Analyzátor klíčových slov

Bc. Martin Ševčík, Masarykova univerzita 26/02/2019

Popis skriptu

Skript načte obsah vstupního textového souboru. Poté na základě algoritmů vyfiltruje z textu tzv. stop slova a provede jeho analýzu. Na základě analýzy definuje pět klíčových slov, která seřadí sestupně dle frekvence výskytu v textu. Tato slova vypíše v terminálu a uloží je do nového textového souboru. Součástí skriptu je i provizorní instalátor, který nainstaluje potřebné CPAN moduly.

Vstup a výstup

Vstupním souborem je **"text.txt"**, který je potřeba naplnit libovolným textem určeným k analýze. Výstupem skriptu je **pět klíčových slov**, které skript zobrazí v terminálu, a poté uloží do samostatného textového souboru s názvem **"klicovaslova.txt"**.

Prerekvizity

Podmínkou pro spuštění skriptu je nainstalavaný Perl verze 5. Skript dále využívá **CPAN** modul **Lingua::EN::Keywords**, jehož funkce je propojena s modulem **Lingua::EN::Tagger** a následně i s **Lingua::Stem::En.** Tyto moduly je nutné mít nainstalované pro správné fungování skriptu.

Instalace

Skript obsahuje provizorní automatický instalátor, který ale dle prozatímního testování nemusí být spolehlivý na všech systémech. Hlavním kamenem úrazu instalátoru je snaha o obejití nutnosti instalovat moduly jako *root* uživatel skrze **local::lib**. Pokud instalátor selže, nainstalujte moduly manuálně dle instrukcí na webových stránkách uvedených níže.

- https://metacpan.org/source/Lingua::Stem::En
- https://metacpan.org/source/Lingua::EN::Tagger
- https://metacpan.org/pod/Lingua::EN::Keywords

Spuštění

Skript se spustí po otevření složky (pomocí terminálu), ve které se nachází a zadání následujícího příkazu:

Listing 1: Příkaz pro spuštění skriptu – Analyzátor klíčových slov.

```
1 perl klicova_slova.pl
```

Skript je rozdělen do několika dílčích kroků a vyžaduje vstupy od uživatele. Prvním důležitým vstupem před spuštěním skriptu je soubor **"text.txt"**, který se musí nacházet ve stejné složce jako skript a musí být naplněný textem určeným k analýze. Další vstupy, které skript po spuštění od uživatele vyžaduje jsou potvrzení průběhu skriptu v jednotlivých popsaných částech pomocí klávesy *enter*.

Jak skript probíhá

1. Skript ověří přítomnost potřebných CPAN modulů, případně se je pokusí nainstalovat
2. Načte se modul Lingua::EN::Keywords a obsah textového souboru "text.txt"
3. Pomocí regulárních výrazů se z textu odfiltrují stop slova
4. Skript zobrazí filtrovaný dokument
5. Text zobrazí klíčová slova a zapíše je do souboru "klicovaslova.txt"
6. Skript se ukončí

Dodatek

Skript pro filtrování stop slov využívá slova ze zdroje:https://countwordsfree.com/stopwords/czech