Objektovo orientované programovanie - *pokročilí*

Učiteľ:
Ing. Jozef Wagner, PhD.

Učebnica:
https://oop.wagjo.com/

OPGP
Pokročilí 27

1. Web Scraper – princíp fungovania
2. Nástroje pre Web Scraping
3. Platforma Apify

Web scraper

Web scraper je názov pre programy, ktoré systematicky prehľadávajú webovú lokalitu a sťahujú z nej dáta

Princíp fungovania:

Scraper začína na štartovacej URL adrese

1. Pošle HTTP požiadavku a stiahne HTML súbor so stránkou.
2. Zparsuje a zanalyzuje HTML kód stránky a nájde URL odkazy na ďalšie stránky alebo súbory
3. Stiahne súbory, o ktoré máme záujem
4. Z nájdených URL odkazov vyberie tie, o ktoré máme záujem a pokračuje bodom 1

Web scraper

Aby scraper nepokračoval do nekonečna, musí si:

- pamätať už navštívené stránky
- ukončiť scrapovanie po určitom počte stránok, resp. súborov
- ukončiť po určitom čase

Pri pokročilom scrapovaní nestačí iba stiahnuť HTML súbor, ale scraper musí vykonať aj Javascript kód na stránke a tváriť sa, že je regulérny prehliadač.

Nástroje pre Web Scraping

Najčastejšie Python a Javascript (Typescript)

Sú to často stredne veľké skripty, ktoré necháte bežať buď lokálne na vašom počítači, alebo na prenajatom serveri. Populárne Python nástroje:

- **Requests** — Najpoužívanejšia knižnica na odosielanie HTTP požiadaviek.
- **Crawlee** - Open-source crawling framework - moderný nástroj na tvorbu robustných scraperov
- **BeautifulSoup** — Knižnica na parsovanie HTML a jednoduchú extrakciu dát.
- **Scrapy** — Plnohodnotný framework na veľké a škálovateľné crawling projekty.
- **Playwright** — Knižnica na automatizáciu prehliadača pri dynamických stránkach.
- **Selenium** — Klasická knižnica na ovládanie reálneho prehliadača

Etika a právne aspekty

- Zbierajte **len verejné dáta**
- Rešpektujte **Terms of Service** stránok
- Pri osobných údajoch dodržiavajte **GDPR**
- Nevytvárajte DDoS útoky – scraping nie je útok
- Buďte transparentní (v User-Agent hlavičke môžete uviesť, že ste scraper)

Web scraping je legálny nástroj, ale jeho zneužitie môže mať právne následky. Vždy postupujte eticky a zodpovedne.

Apify.com - platforma na **web scraping**

Actor	Serverless skripty - boti na scraping alebo automatizáciu	Instagram Scraper, vlastný e-shop crawler
Task	Uložená konfigurácia Actora s konkrétnymi vstupnými parametrami	„Scrapuj denne produkty z Alza.sk“
Run	Jedno konkrétne spustenie Actora alebo Tasku	Spustenie Tasku každý deň o 8:00
Dataset	Hlavné úložisko výsledkov (štruktúrované dáta)	Export do CSV / JSON / Excel / Pandas
Key-value store	Úložisko pre súbory, screenshoty, JSON objekty	Uloženie PDF faktúr alebo obrázkov
Proxy	Automatická rotácia IP adries (datacenter + residential)	Zabrániť blokovaniu pri veľkom scrapingu
Schedule	Automatické spúšťanie Taskov podľa časového plánu (cron)	Denný / hodinový zber dát