

# Processamento de Linguagens – LEI (3ºano)

## Trabalho Prático nº 1 (FLex)

Ano lectivo 12/13

## 1 Objectivos e Organização

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

- aumentar a experiência de uso do ambiente **Linux**, da linguagem imperativa **C** (para codificação das estruturas de dados e respectivos algoritmos de manipulação), 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 *geradores de filtros/processadores de texto*, como o **Flex**

Para o efeito, esta folha contém vários enunciados, dos quais deverá resolver pelo menos um. 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 (3 alunos), 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 e a implementação, deverá conter exemplos de utilização (textos fontes diversos e respectivo resultado produzido). Como é de tradição, o relatório será escrito em **L<sup>A</sup>T<sub>E</sub>X**.

## 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 Processador de Texto para fazer o reconhecimento dos padrões identificados e proceder à transformação pretendida, com recurso ao Gerador **Flex**.

### 2.1 BibTeXPro — Um processador de BibTeX

BibTeX é uma ferramenta de formatação de citações bibliográficas em documentos **L<sup>A</sup>T<sub>E</sub>X**, criada com o objectivo de facilitar a separação da base de dados da bibliografia consultada da sua apresentação no fim do documento **L<sup>A</sup>T<sub>E</sub>X** em edição. BibTeX foi criada por Oren Patashnik e Leslie Lamport em 1985, tendo cada entrada nessa base de dados textual o aspecto que se ilustra a seguir

```

1  @InProceedings{CPBFH07e,
2      author = {Daniela da Cruz and Maria Joao Varanda Pereira
3                and Mario Beron and Ruben Fonseca and
4                Pedro Rangel Henriques},
5      title = {Comparing Generators for Language-based Tools},
6      booktitle = {Proceedings of the 1.st Conference on Compiler
7                   Related Technologies and Applications, CoRTA'07
8                   --- Universidade da Beira Interior, Portugal},
9      year = {2007},
10     editor = {},
11     month = {Jul},
12 }

```

De modo a familiarizar-se com o formato do BibTeX poderá consultar o ficheiro `lp.bib` disponível em <http://www.di.uminho.pt/~prh/lp.bib> e ainda a página oficial do formato referido (<http://www.bibtex.org/>), devendo para já saber que a primeira palavra (logo a seguir ao carácter "@") designa a categoria da referência (havendo em BibTeX pelo menos 14 diferentes).

As tarefas que deverá executar neste trabalho prático são:

- a) Analise o documento BibTeX referido acima e faça a contagem das categorias (`phdThesis`, `Misc`, `InProceeding`, etc.), que ocorrem no documento. No final, deverá produzir um documento em formato HTML com duas partes:
  - um índice de categorias;
  - uma listagem do conteúdo, ou seja, a lista de registos em HTML agrupados por categorias.
- b) Transforme o índice de categorias num índice a 2 níveis:
  - no primeiro nível, apresenta as categorias;
  - num segundo nível e associadas a cada categoria a chave de citação dos registos dessa categoria ( a chave de citação é o primeiro elemento do registo, no exemplo: *CPBFH07e*).
- c) Em complemento, gere um índice de autores, que mapeia cada autor nos respectivos registos; acrescente este novo índice à página HTML.
- d) Uma das linguagens descritivas mais utilizadas na Web Semântica é o *FOAF* ("*Friend Of A Friend*"); pesquise sobre esta linguagem e crie um ficheiro FOAF a partir do BibTeX tendo como base as seguintes relações:
  - co-autoria** Dois co-autores supostamente conhecem-se, assim podemos estabelecer uma relação entre eles; em FOAF utilizamos a propriedade *foaf:knows* para isso;
  - publicação** Em FOAF há também uma propriedade que relaciona um indivíduo com uma publicação, *foaf:publication*, utilize-a criando uma relação entre uma publicação e cada um dos seus autores;
  - outras** Explore o FOAF, há *viewers* que podem ser usados para navegar na informação; tente ver se há mais relações que se possam extrair do BibTeX: datas, instituições, etc.

## 2.2 Processador de ABC (formato textual para música)

Estude o formato ABC ([abcnotation.com](http://abcnotation.com), [abcplus.sourceforge.net](http://abcplus.sourceforge.net) – ABC Plus project) e instale algum dos seus processadores (Exemplo EasyABC, `abcmidi` ou `abcm2ps`)

Para uma documentação completa sobre o formato ABC, veja: [abcplus.sourceforge.net/abcplus\\_en-2012-03-30.zip](http://abcplus.sourceforge.net/abcplus_en-2012-03-30.zip)

Construa um pré-processador em Flex que "reconheça" um texto ABC e que faça pequenas alterações:

1. Conte quantos compassos tem a música;
2. Alterar o instrumento associado a uma voz;
3. Ponha o volume de uma voz a 100% e todas as outras a 20% (para ajudar a aprender essa voz);
4. Ponha o volume de uma voz a 20% e todas as outras a 100% (para ajudar a treinar essa voz);

Para detalhes e dúvidas → contactar `jj@di`

## 2.3 Pré-processador para LaTeX ou HTML

Desenvolver um documento em  $\text{\LaTeX}$  ou mesmo em HTML é uma actividade inteligente e intelectualmente interessante enquanto estruturante das ideias e sistematizante dos processos.

Porém o acto de editar o respectivo documento é por vezes fastidioso devido ao peso das marcas (as *tags*) que tem de ser inseridas para anotar o texto com indicações de forma, conteúdo ou formato.

Por isso apareceram editores sensíveis ao contexto que sabendo que se está a escrever um documento  $\text{\LaTeX}$  ou HTML nos facilitam a vida inserindo as ditas marcas, ou anotações.

Uma alternativa mais simples mas também muito frequente é permitir o uso de anotações mais leves e simples (até de preferência independentes do tipo de documento final) e de pois recorrer ao pré-processamento para substituir essa notação ligeira, abreviada, pelas marcas finais correctas.

Este é o caso do conhecido PPP<sup>1</sup>, desenvolvido há alguns anos por José Carlos Ramalho, ou mesmo do mais actual e não menos conhecido sistema Wiki para construção interactiva e via web de páginas HTML.

O que se lhe pede neste trabalho é que, depois de investigar os tais pré-processadores PPP, Wiki, ou outro análogo, especifique uma sua linguagem de anotação para abreviar a escrita de:

- **formatação:** *negrito*, *itálico*, *sublinhado*;
- vários **níveis de títulos**;
- **listas de tópicos (items)** *não-numerados*, *numerados* ou tipo *entradas de um dicionário*;
- **inclusão de imagens**;
- **inclusão e formatação de tabelas**;
- todos os outros que achar necessário ou a sua imaginação vislumbrar.

Deve, depois e recorrendo à ferramenta Flex, criar um processador que transforme a sua notação em  $\text{\LaTeX}$  ou HTML<sup>2</sup>.

## 2.4 Processamento da Wikipedia

A Wikipedia é actualmente uma base de conhecimento online disponível em várias línguas (entre as quais, o português). Como tal, e por estar disponível num formato passível de transformação (XML) muitas são as ferramentas que trabalham sobre a Wikipedia para produzir diferentes resultados.

---

<sup>1</sup>Consultar detalhes no manual da linguagem em <http://www3.di.uminho.pt/~jcr/TUTORIAL/tutorial-ppp.html>.

<sup>2</sup>O mais interessante mesmo é que fosse possível escolher a saída final no início do próprio texto a pré-processar.

Neste trabalho pretende-se que desenvolva em Flex um filtro para estruturar, num site HTML, um conjunto de informações extraídas da Wikipedia.

Para isso poderá descarregar a versão completa da wikipedia (exemplo wikipedia PT) ou exportar uma ou mais páginas usando o **Special Export** disponível em <http://pt.wikipedia.org/wiki/Especial:Exportar> (ou <http://en.wikipedia.org/wiki/Special:Export> para a versão inglesa).

## Variante1

Estude o seu conteúdo de forma a produzir os seguintes resultados para cada página existente no ficheiro XML obtido:

- Título;
- Autor da última revisão;
- Data da última revisão;
- N° de links internos (e explicitar quais);
- N° de links externos (e explicitar quais);
- N° de secções (e explicitar quais).

Estude ainda a possibilidade de gerar uma página HTML para cada página existente no ficheiro com o respectivo conteúdo original, colocando como cabeçalho as informações referidas acima (escolha um *layout* do seu gosto para a página HTML).

## Variante2

Construa dicionários bilingues através de:

Usando a wikipedia completa, extrair os termos verbetes (ponto de entrada) e a sua tradução para a língua destino pretendida; junte também outra informação que achar relevante (exemplo as categorias).

Construa um dicionário  $\text{\LaTeX}$  com essa informação.

## Variante3

Escolha algumas Infoboxes e centrando-se só nelas, atravesse a wikipedia extrai-as e gere um conjunto de páginas HTML com a informação relevante que conseguir extrair.

Exemplo processar info-boxes de tipo geográficas – coordenadas:

```
1 |{{geocoordenadas|41_43_07_N_8_18_34_W_type:adm2nd_region:PT-16|41° 43' N, 8° 19' W}}
ou Municípios
```

```
1 |{{Info/Município de Portugal|
2 |município           = Terras de Bouro
3 |imagem_brasão       = TBR.png
4 |imagem_geral        = Pacos_do_concelho.jpg
5 |imagem_localização  = LocalTerrasDeBouro.svg
6 |região              = [[Região Norte (Portugal)|Norte]]
7 |subregião           = [[Cávado (sub-região)|Cávado]]
8 |distrito            = [[Distrito de Braga|Braga]]
9 |província           = [[Minho (província)|Minho]]
10 |área                = 277.5
11 |população           = 7253
12 |...
```

## 2.5 Processador de Ficheiros de Genalogia (gedcom)

O formato GEDCOM é muito usado para intercambio de dados genealógicos.

```

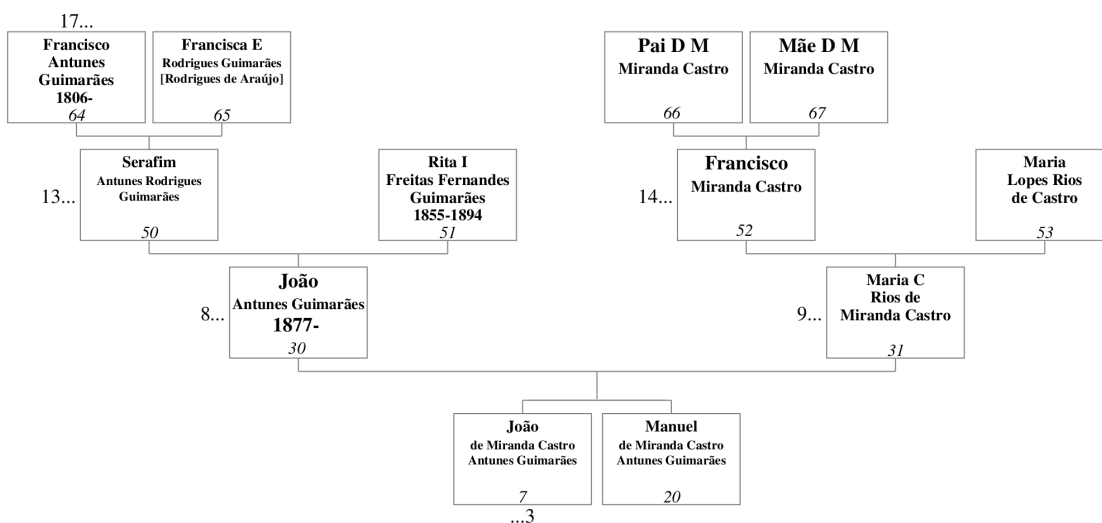
1 | 0 @I10 INDI
2 | 1 NAME Rui Jorge da Costa dos /Reis Borges/
3 | 2 GIVN Rui Jorge da Costa dos
4 | 2 SURN Reis Borges
5 | 1 SEX M
6 | 1 BIRT
7 | 2 DATE 27 APR 1944
8 | 2 PLAC S.Mamede Lisboa
9 | 1 FAMS @F20
10 | 1 FAMS @F30
11 | 1 FAMC @F10
12 | 0 @F20 FAM
13 | 1 HUSB @I10
14 | 1 WIFE @I40
15 | 1 CHIL @I70
16 | 1 MARR
17 | 2 PLAC Angra do Heroísmo Terceira Açores
18 | 0 @F30 FAM
19 | 1 HUSB @I10
20 | 1 WIFE @I50
21 | 1 MARR
22 | 2 PLAC Cascais

```

**Variante 1** Processe um ficheiro GEDCOM e gere uma página (notação Wiki? Zim? HTML?) com a informação de cada pessoa.

**Variante 2** Crie (em L<sup>A</sup>T<sub>E</sub>X?) diagramas de família (filhos+casal+pais+avós); ou crie um livro de família.

### 5. João Antunes Guimarães e Maria Cecília Rios de Miranda Castro



## 1. ASCENDÊNCIA PATERNA

