Twitter TF-IDF

Dokument zawierający informacje wymagane podczas drugiego spotkania z klientem

Sformalizowanie ról w grupie projektowej - przyporządkowanie zadań (z uwzględnieniem punktów z kolejnych deadline'ów)

Paweł Chorążyk

- zarządzanie serwerem CI, deployment aplikacji
- organizacja i struktura projektu
- stworzenie mechanizmów współpracy z Twitter Streaming API
- stworzenie części topologii obliczającej ostateczną wartość tf-idf

Wojciech Grajewski

- stworzenie topologii pobierającej dokumenty podlinkowane na Twitterze
- precomputing dokumentów, np. przypisanie wartości TF do dokumentu i persystencja
- zarządzanie testami jednostkowymi i integracyjnymi, code coverage
- utrzymywanie dokumentacji

Określenie tematu projektu wraz z opisem proponowanego problemu

Przetwarzanie strumienia danych z serwisu Twitter w celu wyliczenia współczynników tf-idf dla wybranych słów. Analiza częstości występowania słów jest przeprowadzana na dokumentach, do których linki zostają umieszczone w "tweetach".

Organizacja sprzętu

Continuous integration
Jenkins, http://bachor.us.to:8383/
serwer: VPS, 3.2GHz, 3GB RAM, Ubuntu 12.04

W początkowej fazie projektu symulacja topologii lokalnie, później być może deployment do rozproszonych węzłów.

Wykorzystanie środowiska Maven/Ant jako ramy do prezentacji projektu (opisu, kodu, dokumentacji, testów itd.) - całość dostępna przez stronę www

Do budowania projektu, deploymentu i uruchamiania testu będzie wykorzystywany Maven 3 we współpracy z Jenkinsem.

Wykorzystanie Subversion/GIT jako repozytorium tworzonego kodu (projekt, dokumentacja, testy)

Projekt będzie umieszczony w prywatnym repozytorium git umieszczonym na serwerze github. Dokumentacja i raport z testów będzie generowana regularnie przez jenkinsa (uruchamiając odpowiednie “maven goals”) i umieszczana na serwerze VPS pod adresem http://bachor.us.to/twitter-tf-idf/

Przemyślenie projektu, analiza systemu: diagramy przypadków użycia

- Czy analizujemy słowa bezpośrednio w tweetach, czy w podlinkowanych dokumentach? (w książce jest ta druga opcja)
- Jaką użytkownik może mieć funkcjonalność? Podaje słowo i dostaje tf-idf, czy jakaś bardziej zaawansowana funkcjonalność jeszcze?