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 nalezy znaleźć artykuły, które opisują dane narzędzie).

2. Założenia projektu

- Wejście - lista narzędzi bioinformatycznych dodanych w polu tekstowym interfejsu webowego;
- Wyjście - lista danych narzedzi posortowana według jakości wraz z wynikiem funkcji oceny;

Wyniki funkcji oceny bedące miarą jakości zadanych narzędzi wyliczany będzie według nastepujących kryteriów:

- **Liczba artykułów, w których wystapiło odniesienie do danego narzędzia;**
- **Liczba cytowań;**
- ***h-indeks* autora zródła, w którym wystapiło odniesienie do narzędzia;**
- ***h-indeks* autorów cytowań;**
- **liczba wyników zwracanych dla zapytania Google „[narzedzie] bioinformatics";**
- **liczba artykułów na Wikipedii odnoszacych sie do narzędzia.**

Ogólny współczynnik jakosci 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

W projekcie wykorzystane zostały następujące źródła wiedzy:

- **pubMed** – internetowa baza danych obejmująca artykuły z dziedziny medycyny i nauk biologicznych;
- **Google Scholar** - usługa firmy Google upraszczająca wyszukiwanie tekstów naukowych. Umożliwia ona wyszukiwanie materiałów z wielu dziedzin i źródeł np.: artykuły recenzowane, prace naukowe, książki, streszczenia i artykuły pochodzące z wydawnictw naukowych, towarzystw naukowych, repozytoriów materiałów zgłoszonych do publikacji, uniwersytetów i innych organizacji akademickich.
- **Wikipedia** - wielojęzyczny projekt internetowej encyklopedii działającej w oparciu o zasadę otwartej treści.

4. Implementacja

Ze względu na uniwersalność użyty został język Python. Doskonale sprawdza się on w przetwarzaniu tekstów i pozyskiwaniu danych ze stron. Interfejs użytkownika zrealizowany został w formie webowej przy uzyciu frameworku pylons.

Działanie programu polega na ustaleniu punktacji rankingowej dla danej listy narzędzi bioinformatycznych poprzez realizację szeregu zapytań do wybranych narzędzi zdobywania wiedzy. W wyniku zapytań uzyskiwane są wartości odnoszące się do jakości każdego z tych narzędzi. Poniżej opisane są procedury zapytań.

4.1. Realizacja zapytań

- **Wikipedia**

  Do serwera wikipedii wysyłane jest zapytanie o treści „<nazwa narzędzia> bioinformatics”. Strona z wynikami zapytania zostaje sparsowana pod kątem liczby wyszukanych pozycji odpowiadających danemu narzędziu.

- **Pubmed**

  Do serwera pubmed wysyłane jest zapytanie z nazwą badanego narzędzia bioinformatycznego w celu uzyskania listy artykułów z nim powiązanych. Pod uwagę branych jest maksymalnie 40 artykułów biorąc pod uwagę numer pozycji w liście wynikowej.

  Dla każdego artykułu z listy wynikowej określana jest liczba cytowań. Wartość ta uzyskiwana jest poprzez zapytanie serwera pubmed o każdy artykuł z listy wynikowej oraz sczytanie liczby pozycji, w których dany artykuł był cytowany.

- **Google_Schoolar**

  Serwer google schoolar używany jest do uzyskania h-indeksów autorów artykułów zwróconych przez serwer pubmed. Alipkacja realizuje zapytania o każdego autora danego artykułu, następnie oblicza h-index każdego z nich i zwraca średnią arytmetyczną tych wartości.

4.2. Model współbieżności

Ze względu na dużą liczbę realizowanych zapytań, w celu polepszenia wydajności czasowej uzyskiwania oceny narzędzi bioinformatycznych zastosowano przetwarzanie wielowątkowe. Podział zadań na wątki realizowany jest w następujący sposób:

- Ocena każdego narzędzia z listy wejściowej realizowana jest w osobnym wątku.
- Osobne wątki przetwarzają zapytania każdego z używanych serwerów wiedzy (google schoolar, wikipedia, pubmed).

- Dla każdego artykułu listy wynikowej zapytania pubmedu uruchamiany jest osobny wątek mający na celu uzyskanie h-indeksu autorów oraz liczbę cytowań.
- Każdy wątek według powyższego podziału wywołuje osobne wątki odpytujące serwer google schoolar o pojedyńczego autora artykułu.

Dodatkowym zabiegiem optymalizacyjnym jest utrzymywanie swego rodzaju cache'a autorów, w którym zapisywane są wartości h-indeksów autorów uprzednio odpytanych. Znacznie redukuje to czas całościowej oceny nardzędzia bioinformatycznego w przypadku gdy pare artykułów wynikowych ma tego samego autora.

Założone wartości maksymalne:

- 40 artykułów przy zapytaniach do pubmedu o powiązane artykuły
- 100 artykułów przy ustalaniu h-indeksu autorów

Powyższe założenia wynikają z ograniczeń serwerów na liczbę zapytań w pewnym okresie czasowym. Wartości te dobrane zostały tak aby nie przekraczały tych limitów oraz aby czasy oceny narzędzi bioinformatycznych były rozsądne, tj. rzędu parudziesięciu sekund.

4.3. Wyznaczenie wartości skumulowanej metryk oceny

Dobranie odpowiedniego wzoru jest bardzo trudnym zadaniem, gdyż ciężko zdefiniować pojęcie jakości narzędzia biorąc pod uwagę jedynie wartości ilościowe napisanych na ten temat artykułów. Zastosowane przez nas podejście ustalone zostało metodą empiryczną a wzór oceny jakosci narzedzia 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 okresla spadek jakości znalezionego artykułu wg. wyszukiwarki strony PubMed.

5. Obsługa

Uruchomienie aplikacji polega na wczytaniu w przeglądarce internetowej adresu localhost:5000.

5.1. Dodawanie listy narzędzi

Dodawanie listy narzędzi polega na wpisaniu w pole tekstowe nazwy bio-narzędzia a następnie naciśnięcie przycisku „add to" powodującego dodanie do listy.

5.2. Zwrócenie wyników

Wywołanie procedury oceny listy narzędzi bioinformatycznych odbywa się przez naciśnięcie przycisku „rank tools”. Wyniki rankingowania narzędzi zwracane są w formie tabeli wyświetlonej na stronie wygenerowanej przez aplikację. Tabela ta zawiera kolumny: nazwę narzędzia oraz liczbę uzyskanych punktów rankingowych, a także poszczególne składowe stanowiące część wyniku. Rekordy są posortowane w kolejności malejącej względem liczby punktów.

Rank place	Name	Google factor	Wikipedia factor	PubMed factor	Total points
1	C-GATE	18900	81	754	19735
2	Info CEnter	4390	148	2288	6826
3	java	3600	152	2131	5883
4	c++	1750	1699	2166	5615
5	RaAS	257	6	4263	4526
6	HMMER	522	19	3467	4008
7	xxx	1850	6	2136	3992
8	ProTIS	2640	0	1264	3904
9	INTEGRALL	3470	1	117	3588
10	sun-engine	768	1	802	1571
11	biopython	62	10	1246	1318
12	Ab Classifier	617	2	0	619
13	Mikropoint	1	1	268	270
14	TAPDANCE	128	0	109	237
15	AAARF	2	0	116	118
16	TESeeker	1	0	106	107
17	WikiPoint	2	0	101	103
18	pirogi	92	1	0	93

6. Lista użytych bibliotek

- pylons - framework do stworzenia interfejsu webowego
- BeautifulSoup - parsowanie stron
- urllib, httplib - zapytanie do Google/Wiki/GoogleScholar
- biopython - API do między innymi PubMedu
- threading – wielowątkowość

7. Literatura

- [1] *Google Scholar* http://scholar.google.pl/
- [2] *PubMed* http://www.ncbi.nlm.nih.gov/pubmed
- [4] *googleApi* https://code.google.com/apis/
- [5] *MediaWiki API* http://www.mediawiki.org/wiki/API:Main_page
- [6] *Pylons* http://www.pylonsproject.org/