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.

4.1 Wielkość indeksu

- **Wielkość słownika** 2073396
- **Zindeksowanych dokumentów** 811205

4.2 Czasy wyszukiwania

Poniżej czasy wyszukiwania dla zbiorów zapytań.

- **Zapytania AND:** całkowity czas przetwarzania: 00:00:47.8452000
- **Zapytania OR:** całkowity czas przetwarzania: 00:02:14.0508000
- **Zapytania frazowe:** całkowity czas przetwarzania: 00:00:55.5048000

4.3 Pomiar czasu i pamięci dla indeksu skompresowanego ze stemmingiem

- Czas wczytywania: 00:00:16.2864000
- Zajmowana pamięć: 437 MB
- Skompresowane listy postingowe: tablice bajtow: 298 MB (68,24%) reszta: 47 MB (10,85%)
- Słownik: 69 MB (15,76%)
- Tablice String[] 7,9 MB(1,81%)
- Tablice PositionalPostingList[] 7,9 MB(1,81%)
- Tablice Int64[] 6,2 MB (1,42%)

4.4 Pomiar czasu i pamięci dla indeksu nieskompresowanego ze stemmingiem

- Czas wczytywania: 00:01:08.5152000
- Zajmowana pamięć: 1.5 GB
- Listy postingowe: 1.4 GB (94.02%) w tym:
  - indeksy dokumentow 253 MB
  - listy pozycji 886 + 261 = 1147 MB

4.5 Pomiar czasu i pamięci dla indeksu skompresowanego bez stemmingu

- Czas wczytywania: 0:00:18.8136000
- Zajmowana pamięć: 490 MB
- Skompresowane listy postingowe: tablice bajtow: 323 MB (65,91%) reszta: 57 MB (11,57%)
- Słownik: 84 MB (17,21%)
- Tablice String[] 9,4 MB(1,93%)
- Tablice PositionalPostingList[] 9,4 MB(1,93%)
- Tablice Int64[] 6,2 MB (1,26%)

Słownik: 69 MB Tablice PositionalPostingList[] 32 MB (2,08%) Tablice Int64[] 6,2 MB