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

30. března 2019

Poděkování

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

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl 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 30. března 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 web scraping, extrakce dat, aplikace,

Abstract

Sem doplňte ekvivalent abstraktu Vaší práce v angličtině.

Keywords web scraping, data extraction, application, non-programmers,

Obsah

Seznam obrázků

Úvod

Cíle práce

Hlavním cílem této práce je návrh a tvorba webové aplikace, která bude umožňovat uživatelům extrahovat požadovaná data z libovolné stránky v reálném čase bez jakékoli nutné znalosti programování. Při specifikaci požadavků tohoto softwaru se přihlídne k analýze stávajících řešení, jež je vedlejším cílem této práce.

Neméně důležitou součást práce tvoří dodržení klasického vývojového cyklu softwarového projektu – analýza, design, implementace a testování.

Klíčovým aspektem aplikace je též *přehlednost a jednoduchost uživatelského rozhraní* – důraz bude kladen na intuitivní a rychlé ovládání.

Naopak v rozsahu této práce není tvorba web crawlera ani žádného jiného podobného mechanismu, jenž by systematicky procházel danou oblast webu.

Motivace

Přínos? Proč je téma důležité/aktuální? Komu tím jako prospějete? Proč tohle téma?

Členění práce

V kapitole 1 této práce je představen

Kapitola 1

Analýza a návrh

1.1 Web scraping

Web scraping (nebo také *web harvesting, web data extraction*) je technika získávání nejrůznějších dat z webových stránek. Nejčastěji se v tomto kontextu jedná o automatizovaný proces strojového zpracování a získávání dat, nicméně může jít i o manuální extrakci zadanou uživatelem skrze nějaký software (jako je tomu právě v našem případě). [citace z Wiki - web scraping]

Často se také v souvislosti s pojmem web scraping používá spojení *web crawler* (nebo také *bot, spider, spiderbot*). Jedná se o automatizovaný software, který systematicky prochází danou oblast webu a během toho extrahuje kýžená data. Jak již bylo řečeno v úvodu, touto částí web scrapingu se práce nebude zabývat.

1.1.1 Krátce z historie

Historie web scrapingu šahá k samým počátkům internetu (World Wide Web, 1989). Prvním webovým robotem, který byl vyvinut na MIT k měření velikosti webu, byl World Wide Web Wanderer (napsaný v jazyce Perl) z roku 1993. [citace z Wiki - World Wide Web Wanderer]

O něco později, v roce 2000, se ve velkém začala používat webová APIs – lidé mohli získávat čistá data přímo od serveru a scraping se tak stal o hodně jednodušším.

Dalším milníkem v historii web scrapingu je rok 2004, kdy byla vydána knihovna pro parsování HTML a XML dokumentů Beautiful Soup pro programovací jazyk Python. Ta je do dnes považována za nejsofistikovanější a nejpokročilejší knihovnu pro web scraping.

Za zmínku stojí určitě i rok 2006, kdy je datován příchod vizuálního web scrapingu, tedy techniky, kdy uživatel skrze rozhraní aplikace označí klikáním myši, z kterých oblastí webové stránky chce extrahovat data. Tímto se otevřely

dveře web scrapingu pro všechny. [citace z https://www.octoparse.com/blog/web-scraping-introduction]

1.1.2 Techniky

Technik, jak z webové stránky získat data existuje mnoho, podívejme se alespoň na některé z nich:

- vyhledávání na základě textové shody – např. pomocí UNIX nástroje grep nebo regulárních výrazů
- HTML parsování – základní a stále ještě nejpoužívanější technika extrakce dat; informace jednoduše získáváme z HTML elementů, popř. pomocí tříd nebo id
- počítačové vidění, strojové učení, zapojení umělé inteligence – snaha napodobit způsob, jakým vidí a zpracovává webovou stránku člověk; podobný přístup zkouší např. projekt Diffbot
- vizuální web scraping – jak již bylo zmíněno výše, požadovaná data se musí ručně naklikat skrze rozhraní nějaké aplikace (značně to však usnadňuje např. hledání podobných prvků na základě prvních pár kliknutí)
- manuální vyhledávání a stahování dat (někdy nazývané také *copy–paste*)

1.1.3 Využití web scrapingu

Podob pro uplatnění scrapování dat z webu je nespočet, a to obzvlášť v dnešní době, kdy se velikost všech dat na celém internetu pohybuje v řádech Zettabajtů (1024^7 B).[citace z https://www.nodegraph.se/big-data-facts/]. Mezi ty hlavní patří:

- získání kontaktních informací (např. e-mail) pro marketingové účely
- indexování webových stránek (jako příklad můžeme uvést GoogleBot)
- data mining – proces hledání vzorců ve velkých datových setech [odkaz na Wiki]
- monitorování různých proměnných (např. sledování cen nebo hodnocení produktů)
- recyklace již někdy použitých dat za účelem vytváření „nového“ obsahu
- analýza a zpracování dat k výzkumným účelům

1.1.4 Právní stránka

Podrobná právní analýza celé problematiky web scrapingu by vydala na samostatnou diplomovou práci, a tak se pokusíme pouze shrnout základní body a poskytnout čtenáři alespoň náhled do této oblasti.

1.1.4.1 Obsah na webových stránkách a jeho možné využití

Dle ...[citace] může být obsah chráněn zejména jako:

- projev osobní povahy
  - zde lze připomenout § 86 občanského zákoníku – *Nikdo nesmí zasáhnout do soukromí jiného, nemá-li k tomu zákonný důvod. Zejména nelze bez svolení člověka narušit jeho soukromé prostory, sledovat jeho soukromý život nebo pořizovat o tom zvukový nebo obrazový záznam, využívat takové či jiné záznamy pořízené o soukromém životě člověka třetí osobou, nebo takové záznamy o jeho soukromém životě šířit. Ve stejném rozsahu jsou chráněny i soukromé písemnosti osobní povahy.*
  - do této kategorie můžou spadat třeba i komentáře uživatelů na internetovém fóru
- autorské dílo (včetně databáze, viz níže); z autorského zákona lze zmínit:
  - volné užití (§ 30 odst. 1 autorského zákona) – *Za užití díla podle tohoto zákona se nepovažuje užití pro osobní potřebu fyzické osoby, jehož účelem není dosažení přímého nebo nepřímého hospodářského nebo obchodního prospěchu, nestanoví-li tento zákon jinak.*
  - citaci (§ 31 odst. 1 autorského zákona) – *Do práva autorského nezasahuje ten, kdo*
    a) užije v odůvodněné míře výňatky ze zveřejněných děl jiných autorů ve svém díle,
    b) užije výňatky z díla nebo drobná celá díla pro účely kritiky nebo recenze vztahující se k takovému dílu, vědecké či odborné tvorby a takové užití bude v souladu s poctivými zvyklostmi a v rozsahu vyžadovaném konkrétním účelem,
    c) užije dílo při vyučování pro ilustrační účel nebo při vědeckém výzkumu, jejichž účelem není dosažení přímého nebo nepřímého hospodářského nebo obchodního prospěchu, a nepřesáhne rozsah odpovídající sledovanému účelu;

    vždy je však nutno uvést, je-li to možné, jméno autora, nejde-li o dílo anonymní, nebo jméno osoby, pod jejímž jménem se dílo uvádí na veřejnost, a dále název díla a pramen.

- osobní údaje (čl. 2 GDPR)

Je tedy možné shrnout, že použití pro osobní účely (resp. domácí činnosti) je v zásadě neomezené. Nutno ale vyzdvihnout větu „. . . jehož účelem není dosažení přímého nebo *nepřímého* hospodářského nebo obchodního prospěchu . . .“ – například když použijeme získaná data na svém osobním blogu, kde ale máme určitou formu výdělku třeba v podobě reklamy, můžeme se již dopouštět protiprávního jednání.

Důležité je dát si velký pozor také při zpracování a využití osobních údajů, které upravuje čl. 2 GDPR – toto téma by samo vydalo na několik desítek stránek, a tak se jím v této práci nebudeme zabývat a je zde zmíněno jen pro úplnost.

V neposlední řadě poznamenejme, že při vytěžování webu není podstatné, jakým způsobem k získání obsahu došlo (zdali prostřednictvím automatizovaného nebo manuálního postupu). V každém případě lze doporučit provádět web scraping se souhlasem jejich provozovatele (poskytovatele).[citace]

1.1.4.2 Obsah chráněný autorským zákonem

Autorské právo chrání na internetu různý obsah, zejména budou chráněny různé články, obrázky, videa atd. Vždy však bude muset naplňovat znaky autorského díla, tj. bude muset být *jedinečným výsledkem tvůrčí činnosti autora a být vyjádřeno v jakékoli objektivně vnímatelné podobě včetně podoby elektronické, trvale nebo dočasně, bez ohledu na jeho rozsah, účel nebo význam* (viz § 2 odst. 1 autorského zákona).

Taková kritéria však může splňovat i obsah, který by se na první pohled vůbec nemusel zdát chráněný autorským zákonem – jako příklad může posloužit celkové rozvržení stránky (neboli *layout*), který ponese určitý prvek originality a bude na první pohled asociovatelný s danou webovou stránkou.

Naopak výše zmíněnou definici určitě nesplňují různá počítačem generovaná data, tedy například logy chráněné autorským zákonem určitě nebudou.

1.1.4.3 Web scraping a zvláštní práva pořizovatele databáze

Žádný zvláštní zákon věnující se výslovně vytěžování webových stránek neexistuje. Z naší právní úpravy je tomu však nejblíže úprava zvláštního práva pořizovatele databáze (hlava III autorského zákona), která definuje, co to je databáze (§ 88 autorského zákona) – *Databází je pro účely tohoto zákona soubor nezávislých děl, údajů nebo jiných prvků, systematicky nebo metodicky uspořádaných a individuálně přístupných elektronickými nebo jinými prostředky, bez ohledu na formu jejich vyjádření.* Za nezávislé se v tomto kontextu považují prvky, „které lze od sebe oddělit, aniž by tím byl dotčen jejich informační, literární, umělecký, hudební nebo jiný obsah“ (POLČÁK, Radim. Právo informačních technologií. Praha: Wolters Kluwer, 2018. Právní monografie (Wolters Kluwer ČR). ISBN 978-80-7598-045-8.).[citace]

Dále jsou upraveny některé způsoby užití databáze v souladu s autorským zákonem:

- § 91 Omezení zvláštního práva pořizovatele databáze – *Do práva pořizovatele databáze, která byla zpřístupněna jakýmkoli způsobem veřejnosti, nezasahuje oprávněný uživatel, který vytěžuje nebo zužitkovává kvalitativně nebo kvantitativně nepodstatné části obsahu databáze nebo její části, a to k jakémukoli účelu, za podmínky, že tento uživatel databázi užívá běžně a přiměřeně, nikoli systematicky či opakovaně, a bez újmy oprávněných zájmů pořizovatele databáze, a že nezpůsobuje újmu autorovi ani nositeli práv souvisejících s právem autorským k dílům nebo jiným předmětům ochrany obsaženým v databázi.*
- § 92 Bezúplatné zákonné licence – *Do práva pořizovatele jím zpřístupněné databáze též nezasahuje oprávněný uživatel, který vytěžuje nebo zužitkovává podstatnou část obsahu databáze*
  - *a) pro svou osobní potřebu; ustanovení § 30 odst. 3 zůstává nedotčeno,*
  - *b) pro účely vědecké nebo vyučovací, uvede-li pramen, v rozsahu odůvodněném sledovaným nevýdělečným účelem, a*
  - *c) pro účely veřejné bezpečnosti nebo správního či soudního řízení.*

Tedy pokud využíváme databázi čistě pro osobní potřebu či pro věděcké nebo vyučovací účely, kdy uvedeme zdroj, je vše v pořádku. Taktéž pokud využíváme pouze nepodstatnou část obsahu databáze, a pokud tak děláme v souladu s běžným, očekávaným a přiměřeným použitím, nikoliv systematicky a opakovaně, je v pořádku využití dokonce k jakémukoliv účelu. Důležité je zde ale spojení *oprávněný uživatel* – v každém případě musíme mít k datům autorizovaný přístup.

1.1.4.4 Terms of Sevrice, browsewrap, clickwrap, vymahatelnost

1.1.4.5 Důležité soudní případy v Evropě

Mezi nejvýznamnější rozhodnutí na půdě Soudního dvora Evropské unie patří:

C-444/02 kdy společnost Fixtures Marketing Limited (dále jen "Fixtures") žalovala společnost Organismos Prognostikon Agonon Podosfairou AE (dále jen "OPAP") kvůli opakovanému vytěžování rozpisů ligových soutěží ve fotbale v Anglii (které vytvářela, sestavovala a zveřejňovala společnost Fixtures) a jejich následnému užití na webových stránkách společnosti OPAP.[http://curia.europa.eu/juris/docu

Výsledkem jednání bylo, že i *rozpis fotbalových utkání je databází ve smyslu čl. 1 odst. 2 směrnice 96/9* a jako takový může být předmětem ochrany, kdy pořizovatel databáze má právo zabránit vytěžování nebo zužitkování dat. (ačkoliv u společnosti Fixtures nebyl prokázán podstatný vklad, jenž by mohl odůvodnit poskytnutí ochrany).[http://curia.europa.eu/juris/document/document.jsf;jsessionid=B163

C-30/14 kdy společnost Ryanair Ltd. žalovala společnost PR Aviation BV kvůli automatizovanému sběru dat o cenách, letech a letových řádech, jež byly volně přístupné spotřebitelům z webových stránek Ryanair Ltd (k přístupu musel uživatel potvrdit souhlas se všeobecnými podmínkami). Tyto data pak společnost PR Aviation BV používala ke srovnávání cen na svém vlastním portálu.[http://curia.europa.eu/juris/document/document.jsf?text=&docid=161388&pageIndex=0&

Soudní dvůr rozhodl ve prospěch PR Aviation BV – *jestliže se autor databáze rozhodne poskytnout souhlas s jejím využitím, nic mu nebrání ve stanovení smluvních podmínek*, jež by omezily používání databáze ze strany třetích osob, aniž by přitom bylo dotčeno použitelní vnitrostátní právo.[http://curia.europa.eu/juris/documen

1.1.4.6 Důležité soudní případy ve Spojených státech

Úprava v USA je postavena na jiné právní úpravě, a závěry tak nejsou automaticky přenositelné do našeho práva (což ovšem nevylučuje možnost použití těchto rozhodnutí pro účely argumentace). Těmi nejdůležitějšími rozhodnutími v právní historii této problematiky jsou:

hiQ v. LinkedIn kdy analytická společnost hiQ automatizovaně extrahovala veřejně dostupná (bez nutnosti registrace) data z profilů uživatelů sítě LinkedIn. Ta společnosti hiQ zaslala *cease and desist letter* a požadovala okamžité ukončení činnosti ze strany hiQ pod pohrůžkou porušení CFAA (*Computer Fraud and Abuse Act*).[https://assets.documentcloud.org/documents/3932131/2017-0814-Hiq-Order.pdf]

K dispozici je zatím pouze předběžné opatření ve prospěch hiQ, které říká, že *nelze zabránit přístupu k veřejně dostupným informacím.* hiQ mimo jiné argumentoval i tím, že sociální sítě jsou novodobá veřejná fóra (místa pro veřejné projevy, zejména politického charakteru) a rozhodování o udělení přístupu k nim (ať už ze strany státu či soukromé společnosti) je porušením svobody slova a vyjadřování. [https://www.forbes.com/sites/forbestechcouncil/2017/09/20/linkedin-vs-hiq-ruling-casts-a-long-shadow-over-the-tech-industry/#73e4bada5e6c]

1.1.4.7 Možné postihy

1.1.4.8 Zodpovědnost v případě protiprávního jednání

1.2 Analýza konkurence

První skupinou, na kterou můžeme při hledání na internetu narazit, jsou společnosti, které nabízejí zákazníkům kompletní péči v rámci extrakce dat. Cílí především na velké korporace, jimž postaví scrapovací nástroj přesně na míru, který poté také hostují a spravují. Zákazník tedy dostane data a o nic víc se již nemusí starat. Jako příklad můžeme jmenovat třeba ContentGrabber, Mozenda a další.

Pro nás mnohem relevantnější kategorií je konkurenční nabídku nástrojů poskytujících uživatelům rozhraní k web scrapingu. Budeme se zaměřovat pouze na takové nástroje, které nevyžadují jakoukoli znalost programování – tedy žádné knihovny, API a nástroje pro budování vlastních scraperů.

Mezi ty největší představitele patří ParseHub, Octoparse, WebScraper, Data Scraper a Dexi.io. Čtyři ze zmíněných nástrojů jsou volně dostupné (které mají však velmi omezenou funkcionalitu a pokročilejší operace se odemknou až s určitým platebním plánem – tzv. freemium model) a jeden poskytuje bezplatně pouze 7denní zkušební verzi.

Předtím, než začneme jednotlivé nástroje porovnávat, musíme si určit kritéria, podle kterých budeme hodnotit kvalitu daného nástroje. Především nám půjde o jednoduchost používání, celkovou přehlednost a rychlost, se kterou se uživatel dostane k požadovaným datům. Také nás bude zajímat způsob výběru dat, možnosti exportu získaných dat, jak aplikace sama dokáže uživatele seznámit s používáním a také, v jaké formě se nástroj vůbec používá a čím se od ostatních odlišuje (ať už v pozitivním či negativním smyslu).

Pojďme se tedy na některé nástroje podívat blíže:

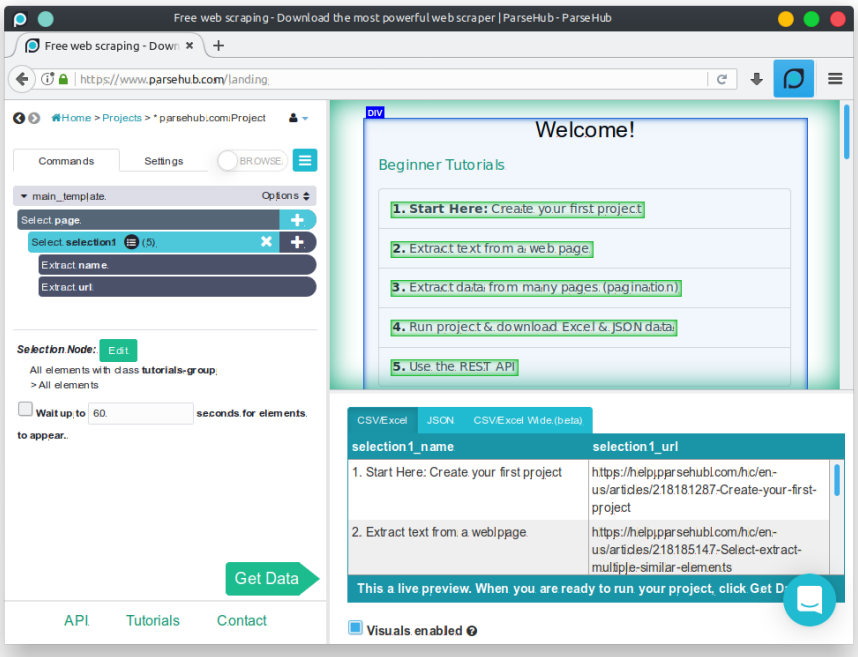
1.2.1 ParseHub

Výhody:

- výběr dat jak pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí), tak pomocí XPath, regulárních výrazů nebo CSS selektorů
- aplikace obsahuje interaktivní tutoriál, který na jednoduchých příkladech ukáže, jak s nástrojem zacházet
- možnost získání dat různými formami - přes API, jako CSV/XLS, do GoogleSheets nebo do Tableau
- různé módy kliknutí (výběr, relativní výběr, kliknutí), zooming in/out na HTML elementy – když se uživatel netrefí (nebo ani trefit nemůže) přesně na požadovaný prvek, lze na něj lehce přejít pomocí této funkce
- automatická rotace IP adresy (tedy nedochází k blokování ze strany serveru)

Nevýhody:

- nutnost stažení aplikace (ale je zde podpora pro Windows, Linux i Mac)
- aplikace je celkově těžkopádná, nemá moc přívětivé uživatelské rozhraní, ovládání působí nepřehledně a přehlceně – na uživatele se vyvalí hodně informací a možností najednou



Obrázek 1.1: ParseHub

1.2.2 Octoparse

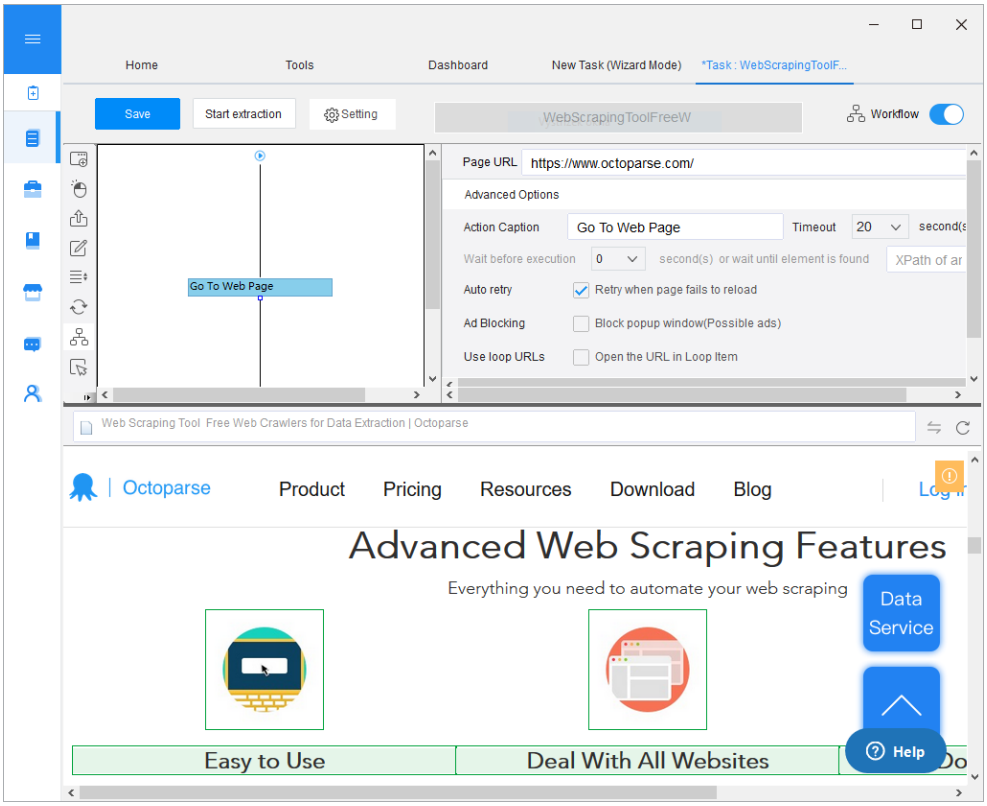
Výhody:

- výběr dat jak pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí), tak pomocí XPath nebo regulárních výrazů
- nástroj obsahuje hotové šablony, které mohou velmi urychlit práci
- pestrá paleta možností (branch judgment, tvoření smyček apod.) – dá se vytvořit téměř jakákoli logika procházení webu a extrakce dat

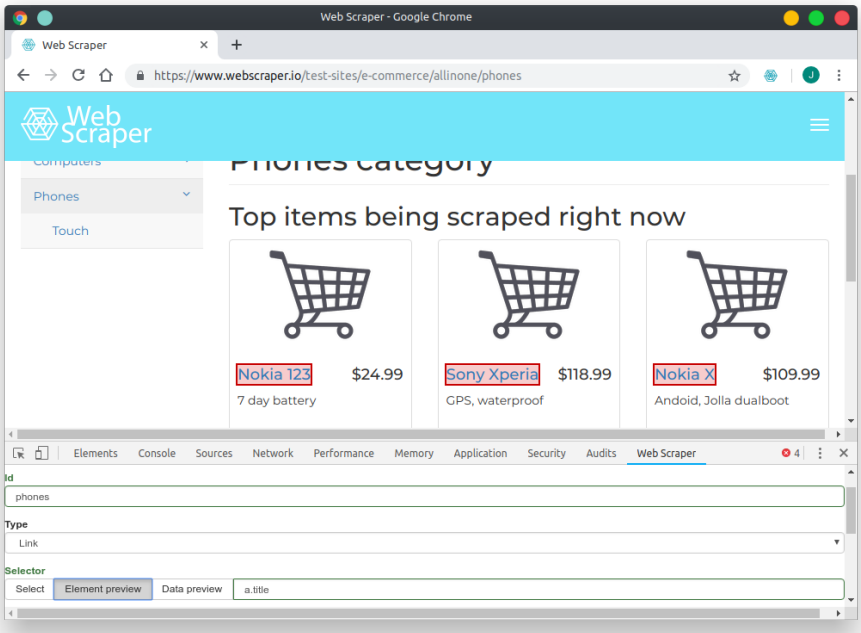
- lehký způsob, jak scrapování automatizovat
- možnost řídit tasky přes API (a získávat tak data taktéž přes API); data jdou nahrát rovnou i do lokální databáze

Nevýhody:

- nutnost stažení aplikace (která je navíc pouze pro Windows)
- těžkopádné a pomalé ovládání, neintuitivní rozhraní
- tutoriál je v podstatě nic neříkající
- připravených šablon je jenom pár a jsou velmi konkrétní



Obrázek 1.2: Octoparse



Obrázek 1.3: WebScraper

1.2.3 WebScaper

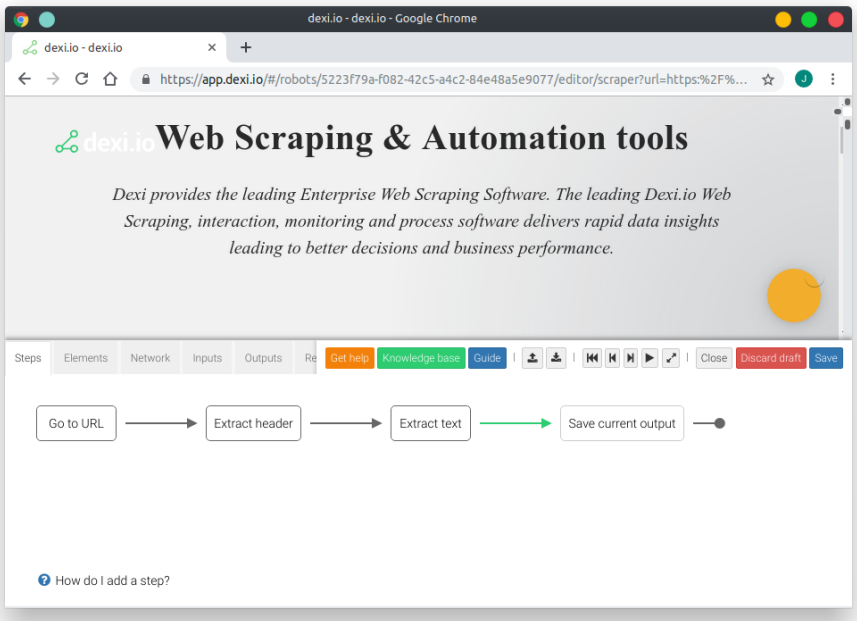
Výhody:

- jednoduchá instalace (jedná se pouze o rozšíření do prohlížeče Google Chrome); scrapování probíhá skrze vývojářskou konzoli
- výběr dat pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí)
- tutoriály jsou formou videí – jednoduché, rychlé a naprosto postačující
- různé typy elementů, které vybíráme (text, odkaz, scroll down), takže lze celkem snadno projít celou doménu
- možnost získání dat různými formami – přes API, jako CSV/XLS nebo do Dropboxu
- klávesové zkratky při výběru elementů velmi usnadňují práci
- možnost využít jejich cloud k automatizaci celého procesu
- oproti konkurenci nabízí přehledné rozhraní, rychlé a jednoduché používání

Nevýhody:

- nutnost používat Google Chrome, což pro některé uživatele může být překážka
- nelze vyhledávat podle klíčových slov ani podle HTML nebo CSS, tudíž všechno se musí manuálně naklikat

1.2.4 Dexi.io



Obrázek 1.4: Dexi.io

Výhody:

- bez nutnosti stahování aplikace – vše se ovládá přes webové rozhraní
- výběr dat jak pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí), tak pomocí HTML, CSS nebo textové shody
- mnoho návodů dostupných na stránkách, interaktivní rádce přímo při scrapování
- všechny možné druhy kliknutí, takže lze lehce projít celou doménu

- možnost exportovat data do CSV, JSON, XLS, získat přes API, poslat do Google Drive, Google Sheets nebo Amazon S3
- různé módy aplikace – scraping, crawler, pipes (skládání menších scrape botů) a autobot (extrahování z více stránek najednou se stejným rozložením); možnost takto automatizovat celý proces.
- nápomocné jsou různé addony (např. na obcházení Captchy)

Nevýhody:

- široká nabídka možností, a tak chvíli trvá, než se člověk zorientuje
- placený nástroj, zadarmo je dostupná pouze týdenní zkušební verze
- úvodní tutoriál je velmi strohý a žádné velké seznámení s nástrojem se nekoná

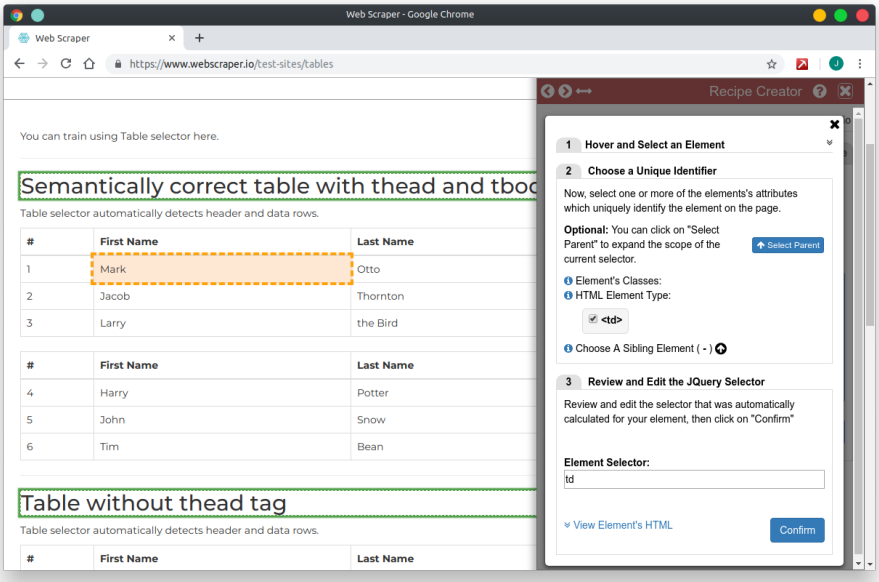
1.2.5 Data Scraper

Výhody:

- jednoduchá instalace (jedná se pouze o rozšíření do prohlížeče Google Chrome).
- velmi jednoduché ovládání a přehledné rozhraní
- výběr dat probíhá pomocí klikání
- klikáním se utváří JQuery selektor, který si uživatel může podle svého upravit a doladit tak drobné detaily, jež by jinak nutně zahltily uživatelské rozhraní (tedy je možné vyhledávat i podle HTML tagů, id, CSS selektorů – zkrátka vše, co umí klasické JQuery)
- různé druhy kliknutí
- možnost spustit na stránce libovolný JavaScriptový kód v rámci scrapování

Nevýhody:

- nutnost používat Google Chrome, což pro některé uživatele může být překážka
- oproti ostatním nástrojům se může zdát velmi chudý na různé funkce



Obrázek 1.5: Data Scraper

1.3 Specifikace požadavků

Jak jsme viděli v předchozí analýze konkurenčních nástrojů, největšími neduhy, které se prolínají napříč valnou většinou aplikací, jsou *těžkopádné uživatelské rozhraní*, *neintuitivní ovládání* a *rychlost* (nebo spíš pomalost), se kterou se uživatel dostane k požadovaným datům. Pro aplikaci, jíž se tato práce zabývá, bude klíčové se výše zmíněným nedostatkům vyhnout a nabídnout jejich přesný opak. Také jsme se přesvědčili, že nejpříjemnější cestou je celou aplikaci ovládat přes webové rozhraní *bez nutnosti stahování a instalace*.

Na druhou stranu se můžeme u konkurence i inspirovat. Za vyzdvižení stojí určitě *různé druhy výběru dat – klikání* přímo na stránce spolu s inteligentním hledáním podobných prvků jistě tvoří mocný mechanismus. Avšak je potřeba zajistit i ostatní způsoby výběru (jako je např. *textová shoda, HTML tagy, CSS selektory*) pro případ, kdy je pouhé klikání zdlouhavé či nevyhovující. Rovněž široký výběr způsobů exportu dat, intuitivní klávesové zkratky a zooming in/out na prvky může uživatelům zpříjemnit práci s nástrojem.

Neméně důležitou vlastností aplikace je také schopnost sebe sama kvalitně, ale svižně představit, *seznámit uživatele s používáním* a poskytnout mu alespoň pro začátek určité vodítko. Pro většinu ovládacích prvků by však mělo platit to stejné, co platí pro správný kód – měly by být tzv. *self-explanatory*. Tedy každému by mělo být na první pohled jasné, co který element dělá.

Pojďme si nyní všechny požadavky shrnout do několika bodů a rozdělit

na funkční a nefunkční:

1.3.1 Funkční požadavky

- uživatelské rozhraní se skládá z hlavní pracovní plochy, kde se bude nacházet uživatelem zadaná stránka a z postranního panelu, obsahující všechny ovládací prvky
- postranní panel skryje tlačítka, formuláře a ostatní elementy k ovládání aplikace do několika záložek – tímto se na uživatele nevyvalí velké kvantum informací a možností najednou; podle potřeby si každý rozbalí tu možnost, kterou potřebuje
- výběr dat bude probíhat těmito způsoby:
  - kliknutím myší na požadované elementy (na základě předchozích kliknutí se program pokusí označit všechny podobné prvky, výběr však půjde uživatelem zrušit)
  - na základě textové shody (uživatel jednoduše zadá text, jenž má být obsažen v extrahovaných datech)
  - pomocí HTML tagů (např. image, header, article), které se budou psát do textového pole
  - pomocí CSS selektorů (třídy, id, hodnota atributu, různé následnosti); k tomu poslouží formulář umožňující vše přehledně zadat
  - na ovládacím panelu nalezneme i tlačítka s hotovými akcemi představující šablonu pro nejpoužívanější operace (stažení všech obrázků ze stránky, všech emailových adres atd.)
- po kliknutí na určitý prvek se tento barevně označí; taktéž všechny již vybrané prvky budou barevně odlišeny, aby bylo jasné, co už je připraveno k extrakci a co ještě ne
- k dispozici bude přibližování/oddalování momentálního výběru pomocí ikony + a − (uživatel klikne na daný element a pomocí této funkce může traverzovat napříč zanořenými prvky oběma směry)
- na základě výběru dat uživatelem se vytvoří určitý filtr (textový řetězec), který může být ručně upraven – půjde tak o alternativu pro zkušenější uživatele, aniž bychom zanesli uživatelské rozhraní přehršlí možností a celé ho tak znepřehlednili
- získaná data půjdou exportovat do formátů JSON, CSV, XLS, pokud se bude jednat o text; v případě obrázků, videí nebo zvukových souborů poskytne aplikace výstup v zabaleném archivu ZIP

1.3.2 Nefunkční požadavky

- půjde o webovou aplikaci běžící v internetovém prohlížeči, tedy nebude nutná žádná instalace
- program se bude skládat ze dvou částí:
  - *frontend* – kód, který poběží u klienta v prohlížeči; představuje celé uživatelské rozhraní aplikace
  - *backend* – kód, který poběží na serveru; bude naslouchat požadavkům a zpracovávat je; zde bude probíhat samotná extrakce dat
- aplikace cílí primárně na celkový zážitek uživatele – grafické rozhraní bude přehledné a co nejjednodušší, ovládání intuitivní
- čas, za který se uživatel dostane k požadovaným datům (tedy čas, který stráví vybíráním dat; nepočítáme čas potřebný ke stažení), bude co nejmenší

1.3.3 Nice to have požadavky

V předchozích dvou sekcích jsme si shrnuli, jaké požadavky by naše aplikace v každém případě měla splňovat a bez nichž by neměla vůbec být uvedena k dispozici uživatelům. Pak tu máme ale také požadavky, které rozhodně zlepšují celkovou kvalitu a pocit z nástroje samotného, avšak nejsou již pro nás vitální a pokud by se jejich implementace nepovedla, aplikace bude stále plně funkční a připravená k použití. Patří sem:

- uživatelské rozhraní aplikace nabídne intuitivní klávesové zkratky pro usnadnění práce – klikání s přidrženou klávesou Ctrl bude automaticky vybírat všechny podobné elementy; Ctrl+ a Ctrl- obstará přibližování/oddalování momentálního výběru; ...
- export dat realizovatelný i do Google Sheets, Google Drive, Dropbox
- interaktivní tutoriál, který v rychlosti představí práci s nástrojem

Kapitola 2

Realizace

Závěr

Literatura

Příloha **A**

Seznam použitých zkratek

HTML HyperText Markup Language

XML eXtensible Markup Language

API Application Programming Interface

MIT Massachusetts Institute of Technology

CSS Cascading Style Sheets

CSV Comma-Separated Values

IP Internet Protocol

JSON JavaScript Object Notation

XLS formát souboru používaný aplikací Microsoft Excel

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