Wprowadzenie do eksploracji danych tekstowych w sieci WWW
Sprawozdanie końcowe
Ocena narzędzi bioinformatycznych

Piotr Jastrzębski
Piotr Król
Rafał Karolewski

1 Szczegółowy opis zadania

Należy opracować program, który dla zadanej listy linków do narzędzi bioinformatycznych oceni ich jakość np. poprzez analizę cytowań danego narzędzia (najpierw należy znaleźć artykuły, które opisują dane narzędzie).

2 Załozenia projektu

Wejście - lista narzędzi bioinformatycznych w formie tekstowej
Wyjście - lista danych narzędzi posortowana według jakości wraz z wynikiem funkcji oceny

Wyniki funkcji oceny będące miarą jakości zadanych narzędzi wyliczany będzie według następujących kryteriów:

- **Liczba artykułów, w których wystąpiło odniesienie do danego narzędzia**
- **Liczba cytowań**
- ***h-indeks* autora źródła, w którym wystąpiło odniesienie do narzędzia**
- ***h-indeks* autorów cytowań**
- **liczba wyników zwracanych dla zapytania Google „[narzędzie] bioinformatics"**
- **liczba artykułów na Wikipedii odnoszących się do narzędzia**

Ogólny współczynnik jakości wyliczony zostanie jako średnia ważona wszystkich kryteriów z odpowiednio dobranymi wagami. Wektor wag zostanie wyznaczony na podstawie ważności poszczególnych kryteriów.

3 Narzędzia do wydobywania wiedzy

- **pubMed** – internetowa baza danych obejmująca artykuły z dziedziny medycyny i nauk biologicznych [2]
- **Google Scholar** [1]
- **Google API** [4]
- **MediaWiki API** [5]
- **web2py** [3]

4 Implementacja

Ze względu na uniwersalność planujemy użycie języka Python. Doskonale sprawdzi się on w przetwarzaniu tekstów i pozyskiwaniu danych ze stron. Interfejs użytkownika zrealizowany zostanie w formie webowej przy użyciu HTML albo w formie aplikacji okienkowej Qt.

4.1 Wczytanie listy narzędzi

Lista narzędzi zostanie wczytana z odpowiednich pól tekstowych poprzez przeglądarkowy interfejs aplikacji albo z pliku zapisanego w formacie CSV. Interfejs webowy będzie zawierał pola do wpisania listy narzędzi oraz przycisk rozpoczynający przetwarzanie informacji.

4.2 Zwrócenie wyników

Wyniki rankingowania narzędzi zostaną zwrócone w formie tabeli wyświetlonej na stronie wygenerowanej przez aplikację. Tabela ta będzie zawierała dwie kolumny: nazwę narzędzia oraz liczbę uzyskanych punktów rankingowych. Rekordy będą posortowane w kolejności malejącej względem liczby punktów.

5 Do wykorzystania!

Wzór oceny jakości narzędzia wyrażony został wzorem: $G = T + \sum_{i=0}^{T-1} \frac{T-i}{T}(\overline{h_{index}} + C)$, gdzie T - to liczba wszystkich znalezionych artykułów, a C to liczba cytowań tego artykułu. Liczba porządkowa i rosnąc określa spadek jakości znalezionego artykułu wg. wyszukiwarki strony PubMed.

Indeks H[1] - dodać wyjaśnienie.

Napisać, że zmieniliśmy z web2py na pylons i dlaczego.

Wrzucić screen z interfejsu (koniecznie 2).

[1] `http://pl.wikipedia.org/wiki/Indeks_H`

Opisać, że Google Scholar nas odcięło.
Że trudno było dobrać sensowną funkcję.
Że były problemy z bombardowaniem pubmedu wieloma wątkami.
Że ciężko ocenić jakość.

Literatura

[1] *Google Scholar* `http://scholar.google.pl/`

[2] *PubMed* `http://www.ncbi.nlm.nih.gov/pubmed`

[3] *web2py* `http://www.web2py.com/`

[4] *googleApi* `https://code.google.com/apis/`

[5] *MediaWiki API* `http://www.mediawiki.org/wiki/API:Main_page`

[6] *Pylons* `http://www.pylonsproject.org/`