Processamento de Linguagens e Compiladores
LCC+MiEFis (3ºano)

Trabalho Prático nº 1 (Gawk)

Ano lectivo 17/18

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 $exe = (NGr \ \% \ 5) \ + \ 1$.
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 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 CETEMPúblico

Ver ficheiro `natura.di.uminho.pt/~jj/pl-18/TP1/CORPORA1/`. Genericamente os corpora agrupam (grandes quantidades) 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 são: palavra, secção, semestre, lema, pos(part of speech), tempoVerbal-modo, num-pessoa, Género, árvore, etc.

Considere o seguinte extracto de CetemPublico:

```
<ext n=1668 sec=pol sem=92a>
<p par=ext1668-pol-92a-1>
<s>
<mwe pos=ADV>
Ontem   pol 92a ontem   ADV 0   0   0   ADVL>   0   0   0
de  pol 92a de  PRP 0   0   0   N<  0   0   0
manhã   pol 92a manhã   N   0   S   F   P<  0   0   0
</mwe>
,   pol 92a ,   PU  0   0   0   PONT    0   0   0
reuniram    pol 92a reunir  V   PS/MQP_IND  3P  0   FMV 0   0   0
na  pol 92a em+o    PRP+DET_artd    0   S   F   <ADVL+>N    0   0   0
sede    pol 92a sede    N   0   S   F   P<  0   0   0
do  pol 92a de+o    PRP+DET_artd    0   S   M   N<+>N   0   0   0
MDP pol 92a MDP PROP    0   S   M   P<  0   0   0
,   pol 92a ,   PU  0   0   0   PONT    0   0   0
na  pol 92a em+o    PRP+DET_artd    0   S   F   <ADVL+>N    0   0   0
Damaia  pol 92a Damaia  PROP    0   S   F   P<  0   0   0
,   pol 92a ,   PU  0   0   0   PONT    0   0   0
as  pol 92a o   DET_artd    0   P   F   >N  0   0   0
comissões   pol 92a comissão    N   0   P   F   <ACC    0   0   0
de  pol 92a de  PRP 0   0   0   N<  0   0   0
redacção    pol 92a redacção    N   0   S   F   P<  0   0   0
e   pol 92a e   KC  0   0   0   CO  0   0   0
de  pol 92a de  PRP 0   0   0   N<  0   0   0
organização pol 92a organização N   0   S   F   P<  0   0   0
do  pol 92a de+o    PRP+DET_artd    0   S   M   N<+>N   0   0   0
movimento   pol 92a movimento   N   0   S   M   P<  0   0   0
.   pol 92a .   PU  0   0   0   PONT    0   0   0
</s>
```

Analise alguns extractos.

Construa um ou mais programas Awk que processem o CETEMPúblico de modo a:

- contar o número de Extratos, Parágrafos e Frases.
- extrair a lista das multi-word-expressions e respectivo número de ocorrências.
- calcule a lista dos verbos PT: (Lema, para palavras com pos=V) e respectivo número de ocorrências.
- determinar o dicionário implícito no corpora – calcule a lista das palavras associando-lhes os possíveis (lema, pos)

2.2 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, usa separa extratos com uma linha em branco, e usa colunas separadas por espaços para a informação morfossintática de cada palavra. As colunas presentes são: num, palavra, lema, pos-tag, pos(part of speech), features, ..., árvore.

Considere o seguinte extracto de texto anotado com freeling:

```
1  1  Além_de       além_de       SP      SP  pos=adposition|type=preposition     - - (S:0(grup-sp:1(prep:1)
2  2  Sócrates      sócrates      NP00000 NP  pos=noun|type=proper                - - (sn:2(grup-nom-ms:2(w-ms:2))))
3  3  ,             ,             Fc      Fc  pos=punctuation|type=comma          - - -
4  4  estão         estar         VMIP3P0 VMI pos=verb|type=main|mood=indicative  - - (grup-verb:4(verb:4))
5  5  acusados      acusar        VMP00PM VMP pos=verb|type=main|mood=pastpartic  - - (s-adj:5(s-a-mp:5(parti-mp:5)))
6  6  o             o             DA0MS0  DA  pos=determiner|type=article|gen=ma  - - (sn:7(espec-ms:6(j-ms:6))
7  7  empresário    empresário    NCMS000 NC  pos=noun|type=common|gen=masculine  - - (grup-nom-ms:7(n-ms:7)
8  8  Carlos_Santos carlos_santos_NP00000 NP  pos=noun|type=proper                - - (w-ms:8)))
9  9  ,             ,             Fc      Fc  pos=punctuation|type=comma          - - -
10 10 amigo         amigo         NCMS000 NC  pos=noun|type=common|gen=masculine  - - (sn:10(grup-nom-ms:10(n-ms:10)))
11 11 de            de            SP      SP  pos=adposition|type=preposition     - - (sp-de:11
12 12 longa         longo         AQ0FS00 AQ  pos=adjective|type=qualificative|g  - - (sn:13(grup-nom-fs:13(s-a-fs:12(a-fs:12
13 13 data          data          NCFS000 NC  pos=noun|type=common|gen=feminine|  - - (grup-nom-fs:13(n-fs:13)))))
14 14 e             e             CC      CC  pos=conjunction|type=coordinating   - - (coord:14)
15 15 alegado       alegado       NCMS000 NC  pos=noun|type=common|gen=masculine  - - (sn:15(grup-nom-ms:15(n-ms:15)))
16 16 testa         testo         AQ0FS00 AQ  pos=adjective|type=qualificative|g  - - (s-adj:16(s-a-fs:16(a-fs:16)))
17 17 de            de            SP      SP  pos=adposition|type=preposition     - - (sp-de:17
18 18 ferro         ferro         NCMS000 NC  pos=noun|type=common|gen=masculine  - - (sn:18(grup-nom-ms:18(n-ms:18))))
19 19 de            de            SP      SP  pos=adposition|type=preposition     - - (sp-de:19
20 20 o             o             DA0MS0  DA  pos=determiner|type=article|gen=ma  - - (sn:22(espec-ms:20(j-ms:20))
21 21 antigo        antigo        AQ0MS00 AQ  pos=adjective|type=qualificative|g  - - (grup-nom-ms:22(s-a-ms:21(a-ms:21))
22 22 primeiro-min  primeiro-min  NCMS000 NC  pos=noun|type=common|gen=masculine  - - (grup-nom-ms:22(n-ms:22)))))
23 23 ,             ,             Fc      Fc  pos=punctuation|type=comma          - - -
24 24 o             o             DA0MS0  DA  pos=determiner|type=article|gen=ma  - - (sn:25(espec-ms:24(j-ms:24))
25 25 ex            ex            NCCN000 NC  pos=noun|type=common|gen=common|nu  - - (grup-nom-ms:25(n-ms:25)))
26 26 -             -             Fg      Fg  pos=punctuation|type=hyphen         - - (F-no-c:26)
27 27 presidente    presidente    NCMS000 NC  pos=noun|type=common|gen=masculine  - - (sn:27(grup-nom-ms:27(n-ms:27)))
28 28 de            de            SP      SP  pos=adposition|type=preposition     - - (sp-de:28
```

Analise alguns extractos.

Construa um ou mais programas Awk que processem corpora freeling de modo a:

- contar o número de Extratos.
- calcule a lista dos personagens do Harry Potter (nomes próprios) e respectivo número de ocorrências.
- calcule a lista dos verbos, substantivos, adjectivos e advérbios PT: e crie um ficheiro com cada uma destas listas.
- determinar o dicionário implícito no corpora – lista contendo os lema, pos e palavras dele derivadas.

2.3 Processador / sincronizador de Legendas

Ver ficheiro `natura.di.uminho.pt/~jj/pl-18/TP1/SUBTITLES/`
Considere o seguinte extracto de legendas formato srt:

```
1
00:00:48,344 --> 00:00:49,500
Chamada: recebida

2
00:00:49,707 --> 00:00:53,128
-Está tudo pronto?
-Você não tinha de me substituir.

3
00:00:53,328 --> 00:00:56,014
Eu sei, mas quero fazer um turno.

4
00:01:06,485 --> 00:01:08,943
Você gosta dele, não?
Gosta de observá-lo.
```

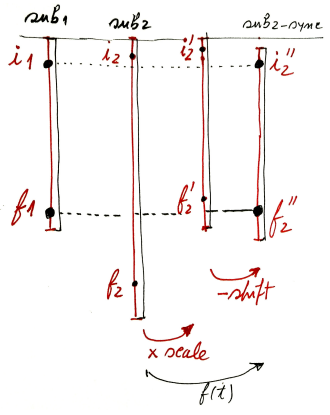
As linhas 1, 5, 10, 14 contêm os identificadores de legenda. As linhas 2, 6, 11, 15 contêm os tempos de início e desaparecimento da legenda. As legendas são separadas por linha em branco.
Considere ainda que dispomos legendas do mesmo filme em várias línguas, mas que frequentemente diferem no tempo inicial e duração do filme.

1. Construa um processador de srt que:
   - retire os identificadores de legenda.
   - coloque as legendas numa única linha juntando-as com "|"
   - marque com traço horizontal os intervalos com mais de 2 segundos de silêncio.

   ```
   00:00:53,328 --> 00:00:56,014 Eu sei, mas quero fazer um turno.
   00:00:56,014 --> 00:01:06,485 ================================
   00:01:06,485 --> 00:01:08,943 Você gosta dele, não?|Gosta de observá-lo.
   ```

2. Construa um sincronizador de legendas:
   (Exemplo de uso: `srtsync i1=12 i2=8 f1=1000 f2=900 sub1.srt sub2.srt > sub2-sync.srt`)
   que recalcule os tempos das legendas de `sub2.srt` de modo que as legendas com números 12 e 1000 de `sub1.srt` fiquem sincronizadas respectivamente com as legendas 8 e 900 de sub2.

   Para tal sugere-se que comece por calcular o factor de scaling e o necessário shift:




```
dur1=f1-i1
dur2=f2-i2
scale=dur1/dur2
shift=i2*scale -i1
f(t)=t*scale-shift
```

Para realizar este sincronizador, poderá ter que percorrer duas vezes o ficheiro sub2.srt. Um modo de o fazer é no início (BEGIN), juntar mais uma cópia do argumento de linha de comando sub2:

```
BEGIN { if(ARGC==7){ ARGC=8;  ARGV[7]=ARGV[6]}
        else       { print "usage: alinha i1=3 i2=6 f1=30 f2=60 a.srt b.srt\n"; }}
```

2.4 Processador de Thesaurus 1

Ver ficheiros em `natura.di.uminho.pt/~jj/pl-18/TP1/THE/` que pode corrigir/completar. Os ficheiros fornecidos `....mdic` descrevem numa sintaxe simples as entradas (triplos `termo1, rel, termo2`) de um Thesaurus que se pretende criar automaticamente.
Cada ficheiro mdict contem:

- comentários a ignorar (do símbolo cardinal, até ao fim da linha)
- directivas gerais:
  - `%dom: alimentação` – todos os termos definidos são do domínio *alimentação*; válido até nova indicação de novo domínio. Dom é uma relação e a sua inversa é voc(vocabulário).
  - `%inv: nt : bt` – indica que a relação *nt* (=narrow term), é a inversa *bt* (=broader term).
- tabelas de relações constituídas por uma linha indicadora de relações (começada por `%THE`), seguida de várias linhas com tuplos.

```
%inv:atravessa : é_atravessado_por
%inv:tem_como_instancia : iof
%THE : nt
bebida    : vinho | sumo
sobremesa : pudim | fruta | leite creme |baba de camelo

%THE<rio : nasce_em : atravessa < localidade : foz
rio Cávado : serra do Larouco: Montalegre|Barcelos : Esposende
```

Note que:

- Cada linha tem 1 ou mais termos.
- Os termos são separados por ':' daqueles com que se relacionam.
- A relação entre o termo da coluna 1 e os termos da coluna n é a indicada na posição n da linha `%THE` (nt, instância, iof, etc.).
- Quando há vários termos com a mesma relação, eles podem ser agrupados com '—'.
- Um campo do cabeçalho pode conter `< classe`, indicando que todos os elementos dessa coluna são instâncias da *classe*. Exemplo `(rio Cávado, iof, rio)(Montalegre, iof, localidade)`

Exemplo de alguns triplos decorrentes do exemplo anterior:

```
(bebida, nt, vinho)(bebida, nt, sumo) (rio Cávado, iof, rio)
(rio, tem_como_instancia, rio Cávado)
(rio Cávado, nasce_em, serra do Larouco) (rio Cávado, atravessa, Montalegre)
(Montalegre, é_atravessado_por, rio Cávado)
```

Escreva programas Gawk que dado um ou mais mdic:

1. determine a lista dos domínios e das relações usadas.
2. mostre os triplos expandidos correspondentes (um triplo por linha)
3. mostre a informação contidas nos triplos, agrupadas pelo termo1 (formato Thesaurus ISO) – Exemplo de um extrato :

```
rio Cávado
iof:   rio
nasce_em:    serra do Larouco
atravessa:   Montalegre
atravessa:   Barcelos
foz:         Esposende

Barcelos
é_atravessado_por:   rio Cávado
iof: localidade
```

2.5 Processador de thesaurus 2

Usando o mesmo formato e os mesmos ficheiros mdic descritos no enunciado anterior:
Escreva programas Gawk que dado um ou mais mdic:

1. determine a lista dos domínios e das relações usadas.
2. mostre os triplos expandidos correspondentes (um triplo por linha)
3. Construa um conjunto de páginas HTML (uma página por cada termo1) em que os termos2 hiperliguem às correspondentes páginas.