# Keywords extraction script

Artem Gorodilov

## 1. Abstract

The script takes a UTF-8 encoded text as standard input and gives a text file with keywords sorted by importance (frequency of occurrence in the text and total number of occurrences) as standard output.

## 2. Description of the script

The script is written in Perl v5.34.0 [1] (works with newer versions also).
The script is divided into several parts:

1. Importing the required modules, setting the UTF-8 encoding and defining the writer delimiter.
2. Reading the file from standard input, converting the text to a word list and changing the case of all words to lowercase.
3. Importing a data file with stop words, filtering the word list from stop words and numbers.
4. Calculating the frequency of occurrence of each filtered keyword in the text, sorting keywords by frequency of occurrence and sorting by total number of occurrences in the text.
5. Output of the list of keywords.

### 2.1. Script settings

The script settings defined in the code by following lines:

```perl
use strict;
use warnings;

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

The script uses the strict and warnings modules to ensure the correct operation of the script and to display warnings in case of errors. The open module is used to set the encoding of the input and output files to UTF-8.

### 2.2. Reading the file

The script reads the text from the standard input and converts it to a word list [2]. The input text is located inside *input/* folder.
The text is divided into words by spaces, punctuation marks, and line breaks. The case of all words is changed to lowercase.

```perl
my $text = <STDIN>;

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

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

### 2.3. Filtering stop words and numbers

The script reads the stop words from the data file [3] and filters the word list from stop words and numbers.

```perl
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. Calculating the frequency of occurrence and sorting

The script calculates the frequency of occurrence of each filtered keyword in the text and sorts the keywords by frequency of occurrence [4] and by total number of occurrences in the text.
Threshold for the number of occurrences of a keyword in defind by the argument given to code initialization line [5] (explained in the chapter "3. Documentation").

```
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. Output of the list of keywords

The script outputs the list of keywords to the standard output. The keywords are presented in the form of a list, each keyword on a separate line.

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

## 3. Documentation

The script is run from the command line with the following command:

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

The script takes one argument - the threshold for the number of occurrences of a keyword in the text. This number depends on the size of the text. If the text is large, the threshold should be set to a higher value. If the text is small, the threshold should be set to a lower value.

The script reads the text from the *input/* folder and writes the list of keywords to the *output/* folder.

## 4. Discussion

The script works correctly and gives the expected result. Nevertheless, the clearence of the keywords are mainly dependent on the stopwords data. The user can write more suitable stopwords for the specific text in the data file or find another source of stopwords.

## References

[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