Innoscripta Data Science Task

Description:

As a business, we work with small and medium German enterprises. To keep our reputation safe, we have to monitor financial background of 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 lots of German company balance sheets (fig. 1 and fig. 2)*. Each balance sheet has Active and Passive parts, which also has its own positions. We are interested in data from these positions - *name and amount/value.*

The idea is to extract data from these tables, but their structure is very flexible and different. That is why we cannot just use a dummy parser to iterate over table rows and extract each position by taking numbers from the row.

Usually, documents have the following level system (fig. 5):
- A/B/C/D
- I/II/III/IV
- 1/2/3/4
- Sometimes they don't have those level indicators

Major obstacles:

- Sometimes position values are presented as sub-positions of that position
- Balance sheets can have flexible amount of columns
- Sometimes they have values not in euros, but in thousands of euros
- Sometimes they look like nightmares (fig. 3 and fig. 4)

Your task:

We do not ask you to write this parser, but we would appreciate if you can write down a few ideas on possible solution, in free format. In general, we need a tool to extract data from positions *name* and *value*, like in fig. 6 and fig. 7. Presumably, the tool must accept files or links to them and returns JSON or CSV file with data from balance sheets**. It must be a fast solution able to run 24/7, for we have LOTS of documents.

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