Skript pro extrakci klíčových slov

Artem Gorodilov

1. Abstrakt

Skript přijímá text v kódování UTF-8 jako standardní vstup a jako standardní výstup poskytuje textový soubor s klíčovými slovy seřazenými podle důležitosti (četnost výskytu v textu a celkový počet výskytů).

2. Popis skriptu

Skript je napsán v jazyce Perl v5.34.0 [1] (funguje i v novějších verzích).
Skript je rozdělen do několika částí:

1. Import požadovaných modulů, nastavení kódování UTF-8 a definování oddělovače writerů.
2. Načtení souboru ze standardního vstupu, převod textu na seznam slov a změna velikosti všech slov na malá písmena.
3. Import datového souboru se stop slovy, filtrování seznamu slov od stop slov a čísel.
4. Výpočet četnosti výskytu jednotlivých filtrovaných klíčových slov v textu, seřazení klíčových slov podle četnosti výskytu a seřazení podle celkového počtu výskytů v textu.
5. Výstup seznamu klíčových slov.

2.1. Nastavení skriptů

Nastavení skriptu je v kódu definováno následujícími řádky:

```
use strict;
use warnings;

use open qw(:std :utf8);
```

Skript používá moduly strict a warnings k zajištění správné činnosti skriptu a k zobrazení varování v případě chyb. Modul open slouží k nastavení kódování vstupních a výstupních souborů na UTF-8.

2.2. Čtení souboru

Skript načte text ze standardního vstupu a převede jej na seznam slov [2]. Vstupní text je umístěn ve složce *input/*.
Text je rozdělen na slova pomocí mezer, interpunkčních znamének a zalomení řádků. Velká písmena všech slov jsou změněna na malá.

```
my $text = <STDIN>;

my @words = $text =~ /(\w+)/g;

foreach my $word (@words) {
    $word = lc $word;
}
```

2.3. Filtrování stop slov a čísel

Skript načte stop slova z datového souboru [3] a odfiltruje seznam slov od stop slov a čísel.

```
my @stopwords;
open my $stopwords_file, '<', 'stopwords-cs.txt' or die "Cannot open stopwords-cs.txt: $!";

@stopwords = split(/\s+/, <$stopwords_file>);
my %stopwords_hash = map { $_ => 1 } @stopwords;

my @filt_words = grep { !$stopwords_hash{$_} && !/\d/ } @words;
```

2.4. Výpočet četnosti výskytu a třídění

Skript vypočítá četnost výskytu každého filtrovaného klíčového slova v textu a seřadí klíčová slova podle četnosti výskytu [4] a podle celkového počtu výskytů v textu.

Prahová hodnota pro počet výskytů klíčového slova v defind podle argumentu uvedeného v inicializačním řádku kódu [5] (vysvětleno v kapitole "3. Dokumentace").

```
my %word_counter;
foreach my $word (
    @filt_words) {
    $word_counter{$word}++;
}

my @sorted_words = sort {
    $word_counter{$b} <=>
    $word_counter{$a} } keys
    %word_counter;

my $frequency_threshold =
    $ARGV[0];

my @keywords = grep {
    $word_counter{$_} >
    $frequency_threshold }
    @sorted_words;
```

2.5. Výstup seznamu klíčových slov

Skript vypíše seznam klíčových slov na standardní výstup. Klíčová slova jsou uvedena ve formě seznamu, přičemž každé klíčové slovo je na samostatném řádku.

```
foreach my $keyword (
    @keywords) {
    print "$keyword\n";
}
```

3. Dokumentace

Skript se spouští z příkazového řádku následujícím příkazem:

```
perl klicova_slova.pl 10 <
input/text.txt > output/keywords.txt
```

Skript přijímá jeden argument - prahovou hodnotu pro počet výskytů klíčového slova v textu. Tento počet závisí na velikosti textu. Pokud je text velký, měla by být prahová hodnota nastavena na vyšší hodnotu. Pokud je text malý, měla by být prahová hodnota nastavena na nižší hodnotu.
Skript načte text ze složky $input/$ a zapíše seznam klíčových slov do složky $output/$.

4. Diskuse

Skript funguje správně a poskytuje očekávaný výsledek. Nicméně jasnost klíčových slov závisí především na údajích o stopwords. Uživatel může do datového souboru napsat vhodnější stopslova pro konkrétní text nebo najít jiný zdroj stopslov.

Odkazy

[1] Perl website. www.perl.com/

[2] PerlScript to take certain words from text file. https://stackoverflow.com/questions/66495616/perlscript-to-take-certain-words-from-text-file

[3] Stopwords ISO CZ. https://github.com/stopwords-iso/stopwords-cs/blob/master/stopwords-cs.txt

[4] How do I sort by frequency of a value? https://stackoverflow.com/questions/4519979/how-do-i-sort-by-frequency-of-a-value

[5] How can I pass command-line arguments to a Perl program? https://stackoverflow.com/questions/361752/how-can-i-pass-command-line-arguments-to-a-perl-program