Eksploracja danych w Internecie

(Kod przedmiotu: 02 64 5054 00)

Laboratorium 3

Celem laboratorium jest zapoznanie się i nabranie umiejętności posługiwania się algorytmami do oceny podobieństwa.

Zadanie: Napisać program, który posiada dwie funkcje:

1) Budowanie bazy danych.
    - W oparciu o adres URL, powinien rekursywnie (implementacja bez rekurencji) "odwiedzić" podstrony i zapisać je na dysku.
    - Strony powinny być "oczyszczone" a tekst przetworzony w podobny sposób jak w zadaniu do laboratorium 1.
    - Do każdego dokumentu w bazie danych powinien być zbudowany zbiór występujących w nim n-gramów.
2) Wyszukiwanie podobnych stron.
    - Wejście: adres URL strony internetowej, której zawartość powinna być przetworzona w podobny sposób jak w punkcie 1).
    - Wyjście: 3 najbardziej podobne (patrz dodatkowe informacje poniżej) strony (np. adresy URL) z bazy danych (utworzonej w punkcie 1) do strony podanej na wejściu.

Dodatkowe informacje:

- Zapoznaj się z terminem N-gram.
- Sprawdź, jak algorytm zachowa się dla różnych wartości N (a jak dla całych słów).
- Jako miarę podobieństwa należy zastosować Indeks Jaccarda.
- Warto sprawdzić także czy zastosowanie technik:
    + bag of words
    + dystans koosinusowy (*consine distance*)

Uwagi:

- Każda oddana praca powinna budować bazę danych dla innego adresu internetowego.

W ocenie zadania, powyższych punktów brane pod uwagę są:

- Sprawdzanie poprawności i weryfikacja wejścia (czy podano link).
- Obsługa błędów i wyjątków (np. Jeżeli adres nie istnieje, albo przekroczono czas odpowiedzi na żądanie).

- Jakość napisanego kodu (formatowanie, nazwy zmiennych, …).