Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Ilkovičova 2 , 842 16 Bratislava

Extrakcia dát z webu

[WebExtraction]

Modul Extraction management

Tím: č. 16, WebX
Vedúci tímu: Ivan Srba
Členovia tímu: Ján Brechtl, Tomáš Juhaniak, Martin Kalužník, Rastislav Krchňavý, Michal Kren, Martin Lacek, Andrej Vaculčiak
Akademický rok: 2016/2017
Autor: Ján Brechtl, Tomáš Juhaniak, Martin Kalužník, Rastislav Krchňavý, Michal Kren, Martin Lacek, Andrej Vaculčiak
Verzia číslo: 1.2
Dátum poslednej zmeny: 15.05.2017

1 Úvod

Základom celého procesu extrakcie je pravidelné spúšťanie vopred definovaného skriptu cralwerom, ktorý vykonáva samotnú extrakciu z konkrétnej stránky na základe spomínaného sriptu. Obsahom tohto dokumentu je opis záležitostí, ktoré súvisia s týmto procesom, spolu so spôsobom zobrazovania extrahovaných dát.

2 Extraction management

2.1 Analýza

V predchádzajúcom module bol opísaný manažment projektov a skriptov. Tie tvoria akúsi predlohu pre extrakciu dát, ale samotná extrakcia sa vykonáva v module extrakcií. Obsahuje najmä extrahovanie dát z webu, t.j. crawlovanie, a spúšťač jednotlivých skriptov, t.j. scheduler.

2.2 Návrh

Z výsledku analýzy sme dospeli k návrhu modulu pre extrakciou dát. Hlavnou súčasťou je crawler, ktorý vykoná extrakciu podľa zadaného skriptu. Crawler je univerzálny pre všetky typy skriptov. Výsledok extrakcie bude uložený v databáze spolu s informáciou o správnom dokončení extrakcie.

Celý systém sme navrhovali s cieľom umožniť používateľovi prispôsobiť si extrakciu svojim potrebám. Skripty sú založené XPath dopytoch, ktoré určujú požadované elementy na stránke, pričom každému takémuto elementu môže používateľ definovať špeciálne správanie, tzv. postprocessing. Týmto spôsobom sa dá definovať navigácia po stránkach, obmedzenie vyakonávanie XPath dopytov len na určitú časť stránky, stránkovanie, vyplnenie prihlasovacieho formulára atď. Používateľ si tieto možnosti môže navoliť prostredníctvom rozšírenia do prehliadača, čím sa vytvorí definícia skriptu v JSON notácií. Skript sa v takejto forme zobrazuje vo webovom rozhraní, kde ho pokročilí používatelia môžu upravovať. Definícia skriptu predstavuje istú formu doménovo špecifického jazyka (angl. DSL), ktorý je zrozumiteľný aj pre bežného človeka a zároveň priamo spracovateľný systémom. Detailný popis postprocessing-ov a konkrétne príklady skriptov predstavíme v kapitole 2.3.

Samotným prvkom je plánovanie extrakcií. Podľa požiadaviek zákazníka si môže používateľ nastaviť viacero časových spúšťaní skriptov. Na začiatku si používateľ zadefinuje čas prvého spustenia a interval, teda časové obdobie za ktoré sa extrakcia zopakuje a periódu teda počet intervalov.

Príklad nastavenia extrakcie:

- čas prvého spustenia: 21.12.2016 15:00
- perióda: deň
- interval: 3

Pri takomto nastavení sa budú extrakcie spúšťať:

- 21.12.2016 15:00; 24.12.2016 15:00; 27.12.2016 15:00; ...

2.3 Implementácia

Vytvorili sme vo webovej aplikácií službu na spustenie konkrétneho skriptu, čím sa stiahnu požadované údaje podľa dátovej schémy projektu z vopred definovanej URL. Využili sme pri tom gem Mechanize, ktorý stiahne zo zadanej URL HTML dokument a následne naň aplikuje Xpath dopyty.

Spracovanie skriptu má na starosti modul Crawler, ktorý na základe inštrukcii zo skriptu navštevuje stránky a sťahuje z nich požadované informácie. Tieto inštrukcie okrem počiatočnej URL a elementov na stiahnutie určujú aj správanie crawlera pomocou vyššie spomenutých postprocessingov. Ide o preddefinované spracovanie dát extrahovaných z daného elementu. Konkrétne možnosti sú:

Nested - extrakcia sa vnorí o úroveň nižšie na URL z daného elementu. Na príklade nižšie vidno XPathom definovaný odkaz v <a> elemente. Po vnorení sa extrahujú elementy špecifikované v časti *data*.

```
"name": "zakazka",
    "xpath": "//*[@id=\"lists-table\"]/tbody/tr/td[1]/a",
    "postprocessing": [
     {
       "type": "nested",
       "data": []
```

Restrict - vykonávanie XPathov sa obmedzí na časť stránky. Ide o podobný princíp ako v prechádzajúcom prípade, ibaže stránky pre extrakcie nižšej úrovne nie sú dané URL ale vymedzením častí aktuálnej stránky. Tento príklad udáva, že každý riadok zvolenej tabuľky sa bude spracovávať ako samostatná entita, podľa ktorej sa budú zoskupovať dáta stiahnuté z daného riadku tabuľky.

```
"name": "document",
    "xpath": "//table[@id='lists-table']/tbody/tr",
    "postprocessing": [
     {
       "type": "restrict",
       "data": []
```

Pagination - tento postprocessing umožňuje definovať element, ktorý sa odkazuje na ďalšiu stránku. Atribútom *limit* si používateľ špecifikuje, koľko krát sa má prejsť na ďalšiu stránku.

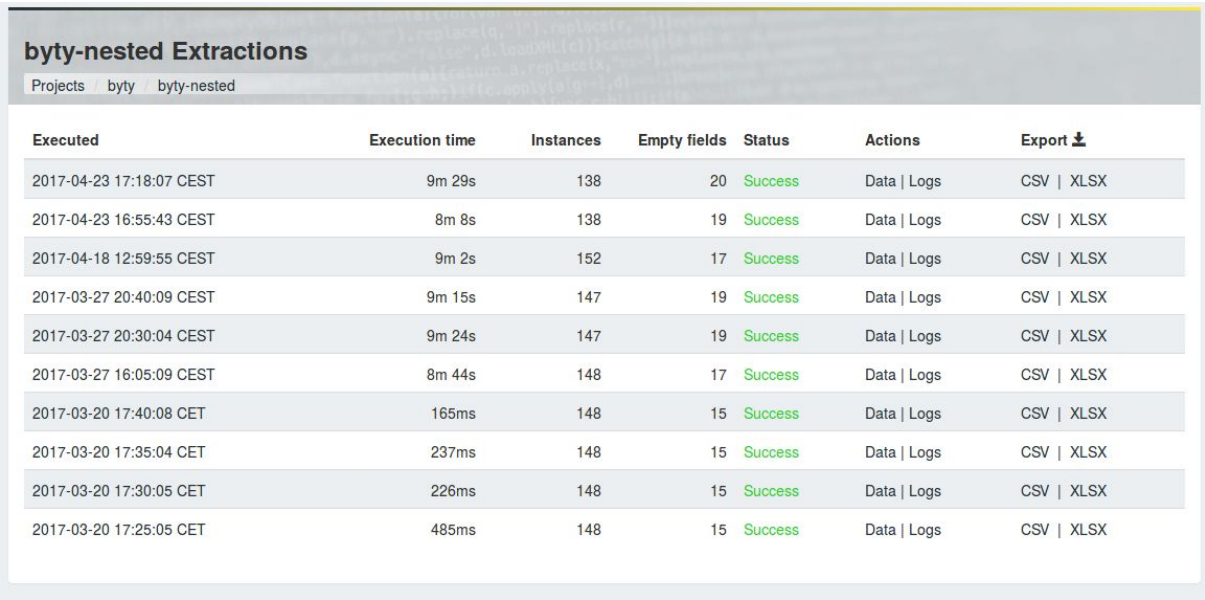
```
"name": "next_page",
    "xpath": "//i[contains(@class, 'fa-angle-right')]/parent::a/@href",
    "postprocessing": [
     {
       "type": "pagination",
       "limit": 46
     }
```

Post - tento postprocessing umožňuje definovať formulár a prihlasovacie údaju, v prípade že prístup k niektorým údajom si vyžaduje prihlásenie. Najprv je potrebné XPathom definovať formulár na konkrétnej stránke. V ďalších možnostiach je potrebné uviesť

redirect_url, určujúce na akej stránke má pokračovať extrakcia, a taktiež je potrebné uviesť prihlasovacie meno a heslo.

```
"name": "form",
    "xpath": "//form[@id='login-form']",
    "postprocessing": [
     {
      "type": "post",
      "redirect_url": "https://stackoverflow.com",
      "fields": [
       {
        "name": "email",
        "value": "prihlasovaci@mail.com"
       },
       {
        "name": "password",
        "value": "heslo",
       }
      ]
     }
```

Spustením skriptu sa vytvorí záznam v tabuľke extrakcií (obr. 1), ktorý predstavuje logy zo spúšťania skriptov. Záznam v tejto tabuľke obsahuje čas spustenia, trvanie vykonávania a informáciu o tom, či bola extrakcia úspešná.



byty-nested Extractions

Projects / byty / byty-nested

Executed	Execution time	Instances	Empty fields	Status	Actions	Export
2017-04-23 17:18:07 CEST	9m 29s	138	20	Success	Data | Logs	CSV | XLSX
2017-04-23 16:55:43 CEST	8m 8s	138	19	Success	Data | Logs	CSV | XLSX
2017-04-18 12:59:55 CEST	9m 2s	152	17	Success	Data | Logs	CSV | XLSX
2017-03-27 20:40:09 CEST	9m 15s	147	19	Success	Data | Logs	CSV | XLSX
2017-03-27 20:30:04 CEST	9m 24s	147	19	Success	Data | Logs	CSV | XLSX
2017-03-27 16:05:09 CEST	8m 44s	148	17	Success	Data | Logs	CSV | XLSX
2017-03-20 17:40:08 CET	165ms	148	15	Success	Data | Logs	CSV | XLSX
2017-03-20 17:35:04 CET	237ms	148	15	Success	Data | Logs	CSV | XLSX
2017-03-20 17:30:05 CET	226ms	148	15	Success	Data | Logs	CSV | XLSX
2017-03-20 17:25:05 CET	485ms	148	15	Success	Data | Logs	CSV | XLSX

Obr. 1 - Zobrazenie tabuľky extrakcií

Z tejto tabuľky sa dá prekliknúť na zoznam extrahovaných dát, ktorý obsahuje páry - pole dátovej schémy a extrahované dáta (obr. 2) alebo na zoznam logov (obr. 3).

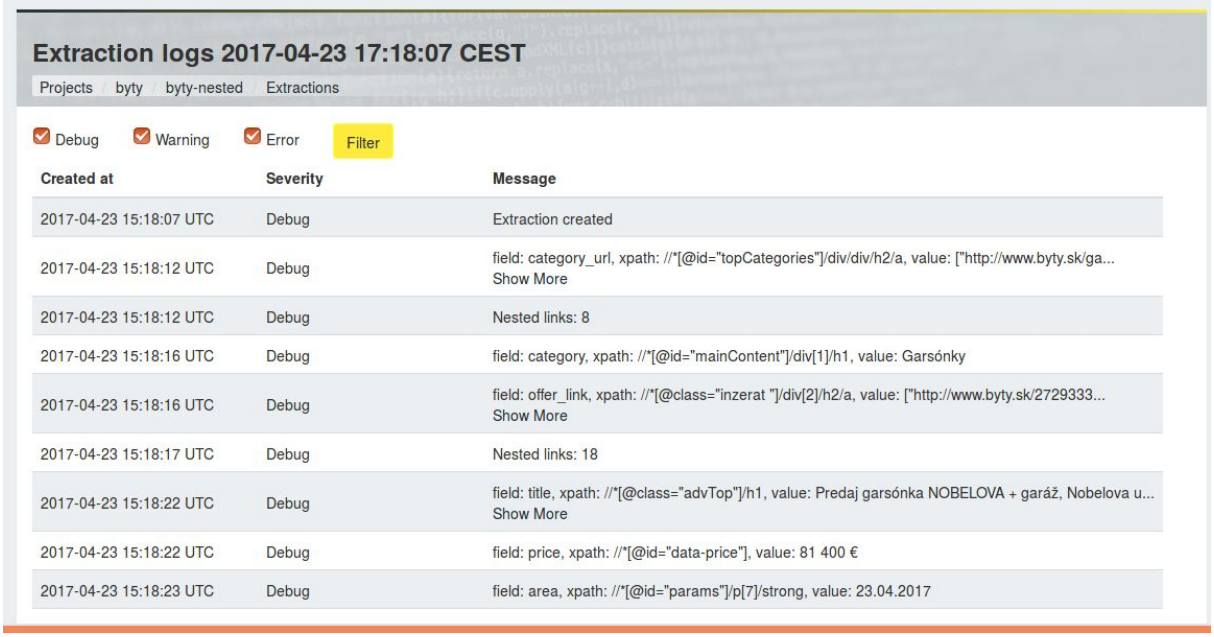
Extraction 2017-04-23 17:18:07 CEST

Projects / byty / byty-nested / Extractions

Export CSV | XLSX

instance_id	url	category_url	category	offer_link	title	price	area	condition
1740	http://www.byty.sk/	http://www.byty.sk/garsonky	Garsónky	http://www.byty.sk/2729333/predaj-garsonka-nobelova-garaz-nobelova-ulica-ba-iii-nove-mesto	Predaj garsónka NOBELOVA + garáž, Nobelova ulica, BA III. Nové Mesto	81 400 €	23.04.2017	40 m²
1741	http://www.byty.sk/	http://www.byty.sk/garsonky	Garsónky	http://www.byty.sk/2772838/olymp-garzonka-na-rovniankovej-s-loggiou-v-zateplenom-dome-na-2-p	OLYMP - Garzónka na Rovniankovej s loggiou v zateplenom dome na 2.p.	69 000 €	23.04.2017	24 m²
1742	http://www.byty.sk/	http://www.byty.sk/garsonky	Garsónky	http://www.byty.sk/2779723/novostavba-ruzinov-polarna-ul-iba-65-000-eur	NOVOSTAVBA, Ružinov, Polárna ul.-----iba 65 000 Eur	65 000 €	23.04.2017	24 m²
1743	http://www.byty.sk/	http://www.byty.sk/garsonky	Garsónky	http://www.byty.sk/2351252/hladam-do-prenajmu-garsonku	Hľadám do prenájmu Garsónku	450 €/mes.	20 m²	Pôvodný stav
1744	http://www.byty.sk/	http://www.byty.sk/garsonky	Garsónky	http://www.byty.sk/2786535/garsonka-na-prenajom-vlcince-ii	Garsónka na prenájom - Vlčince II.	270 €/mes.	20 m²	Kompletná rekonštrukcia
1745	http://www.byty.sk/	http://www.byty.sk/garsonky	Garsónky	http://www.byty.sk/2780272/starbrokers-utulna-podkrovna-garsonka-	StarBrokers - Útulná podkrovná garsónka na	68 500 €	23.04.2017	23 m²

Obr. 2 - Detail extrakcie



Extraction logs 2017-04-23 17:18:07 CEST

Projects / byty / byty-nested / Extractions

Debug Warning Error Filter

Created at	Severity	Message
2017-04-23 15:18:07 UTC	Debug	Extraction created
2017-04-23 15:18:12 UTC	Debug	field: category_url, xpath: //*[@id="topCategories"]/div/div/h2/a, value: ["http://www.byty.sk/ga... Show More
2017-04-23 15:18:12 UTC	Debug	Nested links: 8
2017-04-23 15:18:16 UTC	Debug	field: category, xpath: //*[@id="mainContent"]/div[1]/h1, value: Garsónky
2017-04-23 15:18:16 UTC	Debug	field: offer_link, xpath: //*[@class="inzerat "]/div[2]/h2/a, value: ["http://www.byty.sk/2729333... Show More
2017-04-23 15:18:17 UTC	Debug	Nested links: 18
2017-04-23 15:18:22 UTC	Debug	field: title, xpath: //*[@class="advTop"]/h1, value: Predaj garsónka NOBELOVA + garáž, Nobelova u... Show More
2017-04-23 15:18:22 UTC	Debug	field: price, xpath: //*[@id="data-price"], value: 81 400 €
2017-04-23 15:18:23 UTC	Debug	field: area, xpath: //*[@id="params"]/p[7]/strong, value: 23.04.2017

Obr 3. Zoznam logov z extrakcie

Spúšťanie skriptov vykonáva scheduler resque, ktorý sa spúšťa v 5 minútových intervaloch. Pri každom spustení sa vytvorí zoznam skriptov, ktoré sa majú spustiť - podľa frekvencií skriptu sa vypočíta čas jeho nasledujúceho spustenia. Skript sa pridá do zoznamu, ak čas nasledujúceho spustenia prekročil čas spustenia schedulera.

Skript, ktorý má byť spustený v daný čas sa pridá do inej resque fronty, z ktorej sú skritpy následne pridelované crawleru na spracovanie. V tejto fronte vieme zachytiť prípadné zlyhanie crawlera, kedy crawler nezapíže výsledok o neúspešnosti extrakcie.

Okrem spomínaného zobrazenia dát pomocou tabuľky je možné exportovať výsledky extrakcií aj pomocou API. API sa skladá z dvoch častí a to časti pre zobrazenie vykonaných

extrakcií (list) a časti pre exportovanie inštancií (export). Pre používanie oboch častí je potrebné poznať tzv. API key, ktorý nájde používateľ v profile. API sme implementovali pomocou knižnice Grape.

Časť pre zobrazenie extrakcií (list) sa nachádza na URL `/api/export/list`, s typom volania GET, pričom volanie obsahuje nasledujúce parametre:

- token - povinný parameter s hodnotou rovnou API key z profilu
- script_id - povinný parameter, ID skriptu pre zobrazenie extrakcií
- limit - limit počtu extrakcií, predvolená hodnota je 100
- offset - zarážka pre extrakcie podľa počtu, predvolená hodnota je 0
- last_extraction_id - zarážka od poslednej extrakcie
- since - dátum vo formáte ISO 8601 od ktorého chceme nájsť extrakcie

Časťpre export výsledkov extrakcií (export) sa nachádza na URL /api/export/extraction, s typom volania GET, pričom volanie obsahuje nasledujúce parametre:

- token - povinný parameter s hodnotou rovnou API key z profilu
- id - povinný parameter s ID extrakcie, môže nadobúdať aj hodnotu “last”
- script_id - povinný parameter iba v prípade aj “id=last”, ID skriptu
- limit - limit počtu inštancí, predvolená hodnota je 100
- offset - zarážka pre inštancí podľa počtu, predvolená hodnota je 0
- last_instance_id - zarážka od poslednej inštancí
- since - dátum vo formáte ISO 8601 od ktorého chceme nájsť inštancie

2.4 Testovanie

Správne vykonávanie skriptov sme testovali pomocou reálnej webstránky rubygems.org, keďže extrahovať dáta z lokálneho HTML súboru sme nevedeli bež vážnejšieho zásahu do produkčného kódu. Uvedomujeme si riziko, že stránka môže byť nedostupná alebo sa môže zmeniť, a teda bude potrebné tento test časom opravovať.

V module je potrebné testovať aj samotný scheduler a úlohy na pozadí (background joby). Na ich testovanie budú použité gemy ako resque_spec a timecop.