04 – Information Retrieval – September 12, 2016

Assume a collection of 100 documents, named 101.txt, 102.txt, 103.txt, …, 200.txt.

Assume that we run two different relevance ranking algorithms and get the following ordered lists of documents that are relevant to the user's query.

Algorithm #1		Algorithm #2		Relevant Documents	
1	104.txt	1	106.txt	101.txt	106.txt
2	106.txt	2	120.txt	102.txt	107.txt
3	108.txt	3	104.txt	103.txt	108.txt
4	120.txt	4	102.txt	104.txt	109.txt
5	122.txt	5	108.txt	105.txt	110.txt
6	124.txt				

(1) Assuming that the correct relevant documents in the collection are those listed in the third column above (101.txt … 110.txt), compute the following:

a. Number of **True Positives** for above document collection:

b. Number of **True Negatives** for above document collection:

c. Number of **False Positives** for Algorithm #1:

d. Number of **False Negatives** for Algorithm #1:

e. **Precision** of Algorithm #1:

f. **Recall** of Algorithm #1:

g. Number of **False Positives** for Algorithm #2:

h. Number of **False Negatives** for Algorithm #2:

i. **Precision** of Algorithm #2:

j. **Recall** of Algorithm #2:

Assume that we run two different relevance ranking algorithms and get the following ordered lists of documents that are relevant to the user's query.

Algorithm #3		Algorithm #4	
1	140.txt	1	150.txt
2	145.txt	2	160.txt
3	150.txt	3	140.txt
4	160.txt	4	145.txt

(2) Compute the similarity of the two ranking algorithms (#3 and #4) using the Kendall Tau Coefficient:

SOLUTIONS

(1) Question #1:

a. Number of **True Positives** for above document collection:
10 (number of correct answers)

b. Number of **True Negatives** for above document collection:
90 (number of incorrect answers)

c. Number of **False Positives** for Algorithm #1:
3 (should not have been returned, but were)

d. Number of **False Negatives** for Algorithm #1:
7 (should have been returned, but were not)

e. **Precision** of Algorithm #1:
3 (correct results returned) / 6 (total returned) = 50%

f. **Recall** of Algorithm #1:
3 (correct results returned) / 10 (total correct) = 30%

g. Number of **False Positives** for Algorithm #2:
1 (should not have been returned, but were)

h. Number of **False Negatives** for Algorithm #2:
6 (should have been returned, but were not)

i. **Precision** of Algorithm #2:
4 (correct results returned) / 5 (total returned) = 80%

j. **Recall** of Algorithm #2:
4 (correct results returned) / 10 (total correct) = 40%

(2) Question #2: For simplicity of presentation, we rename the different files as A,B,C,D:

Algorithm #3		Algorithm #4	
1	140.txt – A	1	150.txt – C
2	145.txt – B	2	160.txt – D
3	150.txt – C	3	140.txt – A
4	160.txt – D	4	145.txt – B

Rankings of documents by different algorithms:

Doc	Algorithm #3	Algorithm #4
A	1	3
B	2	4
C	3	1
D	4	2

Combinations based on rankings from Algorithm #3:

- A,B: Concordant (A is ranked higher than B under algo #3 and under algo #4)
- A,C: Discordant (A is ranked higher than C under algo #3, but lower under algo #4)
- A,D: Discordant
- B,C: Discordant
- B,D: Discordant
- C,D: Concordant

Combinations based on rankings from Algorithm #4:

- C, D: Concordant
- C, A: Discordant
- C, B: Discordant
- D, A: Discordant
- D, B: Discordant
- A, B: Concordant

So we have Kendal Tau coefficient = 4/12 C – 8/12 D = - 4/12 = - 33.33%