Temat: Analiza narzędzi wyszukujących zawartość tekstową w systemie Linux

Cel: Znalezienie programu odpowiedniego do wyszukiwania zawartości tekstowej w terminalu, porównanie wydajności pomiędzy językami i implementacjami

Zadanie: Porównanie wydajności i implementacji kilku najpopularniejszych programów wyszukujących pliki i treść w plikach. Sprawdzenie wykorzystanych algorytmów, dostępnych flag (złożoności programu), prędkości wyszukiwania w zależności od podanych flag.

Zakres pracy: Wyszukanie brakujących informacji tabeli. Stworzenie programu mającego możliwość automatycznego sprawdzenia performance każdego z narzędzi, feacherów dostępnych, wielkości programu.

Programy do szukania plików

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Fzf | Fzy | locate | pick | ack | find |
| Popularność |  |  |  |  |  |  |
| Język | Go, | C, |  | C, | rust |  |
| Prędkość |  |  |  |  |  |  |
| Ilość zastosowań |  |  |  |  |  |  |
| Wsparcie |  |  |  |  |  |  |
| Ilość parametrów |  |  |  |  |  |  |
| Złożoność programu (algorytmy) |  |  |  |  |  |  |
| Wielkość programu | <https://codetabs.com/count-loc/count-loc-online.html> |  |  |  |  |  |
| Łatwość użycia |  |  |  |  |  |  |
| Kolejne wystąpienia wyrazu/pliku |  |  |  |  |  |  |
| link | https://github.com/junegunn/fzf | https://github.com/jhawthorn/fzy |  |  |  |  |

Programy do szukania zawartości w plikach

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | grep | ripgrep | ack | baloo | rare | Ag – the silver search |
| Popularność |  |  |  |  |  |  |
| Język |  |  |  |  |  |  |
| Prędkość |  |  |  |  |  |  |
| Ilość zastosowań |  |  |  |  |  |  |
| Wsparcie |  |  |  |  |  |  |
| Ilość parametrów |  |  |  |  |  |  |
| Złożoność programu (algorytmy) |  |  |  |  |  |  |
| Wielkość programu |  |  |  |  |  |  |
| Łatwość użycia |  |  |  |  |  |  |
| Kolejne wystąpienia wyrazu/pliku |  |  |  |  |  |  |
| link |  |  |  | https://github.com/KDE/baloo/blob/master/docs/user/searching.md | https://github.com/zix99/rare | https://github.com/ggreer/the\_silver\_searcher |

Porównanie z własną implementacją?

Wszystkie exeki w tym samym miejscu?

Pliki szyfrowane?

Biblioteka książek do wyszukiwania (plain text)?

Zagrzanie dysku może wpłynąć na perf

Sprawdzenie które narzędzia sa zainstalowane, puszczanie jednego skryptu który będzie wywoływał jeden z programów który jest najszybszy dla danego case’a

Linux reader distInternals

Czy mogą zobaczyć do zipa, rara

Czy uruchamiać na dysku sieciowym czy lokalnym lub do ramu

# Narzędzie do ekstrakcji tekstu z danych:

Mp3 – Project DeepSpeech / <https://www.notta.ai/en/blog/speech-to-text-open-source>

<https://github.com/openai/whisper>

<https://openai.com/research/whisper>

Maybe to hard to do

Archive:

* 7zz,
* <https://operavps.com/docs/install-7zip-in-linux/>
* tar.gz
* zip- unzip
* rar – unrar

Pdf - <https://github.com/search?q=pdf2txt&type=repositories>, pdftotext (poppler-util)

Htm, html – save as is

Ps – don’t support

Cpp

H

Cxx

(empty)

Js

C

Jpg, gif removed

Additional – tr (remove and replace)

Jeden program będzie miał 5 parametrów w szukaniu po nazwach plików (nie trzeba konwertować)

Drugi program będzie miał 5 parametrów w szukaniu po zawartości (trzeba przekonwertować zipy i pdf, djvu itd.)

# Narzędzia użyte w programie

Man [program]

Info [program]

# Pytania

Output zawsze do terminala, może dodać output do pliku?

Jak działa regex w go gdy używamy exec.command?

Jak działa fuzzing?

# Źródła

Context -https://www.sohamkamani.com/golang/exec-shell-command/