Natural Language Processing

Biomedical Engineering

Assignment 1

2022-2023

The first assignment of Natural Language Processing in Biomedical Engineering consists of applying what we have learned in the classes in order to process several medical PDF documents.

The main objective of this project is to retrieve information from the provided files and store that information so it can be used in future works. In order to accomplish this, we challenge you to create parsers to extract relevant information from each PDF file. Then, the extracted data should be preserved, for example, in a JSON file.

The following steps are recommended:

1. Analysis of the PDF files, selecting their relevant information;

2. Creation of syntax to represent the data structure to be extracted;

3. Conversion of the dictionary in PDF format to a format convenient for its manipulation;

4. Cleaning of the data, removing unnecessary elements;

5. Creation of tags to highlight the fields to be extracted;

6. Extraction of the relevant fields into the previously defined data structures;

7. Save the data in the desired file format.

*The processing of the file **dicionario_termos_medicos_pt_es_en.pdf** is mandatory!

*Besides the provided PDF files, you can use other PDF files that you consider relevant to your project.