Project Progress Report

Srujan
Netid:ssg7

November 29, 2020

1 Introduction to Project

I chose to participate in the Information Retrieval competition. I proposed using a Learning-to-Rank methodology for IR as opposed to standalone rankers. I also proposed to first explore evaluating using combination of rankers rather than one ranking methodology to produce rankings. Finally, I proposed using SVM^{map}, a supervised learning based ranking technique for IR.

2 Estimated steps in the project timeline

2.1 Step 1: Choosing metapy for a preliminary analysis

Status: Completed
On: 11/27/2020
I decided to use metapy to get an understanding for the data (CORD-19 set) and how classical rankers we used through the course behave with this enormous dataset. This means using the ranking template from MP2.4 to see how that implementation performs with the new dataset (includes train and test folders from here on) instead of cranfield data.

2.2 Step 2: Converting data into MeTA format

Status: Completed
On: 11/29/2020
Using metadata.csv and documents folder, the dataset needs to be created as a ".dat" file with documents separated by a new line. Queries should be extracted from the given xml document and converted into a text file that metapy.index.IREval can read. Relevance judgement file must have "doc_id"

as uint64 instead of the alphanumeric format given in the CORD-19 dataset. Finally, building an inverted index to be used by metapy rankers.

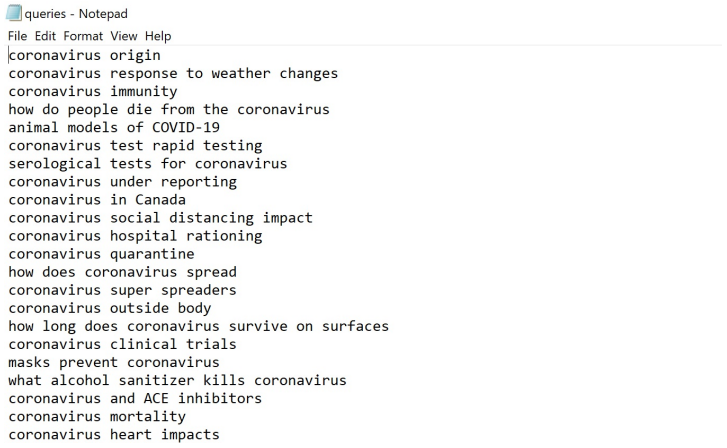

Figure 1: Extracted queries from xml

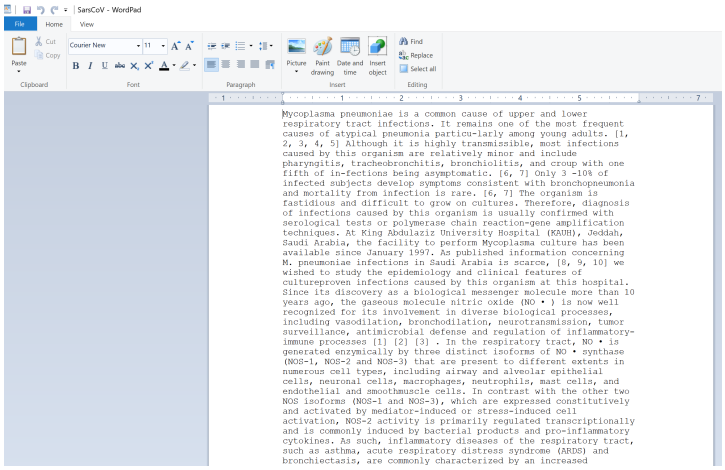

Figure 2: Extracted data in MeTA format

2.2.1 Step 3: Evaluating with known rankers

Status: Pending
Estimated Completion: 12/04/2020
After completing the data conversion, using classical and custom rankers

(BM25+, InL2 ranker etc.) to evaluate their performance with CORD-19 dataset.

2.2.2 Step 4: Evaluate with a combination of rankers

Status: Pending
Estimated Completion: 12/04/2020
Depending on the individual rankers' performance, choose a combination of top rankers based on the data at hand. Theoretically, this implementation changes the combination weights depending on the dataset.

2.3 Step 5: Convert data for SVM^{map} compatibility

Status: Pending
Estimated Completion: 12/09/2020

2.4 Step 6: Evaluate using SVM^{map}

Status: Pending
Estimated Completion: 12/09/2020

2.5 Step 7: Produce final results based on Steps 3, 4, and 6

Status: Pending
Estimated Completion: 12/12/2020

3 Challenges

Data format conversions are pretty challenging owing to lot of exception handling. The final results need to be in the competition format which means the data need to be converted back again. It's difficult to comprehend the embeddings data without going through SPECTER project implementation. So, the data will have to go unused if I rely entirely on introduction section for each paper. There are also a few missing data fields in the dataset collection which need to be handled. The implementation would still work if I ignore missing data but would be incomplete.