# ISI Hakaton – Wyszukiwarka kodów źródłowych

# Raport nr 1

Lider: Anna Hojan

Pozostali członkowie: Emilia Cieszewska

Magdalena Karyś

Maciej Brusiło

## Ogólne założenia projektu:

Planujemy:

1. Napisać crawlera oceniającego wartość repozytoriów (na podstawie liczby watchersów, followersów i gwiazdek) i clonującego te wysoko rankowane.
2. Użyć programu Pygments do rozpoznania języka, jeżeli ta informacja nie będzie zawarta w metadanych projektu.
3. Stworzyć wyszukiwarkę przy użyciu silnika Solr.
4. Stworzyć interfejs użytkownika umożlwiający przeszukiwanie repozytoriów.

Udostępnimy następujące opcje przeszukiwania repozytoriów, po:

* nazwie projektu
* treści commitów
* treści komentarzy
* języku programowania
* loginie autora
* fragmencie kodu.

Umożliwimy przeszukiwanie kilkunastu lub kilkudziesięciu tysięcy repozytoriów (nie wiemy jeszcze jak rozwiązać kwestię zapotrzebowania na dużą przestrzeć dyskową).

Wykorzystywane języki programowania:Python, Perl

Źródła repozytoriów:publiczne repozytoria na Githubie

Wykorzystywane narzędzia: Solr

## Co zrobiliśmy w tym tygodniu:

* utworzyliśmy repozytorium
* zapoznaliśmy się z api githuba (<https://api.github.com>), które będzie naszym punktem wyjścia przy crawlowaniu w poszukiwaniu repozytoriów, poznaliśmy jego ograniczenia (maksymalnie 5000 requestów na godzinę dla zalogowanego użytkownika, 60 dla niezalogowanego)
* poprosiliśmy administratora o udostępnienie dużej przestrzeni dyskowej (niestety bez pozytywnego rezultatu)
* wstępnie przydzieliliśmy sobie zadania
* stworzyliśmy skrypt pobierający metadane repozytoriów i uruchomiliśmy go na pierwszych 10 000 repozytoriów

## Co planujemy zrobić w następnym:

* stworzyć crawlera (obmyślić logikę crawlowania, ustalić dokładne zasady na podstawie których repozytoria będą klasyfikowane jako takie, które warto ściągnąć)
* sklonować dużą porcji repozytoriów
* postawić Solra

Repozytorium:https://github.com/AnnaHojan/Wyszukiwarka-kodow