Meno a priezvysko: Marek Marušic
Login: xmarus05

Dokumentácia k projektu sťahovanie dat z twitteru do predmetu ISJ

1 Základné informácie

Na sťahovanie som si vybral twitter účet @ArchUpdates. Skript je určený hlavne pre python 3.2+.

2 Použité knižnice

Knižnica requests bola použitá pre získavanie html obsahu z url stránok (je obsiahnutá v archíve, nieje ju nutné inštalovať).
Knižnica TwitterSearch bola použitá pre sťahovanie príspevkov z twitteru (je obsiahnutá v archíve, nieje ju nutné inštalovať).

3 Spustenie

Spúšťanie sa vykonáva pomocou príkazum python3 twitter.py.

python3 twitter.py Stiahne 50 posledných príspevkov z @ArchUpdates do defaultného súboru ArchUpdates a html stránky do ArchUpdatesURLs.

3.1 Argumenty

[--output-dir=dir] Definuje výstupný adresár sťahovaných html stránok. Defaultný adresár je ArchUpdatesURLs.

[--output=file] Definuje výstupný súbor, ak nieje zadaný používa sa defaultný subor (ArchUpdates).

[--update] Aktualizuje stiahnutý súbor s twitter príspevkami (v prípade že nieje použitý parameter –input=file aktualizujeme defaultný súbor ArchUpdates).

[--ipnut=file] Určuje aký súbor sa má aktualizovať. Je možné ho použiť iba s parametrom –update.

4 Sťahovanie

Skript vždy zistí či daný príspevok obsahuje odkaz na url a obsah tejto stránky stiahne do zložky ArchUpdatesHtml. Názov tohto html súboru vygenerujeme pomocou regulárneho výrazu z textu tweetu a následnej substitúcie medzier za pomlčku z tejto postupnosti znakov substi.

5 Update

Pri udatovaní už stiahnutých dát som použil funkciu set_since() s parametrom id (číslo posledného stiahnutého príspevku, ktoré je uložené v hlavičke stiahnutého súboru). Výstupné dáta v súbore sú zoradené podobne ako na stránke twitter,tj. od najnovšieho tweetu po najstarší. Kvoli tomuto efektu pri updatovaní načítam celý súbor do pamäti, následne zapíšem všetky nové príspevky a po stiahnutí posledného príspevku zapíšeme na začiatku uložený súbor.