Processamento de Linguagem Natural

Engenharia Biomédica

Trabalho Prático 1

2021-2022

O TP1 de Processamento de Linguagem Natural em Engenharia Biomédica consiste em aplicar os conhecimentos desenvolvidos nas aulas para processar um dicionário de termos médico multilingue.

Primeira Fase

Numa primeira fase, pretende-se fazer um parser para o dicionário médico. Este dicionário encontra-se em formato PDF, e pretende-se agora fazer-se uma extração da sua informação, nomeadamente, dos seus termos médicos, guardando-se todos os campos relevantes aos mesmos. De seguida toda a informação extraída deve ser preservada num ficheiro formatado em JSON.

Para isso são recomendados os seguintes passos:

1. Análise do dicionário médico, termos e seus campos;
2. Criação de uma sintaxe para representar a estrutura de dados a ser extraída;
3. Conversão do dicionário em formato PDF para um formato conveniente à sua manipulação;
4. Limpeza de dos dados, removendo-se elementos desnecessários;
5. Criação de marcas para destacar os campos a serem extraídos;
6. Extração dos campos relevantes para estruturas de dados anteriormente definidas;
7. Guardar dados num ficheiro no formato pretendido.