Univerzita Pavla Jozefa Šafárika v Košiciach

Prírodovedecká fakulta

KOORDINOVANÉ ZÍSKAVANIE A EXTRAKCIA DÁT Z WEBOVÝCH PORTÁLOV CEZ SPOLUPRACUJÚCE ROZŠÍRENIA WEBOVÝCH PREHLIADAČOV

DIPLOMOVÁ PRÁCA

Študijný program:	**Informatika**
Študijný odbor:	**9.2.1. informatika**
Školiace pracovisko:	**Ústav informatiky**
Vedúci záverečnej práce:	**RNDr. Peter Gurský, PhD.**

Košice 2018

Bc. Matej Perejda

Univerzita P. J. Šafárika v Košiciach
Prírodovedecká fakulta

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Matej Perejda
Študijný program: Informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.2.1. informatika
Typ záverečnej práce: Diplomová práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Koordinované získavanie a extrakcia dát z webových portálov cez spolupracujúce rozšírenia webových prehliadačov

Názov EN: Coordinated data acuisition and extraction from web portals via collaborative web browsers' add-ons

Cieľ: 1. Porovnanie súčasných spôsobov extrakcie dát z webových portálov najmä z hľadiska schopnosti extrahovať dáta z dynamicky vytváraných webových stránok cez AJAX volania a schopnosti distribúcie procesu prehľadávania a extrakcie.
2, Obohatenie existujúceho rozšírenia webového prehliadača na anotáciu webových stránok o schopnosť prehľadávania a extrakcie dát z webu aj pre dynamické webové stránky simuláciou správania používateľa.
3, Návrh a vytvorenie škálovateľného servera koordinujúceho spoluprácu viacerých inštancií vytvoreného rozšírenia webového prehliadača z cieľa 2.
4, Otestovanie korektnosti a škálovateľnosti vytvoreného riešenia extrakciou reálnych webových portálov.

Literatúra:
[1] Liu, Bing: Web Data Mining: Exploring Hyperlinks, Contents,and Usage Data. Second Edition, ISBN 978-3-642-19459-7, Springer, 2011
[2] Kushmerick, N.: Wrapper induction: efficiency and expressiveness. Artificial Intelligence, 118:15-68, 2000.
[3] Muslea, I., Minton, S. and Knoblock, C.: A hierarchical approach to wrapper induction. Agents-99, 1999.
[4] Cohen, W., Hurst, M., and Jensen, L.: A flexible learning system for wrapping tables and lists in HTML documents. WWW-2002, 2002.
[5] Hsu, C.N., Dung, M.T.: Generating finite-state transducers for semi-structured data extraction from the Web. Information Systems. 23(8): 521-538, 1998.
[6] Chabaľ, V: Poloautomatická extrakcia komentárov z produktových katalógov. Diplomová práca. Košice 2014
[7] Crescenzi, V., Mecca, G., Merialdo,P.: Roadrunner: Towards automatic data extraction from large web sites. In Proceedings of VLDB 2001, pp. 109-118.

Vedúci: RNDr. Peter Gurský, PhD.
Ústav : ÚINF - Ústav informatiky
Riaditeľ ústavu: prof. RNDr. Viliam Geffert, DrSc.

Dátum schválenia:

Obsah

1 Analýza a prehľad existujúcich riešení **3**

1.1 Nástroje na extrakciu dát z webových stránok 4

1.2 Prehľad vlastností existujúcich nástrojov 17

Kapitola 1

Analýza a prehľad existujúcich riešení

Webový scraper je počítačový softvér alebo skript, ktorý navštevuje a sťahuje (crawluje) viaceré webové stránky, extrahuje z nich špecifické informácie (scrapuje) a ukladá ich napríklad do súboru alebo priamo do databázy. Vyextrahované dáta sa následne prevádzajú do štruktúrovaného, strojovo čitateľného formátu, ktorý počítač dokáže jednoducho prečítať a spracovať. Medzi tieto formáty patria napríklad CSV, JSON, XLS (XLSX) a XML súbory.

Scrapovanie webových stránok je technika pristupovania k dátam, ktorá nevyžaduje opätovné prepisovanie alebo kopírovanie dát. Manuálne kopírovanie údajov z webstránok do textového dokumentu alebo Excelovského hárku funguje správne iba pri malom množstve údajov a vyžaduje si značné množstvo času. Pri zbieraní väčšieho objemu dát je potrebná automatizácia, čo poskytujú práve webové scrapovacie nástroje, ktoré dokážu plnohodnotne simulovať aktivitu užívateľa priamo na webstránke, interagovať s prostredím stránky, klikať na tlačidlá, rolovať stránky, nasledovať iné URL odkazy a podobne.

Proces scrapovania je tvorený niekoľkými časťami. Začína nasledovaním všetkých URL adries, ktoré sú súčasťou hlavnej domovskej stránky, vyslaním HTTP požiadaviek a prijatím odpovedí vo forme HTML kódov celých stránok. Prevzaté a načítané HTML kódy stránok sa spracovávajú za pomoci nástroja (parsera), ktorý analyzuje kód a extrahuje z neho požadované údaje do pološtruktúrovanej formy použitím rôznych techník (regulárne výrazy; konvertovanie HTML kódu do stromovej štruktúry a prechádzanie pomocou XPath a CSS selektorov; DOM parsovanie pomocou webového prehliadača a iné). Po extrakcií sú dáta nie vždy v želanom tvare, preto nasleduje

proces čistenia a transformovania dát. Posledným krokom je serializácia údajov podľa požadovaných dátových modelov, ktorej výsledkom sú dáta v štruktúrovaných formátoch.

Využitie webových scraperov je rôznorodé. Užitočné sú napríklad pri sledovaní cien produktov v eshopoch, analýze a prieskume trhu a zbieraní kontaktných údajov. Ich implementácia je súčasťou rôznych produktových porovnávačov, či webových vyhľadávačov.

1.1 Nástroje na extrakciu dát z webových stránok

Počas testovania scrapovacích nástrojov a tvorby tohto prehľadu sme sa zamerali na extrakciu dát z produktových stránok internetového obchodu Alza.sk a porovnávača Heureka.sk. Používali sme bezplatné verzie, prípadne trial verzie platených produktov s obmedzenou funkcionalitou. Balíčky platených nástrojov sa líšili vo funkcionalite, preto nie je úplne zaručené, že náš prehľad je objektívny.

Relevantnosť extrahovaných dát sme zisťovali na produktových stránkach, kde sme testovali funkčnosť a korektnosť extrakcie názvu produktu, jeho ceny s menovou jednotkou a tabuľky produktovej špecifikácie, kde každý riadok tabuľky obsahoval názov atribútu, jemu prislúchajúcu hodnotu a doplňujúce informácie.

Zo všetkých vlastností nástrojov sme skúmali najmä podporu interaktívnej anotácie elementov na webstránkach, automatickú extrakciu dát, relevantnosť výsledkov, možnosti exportovania dát, pristupovanie k výsledkom pomocou API a podporu extrakcie dynamicky načítaných stránok.

Tabuľka špecifikácie produktu sa v eshope Alza.sk nachádza v časti „*Popis*“ produktovej stránky na adrese `www.alza.sk/`[`nazov_produktu`]`.htm#popis`, kde sa okrem popisu produktu vo forme voľného textu a rôznych obrázkov nachádza aj časť tabuľky so špecifikáciou. Po kliknutí na tlačidlo „*Viac parametrov*“ sa zobrazí celá tabuľka s parametrami daného produktu. Pre získanie relevantných údajov bolo preto potrebné využiť scrapovací nástroj, ktorý umožnil plne simulovať aktivitu užívateľa na stránke. V cenovom porovnávači Heureka.sk sa tabuľka špecifikácie produktu nachádza v časti „*Špecifikácia*“ produktovej stránky na adrese `www.`[`katagoria`]`.heureka.sk/`[`nazov_produktu`]`/specifikace/`, ktorá rovnako ako v predchádzajúcom prípade obsahuje okrem samotnej tabuľky aj voľný text s obšírnejším popisom a obrázkami.

Agenty

Agenty je SaaS platforma s ľahko použiteľnými nástrojmi na automatickú extrakciu dát v cloude na prevažne profesionálne a obchodné účely. Umožňuje prevod neštruktúrovaných dát do strojovo čitateľnej štruktúrovanej tabuľky. Pozostáva z rozšírenia prehliadača Chrome (Advanced Web Scraper) a webovej aplikácie na správu konta používateľa, umožňuje vytvárať nových agentov na scrapovanie a ukladať výsledky extrakcií.

Dostupná je 14-dňová trial verzia tohto nástroja, po uplynutí ktorej si užívateľ môže mesačne predplatiť jeden z programov v hodnote 29 – 99 dolárov. Každý z programových balíkov má určité obmedzenia napr. v počte scrapovaných stránok, počte používateľov, histórie uchovávaných výsledkov a inej funkcionality. Pre náročnejších užívateľov existuje možnosť využitia prispôsobeného enterprise balíka na žiadosť.

Užívateľ sa po registrácii prihlási do webovej aplikácie, v ktorej si buď vyberie existujúceho agenta spĺňajúceho jeho požiadavky na extrakciu dát, alebo si vytvorí vlastného. Tvorbu nového agenta zabezpečuje rozšírenie v prehliadači. Pomocou rozšírenia je možné anotovať elementy na stránke, ktoré chceme extrahovať a ich hodnoty uložiť do vlastných pomenovaných stĺpcov tabuľky. Agent sa uloží a jeho monitorovanie je kontrolované z webovej aplikácie. Po dokončení procesu scrapovania sú dáta pripravené v cloude a výsledky je možné exportovať do jedného z formátov CSV, JSON, TSV a XML. Možný je tiež export dát pomocou FTP, prípadne integrácia s externými aplikáciami cez REST API.

Agenty je nástroj s bohatou funkcionalitou. Medzi jeho top vlastnosti patria napríklad distribuované crawlovanie pomocou viacerých strojov, písanie vlastných skriptov k úprave výsledkov do požadovaného formátu, podpora statických ale aj dynamicky načítavaných webstránok. Ďalej je to napríklad plánovanie úloh, zasielanie notifikácií emailom o dokončených úlohách, história výsledkov, stránkovanie (pagination), IP rotation (využitie proxy servera na zmenu IP adresy a zabránenie blokovaniu), vypĺňanie formulárov, podpora prihlasovania do používateľských účtov a iné.

Anotovanie elementov v tabuľke produktovej špecifikácie na stránke Heureka.sk nie je veľmi intuitívne. Nedokázali sme oddelene vyextrahovať atribúty tabuľky a priradiť k nim prislúchajúce hodnoty z tabuľky. Na stránke Alza.sk extrakcia celej tabuľky produktovej špecifikácie funguje správne a Agenty vracia výborné výsledky.

Content Grabber

Content Grabber je ľahko použiteľný, multifunkčný, škálovateľný nástroj slúžiaci na scrapovanie webových stránok. Pozostáva z platenej desktopovej aplikácie dostupnej v 30-dňovej trial verzii. Cena jednorazového poplatku za licenciu je 995 dolárov a zahŕňa údržbu a podporu programu s pravidelnými aktualizáciami.

Aplikácia má jednoduché používateľské rozhranie so vstavaným internetovým prehliadačom. Užívatelia prehľadávajú webové stránky, vytvárajú agentov a interaktívne vyznačujú elementy na webových stránkach. Po označení daného elementu na stránke nástroj vygeneruje jeden jemu prislúchajúci editovateľný XPath. Agenti zabezpečujú zber dát, prevádzajú ich do štruktúrovaných formátov CSV, JSON, XLS, XML a taktiež ich exportujú do rôznych typov databáz. Výsledok extrakcie je možné prehľadávať (filtrovať) použitím Regex výrazov.

Content Grabber dokáže extrahovať obrázky, text, PDF, CSV a Excel súbory, videá a iné typy údajov. Podporuje extrakciu dynamicky načítavaných dát, plánovanie, spúšťanie z príkazového riadka a okrem iného aj použitie vlastných skriptov na úpravu zozbieraných údajov.

Aplikácia vie veľmi dobre extrahovať všetky produktové informácie z oboch testovaných produktových stránok. Pri tabuľke produktovej špecifikácie na stránke Alza.sk má problémy s identifikáciou prvej časti tabuľky a nedokáže tieto informácie pričleniť k zvyšku tabuľky.

Data Miner

DataMiner je rozšírenie prehliadača Google Chrome, ktoré dopomáha pri extrakcii dát z webových stránok a umožňuje ich exportovanie do rôznych formátov. Na identifikovanie elementov stránky využíva interaktívnu anotáciu použitím XPath, JQuery a CSS selektorov.

Užívateľ vytvára šablóny pre extrakciu dát (tzv. recipes), postupnosti inštrukcií, ktoré sa vykonávajú na danej webovej stránke. K dispozícii je viac než 40 tisíc verejných šablón zdieľaných inými užívateľmi na získavanie štruktúrovaných dát z rôznych stránok. Takáto vytvorená šablóna má svoje meno a obsahuje pozície elementov webových stránok, z ktorých majú byť dáta vyextrahované. Šablóna neuchováva žiadne súkromné informácie o užívateľoch, ani o extrahovaných dátach. Užívateľom vytvorená šablóna funguje iba dovtedy, kým sa štruktúra danej webovej stránky výrazne nezmení.

Výsledné dáta sa neukladajú na serveroch, ale iba lokálne, v dočasných súboroch internetového prehliadača a je ich možné exportovať do formátov CSV, TSV, XLS a XLSX. Osobný účet na stránke slúži iba na správu vytvorených šablón a na zakúpenie mesačných programov.

Dostupná je bezplatná verzia, ale aj platené programy. Bezplatná verzia limituje použitie rozšírenia na 500 scrapovaných stránok mesačne, kde po mesiaci sa počítadlo stránok resetuje. Po prípadnom prekročení počtu scrapovaných stránok sa účet zablokuje na neurčito a odomkne sa iba zakúpením niektorého z platených programov. K dispozícii je viacero balíkov platených programov, ktorých mesačná cena je v rozmedzí 20 – 200 dolárov s limitom 500 – 9000 scrapovaných stránok za mesiac.

Rozšírenie Data Miner podporuje načítanie a extrakciu dynamicky načítaného obsahu pomocou AJAX. Užívateľ môže spúšťať vlastné funkcie písané v jazyku JavaScript, ktoré stiahnuté dáta dokážu vyčistiť a previesť do želanej formy (napr. extrahovanie emailov z textu, oddelenie mien od priezvisk a podobne). Data Miner dokáže automaticky sťahovať veľké kolekcie obrázkov z webových stránok, prechádzať stránkami (auto pagination), dopĺňať a odosielať formuláre použitím údajov z predpripraveného CSV súboru a jednoducho extrahovať zoznamy a tabuľky. Ďalšími podporovanými funkciami okrem iného sú: extrakcia emailových adries použitím regulárnych výrazov, podpora UTF-8 kódovania, scrapovanie dát zo stránok, ktoré si vyžadujú prihlásenie do účtu. Nie je možné napríklad priamo zasielať výsledky extrakcií na server, prípadne do databázy, alebo maskovanie IP adresy použitím proxy serverov.

Spôsob extrakcie je nasledovný. Prvým krokom je stiahnutie rozšírenia z obchodu Chrome a jeho inštalácia. Následne užívateľ v prehliadači zadá URL adresu stránky, ktorú chce scrapovať a klikom aktivuje rozšírenie DataMiner. V užívateľskom rozhraní sa vyberie existujúca šablóna pre extrakciu dát, alebo sa vytvorí nová šablóna, kde využitím interaktívnej anotácie označí elementy na stránke, ktoré chce extrahovať. Šablóna sa uloží a po spustení procesu extrakcie je možné zozbierané dáta vyexportovať.

Bezplatná verzia rozšírenia Data Miner použitím správnej konfigurácie šablóny pre extrakciu vracia veľmi dobré výsledky na oboch testovaných webových stránkach. Rozšírenie nemá problémy s anotáciou elementov na produktových stránkach, s extrakciou názvu produktu, jeho ceny a celej tabuľky so špecifikáciou.

Data Toolbar

Data Toolbar je intuitívny nástroj na scrapovanie webových stránok, ktorý automatizuje extrakciu dát priamo v prehliadači. Voľná aj platená verzia tohto nástroja obsahuje rozšírenie pre tri prehliadače Internet Explorer, Firefox a Chrome. Za licenciu užívateľ zaplatí jednorazový poplatok 24 dolárov. Zakúpenie licencie zahŕňa aktualizácie produktu, podporu a nelimituje export zozbieraných dát.

Rozšírenie zabezpečuje výber extrahovaných elementov stránky pomocou interaktívnej anotácie, umožňuje vytvárať skupiny elementov a pridávať rôzne akcie, ktoré sa majú na stránke vykonávať (klikanie, rolovanie, atď.). Na základe pozície anotovaného elementu nástroj vygeneruje XPath, ktorý je možné manuálne upraviť alebo opätovne zadať. Podporuje paralelný zber údajov z vysoko interaktívnych HTML5 a AJAX webových stránok vo viacerých vláknach, sťahovanie obrázkov, stránkovanie (pagination), automatické prihlásenia a vypĺňanie formulárov, infinite scrolling, plánovanie úloh a filtrovanie výsledkov pomocou regulárnych výrazov. Výsledky scrapovania je možné exportovať do formátov CSV, HTML, XLS a XML.

Data Toolbar dobre extrahuje názov a cenu produktov na oboch testovaných stránkach. Pomocou tohto nástroja sa nám podarilo zo stránky Alza.sk vyextrahovať iba časť tabuľky so špecifikáciou produktu. Výsledky extrakcie tabuľky zo stránky Heureka.sk obsahujú okrem správne vyextrahovaných dát aj popisy k jednotlivým atribútom, ktoré slúžia ako pomôcka pre návštevníkov stránky.

Dexi.io

Dexi.io, formálne známy aj ako CloudScrape je webová aplikácia (SaaS) vo forme editora s interaktívnou anotáciou a veľmi pekným užívateľským rozhraním, ktorá užívateľa navádza krok za krokom pri tvorbe vlastných scrapovacích robotov. Podporuje zhromažďovanie dát z ľubovoľných webových stránok v reálnom čase a nevyžaduje sťahovanie žiadnych aplikácií.

Aplikácia je platená, ale k dispozícii je 60-minútová trial verzia. Užívateľ má možnosť výberu z business programových balíkov s cenou od 119 – 699 dolárov, prípadne enterprise riešenie pre extrakciu veľmi veľkého objemu dát z rôznych webových stránok.

Zozbierané dáta sú ukladané na serveroch po dobu 2 týždňov a je ich možné exportovať do cloudu ako napríklad Google Drive alebo vyexportovať do formátov CSV, JSON, SCSV, XLS, XLSX a XML. Taktiež je možné výsledok scrapovania

exportovať pomocou API, FTP alebo SFTP. Funkcionalitu webovej aplikácie je možné rozšíriť pridaním rôznych rozšírení, napr. PostgreSQL, MySQL, Amazon S3 a iné. Tiež umožňuje anonymne pristupovať k dátam webstránok pomocou proxy serverov na skrytie identity.

Nástroj Dexio.io podporuje použitie cyklu na iterovanie cez anotované elementy na stránke, napríklad riadky tabuľky. Nie je ale jednoduché takýmto spôsobom vyextrahovať tabuľku produktovej špecifikácie vo formáte „atribút – hodnota“ z oboch testovaných stránok.

Easy Web Extract

Easy Web Extract je desktopová aplikácia na extrakciu dát webových stránok prevažne pre obchodné účely. Nástroj je vytvorený pomocou technológie .NET, podporuje viacvláknové prehľadávanie až 24 rôznych webových stránok a dokáže získavať aj dynamický obsah načítavaný cez AJAX a JavaScript. Užívateľ vytvára úlohy (tasks), kde interaktívnou anotáciou definuje pozície elementov webstránky určené na extrakciu a výsledok môže exportovať do formátov CSV, HTML, XML alebo ho zaslať priamo na vzdialený server.

Užívateľ má možnosť vyskúšať si nástroj Easy Web Extract v 14-dňovej trial verzii, prípadne si zakúpiť licenciu za 59,99 dolárov. Medzi ďalšie funkcie patrí okrem iného písanie vlastných skriptov na transformovanie extrahovaných dát, podpora infinite scrolling, vstavaný plánovač a filtrovanie obsahu pomocou Regex výrazov. Easy Web Extract automaticky náhodne prerušuje extrakciu, aby sa zabránilo blokovaniu IP adresy pri nadmernom posielaní požiadaviek na danú stránku.

Aplikácia nedokáže správne vyextrahovať tabuľku produktovej špecifikácie, ktorá sa nachádza na Heureka.sk. Naopak, nemá žiadne problémy s extrakciou tabuľky na produktových stránkach internetového obchodu Alza.sk.

Fminer

Fminer je desktopová aplikácia vytvorená v jazyku Python slúžiaca na extrakciu dát z webových stránok a tiež na ich crawlovanie. Hlavnou charakteristikou tejto aplikácie je vizuálne zobrazenie procesu scrapovania vo forme diagramu, zaznamenávanie makier vo vstavanom webovom prehliadači (Macro Designer) a ich následné spúšťanie, čím sa vykonávajú rôzne akcie na webových stránkach. Súčasťou nástroja je interaktívna anotácia elementov stránky a taktiež podpora manuálneho definovanie výberu

pomocou XPath výrazov. Pre každý klikom anotovaný element sa XPath vygeneruje automaticky. Zozbierané dáta je možné exportovať do formátov CSV, HTML, JSON, XLS, XML a taktiež do databáz MySQL, Oracle, SQL a SQLite.

Cena licencie plnej verzie aplikácie je 168 – 248 dolárov, ale k dispozícii je aj trial verzia na 15 dní zadarmo. Okrem základnej funkcionality scrapovania a simulácie aktivity užívateľa na webe, Fminer ponúka pokročilejšie funkcie ako je napríklad spúšťanie vlastného kódu v jazyku Python, extrakcia dát načítaných dynamicky, parsovanie výsledkov použitím regulárnych výrazov, podpora CAPTCHA, plánovač úloh a emailové reporty.

Fminer nemá väčšie problémy s extrakciou dát z produktových stránok Alza.sk. Z tabuľky produktovej špecifikácie na stránkach Heureka.sk nevie extrahovať správne riadky a k nim prislúchajúce hodnoty.

GetData.IO

GetData.IO je rozšírenie prehliadača Chrome. Pomocou interaktívnej anotácie umožňuje scrapovať elementy webovej stránky a exportovať ich do formátov CSV a JSON. Bezplatná (community) verzia programu zdieľa výsledky scrapovania medzi ostatných registrovaných používateľov, pričom dáta sa odstránia po vymazaní účtu. Platený program stojí mesačne 14,99 dolárov, uchováva súkromné výsledky extrakcie a podporuje plánované scrapovanie. GetData.IO poskytuje API na extrakciu dát pre ich zasielanie do externých aplikácií a taktiež umožňuje zasielať užívateľom notifikácie o extrakcii nových dát. Veľkou nevýhodou tohto rozšírenia je, že priamo nepodporuje dynamicky načítavaný obsah pomocou AJAX, ale príslušné API poskytuje funkcie, ktoré môžu dopomôcť pri načítaní takéhoto obsahu.

Interaktívna anotácia riadkov tabuľky produktovej špecifikácie na stránke Heureka.sk nefunguje správne. Anotovanie riadkov nie je intuitívne a nie je možné správne označiť všetky položky tabuľky. Scrapovanie tabuľky na produktových stránkach Alza.sk vracia veľmi dobré výsledky s veľmi malými (akceptovateľnými) chybami.

Grepsr

Grepsr je rozšírenie prehliadača Chrome, ktoré umožňuje jednoduchú extrakciu dát z webstránok použitím intuitívneho anotovania elementov na stránke. Extrahované dáta môžu byť exportované do formátov CSV, JSON, RSS a XLSX alebo odoslané na cloudové úložisko ako je napr. Dropbox, Google Sheets, Amazon S3, prípadne využiť

FTP prenos. Samotný proces scrapovania spúšťa a spravuje webová aplikácia, v ktorej je možné výsledné záznamy skontrolovať a exportovať.

Voľná verzia rozšírenia dovoľuje načítať max. 500 záznamov počas jedného behu a mesačne vyextrahovať max. 1000 záznamov. Podporuje plánovanie, zasielanie upozornení, prístup k API a neukladá históriu uložených dát. Ak užívateľ využije platený program, získa rovnakú funkcionalitu ako pri voľnej verzii s rozdielom, že nie je limitovaný počet načítaných záznamov a počet vyextrahovaných záznamov mesačne nesmie prekročiť 1 mil. záznamov. Cena platenej verzie aplikácie Grepsr je 20 - 250 dolárov na mesiac.

Grepsr podporuje načítanie obsahu scrollovaním stránok (infinite scrolling), stránkovanie (pagination), kliknutie na tlačidlo „viac obsahu“, plánovanie scrapovania a taktiež je dostupné API na správu dát.

Rozšírenie Grepsr funguje na produktových stránkach Heureka.sk správne a darí sa mu bezproblémovo extrahovať celú tabuľku produktovej špecifikácie. Interaktívna anotácia tabuľky produktovej špecifikácie na Alza.sk zlyháva pri identifikovaní hodnôt, ktoré prislúchajú daným atribútom v tabuľke. Rozšírenie má taktiež problémy s kliknutím na tlačidlo „viac parametrov“, čím nedokáže extrahovať ďalšie položky tabuľky.

Helium Scraper

Helium Scraper je desktopová aplikácia na vizuálnu extrakciu dát použitím interaktívnej anotácie elementov na stránkach, ktorá pre každý označený element vygeneruje editovateľný XPath. Zozbierané údaje je možné exportovať do formátov CSV a XML, prípadne odoslať do MySQL databázy alebo k nim pristupovať pomocou API. Trial verzia nástroja je k dispozícii na 10 dní, cena platených programov je od 99 – 699 dolárov v závislosti od počtu licencií a počtu scrapovaných stránok.

Užívateľ definuje množinu položiek webstránky určených na extrakciu a taktiež množinu akcií, ktoré majú byť na stránkach vykonané (navigácia, export, čakania na dáta, prihlásenie do účtov,...). Podporovaná je extrakcia dynamicky načítavaných údajov a viacúrovňová extrakcia dát, t.j. extrakcia z rôznych stránok a spojenie výsledkov do spoločnej tabuľky.

Anotácia produktových informácií na stránkach Heureka.sk a ich export funguje správne. Výber elementov tabuľky produktovej špecifikácie na stránkach Alza.sk nie je veľmi intuitívny a užívateľ musí manuálne vyznačovať väčšinu atribútov v tabuľke,

aby bola celá tabuľka rozpoznaná. Každopádne Helium Scraper vracia dobré výsledky scrapovaných produktov.

Import.io

Import.io je webová aplikácia (SaaS) s moderným používateľským rozhraním, ktorá dokáže pomocou umelej inteligencie a interaktívnej anotácie elementov na webových stránkach konvertovať množstvo dát do štruktúrovaných strojovo čitateľných foriem, bez nutnosti akéhokoľvek programovania.

Užívateľ zadá URL adresu stránky, ktorú chce scrapovať a aplikácia sa sama pomocou strojového učenia pokúsi o automatickú prípravu extraktora. Ak aplikácia nie je schopná samostatne vyhľadať dôležité informácie na stránke, užívateľ môže nakonfigurovať extraktor tak, že samostatne vyznačí elementy na stránke a priradí im akcie, ktoré sa majú vykonať. Výsledok extrakcie sa následne dá exportovať do formátov CSV, JSON, XLSX, prípadne je možné dáta integrovať do vlastných aplikácií pomocou REST API.

Aplikácia je dostupná na vyskúšanie v 7-dňovej trial verzii, ktorá má veľmi limitovanú funkcionalitu. Okrem podpory scrapovania trial verzia nedokáže napríklad vykonávať akcie na stránkach. Cena platených programov je 299 dolárov mesačne alebo 1999 – 9999 dolárov ročne, ktoré sa líšia najmä v počte dopytovaných webstránok a počte stiahnutých obrázkov/súborov.

Import.io dokáže analyzovať dáta pomocou reportov a vizualizácií, vykresliť ich v grafoch a tabuľkách a vytvárať rôzne štatistické predpovede. Okrem iného ponúka extrakciu dynamicky načítavaných dát pomocou AJAX, ukladanie dát do cloudu, automatické stránkovanie (auto pagination), plánovanie, auto IP rotation, emailové notifikácie a pokročilú špecifikáciu XPath výrazov. Načítanie obsahu scrollovaním (infinite scrolling) nepodporuje priamo, ale túto funkciu je možné vykompenzovať zložitejšou konfiguráciou extraktora.

Rovnako ako pri všetkých testovaných nástrojoch sme aj pri tejto aplikácii skúmali relevantnosť extrahovaných dát pomocou trial verzie. Znamená to teda, že sme nemali plne k dispozícii všetky funkcie, ktoré tento nástroj ponúka. Import.io dokáže automaticky detegovať časť produktových informácií (názov, cenu, peňažnú menu, hodnotenie produktu, kategóriu a ilustračný obrázok) na stránkach Heureka.sk. Po upravení extraktora, vyznačením elementov pomocou interaktívnej anotácie, dokáže vyextrahovať všetky hodnoty tabuľky produktovej špecifikácie správne. S extrakciou

produktových informácií zo stránok Alza.sk nemá aplikácia žiadne problémy.

Instant Data Scraper

Instant Data Scraper je voľne dostupné rozšírenie prehliadača Chrome s jednoduchým užívateľským rozhraním, ktoré dokáže automaticky použitím umelej inteligencie identifikovať tabuľky produktových špecifikácií na stránkach a extrahovať ich.

Užívateľ otvorí webovú stránku vo svojom prehliadači, klikom aktivuje rozšírenie a aplikácia automaticky lokalizuje tabuľky produktových špecifikácií. Pred exportovaním výsledkov do formátov CSV alebo XLSX je možné informácie v detegovanej tabuľke upraviť, vymazať jej stĺpce alebo prepísať/upraviť hodnoty. Ak umelá inteligencia na začiatku odhadne tabuľku zle, je možné stlačením príslušného tlačidla pozmeniť lokalizáciu nájdenej tabuľky. Rozšírenie ponúka funkciu stránkovania (pagination) pomocou interaktívnej anotácie a tiež nastavenia limitu čakania na dynamicky načítané dáta. Nepodporuje export výsledkov na server alebo uloženie do databázy.

Instant Data Scraper dokáže správne identifikovať tabuľku produktovej špecifikácie na stránke Heureka.sk, ale pri jej extrakcii vracia okrem správne extrahovaných hodnôt aj prázdne riadky. Okrem tabuľky nedokáže identifikovať názov produktu a jeho cenu. Rozšírenie vie detegovať iba časť tabuľky nachádzajúcej sa na stránke Alza.sk.

Mozenda

Scrapovací nástroj Mozenda pozostáva z desktopovej aplikácie (Agent Builder) na tvorbu lokálnych projektov (agentov) a webovej aplikácie (Web Console) riadiacej spúšťanie agentov na serveri. Užívateľ pomocou desktopovej aplikácie klikaním definuje extrakčné pravidlá a webová aplikácia túto extrakciu spustí, sleduje ju, organizuje a exportuje zozbierané výsledky do formátov CSV, JSON, TSV, XLSX a XML. Spúšťanie agentov na serveroch zabezpečí predídenie možnému zablokovaniu IP adresy z dôvodu veľkého množstva zasielaných požiadaviek.

Ceny mesačných balíkov tohto nástroja sú 300 – 450 dolárov. Balíky majú rôzne obmedzenia napríklad počtu scrapovaných stránok mesačne, počtu vytvorených agentov a registrovaných používateľov, taktiež obmedzenie prístupu k API, veľkosti priestoru cloudového úložiska a iných funkcií. Zakúpenie prispôsobeného balíka s vysokokapacitným cloudovým úložiskom stojí 40 tisíc dolárov ročne. Na vyskúšanie je k dispozícii 30 dňová trial verzia.

Extrahované údaje dokáže užívateľ prostredníctvom webovej aplikácie analyzovať,

vizualizovať, zasielať emailom, exportovať do iných cloudových úložísk (napr. Dropbox a Azure), alebo k dátam pristupovať pomocou API. Mozenda okrem iného podporuje dynamicky načítaný obsah stránok, infinite scrolling, vypĺňanie formulárov, stránkovanie, plánovanie, použitie XPath výrazov (po anotovaní elementu stránky sa vygeneruje jediný editovateľný XPath) a filtrovanie výsledkov pomocou regulárnych výrazov.

Tento nástroj dokáže automaticky rozoznať niektoré dátové štruktúry nachádzajúce sa na stránkach, ako je napríklad cena, dátum, adresa a telefónne čísla. Výsledkom extrakcie oboch testovaných produktových stránok sú správne produktové informácie vrátane tabuľky produktovej špecifikácie, s ktorou nemá Mozenda problém pri jej identifikovaní.

ParseHub

ParseHub je nástroj na scrapovanie webových stránok, ktorý pozostáva z desktopovej a webovej aplikácie. Klientska desktopová aplikácia predstavuje vstavaný webový prehliadač, pomocou ktorého je možné anotovať elementy stránky (podpora zadávania XPath a Regex výrazov) určené na extrakciu a výsledok exportovať do CSV a JSON súborov, prípadne dáta importovať priamo do Dropbox alebo Google Sheets. K dispozícii je taktiež REST API, ktoré je možné využiť pri integrácii dát s externými aplikáciami. ParseHub pri anotovaní a hľadaní podobných elementov na webových stránkach využíva metódy strojového učenia. Dáta vyextrahované pomocou desktopovej aplikácie sa uchovávajú v cloude a webová aplikácia slúži okrem iného k ich prehliadaniu.

Bezplatná verzia umožňuje tvorbu maximálne piatich verejných projektov a extrakciu dát z 200 stránok v priebehu 40 minút. Dáta sú ukladané v cloude po dobu 14 dní a klienti získavajú iba minimálnu podporu. Pri platenej verzii je možné vybrať si z viacerých programov v cenovej relácii od 149 – 499 dolárov. Tie sa líšia v počte extrahovaných stránok za minútu, v možnosti vytvorenia súkromných projektov, dlhšej doby uchovávania dát (14 – 30 dní), plánovača úloh, podpory IP rotation (zmena IP adresy proxy servermi) a iných funkcií. Existuje taktiež možnosť prispôsobenia programu na základe vlastných požiadaviek.

Veľké množstvo rozličných funkcií robí ParseHub veľmi silným nástrojom. Dokáže získavať dáta načítané dynamicky pomocou AJAX a JavaScript, vypĺňať formuláre, prihlasovať sa do účtov, klikať na mapy, vysporiadať sa s pop-up oknami, podporuje infinite scrolling a stránkovanie (pagination).

Bezplatná verzia sa javí ako skvelý nástroj, ktorého desktopová aplikácia má veľmi pekné, prehľadné a jednoduché užívateľské rozhranie. ParseHub dokáže anotovať a scrapovať všetky dôležité produktové informácie na testovaných webových stránkach a vracia veľmi dobre extrahované výsledky.

Scrapinghub (Portia)

Nástroj Scrapinghub, open-source webová aplikácia známa tiež ako Portia, je bezplatný nástroj, ktorý umožňuje vizuálne scrapovať menšie množstvo dát z webstránok bez akýchkoľvek programovacích znalostí. Užívateľ si vytvorí nový projekt a zadá URL adresu stránky, z ktorej chce dáta extrahovať. Následne vytvorí nový sample, tj. Šablónu, v ktorej pomocou interaktívnej anotácie vyberie elementy stránky k extrakcii a k týmto elementom priradí názvy atribútov, do ktorých sa uložia vyextrahované hodnoty jednotlivých anotovaných elementov. Počas anotovania prvkov stránky Portia jednotlivo generuje XPath výrazy, ktoré nie je možné nijak upravovať alebo prispôsobovať. Na základe definovaných elementov Portia vytvorí tzv. spider, ktorý beží vo webovom prehliadači a riadi extrakciu dát. Výsledok extrakcie je možné exportovať do formátov CSV, JSON, TXT a HTML.

Portia veľmi dobre extrahuje produktové informácie a tabuľku so špecifikáciou zo stránok Alza.sk a Heureka.sk.

Visual Web Ripper

Visual Web Ripper je multifunkčný nástroj v podobe desktopovej aplikácie na automatickú extrakciu dát z webových stránok. Hlavným cieľom je tvorba šablón na vykonávanie rôznych akcií a definovanie obsahu na extrakciu pomocou interaktívnej anotácie vstavaného internetového prehliadača (Visual Designer). Podporuje jednoduché scrapovanie dynamicky načítaného obsahu, jeho export do formátov CSV, JSON, PDF, XLS, XML alebo uloženie do rôznych databáz (MySQL, OleDB, SQL, SQLite). API umožňuje priamy prístup k výsledkom extrakcie, vytváranie, úpravy a spúšťanie projektov.

Užívatelia majú k dispozícii 14 dňovú trial verziu, po jej vypršaní je možnosť zakúpenia polročnej licencie v cene 349 dolárov.

Disponuje funkciou oneskoreného zasielania požiadaviek, ktorá použitím vysoko výkonného proxy servera náhodne prideľuje nové IP adresy a zabraňuje blokovaniu procesu scrapovania. Okrem iného podporuje extrakcie údajov zo stránok chránených

CAPTCHA, vypĺňanie formulárov, automatické generovanie XPath výrazov anotovaného elementu (anotovaný výber je možné manuálne upraviť), detekciu duplikovaných dát a ich odstránenie.

Nástroj dokáže samostatne prechádzať webové stránky a zbierať informácie napríklad z produktových katalógov. Visual Web Ripper dokáže okrem názvu a ceny produktu extrahovať iba časť tabuľky produktovej špecifikácie zo stránok Heureka.sk, pričom výsledok extrakcie obsahuje aj popisy k jednotlivým atribútom, ktoré slúžia ako pomôcka pre návštevníkov stránky. Na produktových stránkach Alza.sk funguje Visual Web Ripper v porovnaní s Heureka.sk lepšie a dokáže zozbierať všetky dôležité informácie.

Web Scraper

Web Scraper je open-source nástroj umožňujúci vytvárať plány scrapovania webstránok (sitemaps), ktoré určujú, ako by mala byť daná stránka prechádzaná a ktoré elementy stránky majú byť extrahované. Nástroj obsahuje bezplatné rozšírenie pre prehliadač Chrome a platenú webovú aplikáciu slúžiacu ako cloud na uchovávanie výsledkov extrakcie s cenou 50 - 250 dolárov, ktorá sa odvíja od počtu scrapovaných stránok.

Okrem iného podporuje načítanie a extrakciu dát generovaných dynamicky, pomocou JavaScriptu a AJAX, dokáže simulovať užívateľskú aktivitu na stránkach, klikať na tlačidlá rozbaľujúce ďalší obsah, rolovať stránky (infinite scrolling) a taktiež podporuje stránkovanie (pagination). Výsledky extrakcie je možné exportovať do formátu CSV.

Užívateľ spúšťa rozšírenie v prehliadači Chrome pomocou vývojárskeho módu (klávesa F12), v ktorom vytvára a spravuje svoje plány extrakcií stránok, konfiguruje ich a vytvára selektory použitím interaktívnej anotácie elementov na stránke. Jednotlivé výbery elementov je možné transformovať do želaného tvaru použitím regulárnych výrazov.

Web Scraper, bezplatné rozšírenie prehliadača Chrome, je veľmi dobrý nástroj, ktorý dokáže bezproblémovo extrahovať názov a cenu produktu zo stránok Alza.sk a Heureka.sk. Extrakcia tabuľky produktovej špecifikácie mu robí isté problémy a to v tom, že nedokáže extrahovať atribúty tabuľky a k nim prislúchajúce hodnoty do samostatných stĺpcov a je ich nutné oddeliť parsovaním. Dokáže extrahovať celú tabuľku ako skupinový (grouped) objekt, ktorý vloží do jedného stĺpca, alebo zvlášť extrahuje

atribúty a hodnoty do nových riadkov, alebo atribúty a hodnoty vloží spoločne do jedného stĺpca.

Web Sundew

Web Sundew je desktopová aplikácia, ktorá umožňuje užívateľom automatizovať celý proces extrakcie a ukladania štruktúrovaných informácií z webových stránok. Užívatelia vytvárajú nové projekty pozostávajúce z agentov zabezpečujúcich navigáciu na stránkach, scrapovanie a ukladanie dát. Jednotlivo definované kroky sa vyobrazujú v diagrame, ktorý predstavuje crawlovaciu sieť. Web Sundew podporuje viacvláknové spúšťanie agentov, viacúrovňovú extrakciu dát z dynamických stránok, export výsledkov do formátov CSV, XLS, XML a ich uloženie do databáz MySQL, Oracle a SQL Server. Pre každý anotovaných element na stránke je automaticky vygenerovaný editovateľný Node Path výraz, pomocou ktorého je možné výber bližšie dodefinovať. K dispozícii je taktiež podpora parsovania výsledkov použitím Regex výrazov.

Cena nástroja je od 89 – 2495 dolárov, pričom k dispozícii je aj 15-dňová trial verzia. Najdrahšia (enterprise) verzia ponúka navyše prístup k API, umožňuje spúšťať proces scrapovania na vzdialenom serveri a publikovať extrahované dáta cez FTP.

Web Sundew nedokáže vôbec načítať stránky produktov na Heureka.sk. Výsledok extrakcie produktových stránok Alza.sk je pomerne dobrý, extrahuje názov, cenu a tabuľku špecifikácie produktu, avšak súčasťou extrahovanej tabuľky sú aj podrobné popisy atribútov.

1.2 Prehľad vlastností existujúcich nástrojov

Tabuľky 1.1, 1.2 a 1.3 predstavujú stručný prehľad vybraných vlastností nájdených scrapovacích nástrojov. Medzi porovnávané vlastnosti v tabuľkách sme zaradili typ nástroja a licencie, otvorenosť zdrojového kódu, podporu interaktívnej anotácie elementov na stránkach, automatickú extrakciu stránok, spôsoby exportovania výsledkov, možnosť extrakcie stránok s dynamicky načítaným obsahom (podpora AJAX, JavaScript, simulácia užívateľskej aktivity, nekonečné rolovanie stránok a iné funkcie) a taktiež možnosť integrácie dát do vlastných aplikácií využitím API.

Výskyty podporovaných vlastností sú v tabuľkách označené fajkami (✓) a ich absenciu značia krížiky (×). Miesta v tabuľkách vyznačené otáznikmi (?) hovoria o nezistených informáciách ku ktorým sa nebolo možné dopátrať.

Počas testovania sme sa zamerali hlavne na nástroje, ktoré sú ľahko použiteľné, s veľmi dobrými výsledkami extrakcií, podporujú dynamicky načítavaný obsah a dokážu plne simulovať užívateľskú aktivitu na webových stránkach. Ak bol nájdený nástroj, ktorý poskytoval iba knižnice na tvorbu vlastného scrapera, prípadne nespĺňal vyššie stanovené podmienky, ďalšie z jeho vlastností ďalej neboli skúmané. Tieto vlastnosti sú v tabuľkách označené pomlčkou (-).

Č.	Názov	Typ	Licencia	Open-source	Interaktívna anotácia	Automatická extrakcia	Export	Dynamicky načítavané dáta	API
1	Agenty	chrome rozšírenie, web app (cloud)	platená (14 dní trial)	×	✓	×	CSV, FTP, JSON, TSV, XML	✓	✓
2	Apify	web app	bezplatná, platená	×	×	×	CSV, HTML, JSON, JSONL, RSS, XML	✓	✓
3	Connotate	desktop app, web app	platená	×	✓	×	CSV, databáza, email, HTML, XLS, XML	✓	?
4	Content Grabber	desktop app	platená (30 dní trial) na žiadosť	×	✓	×	CSV, DOCX, FTP, JSON, MySQL, Oracle, PDF, SQL Server, XLS, XML	✓	✓
5	CrawlMonster	web app	bezplatná, platená	×	×	×	email	?	?
6	Data Miner	chrome rozšírenie	bezplatná, platená	×	✓	×	CSV, TSV, XLS, XLSX	✓	×
7	DataScrapper	firefox, chrome rozšírenie	bezplatná	✓	×	×	CSV	×	×
8	Data Toolbar	IE, firefox, chrome rozšírenie	bezplatná, platená	×	✓	×	CSV, HTML, SQL, XLS, XML	✓	×
9	Dexi.io	web app (SaaS)	platená (60 minút trial)	×	✓	×	Amazon S3, CSV, FTP, Google Drive, Google Docs, JSON, SCSV, SFTP, XLS, XLSX, XML	✓	✓
10	Diffbot	web app, knižnice	platená (14 dní trial)	×	×	✓	CSV, JSON	✓	✓
11	Diggernaut (Excavator - visual extractor)	desktop app, chrome rozšírenie, web app (cloud)	bezplatná, platená	×	✓	×	CSV, JSON, XLS	×	✓
12	Easy Web Extract	desktop app	platená (14 dní trial)	×	✓	×	CSV, HTML, SQL server, XML	✓	×
13	Embed.ly	web app	platená (30 dní trial)	×	×	✓	JSON	-	✓
14	Expired Domain Scraper	chrome rozšírenie	bezplatná	×	×	✓	-	podpora iba Youtube, Google, Bing, Yandex	×
15	Fminer	desktop app	platená (15 dní trial)	×	✓	×	CSV, FTP, HTML, JSON, MS SQL, MySQL, Oracle, SQL, SQLite, XLS, XML	✓	×
16	GetData.IO	chrome rozšírenie	bezplatná, platená	×	✓	×	CSV, JSON	×(iba pomocou API)	✓
17	Grepsr	chrome rozšírenie, web app	bezplatná, platená	×	✓	×	CSV, Dropbox, FTP, Google Docs, JSON, RSS, XLSX, Web hooks	✓	✓
18	Handy Web Extractor	desktop app	bezplatná	×	×	×	×	×	×
19	Helium Scraper	desktop app	platená (10 dní trial)	×	✓	×	CSV, MySQL, XML	✓	✓
20	iMacros	IE, firefox, chrome rozšírenie, desktop app	platená (30 dní trial)	×	✓	×	CSV, databáza, TXT, XML	✓	✓

Tabuľka 1.1: Prehľad vlastností existujúcich nástrojov

Č.	Názov	Typ	Licencia	Open-source	Interaktívna anotácia	Automatická extrakcia	Export	Dynamicky načítavané dáta	API
21	Import.io	web app (cloud)	platená (7 dní trial)	×	✓	×	CSV, JSON, XLSX	✓	✓
22	Instant Data Scraper	chrome rozšírenie	bezplatná	×	×	✓	CSV, XLSX	✓	×
23	KantuX	desktop app	bezplatná, platená	×	✓(OCR)	×	CSV	✓	✓
24	Kido Scraper Generator	chrome rozšírenie	bezplatná	?	✓	×	-	-	×
25	Morph.io	scrapovacia platforma	platená	✓	-	-	CSV	-	✓
26	Mozenda	web app (cloud), desktop app (agent builder)	platená (30 dní trial)	×	✓	×	CSV, JSON, TSV, XLSX, XML	✓	✓
27	myTrama	chrome rozšírenie, web app	platená (trial)	×	✓	×	CSV, HTML, JSON, PDF, XML	-	✓
28	Octoparse	desktop app, cloud	bezplatná, platená	×	✓	×	CVS, HTML, MySQL, Oracle, SQL, TXT, XLS	✓	✓
29	OutWit Hub	desktop app, firefox rozšírenie	bezplatná, platená	×	×	✓	CSV, FTP, HTML, JSON, SQL, TXT, XLS	✓	×
30	ParseHub	desktop app, web app (cloud)	bezplatná, platená	×	✓	×	CSV, Dropbox, Google Sheets, JSON	✓	✓
31	PhantomJS	WebKit, cloud	bezplatná, platená	✓	-	-	-	-	✓
32	QuickCode (ScraperWiki)	web app (IDE)	?	✓	-	-	-	-	✓
33	Rank Scraper	chrome rozšírenie	bezplatná	×	?	?	?	?	×
34	Regex Scraper	chrome rozšírenie	bezplatná	?	×	×	HTML	?	×
35	RegexSearch	firefox rozšírenie	bezplatná	✓	×	×	?	?	×
36	ScrapBook	firefox rozšírenie	bezplatná	×	×	×	-	-	×
37	Scrape.it	chrome rozšírenie, web app (cloud)	platená (7 dní trial)	?	✓	×	?	×	✓
38	ScrapeHero	service	platená	×	?	?	Amazon S3, Cassandra, CSV, databáza, Dropbox, DynamoDB, FTP, Hadoop, Hbase, JSON, MongoDB, MySQL, Oracle, XML	✓	✓
39	Scraper	chrome rozšírenie	bezplatná	?	×(iba XPath referencie)	×	Google Docs	?	×
40	Scraper Crawler	chrome rozšírenie	bezplatná, platená	×	×	✓	?	?	×

Tabuľka 1.2: Prehľad vlastností existujúcich nástrojov

Č.	Názov	Typ	Licencia	Open-source	Interaktívna anotácia	Automatická extrakcia	Export	Dynamicky načítavané dáta	API
41	Scrapinghub (Portia)	web app	bezplatná	✓	✓	×	CSV, JSON, TXT, XML	✓	×
42	Scrapy	framework	bezplatná	✓	-	-	CSV, FTP, JSON, XML	-	✓
43	Screen Scraper	desktop app, chrome rozšírenie	bezplatná, platená (30 dní trial - vyžaduje kreditnú kartu)	×	×	×	CSV, databáza, HTML, JSON, MySQL, SQL, TXT, XML	✓	✓
44	UIPath	desktop app (Studio), chrome, firefox rozšírenie, web app (Orchestrator)	platená (60 dní trial) na žiadosť	×	✓	×	CSV, email, XLS	✓	✓
45	uScraper	web app	bezplatná, platená	-	×	✓	CSV	-	×
46	Visual Web Ripper	desktop app	platená (14 dní trial)	×	✓	×	CSV, JSON, MySQL, OleDB, Oracle, PDF, SQL, SQLite, XLS, XML	✓	✓
47	Web Content Extractor	desktop app	platená (14 dní trial)	×	✓	×	CSV, FTP, HTML, HTTP, MS Access, MySQL, ODBC, SQL, TXT, XLS, XML	✓	×
48	Web Data Extractor (Pro)	desktop app	platená (15 dní trial)	×	✓	×	CSV, TXT, XLSX	×	×
49	WebHarvy Web Scraper	desktop app	platená (15 dní trial)	×	✓	×	CSV, JSON, MySQL, MS SQL, Oracle, TSV, XLS, XML	✓	×
50	Webhose.io	web app	bezplatná, platená	×	×	×	JSON, RSS, XLS, XML	-	✓
51	Web Robots Scraper	chrome rozšírenie (scraping IDE)	bezplatná	×	×	×	CSV, database, server, XLSX	-	×
52	Web Scraper	chrome rozšírenie, web app (cloud)	bezplatná (rozšírenie), platená (cloud)	✓	✓	×	CouchDB, CSV, Dropbox	✓	×
53	WebSundew	desktop app	platená (15 dní trial)	×	✓	×	CSV, email, FTP, MySQL, Oracle, RSS, SQL Server, XLS, XML	✓	✓
54	WinAutomation	desktop app	platená (30 dní trial)	×	✓	×	CSV, FTP, XLS	✓	✓
55	80legs	web app	bezplatná, platená	×	×	×	CSV, JSON	-	✓

Tabuľka 1.3: Prehľad vlastností existujúcich nástrojov

Zoznam použitej literatúry

[1] *Agenty.*[online].Dostupné na: https://www.agenty.com/

[2] *Apify.*[online].Dostupné na: https://www.apify.com/

[3] *Connotate.*[online].Dostupné na: https://www.connotate.com/

[4] *Content Grabber.*[online].Dostupné na: https://contentgrabber.com/

[5] *CrawlMonster.*[online].Dostupné na: https://www.crawlmonster.com/

[6] *Data Miner.*[online].Dostupné na: https://data-miner.io/

[7] *DataScrapper.*[online].Dostupné na: https://github.com/rakot/DataScrapper

[8] *Data Toolbar.*[online].Dostupné na: http://datatoolbar.com/

[9] *Dexi.io.*[online].Dostupné na: https://dexi.io/

[10] *Diffbot.*[online].Dostupné na: https://www.diffbot.com/

[11] *Diggernaut.*[online].Dostupné na: https://www.diggernaut.com/

[12] *Easy Web Extract.*[online].Dostupné na: http://webextract.net/

[13] *Embed.ly.*[online].Dostupné na: http://embed.ly/

[14] *Expired Domain Scraper.*[online].Dostupné na: http://expireddomainscraper.com/

[15] *Fminer.*[online].Dostupné na: http://www.fminer.com/

[16] *GetData.IO.*[online].Dostupné na: https://getdata.io/

[17] *Grepsr.*[online].Dostupné na: https://www.grepsr.com

[18] *Handy Web Extractor*.[online].Dostupné na: http://scraping.pro/handy-web-extractor/

[19] *Helium Scraper*.[online].Dostupné na: http://www.heliumscraper.com

[20] *iMacros*.[online].Dostupné na: https://imacros.net/

[21] *Import.io*.[online].Dostupné na: https://www.import.io/

[22] *Instant Data Scraper*.[online].Dostupné na: https://webrobots.io/instantdata/

[23] *KantuX*.[online].Dostupné na: https://a9t9.com/

[24] *Kido Scraper Generator*.[online].Dostupné na: https://github.com/kidozen/kido-scraper-script-generator

[25] *Morph.io*.[online].Dostupné na: https://morph.io/

[26] *Mozenda*.[online].Dostupné na: https://www.mozenda.com/

[27] *myTrama*.[online].Dostupné na: https://www.mytrama.com

[28] *Octoparse*.[online].Dostupné na: https://www.octoparse.com/

[29] *OutWit Hub*.[online].Dostupné na: http://outwit.com/

[30] *ParseHub*.[online].Dostupné na: https://www.parsehub.com/

[31] *PhantomJS*.[online].Dostupné na: http://phantomjs.org/

[32] *QuickCode*.[online].Dostupné na: https://quickcode.io/

[33] *Rank Scraper*.[online].Dostupné na: https://chrome.google.com/webstore/detail/rank-scraper/hllkjcahamhmchkolednmieeifkimiif

[34] *Regex Scraper*.[online].Dostupné na: https://chrome.google.com/webstore/detail/regex-scraper/akjalgjglcdpomokfhgcmononebebioc

[35] *RegexSearch*.[online].Dostupné na: https://github.com/Mohd-PH/RegexSearch/

[36] *ScrapBook*.[online].Dostupné na: http://www.xuldev.org/scrapbook/

[37] *Scrape.it.*[online].Dostupné na: https://chrome.google.com/webstore/detail/scrapeit-web-scraping-sof/dahgmkmenffebhjdnmlemhbjiocpbbon

[38] *ScrapeHero.*[online].Dostupné na: https://www.scrapehero.com/

[39] *Scraper.*[online].Dostupné na: https://chrome.google.com/webstore/detail/scraper-with-pagination/dbmjjlbalcbgmlniekckllnaiejomiop

[40] *Scraper Crawler.*[online].Dostupné na: http://scrapercrawler.com/

[41] *Scrapinghub (Portia).*[online].Dostupné na: https://scrapinghub.com/

[42] *Scrapy.*[online].Dostupné na: https://scrapy.org/

[43] *Screen Scraper.*[online].Dostupné na: http://www.screen-scraper.com/

[44] *Software for Web Scraping.*[online].Dostupné na: http://scraping.pro/software-for-web-scraping/

[45] *UIPath.*[online].Dostupné na: http://www.uipath.com/

[46] *uScraper.*[online].Dostupné na: https://uscraper.com/

[47] *Visual Web Ripper.*[online].Dostupné na: http://visualwebripper.com/

[48] *Web Content Extractor.*[online].Dostupné na: http://www.newprosoft.com/

[49] *Web Data Extractor.*[online].Dostupné na: http://www.webextractor.com/

[50] *WebHarvy Web Scraper.*[online].Dostupné na: https://www.webharvy.com/

[51] *Webhose.io.*[online].Dostupné na: https://webhose.io/

[52] *Web Robots Scraper.*[online].Dostupné na: https://webrobots.io/

[53] *Web Scraper.*[online].Dostupné na: http://webscraper.io/

[54] *WebSundew.*[online].Dostupné na: http://www.websundew.com/

[55] *WinAutomation.*[online].Dostupné na: http://www.winautomation.com/

[56] *9 FREE Web Scrapers That You Cannot Miss.*[online].Dostupné na: https://www.octoparse.com/blog/9-free-web-scrapers-that-you-cannot-miss/

[57] *10 Web Scraping Tools to Extract Online Data.*[online].Dostupné na: https://www.hongkiat.com/blog/web-scraping-tools/

[58] *80legs.*[online].Dostupné na: http://80legs.com/