Extrakcia informácii 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

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"

Ďalší slajd

- Nejaký text
- Ďalší text – *zvýraznený text*
- *Kľúčová poznámka*
- Bol použitý balík beamer[1]

[1] http://www.tex.ac.uk/tex-archive/macros/latex/contrib/beamer/doc/beameruserguide.pdf

Princíp extrakcie dát

Nejaká poznámka k obrázku, možno zdroj...

Jednoduchý príklad

Vypísanie názvu stránky

- 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...