ISJ – projekt

Sťahovanie dát z fóra a twittera

Fórum

$ redditscript.py

Na sťahovanie dát bol použitý modul **PRAW**(Python Reddit API Wrapper) na jednoduché sťahovanie a parsovanie dát.

PRAW Install:

$ pip install praw

$ easy_install praw

Zameral som sa na subforum(subreddit) – **IamA**. Reddit API vyžaduje autorizáciu pri manipulovaní s dátami, aby mohlo tieto činnosti regulovať. **User_agent** slúži na identifikáciu užívateľa, spravidla tu ma byt krátky popis k čomu sa API využíva.

Skript stiahne najnovšie príspevky - **subreddit.get_new()**, ich štatistické údaje a komentáre. Príspevkov je limitovaný počet **200**. Po prvom spustení vytvorí súbor **RedditPosts**, do ktorého sa uložia všetky stiahnuté príspevky. Pri ďalšom spustení ma užívateľ možnosť stiahnuť iba nove príspevky od posledného sťahovania.

Skript overuje posledný čas modifikácie súboru RedditPosts a porovnáva ho so súčasným. Čas modifikácie zisťujem pomocou **os.stat(súbor).st_mtime**

Parsovanie informácii o príspevku bolo vcelku jednoduché. Boli použité už definovane metódy ako napr. **comments**, ktorá vrátila celu sekciu komentárov v danom príspevku.

Informácie sú uložené v user-friendly formáte na zrozumiteľne čítanie, avšak na následne parsovanie dát môže byt tato štruktúra zložitejšia.

Twitter

$ twitterscript.py

Sťahovanie príspevkov prebiehalo pomocou modulu **tweepy.**

Tweepy Install:

$ pip install tweepy

Twitter API taktiež vyžaduje autorizáciu užívateľa. Vytvoril som si účet na Twitteri cez ktorý som získal 4 identifikačné údaje: **consumer_key, consumer_secret, access_token, access_token_secret.**

Následná autorizácia sa overuje cez **OAuth** pomocou identifikačných údajov.

Skript sťahuje najnovšie príspevky(tweety) vrátane re-tweetov z timelinu daného užívateľa. Príspevky sú uložené do súboru ***Tweets.***

Skript stiahne všetky príspevky(Limit je **200**) a uloží ich do zoznamu **tweetsList,** pričom každý tweet obsahuje **username, tweet.id_str**(ID tweetu)**, tweet.created_at**(dátum) a **tweet.text**

Zároveň hľadá URL odkazy v každom príspevku pomocou regulárnych výrazov **re.findall()**. Pomocou **urllib** sa uložia zdrojové kódy odkazovaných stránok do adresára ***URL/***. Každý súbor so zdrojovým kódom je uložený vo formáte {ID tweetu-addressoffset}