

دورة "استرجاع المعلومات" باللغة العربية - صيف ٢٠٢١ Information Retrieval – Summer 2021



3. Evaluation (LAB)

Tamer Elsayed Qatar University

Suppose a retrieval system produced two ranked lists of 50 documents given two queries Q1 and Q2. The ranks of relevant documents were as follows:

- Q1: relevant docs at ranks 5, 7, 9, 10, and 22
- o Q2: relevant docs at ranks 1, 2, and 50

There are 6 and 4 relevant documents in the entire collection for Q1 and Q2 respectively.

For each query, calculate the following values or indicate that there is not sufficient information to calculate it:

- O Precision
- o Recall
- o F_{0.5}
- o Precision at 10
- Precision when recall is 50%

Suppose a retrieval system produced two ranked lists of 50 documents given two queries Q1 and Q2. The ranks of relevant documents were as follows:

- **o** Q1: relevant docs at ranks 5, 7, 9, 10, and 22
- o Q2: relevant docs at ranks 1, 2, and 50

There are 6 and 4 relevant documents in the entire collection for Q1 and Q2 respectively.

Calculate mean average precision (MAP) for the system.

Consider a query that has 6 relevant documents in some collection: two that are perfect (P), one that is excellent (E), and three that are good (G). The rest are non-relevant (N). Suppose the documents retrieved in response to the query are **rated** as follows, with the start of the ranked list on the left:

G

E

N

Р

Ν

G

N

N

G

Ν

Note that anything rated P, E, or G is considered relevant.

Calculate **nDCG**. Assume that G=1, E=10, and P=100.

In one of TREC tracks, 3 teams T_1 , T_2 , and T_3 have participated and they were asked to retrieve up to 15 documents per query. In reality (with exhaustive judgments), a query Q has 9 relevant documents in the collection: A, B, C, D, E, F, G, H, and I.

The submitted ranked lists are as follows:

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T ₁	Α	M	Υ	R	K	L	В	Z	Ε	N	D	С	W		
T ₂	Υ	Α	J	R	N	Z	M	С	G	В	Χ	Р	D	K	W
T ₃	G	В	Υ	K	Ε	Α	Z	L	N	С	Н	K	W	Χ	

To construct the judging pools for Q:

• Pool A: using only the top 5 documents of each of the submitted ranked lists.

What is the <u>Average Precision</u> of T₁ if Pool A is used?

In one of TREC tracks, 3 teams T_1 , T_2 , and T_3 have participated and they were asked to retrieve up to 15 documents per query. In reality (with exhaustive judgments), a query Q has 9 relevant documents in the collection: A, B, C, D, E, F, G, H, and I.

The submitted ranked lists are as follows:

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T ₁	Α	M	Υ	R	K	L	В	Z	Ε	N	D	С	W		
T ₂	Υ	Α	J	R	N	Z	M	С	G	В	X	Р	D	K	W
T ₃	G	В	Υ	K	Ε	Α	Z	L	N	С	Н	K	W	Χ	

To construct the judging pools for Q:

• Pool **B**: using all retrieved documents of each of the submitted ranked lists.

What is the <u>Average Precision</u> of T₂ if Pool B is used?

In one of TREC tracks, 3 teams T_1 , T_2 , and T_3 have participated and they were asked to retrieve up to 15 documents per query. In reality (with exhaustive judgments), a query Q has 9 relevant documents in the collection: A, B, C, D, E, F, G, H, and I.

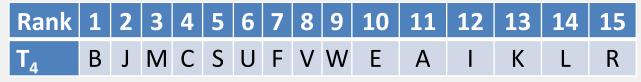
The submitted ranked lists are as follows:

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T ₁	Α	M	Υ	R	K	L	В	Z	Ε	N	D	С	W		
T ₂	Υ	Α	J	R	N	Z	M	С	G	В	Χ	Р	D	K	W
T ₃	G	В	Υ	K	Ε	Α	Z	L	N	С	Н	K	W	Χ	

There are two possible ways to construct the judging pools for Q:

- Pool A: using only the top 5 documents of each of the submitted ranked lists.
- Pool **B**: using all retrieved documents of each of the submitted ranked lists.

Team T4 didn't participate in TREC, but had the following ranked list for Q.



Is T4 penalized for not participating in TREC if pool A was used? Is it if pool B was used? Justify for both. Assume Average Precision is the evaluation measure.





In pooled assessment methodology, if the evaluation measure is precision@5, should we only make the pool depth to be 5? Why?

