COVID-19 QA:

Question Answering Using a Transformer Models

By Salim Haruna (Group 3)
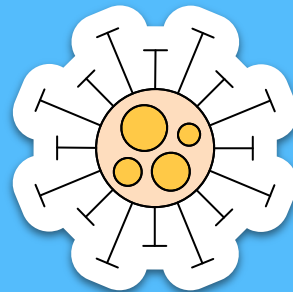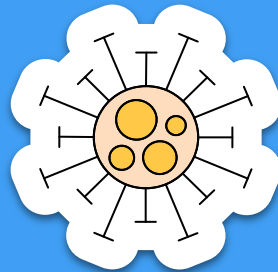
Table of contents

01
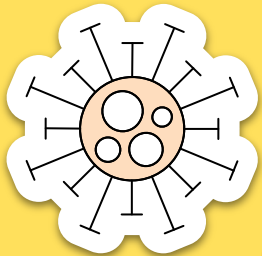# Introduction & Problem Statement

Introduction & Problem

Covid-19 is a coronavirus that was discovered in 2019 at Wuhan, China.

1. Overtime people have trust issues with public health which are based on either cultural or political sources they do not believe in.
2. I will use transformer models to create a question answering model and summarize the answers.
3. When presenting these answers, i will also show the citation of where these answers were retrieved from.
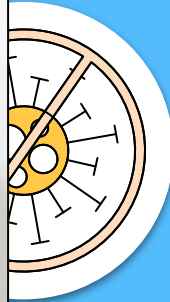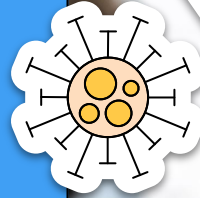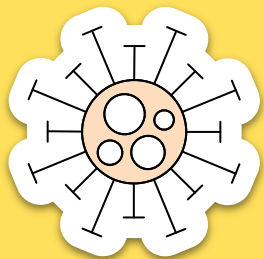
02
Dataset

Dataset Description

Source

The dataset was sourced via Kaggle and was put together by the white house and some of the leading research groups.
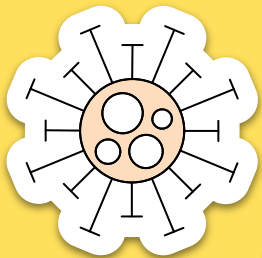
Size

The dataset is 16GB in disk size. This can be classified under big data in the context of text.

Content

The dataset contains csv and json files of over 134,000 scholarly articles, including over 60,000 full text, about COVID-19.

03
Text Preprocessing

Text Preprocessing

To generate the dataset to find the answers to the question we perform the following tasks:

1. Load metadata into pandas dataframe and select only useful columns.
2. Create a new column for the body text and run an iterator.
3. While runing the iterator, get the URL for each row, use beautiful soup to escape the html text and append the final body text to the body text column.

When a user asks a question, there are two types of inputs, the text input and speech input. To preprocess the input, we perform the following tasks:

1. If the user selects the text input, the text is simply passed to the model.
2. If the user selects the speech input, the program uses the sound device package to activate the device default microphone and collects the speech with 10 seconds timeout.
3. The speech is written to memory as audio chunks which is now converted to a WAV file. The WAV file is passed to a google speech-to-text translator and the speech is generated.
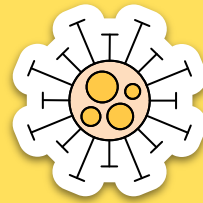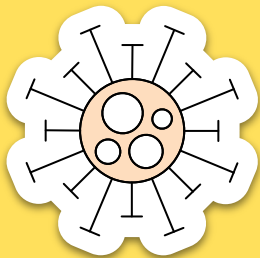
04

Transformer Model (Bert Cased)
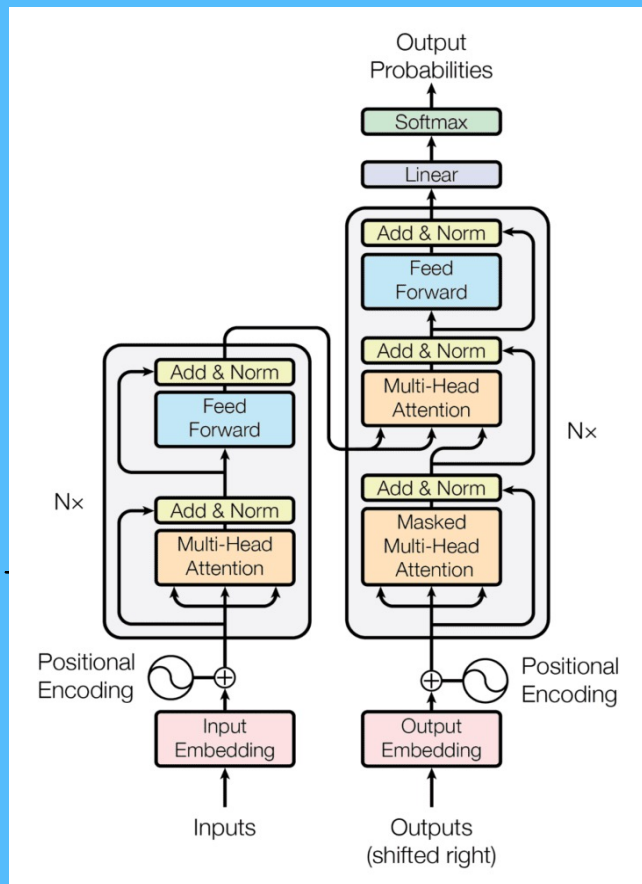
Transformer Model

Definition

The transformer model is a neural network that learns context and meaning by tracking relationships in sequential data such as the words in sentences. These models apply self-attention mathematical techniques, to detect subtle ways even distant data elements in a series influence and depend on each other.

Encoder

The task of the encoder is to map an input sequence to a sequence of continuous representations, which is then fed into a decoder.

Decoder

The task of the decoder is to receive the output of the encoder together with the decoder output at the previous time step, to generate an output sequence.
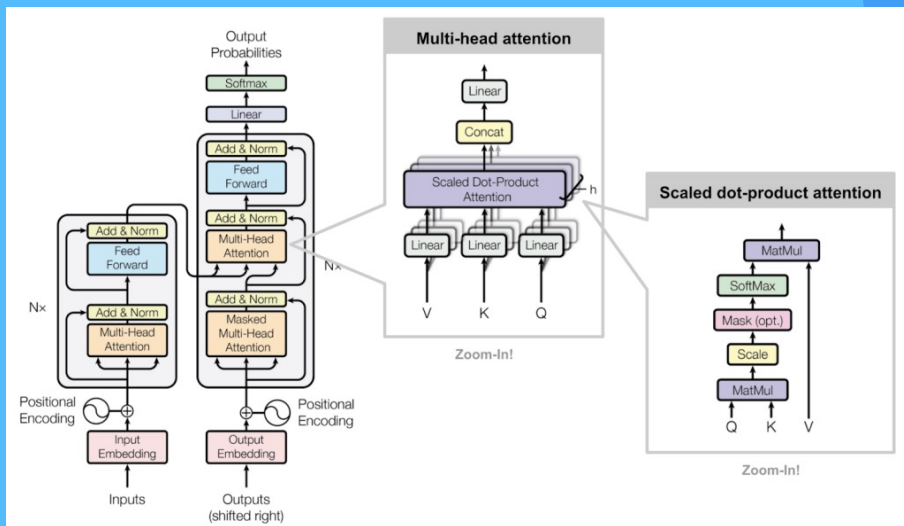
Bert Cased Model

Definition

BERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. It makes use of Transformer, an attention mechanism that learns contextual relations between words in a text. It is considered bidirectional.

Masked Language Modeling (MLM)

The model takes a sentence and randomly masks 15% of the words in the input then run the entire masked sentence through the model and must predict the masked words.

Next Sentence Prediction (NSP)

The models concatenate two masked sentences as inputs during pretraining. The model then has to predict if the two sentences were following each other or not.
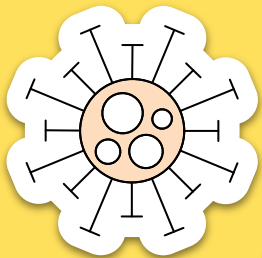
05
Experimental Setup & Results

Experimental Setup

Load Components

The metadata, tokenizer, summarizer and bert transformer model are loaded

TF-IDF

Retrieve top n documents using TF-IDF cosine similarity score using the title, abstract and body text
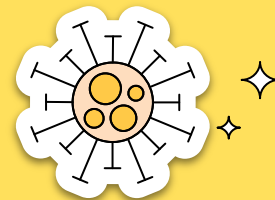
Inputs & Modeling

Texts are tokenized, encoded with special tokens and padding, and tensor inputs are generated, batched and passed to the model in chunks.

Summarization

Start and end logits are returned based on the argmax score . The answer text is now summarized.
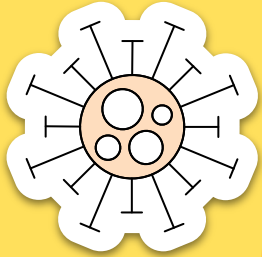
Results

When the program was run, the following question was asked:

What are the treatments for Covid-19?

Title	Source	Authors	URL	Publish Time	DOI	Start Score	End Score	Summarized Answer
Repurposing Therapeutics for COVID-19: Rapid Prediction of Commercially available drugs through Machine Learning and Docking	medrxiv	Sovesh Mahapatra; Prathul Nath; Manisha Chatterjee; Neeladrisingha Das; Deepjyoti Kalita; Partha Roy; Soumitra Satapathi	https://doi.org/10.1101/2020.04.05.20054254	2020-04-07	10.1101/2020.04.05.20054254	303	349	Coronaviruses are classified into four classes designated as alpha, beta, gamma, and delta. 14 SARS - CoV - 2 spikes also bind to receptors on the human cell surface called angiotensin.
Chloroquine and hydroxychloroquine in the treatment of COVID-19 with or without diabetes: A systematic search and a narrative review with a special reference to India and other developing countries	Elsevier	Singh, Awadhesh Kumar; Singh, Akriti; Shaikh, Altamash; Singh, Ritu; Misra, Anoop	https://doi.org/10.1016/j.dsx.2020.03.011	2020-06-30	10.1016/j.dsx.2020.03.011	98	355	People with diabetes and COVID - 19 may need special attention and clinical care. Reports gathered so far have suggested that a number of drugs could be potential candidates for the treatment. The clinical effectiveness of these drugs have not yet been fully evaluated. HCQ has been approved in the treatment of type 2 diabetes in India since 2014.
Associations between immune-suppressive and stimulating drugs and novel COVID-19—a systematic review of current evidence	PMC	Russell, Beth; Moss, Charlotte; George, Gincy; Santaolalla, Aida; Cope, Andrew; Papa, Sophie; Van Hemelrijck, Mieke	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7105343/	2020-03-27	10.3332/ecancer.2020.1022	98	134	The current review focuses on the effects of immune - suppressive or immune - stimulating drugs. Patients may be treated with immune-suppressive drugs or immune-stimulating drugs.
Genetic Profiles in Pharmacogenes Indicate Personalized Drug Therapy for COVID-19	medrxiv	Lei-Yun Wang; Jia-Jia Cui; Qian-Ying OuYang; Yan Zhan; Yi-Min Wang; Xiang-Yang Xu; Cheng-Xian Guo; JiYe Yin	https://doi.org/10.1101/2020.03.23.20041350	2020-03-30	10.1101/2020.03.23.20041350	95	282	The National Health Commission of China released at least seven versions of Guideline of Diagnosis and Treatment of Pneumonitis Caused by COVID - 19. Oxygen therapy, mechanical ventilation and drug therapy are recommended as major treatments, so that several drugs may be utilized simultaneously.
Identification of antiviral drug candidates against SARS-CoV-2 from FDA-approved drugs	biorxiv	Sangeun Jeon; Meehyun Ko; Jihye Lee; Inhee Choi; Soo Young Byun; Soonju Park; David Shum; Seungtaek Kim	https://doi.org/10.1101/2020.03.20.999730	2020-03-20	10.1101/2020.03.20.999730	81	240	Currently, there is no vaccines or therapeutics available for COVID - 19. Drug repositioning could be an effective strategy to respond immediately to emerging infectious diseases. Several drugs are being tested in numerous clinical trials.

06
Conclusion &
Citations

Conclusion

The use of transformer models has provided very great results efficiently. One important benefit of using the transformer model is due to the multi-head attention mechanism that gives the network the ability to pass through multiple words simultaneously.

Citations

- https://huggingface.co/bert-base-cased
- https://huggingface.co/docs/transformers/tasks/question_answering
- https://huggingface.co/docs/transformers/tasks/summarization
- https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformermodel/
- https://machinelearningmastery.com/the-transformer-model/
- https://www.gavi.org/vaccineswork/what-is-covid-19-and-how-does-it-spread?gclid=CjwKCAjw9e6SBhB2EiwA5myr9vipbbHCxlKWGmVhvNYuWIPNexzxpJGhBC2WGjOOhnwaWcVAaGiJjhoCUoMQAvD_BwE
- https://qa.fastforwardlabs.com/pytorch/hugging%20face/wikipedia/bert/transformers/2020/05/19/Getting_Started_with_QA.html

Thanks!

Do you have any questions?