SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Evidenčné číslo: FEI-5384-80051

NÁSTROJE NA ZÍSKAVANIE TEXTOVÝCH DÁTOVÝCH MNOŽÍN PRE POTREBY UMELEJ INTELIGENCIE

DIPLOMOVÁ PRÁCA

2021 **Lukáš Orlický**

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Evidenčné číslo: FEI-5384-80051

NÁSTROJE NA ZÍSKAVANIE TEXTOVÝCH DÁTOVÝCH MNOŽÍN PRE POTREBY UMELEJ INTELIGENCIE

DIPLOMOVÁ PRÁCA

Študijný program:	Aplikovaná informatika
Názov študijného odboru:	Informatika
Školiace pracovisko:	Ústav informatiky a matematiky
Vedúci záverečnej práce:	prof. Dr. Ing. Miloš Oravec
Konzultant:	Ing. Zuzana Bukovčiková

Bratislava 2021 **Lukáš Orlický**

SÚHRN

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Študijný program:	Aplikovaná informatika
Autor:	Lukáš Orlický
Diplomová práca:	Nástroje na získavanie textových dátových množín pre potreby umelej inteligencie
Vedúci záverečnej práce:	prof. Dr. Ing. Miloš Oravec
Konzultant:	Ing. Zuzana Bukovčiková
Miesto a rok predloženia práce:	Bratislava 2021

Tu bude súhrn

Kľúčové slová: kľúčové slovo1, kľúčové slovo2, kľúčové slovo3

ABSTRACT

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF ELECTRICAL ENGINEERING AND INFORMATION TECHNOLOGY

Study Programme:	Applied Informatics
Author:	Lukáš Orlický
Master's thesis:	Nástroje na získavanie textových dátových množín pre potreby umelej inteligencie
Supervisor:	prof. Dr. Ing. Miloš Oravec
Consultant:	Ing. Zuzana Bukovčiková
Place and year of submission:	Bratislava 2021

Abstract to be done!

Keywords: keyword1, keyword2, keyword3

Čestné vyhlásenie

Čestne vyhlasujem, že som túto diplomovú prácu spracoval samostatne na základe získaných teoretických vedomostí a že všetku použitú literatúru a ďalšie pramene som v diplomovej práci vyznačil.

V Cíferi, dňa xx. xx. 2021

.....

Podpis

Poďakovanie

Ďakujem vedúcej diplomovej práce prof. Dr. Ing. Milošovi Oravcovi vedenie mojej Diplomovej práce. Ďalej moje poďakovanie patrí Ing. Zuzane Bukovšíkovej, za jej cenné poznatky, rady a pripomienky, ktorými mi bola nápomocná pri tvorbe tejto práce.

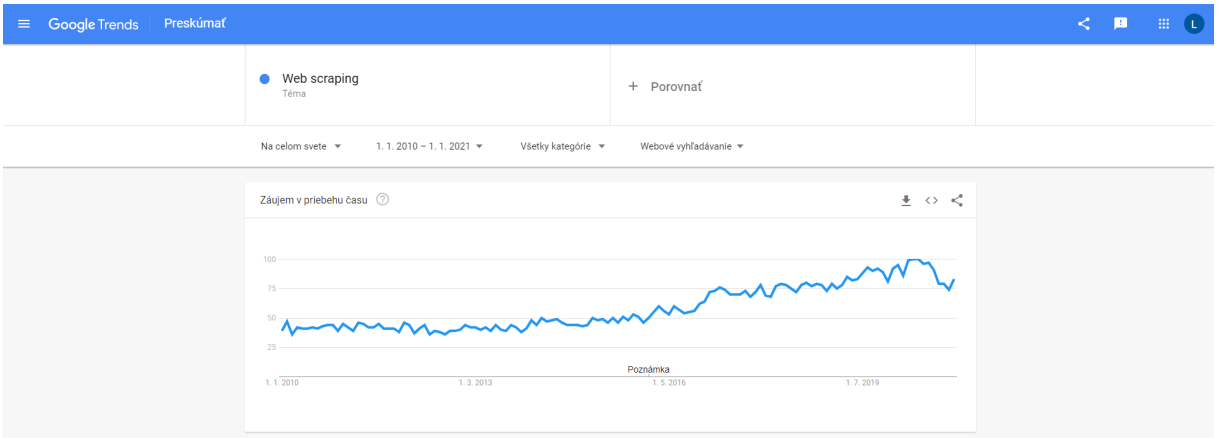
Obsah

Zoznam obrázkov a tabuliek

Úvod

1 Web scraping

Pojem web scraping sa používa pre extrakciu dát z webových stránok a následné uloženie týchto dát na neskoršie využitie. Pre webscraping sa používa najmä softvér, ktorý je manuálne ovládaný používateľom. V poslednej dobe sa začína do popredia dostávať automatizovaný web scraping pomocou robota alebo webového prehliadača. Web scraping je najviac používaný pri sledovaní cien internetových obchodoch, pri sledovaní spravodajstva a tiež na prieskum trhu. Na obrázku č.1 z Google Trends, môžme vidieť, že jeho popularita za posledných 10 rokov rýchlo narastá.



Obr. 1: Nárast popularity Web Scrapingu

Výhodou web scrapingu je aj to, že môže brať aj údaje, ktoré sú ukladané do databázy, tým pádom extrahuje celý webový obsah stránok.
Existuje viacero typov nástrojov na extrakciu dát, tzv. robotov, ktoré si vie používateľ prispôsobit. Poznáme niekoľko rôznych robotov, ktoré sú určené napriklad na:

1. rozpoznávanie štruktúry stránky
2. transformáciu obsahu
3. ukladanie dát
4. extrahovanie dát z API

1.1 Extrakcia dát a bezpečnosť

Extrakcia dát sa používa v oprávnené ale aj v neoprávnené účely zberu dát. Medzi oprávnené účely sa používa hlavne v rôznych digitálnych podnikoch, ktoré potrebujú väčšie množstvo dát. Ak sa web scraping používa na nelegálne účely ide hlavne o kradnutie obahu

s autorskými právami. Keďže všetky nástroje na zber dát majú rovnaký účel, t.j. zbierať dáta zo stránok, ťažko sa rozlišuje medzi legálnym a nelegálnym robotom. Oprávnené použitia web scrapingu sú[1]:

- prehľadávacie nástroje
- roboty na prehľadávanie e-shopov
- získavanie dát z fór a sociálnych médií

2 Softvéry na extrahovanie dát

Pre extrahovanie dát z webových stránok existuje viacero softvérov. Väčšina týchto softvérov pracuje na backende umelá inteligencia.

2.1 Import.io

...[2]

Záver

Zoznam použitej literatúry

1. *Web Scraping: What is web scraping.* impreva.com, 2019. Dostupné tiež z: `https://www.imperva.com/learn/application-security/web-scraping-attack/`.
2. *Top 10 Best Website Crawlers in 2019 (Reviews Comparison: Import.io* [online]. scrapestorm.com, 2019 [cit. 2019-11-03]. Dostupné z: `https://www.scrapestorm.com/tutorial/top-10-best-website-crawlers-in-2019-reviews-comparison/`.

Prílohy

A Štruktúra elektronického nosiča