Eksploracja tekstów
ćwiczenia 2
Zajęcia 8
(pierwsze zajęcia o Wielkanocy)

Zadanie 1. W wyniku awarii dysku zniszczeniu uległa ważna kolekcja dokumentów. Na szczęście ocalał indeks pozycyjny (jako baza danych na innym dysku). Opisz możliwie najbardziej efektywny sposób odzyskania kolekcji dokumentów (jest ona niestety zbyt duża, by zmieścić się w pamięci operacyjnej).

Zadanie 2. Wyjaśnij czy (i dlaczego) poniższe zdania są prawdziwe.

a) Stemming nigdy nie powoduje zmniejszenia precyzji.

b) Stemming nigdy nie powoduje zmniejszenia kompletności.

c) Stemming zwiększa rozmiar słownika.

d) Stemming należy wykonywać w procesie indeksowania dokumentów, ale nie dla treści za- pytania.

Zadanie 3. Rozważamy nieskompresowane listy postingowe. Wyjaśnij, jaką korzyść można uzyskać, umieszczając w pewnych miejscach takiej listy informacje o tym, że za K pozycji znajdzie się dokument o numerze N.

Zadanie 4. Jak zaimplementować efektywnie (pod kontem zużycia pamięci) indeks pozycyjny z dwupoziomowymi listami postingowymi.

Zadanie 5. Zaimplementuj w wybranym języku odległość edycyjną dwóch napisów. Twoja implementacja powinna umożliwiać różnicowanie wag różnych operacji, jak również obsługiwać wybrane operacje zmiany kolejności znaków.

Zadanie 6. Zaprojektuj postać normalną dla słowa (dla języka polskiego), która utożsami dwa słowa, jeżeli da się je otrzymać wykonując dowolną liczbę operacji:

- zamiany polskiego znaku diakrytycznego na jego łaciński odpowiednik (lub odwrotnie, czyli pomyłkowe nienaciśnięcie lub naciśnięcia klawisza ALT)
- błąd ortograficzny (ch-h, rz-ż, ó-u)
- zamiana słowa na brzmiące tak samo (być może nie wszystkie zasady fonetyki da się tu uwzględnić)

Jak taka postać może być użyteczna przy wykonywaniu korekty słowa?

Zadanie 7. SounEx nie jest najlepszym algorytmem. Znajdź w Internecie jakąś alternatywę dla SounEx-u (dla języka angielskiego) i opisz jej działanie.

Zadanie 8. Piszesz wyszukiwarkę do Wikipedii i dostałeś zadanie napisania mechanizmu korygującego zapytania użytkowników. Jak wykorzystać spostrzeżenie, że użytkownicy Waszego serwisu najczęściej pytają o tytuły i fragmenty tytułów z Wikipedii.

Zadanie 9. Jak zmodyfikować algorytm liczenia odległości edycyjnej, żeby zwracał nie liczbę operacji (czy też ich koszt), lecz zbiór operacji o minimalnym koszcie przekształcających jedno słowo w drugie. Jaki problem jest z ww. sformułowaniem zadania? Do czego taki algorytm mógłby być użyteczny?

Zadanie 10. Dysponujesz dużą liczbą par (słowo-wpisane, słowo-poprawne). Jak wykorzystać ten zbiór do ustalenia kosztów następujących operacji:

- Wpisania wariantu polskawego literki
- Omyłkowego podwojenia literki
- Wpisania *dbca* zamiast *abcd* (dla dowolnych a, b, c oraz d)

Zadanie 11. Wyszukiwarka Google prezentuje tzw. snippety, czyli krótkie fragmenty znalezionych dokumentów, które mają uzasadnić istotność tych dokumentów w stosunku do zapytania. Zadaj kilka pytań do Google'a i przeanalizuj dokładnie wygląd snippetów. Opisz, jak wyobrażasz sobie algorytm ich wyznaczania.