Raport

17 maja 2009

Spis treści

1 Opis formatu słownika morfologicznego

Słownik morfologiczny ma fromat zmieniony w stosunku do wersji oryginalnej.

1.1 Format pliku

Plik binarny słownika morfologicznego ma następujący format: Długość listy form bazowych ze słownika jako uint.
Lista form bazowych jako ciąg stringów o dlugości zapisanej wcześniej. Ilość par (słowo, lista numerów form bazowych) ze słownika. Dla każdej pary takiej jak powyżej słowo jako string a potem długość listy numerów form bazowych a potem lista tych numerów.

1.2 Format w programie

W programie słownik morfologiczny po pwczytaniu z pliku jest przechowywany jako seria tablic.

- Tablica słów które można sprawdzić przechowywane.
- Tablica tablic numerów form bazowych. Te numery odpowiadają indeksom w tablicy form bazowych. Natomiast indeksy są takie jak dla tablicy słów powyżej.
- Tablica form bazowych.

Wyszukujemy binarnie.

2 Indeks odwrócony

2.1 Format w programie

W indeksie przechowujemy:

- Tablicę tokenów do wyszukiwania posortowaną.
- Tablicę list postingowych o indeksach odpowiadających indeksom słów.
- Tablicę pozycji początków kolejnych artykułów w pliku.

Wyszukujemy binarnie.

2.2 Format listy postingowej

W lisie postingowej przechowujemy:

- Tablicę pozycji dokumentów (pozycja identyfikuje dokument)
- Tablicę tablic pozycji w dokumencie.

Listy postingowe mogą być zkompresowane przy pomocy kompresji gamma.

2.3 Tworzenie indeksu

Indeks odwrócony tworzymy w dwóch przebiegach.
W pierwszym przebiegu tworzymy posortowany słownik mapujący słowa na listy postingowe. Gdy skończy się miejsce zapisujemy go do pliku tymczasowego i tworzymy nowy słownik. Kiedy przerobimy cały plik łączymy powstałem pliki tymczasowe do jednego.
W drugim przebiegu wyliczamy pozycje dokumentów w pliku z Wikipedią i zapisujemy do pliku.

3 Stemming

Nasz program wykorzystuje stemming przy tworzeniu indeksu oraz wyszukiwaniu.
Najpierw z końca słowa są usuwane samogłoski. Potem rozpoznawane są konkretne końcówki które można usuwać. W przypadku gdy nie ma więcej końcówek do usunięcia lub słowo robi się zbyt krótkie stemming jest zakończony.

4 Wyniki

Przeprowadziliśmy testy naszego programu na wybranych opcjach tworzenia indeksu oraz wybranych zapytaniach.