Homework 2: Evaluation Metrics

> Student ID:
>
> Student Name:
>
> Lectured by: Shangsong Liang
>
> Course: Information Retrieval Course, Sun Yat-sen University

Exercise 1: Rank-based Evaluation Metrics, MAP@K, MRR@K

Assume you have three queries, and the ranking results that a system in response to these three queries are as follows:

Ranking 1 in response to query #1 is: d1, d2, d3, d4, d5, d6, d7, d8, d9, d10. Here only d1, d3, d4, d6, d7, and d10 are relevant (relevance is binary, i.e., either 1 if relevant or 0 if non-relevant) in response to query #1.

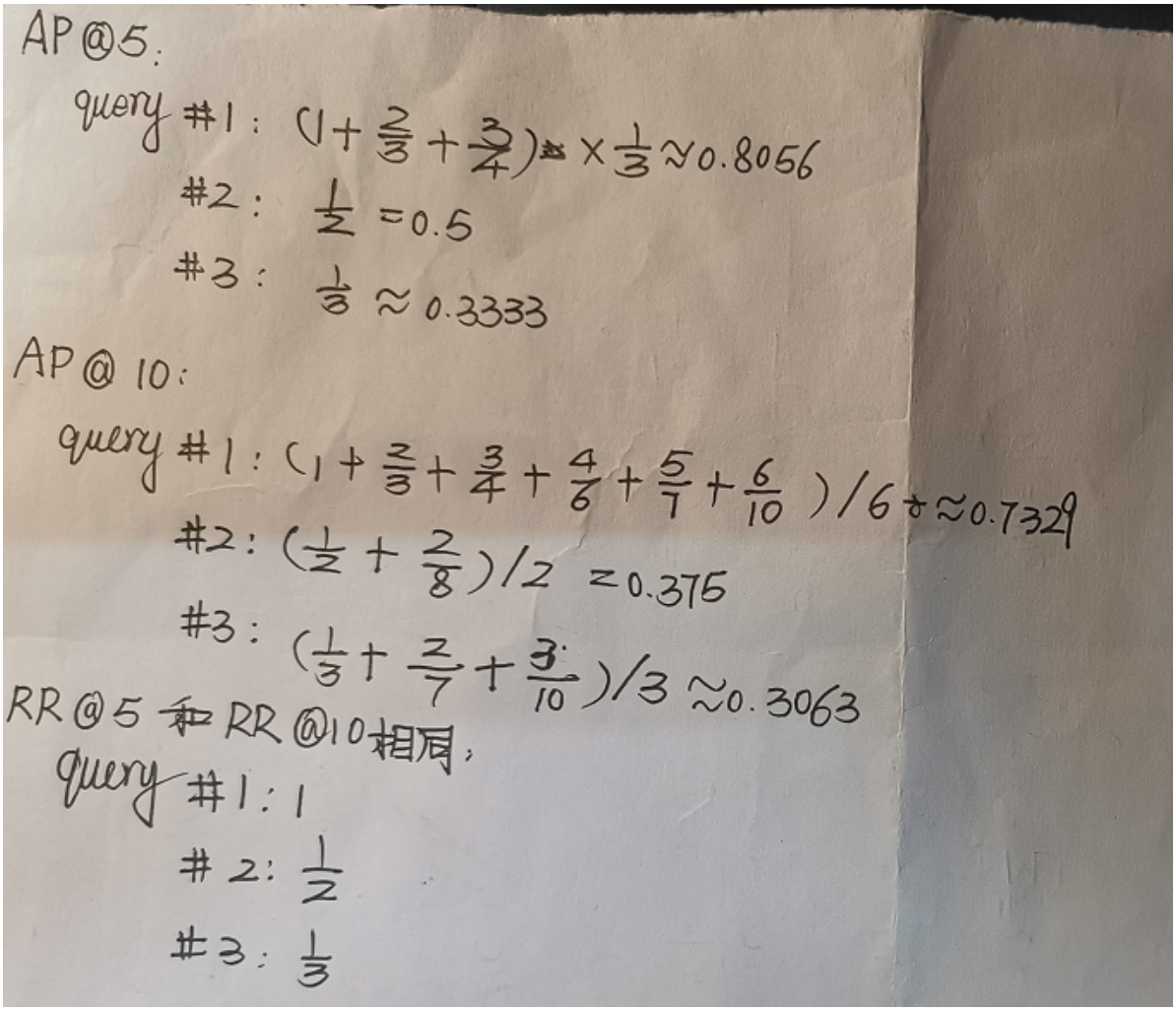
Ranking 2 in response to query #2 is: d3, d8, d7, d1, d2, d4, d5, d9, d10, d6. Here only d8 and d9 are relevant in response to query #2.

Ranking 3 in response to query #3 is: d7, d6, d5, d3, d2, d1, d9, d10, d4, d8. Here only d5, d9, and d8 are relevant in response to query #3.

Answer the questions below.

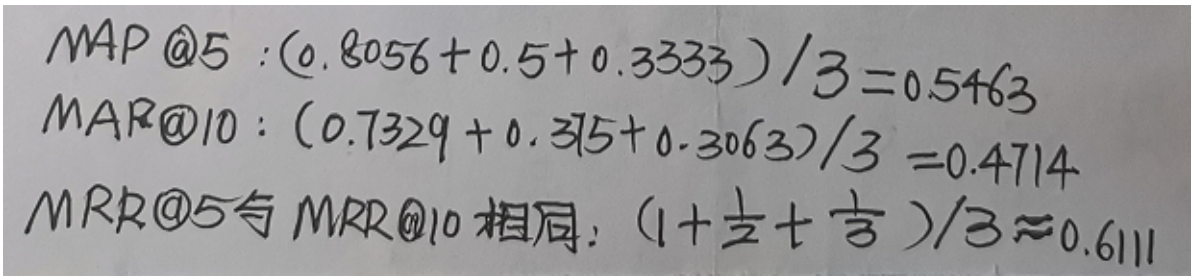
(a) Compute the scores for these metrics: AP@5 (Average Precision @5), AP@10 for each query; RR@5 (Reciprocal Rank score @5), RR@10 for each query.

解：



(b) Compute the scores for these metrics: MAP@5 (Mean Average Precision @5), MAP@10, MRR@5 (Mean Reciprocal Rank score @5), MRR@10 for this system.

解:



Exercise 2: Rank-based Evaluation Metrics, Precision@K, Recall@K, NDCG@K

Assume the following ranking for a given query (only results 1-10 are shown); see Table 1. The column 'rank' gives the rank of the document. The column 'docID' gives the document ID associated with the document at that rank. The column 'graded relevance' gives the relevance grade associated with the document (4 = perfect, 3 = excellent, 2 = good, 1 = fair, and 0 = bad). The column 'binary relevance' provides two values of relevance (1 = relevant and 0 = non-relevant). The assumption is that anything with a relevance grade of 'fair' or better is relevant and that anything with a relevance grade of 'bad' is non-relevant.

Also, assume that this query has only 7 documents with a relevance grade of fair or better. All happen to be ranked within the top 10 in this given ranking.

Answer the questions below. P@K (Precision@K), R@K (Recall@K), and average precision (AP) assume binary relevance. For those metrics, use the 'binary relevance' column. DCG and NDCG assume graded relevance. For those metrics, use the 'graded relevance' column.

Table 1 Top-10 ranking result of a system in response to a query.

rank	docID	graded relevance	binary relevance
1	51	4	1
2	501	1	1
3	21	0	0
4	75	3	1
5	321	4	1
6	38	1	1
7	521	0	0
8	412	1	1
9	331	0	0
10	101	2	1

(a) Compute P@5 and P@10.

解:

P@5	P@10
0.8	0.7

(b) Compute R@5 and R@10.

解:

R@5	R@10
4/7	1

(c) Provide an example ranking for this query that maximizes P@5.

解:

> max P@5 = 1

rank	docID	binary relevance
1	51	1
2	501	1
4	75	1
5	321	1
6	38	1

(d) Provide an example ranking for this query that maximizes P@10.

解:

> max P@10 = 0.7

rank	docID	binary relevance
1	51	1
2	501	1
4	75	1
5	321	1
6	38	1
8	412	1
10	101	1
3	21	0
7	521	0
9	331	0

(e) Provide an example ranking for this query that maximizes R@5.

解:

> max R@5 = 5/7

rank	docID	binary relevance
1	51	1
2	501	1
4	75	1
5	321	1
6	38	1

(f) Provide an example ranking for this query that maximizes R@10.

解:

> max R@10 = 1

rank	docID	binary relevance
1	51	1
2	501	1
4	75	1
5	321	1
6	38	1
8	412	1
10	101	1
3	21	0
7	521	0
9	331	0

(g) You have reason to believe that the users of this system will want to examine every relevant document for a given query. In other words, you have recent to believe that **users want perfect recall**. **You want to evaluate based on P@K**. Is there a query-specific method for setting the value of K that would be particularly appropriate in this scenario? What is it? (**Hint**: there is an evaluation metric called R-Precision, which we did not talk about in the lectures. Your answer should be related to R-Precision. Wikipedia/Google might help.)

解：

> R-Precision是给定序列中前 R 个位置的准确率；可以使用R-Precision，并尽可能让 R-Precision 变大

(h) Compute average precision (AP). What are the difference between AP and MAP (Mean Average precision)?

解：

$$AP = \frac{(1 + \frac{2}{2} + \frac{3}{4} + \frac{4}{5} + \frac{5}{7} + \frac{6}{8} + \frac{7}{9})}{7} = 0.8333$$

> 区别：AP 是对一个查询算得的平均准确率， 而MAP 则是针对多个查询的 AP 取平均值所得的值

(i) Provide an example ranking for this query that maximizes average precision (AP).

解：

> max AP = 1

rank	docID	binary relevance
1	51	1
2	501	1
4	75	1
5	321	1
6	38	1
8	412	1
10	101	1
3	21	0
7	521	0
9	331	0

(j) Compute *DCG*5 (i.e., the discounted cumulative gain at rank 5)

解:

$$DCG_5 = \sum_{i=1}^{5} \frac{rel_i}{\log_2(i+1)} = 4 + 0.6309 + 0 + 1.2920 + 1.5474 = 7.4703$$

(k) $NDCG_5$ is given by

$$NDCG_5 = \frac{DCG_5}{IDCG_5}$$

where *IDCG*5 is the *DCG*5 associated with the *ideal* top-5 ranking associated with this query. Computing *NDCG*5 requires three steps.

(i) What is the *ideal* top-5 ranking associated with this query (notice that the query has 2 *perfect* documents, 1 *excellent* document, 1 *good* document, 3 *fair* documents, and the rest of the documents are *bad*)?

解:

rank	docID	graded relevance
1	51	4
5	321	4
4	75	3
10	101	2
2	501	1

(ii) *IDCG*5 is the *DCG*5 associated with the *ideal* ranking. Compute *IDCG*5. (**Hint:** compute *DCG*5 for your ranking proposed in part (i).)

解：

> $$IDCG_5 = 4 + 2.523 + 1.5 + 0.8614 + 0.3868 = 9.2712$$

(iii) Compute *NDCG*5 using the formula above.

解：

$$NDCG_5 = \frac{DCG_5}{IDCG_5} = \frac{4 + 0.6309 + 0 + 1.2920 + 1.5474}{4 + 2.523 + 1.5 + 0.8614 + 0.3868} = 0.8056$$

(l) Are there other evaluation metrics to be used to evaluate the performance of the rankings in the table? What are the evaluation scores obtained by these metrics?

解：

(1) Fall-out

- The proportion of non-relevant documents that are retrieved, out of all non-relevant documents available:

$$\text{fall-out} = \frac{|\{\text{non-relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{non-relevant documents}\}|}$$

- evaluation scores obtained by this metric：
  假设无关文档的总数为n，则算得的 fall-out 值为 $fallout = 3/n$

(2) balanced F-score

- The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

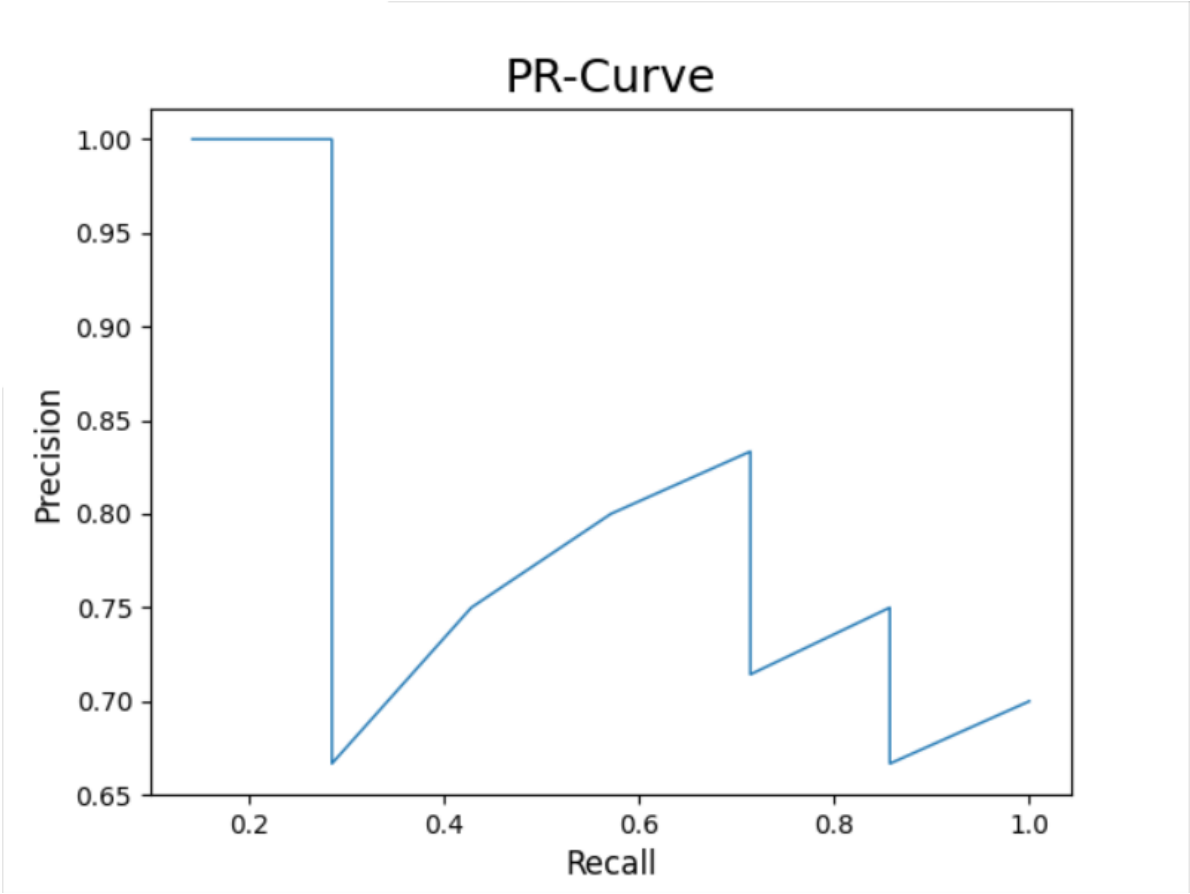
- evaluation scores obtained by this metric：
  $F@10 = 2 * 0.7 * 1/(0.7 + 1) = 14/17$

Exercise 3: Precision-Recall Curves

A Precision-Recall (PR) curve expresses precision as a function of recall. Usually, a PR-curve is computed for each query in the evaluation set and then averaged. For simplicity, the goal in this question is to **draw a PR-curve for a *single* query. Draw the PR-curve associated with the ranking in Exercise 2 (same query, same results)**. (**Hint:** Your PR curve should always go down with increasing levels of recall.)

解：

> 算得的Precision与Recall值：
>
> ```
> Precision = [1, 1, 2/3, 0.75, 4/5, 5/6, 5/7, 6/8, 6/9, 7/10]
> Recall = [1/7, 2/7, 2/7, 3/7, 4/7, 5/7, 5/7, 6/7, 6/7, 1]
> ```



Exercise 4: Other Evaluation Metrics

Except the metrics we have in our lecture slides, are there **other evaluation metrics** that can be used to evaluate the performance of specific tasks in data mining? **What are the tasks and how do to compute such evaluation metrics?** (**Hint:** Use the internet to find your answers.)

解:

(1) Spearman's rank correlation coefficient

- tasks: spearman相关系数。常用希腊字母ρ表示。 它是衡量两个变量的依赖性的非参数指标。 它利用单调方程评价两个统计变量的相关性。 如果数据中没有重复值， 并且当两个变量完全单调相关时，斯皮尔曼相关系数则为+1或−1。
- how to compute:

$$\frac{\sum_{(i,j)\in\Omega^{test}}(S_{ij}^{*}-\bar{s}^{*})(y_{ij}^{*}-\bar{y}^{*})}{\sqrt{\sum_{(i,j)\in\Omega^{test}}(S_{ij}^{*}-\bar{s}^{*})^{2}\sqrt{\sum_{(i,j)\in\Omega^{test}}(y_{ij}^{*}-\bar{y}^{*})}}}$$

其中， s_{ij}^{*} 表示模型预测中，物品 j 在用户 i 的推荐列表上的排序位置；
y_{ij}^{*} 表示按实际用户 i 对物品的评分来排序时物品 j 在 i 的推荐列表上的排序位置；
$\bar{s}^{*}$ 是 s_{ij}^{*} 的平均值；
$\bar{y}^{*}$ 是 y_{ij}^{*} 的平均值。

(2) Kendall tau distance

- tasks: 用与比较两个排序之间的相似度。

- how to compute:

$$K(\tau_1, \tau_2) = |\{(i,j) : i < j,\ (\tau_1(i) < \tau_1(j) \wedge \tau_2(i) > \tau_2(j))\ \vee\ (\tau_1(i) > \tau_1(j) \wedge \tau_2(i) < \tau_2(j))\}|$$

  其中 $\tau_1(i)$ 和 $\tau_2(i)$ 分别表示元素 i 在两个排序中的位置。