

Extrakcia informácií z webu pre analýzu dát pomocou Pythonu

Metódy inžinierskej práce 2023/2024

Štefan Kučerák

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

6. november 2023

Úvod

- automatizované zbieranie dát z web stránok
- časová náročnosť

Prehľad

1 Príklady reálneho využitia

2 Knižnice

3 princíp

4 príklad

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"

Knížnice

- Requests¹
- BeautifulSoup²
- Scrapy³
- Selenium⁴

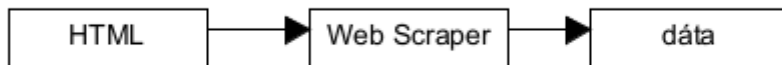
¹ <https://pypi.org/project/requests/>

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 BeautifulSoup

■ 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...