Web scraping and social media scraping: Scrapy

Przemysław Kurek, Maciej Wysocki

Chair of Political Economy
Faculty of Economic Sciences
University of Warsaw

Class 06

We already know:

- How to make a simple scraper using Seautiful Soup.

Today we will learn:

- How to do the same thing using Scrapy.

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

What is Scrapy?

- Scrapy (`scrapy.org`) is a **framework** for writing web scrapers and crawlers.
- It means, that it is **whole environment** for making scrapers.
- When you approach a framework, you stop learning python, and you start learning framework.
- Frameworks are different from libraries (still, no hard definition), they force users to live their ways.
- When you start using scrapy, you should forget old habits, and start doing things **the scrapy way**. *This is the way.*

Why scrapy?

- Scrapy is real titan in case of efficiency. It performs much better than Beautiful Soup and Selenium.
- It is fully automated from the beginning. You have full power just after writing a few lines.
- You have access to tons of automagical stuff by the optios.
- It is also very extensible. You can customize everything.
- However! At the beginning it may be very confusing what is going on around and how to live the scrapy way.

First Scrapy steps. To start with Scrapy we need to cover three areas:

- Managing Scrapy projects.
- Basic construction of a spider.
- Scrapy selection tool: Xpath.

Managing Scrapy projects:

- Creating a project.
- Creating or copying spiders.
- Running spiders.
- Managing options.

Uniwersytet Warszawski
Wydział Nauk
Ekonomicznych

Basic construction of a spider:

- What to scrap.
- Where to scrap.
- How to scrap.

Uniwersytet Warszawski
Wydział Nauk
Ekonomicznych

Xpaths:

- Xpath is Scrapy (and Selenium) query language to parse HTML and retrieve data.
- They do generally the same, as Beautiful Soup:
  - **Tree navigation:**

    `'/html/body/h1'`
  - **Searching by tag name and/or attributes:**

    `'//title[@lang='en']'`
  - **Searching by regex:**

    `'//a[re:test(@title, "List of painters.*")]'`
  - **Access to text:**

    `'../text()'`
  - **Access to attribute value:**

    `'../@href'`

How to learn Scrapy:

- Unfortunately, we do not have time here for a lot of Scrapy.
- If you wish to extend your knowledge about this framework, and at some point write really powerful scrapers and crawlers, some external resources (tutorials, courses) would be needed.
- Xpaths are other thing. They are needed even for simple projects. You might want to know them also for Selenium.
- If you want to experiment with them at a website of your choice, use and modify file: `06a_xpath.py`.
- There are also tons of Xpath content in the web. For example: `https://www.w3schools.com/xml/xpath_syntax.asp`

Uniwersytet Warszawski
Wydział Nauk
Ekonomicznych

Classroom activity:

- Run and analyze provided Scrapy scraper.
- Solve exercises at `06_exercises.py` file.

Suggested reading:

- Chapter 5 gives some additional information about Scrapy.

Homework:

- For the next classes we will need Selenium and geckodriver
- Make sure you can run it at a computer of your choice. Test it with a file provided before.
- **You will not be able to follow next class without it!**