TÍM č. 16 IS-SI

WebX

Názov projektu:

Extraction - Extrakcia dát z webu

Členovia tímu (študenti):

Ján Brechtl, Tomáš Juhaniak, Martin Kalužník, Rastislav Krchňavý, Michal Kren, Martin Lacek, Andrej Vaculčiak

Ved. tímu (pedagóg):

Dr. Ivan Srba

Motto tímu:

„Peňazí, ani dát nie je nikdy dosť.“

O ČOM JE NÁŠ PROJEKT?

Cieľom projektu je vyriešenie problému, či už v akademickej (napr. analýza správania), ale aj komerčnej sfére (analýza aktivity konkurencie), v podobe potreby automatizovaného zberu dát, ktorý je v dnešnej dobe riešený rôznymi ad-hoc parsermi, čím zbytočne zaberá čas a stojí peniaze.

Náš systém automaticky, na základe používateľom stanovených intervalov sťahuje dáta zo stránok podľa vopred definovaného skriptu.

Skladá sa z dvoch základných častí, a to webovej aplikácie a rozšírenia do prehliadača Google Chrome.

Webová aplikácia poskytuje správu používateľov (prihlásenia a registráciu), manažment projektov používateľa (projekty predstavujú určitú doménu, do ktorej extrahované dáta zaraďujeme), zahŕňajúci definíciu a správu dátových polí.

Týmto spôsobom sa určí schéma extrakcie. Pomocou dátových polí sa vytvorí skript, ktorý sa aplikuje na príslušnú web stránku a získa požadované dáta. Vykonávanie skriptu je závislé na požiadavke používateľa, ktorý si definuje ako často sa bude skript vykonávať. Okrem toho aplikácia poskytuje prehľad vykonaných extrakcií, spolu so štatistickými informáciami a výsledným stavom extrakcie. Dáta získané počas extrakcie sú tak isto k dispozícií, pričom používateľ si ich môže aj stiahnuť pomocou API alebo vo formáte CSV.

Rozšírenie do prehliadača Chrome poskytuje rozhranie, ktorým si používateľ pre požadovanú stránku zvolí elementy, ktoré chce extrahovať. Dáta, ktoré získa je možné po extrahovaní spracovať pomocou post-procesorov (napr. vykonať vnorenú extrakciu, vyčistiť text od prebytočných znakov na začiatku a na konci a pod.)

Po "vyklikaní" elementov extrakcie sa vytvorí spomínaný skript, čo teda enormne uľahčuje jeho vytvorenie a používateľ vôbec nemusí vedieť, ako skript vyzerá alebo aká je jeho štruktúra.

ČO NÁM DÁVA PRÁCA NA TOMTO PROJEKTE?

Okrem osvojenia si metódy vývoja softvéru v podobe scrumu, sa každý člen učí nové technológie a postupy, ako riešiť problémy v danej doméne. Práca v tíme je tiež výzvou, ktorá dáva príležitosť každému z členov objaviť, aké má predpoklady na určitú pozíciu v rámci tímu. Tímová spolupráca je nevyhnutnosťou, nakoľko ani jeden z členov nie je expertom v danej oblasti, ani čo sa týka problematiky, ani technológií.

S technológiami sa spája aj zavedenie si určitej metódy práce v rôznych oblastiach (dokumentácia, vývoj, verzionovanie, komunikovanie, atď.).

Z toho vyplýva dôležitosť zavedenia metodík pre relevantné oblasti, ale aj ich dodržiavanie.

PREČO JE NÁŠ PROJEKT ZAUJÍMAVÝ?

Pracujeme s modernými technológiami, či už čo sa týka vývoja alebo aj zameraním projektu, nakoľko je extrakcia dát v dnešnej dobe pomerne žiadaným a aktuálnym problémom, a to nie len v akademickej sfére.
Systém, ktorý vyvíjame kladie dôraz na intuitívnosť a jednoduchosť používania. Ponúka oproti konkurencií viacero výhod v podobe jednoduchej možnosti výberu elementov extrakcie, cez prehľadné zobrazenie výsledkov s možnosťou stiahnutia a správu vlastných projektov. Extrahovanie je možné naplánovať, a teda nie je potrebné manuálne extrakcie spúšťať (aj keď existuje aj táto možnosť).

POUŽITÉ TECHNOLÓGIE:

Ruby on Rails, HTML, CSS, JavaScript, Angular, PostgreSQL, ElasticSearch, Redis

O ČOM TO VLASTNE JE?

Ponúknutie alternatívneho riešenia problému potreby extrakcie dát z rôznych zdrojov spôsobom, ktorý je prijateľný tak pre odborne skúsených, ako aj menej skúsených používateľov.

Vytvorenie systému, pomocou ktorého jednoducho, s použitím niekoľkých krokov, dokáže používateľ určiť, ktoré elementy majú byť extrahované, ako sa extrahované dáta spracujú a ponúknuť ich používateľovi v prehľadnom zobrazení alebo ponúknuť možnosť stiahnutia v čitateľnom formáte.