innoscripta

Innoscripta Data Science Task [2]

Description:

As a business, we work with small and medium German enterprises. To replenish our own database, we have to monitor and scrap out general info about our counterparties. Having this kind of information at hand will make our working routine easier, so we launched an initial data gathering process.

Data types:

We have a lot of documents with general info about top managers of German companies, including their:

- Name and surname
- City of residence
- Birth date
- Current title

Sometimes these documents contain such info as the country of managers' location and details about their foregoers, partners or deputies.

The idea is to extract data from these documents, but their structure is very flexible and different.*

For each piece of data, the key "text" stores the source text, the key "labels" stores what we have to learn to look for. Each label is represented as [start character, end character, signature]. **

Your task:

You are given a dataset (websites of German companies) and a JSON-file with ready-made pattern. Your goal is to train the NN.

Please, note that this task will be discussed at the final interview.
The follow-up questions will be related to possible data classification and preparation of raw datasets, so you may think about further steps and strategy in advance.

*Please see separate file with figures attached.
**Please see separate JSON files with ready-made patterns attached.