Wyszukiwanie informacji w sieci, Lista 4 **12.05.2007**

Uwaga. W nawiasach okrągłych przy zadaniach podany jest numer danego zadania (lub do niego podobnego) w podręczniku IIR lub numer rozdziału, którego dotyczy zadanie.

1. [1] (6.2, 6.3)

   (a) Jaka jest wartość idf_t dla termu występującego w każdym dokumencie? Porównaj efekt skalowania uzyskany przez zastosowanie idf z tworzeniem list stop words.

   (b) Jaki efekt dla wartości wag tf.idf ma zmiana podstawy logarytmu w wyrażeniu definiującym idf? Czy może ona zmienić uporządkowanie dokumentów w rankingu?

2. [1] (7.1)

   (a) Pokaż, że dla znormalizowanych wektorów w przestrzeni R^n, odległości kątowe i odległości Euklidesowe dają takie same uporządkowanie.

   (b) Podaj krótki dowód faktu, że iloczyn skalarny znormalizowanych wektorów jest równy kosinusowi kąta między nimi.

3. [1] (7.4)

   (a) Załóżmy, że jeden z termów występujących w zapytaniu nie występuje w słowniku naszej kolekcji danych. Zaproponuj sposób reprezentacji takich sytuacji w modelu przestrzeni wektorowej, efektywny z punktu widzenia jakości otrzymywanego rankingu dokumentów i czasu obliczeń.

   (b) Uzasadnij dlaczego w modelu przestrzeni wektorowej celowe jest wpisywanie w zapytaniu niektórych termów więcej niż jeden raz. Podaj przekonujące przykłady.

4. [0.5] (7) Uporządkuj pary dokumentów o poniższych charakterystykach w kolejności malejącego podobieństwa w modelu przestrzeni wektorowej z wagami tf.idf.

   (a) dwa dokumenty, dla których wszystkie wspólne termy są słowami bardzo często występującymi w danych;

   (b) dwa dokumenty, które nie mają żadnych wspólnych słów;

   (c) dwa dokumenty, których wspólne słowa występują bardzo rzadko w kolekcji danych (i tych wspólnych słów jest dużo).

5. [1] (7) Dla przyspieszenia rangowania dokumentów w modelu wektorowym, można stosować technikę "cluster pruning". W procesie preprocessingu wybieramy $\sqrt{n}$ dokumentów z kolekcji (gdzie n to liczba wszystkich dokumentów z kolekcji) nazywanych "liderami". Dla każdego nielidera wyznaczamy najbliższego mu lidera. Obozem lidera d nazywać będziemy dokumenty, dla których d jest najbliższym liderem

   Realizacja zapytania q wygląda następująco

   - znajdź najbliższego lidera d_q;
   - sporządź ranking dokumentów dla zapytania q, ale tylko w obozie lidera d_q.

   Zadanie

   (a) Oceń efektywność (i czas działania) tej techniki, gdy kolekcja dokumentów jest reprezentowana w postaci list sąsiadów, a zapytania są krótkie.

   (b) Oceń efektywność i czas działania, gdy zapytania to całe dokumenty (p. "Podobne strony") i kolekcja dokumentów jest reprezentowana w postaci list sąsiadów.

(c) Zaproponuj sposób wykorzystania tej metody w zadaniu klasyfikacji dokumentów.

6. [1] (7) Załóżmy, że dokumenty i zapytania reprezentowane są w przestrzeni euklidesowej (tylko!) dwu-wymiarowej. Podaj strukturę danych, która przy tym założeniu pozwoli szybko(!) realizować metodę "cluster pruning". (Przez "szybko" rozumiemy czas oczekiwany istotnie mniejszy od $\sqrt{n}$).

7. [1] (8) Zdefiniujmy wykres zależności precision/recall w następujący sposób. W układzie współrzędnych oś OX odpowiada wartości recall, a oś OY odpowiada precision. Po ułożeniu dokumentów w ranking tworzymy wykres, w którym zaczynając od pierwszego dokumentu w rankingu dodajemy dokumenty po jednym. Po dodaniu kolejnego dokumentu dodajemy nowy punkt na wykresie ilustrujący aktualne wartości precision i recall, oraz łączymy ten punkt odcinkiem z poprzednim.

   Punktem równowagi takiego wykresu nazywamy punkt, w którym wartości precision i recall są równe.

   (a) Czy wykres może mieć więcej niż jeden punkt równowagi? Jeśli odpowiedź jest pozytywna, podaj przykład. W przeciwnym razie wykaż, że jest to niemożliwe.

   (b) Czy wykres zawsze ma co najmniej jeden punkt równowagi?

   Powyższe pytania rozważ dla wykresu zmodyfikowanego, w którym każdej wartości x (recall) odpowiada jedna wartość y (precision), równa maksimum po wartościach precision (y_i) dla wszystkich punktów (x_i, y_i) z oryginalnego wykresu, w których $x_i \geq x$ (p. rozdział 8).

8. [1] (8.1) Pokaż, że dwie poniższe formuły dla miary jakości wyszukiwania F są równoważne:

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

$$F = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

   gdzie P i R oznaczają precision i recall, $\alpha \in [0,1]$ oraz $\alpha = 1/(\beta^2+1)$.

9. [0.5] (8.4) Niech $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R}$.

   (a) Określ zależność między wartościami miary F_1 dla punktów z wykresu zależności precision/recall a punktem równowagi.

   (b) Jak będzie różnić się działanie dwóch wyszukiwarek, optymalizowanych ze względu na F_2 i $F_{1/2}$.

10. [0.5] (8.2) Jednym ze sposobów oceny rangowania przez wyszukiwarkę jest sporządzenie wykresu ilustrującego wartości precision dla recall równych $0; 0.1; 02; \ldots; 0.9; 1.0$. Jednak wartość precision dla recall równego 0 nie jest zdefiniowana jednoznacznie. Podaj sensowny sposób jej "aproksymowania" i uzasadnij przydatność Twojego sposobu.

11. [0.5] (9.1) Uzasadnij (nie)słuszność następujących opinii w odniesieniu do relevance feedback z użyciem metody Rocchio.

    (a) Przykłady pozytywne (tzn. dokumenty istotne dla zapytania) są bardziej przydatne od negatywnych.

    (b) Korzystniejsze jest wykorzystanie jednego dokumentu nieistotnego (spośród wskazanych jako nieistotne) niż wielu.

    (c) Lepiej uznać za nieistotne wszystkie dokumenty, które nie zostały zaznaczone jako istotne (niż tylko wskazane jako nieistotne).

12. [0.5] (9)Chcemy zastosować metodę Rocchio jako narzędzie do uzyskanie listy dokumentów po kliknięciu w wyszukiwarce na przycisk „Znajdź podobne strony" dla konkretnej strony. Jak zmodyfikujesz algorytm w tym celu, jakie wartości współczynników α, β i γ najlepiej przyjąć?

13. [1] (9.3) W celu automatycznego tworzenia tezaurusa w oparciu o kolekcję danych, definiuje się macierz A zależności między termami. Rozważmy dwa sposoby definiowania tej macierzy:

    (a) $a_{i,j} = 1$ gdy termy t_i występuje w dokumencie d_j, $a_{i,j} = 0$ w przeciwnym razie;

    (b) $a_{i,j}$ jest równe liczbie wystąpień termu t_i w dokumencie d_j.

    W obu przypadkach podaj interpretację zawartości macierzy $C = AA^T$. Zaproponuj alternatywną definicję macierzy A, która lepiej pozwoli wyznaczać korelacje między termami (uzasadnij swój wybór).

    Podaj przykłady sytuacji, w których powyższa metoda nie wykryje korelacji między termami, które są synonimami.

14. [0.5] (9) Zapoznaj się z metodą relevance feedback stosowaną w DirectHit. W oparciu o stosowany tam pomysł wyjaśnij różnicę między lokalnym i globalnym relevance feedback i omów zalety/wady obu podejść.