Meno a priezvysko: Marek Marušic
Login: xmarus05

Dokumentácia k projektu sťahovanie dát z fóra do predmetu ISJ

1 Základné informácie

Na sťahovanie som si vybral fórum www.zabublame.cz. Skript je určený hlavne pre python 3.2+.

2 Použité knižnice

Pre implementáciu sťahovania dát som použil knižnicu requests a knižnicu lxml.
Knižnica requests bola použitá pre získavanie dáta z url stránok (je obsiahnutá v archíve, nieje ju nutné inštalovať).

Knižnicu lxml som použil pre parsovanie, prácu s html kódom a pre formátovanie výstupu (**inštalácia je nutná**). Pre inštaláciu treba použit:

pip install lxml

3 Spustenie

Spúšťanie sa vykonáva pomocou príkazum python3 forum.py.

python3 forum.py Stiahne fórum do defaultného súboru zabublame.cz.xml.

3.1 Argumenty

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

[--update] Aktualizuje stiahnutý súbor s fórom (v prípade že nieje použitý parameter –input=file aktualizujeme defaultný súbor zabublame.cz.xml).

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

4 Sťahovanie

Skript posutpne z hlavnej stránky prehľadáva každé fórum a sťahuje príspevky v jednotlivých témach. Výstup sa ukladá do výsledného súboru s xml štruktúrou.

5 Update

Pri updatovaní som využil dátum poslednej úpravy fór. V každej úrovni fóra je možné zistit do ktorého fóra a do ktorej témy je nutné vstúpiť kvôli aktualizácii a preto nieje nuté prechádzať aj fóra, témy a príspevky, ktoré nie sú upravené.