BÀI TẬP ÔN

Lưu ý: Bài tập ôn chỉ để nhắc lại những vấn đề có liên quan, các bạn phải ôn lại tất cả những bài đã học.

BÀI TẬP 1.

1) Từ matrận Term-Doc, xác định được từ điển và danh sách posting như sau:

Term	nDoc
bank	3
import	2
price	3
transport	3

ID	tf
1	3
2	1
3	2
2	6
3	5
1	4
2	1
3	3
1	1
2	8
3	4

Tính giá trị idf và w=tf*idf như sau:

Term	nDoc	IDF
bank	3	0.33
import	2	0.5
price	3	0.33
transport	3	0.33

ID	tf	w
1	3	0.99
2	1	0.33
3	2	0.66
2	6	3
3	5	2.5
1	4	1.32
2	1	0.33
3	3	0.99
1	1	0.33
2	8	2.64
3	4	1.32

Chuẩn hóa w, kết quả lập chỉ mục là:

Term	nDoc	IDF
bank	3	0.33
import	2	0.5
price	3	0.33
transport	3	0.33

ID	tf	w
1	3	0.588348
2	1	0.082021
3	2	0.215176
2	6	0.745646
3	5	0.815059
1	4	0.784465
2	1	0.082021
3	3	0.322763
1	1	0.196116
2	8	0.656168
3	4	0.430351

2) Quá trình xử lý truy vấn transportation pricing:
   - Tiền xử lý gồm lọc stopword và stemming câu truy vấn, được: transport price
   - Truy xuất chỉ mục có $idf_{transport}$ = 0.33, idf_{price} = 0.33.

- Chuẩn hóa: $w_{transport} = 0.71$, $w_{price} = 0.71$
- Xử lý term transport, $w_{transport} = 0.71$:
  R = {}
  Posting = {(1, 0.2), (2, 0.66), (3, 0.43)}
  ➔ R = {(1, 0.14), (2, 0.47), (3, 0.31)}
- Xử lý Term price, $w_{price} = 0.71$
  R = {(1, 0.14), (2, 0.47), (3, 0.31)}
  Posting = {(1, 0.78), (2.0.08), (3, 0.32)}
  ➔ R = {(1, 0.69), (2, 0.53), (3, 0.55)}

  Kết quả truy vấn được thứ tự tài liệu như sau:

  d1, d3, d2.

3) Các vector tài liệu ở đây có thể lấy từ ma trận Term-Doc hoặc lấy từ kết quả lập chỉ mục. Thông thường, mình sẽ lấy vector nào tốt nhất, còn vector dạng count thì không được đánh giá cao.
   Vector từ ma trận Term-Doc là:
   d1 = (3, 0, 4, 1)
   d2 = (1, 6, 1, 8)
   d3 = (2, 5, 3, 4)
   Vector từ kết quả lập chỉ mục là:
   d1 = (0.59, 0, 0.78, 0.2)
   d2 = (0.08, 0.75, 0.08, 0.66)
   d3 = (0.22, 0.82, 0.32, 0.43)
   Các vector tài liệu là:
   d1 = (0.59, 0, 0.78, 0.2) (Ngân Hàng ⇔ +)
   d2 = (0.08, 0.75, 0.08, 0.66) (Nhập khẩu ⇔ -)
   d3 = (0.22, 0.82, 0.32, 0.43) (Nhập khẩu ⇔ -)
   Lưu ý: Có thể chọn lớp Ngân hàng là - và lớp Nhập khẩu là +, kết quả sẽ khác nhưng vẫn đúng.
   Perceptron có dạng: $f(X) = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + b = 0$
   Bộ trọng số ban đầu: W = (0, 0, 0, 0, 1)
   Chọn δ=0.5
   Lưu ý: Có thể chọn δ với giá trị khác, kết quả cũng khác nhưng vẫn đúng.
   - Lần lặp 1:
   - f(d1) = 1 > 0 thỏa.
   - f(d2) = 1 > 0 không thỏa, cập nhật W = (0 - 0.5 * 0.08, 0 - 0.5 * 0.75, 0 - 0.5 * 0.08, 0 - 0.5 * 0.66, 1 - 0.5 * 1) = (-0.04, -0.38, -0.04, -0.33, 0.5)
   - f(d3) = 0.03 > 0 không thỏa, cập nhật W = (-0.04 - 0.5 * 0.22, -0.38 - 0.5 * 0.82, -0.04 - 0.5*0.32, -0.33 - 0.5 * 0.43, 0.5 - 0.5 * 1) = (-0.15, -0.79, -0.2, -0.55, 0)

   Do có cập nhật W nên tiếp tục bước lặp.

   - Lần lặp 2:
   - f(d1) = -0.35 < 0 không thỏa, cập nhật W = (-0.15 + 0.5*0.59, -0.79 + 0.5 * 0, -0.2 + 0.5 * 0.78, -0.55 + 0.5 * 0.2, 0 + 0.5 * 1) = (0.15, -0.79, 0.19, -0.45, 0.5)
   - f(d2) = -0.36 < 0 thỏa.

- f(d3) = -0.25 < 0 thỏa.

Do có cập nhật W nên tiếp tục bước lặp:

- Lần lặp 3:

- f(d1) = 0.65 > 0 thỏa.
- f(d2) = -0.36 < 0 thỏa.
- f(d3) = -0.25 < 0 thỏa.

Vậy, bộ trọng số của perceptron là: W = (0.15, -0.79, 0.19, -0.45, 0.5)

BÀI TẬP 2

Kết quả thực hiện các truy vấn như sau (N – không liên quan, R – liên quan):

Truy vấn	Kết quả truy vấn	Số tài liệu liên quan trong tập Gold
1	NRNRN NNNNR NNRNR	5
2	NNRNR NNNRN NRNN	7

1)

Truy vấn 1:

r = 1/5 = 0.2, p = 1/2 = 0.5

r = 2/5 = 0.4, p = 2/4 = 0.5

r = 3/5 = 0.6, p = 3/10 = 0.3

r = 4/5 = 0.8, p = 4/13 = 0.31

r = 5/5 = 1, p = 5/15 = 0.33

→ Nội suy 11 điểm:

r = 0.0, p = 0.5

r = 0.1, p = 0.5

r = 0.2, p = 0.5

r = 0.3, p = 0.5

r = 0.4, p = 0.5

r = 0.5, p = 0.33

r = 0.6, p = 0.33

r = 0.7, p = 0.33

r = 0.8, p = 0.33

r = 0.9, p = 0.33

r = 1.0, p = 0.33

➔ AP_1 = (5*0.5 + 6 * 0.33) /11 = 0.41

Truy vấn 2:

r = 1/7 = 0.14, p = 1/3 = 0.33

r = 2/7 = 0.29, p = 2/5 = 0.4

r = 3/7 = 0.43, p = 3/9 = 0.33

r = 4/7 = 0.57, p = 4/12 = 0.33

→ Nội suy 11 điểm:

r = 0.0, p = 0.4

r = 0.1, p = 0.4
r = 0.2, p = 0.4
r = 0.3, p = 0.33
r = 0.4, p = 0.33
r = 0.5, p = 0.33
r = 0.6, p = 0.0
r = 0.7, p = 0.0
r = 0.8, p = 0.0
r = 0.9, p = 0.0
r = 1.0, p = 0.0

➔ $AP_2 = (3*0.4 + 3 * 0.33 + 5 * 0.0) /11 = 0.2$

➔ $MAP = (AP_1 + AP_2) / 2 = (0.41 + 0.2) / 2 = 0.31$

2)

Truy vấn 1:

$P_1 = (P_1@5 + P_1@10) = (2/5 + 3/10)/2 = 0.35$

Truy vấn 2:

$P_2 = (P_2@5 + P_2@10) = (2/5 + 3/10)/2 = 0.35$

➔ Độ chính xác trung bình tại các mức cắt P@5 và P@10:
$P = (P_1 + P_2)/2 = 0.35$