Natural Language Processing Project Proposal

Salim Haruna

The problem I selected for this project is related to Covid-19. During the pandemic up till now people find information confusing and sometimes do not know what to believe. Therefore, I imagined having a question-and-answer model citing scientific research will go a long way to answer a lot of questions people have related to Covid-19 and prove the scientific basis for those answers.

The dataset that we will use is sourced via Kaggle and was put together by the white house and some of the leading research groups. The dataset is a resource of over 134,000 scholarly articles, including over 60,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses

I plan to use the pretrained question-and-answer transformer model to obtain answers from questions and use the pipeline summarizer to summarize the scientific answers for easy understanding. Furthermore, the questions cannot only be asked by passing text values but by also passing speech that is converted to text and passed to the model.

The framework used will be the Pytorch framework using the Bert pretrained Cased model from huggingface.

The reference materials we will use to obtain sufficient background on applying the QA transformer model to the covid-19 questions are the Natural Language Processing course materials and the Pytorch transformer documentation website.

We will judge the performance of our models making sure the models are able to provide answers to questions from the scientific articles based on the context of the question.

Since I am working alone, I will mostly work on the project any slight opportunity I have to.

Step	Target Date
Decide on dataset	2022-03-24
Text Preprocessing	2021-04-10
Build QA model	2021-04-11
Optimize and finalize QA model	2021-04-13
Draft of final report	2022-04-21
Finalize report	2022-04-28

Create and finalize presentation	2022-04-31