Metody extrakce dat z webových stránek

Lukáš Perina

Brno University of Technology, Faculty of Information Technology
Božetěchova 1/2. 612 66 Brno - Královo Pole
xperin11@fit.vutbr.cz

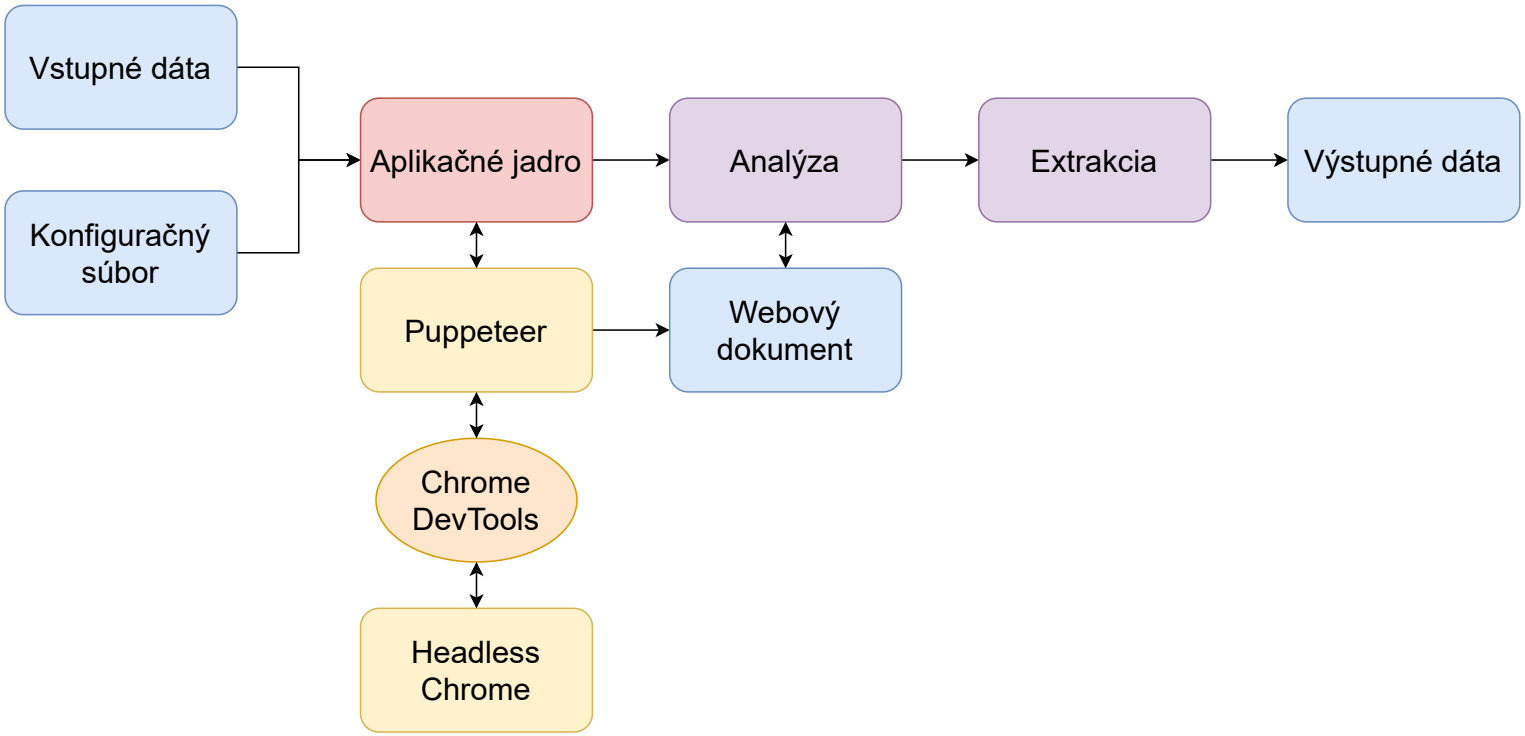
Jún 16, 2021

Motivácia

- Extrakcia dát z webu stále nieje dokonalá
- Možnosť extrakcie na základe uvedenia obsahu webového dokumentu
- Automatizácia

Technológie súčasnosti

- Hotové riešenia
  - Scrapping-bot
  - Octoparse
  - Import.io
- Platformy a knižnice
  - Apify SDK
  - Cheerio
  - Scrapy
  - Puppeteer



Obr.: Diagram navrhnutej architektúry

Implementácia

- Konfiguračný súbor
- Vstupné dáta
- Analýza a extrakcia
  - Analýza štylizačných atribútov class
  - Zoradenie na základe početnosti
  - Detekcia primárneho prvku
  - Zostavenie stromovej štruktúry a následná extrakcia dát

Tabuľka: Výsledné hodnoty testovaných dátových sád

Dátová sada	**Presnosť**	**Úplnosť**	**Čas**
Futbalové výsledky	97,9 %	86,5 %	13627 ms
Eshop TSBohemia	100,0 %	100,0 %	10876 ms
Eshopy	71,42 %	68,41 %	11228 ms
Správy a novinky	72,73 %	66,23 %	12885 ms
Priemerné výsledky	85,51 %	80,28 %	12154 ms

Pozn.: Dátová sada obsahuje 10 webových dokumentov.

Otázky oponenta

1. U knihovny Puppeteer jste zmiňoval především výhody, můžete se zamyslet i nad nevýhodami použití této knihovny?

2. Chápu, že v současné době není k dispozici žádná srovnávací testovací sada pro alternativní nástroje, ale dle zmínky v textu byl vytvořen jiným studentem alternativní nástroj s jiným přístupem k extrakci, ale využívající stejné testovací sady. Můžete provést alespoň srovnání s kolegou z hlediska přesnosti a časové náročnosti pro jednotlivé sady?

Dátová sada	Moja aplikácia			Kolegova aplikácia		
	Presnosť	Úplnosť	Čas	Presnosť	Úplnosť	Čas
Futbalové výsledky	97,9 %	86,5 %	13627 ms	93,97 %	75,12 %	26705 ms
Eshop TSBohemia	100,0 %	100,0 %	10876 ms	88,72 %	87,19 %	21383 ms
Eshopy	71,42 %	68,41 %	11228 ms	75,18 %	66,58 %	20090 ms
Správy a novinky	72,73 %	66,23 %	12885 ms	82,58 %	81,79 %	21842 ms
Priemerné výsledky	85,51 %	80,28 %	12154 ms	85,11 %	77,67 %	22505 ms

Obr.: Porovnanie dvoch odlišných prístupov k extrakcii

Ďakujem za pozornosť!