

# Dokumentace

Jan Rychlý

25. 11. 2020

## úvod

Klicova\_slova je skript na extrahování klíčových slov z textových souborů. To dělá podle četnosti výskytu slov. Spočítá výskyt všech slov, některé ignoruje a pak vypíše pouze ty nejfrekventovanější. V základním běhu bez přepínačů zpracuje text na standardním vstupu a pak vypíše slova s četností 95. percentilu a vyšší v pořadí od nejfrekventovanějších po ty méně frekventovaná na standardní výstup. Zároveň při tom ignoruje slova o délce menší než 3 a jakákoliv slova ze souboru ./ignore\_list.txt pokud soubor existuje.

Pomocí přepínačů je možné určit minimální délku slov, jejich minimální četnost, konkrétní percentil nebo danou četnost u každého slova zobrazit. Lze tedy i např. získat četnost každého slova v textu.

Je vhodné zadat konkrétní percentil podle rozsahu souboru.

## spuštění

synopse:

```
./klicova_slova.pl [-h] [-lN] [-cN] [-pN] [-v] [FILEPATH]
```

V základu skript čte ze standardního vstupu. Pokud zadáme FILEPATH skript se pokusí otevřít a číst soubor na této adrese.

## přepínače

- lN**        Zobraz pouze slova o délce **N** nebo více
- cN**        Zobraz pouze slova s **N** nebo více výskyty
- pN**        Zobraz pouze slova s výskytem **N** percentilu nebo více (přebíjí možnost **-cN**)
- v**        Zobraz u každého slova počet výskytů
- h**        Vypiš zprávu o použití.

## ignorovací slovník

Skript se pokaždé pokouší načíst slovník slov, které by měl ignorovat, ze souboru ./ignore\_list.txt. Přesné shody s těmito slovy se poté budou při analýze přeskakovat.

Každé slovo musí být v souboru samo na vlastním řádku bez bílých znaků.