Web Scraping BCN-PA-Pr Gig Description

This gig's delivery is

- a cleanly written Python 3.7 or 3.8 script or set of scripts using English for naming with
- a simple documentation in English and
- installation guide for a clean linux machine in the form of executable terminal commands or shell script.
- requirements.txt file with the python packages to be installed
- (optional) dockerfile
- Any other file(s) the script needs to run

Inputs

The main script will expect 2 inputs:

- Name (string)
- URL (string)
    - 5 pages:
        - http://www.governotransparente.com.br/transparencia/4382489/consultarempenho/resultado?ano=7&inicio=01%2F01%2F2020&fim=25%2F01%2F2020&unid=-1&valormax=&valormin=&credor=-1&clean=false
        - http://www.governotransparente.com.br/transparencia/4382489/consultarliqdesporc/resultado?ano=8&inicio=01%2F01%2F2021&fim=24%2F01%2F2021&orgao=-1&elem=-1&unid=-1&valormax=&valormin=&credor=-1&clean=false
        - http://www.governotransparente.com.br/transparencia/4382489/consultarliqrestpag/resultado?ano=7&inicio=01%2F01%2F2020&fim=25%2F01%2F2020&unid=-1&valormax=&valormin=&credor=-1&clean=false
        - http://www.governotransparente.com.br/transparencia/4382489/consultarpagdesporc/resultado?ano=7&inicio=01%2F01%2F2020&fim=25%2F01%2F2020&unid=-1&orgao=-1&elem=-1&valormax=&valormin=&credor=-1&clean=false
        - http://www.governotransparente.com.br/transparencia/4382489/consultarpagrestpag/resultado?ano=7&inicio=01%2F01%2F2020&fim=25%2F01%2F2020&unid=-1&valormax=&valormin=&credor=-1&clean=false

Script

The script will go to the URL and extract the data (from all pages) from the available table, including the header. If any of the columns contain links, follow each link to download additional table(s). The link URL and text value will be used as the first two columns in every line in these additional table(s).

Following the links, there are 1-4 sections. All of the available sections will be downloaded into separate csv files. Tabular data will go as lines into a csv. Text data (bullet points) will be kept as a single string with a clear separation between bullet points (you can use the bullet point character or some other special character, configured as a constant) also as a csv record. Data from every column link will be appended as new rows to the existing csv files.

Outputs

Outputs will be located in the code directory or in the data directory. Csv files will have the .csv extension, will be UTF-8 encoded and will be separated by comma.

- 1 main csv file with the data from the main table at the provided URL named Name.csv
- For each column that contains links, create 1 or more csv file, using the Column name as the filename as follows: Name.ColumnName.Section.csv. Remember to use the cell value and link as values in the first two columns of every record.

Requirements

- If you use a web driver (such as chrome browser), it needs to run in a headless mode, without the visual part.
- Environment variables and constants will be all defined in the beginning of each file.
- Concurrent downloads/workers are not a priority, as the code will be run on Google Cloud Run in a docker container. You can implement or suggest implementing concurrency if you are sure Cloud Run supports it.
- If URL's web certificates need to be accepted, you can accept the one's from the URL and its parents. Download the certificate as part of the solution, if necessary.
- If you can also provide a functional dockerfile to deploy our solution, we will pay extra, as discussed in the chat.

Examples

1. Name: consultarliqdesporc,
   URL:
   http://www.governotransparente.com.br/transparencia/4382489/consultarliqdesporc/resultado?ano=8&inicio=01%2F01%2F2021&fim=24%2F01%2F2021&orgao=-1&elem=-1&unid=-1&valormax=&valormin=&credor=-1&clean=false
   The output will contain
   - consultarliqdesporc.csv - the main table with header and 179 rows.
   - Following the links in the Documento column, add rows to the following tables:
     - consultarliqdesporc.documento.liquidacao.csv
     - consultarliqdesporc.documento.pagamento.csv
     - consultarliqdesporc.documento.empenho.csv
     - consultarliqdesporc.documento.movimentos_empenho.csv
   - Following the links in the Empenho column, add rows to the following tables:
   - consultarliqdesporc.empenho.csv

- consultarliqdesporc.empenho.movimentos_empenho.csv

Final Notes

As stated previously, the code will run on Google Cloud Run. If you have experience with it, feel free to make suggestions.

Any questions on your part or anything else you think we should know beforehand is encouraged, so we make everything clear between each other and can make you more productive.