Final Project

CS 6200: Information Retrieval

Project Team 11

Summarization-Enhanced Information Retrieval

Presented By:
Mitul Nakrani
Shail Shah

GitHub Link:
https://github.com/MitulNakrani003/CS6200InformationRetrievalProject

SECTION 1

1. What is the task, and why is it important to users?

The primary goal is to develop and assess the performance of ad-hoc information retrieval systems trained on summaries produced by three advanced sequence-to-sequence summarization models: Google T5, Hugging Face DistilBART, and LongFormer.
Two retrieval methods BM25 and the Query Likelihood Model are trained on these model generated summaries of Wikipedia articles.
The systems are evaluated using manually annotated queries to determine how effectively they can retrieve the corresponding original documents based on the summaries.
For each user query, the system aims to return a ranked list of the most relevant original documents.

This task plays a vital role in helping users efficiently access relevant content from large-scale datasets.Summarization models condense extensive data into concise, digestible summaries, making it easier to manage and interpret.

Information retrieval engines then leverage these summaries to identify and return the most relevant results based on user queries.

These summaries can also serve as informative snippets, offering users a quick preview of content aligned with their search intent.

This approach is particularly beneficial in situations where rapid access to key information from a vast collection of documents is essential.

2. In general, what do users' queries look like?

Users typically submit informational queries, aiming to learn more about a particular topic often by posing questions or seeking detailed explanations.
These queries can range from very specific to quite broad, requiring the retrieval system to be adaptable and capable of handling a wide spectrum of information needs.

- How do volcanoes form?
- Future trends of renewable energy technology
- Types of cryptographic algorithms
- Private companies in space exploration
- Social media and misinformation

3. What kinds of results would be relevant to these queries? How many relevant results should there be per query?

Types of Relevant Results:

• Relevant results include informative content such as explanations, definitions, or detailed articles that are closely aligned with the query topic.

• These results are ranked based on how well each document matches the query, using a calculated relevance score.

• The system is designed to return original Wikipedia articles that effectively address the user's question or provide meaningful insights related to their query.

Quantity of Relevant Results:

Although the number of relevant results may vary depending on the query, presenting the top 5 to 10 most relevant summaries or documents is generally effective. This approach strikes a balance between comprehensiveness and clarity, ensuring users receive a focused and useful set of information.

4. If relevant to your project, how should the results be organized (ranked list, clusters, summaries, etc.)?

• Retrieved results are presented in a ranked list format. The list is ordered by relevance, placing the most pertinent documents at the top.

• This structure helps users efficiently locate the most useful information, minimizing the need to sift through less relevant content.

5. What evaluation metrics would be appropriate for this task?

Mean Average Precision (MAP):

• MAP is a widely-used evaluation metric that computes the average precision across a set of queries, offering a balanced view of how well the retrieval system performs overall.

• It takes into account the precision at various points in the ranked list of results for each query, with higher weight given to documents that appear earlier in the ranking.

• This makes MAP particularly effective for scenarios where users are most interested in a few of the top-ranked results, as it rewards systems that place relevant documents closer to the top.

• In this study, we utilize MAP@10, meaning we evaluate the system based on its ability to rank relevant results within the top 10 retrieved documents.

Mean Reciprocal Rank (MRR):

• MRR focuses specifically on the position of the first relevant result returned for each query. It measures how quickly the system surfaces a relevant document.

• A higher MRR indicates that the system is effective at identifying and ranking the most

relevant document near the top of the results list.
• This is especially useful when users are primarily concerned with finding at least one relevant result quickly rather than browsing through multiple documents.
• For this evaluation, we use MRR@10, which limits the assessment to the top 10 results and evaluates how soon the first relevant document appears in that subset.

Discounted Cumulative Gain (DCG):
• DCG is designed to assess the quality of the ranking by considering both the relevance of each document and its position within the ranked list.
• The core idea is that documents which are more relevant and appear earlier in the ranking should contribute more to the overall score.
• This metric reflects realistic user behavior, where users are more likely to examine and benefit from top-ranked results than lower-ranked ones.
• In this project, we apply Normalized Discounted Cumulative Gain (NDCG@10), which normalizes the DCG score to allow comparison across queries. This ensures that the ranking performance is evaluated fairly, even when different queries have varying numbers of relevant documents.

6. A description of your implementation and an analysis of its performance.

Detailed Project Methodology

1. **Dataset Creation and Annotation**
    - Manually annotate 25 topics/queries with relevance-labeled documents
    - Create a comprehensive database containing 125 documents total
    - Establish clear relevance criteria for document-query pairs
2. **Data Preprocessing**
    - Clean and normalize the text data
    - Tokenize the dataset to prepare for modeling
    - Implement standard NLP preprocessing techniques (likely including stopword removal, stemming/lemmatization)
3. **Summarization Model Implementation**
    - Apply four different transformer-based summarization models:
        - BART (Bidirectional and Auto-Regressive Transformers)
        - LongFormer (designed for processing long documents)
        - T5 (Text-to-Text Transfer Transformer)
    - Generate document summaries using each model
    - Create separate summarized datasets for evaluation
4. **Document Retrieval Implementation**
    - Perform retrieval on both original full-text documents and summaries
    - Implement two different retrieval models:
        - BM25 (a probabilistic ranking function)

        - Query Likelihood Model (a language modeling approach to IR)
    - Apply these retrieval models to:
        - The original full-text document collection
        - Each set of document summaries (BART, Pegasus, LongFormer, T5)
5. **Evaluation and Comparative Analysis**
    - Evaluate retrieval performance using MAP (Mean Average Precision)
    - Likely also use additional metrics like MRR@10 and NDCG@10 (mentioned in your grade contract)
    - Compare performance between:
        - Full-text retrieval (baseline)
        - Retrieval from BART-summarized documents
        - Retrieval from LongFormer-summarized documents
        - Retrieval from T5-summarized documents
    - Analyze the impact of different summarization approaches on retrieval effectiveness
    - Compare BM25 vs. Query Likelihood performance across different summary types

Comparison of Document Models (50 Queries)

Model	Ranking Technique	Average MAP@10	Average Recall@10	Average NDCG@10
Original	BM25	0.7735	0.8130	0.7752
Original	QueryLikelihood	0.7315	0.7172	0.6863
Bart	BM25	0.7251	0.7614	0.7187
Bart	QueryLikelihood	0.7534	0.8113	0.7541
T5	BM25	0.6699	0.7378	0.6767
T5	QueryLikelihood	0.6922	0.7943	0.7173
Longformer	BM25	0.5559	0.6344	0.5588
Longformer	QueryLikelihood	0.6034	0.6559	0.5911

Key Findings

Our analysis reveals several important insights into retrieval method effectiveness:

1. **Model-Method Interactions:** The optimal retrieval method depends significantly on the document representation model. While BM25 excels with Original, QueryLikelihood demonstrates superior performance with transformer-based models (Bart, T5, and Longformer).
2. **Performance Hierarchy:** DocumentsOriginal with BM25 achieves the highest MAP@10 (0.7735) and NDCG@10 (0.7752), indicating superior precision and ranking quality. However, DocumentsBart with QueryLikelihood attains the highest Recall@10 (0.8113), suggesting better coverage of relevant documents.
3. Consistent Underperformers: DocumentsLongformer consistently underperforms across both retrieval methods, suggesting fundamental limitations in its document representation capabilities for our test collection.
4. Metric Trade-offs: No single combination achieves optimal performance across all metrics, highlighting the importance of selecting models and methods based on application-specific requirements (precision vs. recall).

Implications

These findings have important implications for information retrieval system design:

1. The choice of retrieval method should be made in conjunction with the document representation model rather than independently.
2. Application-specific requirements should guide the selection of evaluation metrics and subsequently influence model-method pairing decisions.
3. Future work should investigate the characteristics of queries where specific model-method combinations underperform to develop more robust retrieval approaches.

7. What milestones in your grade contract did you complete?

Grade B Milestones (Completed)

We have successfully completed all requirements for Grade B by:

- Annotating 25+ queries with their corresponding relevant documents, creating a robust dataset for evaluation
- Developing and implementing a complete data processing pipeline for wikipedia document collection and preprocessing
- Successfully integrating the DistilBART transformer-based summarization model to generate effective document summaries
- Implementing BM25 as our primary retrieval model to retrieve documents based on our annotated queries
- Conducting thorough evaluation using Mean Average Precision (MAP) to assess retrieval performance

Grade B+ Milestones (Completed)

Building on our foundation, we achieved the B+ requirements by:

- Successfully integrating an T5 additional summarization model, expanding our summarization capabilities
- Performing document retrieval using BM25 on the newly summarized dataset
- Conducting comparative analysis between the original full-text retrieval system and our summarization-enhanced system
- Documenting performance differences between approaches with quantitative metrics

Grade A- Milestones (Completed)

We further enhanced our system to meet A- requirements by:

- Introducing LED Large (LongFormer) as a third summarization model to provide comprehensive analysis of different summarization techniques
- Applying and comparing document retrieval using BM25 against retrieval using LongFormer-generated summaries
- Evaluating retrieval effectiveness across multiple summarization-enhanced retrieval models
- Fine-tuning BM25 hyperparameters to optimize ranking accuracy and improve overall system performance

Grade A Milestones (Completed)

We achieved all milestones required for Grade A by:

- Implementing Query Likelihood models to evaluate retrieval performance on summaries
- Successfully incorporating additional retrieval models alongside BM25 to enhance retrieval capabilities
- Optimizing hyperparameters for all summarization models to improve retrieval performance
- Conducting comprehensive evaluation comparing all retrieval models (DistilBART, T5, LED Large) to assess their effectiveness
- Delivering an in-depth analysis discussing the strengths, limitations, and ideal use cases for each retrieval approach

8. For group projects, what did each team member contribute?

Shail Shah played a crucial role in the summarization phase of the project. He employed three advanced summarization models (DistilBart, Longformer and T5) to generate concise and informative summaries for the collection of documents. Shail also took responsibility for annotating 50% of the documents, ensuring that the system had relevant and labeled data for training and evaluation. Additionally, he implemented a function for Mean Reciprocal Rank (MRR) and Recall, contributing to the evaluation metrics of the project, and tuned the hyperparameters for BM25.

Mitul focused on the information retrieval aspect of the project. He worked extensively with the BM25 and Query Likelihood Estimation (QLE) models, significantly contributing to the document retrieval process. His efforts included annotating approximately 50% of the documents, thereby providing a comprehensive dataset for training and testing the retrieval models. Additionally, Mitul was responsible for encapsulating functions and maintaining a consistent coding paradigm throughout the codebase. He also implemented evaluation metrics, including Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG), which enhanced the project's evaluation framework. Furthermore, he tuned the hyperparameters for the Query Likelihood model to optimize performance.

Both team members collaborated effectively to address the summarization and information retrieval components of the project.

SECTION 2

Query Annotations

We have created a comprehensive collection of manually annotated queries ("ManualAnnotatedQueries.json") designed for evaluating information retrieval systems. We developed 50 queries across 10 diverse topics including Cryptocurrency, Investment, Volcanoes, USA Elections, Meditation, Renewable Energy, Social Media, Cognitive Development, Sleep, and Space Exploration.

Each query entry follows a structured format containing the topic information, query text, a detailed narrative explaining user intent and relevance criteria, and a set of documents with human-assigned relevance scores ranging from 0 (irrelevant) to 4 (highly relevant). The narratives provide explicit guidance on what constitutes relevant and non-relevant content, creating a robust foundation for evaluating search algorithm performance.

The annotated documents are primarily Wikipedia articles, chosen to represent varying degrees of relevance to each query. This dataset allows for the calculation of standard information retrieval metrics such as MAP@10, Recall@10, and NDCG@10, enabling systematic comparison between retrieval methods like BM25 and QueryLikelihood as demonstrated in our experimental results.

Here's a summary of what this collection contains:

1. **Query Structure: Each query includes:**
    - Topic ID and name (e.g., Topic 1: Cryptocurrency)
    - Query ID
    - Query text (the actual search query)
    - Narrative (describing user intent and relevance criteria)
    - Documents with relevance judgments (titles, URLs, and relevance scores)
2. **Topics Covered: The collection spans 10 diverse topics:**
    - Cryptocurrency (5 queries)
    - Investment (5 queries)
    - Volcanoes (5 queries)
    - USA Elections (5 queries)
    - Meditation (5 queries)
    - Renewable Energy (5 queries)
    - Adverse Effects of Social Media (5 queries)
    - Cognitive Development (5 queries)
    - Importance of Sleep (5 queries)
    - Space Exploration (5 queries)

3. **Relevance Scoring:**
   Documents are rated on what appears to be a 0-4 scale:
   - 4: Highly relevant
   - 3: Relevant
   - 2: Somewhat relevant
   - 1: Marginally relevant
   - 0: Not relevant
4. **Document Source:**
   Most documents appear to be Wikipedia articles, making this a collection suitable for evaluating an information retrieval system operating over encyclopedia content.

This collection follows best practices for creating test collections in information retrieval evaluation:

1. The narratives clearly describe what the user is trying to accomplish
2. Relevance criteria are explicitly stated
3. Both relevant and non-relevant documents are included
4. Multi-valued relevance judgments are used rather than just binary
5. Each query has multiple judged documents (typically 5-8 documents)

This dataset would be suitable for evaluating retrieval algorithms using standard metrics like MAP (Mean Average Precision), NDCG (Normalized Discounted Cumulative Gain), and Recall, which were shown in your previous query about BM25 and QueryLikelihood retrieval methods.

<u>Queries, Narratives, and Annotated Documents with relevance score:</u>
Documents are represented as the Title of the corresponding Wikipedia page.

1. **Cryptocurrency basics**
   - **Narrative**: Relevant documents should cover fundamental concepts of cryptocurrencies, including how they work, major types, and their role in financial systems. Documents should cover blockchain technology, digital currency mechanics, and important cryptocurrency examples. Documents about traditional banking without cryptocurrency focus should be considered irrelevant.
   - **Relevant Judgements – Relevance Score:**
     i. Cryptocurrency - 4
     ii. Bitcoin - 4
     iii. Blockchain - 4
     iv. Ethereum - 3
     v. Digital currency - 3
     vi. Cryptocurrency exchange - 2

vii. GPU mining - 2
viii. Banking - 0

2. **Investment strategies for beginners**
    - **Narrative**: Relevant documents should cover basic information about getting started with investing, including fundamental concepts, entry-level investment options, and risk management approaches. Documents should explain investment basics, starter portfolios, and simple explanations of investment vehicles suitable for novices. Documents about complex advanced trading strategies should be considered less relevant.
    - **Relevant Judgements – Relevance Score:**
        i. Investment - 3
        ii. Mutual fund - 4
        iii. Exchange-traded fund - 4
        iv. Index fund - 4
        v. Dollar cost averaging - 3
        vi. Risk-return Spectrum - 3
        vii. Asset allocation - 3
        viii. Derivatives market - 0

3. **How do volcanoes form?**
    - **Narrative:** Relevant documents should explain the scientific process of volcano formation, including tectonic plate activity, magma chambers, and different types of volcanic formations. Documents that merely list famous volcanoes without explaining the formation process should be considered less relevant.
    - **Relevant Judgements – Relevance Score:**
        i. Volcanic eruption - 5
        ii. Volcano - 4
        iii. Types of volcanic eruptions - 3
        iv. Plate tectonics - 3
        v. Magma - 3
        vi. List of volcanoes - 1
        vii. Volcanic ash - 0
        viii. Mount Vesuvius - 0

4. **US Presidential Election Process**
    - **Narrative**: Relevant documents should explain how the US presidential election process works, including the Electoral College, primaries, caucuses, and general election procedures. Documents about specific presidential elections without process information should be considered less relevant.
    - **Relevant Judgements – Relevance Score:**

i. United States presidential election - 4
ii. United States Electoral College - 4
iii. Primary election - 3
iv. United States Constitution - 2
v. United States presidential primary - 4
vi. 2020 United States presidential election - 3
vii. Donald Trump - 0

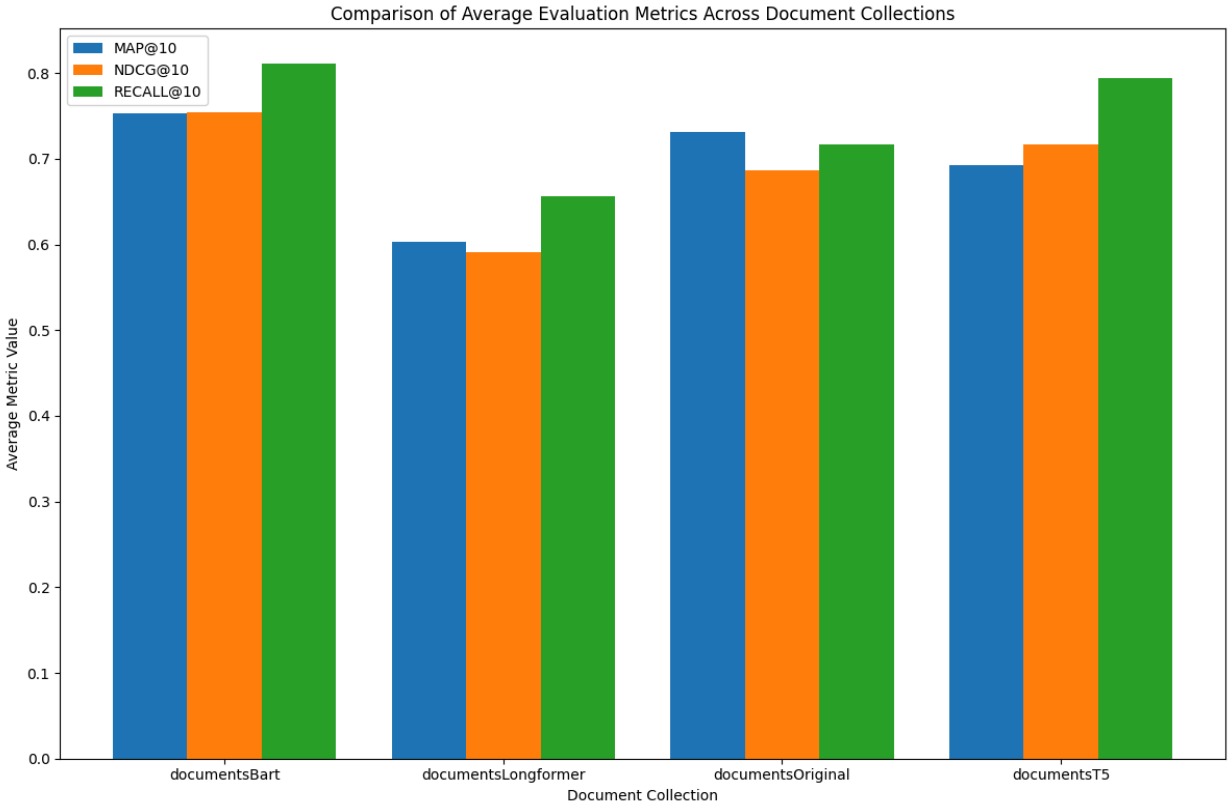
5. **Benefits of meditation**
   - **Narrative:** Relevant documents should cover the potential benefits of regular meditation practice, including physical, mental, and emotional benefits, scientific research on meditation effects, and different meditation techniques. Documents focusing only on spiritual or religious aspects without discussing benefits should be considered less relevant.
   - **Relevant Judgements – Relevance Score:**
     i. Meditation - 4
     ii. Stress management - 2
     iii. Transcendental Meditation - 2
     iv. Buddhism - 1
     v. Psychology - 0

6. **Renewable energy sources overview**
   - **Narrative:** Relevant documents should contain information about various renewable energy sources. Documents that do not mention any renewable energy sources like solar, wind, hydroelectric, geothermal, or biomass energy should be considered irrelevant.
   - **Relevant Judgements – Relevance Score:**
     i. Renewable energy - 4
     ii. Solar energy - 3
     iii. Wind power - 3
     iv. Hydroelectricity - 3
     v. Fossil fuel - 1
     vi. Energy conservation - 1
     vii. Energy - 0

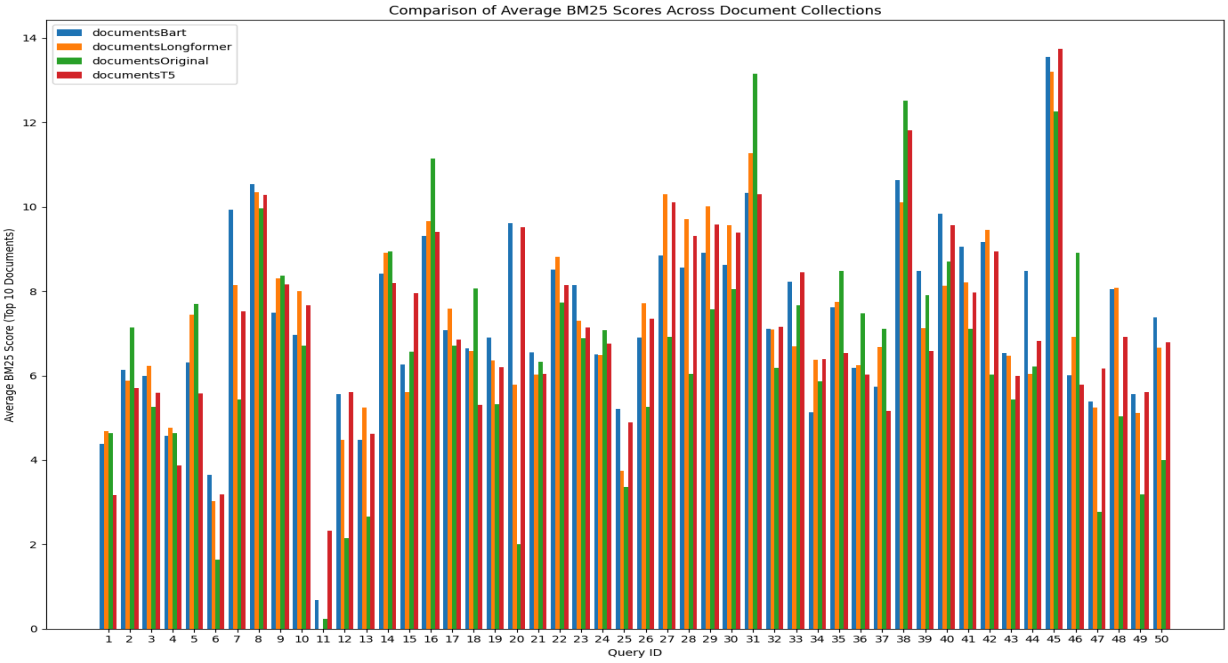
Results:

1. BM25 -

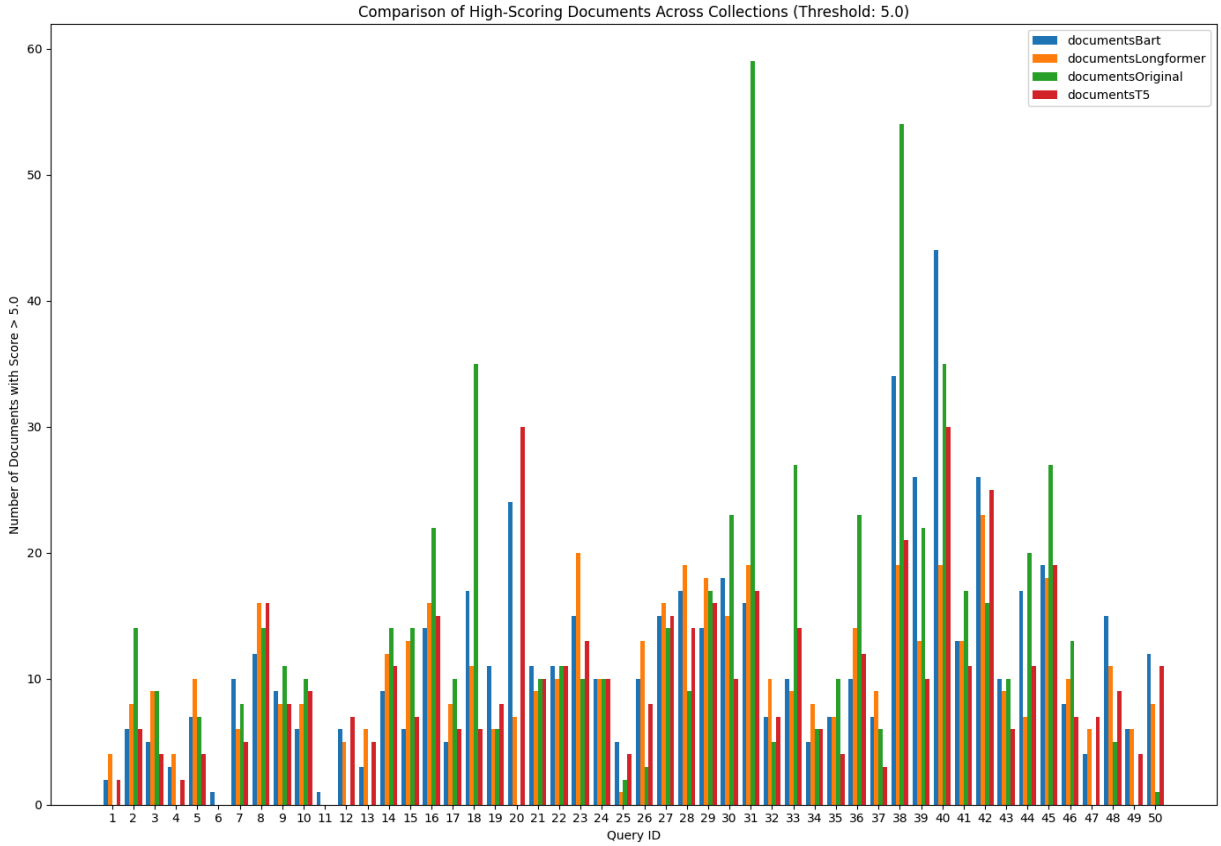
Model Comparison:



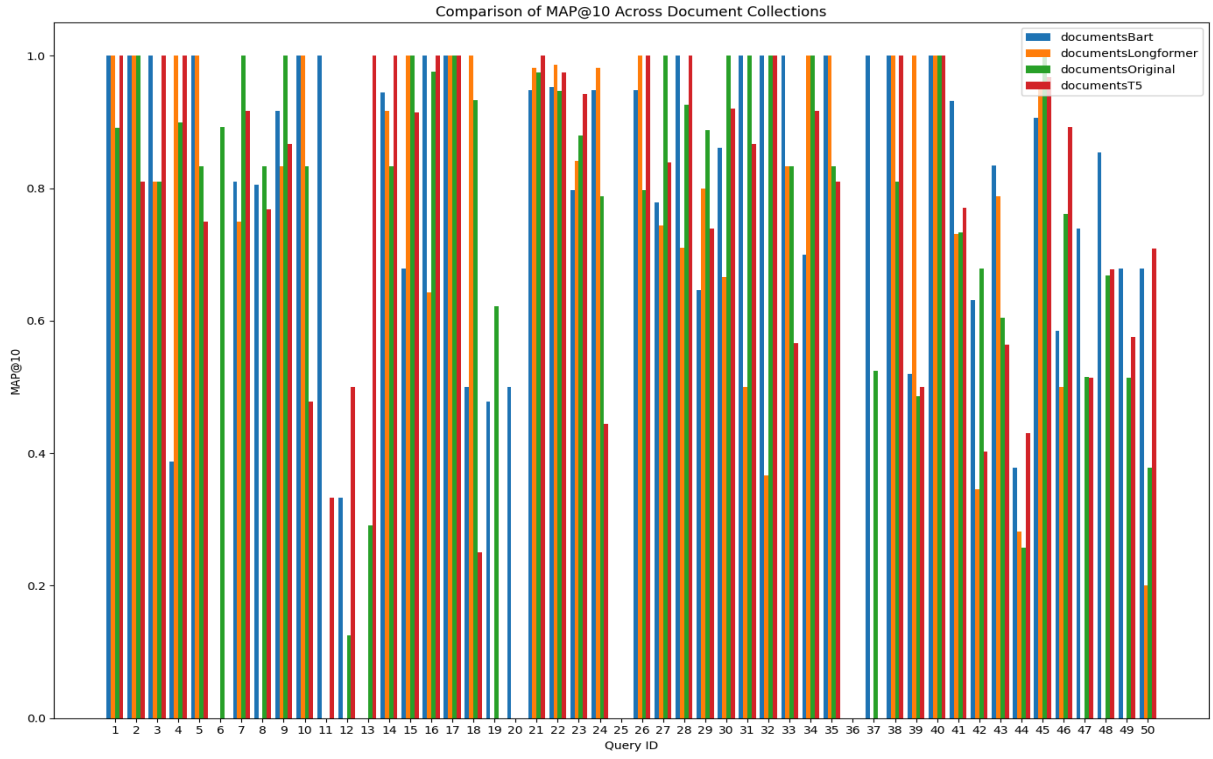
Avg Scores:



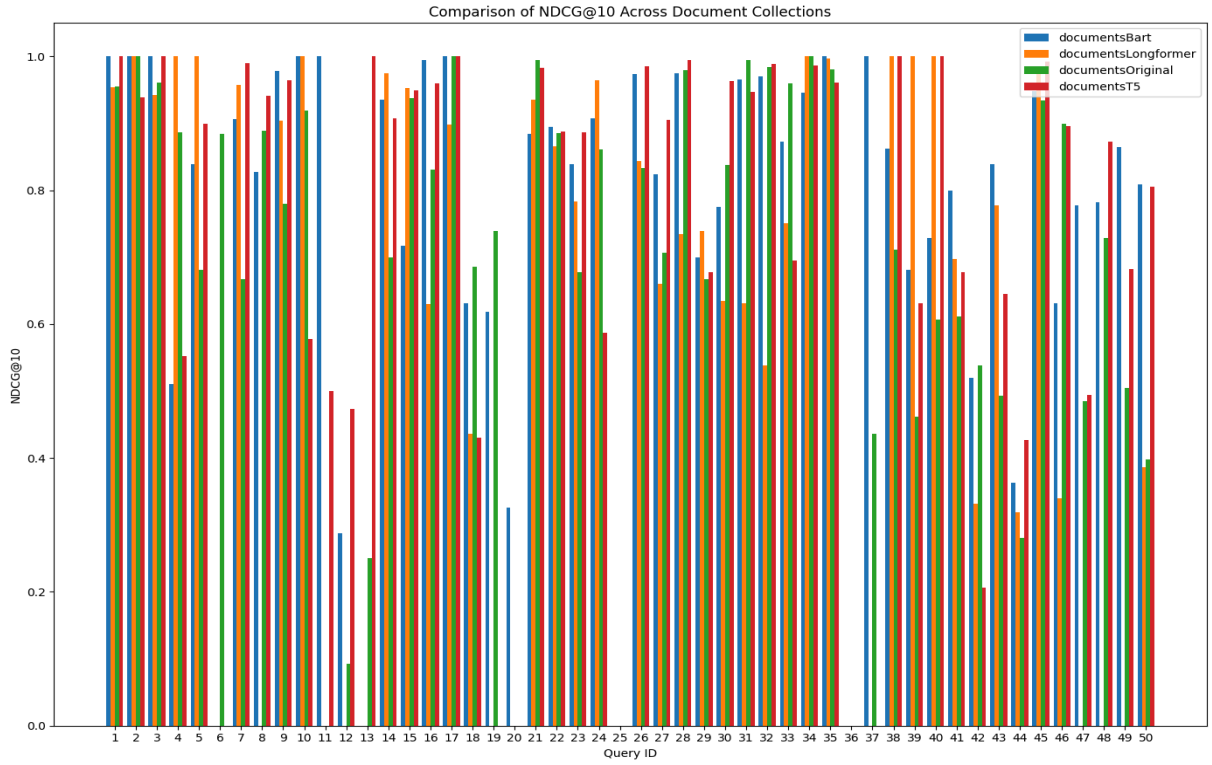
High Scoring Docs:



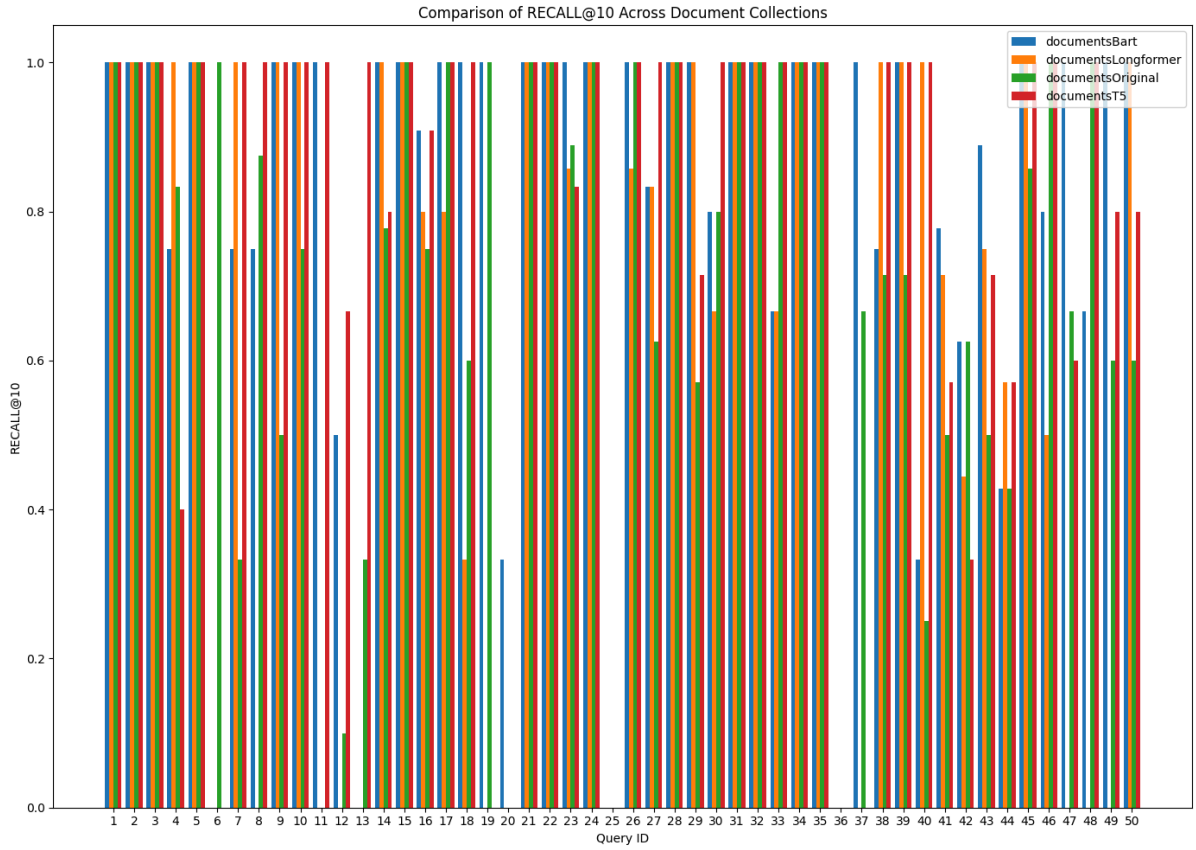
MAP@10:



NDCG@10:

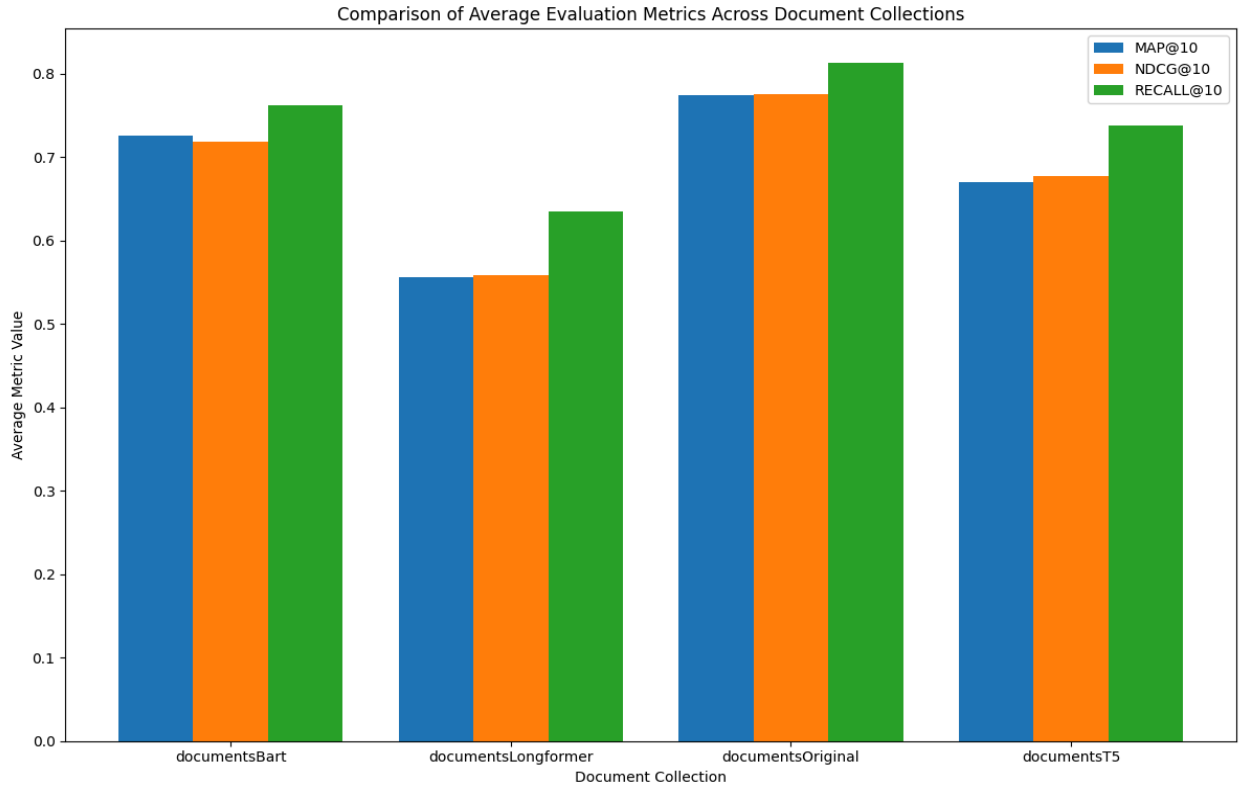


Recall@10:

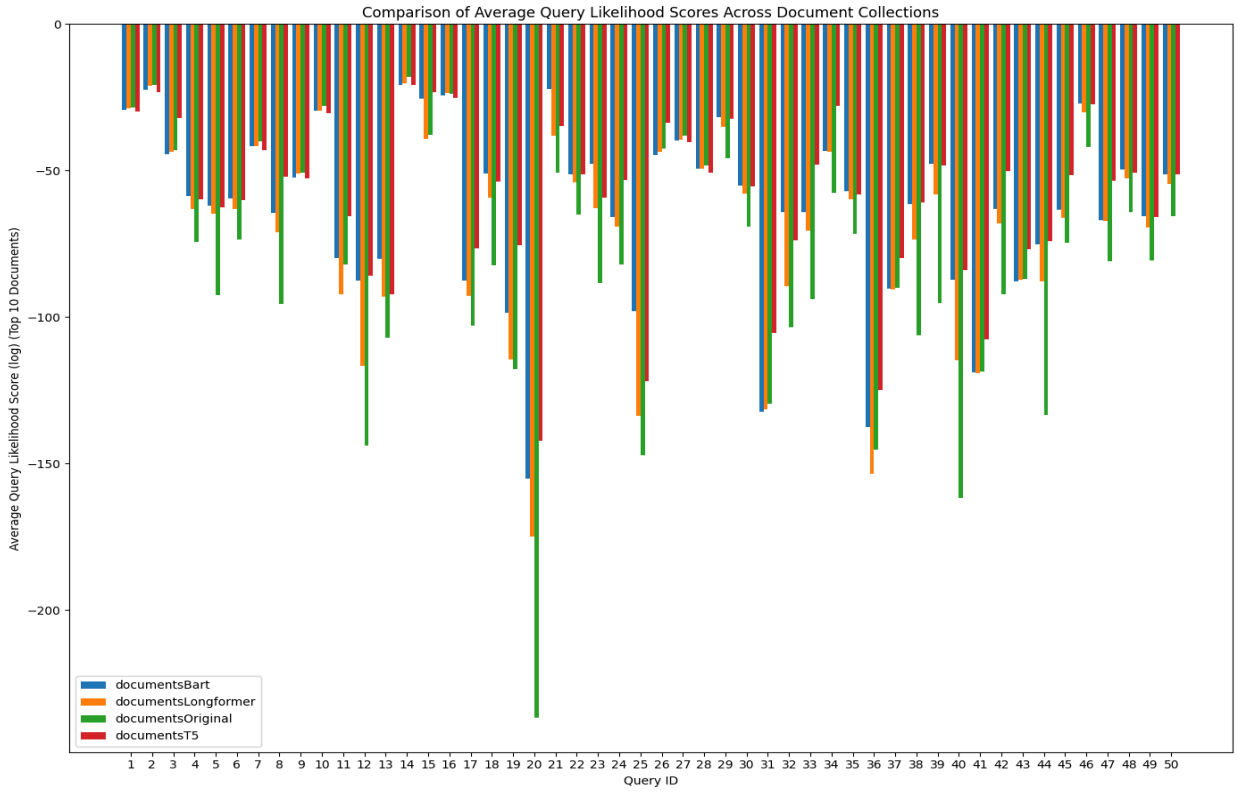


2. Query Likelihood

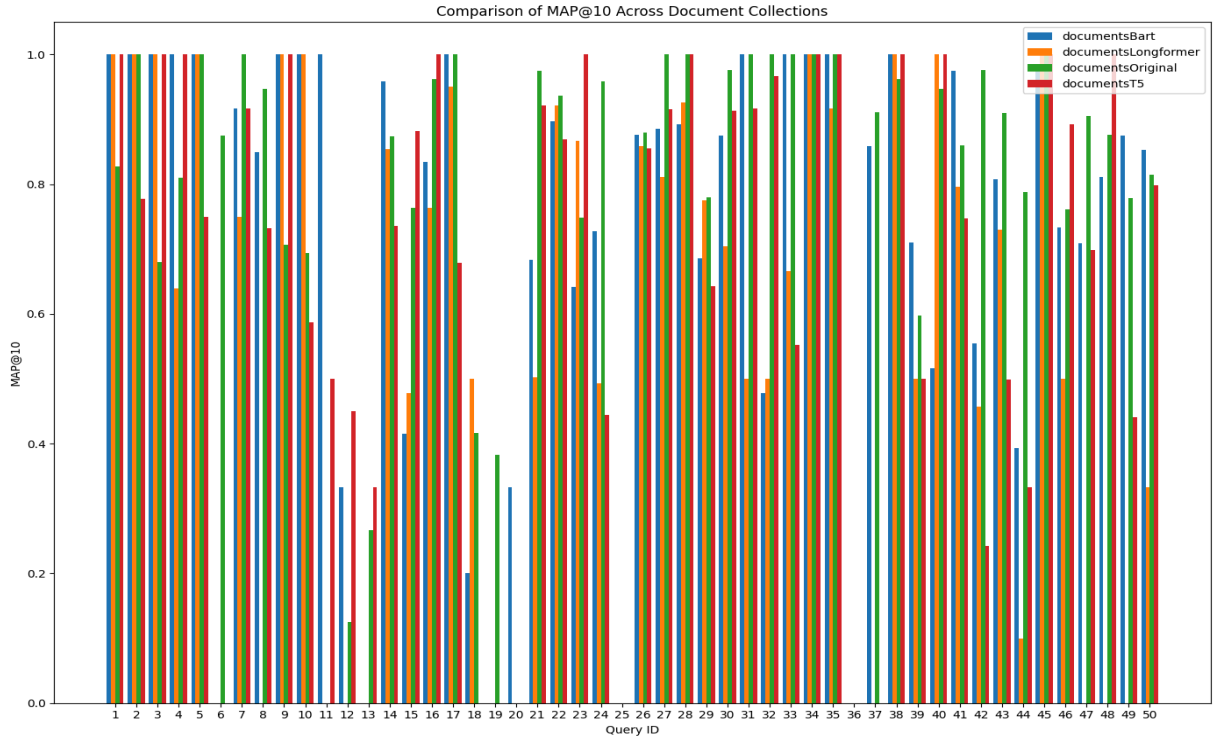
Model Comparison:



Avg Scores:

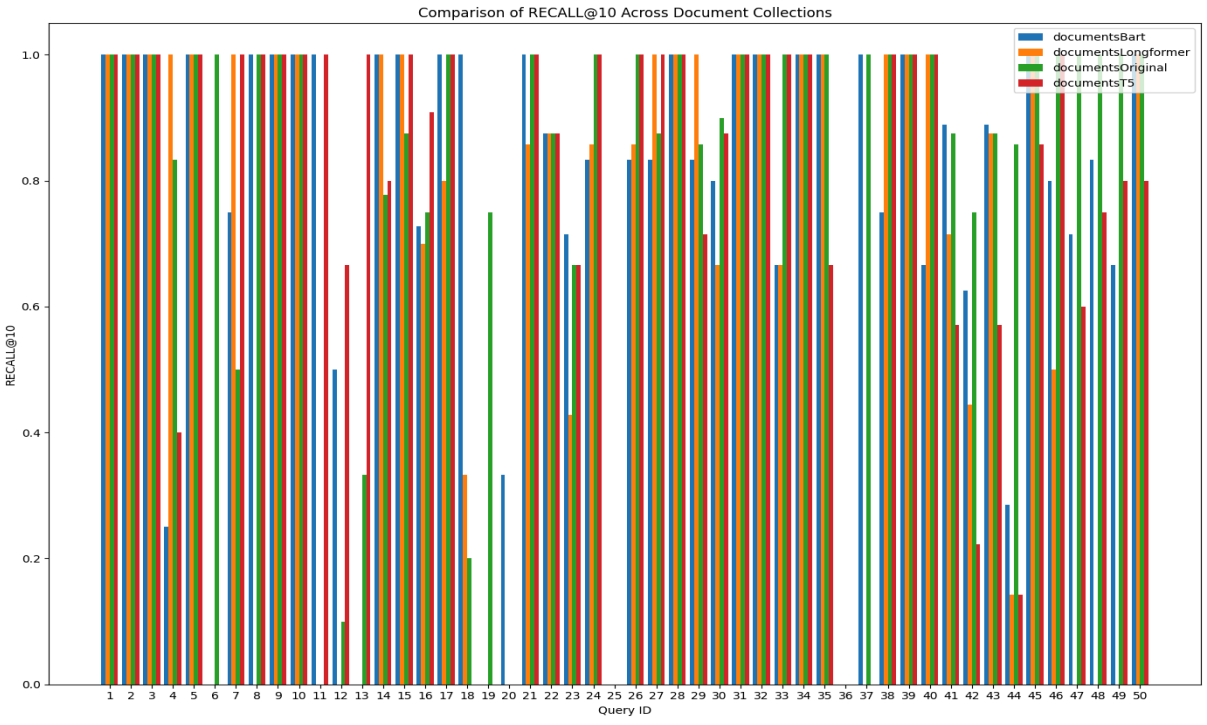


MAP@10:



NDCG@10:

Recall@10:



References:

1. Search Engines: InformaZon Retrieval in PracZce, Crot, Metzler and Strohman (Addison-Wesley)
2. Pegasus: https://huggingface.co/docs/transformers/en/model_doc/pegasus
3. Bart: https://huggingface.co/transformers/v2.11.0/model_doc/bart.html
4. T5: https://huggingface.co/docs/transformers/en/model_doc/t5
5. LongFormer: https://huggingface.co/docs/transformers/en/model_doc/longformer