Chapter No. 8 - Evaluation in Information Retrieval

<Food for Thoughts>

1. What is the common scheme for evaluation of information retrieval experiments? explain.
2. What is the relationship between the value of F1 and the break-even point?
3. Consider an information need for which there are 4 relevant documents in the collection.
   Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):
   System 1 R N R N N N N N R R
   System 2 N R N N R R R N N N
   a. What is the MAP of each system? Which has a higher MAP?
   b. Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?
   c. What is the R-precision of each system? (Does it rank the systems the same as MAP?)
4. The balanced F measure (a.k.a. F1) is defined as the harmonic mean of precision and recall. What is the advantage of using the harmonic mean rather than "averaging" (using the arithmetic mean)?
5. What is Fall-out for a system? when it is good to evaluate with this measure?
6. Why evaluation of ranked retrieval is more challenging? Explain.
7. What are some of the drawbacks of cumulative gain?
8. How Normalized Discount Cumulative Gain (NDGC) overcome all the drawbacks of a ranked retrieval? Explain.
9. What is so good about A/B Testing for IR systems? Discuss it use as an instrument for improving IR systems.
10. Differentiate between dynamic summaries and static summaries over search results snippets.