ESCOLA SUPERIOR DE TECNOLOGIA IPCA

Trabalho Prático I

Processamento de Linguagens — 2019/2020

Licenciatura em Engenharia de Sistemas Informáticos

Versão 0.1 — October 11, 2019— Avaliação Contínua

Instruções

- Estes três enunciados correspondem ao **primeiro** trabalho prático de Processamento de Linguagens, regimes diurno e pós-laboral, no ano letivo 2019/2020, para a **avaliação contínua**;
- O trabalho prático será realizado em Grupo com um **máximo de 3 alunos**;
- A data limite para a entrega do primeiro trabalho prático é o **dia 28 de Outubro**, e a entrega será aceite apenas usando o formulário correspondente disponível no Moodle;
- As apresentações dos trabalhos práticos é feita por **todos os elementos do grupo** em data a marcar na **primeira semana de Novembro**;
- Para além da implementação em C + flex do projeto, deverá ser preparado um pequeno relatório que explique de que forma o enunciado foi interpretado, e quais as decisões tomadas na sua implementação.
- Qualquer detalhe que não esteja claro no enunciado deve ser extrapolado pelos alunos, optando pela interpretação que lhes parecer mais lógica/funcional.
- O **enunciado a realizar** por cada grupo é decidido realizando o resto da divisão inteira por três, da soma dos números de alunos do grupo. Por exemplo, considerando que o grupo é constituído pelos alunos número 13523 e 15324, cuja soma é 28847, então o enunciado a realizar será o 2 ($28847 = 3 \times 9615 + 2$).

0. IMDB

Pretende-se uma ferramenta que permita consultar a Base de Dados do International Movie DataBase, disponível em `https://my.pcloud.com/publink/show?code=XZI004kZKpzNtJCI447aLTY6CYvd6Qkp0OYy`. Este ficheiro é constituído por linhas que podem ser:

- vazias (apenas com espaços);
- comentários (iniciam com o símbolo `#`;
- dados, em formato tabular (separados por um carater de tabulação: `\t`.

Considere-se o seguinte exemplo:

```
### tconst  titleType  primaryTitle            originalTitle           isAdult startYear  endYear runtimeMinutes  genres
tt0000001   short      Carmencita              Carmencita              0       1894       \N      1               Documentary,Short
tt0000002   short      Le clown et ses chiens  Le clown et ses chiens  0       1892       \N      5               Animation,Short
tt0000003   short      Pauvre Pierrot          Pauvre Pierrot          0       1892       \N      4               Animation,Comedy,Romance
```

O comentário indica a correspondência a cada coluna. A primeira é o identificador de filme (inicia sempre por `tt`), o tipo de filme (curto, longo), o título principal, o título original, um booleano que indica se é um filme para adultos, duas colunas com informação do ano de início e término, a duração do filme em minutos, e os tipos de filmes (separados por vírgulas).

Do mesmo modo, existe outra secção correspondente a atores, de acordo com o seguinte exemplo:

```
###  nconst primaryName     birthYear  deathYear  primaryProfession               knownForTitles
nm0000001   Fred Astaire    1899       1987       soundtrack,actor,miscellaneous  tt0072308,tt0043044,tt0050419,tt0053137
nm0000002   Lauren Bacall   1924       2014       actress,soundtrack              tt0037382,tt0117057,tt0038355,tt0071877
nm0000003   Brigitte Bardot 1934       \N         actress,soundtrack,producer     tt0059956,tt0049189,tt0054452,tt0057345
```

Neste caso temos o identificador do ator (inicia sempre por `nm`), o nome do ator, ano de nascimento e morte, principal profissão, e lista dos quatro filmes mais conhecidos desse ator.

Pretende-se uma aplicação que seja capaz de ler toda esta informação, e apresentar no ecrã uma lista formatada com o nome de atores, ordenados alfabeticamente, bem como a lista dos filmes em que participaram. Por exemplo:

```
Brigitte Bardot
   - Viva Maria!
   - ...And God Created Woman
   - La Vérité
   - Contempt
Fred Astaire
   - The Towering Inferno
   - Three Little Words
   - Funny Face
   - On the Beach
```

1. Dicionário Aberto

Pretende-se uma ferramenta que permita processar um ficheiro XML com todas as entradas do Dicionário Aberto, disponível em `https://my.pcloud.com/publink/show?code=XZo004kZhbicrCbekU8DxEXelaFIz779rvKX`. Este ficheiro é constituído por várias entradas, como a que a seguir se apresenta:

```
<entry id="achafundar" ast="1">
<form>
<orth>Achafundar</orth>
</form>
<sense>
<gramGrp>v. t.</gramGrp>
<usg type="style">Pop.</usg>
<def>
Enterrar no lodo; meter no fundo da água.
</def>
</sense>
</entry>
```

Cada entrada inclui a palavra do dicionário (etiqueta `<orth>`), um conjunto de etiquetas de uso (`<usg...>`), bem como uma definição (etiqueta `<def>`). O que se pretende é uma lista ordenada, de tipos de uso, que inclua as palavras e definições desse tipo. Note que uma palavra pode pertencer a mais que um tipo.

Como exemplo de output:

```
ant:
- Alampião: O mesmo que _lampião_. Cf. B. Pereira, _Prosódia_, vb. _polymixus_.
- Anêspera: O mesmo que _nêspera_. Cf. B. Pereira, _Prosodia_, vb. _pytmena_.
- Arrabaça: Planta, o mesmo que _rabaça_.
...
Pop.
- Achafundar: Enterrar no logo; meter no fundo da água.
- Acarditar: O mesmo que _acreditar_.
- Alampião: O mesmo que _lampião_. Cf. B. Pereira, _Prosódia_, vb. _polymixus_.
...
```

2. WordNet

Pretende-se uma ferramenta que permita processar um *dump* da base de dados da WordNet Portuguesa, disponível em `https://my.pcloud.com/publink/show?code=XZk504kZjRQdCTnKFNXokAGjcDt96fd9628k`. A WordNet organiza termos (também conhecidos por *variantes*) em grupos (designados *synsets*). Dois termos pertencem ao mesmo grupo se são sinónimos[1].

Neste exemplo,

```
"VARIANT: ",alojamento,por-30-03259505-n
"VARIANT: ",andar,por-30-03259505-n
"VARIANT: ",casa,por-30-03259505-n
"VARIANT: ",chalé,por-30-03259505-n
"VARIANT: ",domicílio,por-30-03259505-n
"VARIANT: ",habitação,por-30-03259505-n
"VARIANT: ",lar,por-30-03259505-n
"VARIANT: ",morada,por-30-03259505-n
"VARIANT: ",moradia,por-30-03259505-n
"VARIANT: ",pousada,por-30-03259505-n
"VARIANT: ",residência,por-30-03259505-n
"VARIANT: ",vivenda,por-30-03259505-n
"SYNSET: ",por-30-03259505-n,"habitação que alguém está vivendo em
```

existem 12 variantes, em que todas pertencem ao mesmo grupo.

O resultado do processamento deve agrupar as variantes, e ordená-las alfabeticamente, como no exemplo seguinte:

```
por-30-03259505-n: habitação que alguém está vivendo em
- alojamento
- andar
- casa
- chalé
- domicílio
- habitação
- lar
- morada
- moradia
- pousada
- residência
- vivenda
```

[1] Note que o ficheiro disponibilizado tem erros, já que foi construido automaticamente.