COL-106

ASSIGNMENT-7

Name:	Bhyri Siddharth Roy	Entry number:	2022CS11105
Name:	Ganji Srinath	Entry number:	2022CS11092
Name:	Pola Venkata Revanth	Entry number:	2022CS51650
Name:	Sarthak Panda	Entry number:	2022CS51217

INTRODUCTION:

- Data-Structures Utilized
- Brief explanation of Porter Stemmer
- Explanation of the usage of Porter Stemmer in our code
- Explanation of Scoring Strategy
- Limitations of our Scoring strategy
- Limitations of Part-1 that gave way to this Scoring Method
- Explanation of extra header file of c++ STL
- Test cases
- Scope of improvement and conclusion

Data-Structures Utilized

We used the Trie Data structure to maintain an Adjacency List Kind of structure. While inserting the book we inserted the word and simultaneously kept track of paragraph by keeping them in a vector at word's end Trienode. Further we kept the paragraphs as string in vector<vector<vector<string>>>. Where the first vector keeps track of using Book_id. The second one using page_id. The third one using paragraph_id. So, each paragraph can be accessed using notation Corpus [Book_id][Page_id][Paragraph_id]. A min heap was used for maintain current top -k paragraphs.

Brief explanation of Porter Stemmer

The **Porter Stemmer** is a well-known tool used in text processing to reduce words to their root form, or stem. This process, known as stemming, is crucial in many natural language processing tasks, including information retrieval and text classification.

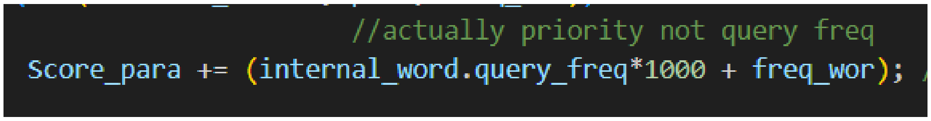
1. **Contextual Understanding**: The algorithm treats each word in isolation, without considering its context. This can lead to errors, as the same word can have different meanings in different contexts. For example, the word "apple" could refer to a fruit, a technology company, or a record label, depending on the context.
2. **Handling of Irregular Forms**: The algorithm is based on a set of predefined rules and does not handle irregular forms well. For example, it might not correctly stem irregular verbs or words with unusual plural forms.
3. **Over stemming and Under stemming**: The algorithm can sometimes be too aggressive (over stemming, where words are reduced too much and lose their original meaning) or too conservative (under stemming, where words are not reduced enough and different forms of the same word are treated as separate words).

Explanation of the usage of Porter Stemmer in our code

In our assignment, we utilized the Porter Stemmer to simplify our text data. By reducing words to their base form, this process enhanced the precision of our search results by treating different forms of the same word as the same. So, During Pre-Processing we stored the corpus words (in a Trie Data Structure) and corpus paragraph in the stemmed form. Further when the Query is given, we convert each word of it to root form and then search the corresponding words in the Trie to get adjacent Paragraph. Hence Porter stemmer just helps in pre processing the Corpus and Query.

Explanation of Scoring Strategy

In the second part of the assignment, a scoring strategy was implemented that prioritized relevance. This means that words that were more closely related to the search query were given a higher score. From the query given to us after stemming we found the frequency of each word in the query at this point we neglected common words that are usually used during paraphrasing the question (E.g. what,was,how…etc.). The one having the lowest frequency in the corpus was termed as the critical word and we only go to paragraphs that contain this critical word. Further there can be multiple critical word so beyond minimum frequency word if any other word in query has a frequency less than 1140(maximum number of words in any paragraph of MK-Gandhi corpus) is too considered a critical word. Now among these paragraph that contain the critical word the paragraph that had the occurrence of words of query were given more priority over frequency of individual word .i.e. more numbers of words appearing query present in paragraph higher was score rather than frequency of critical word. This is because if say query is "Mahatma's views on manner to eat the food" and a paragraph has "soldiers were eaten away" although eat was stemmed equivalent to eaten but this paragraph is about war and not food so occurrence of query words is priotized than frequency[achieved by multiplying 1000 again frequency can't be larger than 1140 for a paragraph).so the scoring formula is shown below.



```
                    //actually priority not query freq
Score_para += (internal_word.query_freq*1000 + freq_wor);
```

(Here query freq actually priority of word in query which we get by sorting them in increasing order of frequency of occurrence in paragraph and

priority of query word=total number of non-common words-index in sorted order)

Note: we created our own CSV file containing frequently occurring words. These words, are typically common words that do not carry much meaning and are therefore ignored in the scoring process. By excluding these words, we can focus on the words that truly matter, thereby improving the accuracy of our search results.

Limitations of our Scoring strategy

Although we gave priority to the occurrence of the word but still it the paragraph found may not be logically connected. Reconsider the example “Mahatma’s views on manner to eat the food” one of the returned paragraph using this strategy had “soldiers were eaten away ….. The people should not kill each other in this Manner” Although it has eat and Manner still irrelevant to the context. Further this strategy does not handle synonyms (E.g. which food mahatama likes to eat? Which food mahatma loves to eat?.

Limitations of Part-1 that gave way to this Scoring Method

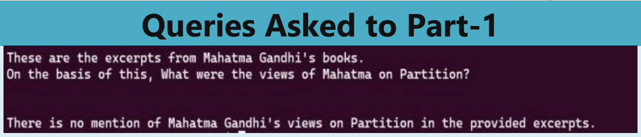
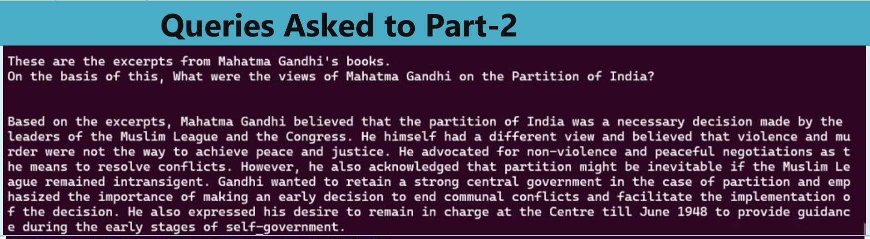
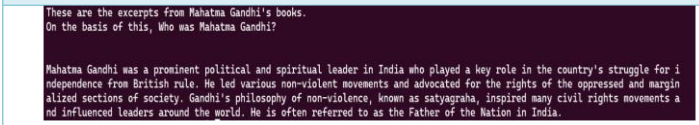
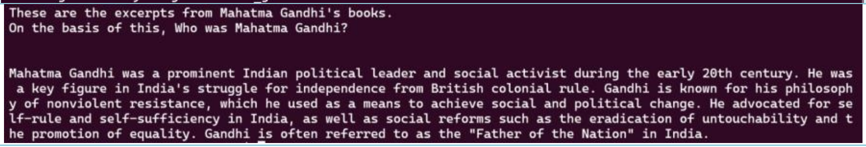
1. **Equal Word Weight**: All words in the query are treated equally, which may not reflect their actual importance in the context of the query.
2. **Lack of Semantic Understanding**: The algorithm doesn’t recognize synonyms, antonyms, or other semantic relationships between words, potentially missing relevant paragraphs.
3. **Frequency Over Relevance**: The algorithm prioritizes word frequency over relevance, which can lead to less relevant paragraphs being scored higher.
4. **Common Words**: The algorithm doesn’t filter out common words that usually don’t contribute much to the overall meaning.
5. **No Score Normalization**: The algorithm doesn’t normalize scores, potentially favoring longer paragraphs simply because they contain more words.

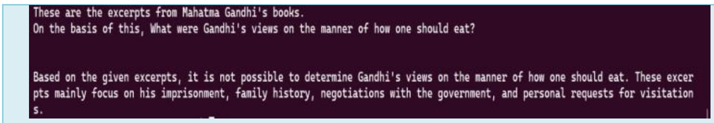
Explanation of extra header file of c++ STL

- <string.h> Since the Porter stemmer algorithm was outsourced from online resources the memset operation used in it required this library
- <algorithm> It was used because we needed to sort the words based on frequency

Test-cases

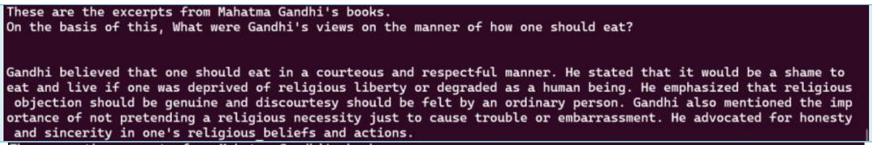
To check relevance of result we tested it against the queries in Grade scope and we can see the difference from part-1 algorithm directly being fed.

Queries Asked to Part-1	Queries Asked to Part-2
 These are the excerpts from Mahatma Gandhi's books. On the basis of this, What were the views of Mahatma on Partition? There is no mention of Mahatma Gandhi's views on Partition in the provided excerpts.	 These are the excerpts from Mahatma Gandhi's books. On the basis of this, What were the views of Mahatma Gandhi on the Partition of India? Based on the excerpts, Mahatma Gandhi believed that the partition of India was a necessary decision made by the leaders of the Muslim League and the Congress. He himself had a different view and believed that violence and mu rder were not the way to achieve peace and justice. He advocated for non-violence and peaceful negotiations as t he means to resolve conflicts. However, he also acknowledged that partition might be inevitable if the Muslim Le ague remained intransigent. Gandhi wanted to retain a strong central government in the case of partition and emp hasized the importance of making an early decision to end communal conflicts and facilitate the implementation o f the decision. He also expressed his desire to remain in charge at the Centre till June 1948 to provide guidanc e during the early stages of self-government.
 These are the excerpts from Mahatma Gandhi's books. On the basis of this, Who was Mahatma Gandhi? Mahatma Gandhi was a prominent political and spiritual leader in India who played a key role in the country's struggle for i ndependence from British rule. He led various non-violent movements and advocated for the rights of the oppressed and margin alized sections of society. Gandhi's philosophy of non-violence, known as satyagraha, inspired many civil rights movements a nd influenced leaders around the world. He is often referred to as the Father of the Nation in India.	 These are the excerpts from Mahatma Gandhi's books. On the basis of this, Who was Mahatma Gandhi? Mahatma Gandhi was a prominent Indian political leader and social activist during the early 20th century. He was a key figure in India's struggle for independence from British colonial rule. Gandhi is known for his philosoph y of nonviolent resistance, which he used as a means to achieve social and political change. He advocated for se lf-rule and self-sufficiency in India, as well as social reforms such as the eradication of untouchability and t he promotion of equality. Gandhi is often referred to as the "Father of the Nation" in India.



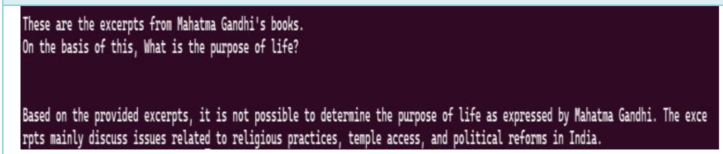
```
These are the excerpts from Mahatma Gandhi's books.
On the basis of this, What were Gandhi's views on the manner of how one should eat?


Based on the given excerpts, it is not possible to determine Gandhi's views on the manner of how one should eat. These excer
pts mainly focus on his imprisonment, family history, negotiations with the government, and personal requests for visitation
s.
```



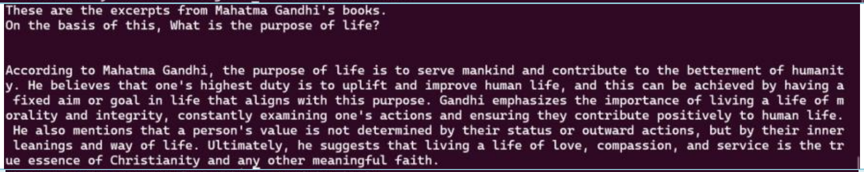
```
These are the excerpts from Mahatma Gandhi's books.
On the basis of this, What were Gandhi's views on the manner of how one should eat?


Gandhi believed that one should eat in a courteous and respectful manner. He stated that it would be a shame to
eat and live if one was deprived of religious liberty or degraded as a human being. He emphasized that religious
 objection should be genuine and discourtesy should be felt by an ordinary person. Gandhi also mentioned the imp
ortance of not pretending a religious necessity just to cause trouble or embarrassment. He advocated for honesty
 and sincerity in one's religious beliefs and actions.
```



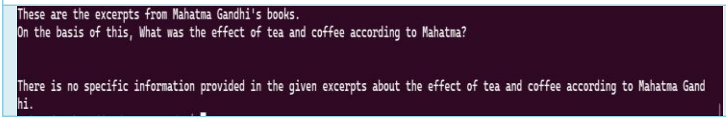
```
These are the excerpts from Mahatma Gandhi's books.
On the basis of this, What is the purpose of life?


Based on the provided excerpts, it is not possible to determine the purpose of life as expressed by Mahatma Gandhi. The exce
rpts mainly discuss issues related to religious practices, temple access, and political reforms in India.
```



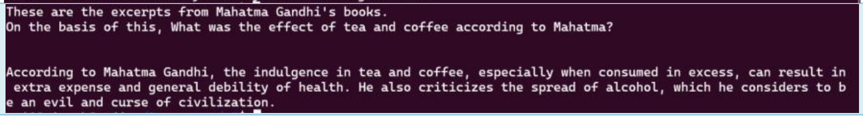
```
These are the excerpts from Mahatma Gandhi's books.
On the basis of this, What is the purpose of life?


According to Mahatma Gandhi, the purpose of life is to serve mankind and contribute to the betterment of humanit
y. He believes that one's highest duty is to uplift and improve human life, and this can be achieved by having a
 fixed aim or goal in life that aligns with this purpose. Gandhi emphasizes the importance of living a life of m
orality and integrity, constantly examining one's actions and ensuring they contribute positively to human life.
 He also mentions that a person's value is not determined by their status or outward actions, but by their inner
 leanings and way of life. Ultimately, he suggests that living a life of love, compassion, and service is the tr
ue essence of Christianity and any other meaningful faith.
```



```
These are the excerpts from Mahatma Gandhi's books.
On the basis of this, What was the effect of tea and coffee according to Mahatma?


There is no specific information provided in the given excerpts about the effect of tea and coffee according to Mahatma Gand
hi.
```



```
These are the excerpts from Mahatma Gandhi's books.
On the basis of this, What was the effect of tea and coffee according to Mahatma?


According to Mahatma Gandhi, the indulgence in tea and coffee, especially when consumed in excess, can result in
 extra expense and general debility of health. He also criticizes the spread of alcohol, which he considers to b
e an evil and curse of civilization.
```

Scope of Improvement and Conclusion

To handle synonyms and logical equivalence we word require either a predefined library or some sort of trainable models. In scope of DSA it was the best of our current thinking capability.