UOC M2.851 Tipología y ciclo de vida de los datos - PR 1

Target

The target is to create a dataset with data scraped from a web and publish the dataset in Zenodo. The web site targeted is https://www.filmaffinity.com/es/ranking.php?rn=ranking_2024_topmovies

Target preliminar analysis

WhoIs

```
{
    'domain_name': ['FILMAFFINITY.COM', 'filmaffinity.com'],
    'registrar': 'Arsys Internet, S.L. dba NICLINE.COM',
    'whois_server': 'whois.nicline.com',
    'referral_url': None,
    'updated_date': datetime.datetime(2023, 3, 23, 18, 20, 32),
    'creation_date': datetime.datetime(2001, 6, 20, 14, 23, 27),
    'expiration_date': datetime.datetime(2026, 6, 20, 14, 23, 27),
    'name_servers': ['HAL.NS.CLOUDFLARE.COM', 'JULE.NS.CLOUDFLARE.COM'],
    'status': ['ok https://icann.org/epp#ok', 'ok
https://www.icann.org/epp#ok'],
    'emails': 'abuse@nicline.com',
    'dnssec': ['unsigned', 'Unsigned'],
    'name': 'REDACTED FOR PRIVACY',
    'org': None,
    'address': 'REDACTED FOR PRIVACY',
    'city': 'REDACTED FOR PRIVACY',
    'state': 'Madrid',
    'registrant_postal_code': 'REDACTED FOR PRIVACY',
    'country': 'ES'
}
```

Terms of use of the target web

The use policy is described in https://www.filmaffinity.com/es/private.php In the use policy of the target web there is no restriction to using scraping tools on the web

BuiltWith

```
{
  'cdn': ['CloudFlare'],
  'hosting-panels': ['cPanel'],
```

```
  'analytics': ['comScore'],
  'javascript-frameworks': ['jQuery']
}
```

Robots

```
[
  'User-agent: *',
  'Disallow: /*?FASID',
  'Disallow: /*&FASID',
  'Disallow: /*/sharerating',
  'Disallow: /flash/rats.swf'
]
```

All agents are allowed, nevertheless we will use

```
Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.7)
Gecko/2009021910 Firefox/3.0.7
```

Sitemap

This web site does not provide a sitemap

Names

- Andoni Iribarren González
- Juan Pedro Rodríguez Nieto

Repository files

- source
  - main.py: entry point for the program
  - common.py: common functionality and constants
  - movie_scraper.py: module to obtain top100_2024 films
  - publication: module to publish the generated dataset to Zenodo

Code instructions

Environment

The file requirements.min.tx contains the minimum requirements to execute the code The minimum python version is 3.9; it is recommended to use python3.11 You can create an environment using venv

```
python3.11 -m env venv311
```

Then activate the virtual environment with

```
source venv311/bin/activate
```

Install the required libraries

```
python3.11 -m pip install --upgrade pip -r requirements.min.txt
```

Execution of the scraping

The code is designed to be executed from the base directory. Source is in the source directory and results will be stored in the dataset directory To execute the program use:

```
python source/main.py
```

Publication instructions

Once the dataset has been created, it can be automatically published in a new Zenodo deposition with

```
python source/publish.py
```

Dataset DOI

We have already published the dataset generated, with DOI https://doi.org/10.5281/zenodo.14078918

License information