Municipalities Crawler

Python project to crawl municipality websites and analyze interactivity.

Installation

First open a terminal in the project folder. Then run the following commands to create an environment with conda:

```
conda create -n muni-crawler
conda activate muni-crawler
conda install pip
pip install -r requirements.txt
```

Maybe I missed some packages in the requirements.txt file. If you get an error, try installing the missing package with conda and message me so I can add it.

Usage (Use terminal with conda environment in project folder)

```
cd src
python -m preprocess_data.py
python -m crawl.py
```

Arguments:

- --num_samples: The number of samples to crawl. (The number of samples to fetch (if -1, fetch all). Random state is set to 1.)
- --depth: The number of depth to crawl sublinks, 0 means only the main page.
- --num_workers: The number of workers to use. The number of workers to use (if depth > 0, 75% of the workers will be used for sublinks. if depth == 0, all workers will be used for the main pages)

PROF

TODO:

- resume crawl option
- Add retry mechanism for failed requests
- Some websites dont have utf-8 encoding and are not saved correctly
    - http://www.outines.fr/ `<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">`

Bugs:

- ClientConnectorError occurred while trying to connect to https://ezysureure.fr: Cannot connect to host ezysureure.fr:443 ssl:False [Connect call failed ('46.105.156.177', 443)]

Retrying with SSL = True: https://ezysureure.fr

- Unhandled exception occurred while trying to connect to https://assets.jimstatic.com/ownbgr.css.72b304e248c5b0dc046b611c132c3ad2.css: Cannot connect to host assets.jimstatic.com:443 ssl:True [Connect call failed ('151.101.38.2', 443)] after trying SSL = True and changing to https
- Unhandled exception occurred while trying to connect to https://www.comune.castrocarotermeeterradelsole.fc.it/: Cannot connect to host www.comune.castrocarotermeeterradelsole.fc.it:443 ssl:True [[SSL: DH_KEY_TOO_SMALL] dh key too small (_ssl.c:1002)] after trying SSL = True and changing to https