Exploring PUBMED data

Yoon Choi, Chris McGowan, Sunny Patel

# **Introduction**

PubMed data is an important piece of information that we can utilize in order to analyze the health data. Our objective in this paper is to determine whether we can gain insights from the speculative dataset and covid-19 dataset in regard to correctly determine the data as speculative or not, and also to extract various topics involved in the dataset. In this paper, we extracted relevant information for the health data analysis, and finally created charts and figures to analyze it.

# **Methods**

We were given two datasets for Speculative data and COVID-19 data, in TOK and CSV format, respectively. The Speculative dataset contain four files: speculative data for training and testing, and non-speculative data for training and testing. It contains rows of free-text information that is delimited by a dot and a new line. The COVID-19 dataset contains various information in free-text format such as fields in titles, abstracts, and conclusions. With the data in our hands, our group decided the most efficient method for parsing and analyzing the dataset was to use Python, and especially Jupyter notebook, as our programming language and write scripts that iterate through all the relevant information.

For analyzing the sentences in the testing speculative / non-speculative data, we first utilized the sentences in the respective training datasets. Training the dataset in the training file allowed us to sufficiently build a model to test on the testing dataset and come up with the best F-1 score as possible.

For analyzing the topics covered in the COVID-19 dataset, we utilized Jupyter Notebook in order to divide up the algorithms in places it takes a long time, i.e., building a LDA model. For identifying the topics covered (question B.1) we utilized LDA. We chose title and abstract because many researchers try to condense their big picture of the research paper into these two categories frequently, allowing us to sufficiently identify the topics covered with these two fields. In order to extract and analyze the drugs and diseases mentioned in the COVID-19 dataset, we utilized a popular open-source software library called spaCy. More specifically, we utilized two models: en\_ner\_bc5cdr\_md and en\_ner\_bionlp13cg\_md.

# 2.1 Question A: Determining a Speculative Sentence with a Classifier

# 2.2 Question B.1: Identifying Topics in COVID-19 Data

After we loaded the COVID-19 Dataset, we pre-processed the data by lemmatizing using NLTK wordnet and removing a list of stop words. We also utilized a spacy model “en\_core\_web\_sm”, which is a NLP model for English languages and more efficient than “en\_core\_web\_trf”. Then we iterated the CSV data by extracting the title and the abstract from each observation and saving it in the variable ‘text\_data’ as a list of list of words. We encapsulated the mapping between normalized words and their integer ids and saved it in the variable ‘dictionary’. We then transformed ‘text\_data’ into the bag-of-words format (a list of tuples with token\_id and token\_count as the first and second element of each tuple pair). Bag of Words method allows us to extract features from text documents, in this case a combination of titles and abstracts. Using these data, we trained the model on 5 to 10 number of topics in order to determine the best number of topics to subdivide the data into. We determinined the number of topics as 10 because of high coherence score and low perplexity score in the graph below. Coherence scores measure the relative distance between words within a topic and analyzing the COVID-19 model with 10 topics demonstrated the most promising high score. Perplexity score measures how surprised a model is of new data it has not seen before and 10 topics also showed the lowest score out of all trials. We got respective scores by applying LDA model from the genism library to save a model, then outputting perplexity and coherence scores on each of them.

|  |  |
| --- | --- |
| number of topics | coherence score |
| 5 | 0.5534 |
| 6 | 0.5466 |
| 7 | 0.5583 |
| 8 | 0.5767 |
| 9 | 0.5861 |
| 10 | 0.584 |

|  |  |
| --- | --- |
| number of topics | perplexity score |
| 5 | -8.4009 |
| 6 | -8.4365 |
| 7 | -8.5074 |
| 8 | -8.5939 |
| 9 | -8.7137 |
| 10 | -8.8793 |

# 2.3 Question B.2: Extracting Drugs and Diseases from the COVID-19 Data

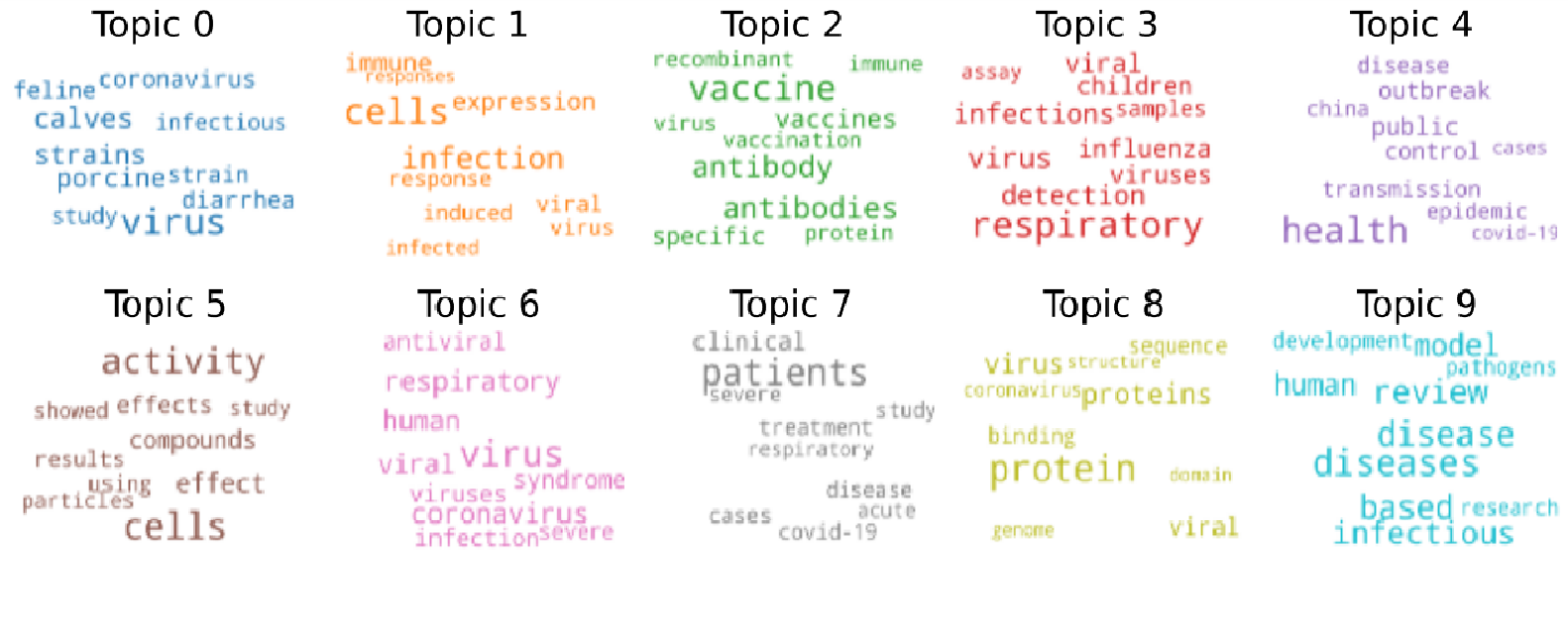
# 2.4 Question B.3: Another Topic

# 2.5 Question B.4: Another Topic

# **Results**

# 3.1 Question A Result:

# 3.2 Question B.1 Result: Identifying Topics in COVID-19 Data

The figure below shows the results we got from having 10 topics on the COVID-19 Dataset.

There are general words mixed into the results such as ‘study’, ‘activity’, ‘effects’, ‘showed’, and etc. Not only that, there are clear overlaps between the words across the topics such as ‘disease’ and ‘corona virus’. To the best of our ability, we tried to describe what each topic may be representing in the table below by highlighting the words that seemed most significant. We believe while there are clear similarities between the topics due to inseparableness of similarity of vocabularies used in the data.

|  |  |
| --- | --- |
| Topics Name | Words and Scores |
| Study on Animals (Topic 0) | 0.020\*"virus" + 0.012\*"calves" + 0.011\*"strains" + 0.010\*"**porcine**" + 0.008\*"coronavirus" + 0.008\*"diarrhea" + 0.008\*"**feline**" + 0.008\*"strain" + 0.007\*"study" + 0.007\*"infectious" |
| How infection induces Certain Responses (Topic 1) | 0.033\*"cells" + 0.023\*"**infection**" + 0.015\*"immune" + 0.014\*"expression" + 0.012\*"viral" + 0.011\*"**response**" + 0.011\*"**induced**" + 0.011\*"virus" + 0.009\*"infected" + 0.007\*"**responses**" |
| Study on treatments or Cures through vaccination (Topic 2) | 0.036\*"**vaccine**" + 0.024\*"**antibody**" + 0.024\*"**antibodies**" + 0.018\*"vaccines" + 0.015\*"specific" + 0.012\*"recombinant" + 0.012\*"protein" + 0.012\*"**vaccination**" + 0.012\*"virus" + 0.011\*"immune" |
| Virus impact in on children’s body (Topic 3) | 0.033\*"**respiratory**" + 0.020\*"virus" + 0.018\*"infections" + 0.018\*"detection" + 0.017\*"viral" + 0.015\*"influenza" + 0.015\*"viruses" + 0.014\*"**children**" + 0.012\*"samples" + 0.010\*"**assay**" |
| The world’s response and interpretation of the virus (Topic 4) | 0.024\*"**health**" + 0.009\*"**public**" + 0.009\*"**outbreak**" + 0.008\*"control" + 0.008\*"transmission" + 0.007\*"disease" + 0.007\*"**epidemic**" + 0.007\*"**china**" + 0.005\*"cases" + 0.005\*"covid-19" |
| The Study on Virus (very general) (Topic 5) | 0.018\*"cells" + 0.017\*"activity" + 0.008\*"effect" + 0.007\*"compounds" + 0.007\*"effects" + 0.006\*"results" + 0.006\*"using" + 0.006\*"particles" + 0.006\*"showed" + 0.005\*"study" |
| Virus itself (very general) (Topic 6) | 0.054\*"virus" + 0.029\*"coronavirus" + 0.029\*"respiratory" + 0.026\*"viral" + 0.026\*"human" + 0.022\*"syndrome" + 0.021\*"viruses" + 0.021\*"infection" + 0.020\*"antiviral" + 0.017\*"severe" |
| Virus’s impact on public (Topic 7) | 0.036\*"**patients**" + 0.013\*"clinical" + 0.010\*"cases" + 0.010\*"**treatment**" + 0.010\*"covid-19" + 0.009\*"disease" + 0.009\*"severe" + 0.009\*"study" + 0.008\*"acute" + 0.008\*"respiratory" |
| Virus’s biochemical structure (Topic 8) | 0.030\*"protein" + 0.017\*"**proteins**" + 0.016\*"virus" + 0.012\*"viral" + 0.009\*"**binding**" + 0.009\*"**sequence**" + 0.007\*"coronavirus" + 0.006\*"**genome**" + 0.006\*"domain" + 0.006\*"structure" |
| How researchers are trying to understand the virus (Topic 9) | 0.011\*"diseases" + 0.010\*"disease" + 0.009\*"**review**" + 0.009\*"based" + 0.008\*"infectious" + 0.007\*"**model**" + 0.007\*"human" + 0.006\*"development" + 0.006\*"**research**" + 0.006\*"pathogens" |

# 3.3 Question B.2 Result: Extracting Drugs and Diseases from the COVID-19 Data

d.

# Limitations

# Conclusions

# Bibliography