Information Retrieval

HW 1

2013210111 남세현

Problem 1

1.a "fools rush in"

Answer : Doc 2(1,2,3), Doc4(8,9,10), Doc7{(3,4,5), (13,14,15)}

fools: 2: <1, 17, 74, 222>; 4: <8, 78, 108, 458>; 7: <3, 13, 23, 193>;

rush: 2: <2, 66, 194, 321, 702>; 4: <9, 69, 149, 429, 569>; 7: <4, 14, 404>;

in: 2: <3, 37, 76, 444, 851>; 4: <10, 20, 110, 470, 500>; 7: <5, 15, 25, 195>;

1.b "where angles fear"

Answer : Doc4(11,12,13), Doc7(16,17,18)

where: 2: <67, 124, 393, 1001>; 4: <11, 41, 101, 421, 431>; 7: <16, 36, 736>;

angels: 2: <36, 174, 252, 651>; 4: <12, 22, 102, 432>; 7: <17>;

fear: 2: <87, 704, 722, 901>; 4: <13, 43, 113, 433>; 7: <18, 328, 528>;

1.c "fools rush in" AND "angels fear to tread"

As 1.a, the documents that contain "fools rush in" are Doc2, Doc4, Doc7.

So, there is 'AND', the document that contain "angles fear to tread" in (Doc2, Doc4, Doc7) is Doc4(12,13,14,15).

Answer : Doc4(12,13,14,15).

angels: 2: <36, 174, 252, 651>; 4: <12, 22, 102, 432>; 7: <17>;

fear: 2: <87, 704, 722, 901>; 4: <13, 43, 113, 433>; 7: <18, 328, 528>;

to: 2: <47, 86, 234, 999>; 4: <14, 24, 774, 944>; 7: <199, 319, 599, 709>;

tread: 2: <57, 94, 333>; 4: <15, 35, 155>; 7: <20, 320>;

Problem 2

2.a Why is the idf of a term always finite?

By definition, idf_t = log_10(N/df_t).

N is total number of documents and df_t is the document frequency of the term t. then df_t <= N. Except N = 0 or df_t = 0, the greatest value of idf_t is when df_t is 1(idf_t = log_10(N)). And the least value of idf_t is when df_t is N (idf_t = log_10(1) = 0).

0 <= idf_t <= log_10(N) So, idf_t is finite value

2.b What is the idf of a term that occurs in every document? Compare this with the use of the stop word lists.

Idf_t = log_10(N/df_t), and df_t is the document frequency of the term. If the term occurs in every document, then the df_t is equal with N. Then idf_t = log_10(N/N) = log_10(1) = 0.

The stop word is some kind of word lik the preposition or article(a, an, the), etc. this is not useful to search the documents because it is meanless and too many. So the stop word lists might be has value of 0 because them occur in almost documents.

2.c Compare the tf-idf weights for the four terms, where the total number of documents N = 806, 791.

By definition, W_t_d = log(1 + tf_t_d) * log_10(N / df_t). The tf-idf wight table is next.

Term	w_Doc1	w_Doc2	w_Doc3
Car	2.38423	1.151571	2.303142
moto	1.251802	3.184248	0
insurance	0	2.484876	2.396679
rent	1.769732	0	1.888881

Table 1tf-idf Weight Table for 2.c

2.d Can the tf-idf weight of a term in a document exceed 1? If so, explain why, but if not, give a counter example.

The answer is yes. Assume the term occurs on only the one document and there are so many words in that document. Just think someone copies and pastes one word on one document.

Then term frequence of therm in that document is very high over at least 10. And value of idf is log_10(N).

If N is greater than 10 or tf_t_d is enoughly greater, the value of w_t_d could exceeds 1 easily..

Problem 3. Calculate precision, recall, F1 Measure (harmonic mean), Average Precision for each row of the above table. Also calculate the Mean Average Precision for S1 and S2, respectively (no need to consider interpolation). Which values consider ranks? Which don't?

	1	*2*	*3*	*4*	*5*
S1, Q1	d3	d5	d8	d10	d11
Relevant	1	0	1	0	0
Precision	1	0.5	0.67	0.5	0.4
recall	0.25	0.25	0.5	0.5	0.5
F1	0.4	0.33	0.57	0.5	0.44
Average Precision			0.61		
S1, Q2	d1	d2	d7	d11	d13
Relevant	1	0	0	1	0
Precision	1	0.5	0.33	0.5	0.4
recall	0.33	0.33	0.33	0.67	0.67
F1	0.5	0.4	0.33	0.57	0.5
Average Precision			0.55		
S2, Q1	d6	d7	d2	d9	d8
Relevant	1	1	0	0	1
Precision	1	1	0.67	0.5	0.6
recall	0.25	0.5	0.5	0.5	0.75
F1	0.4	0.67	0.57	0.5	0.67
Average Precision			0.7533333		
S2, Q2	d1	d2	d4	d11	d14
Relevant	1	0	1	1	0
Precision	1	0.5	0.67	0.75	0.6

recall	0.33	0.33	0.67	1	1
F1	0.5	0.4	0.67	0.86	0.75
Average Precision	0.70				

Mean Average Precision	
S1	0.58
S2	0.73

The Average Precision consider rank. With above table, it could be like normal precision consider rank but it is just expression, not truely considering the rank.

Because A.P consider rank, so it is could said that Mean Average Precision consider rank.

Problem 4

4.a The balanced F measure (a.k.a. F1) is defined as the harmonic mean of precision and recall. What is the advantage or characteristic of using the harmonic mean rather than "averaging" (using the arithmetic mean)?

As we had a discussion on class, the harmonic mean consider the difference(delta value) between values. If some value are SO GREAT and others are not, then the harmonic mean might be lesser. But the arithmetic mean dosen't care about that.

4.b What are the possible values for interpolated precision at a recall level of 0?

The answers are 0 ~ 1. Logically the value of precision at a recall level of 0 is always 0 because there are no retrieval document at level of 0. So we make a value of interpolated precisoin at recall level of 0 is first value occurs on least recall level.