Taxes package documentation

Kacper Cybinski

February 11, 2022

1 Installation

2 Code Overview

2.1 `download.py`

`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 characterically-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.
- Returns:
  - A tuple, with structure:
    tuple(`list_of_sheet_links`, `list_of_corresponding_sheet_names`)

`download_sheet_series`

> This is the function that actually downloads the `.xslx` sheets from this government stat website and puts it in directory $\sim$ */data/* where all source `.xlsx` files will be stored.

The function's inner working:

- Keyword arguments:
  - `sheets`, is a tuple(`list_of_sheet_links`, `list_of_corresponding_sheet_names`) you would expect this to be extracted a result of function `get_sheet_links_names`
- Procedure:
  - The program iterates through the list of links and corresponding names, and tries to download the files using `requests` package.
  - The government website sometimes crashes and returns `.html` page instead of desired excel sheet, so that's why the needed files have been placed on external server, which serves as a mirror download source.
- Returns:
  - Path to a directory where all the downloaded files are located.

`get_gus_stats`

This is the function that fetches, downloads, and unzips the archive from GUS website, which we need for further program works

The function's inner working:

- Keyword arguments:
  - No kwargs
- Procedure:
  - Parse the gus website looking for desired `.zip` archive.
  - Download and unzip the archive using package `dload`. The archive is unpacked to a subdirectory of ∼ */data/*
  - Rename the downloaded directory to `gus`
- Returns:
  - Path to directory ∼ */data/gus/*, where desired `.xlsx` files can be found