Zadanie 1

Opis

Należy napisać aplikacje umożliwiające parsowanie i indeksowanie dokumentów zgromadzonych lokalnie oraz realizację zapytań z tworzeniem rankingu. Dokładniej, realizowane powinny być następujące funkcje:

1. Parsowanie kolekcji dokumentów (umieszczonej we wskazanym folderze)
    a. Wydzielenie termów
    b. W przypadku plików html: wydzielenie linków i tekstów na nie wskazujących
2. Tworzenie słownika termów (opcjonalnie – proste formy stemmingu, usuwanie stopwords).
3. Tworzenie indeksu odwróconego termów w różnych wariantach:
    a. uporządkowane listy dokumentów;
    b. uporządkowane listy dokumentów w postaci skompresowanej;
    c. indeks pozycyjny (w wersjach a. i b.)
4. Realizacja zapytań typu:
    a. boolowskiego dla termów (bez negacji),
    b. "free text query" (ciąg słów bez spójników logicznych),
    c. frazy,
    d. interfejs użytkownika dla zapytań i wyników.

   Ranking wyników (tfidf oraz inna funkcja, wybrana spośród proponowanych w literaturze)
5. Opcjonalnie: słownik linków i struktura danych umożliwiająca uzyskanie linków wskazujących na dany dokument/linków wychodzących z danego dokumentu.
6. Porównywanie zgodności wyników (funkcja kappa) wg różnych strategii (na przykład różne funkcje oceniające).
7. Dokumentacja
    a. Wstępny projekt: opis używanych narzędzi, struktur danych, kluczowych rozwiązań, itp.
    b. Dokumentacja kodu, instrukcja użytkownika.
    c. Końcowy raport opisujący:
        i. algorytmy, struktury danych, sposoby ich implementacji;
        ii. informacje o wykorzystanych gotowych pakietach/funkcjach itp., własnym wkładzie, problemach implementacyjnych i ich rozwiązaniach, ...
        iii. dobór danych testowych, uzasadnienie tego doboru, źródła tych danych; ograniczenia rozmiaru danych;
        iv. opis przeprowadzonych eksperymentów, wyniki i wnioski z nich (w szczególności wpływ kompresji na pamięć i czas, porównanie zgodności dwóch różnych funkcji oceniających).

Zasady

1. Projekty należy wykonywać w 2-3 osobowych grupach.
2. Dane
    a. kolekcje dokumentów z polskich akademickich stron internetowych (np. z www.uni.wroc.pl)
    b. Rozmiar całej kolekcji dokumentów indeksowanych rzędu <= 5GB [indeksujemy

tylko dokumenty o wybranych formatach, z zawartością tekstową]
    c. Dane w formacie uzyskanym przy użyciu wget (indeksujemy dane umieszczone w katalogu będącym mirrorem domeny o nazwie odpowiadającej nazwie katalogu)
3. Narzędzia
    a. Dowolny język programowania.
    b. Dowolne biblioteki (z dokładnym wskazaniem wykorzystania gotowych bibliotek służących implementacji indeksu odwróconego i innych zagadnień projektu!).
4. Wyniki:
    a. Wyniki wyszukiwania w postaci pliku tekstowego, w którym kolejne wiersze to pełne ścieżki dostępu do plików.
    b. Wyniki testów: opisane w raporcie końcowym.
5. Kryteria oceny
    a. Poprawność wyników (o ile jednoznaczne), jakość i efektywność (czas/pamięć/niezawodność)
    b. Dokumentacja (dokumentacja kodu, instrukcja użytkownika),
    c. Raport końcowy (bez niego projekt w ogóle nie będzie oceniany).

Terminy

- Wstępny projekt i podział na grupy: 13 kwiecień 2007
- Końcowe rozwiązanie: 8 maj 2007 (programy wraz z dokumentacją i raportem w postaci elektronicznej, przesłane na moje konto email).