[DESAFÍO #2] Técnicas para el procesamiento del lenguaje (NLP + LLMs)

by:
Daniel Fernando Barrera Armendaris

Index

Introduction

In this challenge it is requested to program a field identifier for CVs in format of PDF. We are going to use several libraries for that including NLTK to manage the corpus of the PDFs.

Structure of the notebook

This notebook can be divided into 4 main sections. The extracting of the text from the PDFs, the cleaning of the text to improve the corpus quality and the extracting of the requested fields to write them in JSON format.

Extracting of the text

In this notebook we're going to use the Fitz library that helps us to extract the text of the PDF. It's needed to mention that this action has some limitations when the structure of the PDF is complex with difficult positions and sizes of the letter.

Cleaning the corpus of the PDFs

The cleaning of the corpus of the pdfs is important because it helps us to reduce the noise and make the identification of the patterns. In this case I used the following operations: Tokenization, removing of stopwords, lemmatization and finally the join of the word to a single one, which helps the process to identify.

Extracting desired fields

This is the final step and it aims to identify the fields that are needed. In this case for the names is used the library Spacy that help us already to identify real names in english and for the other fields like email, phone number, years of experience and knowledge in artificial intelligence are used regular expressions

Technical debt

In this approach I considered the best way to use NLP techniques because it is necessary to extract the information needed. However I know that there are some other techniques than regular expressions. However it requires more analysis and more regularity in the formats of the CVs. We can train for example a NER model (Reconocimiento de Entidades Nombradas). As well the structure of the code can be improved by declaring the whole functions in other parts to reuse the code and finally it's needed to do some unit testing to evaluate the performance of the model and to know which of the approaches is the most suitable.