Raport

19 czerwca 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 długości zapisanej wcześniej. Ilość par (słowo, lista numerów form bazowych) ze słownika. Dla każdej pary takiej jak powyżej słowo jako string, następnie długość listy numerów form bazowych, zaś dalej lista tych numerów.

1.2 Format w programie

Po wczytaniu z pliku (w trakcie działania programu) słownik morfologiczny jest przechowywany jako seria tablic.

- Posortowana tablica słów, dla których można odnaleźć formy bazowe (wyszukiwanie binarne).
- 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.

Wczytywanie słownika morfologicznego trwa: 00:00:04.3992.
Użyta pamięć: 229119604 bajtów.

2 Indeks odwrócony

2.1 Format w programie

W indeksie przechowujemy:

- Posortowaną tablicę zindeksowanych tokenów (słownik).
- Tablicę list postingowych o indeksach odpowiadających indeksom słów.
- Tablicę pozycji początków kolejnych artykułów w pliku.

Słowa wyszukujemy binarnie w tablicy tokenów.

2.2 Format listy postingowej

W liście postingowej przechowujemy:

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

W wersji skompresowanej lista postingowa składa się pojedynczej wratości int - oznaczającej liczbę dokumentów, w których występuje dany term oraz z tablicy bajtów. W tablicy zapisujemy przy pomocy kompresji gamma kolejno: identyfikator dokumentu *id*, liczbę wystąpień danego termu w dokumencie *id*, kolejne pozycje termu w tym dokumencie.

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. Po przetworzeniu całego pliku źródłowego łączymy powstałe pliki tymczasowe do jednego wynikowego.
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 obcinane są z końca słowa rozpoznane końcówki które można usuwać. Następnie z końca słowa usuwamy samogłoski. W przypadku gdy nie ma więcej końcówek do usunięcia lub słowo robi się zbyt krótkie stemming jest zakończony.

Stemming pozwala na zmniejszenie słownika o ok. 16%.

3.1 Wielkość indeksu

- **Wielkość słownika** ze stemmingiem 1995890.
- **Zindeksowanych dokumentów** 811205

4 Realizacja zapytań

Program pozwala na realizację zapytań typu "free text query". Zapytanie takie jest dowolnym ciągiem termów. Wyszukiwarka stara się odnaleźć dokumenty najbardziej pasujące na podstawie funkcji rangujących, opartych na porównywaniu odległości Euklidesowej wektoru zapytania i dokumentów.

Przy wykorzystaniu indeksu pozycyjnego, obliczana jest punktacja każdego dokumentu. Dla kolejnych termów z zapytania przeglądane są ich listy postingowe i dla każdego dokumentu zawierającego term zwiększana jest jego punktacja o wartość $wf \cdot idf$, przy czym:

$$wf = 1 + \log tf_{t,d}$$

$$idf = \log \frac{N}{df_t}$$

gdzie:

- $tf_{t,d}$ - liczba wystąpień termu t w dokumencie d
- N - rozmiar słownika
- df_t - liczba dokumentów zawierających term t

Następnie wartości dokumentów są normalizowane poprzez podzielenie zyskanych punktów przez długość dokumentu. Ostatecznie dokumenty są sortowane według ich rang i zwracane jako odpowiedź w kolejności malejących rang.
Ponadto punkty dokumentu są odpowiednio zwiększane (przydzielany jest bonus) w pewnych określonych, uznanych za pożądane z punktu widzenia istotności dokumentu, sytuacjach:

- Jeśli term z zapytania występuje jako jedno z 5 pierwszych słów, jego punkty są mnożone przez pewien określony stały czynnik.
- Dla każdych dwóch kolejnych termów z zapytania, jeśli występują one jako fraza w pewnym dokumencie, jego punkty są mnożone przez ustaloną wartość bonusową.