ + \ 1$.
Neste 2º TP, que se pretende que seja resolvido em pouco tempo, aprecia-se a imaginação/criatividade dos grupos ao incluir outros processamentos pra além dos pedidos!
Deve entregar a sua solução **até FIXME**
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 Processos de Emigração

Analise com todo o cuidado o ficheiro `natura.di.uminho.pt/~jj/pl-19/TP2/emigra.csv` o qual contém informação sobre processos de pedido de passaporte para emigrar, registados no início do sec. XX.
Construa então um ou mais programas Awk que processem esse arquivo de modo a:

a) contar o número de processos registados por Concelho e Freguesia.

b) calcular a frequência de processos por ano e relacionar com os concelhos.

c) estudar a ocorrência de nomes próprios (não considere só os requerentes, mas considere também seus parentes).

d) desenhar um grafo (em DOT) que a partir dos requerentes os relacione com os pais e o cônjuge.

2.2 Processador de Processos de Formação

Analise com todo o cuidado o ficheiro `natura.di.uminho.pt/~jj/pl-19/TP2/formacao.csv` o qual contém informação detalhada sobre a criação e gestão de cursos de formação profissional.
Construa então um ou mais programas Awk que processem esse arquivo de modo a:

a) limpar o ficheiro dado (criar um segundo ficheiro pre-processado) retirando linhas vazias extra e linhas cujos campos estão todos vazios; sempre que o `'Estado'` esteja vazio, esse campo deve tomar o valor `'NIL'`.
Este ficheiro pre-processado deve ser usado nas alíneas seguintes!

b) contar o número de registos com apresentam um `Codigo` numérico e mostrar para esses, num formato legível, o dito código da ação de formação, o seu título, descrição e notas.

c) identificar os tipos diferentes e calcular o número de processos por tipo.

d) desenhar um grafo (em DOT) que relacione cada ação de formação (identificada pelo seu código) com todos os Diplomas jurídico-administrativos usados.

2.3 Processador de Cartas setecentistas da Etiópia

Analise com todo o cuidado o ficheiro `natura.di.uminho.pt/~jj/pl-19/TP2/cartasetiopia.csv` o qual contém informação diversa sobre uma coleção de cartas trocadas nos anos de mil seiscentos aquando da viagem dos navegadores portuguesas à Etiópia.
Construa então um ou mais programas Awk que processem esse arquivo de modo a:

a) contar o número de cartas por local (considere o local NIL quando se desconhece) relacionando-as com o ano de escrita.

b) criar um index HTML com todos os anos, em que cada ano deve ligar a outra página HTML onde conste, para cada carta desse ano, o título da carta e o seu resumo.

c) mostrar a lista das cartas—cada uma identificada pelo número, devidamente associada (em pares num-nome) aos Apelidos das pessoas envolvidas no assunto relatado.

d) desenhar um grafo (em DOT) que relacione cada autor (identificado pelo seu nome) com o destinatário (também identificado pelo nome).

2.4 Processador de CETEMPúblico

Ver ficheiro `natura.di.uminho.pt/~jj/pl-18/TP1/CORPORA1/`. Genericamente os corpora agrupam (grandes quantidade) de textos aos quais adicionam informação de anotação frásica (parágrafos(`<p>`), frases(`<s>`), multi-word-expressions (`<mwe>`)), e morfossintática (exemplo: lema, categoria gramatical, etc).
O formato CETEMPúblico, usa tags xml para a anotação frásica, e colunas separadas por `'tab'` para a informação morfossintática de cada palavra. As colunas presentes no ficheiro indicado e relevantes para este trabalho são:
`palavra, secção, semestre, lema, pos(part of speech), tempoVerbal-modo, num-pessoa, Género, árvore, ....`
Construa um ou mais programas Awk que processem o CETEMPúblico de modo a:

a) contar o número de Extratos, Paráfrafos e Frases.

b) extrair a lista das multi-word-expressions e respectivo número de ocorrências.

c) calcular a lista dos verbos PT: (Lema, para palavras com pos=V) e respectivo número de ocorrências.

d) determinar o dicionário implícito no córpora – calcule a lista das palavras associando-lhes os possíveis (lema, pos)

2.5 Processador de textos preanotados com Freeling

Ver ficheiro `natura.di.uminho.pt/~jj/pl-18/TP1/CORPORA2/`. Genericamente os corpora agrupam (grandes quantidade) de textos aos quais adicionam informação de anotação frásica e morfossintática (exemplo: lema, categoria gramatical, etc).

O formato Freeling, separa extratos com uma linha em branco, e separa colunas por espaços para a informação morfossintática de cada palavra.

As colunas presentes no ficheiro indicado e relevantes para este trabalho são:
`num, palavra, lema, pos-tag, pos(part of speech), features, ..., árvore`.

Analise alguns extractos e depois construa um ou mais programas Awk que processem corpora Freeling de modo a:

a) contar o número de Extratos.

b) calcular a lista dos personagens do Harry Potter (nomes próprios) e respectivo número de ocorrências.

c) calcular a lista dos verbos, substantivos, adjectivos e advérbios PT e criar um ficheiro HTML com cada uma destas listas.

d) determinar o dicionário implícito no córpora – lista contendo os lema, pos e palavras dele derivadas.