VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Skriptovacie jazyky 2014/2015

Stahování dat z Twitteru a diskusního fóra

13. května 2015 Peter Tisovčík

1 Zadanie

Projekt se skládá ze dvou samostatných částí – zpracování dat z internetového fóra a zpracování dat z Twitteru. Každá je hodnocena maximálně 20 body. Součástí odevzdaného archivu má být i alespoň půlstránková dokumentace, popisující stručně způsob řešení, použité externí knihovny a výslednou rychlost a úspěšnost systému.

Do poloviny března je potřeba si vybrat vhodné fórum či diskusní skupinu, kde za poslední týden přibyl alespoň jeden příspěvek a účet na Twitteru, na němž byl za poslední dva týdny publikován alespoň jeden příspěvek s odkazem na nějakou webovou stránku. Svoji volbu je potřeba zapsat spolu se svým loginem do tabulky na Google Docs. Nikdo však nesmí zpracovávat žádný zdroj, který už si v tabulce zarezervoval jiný student.

Prvním úkolem je napsat v Pythonu nebo Ruby skript, který stáhne všechny dosavadní příspěvky vybraného diskusního fóra a uloží je do souboru spolu s relevantními metainformacemi – označením autora, času, vlákna diskuse apod. Skript také dokáže zjistit nové příspěvky od posledního stažení a spustit jejich stahování.

Druhým úkolem je napsat v Pythonu nebo Ruby skript pro sledování vybraného účtu na Twitteru - viz *https://dev.twitter.com/streaming/overview* – lokální ukládání všech nových příspěvků, nalezení případných odkazů na odkazované webové stránky a stažení jejich obsahu.

2 Upresnenie zadania

2.1 Diskuzné fórum

Pre implementáciu som si vybral diskusné fórum *https://www.hojko.com*. Toto fórum je diskusné fórum, na ktorom sa nenachádza nelegálny obsah a je voľne dostupné. Obsahuje vyše 1.6 milióna príspevkov, v približne 108 tisíc témach, ktoré vytvorilo vyše 50 tisíc užívateľov. Čo je priemerne 15 príspevkov na jednu tému.

Výstupným formátom je *xml* súbor, v ktorom sa nachádza výpis metainformácií o jednotlivých príspevkoch – autor, dátum a čas, text príspevku a vlákno do, ktorého príspevok patrí.

Pri aktualizácií fóra kontolujem, či sa každá téma stiahla správne, či obsahuje dátum posledného príspevku a ak nie stiahnem ju. Počítam s tým, že pri tak veľkom fóre je možné, že počas sťahovania dôjde k chybe a subfórum alebo téma sa nestiahne.

2.2 Twitter

Vybral som si *MIUI* učet na Twitrri. Jedná sa o grafickú nadstavbu Androidu, ktorú vytvára firma XIAOMI. Každý deň pribudne niekoľko nových tweetov, ktoré obsahujú krátky popis a odkaz na stránku. Pri aktualizácií sťahujem len tweety od posledného príspevku a ostatné zostávajú uložené v databáze.

Výstupným formátom je *xml* súbor, v ktorom sú uložené informácie o tweete ako je jeho ID, text tweetu a odkazy, ktoré sa nachádzajú v tweete. Každý odkaz je stiahnutý do súboru s príponou **.html*. Tweety vo výstupnom formáte sú zoradené od najnovšieho tweetu.

3 Implementácia fóra

3.1 Použité knižnice

V projekte pre fórum som použil niekoľko knižníc. Na sťahovanie obsahu webovej stránky som použil knižnicu *urllib2*, ktorá stiahne požadovanú stránku a uloží ju do premennej. Je s ňou podobná práca ako pri otváraní súborov, kde sa stránka otvorí uloží a následne treba zavrieť spojenie. Knižnica *BeautifulSoup* rozkúskuje stránku a vytvorí strom, ktorý sa dá prechádzať pomocou metód. Napríklad nájde všetky odkazy v zadanej triede. Na ukladanie dát som použil databázu *sqlite3*, ktorá využíva SQL syntax. Následne uložené dáta som pomocou knižnice *lxml.etree* uložil do XML formátu. Používam ešte knižnice *sys*, ktorá mi nastaví kódovanie a knižnicu *time* pomocou, ktorej parsujem dátum.

3.2 Princíp fungovania

Na začiatku súboru si definujem kódovanie, v ktorom bude program pracovať. Je to z toho dôvodu, že vybrané fórum je v slovenčine. Ak ide o prvé spustenie scriptu vytvori sa databáza s definovanými tabuľkami. Ak ide o opakované spustenie tak sa otvori databáza, ak zadaná databáza neexistuje tak sa vytvori.

Stiahnem hlavnú stránku fóra, z ktorej si vyberiem odkazy na subfóra. Následne ich po jednom kontrolujem, či existujú v databáze, ak nie pridajú sa. Uložím si ID fóra aby som pri generovaní XML súboru vedel spojiť jednotlivé témy so subfórami. Následne prehľadávam jednotlivé stránky subfóra a v nich hľadám odkazy na témy. Pri ukladaní témy do databáze, kontrolujem, či dátum posledného príspevku je rovnaký ako v databáze ak áno túto tému preskočím. Pri príspevku ukladám príspevok, autora a čas vytvorenia príspevku. V prípade, že príspevok existuje v databáze tak ho už nepridávam ale preskočím ho. Po prejdení každého subfóra vygenerujem XML súbor pomocou funkcie `generate_xml()`, do ktorého uložím všetky témy a príspevky z daného subfóra.

3.3 Problémy pri implementácií

Pri implementá som narazil na niekoľko problémov. Dátum obsahoval české názvy mesiacov a bolo ich treba upraviť do formátu, ktorý je možné previesť na formát, s ktorým sa lepšie pracuje. Na to som vytvoril funkciu `check_date()`. Ktorá dostane na vstup reťazec obsahujúci dátum, následne nahradí české mesiace za anglické a prevedie ich na požadovaný formát. Riešenie s parsovaním času by bolo nastaviť lokalizáciu systému ale tu som narazil na problém, že musia byť dostupné lokalizačné súbory inak by script nefungoval, preto som od tohto riešenia upustil. Pri vkladaní do databáze som zistil, že texty(príspevok, autor) môžu obsahovať apostrofy a preto som nahradzoval apostrofy za obyčajné úvodzovky a až tak som ich ukladal. Asi najväčším problémom bola rýchlosť sťahovania, ktorá sa v mojom prípade nedala urýchliť vláknami z nejakého obmedzenia na strane servera. Server povoľoval len jedno aktívne spojenie na IP adresu, čo znamenalo, že rýchlosť pri meraní s a bez sťahovania s vláknami bola skoro totožná a líšila sa len veľmi málo. Preto som upustil od tohto riešenia, pretože by to nespĺňalo požadovaný výsledok.

3.4 Časová náročnosť

Z problému, že sťahovanie nemohlo byť vykonávané vo vláknach, tak sa celé fórum stiahne za 6.4 dňa.

4 Implementácia Twitteru

4.1 Použité knižnice

V projekte pre Twitter som použil knižnicu *tweepy* pre sťahovanie dát z Twitteru, kde bolo potrebné si zaregistrovať aplikáciu na Twitteri a nechať si vygenerovať prístupové kľúče. Knižnicu *sqlite3* som použil pre ukladanie tweetov a odkazov do databáze. Jednotlivé obsahy odkazov som stiahol pomocou knižnice *urllib2*. Následne som tieto tweety a ich informácie uložil do XML súboru pomocou *lxml.etree* knižnice.

4.2 Princíp fungovania

Na začiatku sa script autentifikuje pomocou prístupových kódov. Ak neexistuje databáza s tweetmi, tak sa vytvori. Počet tweetov, ktoré sa majú stiahnuť sú definované v premennej `pocet_tweetov`. Následne si jednotlivé informácie o tweete uložím do databázi. Ak existujú odkazy v danom tweete, tak sa ich obsah stiahne do súboru. Následne sa vyexportuje výsledný XML súbor s tweetmi.