Evaluation:-

1. **Recall@k**: Measures the fraction of relevant documents retrieved among the top k results. High recall means the relevant documents are among the top k results, indicating effective retrieval.
2. **Precision@k**: Calculates the fraction of relevant documents in the top k results. Precision focuses on the quality of the top k results.
3. **Mean Average Precision (MAP)**: Averages the precision values obtained after each relevant document is retrieved, providing a single-figure measure of quality across recall levels.
4. **Normalized Discounted Cumulative Gain (NDCG)**: Considers the position of relevant documents in the retrieved list, giving higher scores for relevant documents appearing earlier.
5. **Mean Reciprocal Rank (MRR)**: Evaluates the rank position of the first relevant document. Higher MRR indicates relevant documents are found earlier in the ranking.

**Generation Evaluation Metrics**

1. **BLEU (Bilingual Evaluation Understudy)**: Measures the overlap between the generated text and reference text using n-gram precision. Commonly used in machine translation and text generation.
2. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: Evaluates text by comparing the overlap of n-grams, particularly focusing on recall. Common variants are ROUGE-N, ROUGE-L (Longest Common Subsequence), and ROUGE-S (Skip-bigram).
3. **METEOR (Metric for Evaluation of Translation with Explicit ORdering)**: Considers precision, recall, synonymy, stemming, and paraphrasing. Designed to address some limitations of BLEU and ROUGE.
4. **CIDEr (Consensus-based Image Description Evaluation)**: Measures the similarity of generated sentences to reference sentences, giving higher scores for sentences that are more similar to multiple references.
5. **BERTScore**: Utilizes contextual embeddings from BERT to evaluate the similarity between generated text and reference text, considering semantic meaning rather than just surface-level n-gram overlap.
6. **Quality, Informativeness, and Relevance (QIR)**: Human-evaluated metrics where annotators assess the quality of the response, the informativeness of the content, and the relevance to the query.

**Joint Evaluation Metrics**

1. **Exact Match (EM)**: Measures the percentage of exact matches between the generated answers and the reference answers. It is strict and typically used in question-answering tasks.
2. **F1 Score**: Evaluates the balance between precision and recall, especially useful when partial matches are important. It’s often used in tasks like question answering where the exact wording might differ.
3. **Answer Overlap Metrics**: Measures the overlap between the generated answers and reference answers, considering both precision and recall.
4. **Coverage and Diversity Metrics**: Ensures the generated responses cover a wide range of topics (coverage) and are not repetitive or overly similar (diversity).

Code:-

Attached in py file:-

Lib:-

 pip install numpy scikit-learn nltk rouge-score bert-score