Extrakcia informácii z webu pre analýzu dát pomocou Pythonu

Metódy inžinierskej práce 2023/2024

Štefan Kučerák

xkucerak@stuba.sk

Ústav informatiky, informačných systémov a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave

27. november 2023

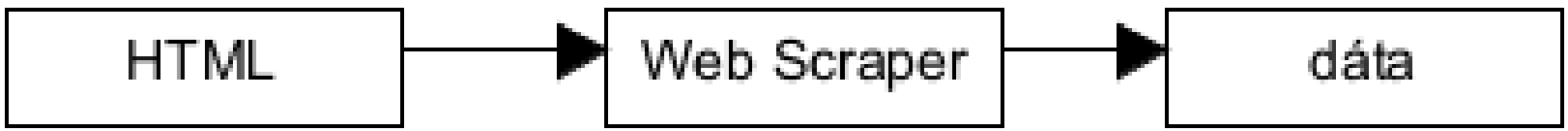
Introduction

Web scraping:

- automated data collection from websites
- obtaining a large amount of data in a short time

Overview

The principle of data extraction



Principle of data extraction.

Libraries

Library	Description
Requests[1]	HTTP requests using POST and GET
Beautiful Soup[2]	the library contains tools for viewing HTML or XML code
Scrapy[3]	is a fast framework for high-level web browsing
Selenium[4]	is a library that is a web driver

[1] https://pypi.org/project/requests/
[2] https://pypi.org/project/beautifulsoup4/
[3] https://pypi.org/project/Scrapy/
[4] https://pypi.org/project/selenium/

Examples

The following projects used Python and one of the libraries listed:

- Phishing Web Page Detection using web Scraping
- Food Genie, Recipe Search Algorithm using Web Scraping
- NEWSONE- AN AGGREGATION SYSTEM FOR NEWS USING WEB SCRAPING METHOD

Simple example

Listing the page title using the Beautiful Soup library

- CODE

```
import requests
from bs4 import BeautifulSoup

page = requests.get("https://www.google.com/")
soup = BeautifulSoup(page.content, "html.parser")

print(soup.title.string)
```

- Output

```
Google
```

This solution can be used on multiple sites

Vstup	**Výstup**
google.com	Google
instagram.com	Instagram
youtube.com	YouTube
wikipedia.org	Wikipedia
www.stuba.sk	Slovenská technická univerzita v Bratislave (STU)
www.fiit.stuba.sk	Fakulta informatiky a informacnych technologii v Bratislave

Conclusion

- the Python programming language is widely used in this area
- we can use one of the listed libraries to extract data from the web
- we can then use the data for further processing