Information Retrieval Quiz 2 Rubric **Quiz-2 [21 Feb]**

Section 1: MCQs [2 marks each]

Question 1: The size of the dictionary in inverted index decreases when doing stemming

A. True B. False

Answer: True

Stemming decreases the size of the dictionary as multiple words are now mapped to their root word.

Question 2: Which of the following is TRUE about Recall?

A. Recall is a non-increasing function of the number of relevant docs retrieved.
B. Recall is a non-decreasing function of the number of relevant docs retrieved.
C. A system that returns all relevant docs has 100% recall.
D. A system cannot have 100% recall.

Answer: B, C

2 marks if B and C ticked. Only 1 mark if EITHER B OR C ticked. No marks if any other option is ticked.

Question 3: While indexing a large set of documents, what might be a better option?

A. Add one document at a time to the index
B. Index multiple documents at once
C. Depends on the document content and length
D. Can't say

Ans B) Index multiple documents at once

Bulk indexing can be much faster than indexing individual documents because less number of merges (and hence disk seeks) may be required to construct the final index.

Question 4: Which of the following are TRUE?

A. BM25 is sensitive to term frequency only
B. BM25 is sensitive to document length only
C. BM25 is sensitive to both term frequency and document length
D. BM25 is not sensitive to both term frequency and document length

Answer: A, B

2 marks if both A and B ticked, only 1 mark if either A or B ticked, 0 if any other option ticked.

Question 5: About Cumulative gain(CG) and Discounted cumulative gain(DCG) select options which are true:

A. CG is affected by changes in ordering of search results and hence, DCG is used for more accurate measure.
B. Highly relevant documents are more useful if appearing earlier in search results.

C. CG includes the position of a result in the calculation of gain of the result set.
D. DCG fluctuates a lot based on the length of the corpus.

Answer: B), D)

2 marks if both B and D ticked, 1 mark if either B or D ticked, 0 marks if any other option ticked.

Section 2: Descriptive and Numerical Questions [10 marks each]

Question 6: What are the benefits (minimum 3) of Probabilistic Information Retrieval over traditional Boolean Information Retrieval? Explain with an example. What evaluation metrics are used for both IR techniques? Which evaluation metric is most appropriate for the following applications: (i) Google search, (ii) User retrieval on Social Media, say X, and (iii) Finding relevant papers on some topic from a Digital library, say Google Scholar.

Answer:

Benefits (at least 3, one mark for each, upto 3 marks)

- Ranking Results by Relevance
- Handling Ambiguity and Synonymy
- Scalability and Flexibility
- Partial Matching
- Reduced Information Overload
- Effective in Noisy Environments

Example: 2 marks

What evaluation metrics are used for both IR techniques?:

Any 2, 1 mark each, total 2 marks:

- Precision
- Recall
- F1 Score
- Mean Average Precision (MAP)
- Normalized Discounted Cumulative Gain (nDCG)
- Discount Cumulative Gain (DCG)

Most appropriate evaluation metric:

i) Google Search: Normalized Discounted Cumulative Gain (nDCG)

Reason: Google's primary objective is to provide users with highly relevant results quickly. nDCG is suitable because it takes into account both the relevance of documents and their position in the ranked list, aligning with Google's goal of presenting the most relevant results at the top.
Just need to write nDCG for 1 mark, no explanation required.

ii) User retrieval on Social Media: Mean Average Precision (MAP)

Reason: Social media platforms prioritize user engagement and satisfaction. MAP is appropriate because it calculates the average precision across all queries, indicating how well the platform retrieves relevant content for various user searches.

Just need to write MAP for 1 mark, no explanation required.

iii) Finding Relevant Papers on a Digital Library: Precision at k (P@k)

Reason: In academic contexts, users often want precise and relevant results, especially when searching for scholarly articles. Precision at k, where k represents the number of top-ranked results, is suitable because it focuses on the precision of the initial results. Users searching on Google Scholar typically expect highly relevant papers among the top results.

Need to write Precision at K for 1 mark, but Precision should be fine as well.