**Analytical tool for reading and compiling PDF reports from capacity auctions run by Polskie Sieci Elektroenergetyczne**

This repository contains an algorithm for reading and compiling data from PDFs published by the Polish national grid operator, Polskie Sieci Elektroenergetyczne (PSE), regarding capacity auctions.

PSE organizes capacity auctions since 2018, the first delivery year being 2021.

The currently maintained version of the website in English covers capacity auctions for delivery years 2026 - 2028

<https://www.pse.pl/web/pse-eng/areas-of-activity/capacity-market/capacity-market-news>

The here-presented compilation in Python is based on the full documentation of capacity auctions for delivery years 2021 – 2028, maintained in the Polish version of the website at <https://www.pse.pl/aktualnosci-rynku-mocy>

The basic analytical problem with that data is that the full results of each auction are provided in a PDF document, such as that available, for example, at <https://www.pse.pl/documents/31287/5147119783/2_Ostateczne_wyniki_aukcji_glownej_na_2028.pdf> .

The algorithm presented in this repository uses a library called TABULA in order to extract tables with numerical data from those documents in PDF.

**Remark**: normally, Tabula allows reading tables from many consecutive pages in a PDF document. For some reason, the PDF’s published by PSE do not allow translating multiple pages. It has to be done page by page. This is why the code contains separate line for each page.

This repository contains two files in Python:

* The file ‘**Read PDF auctions EN.ipynb**’ is the algorithm to read the source PDFs, as published by PSE, and transform them into Excel files. Those Excel spreadsheets require some cleansing and formatting work. TABULA is a bit messy with columns.
* The file ‘**Auctions PSE 2.ipynb**’ is an analytical tool for the cleansed and preprocessed data from Excels made with **Read PDF auctions EN.ipynb.**

The repository, besides the code in Python, contains the source PDFs from PSE, and the Excels created with both algorithms.