TALK-ABOUT-COVID (TAC)

The A team

Optum Global Solutions

Table of Contents

The Team

Team Name: The A-Team

Organization: Optum Global Solutions

Team Members:

Name	:	Snigdha Borra
Designation	:	Sr. Data Scientist
_Linkedin	:	https://www.linkedin.com/in/snigdha-sree-borra-12244016/
Email	:	snigdha.borra@optum.com

Name	:	Shivani Aggarwal
Designation	:	Associate Data Scientist
_Linkedin	:	https://www.linkedin.com/in/shivaniaggarwal001/
Email	:	shivani_aggarwal@optum.com

Name	:	Dipali Agarwal
Designation	:	Data Scientist
Linkedin	:	https://www.linkedin.com/in/dipaliagarwal/
Email	:	agarwal.dipali@optum.com

Name	:	Vineela Swathi Karasala
Designation	:	Data Engineering Consultant
Linkedin	:	https://www.linkedin.com/in/swathi-vineela-karasala-5690351ab
Email	:	swathi.karasala@optum.com

Problem Statement

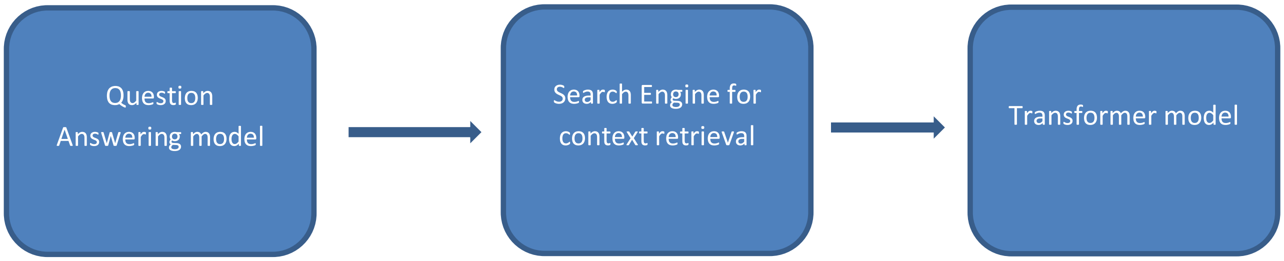
We realized that there is an enormous challenge in terms of finding the right kind of information quickly for the purpose of finding useful and actionable information on today's COVID-19 pandemic. As we progress in times of Covid19 spread, we want to help the research community **find answers** to their queries to **deeply understand coronavirus infectious disease** in the most effective way. A large amount data is available, and it is difficult for researchers to go through each paper every time a query comes up. Our solution aims at making this task easier for the researchers, scientists and also non-experts to improve their understanding of the current situation. We propose to develop a solution that uses research papers datastore on Covid-19 to answer those queries in the most relevant way.

Goal:

Develop a question-answering system able to answer (almost) any kind of question related to coronavirus using a pool of 50k+ research papers on Covid19.

Solution details

High level Architecture –



Question-Answering (QA) model

In machine learning, a question-answering model is composed of three sources: the question, the context and the answer. The model inputs are the question and the context and the model output is the answer. In most cases, but not all, the answer is contained in the context.

It exists many datasets used to train the QA model. One of the most popular is she Stanford Question Answering Dataset, also known as SQuAD. It contains thousands of tuples of the type (question, context, answer) used to teach the model what does it means to both **find** and **return** a question. During training, the model exploits and learn linguistical properties of the language.

Using a search engine to produce the context

In general, the context is quite limited, about one page. In our case, instead, we are dealing with more than 40k papers. **We need therefore to reduce the size of the context**. We do so by selecting all the papers that are most similar to the answer. In the code, a very simple algorithm, Okapi BM25, is used. Okapi BM25 is quite old (from 1980), but it does a great job. In future, I plan to compare the Okapi solution against other most recent approaches and solutions such as transformers.

From (context, question) to answer with transformers

Given a question q , the previous section gives us a list of context. Now, for each context and for the same query q, we ask to a pre-trained and pretty-powerful transformer model what is the part of the context that **better represent** the query.

The data obtained now, are dirty and hard to read. That's why for each task and for each question we visualize the context and the highlighted answer in a friendly way.

Data Set Details

We have used CORD-19 Research Database to train our model. It is a growing collection of 50k+ scientific papers relating to information on variants of coronavirus.

Data link - https://innovation.mit.edu/cord19/

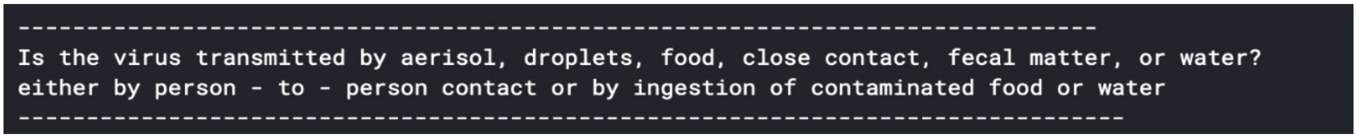
Model Building

Technologies used - Python, NLP, Transformers, IBM Watson Assistant

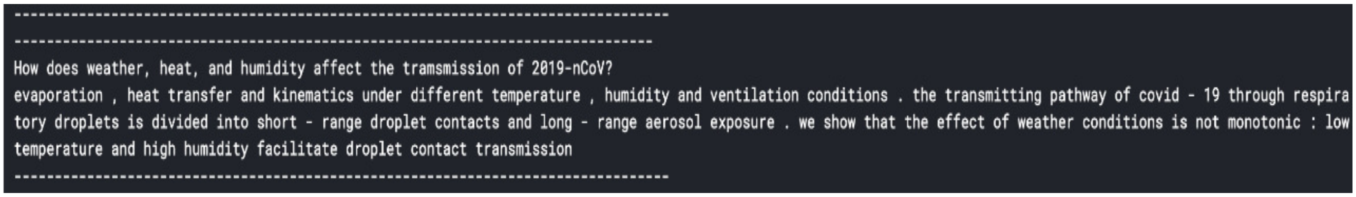
Github link - https://github.optum.com/saggar26/WiT_Hackathon_The_A_Team

Results

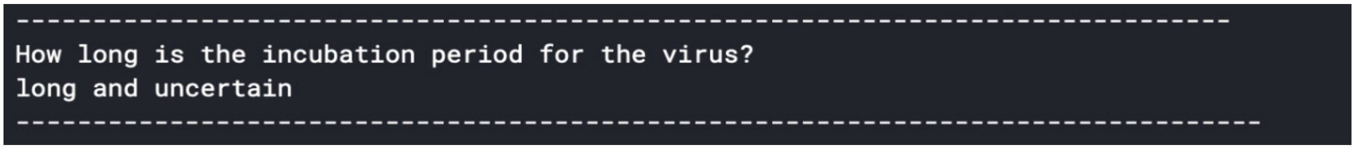
We came up with a list of questions and tested our model on them. Below are some of the results we managed. First lines are the questions and rest are the answers fetched from data source:



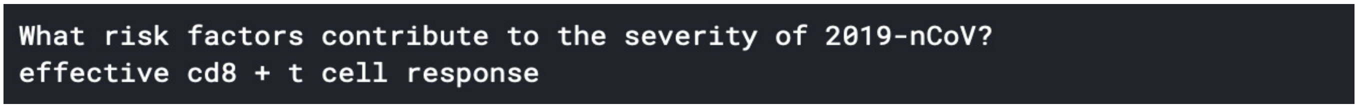
```
--------------------------------------------------------------------------------
Is the virus transmitted by aerisol, droplets, food, close contact, fecal matter, or water?
either by person - to - person contact or by ingestion of contaminated food or water
----------------------------------------------------------------------------------
```



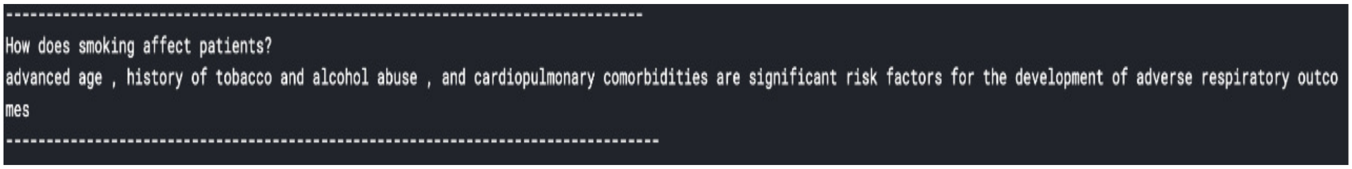
```
--------------------------------------------------------------------------------
------------------------------------------------------------------------------
How does weather, heat, and humidity affect the tramsmission of 2019-nCoV?
evaporation , heat transfer and kinematics under different temperature , humidity and ventilation conditions . the transmitting pathway of covid - 19 through respira
tory droplets is divided into short - range droplet contacts and long - range aerosol exposure . we show that the effect of weather conditions is not monotonic : low
temperature and high humidity facilitate droplet contact transmission
--------------------------------------------------------------------------------
```



```
--------------------------------------------------------------------------------
How long is the incubation period for the virus?
long and uncertain
----------------------------------------------------------------------------------
```



```
What risk factors contribute to the severity of 2019-nCoV?
effective cd8 + t cell response
```



```
------------------------------------------------------------------------------
How does smoking affect patients?
advanced age , history of tobacco and alcohol abuse , and cardiopulmonary comorbidities are significant risk factors for the development of adverse respiratory outco
mes
--------------------------------------------------------------------------------
```

Chat Bot Integration using Watson Assistant

We have integrated our results with Watson Assistant Chatbot using the starter kit provided by IBM. This would help provide more user-friendly interface to the research community as well as the common man to understand Covid19 and related aspects from more reliable data sources. The have significantly enhanced the Community Crisis Chatbot provided by IBM by adding more intents (trending questions from the Cord 19 dataset) and dialogues (answers to those trending questions from our NLP based model) to entertain a wider variety of topics.

Chatbot link:

https://web-chat.global.assistant.watson.cloud.ibm.com/preview.html?region=eu-gb&integrationID=8088981a-82a0-4a70-a786-72c35af3bbbd&serviceInstanceID=2b2c3786-50d4-4d49-9f71-8ead6a1ede2a

Future scope

Here are some of the things we're hoping to explore in the future alongside other experiments to better answer the user query:

- Find the articles that represent the "supporting knowledge" for a given article.
- Return a summarization abstract for each question for a detailed answer.
- Determine unique work given the context of all other works.
- Finding a network of authors contributing to the same domain.
- Attribute scientific rigor to certain articles, authors and apply that "trust" in the final aggregation step of the solution.
- Represent articles with their body text and identify similar content in other articles to help find parallel / adjacent / orthogonal work?
- Represent articles by the language used to cite them