Sem vložte zadání Vaší práce.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Webová aplikace pro online web scraping

Jakub Drahoš

Katedra softwarového inženýrství
Vedoucí práce: Martin Podloucký

2. února 2019

Poděkování

Doplňte, máte-li komu a za co děkovat. V opačném případě úplně odstraňte tento příkaz.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo"), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 2. února 2019 .....

Odkaz na tuto práci

Drahoš, Jakub. *Webová aplikace pro online web scraping*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2019.

Abstrakt

V několika větách shrňte obsah a přínos této práce v češtině. Po přečtení abstraktu by se čtenář měl mít čtenář dost informací pro rozhodnutí, zda chce Vaši práci číst.

Klíčová slova Nahraďte seznamem klíčových slov v češtině oddělených čárkou.

Abstract

Sem doplňte ekvivalent abstraktu Vaší práce v angličtině.

Keywords Nahraďte seznamem klíčových slov v angličtině oddělených čárkou.

Obsah

Seznam obrázků

Úvod

Kapitola 1

Cíl práce

Kapitola 2

Analýza a návrh

2.1 Analýza konkurence

2.1.1 ParseHub

Celkový dojem - je to trošku sloppy, během 5 minut používání to stihlo crashnout. Rozhraní mi nepřijde super přívětivé, je toho na uživatele moc najednou, nepůsobí to na mě moc přehledně. Všechno se musí naklikat, což není úplně ideální. Na druhou stranu je klikání někdy tou nejlepší možností. V kombinaci s chytrým hledáním patternů (např. kdy uživateli stačí označit dvě položky a všechny stejné se automaticky vyberou také) je to rozhodně super feature.

- Nutnost stažení aplikace - vypadá jako upravená Mozilla Firefox (nejspíš Add-on). Podpora Window, Linux, Mac.
- Výběr dat ke scrapování probíhá pomocí klikání. Hezké je, že aplikace není úplně hloupá a probíhá zde určitý autoselect na základě našeho výběru (takže se člověk neukliká). Toto inteligentní hledání vzorců, podle kterých uživatel pravděpodobně chce vybrat data je hodně cool.
- Aplikace obsahuje moc pěkný tutorial, jak s nástrojem zacházet...not bad.
- Možnost získání dat různými formami - přes API, jako CSV/Excel, do GoogleSheets nebo do Tableau.
- Různé módy kliknutí (select, relative select, click). Zooming in/out na HTML elementy.
- Nelze vyhledávat podle klíčových slov nebo hledání (vše se musí naklikat).

2.1.2 Octoparse

Celkový dojem - opět velmi těžkopádná aplikace, nepřehledná a složitá. Je zde ale možnost data jak naklikat, tak vybrat na základě nějaké shody (XPath, RegEx). U klikání taktéž chytré hledání patternů. Možnost nakonfigurovat logiku procházení na webu, nastavení stovky a tisíce tasků, jejich spouštění - skvělá příležitost pro automatizaci a možnost, jak prolézt nějakou celou doménu.

- Nutnost stažení aplikace. Podpora pouze pro Windows.
- Možnost data jak naklikat, tak vybírat na základě hledání pomocí XPath nebo RegExpů.
- Tutoriál v podstatě nic neříkající.
- Předpřipravené šablony (může velmi urychlit práci, ale je jich jenom pár a jsou velmi konkrétní).
- Chytré hledání podobných dat (takže se uživatel neukliká)
- Pestrá paleta možností (branch judgment, tvoření smyček apod.) - dá se vytvořit téměř jakákoli logika
- Možnost řídit tasky přes API (a získávat tak data taktéž přes API). Data jdou nahrát rovnou i do lokální databáze.

2.1.3 WebScaper

Celkový dojem - zatim rozhodně nejlepší aplikace. Jednoduchá instalace (pouze rozšíření do Chromu) a hlavně JEDNODUCHÉ a PŘEHLEDNÉ ovládaní. Nastavování probíhá skrze dev console v Chromu, všechno běží hladce a plynule. Klikání je jednoduché a intuitivní, tutoriály vysvětlí vše potřebné během pár minut. Samozřejmostí je označení všech podobných prvků na základě dvou/tří ručně označených.

- Nutnost stažení pouze rozšíření do Chromu (ideální)
- Možnost si hledaná data jak naklikat, tak hledat na základě HTML, CSS
- Tutoriály formou videí - jednoduché, rychlé, naprosto postačující
- Různé módy kliknutí, takže lze lehce prolézt celý web
- Možnost exportovat data do CSV, získat přes API nebo do Dropboxu
- Možnost využít jejich cloud k automatizaci

2.1.4 Dexi.io

Celkový dojem - není potřeba žádná instalace, vše funguje jako webová aplikace, což je v tomto ohledu ta nejlepší varianta. Ovládání velmi podobné jako u WebScraperu, vypadá to v podstatě jako dev console v prohlížeči. Výběr dat ke scrapování probíhá opět přes klikání, je zde ale i možnost vybírat data na základě HTML, CSS, textu. Ovládání není tak moc intuitivní, základní tutorial je hodně povrchní, na druhou stranu mají na stránkách návodů dostatek. Funguje i jako crawler, umožňuje scrapovací boty napojovoat za sebe, skládat z menších větší a automatizovat celý proces.

- Bez nutnosti stahování aplikace - velká výhoda
- Možnost si hledaná data jak naklikat, tak hledat na základě HTML, CSS, textu
- Úvodní tutoriál dost povrchní, ale dostatek návodů na stránkách
- Různé módy kliknutí, takže lze lehce prolézt celý web
- Možnost exportovat data do CSV, získat přes API
- Různé módy aplikace - scraping, crawler, pipes a autobot

Kapitola 3

Realizace

Závěr

Literatura

Příloha **A**

Seznam použitých zkratek

GUI Graphical user interface

XML Extensible markup language

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `exe` ..... adresář se spustitelnou formou implementace
- `src`
  - `impl` ..... zdrojové kódy implementace
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF
  - `thesis.ps` ..... text práce ve formátu PS