| **File** | **what it does (one-liner)** | **you usually run it with…** |
| --- | --- | --- |
| **scrappy.py** | Scrapy spider that visits every link on the site (HTML, PDF, TXT) and stores **one JSON file per page** under data/<domain>\_<timestamp>/. | bash scrapy crawl fullsite \ -a domain=ogtic.gob.do \ -a start\_url=https://ogtic.gob.do/ |
| **realtimestats.py** | Scrapy *extension* that keeps live counters and, when the spider finishes, writes crawl\_summary.json (pages, words, bytes, duration, etc.) into the same folder. | No separate command – it is auto-loaded by Scrapy once the extension is listed in settings.py: python EXTENSIONS = {"site\_crawler.extensions.realtimestats.RealTimeStats": 500} |
| **filter.py** | Post-processing utility: drops duplicate URLs & duplicate texts, keeps only a chosen language, outputs a clean **JSONL** file. | bash python filter.py data/ogtic.gob.do\_YYYYMMDD\_HHMMSS \ --out ogtic\_clean\_es.jsonl \ --lang es |

**Typical “fire-and-forget” crawl inside *screen***

# run detached so it survives logout

screen -S crawl\_ogtic -dm bash -c '

scrapy crawl fullsite \

-a domain=ogtic.gob.do \

-a start\_url=https://ogtic.gob.do/ \

-s CLOSESPIDER\_PAGECOUNT=0 \

2>&1 | tee crawl\_ogtic\_$(date +%Y%m%d\_%H%M%S).log

'

echo "attach with: screen -r crawl\_ogtic (detach: Ctrl-A D)"

When it finishes you’ll find:

bash

data/ogtic.gob.do\_YYYYMMDD\_HHMMSS/ # per-page JSONs

└─ crawl\_summary.json # auto-generated by realtimestats.py

Process/clean the run (example for Spanish pages):

bash

python filter.py data/ogtic.gob.do\_YYYYMMDD\_HHMMSS \

--out ogtic\_clean\_es.jsonl \

--lang es

Top of Form

Bottom of Form

\\\