WEB SCRAPING WITH PYTHON

BY SUSAN NDAGI

EBU European Business University
Luxembourg

LECTURE AIMS AND OBJECTIVES

- What is Web Scraping
- Benefits
- Practical

Web Scraping

- Web scraping is used to collect a large amount of data from websites.

- It is automated, thus saves you time (hours or days) from doing it manually.

- The data collected is **unstructured** and web scraping helps collect this data and store it in a structured format.

Benefits

- **Price Comparison** – e.g. online shopping websites like Amazon, or stock market prices
- **Email address gathering** – bulk emails for marketing, hacking
- **Social media scraping** – e.g. Twitter for trending topics, Memes
- **Research and Development** – many datasets from different sites can be analyzed and used for surveying, R&D
- **Job listings**
- **Online reviews** – check for customer satisfaction

How to scrape websites

- Write your own program
- APIs – Application Programming Interfaces
- Online Services – free or paid

Web Scraping with Python

- You can find this file by appending “/robots.txt” to the URL that you want to scrape – to see if legal to do so
  - Allow
  - Disallow – e.g. YouTube

BENEFITS

- Ease of Use
- Many libraries for extraction and manipulation of data
- Easy to Understand
- Small code, large task – due to use of packages
- Online Support Community

How to Scrape Data

1. Find the URL that you want to scrape
2. Inspecting the Page – Inspect Element
3. Find the data you want to extract
4. Write the code
5. Run the code and extract the data
6. Store the data in the required format

Libraries used for Web Scraping

- **BeautifulSoup:** Beautiful Soup is a Python package for parsing (pull data out of) HTML and XML documents.
  - It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree.

- **Selenium:** It is a headless browser. It is used to automate browser activities. Selenium is normally used for web testing and scraping.

- **PhantomJS:** It is a headless browser used for automating browser activities.

- **Scrapy:** It is a fast high-level web crawling and web scraping **framework**, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing.

- **Pandas:** Pandas is a library used for data manipulation and analysis.
  - It is used to extract the data and store it in the desired format.
  - Used after the scraping of the website for structuring the data

Python Libraries for Data Extraction

- import requests
  - It allows you to send HTTP requests using Python.
  - The HTTP request returns a Response Object with all the response data (content, encoding, status, etc).
  - **Installation:** pip install requests
- import bs4
  - **Installation:** pip install bs4
- import selenium
  - **Installation:** pip install selenium

Python Libraries for Data Manipulation

- NumPy – Numerical Python
  - N-dimensional Arrays – Rows and Columns

- Pandas – Data Analysis
  - Tabular data
  - Series - Columns
  - DataFrame – Rows and Column

- Matplotlib – Data Visualisation
  - 2D graphics/ charts

PRACTICAL

- Get S&P500 list of companies tickers from NYSE
- Product pricing comparison

THANK YOU!

ANY QUESTIONS?

EBU European Business University
Luxembourg