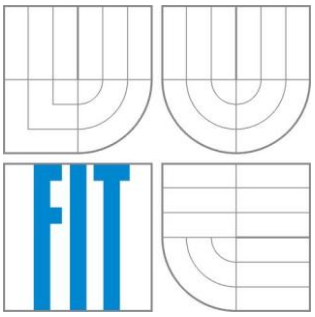

Dokumentace k projektu pro předmět ISJ

Varianta - Automatické stahování a srovnávání titulků k filmům

Martin Kačmarčík

Fakulta informačních technologií

Vysoké učení technické v Brně

Obsah

1. Použité knihovny a verze Pythonu

Můj program používá Python verze:

Python 2.7.4.

Při svém projektu jsem použil některé knihovny. Zde je jejich seznam:

"import sys, getopt, urllib2, re, os, zipfile, xmlrpclib, os.path, collections, shutil"

2. Režim spouštění

Můj skript (psaný v Pythonu) se spouští následujícím příkazem:

python xkacma03.py -u "url", kdy url obsahuje odkaz na stranku s titulky z opensubtitles (pozor zadava se stranka s titulky, ne primo download).

Omlouvám se, že nepoužívám (pochopil jsem, že to je doporučené, nikoliv nutné)

české promluvě neodpovídá anglická, bude na výstupu:
Česky \t
Pokud je to naopak, pak:
\t Anglicky,

moje řešení se slovníkem se mi zdálo přehlednější. Hned na první pohled je zřejmé, zda-li to je čeština nebo angličtina a slovník promluvy hezky ohraničí závorkami. Doufám tedy, že tato změna nepovede k horšímu hodnocení.

3. Algoritmy

3.1 Získání ID a download titulků

Pro získání ID titulků používá můj skript jednoduché regulární výrazy, pomocí kterých získá ze zadané stránky ID anglických titulků, které následně s pomocí urllib2 knihovny stáhne, pokud jsou to ZIP soubory, tak je také rozbalí. Dále stáhne české titulky zadané v argumentu. Pokud jsou ZIP rozbalí je, pokud ne udělá z nich .srt soubor. Vse se deje ve slozce ENG a CZE dle příslušného jazyka.

3.2 Porovnání

Nyní mám všechny titulky na disku a můžu s nimi pracovat. Nejprve si uložím názvy všech titulků do listu a dle těchto názvů ukládám do nového listu velikosti souborů. Následně provádím porovnání českých titulků s každýma titulkama z anglických zástupců a hledám nejbližší shodu velikosti. Ve chvíli jak najdu, vypíšu na stdout jaká je shoda a které titulky shodu obsahují.

Dalším krokem je zapsání pravděpodobných dvojic slov do souboru. Proto projdu svým algoritmem oba soubory s titulky a do slovníku si uložím vždy dvojice čas:promluva, kdy čas je číslo, které vzniklo pro převodu formátu času v titulkách na sekundy. Promluva je řetězec odpovídající dannému času. Pokud už danný čas ve slovníku je, přidá se pouze promluva k dannému času (řetězec s promluvou je ve skutečnosti list).

Nyní už jen stačí dalším algoritmem projít celý slovník a vypisovat vždy promluvy popř. dvojice promluv s nejnižším do souboru, který je nazván dle názvu filmu a příponou _compare.out.

4. Novinky 19.5.2014

Pred deadlinem jsem ještě dodělal nějaké kontroly (spravne URL z hlediska jazyka, jestli nahodou nestahuji uz rovnou .srt a nemam teda rozbalovat etc.).

Dále jsem dodělal nástroj pro úklid. Na konci se program zeptá, jestli chcete po sobě uklidit, což znamená smazat složky ENG a CZE.

Také jsem dodělal nástroj pro obrácené porovnávání (prakticky kosmetické změny v URL souborů), bohužel jsem neměl šanci ho vyzkoušet kvůli překročení limitu stažených souborů z opensubtitles. Ovšem zjistil jsem, že nástroj si správně nalezne ID českých titulků a download prakticky funguje na principu, že jen do URL dává tyto ID a stahuje soubory. Neměl by být tudíž problém.

Bohužel se bez testu neodvažuju se na tuto verzi programu spoléhat (i když jsem vše testoval offline s původními soubory, na 99% bude fungovat, ovšem pro jistotu...), proto přikládám původní soubor (který neobsahuje výše uvedené změny, ale 100% funguje). Pokud tedy download extrémě (což by snad neměl, maximálně v případě obráceného downloadu), prosím o testování původního souboru. Děkuji.