ESCOLA SUPERIOR DE TECNOLOGIA IPCA

Trabalho Prático I

Processamento de Linguagens — 2021/2022

Licenciatura em Engenharia de Sistemas Informáticos

Versão 0.2 — 4 de Novembro de 2021

Instruções

- A não entrega deste trabalho prático implica a reprovação à unidade curricular no ano letivo 2021/2022, **não sendo possível a sua realização em nenhuma época de avaliação**.
- O trabalho prático será realizado em Grupo com um **máximo d 2 alunos**;
- A data limite para a entrega do primeiro trabalho prático é o **dia 27 de Novembro**[a]. A submissão de trabalhos será feita apenas usando o formulário correspondente disponível no Moodle. Não serão aceites trabalhos práticos via e-mail;
- A apresentação do trabalho é feita por **todos os elementos do grupo**;
- Para além da implementação em Python + Ply 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.
- O enunciado é, propositadamente, vago. O objetivo é que cada grupo leia o enunciado, consulte a documentação fornecida, e **desenvolva a aplicação do modo** que lhe pareça que é mais **útil**.

[a] Data alterada, relativamente à indicada inicialmente, considerando a participação do docente no SAP Game Changing Games

Este enunciado apresenta três formatos textuais. Cada grupo deverá escolher um dos formatos, e criar uma ferramenta capaz de transformar o formato amrepsentado e ficheiros HTML e LaTeX. Para cada formato são apresentadas algumas funcionalidades extra que a ferramenta deve implementar.

A. NatTerm: Notação para terminologias

As terminologias são documentos semelhantes a dicionários, mas tipicamente multilingues, focados apenas numa área específica do conhecimento (como música, mineralogia ou matemática) e, muitas vezes com relações ontológicas. Estes recursos são orientados ao conceito e não à palavra. Isto significa que cada registo (ou entrada) da terminologia pode ter vários termos que representam um mesmo conceito.

Para agilizar a escrita deste tipo de recurso, foi definido um formato textual, inspirado noutros formatos, como o Markdown ou Wikimedia.

De seguida apresenta-se e explica-se o formato definido. A sua tarefa será a criação de um conversor que seja capaz de transformar uma terminologia nos formatos HTML e LaTeX.

Cada registo deste ficheiro é dividido por uma linha em branca, e contém termos numa ou mais línguas. Para cada língua (escrita em maiúscuas) é possível definir um conjunto de propriedades (com prefixo +). Também é possível definir proprieades para o registo, usando para isso um termo sem o prefixo. Finalmente, é possível definir texto multi-linha usando chavetas. Seguem-se alguns exemplos:

```
PT = arte
EN-GB = art
DE = Kunst

PT = som
EN-GB = sound
DE = Ton

PT = ritmo
EN-GB = rithm
DE = Rythmus

EN = house
ES = casa
+genero = Feminino
PT = casa
+genero = Feminino
+definicao = edifício próprio para habitação

PT = escritório
+genero = Masculino
+definicao = {compartimento onde se esreve e
onde se realizam também o mais variado tipo de trabalhos}
EN = office
ES = despacho
+genero = Masculino

PT = quarto
+genero = Masculino
+definicao = compartimento da casa, onde se dorme
EN = bedroom
ES = dormitorio/habitación
+observacoes = {Neste caso o género varia no espanhol, apesar da
  palvra mais adequada ser "habitación" (neste caso feminino), o
  uso de "dormitorio" também poderá surgir (neste caso masculino),
  daí não existir uma definição específica para o género.}

descricao = camisa desportiva de gola e manga curta ou comprida, de algodão.
PT = Pólo
+genero = masculino
+numero = singular
EN = Polo shirt
+genero = neutro
+numero = singular
```

B. CSV: Comma separated values

O CSV é um formato textual para a definição de estruturas tabulares. Neste ficheiro, que poe ser visto como uma grelha excel, é composto por várias linhas e por várias colunas. Neste formato, a linhas são separadas pela mudança de linha (*new line*) e as colunas por uma vírgula.

Este formato considera que a primeira linha corresponde ao cabeçalho da tabela. Por sua vez, as linhas que iniciam pelo símbolo cardinal (`#`), são considerados comentários.

O valor de uma coluna pode estar delimitado por aspas. Neste caso, o seu conteúdo é considerado o valor da célula, sem as aspas. Esta notação é especialmente útil porque permite a inclusão de vírgulas no meio do texto de uma célula.

A ferramenta a desenvolver deverá ser capaz de gerar documentos HTML e LaTeX com as tabelas obtidas do documento CSV. Também deverá ser possível indicar, na linha de comandos, quais as colunas que serão incluídas no *output* (usando para isso o título das colunas a apresentar).

```
Country Name,Capital,Currency,Official Languages,Head of Government
Afghanistan,Kabul,Afghani,"Dari Persian, Pashto",President - Amrullah Saleh
Albania,Tirane,Lek,Albanian,Prime Minister - Edi Rama
Algeria,Algiers,Dinar,"Arabic, Tamazight, French",Prime Minister - Aymen Benabderrahmane
Andorra,Andorra la Vella,Euro,Catalan,Prime Minister - Xavier Espot Zamora
Angola,Luanda,New Kwanza,Portuguese,President - João Lourenço
Antigua and Barbuda,Saint John's,East Caribbean dollar,English,Prime Minister - Gaston Browne
Argentina,Buenos Aires,Peso,Spanish,President - Alberto Fernández
Armenia,Yerevan,Dram,Armenian,President - Armen Sarksyan
Australia,Canberra,Australian dollar,English,Prime Minister - Scott Morrison
Austria,Vienna,Euro (formerly schilling),German,President - Alexander Van der Bellen
Azerbaijan,Baku,Manat,Azerbaijani,Prime Minister - Ali Asadov
The Bahamas,Nassau,Bahamian dollar,English,Prime Minister - Hubert Minnis
Bahrain,Manama,Bahrain dinar,Arabic,Prime Minister - Salman bin Hamad Al Khalifa
Bangladesh,Dhaka,Taka,Bangla,Prime Minister - Sheikh Hasina
Barbados,Bridgetown,Barbados dollar,English,Prime Minister - Mia Mottley
Belarus,Minsk,Belorussian ruble,"Belarusian, Russian",President - Alexander Lukashenko
Belgium,Brussels,Euro (formerly Belgian franc),"Dutch, French, German",Prime Minister - Alexander De Croo
Belize,Belmopan,Belize dollar,English,Prime Minister - Johnny Briceño
Benin,Porto-Novo,CFA Franc,French,President - Patrice Talon
Bhutan,Thimphu,Ngultrum,Dzongkha,Prime Minister - Lotay Tshering
Bolivia,"La Paz (administrative), Sucre (judicial)",Boliviano,"Spanish, Quechua, Aymara",President - Luis Arce
Bosnia and Herzegovina,Sarajevo,Convertible Mark,"Bosnian, Croatian, Serbian",Chairman of the Council of Ministers - Z
Botswana,Gaborone,Pula,"English, Tswana",President - Mokgweetsi Masisi
Brazil,Brasilia,Real,Portuguese,President - Jair Bolsonaro
Brunei,Bandar Seri Begawan,Brunei dollar,Malay,Sultan and Prime Minister - Hassanal Bolkiah
Bulgaria,Sofia,Lev,Bulgarian,Prime Minister - Boyko Borisov
```

C. Formato TEI para Dicionários

O formato TEI, definido pelo consórcio Text Encoding Initiative, é um dialeto baseado em XML (eXtensible Markup Language) para a codificação de documentos textuais, e tem vindo a suportar o trabalho de muitos projetos na área das Humanidades Digitais.

Para este enunciado iremos focar apenas num dos capítulos deste *standard*, em particular, o Capítulo 9, que define a codificação de dicionários. Como caso de estudo, será usado o Dicionário Aberto, um dicionário de 1913, transcrito por voluntários e disponível numa versão simplificada do capítulo 9 do TEI.

O que se pretende é a transformação deste XML nos formatos HTML e LaTeX. Sugere-se que se analise uma forma de apresentar o documento que seja legível, clara e, acima de tudo, prática.

```
<dic>
 <head>A</head>

<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>


<entry id="abacamartado" ast="1">
<form>
<orth>Abacamartado</orth>
</form>
<sense>
<gramGrp>adj.</gramGrp>
<def>
Parecido com um bacamarte:
<cit type="example"><quote>≪_uma cravina abacamartada._≫</quote></cit>
(De um testamento de 1693)
</def>
</sense>
</entry>


<entry id="abafo">
<form>
<orth>Abafo</orth>
</form>
<sense>
<gramGrp>m.</gramGrp>
<def>
Roupa ou pelles, que servem de agasalho.
</def>
</sense>
<etym>(De _abafar_)</etym>
</entry>

...
```