Triatmoko/2900127

Exercise 1

(1)

Euclidean distance → distance between two points is the length of the path connecting them. It's like given by the Pythagorean Theorem.

In general, the distance between point x and y in Euclidean is Two vectors x,y € R^+

d(x,y) = $\sqrt{(x2-x1)2+(y2-y1)2}$, so, is $|x|^2 = |y|^2 = 1$;

= $\sqrt{|x|2+|y|2}$

= $\sqrt{\sum_{k=1}^{n}|Xk-Yk|2}$

Formula for Euclidean Distance = $\sqrt{\sum_{k=1}^{n}(dk-qk)}\ ^{2}$

(2) Does this statement remain true when the query vector is not normalized?

No, because if they are not normalized some vector may swamp the contribution of the others.

(3)

Exercise 2

Please see on zip file and you can open that file with notepad.

Exercise 3

Why highly-ranked non-relevant documents returned?

Because those document have a same number of time with other document which is relevant and also shorter than other relevant document.

And why we found some relevant documents not showed, because those documents longer than non relevant document despite they have same number of time.

Its little bit complicated to uses query to decide some document is relevant or not, because we can't decide one document is relevant or not, only from how many number of time in 1 document or if one document shorter will be referred. But to decide one document relevant or not is by human factor.

Exercise 4

(1)

Assume we have 3 query words like: some where else

Some		Where		Else	
Doc5	0.5	Doc5	0.1	Doc4	0.1
Doc10	0.2	Doc6	0.3	Doc6	0.1
Doc15	0.1	Doc8	0.1	Doc10	0.2
Doc 20	0.1	Doc10	0.1	Doc15	0.1

They should ordered by document id to be true

Some		Where		Else	
Doc5	0.5	Doc6	0.3	Doc10	0.2
Doc10	0.2	Doc8	0.1	Doc15	0.1
Doc15	0.1	Doc10	0.1	Doc4	0.1
Doc 20	0.1	Doc5	0.1	Doc6	0.1

(2)

1. List sorted by ID

Some		Where		Else	
Doc5	0.5	Doc5	0.1	Doc4	0.1
Doc10	0.2	Doc6	0.3	Doc6	0.1
Doc15	0.1	Doc8	0.1	Doc10	0.2
Doc 20	0.1	Doc10	0.1	Doc15	0.1

2. Because we need document by maximum, so we sorted list by score.

Some		Where		Else	
Doc5	0.5	Doc6	0.3	Doc10	0.2
Doc10	0.2	Doc8	0.1	Doc15	0.1
Doc15	0.1	Doc10	0.1	Doc4	0.1
Doc 20	0.1	Doc5	0.1	Doc6	0.1

3. Merger the list

Doc4	Doc5	Doc6	Doc8	Doc10	Doc15	Doc20
0.1	0.6	0.4	0.1	0.4	0.2	0.1

4.

K=1
Doc5 [0.5, 1]
Doc6 [0.3, 1]
Doc10[0.2, 1]

K=2
Doc5 [0.5, 0.7]
Doc6 [0.3, 0.6]
Doc10[0.2, 0.5]
Doc10[0.2, 0.4]
Doc8 [0.1, 0.4]
Doc15[0.1, 0.4]

K=3
Doc5 [0.5, 0.7]
Doc6 [0.3, 0.5]
Doc10[0.2, 0.4]
Doc10[0.2, 0.4]
Doc8 [0.1, 0.3]
Doc15[0.1, 0.3]
Doc15[0.1, 0.3]
Doc10[0.1, 0.3]
Doc4 [0.1, 0.3]

From highlight we know that document 10 could be at least 0.2 point and maximal 0.4, but if we look directly in the list, we can find that document 10 have maximal 0.5 point.