Extrakcia informácii z webu pre analýzu dát pomocou Pythonu

Metódy inžinierskej práce 2023/2024

Štefan Kučerák

xkucerak@fiit.stuba.sk

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

26. november 2023

Úvod

- automatizované zbieranie dát z web stránok
- časová náročnosť

Prehľad

Príklady reálneho využitia

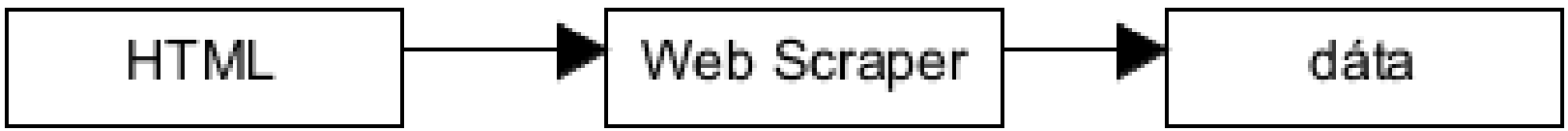
- "Phishing Web Page Detection using web Scraping“
- "Food Genie, Recipe Search Algorithm using Web Scraping"
- "NEWSONE- AN AGGREGATION SYSTEM FOR NEWS USING WEB SCRAPING METHOD"

Knižnice

Knižnica	Popis
Requests[1]	HTTP požiadavky a to POST aj GET
Beautiful Soup[2]	táto knižnica obsahuje nástroje na html alebo xml
Scrapy[3]	je rýchly framework pre prehľadávanie webu na vysokej úrovni
Selenium[4]	je knižnica, ktorá je 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/

Princíp extrakcie dát



Nejaká poznámka k obrázku, možno zdroj...

Jednoduchý príklad

Vypísanie názvu stránky pomocou knižnice Beautiful Soup

- Program

```
import requests
from bs4 import BeautifulSoup

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

print(soup.title.string)
```

- Výstup

```
Google
```

Zhodnotenie a ďalšia práca

- Každá prezentácia musí byť nejako uzavretá
- Ale vždy je čo robiť ďalej...