CS 5830 - Final Project

Leo Beal, Thomas Brower, Chris Brown

Introduction

COVID-19 has drastically changed the lives of everyone on the planet within the past few months and has produced fear, panic, and misinformation. The medical and data science communities are working around the clock to learn more about this virus, to create possible treatments, and reduce the number of cases. In order to quickly save lives, professionals are publishing their work online as quickly as they can and while this can be helpful in getting information out there, it circumvents the tried and true practice of peer-review which can lead to false and redundant information. In an effort to assist in providing more accurate information, we propose a method of analysis that will capture the latent knowledge from prior coronavirus scientific literature. We will apply this knowledge to find missing links in COVID-19 literature. We hope to find similarities between the diseases' properties and find possible treatments that can be also used by medical practitioners to treat COVID-19.

Dataset

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19) [https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge]. This resource has over 33,000 scholarly articles with full text about coronaviruses. The dataset includes an article ID, source, title, abstract, full text, author list, and publish time.

Using this dataset, we plan on constructing an automatically annotated version that finds 10 biomedical entities of interest, namely Disorder, Species, Chemical or Drug, Gene or Protein, Enzyme, Anatomy, Biological Process, Molecular Function, Cellular Component, Pathway and microRNA. We will use the tool Neji for Named Entity Recognition and normalization, which is optimized for biomedical scientific articles.

Analyses

By the end of the first week, we will have some basic analyses done of the data to direct us on how to move forward with the project. By the end of the second week, we will have used the preliminary analyses to have created a few larger analyses that provide better results. The third week will be used for collecting the results and creating any graphs or charts that will be used in the report and presentation.

Chris will be dealing primarily with the creation of the transformer. A transformer will be used as it is a deep machine learning model used primarily in the field of natural language processing. Since their introduction, transformers have become the most state-of-the-art

architectures in NLP and we hope that their utilization will lead to great results. Leo will be cleaning the Kaggle dataset to be used in the transformer. To clean the dataset, excess data such as citations and metadata like authors will be removed. Sections of the papers that are split into multiple pieces would be brought together into one easy to parse piece. Thomas will assist in the cleaning of the data and perform the visualization of the results. Some of the visualizations will include word maps of treatments that are associated with specific terms or phrases while others will be of how the data was constructed or how a transformer works.

A potential problem that could be encountered during the course of this analysis would be difficulty in getting natural language processing running properly as transformers can become computationally intensive. Another problem could be that of all the data we have, very little will have good similarities between itself and COVID-19. If more data is needed or becomes available in the future, we will query publications from PubMed that relate to coronaviruses.

Expected Results

Any preliminary results - NOPE!

Resources:

https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
https://www.nature.com/articles/s41586-019-1335-8
https://hands-on-tech.github.io/2020/03/28/covid19-corpus.html
https://medium.com/huggingface/introducing-fastbert-a-simple-deep-learning-library-for-bert-models-89ff763ad384

Slides

https://docs.google.com/presentation/d/1iRYtwyKo3nDe-tK5luAL38Ig0HY5Yr9wRxfDc6gvfLs/edit?usp=sharing