

Assignment 7: Corpus Q&A Tool

Abstract

The central theme of this assignment is information retrieval, focusing on responding to queries within a specified corpus. While Large Language Models (LLMs) such as ChatGPT are highly capable, they face limitations in processing an entire corpus due to various constraints. Consequently, this project explores methodologies for selective data input and algorithm refinement to improve the efficiency and accuracy of LLM-based information retrieval, despite these inherent limitations.

Brief on Information Retrieval

Information retrieval involves the systematic process of obtaining information from a large pool of data or a corpus in response to a user query. The goal is to retrieve the most relevant and useful information given a specific request.

Within this field, there's a distinction between sparse and dense retrieval methods. Sparse retrieval focuses on traditional keyword-based systems, where documents are indexed and matched based on specific terms or phrases present in the query. It operates on the notion of exact term matching, ranking documents according to their relevance to the search terms.

On the other hand, dense retrieval techniques, often associated with neural language models, embrace a more nuanced approach. These methods leverage dense representations that encapsulate the semantic meaning of words and phrases in a continuous vector space. Rather than relying solely on exact keyword matches, dense retrieval systems utilize embeddings that capture the contextual and semantic relationships between words, allowing for a deeper understanding of the query's intent and document content.

Sparse Retrieval:

Pros:

1. **Simplicity:** It's straightforward and efficient in matching documents based on exact keyword or term matches.
2. **Interpretability:** Results are relatively interpretable as they are based on explicit term matches.
3. **Scalability:** Typically, sparse retrieval systems can handle large-scale indexing and search tasks.

Cons:

1. **Limited Context Understanding:** It struggles to capture the context or nuances of language beyond the specific keywords used, potentially missing relevant documents that don't precisely match the query terms.
2. **Vocabulary Mismatch:** When a user query uses different terms or phrases from those in the documents, relevant content might not be retrieved.
3. **Insensitive to Synonyms:** Sparse retrieval doesn't naturally account for synonyms or related terms, leading to potential information loss.

Dense Retrieval:

Pros:

1. **Semantic Understanding:** Dense retrieval methods have the ability to understand the context and semantics behind words and phrases, capturing more nuanced meanings and relationships.
2. **Robustness to Synonyms:** The use of embeddings allows for a better understanding of related terms, synonyms, and contextual meanings, reducing the impact of vocabulary mismatch.
3. **Improved Relevance:** By understanding the context and meaning, dense retrieval methods often retrieve more relevant documents.

Cons:

1. **Complexity:** The systems are generally more complex, requiring advanced models and computational resources, potentially impacting efficiency.
2. **Resource Intensiveness:** Training and utilizing these models can be resource-intensive, both in terms of computational power and data requirements.
3. **Interpretability Challenges:** The interpretation of results might be more complex, as dense models generate outputs based on complex mathematical representations rather than explicit term matches.

In this assignment we are working with somewhat of a hybrid where we retrieve one set of relevant information using a sparse method then feeding to a LLM (ChatGPT) that is trained with lot of infrastructure on its backend to answer a query from the information chosen by our algorithm.

Therefore, the subsequent report will detail the algorithm implemented for retrieving pertinent information. The chosen sparse method is the BM25+ algorithm, widely recognized as a standard in this domain.

On BM25+ Algorithm

BM25+ calculates the relevance score of a document to a query by considering the frequency of query terms in the document and the entire collection, along with document length normalization and term saturation. It's based on the idea that term frequency (TF) alone might not adequately represent relevance.

1. **Term Frequency (TF):** BM25+ adjusts the term frequency component by considering the frequency of the term within a document while also normalizing this frequency against the document's length. It mitigates the bias toward longer documents by employing a scaling factor.
2. **Inverse Document Frequency (IDF):** This component reflects the rarity of a term in the entire corpus. Terms that occur in fewer documents have higher IDF values, emphasizing their importance.
3. **Document Length Normalization:** To prevent longer documents from having an advantage due to higher term frequencies, BM25 normalizes the term frequency with a function that reduces the impact of longer documents.
4. **Term Saturation:** Unlike some other models, BM25 introduces a saturation term to prevent overemphasizing the importance of extremely frequent terms within a document.

Below is the formula for the BM25 score :

$$BM25(P, Q) = \sum_{i=1}^n IDF(q_i) \left(\frac{f(q_i, D)(k + 1)}{f(q_i, P) + k(1 - b + \frac{b|P|}{L})} + \delta \right)$$

$$IDF(q_i) = \ln \left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1 \right)$$

Where

- q_i is the i^{th} query term.
- n is the total number of terms in the query.
- $f(q_i, D)$ is the frequency of term q_i in paragraph P.
- $|D|$ is the length of paragraph P.
- L is the average document length in the corpus.
- k and b are tuning parameters that control term frequency saturation and document length normalization, respectively.
- δ is another tuning parameter introduced in BM25+ to prevent over normalization of the document length.
- $IDF(q_i)$ is the inverse document frequency term for q_i .
- N is the total number of paragraphs in the document
- $n(q_i)$ is the total number of paragraphs containing q_i .

The Inverse Document Frequency (IDF) is a measure used in information retrieval and text mining to determine the importance of a term within a collection of documents. It evaluates the uniqueness or rarity of a term across the entire corpus. The core concept behind IDF is to give more weight to terms that are less common across the corpus and less weight to terms that are more prevalent. The intuition is that terms occurring in fewer documents are often more discriminative and carry more significance in identifying relevant documents for a particular query.

Key points to understand about IDF:

1. **Logarithmic Scale:** The IDF is often computed on a logarithmic scale to prevent the value from becoming too large for terms that are infrequent. Logarithmic scaling ensures a smoother and more balanced weight distribution across different terms.
2. **High IDF for Rare Terms:** Terms that appear in a small number of documents will have a higher IDF value. For instance, if a term occurs in only a few documents within a large corpus, its IDF will be high, indicating its potential significance.
3. **Low IDF for Common Terms:** Conversely, terms that are widespread across the corpus will have a lower IDF. Common terms like "the," "and," or "is" typically appear in most documents and hence have a lower IDF, as they provide less discriminatory power in distinguishing relevant documents.

IDF plays a crucial role in conjunction with TF (Term Frequency) in many information retrieval algorithms, such as TF-IDF and BM25. It helps in ranking and scoring documents, effectively capturing the importance and discriminative power of terms within a corpus

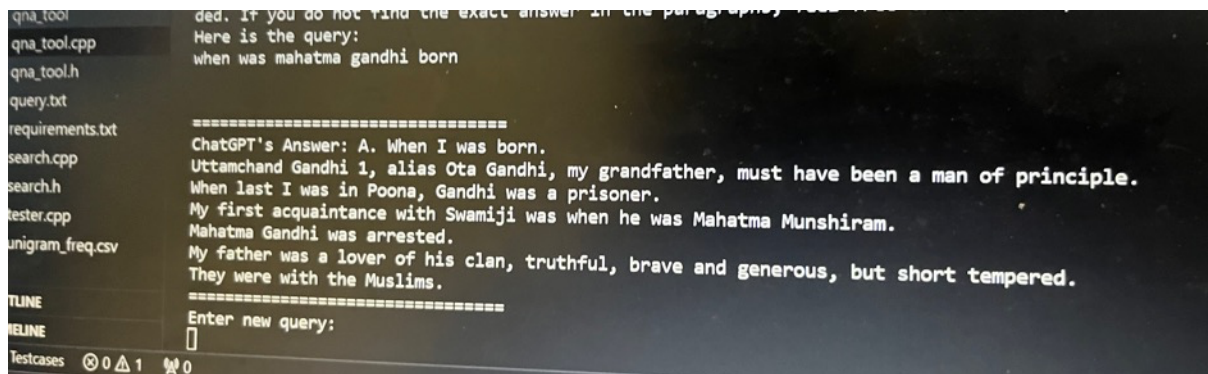
In our approach, we integrated data structures taught in the course, notably the heap, to enhance our Information Retrieval (IR) system. Specifically, we implemented a min-heap to efficiently identify the top 60 most relevant paragraphs. Our process involves initially populating the heap with the first 60 paragraphs. Subsequently, for each new paragraph, we compare its relevance score with the score of the top paragraph in the heap. If the new paragraph's score is higher, we replace the top element of the heap with this new paragraph. This procedure is repeated for all paragraphs, culminating in a selection of the top 60 relevant paragraphs.

To accommodate the peculiarities of our corpus, which includes very short paragraphs, we made a strategic modification. Our system allows the user to input the number of paragraphs (k) to consider during execution. From these, we prioritize longer paragraphs. Additionally, any paragraph with fewer than 15 words is automatically added to the context, not counting towards the k selected paragraphs. This adjustment ensures our IR system remains efficient while being sensitive to the varied lengths of paragraphs in our corpus

Our Prompt Engineering /Testing :

Below are some of the responses that we got while optimizing our BM25+ algorithm

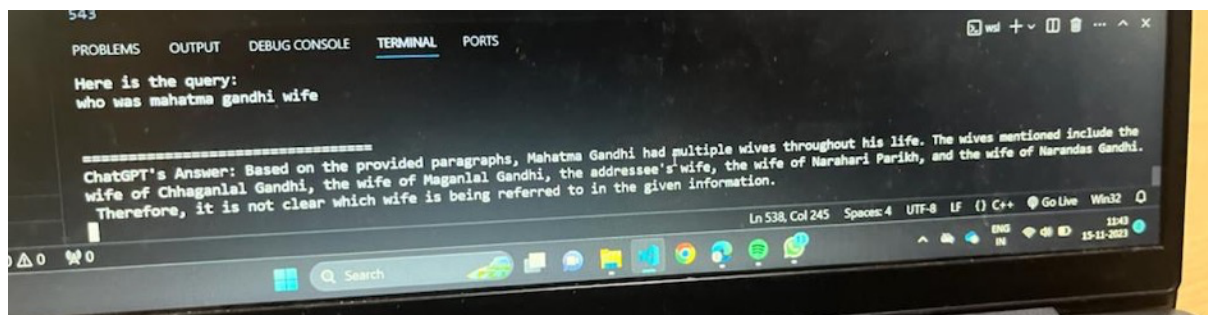
AI Hallucination due to incorrect context



```
qna_tool.cpp      Here is the query:
qna_tool.h        when was mahatma gandhi born
query.txt
requirements.txt
search.cpp
search.h
tester.cpp
unigram_freq.csv

=====
ChatGPT's Answer: A. When I was born.
Uttamchand Gandhi 1, alias Ota Gandhi, my grandfather, must have been a man of principle.
When last I was in Poona, Gandhi was a prisoner.
My first acquaintance with Swamiji was when he was Mahatma Munshiram.
Mahatma Gandhi was arrested.
My father was a lover of his clan, truthful, brave and generous, but short tempered.
They were with the Muslims.
=====
Enter new query:
[ ]
```

Multiple wives of Mahatma Gandhi



```
543
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

Here is the query:
who was mahatma gandhi wife

=====
ChatGPT's Answer: Based on the provided paragraphs, Mahatma Gandhi had multiple wives throughout his life. The wives mentioned include the
wife of Chhaganlal Gandhi, the wife of Maganlal Gandhi, the addressee's wife, the wife of Narahari Parikh, and the wife of Narandas Gandhi.
Therefore, it is not clear which wife is being referred to in the given information.

Ln 538, Col 245  Spaces: 4  UTF-8  LF  C++  Go Live  Win32
```

After numerous trials and iterations, we successfully optimized our algorithm to provide the most relevant context to ChatGPT.

Kasturba Gandhi as the wife of Mahatma Gandhi

```
root@Arpit:/mnt/d/OneDrive - IIT Delhi/academics/Sem 3/COL 106/Assignments/A7 Starter Code 2/A7 Starter Code# ./qna_tool.py
Inserting books
Enter query now:
who was mahatma gandhi wife

Enter 1 to see paragraphs else press anything:
=====
ChatGPT's Answer: From the given paragraphs, it is mentioned that Mahatma Gandhi had a wife named Kasturba Gandhi.
=====
Enter new query:

```

On Partition of India

```
Enter 1 to see paragraphs else press anything:
=====
ChatGPT's Answer: From the given paragraphs, it is mentioned that Mahatma Gandhi had a wife named Kasturba Gandhi.
=====
Enter new query:
what were mahatma gandhi's views on partition of India?

Enter 1 to see paragraphs else press anything:
=====
ChatGPT's Answer: A: Mahatma Gandhi strongly opposed the partition of India and considered it to be a sin. He believed that the Congress should not be party to partition and should instead demand unconditional departure of the British. Gandhi argued that if Pakistan were to be conceded, justice should also be done to non-Muslim majority areas and the Sikhs, but he disagreed with the partitioning of provinces based on religious lines. He believed that the people of both states should voluntarily wish to be one for there to be a reunion, and he considered partition to be incompatible with the greatness, safety, and economic development of the country.
=====

```

Mahatma Gandhi's Life in South Africa

```
Enter new query:
describe mahatma gandhi's life in south africa?

Enter 1 to see paragraphs else press anything:
=====
ChatGPT's Answer: Mahatma Gandhi's life in South Africa was marked by his active involvement in the struggle for the rights of Indians in the country. He firmly believed in the power of self-sacrifice and encouraged Indians in South Africa to have faith in themselves. Gandhi measures taken against Indians in South Africa as an assault on their property rights and their status as free men. They desired to be treated as equals to the European settlers in the country.

During his time in South Africa, Gandhi faced various challenges and injustices. One such instance was when the validity of Indian marriage was conducted according to Hindu, Muslim, and Zoroastrian rites was called into question. This led to Indian women being degraded from the status of wives to concubines and their children being deprived of their inheritance rights. This judgment deeply agitated the Indian community.

South Africa, being a continent of great size, presented different climates and landscapes. Gandhi highlighted that South Africa's climate was varied, with some regions being suitable for European settlers to live comfortably. He also mentioned the injustice of indentured immigration, a system that many Indians in South Africa opposed.

Gandhi considered his work in South Africa as crucial and emphasized the need for Indians in India to have realistic expectations about their achievements there. He urged not to overrate what was done in South Africa and emphasized the importance of focusing on the injustices faced by Indians in India itself.

Overall, Gandhi's life in South Africa revolved around his fight for the rights and equality of Indians, and his experiences there played a significant role in shaping his philosophy and methods of nonviolent resistance, known as Satyagraha.
=====

```

Another accurate response by our algorithm

```
Enter query now:
What were Gandhi's economic ideologies, and how did he advocate for self-reliance and sustainable development?
Enter least number of paragraphs to be considered:
17

Enter 1 to see paragraphs else press anything:
=====
ChatGPT's Answer: Gandhi's economic ideologies can be seen through his emphasis on self-reliance and his belief in sustainable development. He acknowledged that education was largely self-supporting in mediaeval times, but in today's democratic and socialistic society, the state must take on the responsibility of education. Gandhi believed that the only organized power with the resources to drive economic development was the state. He saw the shift of social power to the economic and political sphere and recognized the need for a modern, unitary, and territorial approach to economics.

Gandhi also advocated for self-reliance as a means to achieve economic independence. He encouraged individuals to boycott foreign goods and promote the use of locally produced goods. This concept of self-reliance extended to various aspects of life, including education, legal practice, and military administration. Gandhi believed in the importance of making efforts to become self-sufficient and not relying on external sources.

Furthermore, Gandhi's advocacy for sustainable development can be observed in his focus on the ideals of truth and dharma. He emphasized the need to work for the upliftment of the community and to address the urgent needs and grievances of the people. Gandhi recognized the interdependence of economic development and social well-being, and he urged individuals to work towards their goals with self-sacrifice, selfless work, and discipline.

Overall, Gandhi's economic ideologies centered around the principles of self-reliance, sustainable development, and the pursuit of truth and dharma. He believed that by empowering individuals and communities to take control of their economic destiny and promoting sustainable practices, true economic independence could be achieved.
=====

```

Final values of the free parameters that we chose are $b = 0.5$ $k_1 = 1.2$ and $\delta = 1$

Our Final Prompt to ChatGPT :

"You will be given a set of paragraphs, not necessarily in any logical order or semantically correct. These paragraphs are excerpts from Mahatma Gandhi's books and in decreasing priority of relevance. Now I will give you a query. Answer the query using the paragraphs I have provided. If you do not find the exact answer in the paragraphs, feel free to make assumptions.\n Here is the query: \n";

Additional Concepts Explored but Not Fully Realized:

1. **Query Expansion:** Implemented as a solution to the Vocabulary Mismatch Problem, query expansion broadens the scope of search queries by incorporating synonymous terms and related concepts. One effective approach is utilizing external resources such as WordNet or Knowledge Graphs. This integration ensures that queries with abbreviations or synonyms, like "IITD", are expanded to include more descriptive terms such as "Indian Institute of Technology Delhi." Such expansion is crucial for capturing the full spectrum of relevant documents, but it also introduces complexities in maintaining query intent and avoiding irrelevant results.
2. **Stemming** is a crucial process in Information Retrieval (IR) systems, where words are simplified to their root or stem form. This technique enables the IR system to recognize different forms of a word as essentially the same, for example, reducing "running" to "run." By converting words to their stems, stemming increases the likelihood of matching various forms of a word, thereby enhancing the retrieval of relevant documents. It effectively broadens search results by including variations of the stem word, which in turn increases recall—a significant metric in IR. However, stemming also has its challenges. Over-stemming, which is too aggressive, can lead to a loss of meaning, while under-stemming, being too conservative, might result in missing relevant matches. Despite these challenges, stemming is commonly used in many search engines and IR systems due to its simplicity and effectiveness in managing word variations. A notable example of a stemming algorithm is the Porter Stemming Algorithm.
3. **Vector Space Model:** Documents and queries are represented as vectors in a multidimensional space. Each dimension corresponds to a unique term in the document corpus. Document and query vectors are constructed by assigning weights to these terms, often based on term frequency and inverse document frequency. The similarity between a document and a query is then quantified by calculating the cosine of the angle between their respective vectors. This cosine similarity score determines the relevance of the document to the query. VSM's strength lies in its ability to measure semantic closeness, making it a vital tool for information retrieval tasks.

We also worked on our own vector state model algorithm.

Our Own VSM algorithm :

The main Idea is to utilise cosine similarity whose formula is :

$$\text{cosine}(A, B) = \frac{A \cdot B}{|A||B|}$$

Where cosine is \propto *similarity*

The algorithm, named HAAP, derives its title from the initial letters of our names ;)

- **Initial Step:** Pre-computation of vectors:
 - Consider a document with n relevant unique words (relevance determined by functions like IDF or TF).
 - Assume the document contains m paragraphs.
 - For each paragraph, construct a vector of size n.
 - Map unique words to vector indices, assigning a score or relevance (like frequency or a specific scoring function) at each index.
 - This process is replicated for all paragraphs.
- **Query Processing:**
 - Vectorize the query in a similar fashion (size n).
 - Employ cosine similarity to identify the top k most relevant paragraphs.
- **Performance Insights:**
 - In testing, this algorithm underperformed compared to our tweaked BM25 algorithm, requiring more computation and space.
 - Despite its current limitations, the HAAP algorithm has significant potential for independent tweaking at various stages (e.g., selecting n relevant words and corresponding scoring functions within a paragraph).
- **Conclusion:**
 - Due to time constraints in this assignment, optimization to surpass the BM25 algorithm's performance was not feasible. Hence, the tweaked BM25 algorithm was preferred over HAAP.