RSS parser

Ivan Havasi
ČVUT–FIT
havasiva@fit.cvut.cz

January 1, 2022

1 Úvod

V rámci tejto semestrálnej práce som sa rozhodol vytvoriť parser, ktorý dokáže prečítať RSS feed (teda články) z webových stránok rôznych médií.

Parser som vytvoril ako webovú aplikáciu, ktorá dokáže dané články zobraziť. Tak isto dokáže pridať nové stránky na získavanie ďalších článkov a tieto čĺanky následne zobraziť. Všetky stránky a články si aplikácia ukladá do databázy.

Nakoniec som sa ešte rozhodol pridať jednoduchú stránku so štatistikami. Táto stránka zobrazuje dve štatistiky - počet článkov z každého zdroja a najpoužívanejšie slová. Cieľom týchto štatistík nebolo vytvoriť, čo najdetailnejšie štatistiky, ale jednoducho ukázať, že aj pri projekte ako je RSSParser sa dajú zobraziť zaujímavé štatitiky.

2 Front-end

Na dizajn front-endu som využil Bootstrap.

Pomocou backendového frameworku Flask sa renderujú všetky HTML stránky, ale taktiež je vytvorené API, ktoré dokáže vytvoriť, vymazať, zobraziť... stránky a články.

Pre všetky činnosti, ktoré som uznal za dôležité je vytvorená stránka, na ktorej sa dajú buď zobraziť dané dáta, alebo sa dá poslať HTTP request na back-end. Http requesty sa na back-end posielajú pomocou kniznice jQuery.

3 Back-end

Parsovanie článkov prebieha v triedach Parser a Reader, ktoré dokážu načítať url stránku s RSS obsahom, tento obsah spracovať a následne vrátiť vo forme dictionary. Takto spracované dáta potom aplikácia ukladá do databázy alebo ich vráti ako odpoveď na API request.

Dáta dokáže aplikácia poslať na front-end buď priamo počas renderovania HTML templatu alebo pomocou API.

Štatistiky sú vytvárané pomocou knižnice Pandas a následne odosielané na front-end. Pri štatistikách je zaujímavý súbor excluded-words.txt, ktorý obsahuje na každom riadku iné slovo, ktoré sa nemá brať do štatistiky najčastejšie používaných slov. Často sa totiž stávalo, že najbežnejšími slovami boli spojky, predložky...

Aplikácia ako databázu používá in-file súbor, do ktorého sa ukladajú všetky články a stránky. Na prácu s databázou používa framework Flask, konkrétne jeho súčasť flask-sqlalchemy. Boli vytvorené modely pre článok (article) a stránku (site), ktoré sú zároveň aj tabuľkami v databáze.

4 Výsledky

Aplikácia spĺňa hlavnú požiadavku - zvládnuť parsovať články z rôznych médií. Tieto čĺanky napokon zobrazí na hlavnej stránke aplikácie.

Pre pridávanie nových článkov, z už existujúcich stránok, stačí prejsť na webstránke do sekcie All sites a získať nové články.

Rovnajko jednoduché je pridanie novej stránky v sekcii Add site.

Ak človeka zaujimájú štatistiky, môže prejsť do sekcie Stats, kde sú zobrazené najpoužívanejšie slová. Ak by niekto chcel odstrániť niektoré slová, stačí do súboru excluded-words.txt v kóde aplikácie pridať nové slovo na nový riadok.

5 Budúcnosť

Pre webovú aplikáciu si viem určite predstaviť prepracovanie a zlepšenie dizajnu.

Ohľadom back-endu - určite by bolo dobré pridať podporu pre rozšírené parsovanie, ktoré by sa potom dalo využiť na filtrovanie článkov, ako napr. kategórie, autori atď.

Ďalšia vec, ktorá by bola zaujímavá je pridanie viacerých štatistík a zobrazovanie týchto štatistík pomocou grafov.

6 Závěr

Vytváranie webovej aplikácie pomocou Pythonu bola určite zaujímavá skúsenosť. Ale musím povedať, že lepšie sa mi s ním pracovalo, napr. pri získavaní štatistík alebo samotnom parsovaní článkov.