# Web Scraping 101

## Tu data es mi data

Mihail Gavrilita

May 2, 2024

# The Internet

## A near-infinite source of data

- Your data is there but how to collect it?
- Use it in a Hackathon presentation, ML project or a data-processing startup idea.

# Web Scraping

## Mechanical copy-pasting monkey

- Automatic data collection (scripts);
- For search engines or business market research.

# Web Scraping

## Traditional approach

- Crawling
- Parsing

# Web Scraping

## Traditional approach

- Crawling
  - Visit a link;
  - Find all links on that page;
  - Save the links;
  - Repeat.
- Parsing
  - Visit a link;
  - Detect you are on a page you wanna scrape;
  - Collect data you are interested in;
  - Save the link.

# Web Scraping

## Scrapy – a Python framework

Scrapy allows it's users to crawl websites using spiders and parse pages containing data, saving it into items. On the fly processing of the data we collect happens in itemloaders. The processing of the items themselves is handled by pipelines.

# Web Scraping

## Scrapy – a Python framework

Scrapy allows it's users to crawl websites using spiders and parse pages containing data, saving it into items. On the fly processing of the data we collect happens in itemloaders. The processing of the items themselves is handled by pipelines.

# Web Scraping

## Other options

- APIs provided by websites (often times paid);
- GUI tools for scraping;
- Cloud solutions (e.g. Zyte);
- LLMs and other AI tools.

# Web Scraping 201

## Topics for another time

- JavaScript rendering (e.g. Playwright, Selenium);
- Techniques used by websites to slow down / stop scraping (e.g. Cloudflare, ReCaptcha);
- Legal matters (robots.txt, accounts).

# Web Scraping

## Links 2 3 4

For the live part:

- `https://alternativeto.net/software/scrapy/`
- `https://docs.python.org/3/library/venv.html`
- `https://docs.scrapy.org/en/latest/intro/install.html`
- `https://docs.scrapy.org/en/latest/intro/tutorial.html`
- `https://github.com/TUM-FAF/Lectures`