Processamento de Linguagem Natural

Engenharia Biomédica

Trabalho Prático 1

2024-2025

1 Introdução

O TP1 de Processamento de Linguagem Natural em Engenharia Biomédica consiste em aplicar os conhecimentos desenvolvidos nas aulas para processar vários documentos em formato PDF. Por conseguinte, pretende-se que sejam desenvolvidos vários parsers para extrair a informação que considere relevante. De seguida, a informação extraída deve ser preservada num ficheiro em formato JSON. Este ficheiro deve ser estruturado com o intuito de ser utilizado em trabalhos futuros.

2 Etapas

Para o desenvolvimento do projeto são recomendados os seguintes passos:

1. Análise do documentos em formato PDF, selecionando a informação relevante a ser extraída. Os documentos estão disponíveis no GitHub da disciplina;
2. Criação de uma sintaxe para representar a estrutura de dados a ser extraída;
3. Conversão dos ficheiros em formato PDF para um formato conveniente à sua manipulação;
4. Limpeza dos dados e criação de "marcas"de modo a remover elementos desnecessários e evidenciar a informação útil.
5. Extração de informação relevantes para estruturas de dados anteriormente definidas;
6. Guardar os dados num ficheiro no formato pretendido (JSON).

3 Restrições

- O processamento dos ficheiros **diccionari-multilinguee-de-la-covid-19.pdf** e **glossario_ neologismos_saude.pdf** é obrigatório! Para além dos ficheiros fornecidos, pode utilizar outros ficheiros em formato PDF que considere relevantes ao seu projeto;
- Devem ser processados pelo menos 3 ficheiros PDFs (incluindo o ficheiro obrigatório);

- O projeto deve ser desenvolvido na linguagem de programação Python;
- O relatório técnico deve ser produzido em LaTeX;
- O projeto deve ser desenvolvido por grupos de alunos com 3 elementos.
- Este trabalho prático deve ser entregue até ao dia anterior à data da apresentação do mesmo.

4 Entregáveis

Na entrega deste projeto deve entregar os seguintes elementos:

- Código desenvolvido no âmbito deste projeto;
- Ficheiro JSON com a informação extraída;
- Relatório técnico do sistema desenvolvido.
- Conjunto de slides para apresentação oral do projeto (entre 10 a 15 minutos).