ISI Hakaton – Wyszukiwarka kodów źródłowych

Raport nr 2

Lider:	Anna Hojan
Pozostali członkowie:	Emilia Cieszewska
	Magdalena Karyś
	Maciej Brusiło

Co zrobiliśmy w tym tygodniu:

1. Opracowanie algorytmu crawlowania: crawlujemy repozytoria kolejno z listy: https://api.github.com/repositories odsiewając po drodze te, które nie będą spełniały określonych kryteriów.
2. Określenie kryteriów zakwalifikowania repozytoriów do ściągnięcia lub nie. Parametry brane pod uwagę to:
   - bycie forkiem (warunek odrzucający repozytorium),
   - rozmiar repozytorium: nie będziemy pobierać gigantycznych repozytoriów,
   - liczba gwiazdek, „watchersów" i „followersów" (z większą wagą gwiazdek względem pozostałych).
3. Rozpoczęcie prac nad stworzeniem skryptu clonującego repozytoria.
4. Lokalne postawienie instancji Solra: stworzenie nowego core'a wraz ze zdefiniowaną strukturą indeksu w pliku schema.xml, odpowiadającą indeksowi z danymi jakich będziemy potrzebować, tzn:
   - nazwa repozytorium,
   - autor,
   - komentarz,
   - commit,
   - treść kodu,
   - język,
5. Zaindeksowanie kilku testowych dokumentów i sprawdzenie jak można po nich wyszukiwać.

Co planujemy zrobić w następnym tygodniu:

1. Dokończenie niedziałających jeszcze funkcjonalności skryptów i ich poprawki.
2. Integracja stworzonych elementów (crawlera, clonera oraz Solra).
3. Rozpoczęcie crawlowania dużych danych za pomocą ostatecznych wersji skryptów crawlujących i clonujących.
4. Połączenie z otrzymają maszyną wirtualną i próba pracy na niej.

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