Zagadnienia na egzamin z wyszukiwania informacji

Uwaga. Każdy może zabrać na egzamin ściągę – jedną kartkę formatu A4 o dowolnej treści, podpisaną własnym imieniem i nazwiskiem.

1. Tworzenie i optymalizacja indeksu odwróconego:
    a. Tokenizacja, lematyzacja, normalizacja, stop words;
    b. Indeks standardowy i pozycyjny
2. Zapytania boolowskie i optymalizacja ich realizacji.
3. Zapytania tolerujące błędy (zapytania z wildcards, k-gramy, odległość edycyjna, Jaccard coefficient, …).
4. Konstrukcja indeksu:
    a. parser, inwerter, przydział zadań, …
    b. indeks dynamiczny, „logarytmiczne" scalanie (4.3)
5. Kompresja indeksu:
    a. szacowanie rozmiaru słownika (prawo Heap'a) i kompresja słownika,
    b. kompresja list: kody gamma, uogólnione kody gamma, kody Golomba, szacowanie rozkładu słów i stopnia kompresji (prawo Zipf'a).
6. Rangowanie w modelu przestrzeni wektorowej:
    a. odległość kątowa a odległość Euklidesowa po normalizacji,
    b. współczynniki tf.idf,
    c. rangowanie przy reprezentacji danych w postaci indeksu odwróconego,
    d. metody przybliżone (cluster pruning).
7. Miary oceny wyszukiwania i rangowania
    a. precision, recall, miara F_β, precision-recall graph,
    b. porównywanie zgodności wyników – miara kappa.
8. Relevance feedback:
    a. algorytm Rocchio,
    b. na czym polega pseudo-relevance feedback, implicite relevance feedback;
    c. rozszerzanie zapytań,
    d. automatyczne tworzenie tezaurusa.
9. Rangowanie dokumentów w oparciu o model probabilistyczny i regułę Bayesa:
    a. Reprezentacja dokumentów i zapytań jako wektorów z $\{0,1\}^m$, gdzie m odpowiada liczbie termów (r. 11, **binomial** model)
    b. Ranking dokumentów w powyższym modelu, w oparciu o naiwne założenie Bayesa, sposób wyznaczania rankingu w praktyce (r. 11)
    c. Probabilistyczny model języka w oparciu o dokument, gdzie dokument/zapytanie to ciąg zmiennych losowych o wartościach w zbiorze termów (**multinomial** model). Rangowanie dokumentów w oparciu o ten model, założenie o niezależności i regułę Bayesa. Wygładzanie prawdopodobieństw. (r. 12).
10. Klasyfikacja
    a. Naiwna metoda Bayesa w modelu binomial i multinomial (patrz powyższy punkt), wygładzanie prawdopodobieństw (r. 13).
    b. Wybór cech istotnych dla klasyfikacji w oparciu o dane treningowe, miarę mutual information i test chi-kwadrat (r. 13).
    c. Ocena jakości klasyfikatora przy pomocy miary F_1 z wykorzystaniem uśredniania lokalnego i globalnego (r. 13)
    d. Klasyfikacja w modelu przestrzeni wektorowej: metoda Rocchio, k najbliższych sąsiadów. Metody liniowe (w tym liniowość naiwnego klasyfikatora Bayesa) (r. 14).

    e. Support Vector Machines: sformułowanie problemu optymalizacyjnego (wyboru „najlepszego” separatora), wyrażenie go jako zadania programowania kwadratowego, rozszerzenie dla przypadku, gdy dane nie są liniowo separowalne.
11. Klasteryzacja
    a. Metoda *k*-średnich i jej zbieżność do lokalnego optimum
    b. Hierarchiczna klasteryzacja bottom-up: metody single-link, complete link, centroid, group average.
    c. Latent Semantic Indexing – tylko idea, tylko w przypadku dopytki ustnej.
12. Link analysis:
    a. PageRank: ergodyczność, istnienie unikalnego rozwiązania, interpretacja w terminach algebry liniowej.
    b. HITS i jego rozwiązanie w terminologii algebry liniowej.
13. Porównywanie rozmiaru danych zgromadzonych przez dwie różne wyszukiwarki.
14. Wykrywanie dokumentów podobnych („prawie duplikatów”).