

Popis řešení

Shrnutí

Během řešení jsem používal Perl poprvé, takže jsem musel mnoho vyhledávat. A i proto jsem se rozhodl problém hledání klíčových slov řešit jednoduše možná by se dalo říci až primitivně.

Vycházím z toho, že klíčová slova by pravděpodobně měla patřit mezi ta nejvíce používaná. Daný text tedy postupně procházím po slovech a pomocí asociativního seznamu počítám výskyty jednotlivých slov.

Nakonec iteruji seznamem klíčů (tedy slov) seřazených právě podle jejich četnosti (tedy hodnoty pod klíčem v asoc. seznamu) a vypíšu pouze několik nejfrekventovanějších.

V průběhu implementace jsem se rozhodl umožnit některé parametry nastavit pomocí přepínačů, i když to nebylo v zadání. Skript je tak mnohem užitečnější.

Detailnější popis

Především na začátku skriptu zparsuji argumenty a nastavím si podle toho proměnné - subprocedura **parse_args**. Určitě existují funkce, které zpracují argumenty podle standardu a které by mi dovolily zpracovat sofistikovanější syntax přepínačů, ale já jsem pro takovýto jednoduchý skript zůstal u "ručního" regex porovnávání a klasické rovnosti.

Poté si načtu slovník slov, které chci ignorovat v **load_ignores**. To dělám tak, že se pokusím otevřít určený soubor a pokud neexistuje tak se vrátím a nechám **%ignore** hash prázdný. Pokud se mi podaří soubor otevřít, tak z každého řádku odstraním newline znak a použiji ho jako klíč v **%ignore**.

Následně přichází o něco méně elegantní větvení. Nepodařilo se mi totiž zjistit jak file handle uložit do proměnné, a tak dochází k nepěkné duplikaci kódu. Pokud totiž byla předána cesta k souboru, nejdříve soubor otevřu (v režimu s UTF8 kódováním), pokud ne tak pouze nastavím standardnímu vstupu UTF8 kódování, a pak v obou větvích iteruji po řádcích, které rozdělují pomocí **split** na slova. Pokud je slovo kratší než požadovaná délka (daná buď přepínačem nebo výchozí hodnotou), tak jej přeskočím. Jinak slovo v malých písmenech použiji jako klíč a zvýším hodnotu.

Dále pokud byl použit přepínač **-pN** nebo nebyl použit přepínač **-cN** najdu **N**. respektive výchozí 95. percentil výskytů a tuto hodnotu použiji jako hranici výskytu.

Pak už jenom vypíši na standardní výstup všechna slova - tedy klíče z **%key_words** - v pořadí od nejčastějších po méně častá, přičemž přeskakují slova, která jsou klíčem v **%ignore**, a skončím vypisovat, pokud narazím na slovo s menší hodnotou než je hranice.

Reflexe a komentář