605.744: Information Retrieval
Problem Set (Module 5)

Sabbir Ahmed

October 3, 2022

1. (10%) Briefly describe the three key assumptions of the Cranfield paradigm for information retrieval evaluation.

   Answer: Relevance judgments may vary due to different systems with different pools where some systems or variants may not have even contributed to pools. However, these potential errors may not be significant if the pool size is sufficient, the collection is of a fixed size, and the evaluation is done to score relative rankings as opposed to absolute measures of performance.

2. (10%) What is pooling and why is it used in large-scale text retrieval evaluations?

   Answer: Pooling is a method used to assess the relevance of documents where only a subset of the corpus are considered for review. The subset is created by gathering the top k ranked documents by a number of different IR systems. This method of evaluation is usually performed on large corpuses where evaluating the number of documents may become infeasible.

3. (50%) Consider a query with 10 relevant documents whose docids are: D3, D27, D30, D39, D51, D54, D69, D72, D81, and D96. Assume that all other documents are not relevant. On this query two retrieval systems *FastSearch* and *Telescope* produce the following ranked lists. (Note: D17 is the 1st ranked doc by *FastSearch*; D4 is its 2nd ranked doc, etc ...)

   FastSearch: D17, D4, D69, D54, D37, D41, D89, D85, D3, D5, D91, D39

   Telescope: D3, D1, D94, D27, D50, D54, D16, D7, D72, D39, D95, D62

   (a) How many relevant documents are found by each system?

   Answer: *FastSearch* retrieved 4 relevant documents ({D69, D54, D3, D39}) and *Telescope* retrieved 5 relevant documents ({D3, D27, D54, D72, D39}).

   (b) For both systems what is P@10 (precision at 10 documents) for this query?

   Answer: P@10 can be computed by the expression $\frac{r}{10}$, where r is the number of relevant documents retrieved at 10. *FastSearch* scores 0.3 with {D69, D54, D3} while *Telescope* scores 0.5 on this metric.

   (c) For *FastSearch* what is the uninterpolated precision at 30% Recall?

   Answer: At 30% recall in *FastSearch*, the retrieved documents are {D17, D4, D69, D54, D37, D41, D89, D85, D3}, which is 9 documents. This scores the precision to $\frac{9}{12}$ or 75%.

(d) Assuming that *FastSearch* returns no other documents other than this top-12 ranked list, what is *FastSearch*'s Recall for this query?

Answer: Recall is computed with the expression: $\frac{A}{A+C}$ where A is the number of retrieved relevant documents and C is the number of relevant documents not retrieved. *FastSearch* retrieved 4 relevant documents and did not retrieve the other 6 documents, which scores its recall to $\frac{4}{10}$ or 0.4.

(e) For both systems what is average precision on this query?

Answer: The average precision is computed by summing the precisions at the retrieved relevant documents in the system. For *FastSearch* the retrieved documents are ranked {3, 4, 9, 12} with precisions of {1/3=0.33, 2/4=0.5, 3/9=0.33, 4/12=0.25}. This totals to 1.41 for *FastSearch*.

For *Telescope*, the retrieved documents are ranked {1,4,6,9,10} with precisions of {1/1=1, 2/4=0.5, 3/6=0.5, 4/9=0.44, 5/10=0.5} which totals to 2.94.

4. (15%) Given two retrieval systems (called A and B), is it possible for System A to be better than System B in average precision, but for System B to have higher P@10 than System A? Briefly justify your response.

Answer: Yes, it is possible for a system to have a higher P@10 even with a lower average precision. For example, System B may have retrieved 5 relevant documents in its top 10 ranks where System A retrieved 4, setting the P@10 scores to 0.5 and 0.4 respectively.

For simplicity, we can let all the relevant documents retrieved by System B be placed on even ranks until the top 10 ranks. This placement will attribute an average precision of System B to $0.5 \times 5 = 2.5$. If we let the 4 relevant documents retrieved by System A be placed on its top 4 ranks, then the average precision will total to $1 \times 4 = 4$.

5. (15%) Consider the contingency tables below for the word pairs (bicycle, helmet) and (bicycle, repairs). Suppose we are looking to expand a query containing the word bicycle by adding some potentially useful search terms. Using pointwise mutual information (PMI) to score candidate terms, calculate scores for both helmet and repairs, and indicate which of the two would be the better expansion term. N = 15,000 documents. Use base 2 logs.

$$PMI(x,y) = log_2\left(\frac{P(x,y)}{P(x)P(y)}\right) = log_2\left(\frac{N \times a}{(a+b)(a+c)}\right)$$

A: docs with both terms together	B: docs with first term, but not second
C: docs with second term, but not first	D: docs that contain neither term

	has helmet	missing helmet
has bicycle	22	54
missing bicycle	87	14837

	has repairs	missing repairs
has bicycle	31	45
missing bicycle	164	14760

Answer: The PMI for the first table is:

$$\begin{aligned} log_2\left(\frac{N \times a}{(a+b)(a+c)}\right) &= log_2\left(\frac{15000 \times 22}{(22+54)(22+87)}\right) \\ &= log_2\left(\frac{82500}{2071}\right) \\ &= 5.316 \end{aligned}$$

Answer: The PMI for the second table is:

$$\begin{aligned} log_2\left(\frac{N \times a}{(a+b)(a+c)}\right) &= log_2\left(\frac{15000 \times 31}{(31+45)(31+164)}\right) \\ &= log_2\left(\frac{7750}{247}\right) \\ &= 4.972 \end{aligned}$$

Since the PMI for the contingency table of the word pairs (bicycle, helmet) is higher, “helmet” should be considered as the better expansion term.