Dokumentácia k projektu z ISJ

Automatické sťahovanie a porovnávanie titulkov

Autor: Daniel Klimaj, xklima22@stud.fit.vutbr.cz
Fakulta informačních technologií
Vysoké učení technické v Brně

Obsah

1. Zadanie....3
2. Implementácia....3
2.1 Jazyk....3
2.2 Prepínače....3
2.3 Sťahovanie titulkov....3
2.4 Porovnanie titulkov....4
2.5 Výsledky....4
3 Príklady použitia....4

1. Zadanie

Úlohou projektu bolo napísať skript, ktorý v multijazyčnej databázi titulkov vyhľadá titulky k zadanému filmu, stiahne českú a anglickú verziu titulkov a tieto titulky porovná. Ako databázu titulkov som si zvolil www.subscene.com

2. Implementácia

2.1 Jazyk

Ako implementačný jazyk som si zvolil jazyk Ruby (skript bol testovaný na verziách 1.9.3 a 2.1.0) a využil som gemy:

- open-uri
- mechanize
- nokogiri
- rubyzip
- fileutils

Ku skriptu som pridal Gemfile, a tak všetky gemy možno jednoducho nainštalovať pomocou bundleru.

2.2 Prepínače

Spôsob získavania titulkov závisí od použitého prepínača. Skrip možno spustiť pomocou 3 prepínačov:

- ./app.rb -t title
- ./app.rb -u cs_subtitle_url
- ./app.rb -i imdb_url

2.3 Sťahovanie titulkov

2.3.1 Prepínač -t

Pri použití prepínača -t sa s pomocou gemu Mechanize vyplní vyhľadávacie pole na www.subscene.com a potom, ak existuje presná zhoda názvov, t.j. film sa musí objaviť v poli Exact results sa pomocou open-uri a nokogiri získa URL na stránku s titulkami.

Na stránke s titulkami sa vyhľadajú titulky v angličtine a získajú sa URL na ich stiahnutie. Titulky sa sťahujú v .zip archívoch a ukladám ich do priečinkov ./subtitles/[cs/en]. Po dokončení sťahovania sa .zip archívy rozbalia alebo ak sú poškodené alebo iného formátu (titulky môžu byť aj v .rar archívoch) sú vymazané. Po rozbalení sú všetky archívy vymazané.

2.3.2 Prepínač -u

Zo stránky, na ktorú odkazuje URL k titulkom sa vyparsuje názov filmu a stiahnú titulky, ktoré ukladám do ./subtitles/cs/original.zip. Potom pomocou Mechanize sa vyplní vyhľadávacie pola a práca pokračuje ako v predoľlom prípade.

2.3.3 Prepínač -i

Z IMDB stránky filmu sa získa originálny názov a zvyšné spracovanie prebieha ďalej ako u prepínača -t.

2.4 Porovnanie titulkov

Na porovnávanie som implementoval 2 triedy:

- Subtitles (./lib/subtitles.rb)
- SubtitleItem (./lib/subtitle_item.rb)

Prvá trieda predstavuje súbor s titulkami a definuje niekoľko operácií nad ním, predovšetkým metódu pre porovnanie 2 súborov s titulkami, počet riadkov, atď.. Druhá trieda predstavuje jeden „blok“ titulkov a jeho vlastnosti (začiatok, koniec, počet slov, …).

Pri porovnávaní berim v úvahu 3 faktory:

1. Počet blokov s rovnakým začiatočným časom (+1s)
2. Počet blokov s rovnakým počtom slov
3. Počet blokov s rovnakou dĺžkou trvania (čas od zobrazenia po skrytie)

Celkovú zhodu potom určujem ako súčet vyššie uvedených faktorov delených celkovým počtom blokov a tento súčt delený 3.

2.5 Výsledky

Titulky s najvyššou mierou zhody ukladám do priečinka results, ktorý ďalej obsahuje 3 podpriečinky:

- ./cs – obsahuje české titulky
- ./en – obsahuje anglické titulky, ktoré sa najviac zhodovali s cs verziou
- ./log – obsahuje súbor, v ktorom sú zobrazené zhody/nezhody medzi súbormi v ./cs a ./en

3 Príklady použitia

./app -t 'The Mist'

./app -u 'http://subscene.com/subtitles/the-mist/czech/174834'

./app -i 'tt0884328'