Sheetal Parikh
Problem Set – Module 5
605.744.81 Information Retrieval
Fall 2021

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

The Cranfield tests were conducted to evaluate indexing techniques in information retrieval and made three key assumptions. The first assumption was that relevance can be approximated by topical similarity meaning that all relevant documents are equally important and that the relevance of a document is independent to all other documents. The second assumption was that we have of a set of topics in which a single set of judgements can represent the entire user population. The last assumption was that all relevant documents are known.

Reference:
Pages 2- 3: https://www.inf.ed.ac.uk/teaching/courses/tts/handouts2017/VoorheesIREvaluation.pdf

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

Pooling takes the top ranked lists from n retrieval systems for a fixed set of topics and documents. The top ranked documents are combined into a big set called a pool. We can assume the documents outside the pool are non-relevant. Pooling is used in large-scale text retrieval evaluations in order to optimize how ground truth relevance judgements are collected in a test collection. Test collections are widely used to evaluate the effectiveness of an IR system and consists of documents, search topics and relevance judgements indicating what is relevant for each topic. Generally, when collecting relevance judgements in a test collection, all documents for each search would be judged to make a list of all the relevant documents. However, this would not be efficient and would be very costly for a large-scale system that may contain thousands of documents. Therefore, pooling greatly optimizes the process of assessing what document is relevant since only the top documents that are retrieved are pooled and evaluated.

References:
Lecture 5B
Pages 1 – 2: *https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=51236*

3.) *Suppose a query has 8 relevant documents with docids: D17, D18, D21, D39, D51, D54, D67, and D79, and assume that all other documents are not relevant. On this query two retrieval systems Bang and Giggle produce the following top-15 ranked lists. (Note: D2 is the 1st ranked doc by Bang; D13 is its 2nd ranked doc, etc...)*

Bang: D2, D13, D94, D67, D14, D20, D54, D18, D7, D79, D39, D99, D17, D27, D85
Giggle: D18, D3, D67, D54, D37, D21, D99, D82, D1, D5, D91, D34, D63, D17, D40

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

Bang: D2, D13, D94, D67, D14, D20, D54, D18, D7, D79, D39, D99, D17, D27, D85
6 relevant documents are found by the Bang system

Giggle: D18, D3, D67, D54, D37, D21, D99, D82, D1, D5, D91, D34, D63, D17, D40
5 relevant documents are found by the Giggle system

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

Bang: 1.)D2, 2.)D13, 3.)D94, 4.)D67, 5.)D14, 6.)D20, 7.)D54, 8.)D18, 9.)D7, 10.)D79
P@10 = 4 relevant documents/10 documents = 0.40

Giggle: 1.)D18, 2.)D3, 3.)D67, 4.)D54, 5.)D37, 6.)D21, 7.)D99, 8.)D82, 9.)D1, 10.)D5
P@10 = 4 relevant documents/10 documents = 0.40

*(c) For **Bang** what is the uninterpolated precision at 50% Recall*

System - BANG		
Document	**Precision**	**Recall**
D2	0	0
D13	0	0
D94	0	0
D67	¼ = 0.25	1/8 = 0.125
D14	1/5 = 0.25	1/8 = 0.125
D20	1/6 = 0.1667	1/8 = 0.125
D54	2/7 = 0.2857	2/8 = 0.25
D18	3/8 = 0.375	3/8 = 0.375
D7	3/9 = 0.333	3/8 = 0.375
D79	4/10 = 0.40	4/8 = 0.50
D39	5/11 = 0.4545	5/8 = 0.625
D99	5/12 = 0.4167	5/8 = 0.625
D17	6/13 = 0.4615	6/8 = 0.75
D27	6/14 = 0.4286	6/8 = 0.75
D85	6/15 = 0.40	6/8 = 0.75

Uninterpolated precision at 50% Recall = 40%

(d) For Giggle what is the interpolated precision at 20% Recall?

UNINTERPOLATED

System – GIGGLE		
Document	**Precision**	**Recall**
D18	1/1 = 1	1/8 = 0.125
D3	½ = 0.50	1/8 = 0.125
D67	2/3 = 0.6667	2/8 = 0.25
D54	¾ = 0.75	3/8 = 0.375

We don't have the standard recall values of 0%, 10%, 20%, 30%, etc. Therefore, in order to get those values we could use interpolation.

The first recall value of a relevant document that we have is 12.5% which has a precision value of 100%. The standard recall values of up to 12.5%, which would be 0% and 10% would use the precision value of 100%. The next standard value we have is 20%, the only value over 12.5% but below 25%, which would use the precision of 66.67% which has an actual recall value of 25%. The next standard recall value would be 30%. We would use the precision value of 75% for the 30% recall since 30% is more than 25% and less than 37.5%.

Standard Recall values using interpolation as described above:

0%: Precision = 100%
10%: Precision = 100%
20%: Precision = 66.67%
30%: Precision = 75%

Interpolated Precision at 20% Recall = 66.67%

(e) *For Bang what is average precision on this query?*

Sum of the precision at each rank of a relevant document(as seen in part C) = 0.25 + 0.2857 + 0.375 + 0.40 + 0.4545 + 0.4615 = 2.2267

Average Precision = 2.2267 / 8 relevant documents = 0.2783

References:
Lecture 5C

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

It is possible for a System A to be better than System B in P@10 but for System B to have a higher mean average precision than System A. When computing the precision at a specific number of document such as P@10 (10 documents), we are only looking at the precision at one query. It is possible that a ranking was produced by a query that was overly simple and so System A performed better in P@10. Only evaluating precision can be misleading because two systems could even be the same system but running at different thresholds affecting the set that is retrieved or not retrieved which overall impacts the precision. The mean average precision is a single value that represents the mean of the average precision values at the ranks where you have relevant document for different queries. Therefore, P@10 can be higher for System A for a certain query but System B can have an overall better mean precision for multiple queries, causing System B to have a higher mean average precision.

References:
Lectures 5A and 5B
Pages 158-163: https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

5.) *Suppose we are experimenting with Rocchio's method for automated relevance feedback and taking the top 50 documents as presumed relevant documents and taking documents initially ranked 951 to 1000 as presumed nonrelevant documents. What differences in behavior would you expect in Condition 1: (α=0.5, β=0.5, and γ=0) and Condition 2: (α =0, β =1, and γ=0)?*

Using Rocchio's method, the parameters α, β, and γ will control the amount of movement of the modified query vector towards or away from the original query, relevant documents and non-relevant documents. Both conditions have γ = 0, which means that the non-relevant documents are not considered important and have no effect on the movement of the modified vector. The remaining parameters α and β represent the weight on the original query and the relevant documents, respectively. We can see that Condition 1 has a higher weight on the original query and Condition 2 has a higher weight on the relevant documents. Condition 1 has a higher weight on the original query at α = 0.5, versus Condition 2 which has no weight at α = 0. On the other hand, condition 2 has a higher weight on the relevant documents at β = 1 versus Condition

1 which has $\beta = 0.5$. Therefore, we would expect the modified query at Condition 2 to move in a direction that is closer to the relevant document centroid compared to the modified query at Condition 1 which would move in the same direction of the relevant document centroid but would still be closer to the original query than Condition 2. Because condition 2 places no weight on the original query, it has a possibility of overfitting compared to condition 1.

References:

Lecture 5E
Pages 178 – 186: https://nlp.stanford.edu/IR-book/pdf/09expand.pdf