Dokumentace projektu pro předmět ISJ: Stahování dat z Twitteru a diskusního fóra

Jméno a příjmení: Daniel Žůrek

Login: xzurek12

Twitter: @gamescz

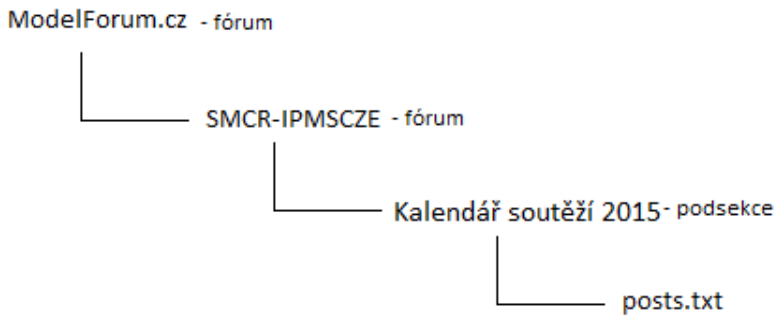
Fórum: www.modelforum.cz

Python: 2.7

Stahování příspěvků z fóra

Jelikož mnou zvolené fórum neposkytuje žádné API pro snadnější manipulaci s příspěvky, bylo nutné toto fórum zpracovávat pomocí parsování html obsahu stránky. Pro snadnější zpracování zdrojového kódu stránky byla použita knihovna *BeautifulSoup*, jedná se o Pythonovský HTML/XML parser (ke stažení zde). Pro optimalizaci rychlosti načítání jednotlivých stránek byla použita knihovna *urllib3*, která využívá stejný socket pro vícenásobné požadavky (ke stažení zde).

Skript vytváří adresářový strom, který odpovídá struktuře fóra. Obsahuje-li některá podsekce příspěvky, jsou tyto příspěvky uloženy do souboru `posts.txt` spolu s metainformacemi o příspěvku (ID, autor, čas, text). Jednotlivá fóra mohou obsahovat odkazy na další fóra nebo odkazy na podsekce.



Obrázek 1: Adresářový model dle fóra

Na stránce jsou vyhledávány určené html značky, které určují typ stránky (fórum, podsekce, příspěvky) spolu s jejich odkazy. Každý nově nalezený odkaz je předán funkci `ZpracujForum()`, přičemž první odkaz je odkaz na hlavní stránku fóra. Podle typu stránky se volají příslušné funkce:

- Fórum – `GetForumAB()` – nalezne odkazy fór, vytvoří adresáře
- Podsekce – `GetForumBG()` – nalezne odkazy podsekcí, vytvoří adresáře

Pokud podsekce již obsahuje soubor `posts.txt` je spuštěn aktualizační mód. Aktualizaci příspěvků zajišťuje funkce `CheckForNews()`. Na poslední stránce příspěvků dané podsekce je nalezeno ID posledního příspěvku. Pokud toto ID soubor s příspěvky již obsahuje, skript pokračuje ve zpracování dalších podsekcí, jinak jsou příspěvky staženy znovu.

Skript zpracovává jednotlivá fóra sekvenčně, což může zapříčinit delší dobu provádění programu. Vylepšením by bylo využití Pythonovských *threadů*.

Stahování dat z Twitteru

Twitter poskytuje REST API, pro jednoduchý programový přístup k datům jednotlivých účtů. Nutností je registrace nové aplikace pracující s daty Twitteru. Po této registraci jsou vygenerovány bezpečnostní klíče, které jsou použity při autorizaci. Pro snadnější OAuth autorizaci je využita knihovna *Tweepy* (ke stažení zde).

Objekt vytvořený pomocí této knihovny, poskytuje přístup k celému REST API a jednotlivé data jsou získávána pomocí knihovních funkcí. Například získání všech tweetů daného účtu, se provede zavoláním metody `user_timeline()`, bez nutnosti zpracování v cyklu. Jeden z parametrů této funkce je počet stahovaných tweetů na jeden request, přičemž maximum je 200 tweetů. Z účtu je možné stáhnout maximálně 3200 tweetů.

Každý tweet je reprezentován JSON strukturou, která obsahuje množství informací od ID tweetu po URL odkazy v jeho textu. Z každého tweetu jsou získány jeho metainformace (ID, text, URL adresy a časové razítko) a vloženy do slovníku. Následně je tento slovník převeden pomocí knihovny *dicttoxml* (ke stažení zde) do formátu XML a uložen do souboru `gamescz-tweets.xml`.

Z URL adres obsažených v tweetech je vytvořen seznam. Z tohoto seznamu jsou načítány jednotlivé adresy. Obsah stránek na těchto adresách je uložen do souboru v adresáři `Stazene stranky`. Jméno souboru je odvozeno od ID tweetu ve kterém se tato adresa nacházela. Pokud tweet obsahoval více odkazů, je jméno souboru obohaceno o číslovku, která určuje pozici odkazu. Např.:

593326743437606912 - 1.html

593326743437606912 - 2.html

Aktualizační mód využívá jednoho z parametrů funkce `user_timeline()`, jedná se o parametr *id_since*. Pokud je tento parametr zadán, jsou staženy jen ty tweety, jejichž ID je větší než hodnota tohoto parametru. ID posledního staženého tweetu je uloženo do souboru `gamescz-last_tweed.txt`. Z tohoto souboru je při každém dalším spuštění skriptu načteno uložené ID a následně použito jako hodnota parametru *id_since*. Jedná-li se o první spuštění programu, je staženo prvních 50 tweetů daného účtu.

Oba programy byly testovány na systému Windows 7 64bit.