Dokumentácia k projektu do predmetu Skriptovacie jazyky

Autor: Filip Gulán (xgulan00)

1 Úvod

Úlohou tohoto projektu bolo vytvoriť dva skripty. Prvý skript má za úlohu stiahnuť všetky príspevky daného vybraného diskusného fóra a uložiť ich o súboru spolu s príslušnými metainformáciami. Skript má byť tiež schopný zistiť nové príspevky od posledného stiahnutia a uložiť ich. Druhý skript má za úlohu stiahnuť všetky nové príspevky daného Twitter účtu a tiež stiahnuť obsah webovej stránky, ktorej link sa v tweete nachádza. Obidva skripty boli tvorené v programovaciom jazyku Python3.4.0.

2 Fórum skript

2.1 Spôsob riešenia

Skript najprv otvorí stránku fóra ceske-hry.cz/forum/index.php pomocou knižnice *urrlib3* a jej metódy *connection_from_url()*. Spojenie sa uloží pre prípadné neskoršie použitie, takže oproti knižnici *urllib2*, ktorú skript používal najskôr, sa zrýchlilo sťahovanie približne 5x. Po otvorení získa obsah stránky a zistí aktuálny dátum fóra, ktorý sa uloží do logu, pre aktualizačný mód. Potom sa postupne vytvorí zoznam mien subfór a zoznam odkazov na tieto subfóra. V cykle sa začne postupne prechádzať každý odkaz na subfóra a skript vola funkciu *ziskajInformacie()*, v ktorej nastáva spracovanie daného subfóra. Najprv sa stiahne obsah prvej stránky subfóra, kde sa nájdu odkazy a mená tém. Potom skript začne postupne spracovať prve stránky danej témy a následne sa prejde na ďalšie stránky danej témy. Po spracovaní všetkých tém na prvej stránke subfóra, skript začne spracovávať ďalšie stránky subfóra. Všetky informácie zo stiahnutej stránky (linky, mená tém, mená používateľov, príspevky…) sa získavajú pomocou regulárnych výrazov. Skript dané informácie po získaní ihneď zapisuje do súboru. Každému subfóru prislúcha jeden súbor uložený v priečinku forum/ceske-hry. V prípade spustenia aktualizačného módu a zistenia nových príspevkov, sa tieto príspevky uložia na koniec príslušného súboru.

2.2 Dostupné prepínače skriptu

Skript je možné spúšťať s niekoľkými argumentami.

- -firstpage: prepínač spôsobí, že sa spracujú iba prvé stránky všetkých subfór.
- -metaonly: prepínač, ktorý spôsobí, že sa nebudú ukladať texty príspevkov, ukladať sa budú iba metainformácie.
- -repair: prepínač, ktorý sa používa po havárií skriptu (napríklad pri nedostupnosti serveru...) a ktorý spôsobí to, že vymaže adresárovú štruktúru spolu so všetkými stiahnutými informáciami a log súborom.

2.3 Požadované knižnice

Skript požaduje na to, aby fungoval nasledujúce knižnice:

- urllib3 - sťahovanie stránok
- re - spracovanie stránok
- os.path - vytvorenie adresárovej štruktúry na uloženie príspevkov a logu.
- sys - na zistovanie existencie súborov
- shutil - na odstránenie celej adresárovej štruktúry

3 Twitter skript

3.1 Spôsob riešenia

Skript na získanie tweetov daného užívateľa používa knižnicu TwitterSearch, pomocou ktorej je získanie tweetov triviálna záležitosť. Pri spustení sa vždy uloží prvý nájdený tweet spolu s dátumom do log súboru a pri aktualizácií sa už iba kontroluje získaný tweet s posledným tweetom v logu. Ak sú tweety totožné, tak sa už nepokračuje čo znamená, že sme našli všetky nové tweety. Tweety sa ukladajú do súboru tweety v zložke twitter. V prípade aktualizácie sa nové tweety uložia na koniec súboru. Prípadné odkazy na webové stránky sa z tweetov získavajú pomocou regulárnych výrazov a obsah stránok sa ukladá do zložky twitter/stranky. Každá uložená stránka je pomenovaná podľa nadpisu stránky získaného z tagu *<title></title>*.

3.2 Dostupné prepínače skriptu

Skript je možné spúšťať s niekoľkými argumentami.

- -nolink: prepínač spôsobí, že sa nebudú sťahovať obsahy stránok, ktorých link sa v tweete nachádza.
- -repair: prepínač, ktorý sa používa po havárií skriptu (napr pri nedostupnosti serveru...) a ktorý spôsobí to, že vymaže adresárovú štruktúru spolu so všetkými stiahnutými informáciami a log súborom.

3.3 Požadované knižnice

Skript požaduje na to, aby fungoval nasledujúce knižnice:

- TwitterSearch (na svoje správne fungovanie potrebuje knižnice: requests a requests_oauthlib) - sťahovanie tweetov
- urllib3 - sťahovanie obsahu stránok
- re - spracovanie tweetov a obsahu stránok
- os.path - vytvorenie adresárovej štruktúry na uloženie tweetov, stránok a logu.
- sys - na zisťovanie existencie súborov
- date time - zistenie aktuálneho dátumu a času
- shutil - na odstránenie celej adresárovej štruktúry

4 Záver

Skripty boli riadne otestované na systéme Linux Deepin 2014.2 a na školskom systéme FreeBSD (eva). Fórum skript dosahoval na systéme Linux Deepin pri prvom spustený, kedy sa sťahuje celé fórum, cca 45 000 príspevkov, časy okolo 17 minút. Na systéme eva zase okolo 5 minút. Aj napriek tomu, že na fóre je napísané, že obsahuje ku dnešnému dňu okolo 48 000 príspevkov, som názoru, že úspešnosť skriptu je 100%, keďže som zistil, že v roku 2007 nastala havária serveru, ktorá mala za následok stratu dát (http://www.ceske-hry.cz/forum/viewtopic.php?t=3). Skript Twitter dosahoval na Deepine časy pri prvom spustený kedy sa sťahuje 50 príspevkov + obsahy stránok cca 1 minútu, 30 sekúnd a totožný čas na eve. V archíve sú k obidvom skriptom pribalené tiež všetky potrebné knižnice, ktoré sú nad rámec základných. Skripty boli tvorené vo vývojovom prostredí PyCharm.