Tím 16 - WebX

Predstavenie tímu

Ján Brechtl - podpora vývoja
Tomáš Juhaniak - dizajn
Martin Kalužník - technológie
Michal Kren - správa verzií
Rastislav Krchňavý - testovanie
Martin Lacek - webová stránka
Andrej Vaculčiak - scrum master, dokumentácia

Ivan Srba - vedúci tímu, product owner

kontakt - tim.16.fiit.1617@gmail.com

Extrakcia dát z webu

Množstvo informácií na webe každým dňom narastá a manuálne postupy ich zbierania už nepostačujú. Systémy pre automatickú extrakciu dát majú veľký potenciál vo vedeckej aj komerčnej sfére, najmä z pohľadu analýzy správania používateľov, alebo dokonca až jeho predpovedanie. Cieľom tohto projektu je vytvoriť systém, ktorý umožní jednoduchú, poloautomatickú extrakciu dát a následnú konverziu do štruktúrovanej podoby. Používateľ má prostredníctvom rozšírenia do internetového prehliadača možnosť definovať extrahovacie skripty alebo označiť požadované elementy na stránke, ktoré náš systém stiahne a spracuje.

Hlavné časti projektu:

- webová aplikácia pre správu používateľov, projektov a extrakcií
- rozšírenie do prehliadača (Chrome) na anotáciu dát
- REST API na komunikáciu rozšírenia s aplikáciou

Prečo extrahovať dáta z webu?

Množstvo dát na webe denno-denne narastá a mnoho ľudí nemá vo svojich dátach prehľad. Používatelia nemajú možnosti neustále sledovať aktuálnosť stránok, ktoré spravujú, alebo sa o ne zaujímajú. Práve náš nástroj by mohol používateľom pomôcť previesť do štruktúrovanej podoby dáta, ktoré je dnes možné zobraziť len ako dokumenty.

Naše riešenie má potenciál pomôcť aj v štátnych IT projektoch, keďže plánujeme extrahovať dáta aj z iných ako html dokumentov. Takto by sme napríklad vedeli zo sérií zmlúv a iných dokumentov ľahko vytvoriť štruktúrované dáta, v ktorých môžu všetci ľahko vyhľadávať. Pri extrakcií dát sa budeme riadiť všetkými zákonnými a etickými pravidlami, keďže si uvedomujeme, že náš nástroj má slúžiť na extrakciu dát a nie na preťažovanie ostatných systémov.

Ako plánujeme náš vývoj

V prvých šprintoch vytvárame prostredie pre registráciu používateľov, vytváranie projektov a definíciu skriptov. V ďalšej fáze sa zameriame na rozšírenie do prehliadača, vďaka ktorému bude možné anotovať informácie, ktoré zaujímajú používateľa a želá si ich spracovanie. Neskôr pribudne spracovanie skriptov na pozadí a možnosť sťahovania dát.

Projekt vyvíjame v Ruby on Rails, keďže podstatnú časť projektu tvorí webová aplikácia a zároveň niektorí členovia tímu majú s týmto jazykom skúsenosti z predošlých projektov. Čo najvyššiu spoľahlivosť nám zabezpečuje testom riadený vývoj - už teraz máme takmer 100 percentné pokrytie. Dbáme tiež na vysokú kvalitu kódu a každá nová funkcionalita je podrobená code-review. V tomto nám pomáha aj systém kontinuálnej integrácie - Travis-CI.

Tím

Náš tím je zložený zo siedmich členov pričom traja študujú odbor informačné systémy a zvyšný štyria odbor softvérové inžinierstvo. V našom tíme majú teda zastúpenie ľudia, ktorí sa v minulosti venovali najrôznejším technológiám. Ako tím máme motiváciu dotiahnuť tento projekt do čo najúspešnejšieho konca, a bez ohľadu na umiestnenie v súťaži by sme naše riešenie chceli nasadiť aj do reálnej prevádzky.

Účasť v súťaži

Súťaž TP Cup berieme najmä ako možnosť prezentovať náš projekt pred ľuďmi z praxe. Účasť v súťaži nám dáva priestor získať hodnotnejšie ocenenie ako iba známku v indexe. Uvedomujeme si, že mnoho tímových projektov dosiahlo významné úspechy aj mimo fakulty a že konkurencia bude tento rok tiež veľká. To nás však vôbec neodrádza od účasti a snahe umiestniť sa.