Documentation of NPD final assignment

Kacper Cybiski

February 9, 2022

1 Installation

2 Code Overview

2.1 `download.py`

2.1.1 `get_sheet_links_names`

This functions is fitted to crawl and return implicit links and names of the `.xlsx` files from the government stat website used in this program.

The function's inner working:

- Keyword arguments:
  - `year`, default value: 2019. This kwarg is responsible for choosing which year you want to download the files from. It is a simplified application, only restricted to scope of this program, because only sites with data for 2019 and 2020 have so similar urls.
- Procedure:
  - The procedure crawls webpage and collects all html segments containing href links
  - Then we cut this list looking for our charactericaly-shaped file links (36-character coded), as only 5 of those are among the links
  - Lastly we parse for according names, as we need those for saving the downloaded `.xlsx` files.