EPIQ Q/A REPORT

Passage answer retrieval using Tf-iDF,BERT,Doc2Vec,U.S.E

Lahari Sreeja - 12041570

Riyanshi Goyal - 12041240

Yedla Usha Sri - 12041780

Ankita Kumari - 12040220

INTRODUCTION

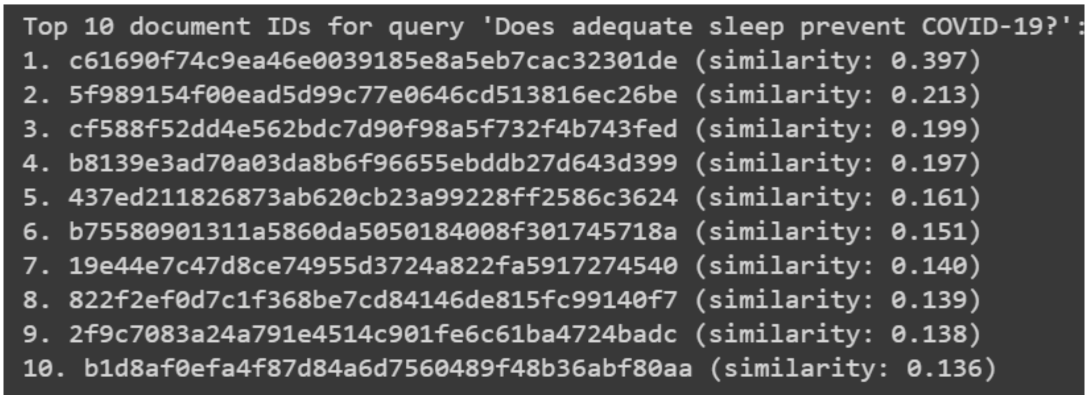
EPIC-QA is a project made to answer questions about COVID-19 in a passage.

There are two tasks: Task A provides expert-level answers to questions asked by researchers, scientists, or clinicians, while Task B provides simpler, consumer-friendly answers for the general public. The goal is to make reliable information more accessible.

The system uses several models, including TfidfVectorizer, BERT, Universal Sentence Encoder, and Doc2Vec, for ranking the documents and generating answers. These models are used to calculate the similarity between the query and the sentences in the documents. The top 5 most relevant sentences are then extracted and merged into a paragraph to form the passage answer.

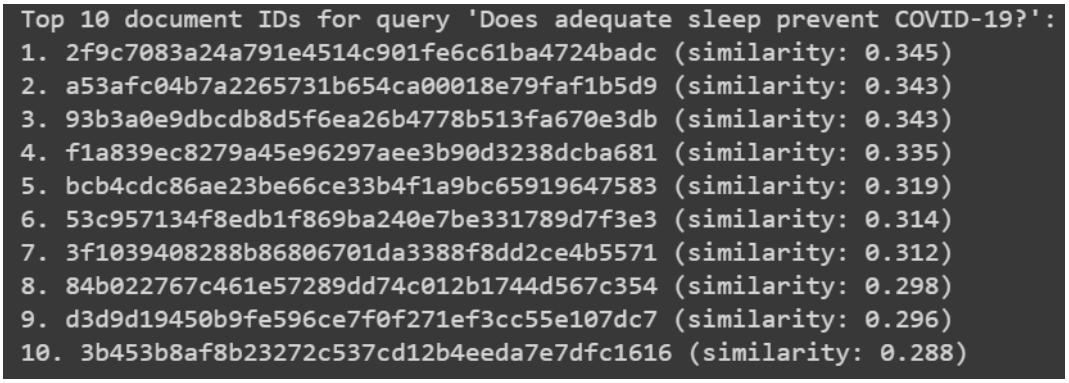
Performance and results of different models:

- ### Tf-iDF Vectorizer



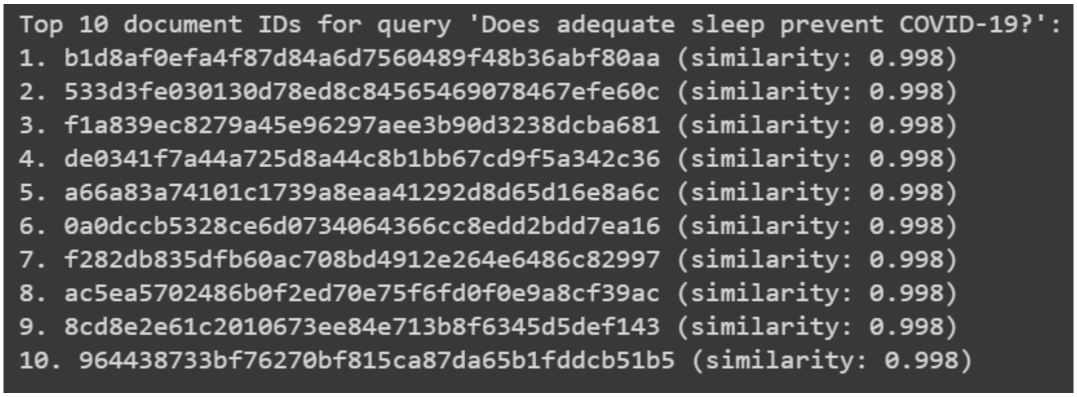
```
Top 10 document IDs for query 'Does adequate sleep prevent COVID-19?':
1. c61690f74c9ea46e0039185e8a5eb7cac32301de (similarity: 0.397)
2. 5f989154f00ead5d99c77e0646cd513816ec26be (similarity: 0.213)
3. cf588f52dd4e562bdc7d90f98a5f732f4b743fed (similarity: 0.199)
4. b8139e3ad70a03da8b6f96655ebddb27d643d399 (similarity: 0.197)
5. 437ed211826873ab620cb23a99228ff2586c3624 (similarity: 0.161)
6. b75580901311a5860da5050184008f301745718a (similarity: 0.151)
7. 19e44e7c47d8ce74955d3724a822fa5917274540 (similarity: 0.140)
8. 822f2ef0d7c1f368be7cd84146de815fc99140f7 (similarity: 0.139)
9. 2f9c7083a24a791e4514c901fe6c61ba4724badc (similarity: 0.138)
10. b1d8af0efa4f87d84a6d7560489f48b36abf80aa (similarity: 0.136)
```

- Universal Sentence Encoder



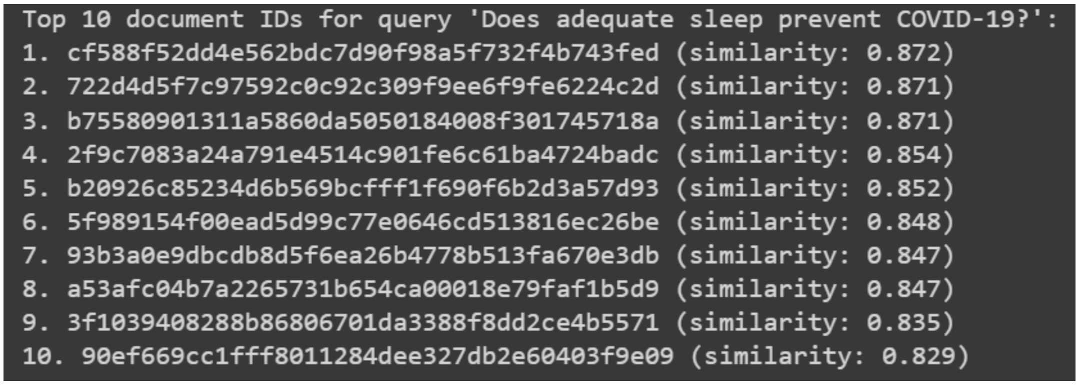
```
Top 10 document IDs for query 'Does adequate sleep prevent COVID-19?':
1. 2f9c7083a24a791e4514c901fe6c61ba4724badc (similarity: 0.345)
2. a53afc04b7a2265731b654ca00018e79faf1b5d9 (similarity: 0.343)
3. 93b3a0e9dbcdb8d5f6ea26b4778b513fa670e3db (similarity: 0.343)
4. f1a839ec8279a45e96297aee3b90d3238dcba681 (similarity: 0.335)
5. bcb4cdc86ae23be66ce33b4f1a9bc65919647583 (similarity: 0.319)
6. 53c957134f8edb1f869ba240e7be331789d7f3e3 (similarity: 0.314)
7. 3f1039408288b86806701da3388f8dd2ce4b5571 (similarity: 0.312)
8. 84b022767c461e57289dd74c012b1744d567c354 (similarity: 0.298)
9. d3d9d19450b9fe596ce7f0f271ef3cc55e107dc7 (similarity: 0.296)
10. 3b453b8af8b23272c537cd12b4eeda7e7dfc1616 (similarity: 0.288)
```

- Doc2Vec



```
Top 10 document IDs for query 'Does adequate sleep prevent COVID-19?':
1. b1d8af0efa4f87d84a6d7560489f48b36abf80aa (similarity: 0.998)
2. 533d3fe030130d78ed8c84565469078467efe60c (similarity: 0.998)
3. f1a839ec8279a45e96297aee3b90d3238dcba681 (similarity: 0.998)
4. de0341f7a44a725d8a44c8b1bb67cd9f5a342c36 (similarity: 0.998)
5. a66a83a74101c1739a8eaa41292d8d65d16e8a6c (similarity: 0.998)
6. 0a0dccb5328ce6d0734064366cc8edd2bdd7ea16 (similarity: 0.998)
7. f282db835dfb60ac708bd4912e264e6486c82997 (similarity: 0.998)
8. ac5ea5702486b0f2ed70e75f6fd0f0e9a8cf39ac (similarity: 0.998)
9. 8cd8e2e61c2010673ee84e713b8f6345d5def143 (similarity: 0.998)
10. 964438733bf76270bf815ca87da65b1fddcb51b5 (similarity: 0.998)
```

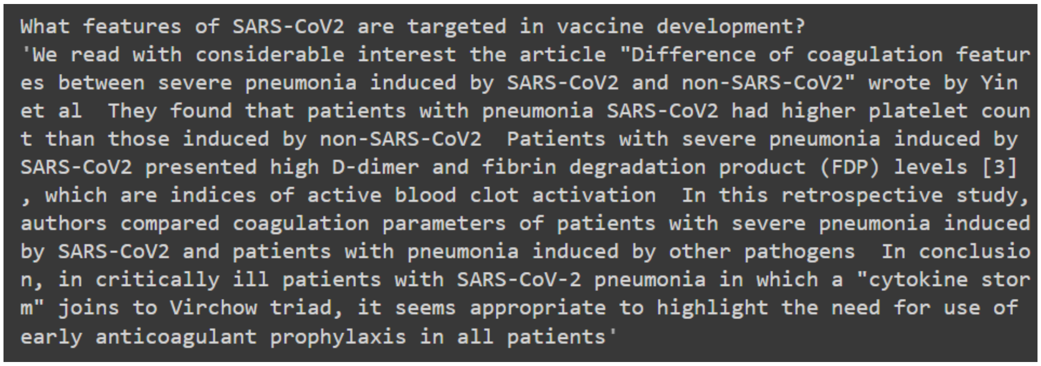
- BERT



```
Top 10 document IDs for query 'Does adequate sleep prevent COVID-19?':
1. cf588f52dd4e562bdc7d90f98a5f732f4b743fed (similarity: 0.872)
2. 722d4d5f7c97592c0c92c309f9ee6f9fe6224c2d (similarity: 0.871)
3. b75580901311a5860da5050184008f301745718a (similarity: 0.871)
4. 2f9c7083a24a791e4514c901fe6c61ba4724badc (similarity: 0.854)
5. b20926c85234d6b569bcfff1f690f6b2d3a57d93 (similarity: 0.852)
6. 5f989154f00ead5d99c77e0646cd513816ec26be (similarity: 0.848)
7. 93b3a0e9dbcdb8d5f6ea26b4778b513fa670e3db (similarity: 0.847)
8. a53afc04b7a2265731b654ca00018e79faf1b5d9 (similarity: 0.847)
9. 3f1039408288b86806701da3388f8dd2ce4b5571 (similarity: 0.835)
10. 90ef669cc1fff8011284dee327db2e60403f9e09 (similarity: 0.829)
```

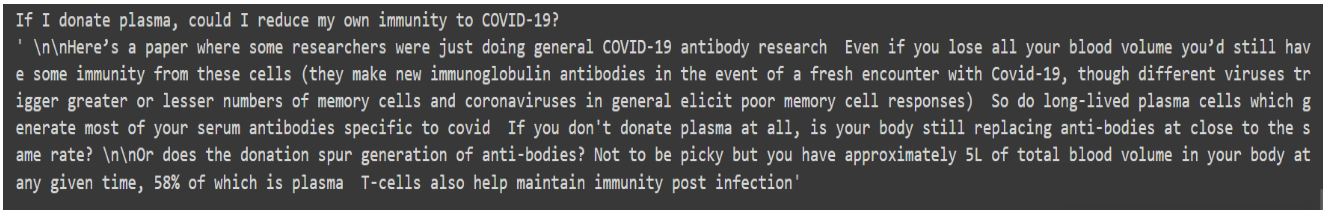
PASSAGE ANSWER RETRIEVED(Example):

- ### EXPERT QA



```
What features of SARS-CoV2 are targeted in vaccine development?
'We read with considerable interest the article "Difference of coagulation featur
es between severe pneumonia induced by SARS-CoV2 and non-SARS-CoV2" wrote by Yin
et al  They found that patients with pneumonia SARS-CoV2 had higher platelet coun
t than those induced by non-SARS-CoV2  Patients with severe pneumonia induced by
SARS-CoV2 presented high D-dimer and fibrin degradation product (FDP) levels [3]
, which are indices of active blood clot activation  In this retrospective study,
authors compared coagulation parameters of patients with severe pneumonia induced
by SARS-CoV2 and patients with pneumonia induced by other pathogens  In conclusio
n, in critically ill patients with SARS-CoV-2 pneumonia in which a "cytokine stor
m" joins to Virchow triad, it seems appropriate to highlight the need for use of
early anticoagulant prophylaxis in all patients'
```

- ### CONSUMER QA



```
If I donate plasma, could I reduce my own immunity to COVID-19?
' \n\nHere’s a paper where some researchers were just doing general COVID-19 antibody research  Even if you lose all your blood volume you’d still hav
e some immunity from these cells (they make new immunoglobulin antibodies in the event of a fresh encounter with Covid-19, though different viruses tr
igger greater or lesser numbers of memory cells and coronaviruses in general elicit poor memory cell responses)  So do long-lived plasma cells which g
enerate most of your serum antibodies specific to covid  If you don't donate plasma at all, is your body still replacing anti-bodies at close to the s
ame rate? \n\nOr does the donation spur generation of anti-bodies? Not to be picky but you have approximately 5L of total blood volume in your body at
any given time, 58% of which is plasma  T-cells also help maintain immunity post infection'
```

RESULTS

We can observe in the above images that the similarity between the queries and documents are retrieved for the top ten documents.Each model retrieves different documents having different similarities accordingly.

CONCLUSION

The project used different models like TfidfVectorizer, Universal Sentence Encoder, Doc2Vec, and BERT to retrieve relevant passages from the given data sets. The system was able to extract satisfactory answers from these passages based on the input queries. Both Consumer and Expert Data sets were explored to extract answers as required. Overall, the project was successful in using different models to find and extract relevant information.