Text Summarization for COVID Documents

Team member(s): Jack Grossi

Can we create a summary of a complex document with the major points of the original document? For this project we will create a text summarization model that will return a summary of a complex medical document for the use by medical professionals. There are two main methods of text summarization: abstractive and extractive. Extractive reuses pieces from the original text to create a summary while abstractive writes an original summary and contains some understanding of the original document. For this task we will be creating an abstractive summary using Seq2Seq; additional work may include comparing Seq2Seq with a more cutting-edge model such as BART.

Our dataset will be the COVID-19 Open Research Dataset Challenge (CORD-19). This dataset was created in response to the COVID-19 pandemic by the White House and a coalition of leading research groups. It contains over 1,000,000 scholarly articles, with over 400,000 of those articles containing full text. The dataset is updated and maintained on a regular basis, adding and removing full text documents. A large amount of effort will primarily be directed at the data preprocessing steps for this dataset as it is incredibly large and requires adjustments before any analysis or modeling is done. As the included documents included in the dataset range from full scholarly papers to short abstracts, they must be sorted, and relevant documents selected.