Prakash Dhimal
Manav Garkel
George Mason University
CS 657 Mining Massive Datasets
Assignment 3: Topic Modeling

Introduction

The goal of this assignment is to use Topic Modeling implementation in Spark to find topic composition and membership for the CORD-19 dataset. The data is provided with COVID-19 Open Research Dataset Challenge (CORD-19) challenge in Kaggle. The CORD-19 dataset represents the most extensive machine-readable coronavirus literature collection available for data mining to date. This dataset is approximately 23GB at the time of this writing. It contains an extensive amount of research papers related to COVID-19. Since we are processing text data, we will be using text pre-processing steps to get our data ready for topic modeling.

Data

After data acquisition, all of the documents saved as json files were read at once using spark as follows:

```
data = spark.read.json(all_json, multiLine=True)
```

The text from the body of each document above is extracted. Although the topic, abstract, and other metadata may have been used for this project, a decision was made to only use the body text and nothing else for the scope of this assignment.

```
    body_text_only_data = spark.sql(
        """
            SELECT
                body_text.text AS body_text,
                paper_id
            FROM data
            """)
```

Data preparation

Several data [text] preprocessing techniques were applied on the body text to clean up the text. Pre-processing was done on the words level and as a result any word that is not alphanumeric was removed. All the words are normalized by turning them into lowercase. Words that are in the english stopword list are removed. Finally, any word less than three characters were removed from the body text. This was done to save the computation time and to prevent things like punctuation marks from appearing on the topics list. PorterStemmer was used to stemm the words and get their base form initially. After discovering topics in different languages, the stemmer was removed.

Feature extraction

In order to get our text data ready for topic modelling using Latent Dirichlet allocation LDA, we had to do a series of transformations to our pre-processed data using the pyspark libraries below:

```
from pyspark.ml.feature import CountVectorizer, Tokenizer,
StopWordsRemover, IDF
```

The end goal here is to get a vector representation that will serve as the input for the LDA model. The next obvious step was to tokenize our text data and turn them into individual words.

```
tokenizer = Tokenizer(inputCol="body_text_cleaned", outputCol="words")
token_DataFrame = tokenizer.transform(data)
```

After tokenization, a Countvectorizer was used to count the occurrence of each word in each document. This gives a Bag-of-Words document representation where each row in the matrix is a document and each column/feature is a specific word. The value in a specific row i and column j is the count of term j in document i. In order to provide some form of normalization, the counts matrix data was converted to a TFIDF matrix. This data was used as input to the LDA model.

[73]:

	words	filtered	count_features
0	[[severe, acute, respiratory, syndrome, corona...	[[severe, acute, respiratory, syndrome, corona...	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
1	[[widely, distributed, activated, either, nore...	[[widely, distributed, activated, either, nore...	(14.0, 72.0, 184.0, 328.0, 375.0, 99.0, 80.0, ...
2	[[cases, cystic, fibrosis, caused, loss, cftr,...	[[cases, cystic, fibrosis, caused, loss, cftr,...	(1985.0, 346.0, 431.0, 576.0, 388.0, 532.0, 55...
3	[[wenn, paracetamol, fiebersenkung, oder, infe...	[[wenn, paracetamol, fiebersenkung, oder, infe...	(0.0, 6.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, ...
4	[[cystic, fibrosis, transmembrane, conductance...	[[cystic, fibrosis, transmembrane, conductance...	(1834.0, 359.0, 358.0, 548.0, 332.0, 536.0, 53...
5	[[pandemic, influenza, h1n1, virus, causes, mi...	[[pandemic, influenza, h1n1, virus, causes, mi...	(198.0, 273.0, 292.0, 296.0, 255.0, 240.0, 271...
6	[[regarding, subjective, currently, evidence, ...	[[regarding, subjective, currently, evidence, ...	(1291.0, 300.0, 101.0, 418.0, 281.0, 316.0, 11...
7	[[promuovere, gestire, cambiamenti, necessari,...	[[promuovere, gestire, cambiamenti, necessari,...	(45.0, 38.0, 5.0, 35.0, 33.0, 76.0, 58.0, 53.0...
8	[[atherectomy, combined, angioplasty, demonstr...	[[atherectomy, combined, angioplasty, demonstr...	(1966.0, 196.0, 23.0, 620.0, 316.0, 162.0, 316...
9	[[high, experimental, structures, complete, pu...	[[high, experimental, structures, complete, pu...	(1245.0, 286.0, 517.0, 419.0, 250.0, 382.0, 48...

LDA

LDA is a probabilistic generative model used in this report for the purpose of topic modelling. The goal is to fit the model to the entire corpus of CORD-19 dataset and answer the following two questions:

1. For each document, what is the topic distribution. This allows us to analyze how many topics can be used to identify a given document.
2. For each topic, which words are best used to describe each topic.

In theory, our goal is to identify each document with a concise list of topics and each topic with a concise list of words. These two goals are at odds with each other. This is because, in order to maximize the first one, you want to use as many words as possible to define a given topic, so that topic may accurately summarize the document. However, we want to keep the topic concise as well and not make it too broad and hence want to restrict the number of words used in each topic.

Model tuning

In order to obtain the best results, the main parameters tuned were the number of topics extracted from the corpus, the maximum iterations the model was run for and the number of words used to describe each topic. The optimizer used for all cases was EM.

The results shown in this report are for 5 topics and 10 topics with the number of iterations equal to 20.

Extracting and Visualizing results from LDA

Once the optimal parameters were found and the model was trained, the model returned two matrices as described above. The first is the document - topic distribution and the second is the topic-term distribution.

We utilize pyLDAvis, a python library for interactive topic model visualization to visualize the topics uncovered using LDA. This tool extracts information from a fitted LDA topic model to inform an interactive web-based visualization that can be saved to a stand-alone HTML file for portability.

LDA 5 Topics

When running the LDA model looking for 5 topics, the model was able to uniquely identify five distinct regions as seen below

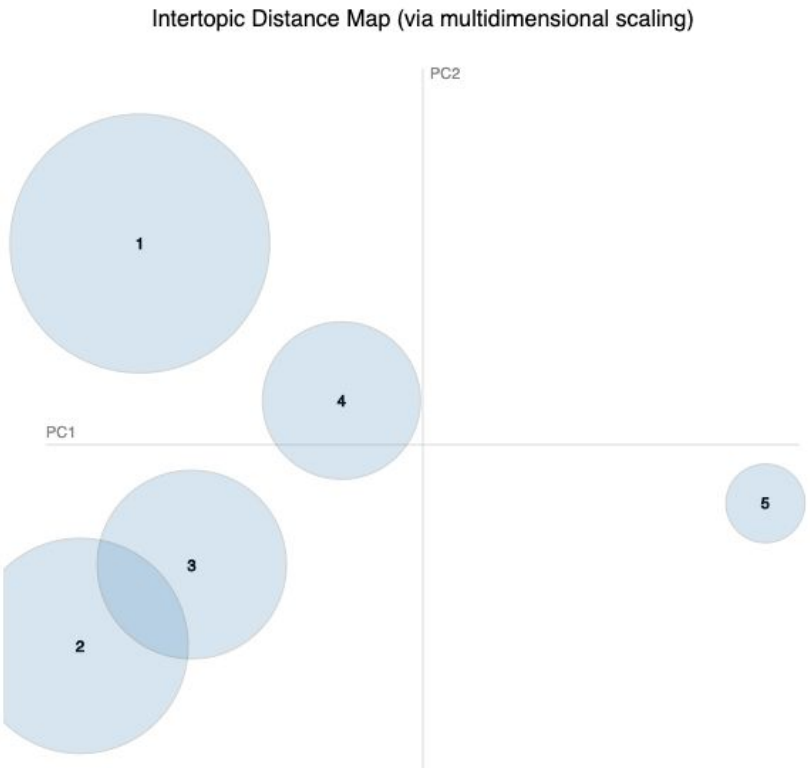

Fig 1: Five Distinct Regions found by LDA model

The ten most commonly occurring words in each topic is given below:

Topic	Words
1	Cells, Protein, Cell, Expression, Mice, Proteins, Viral, Binding, Gene, Immune
2	License, Preprint, Medrxiv, Social, Holder, Copyright, Granted, Health, Preprint, Model
3	Patients, Children, Risk, Influenza, Care, Food, Patient, Respiratory, Vaccine, Water
4	Patients, Blood, Imaging, Patient, Para, Clinical, Surgery, Treatment, Diagnosis, Therapy
5	Eine, Einer, Dans, Sich, Auch, Oder, Nicht, Durch, Nach, Sind

As can be seen above, all the terms are related to healthcare, medicine, and research. This makes sense since the model was run on COVID-19 research papers that cover a broad range of topics in these areas.

When analyzing the top 10 terms in topic 1, it can be concluded that topic 1 is a collection of terms associated with Research. Documents that have a high topic distribution probability for topic 1 would indicate that the document focuses mostly with research on the COVID-19 virus itself. This is due to the presence of words such as Cells, Proteins, Mice, Gene, Binding, and Expression.

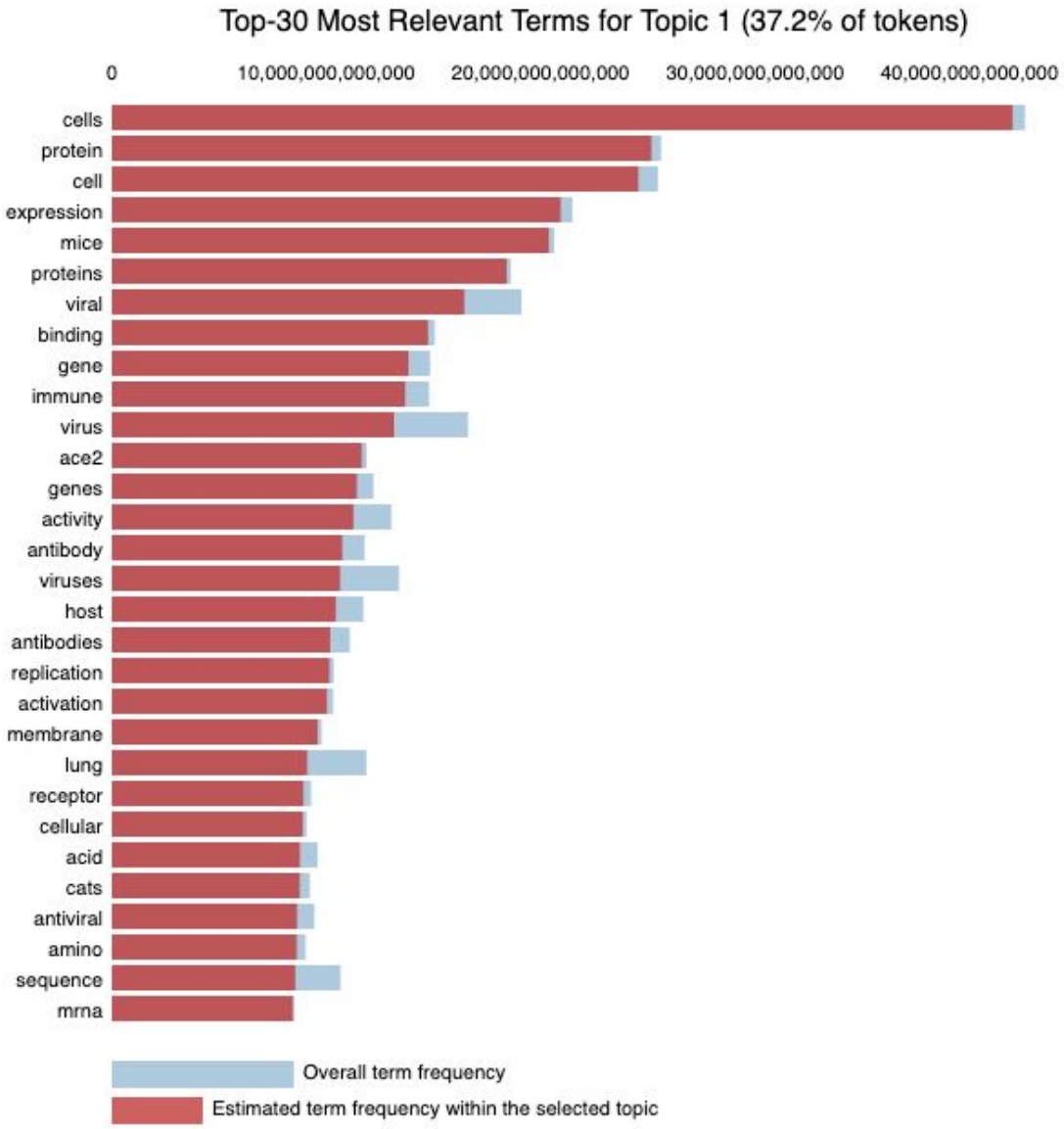

Fig 2: Overall term frequency and Estimated Term Frequency within topic 1

It can be seen in figure 2 that the most relevant features in topic1 have a ratio of the overall term frequency to the estimated term frequency within the selected topic close to 1. This implies that the most relevant terms in topic 1 will only be found in topic 1. This is further verified with topic 1 having covered the largest region in figure 1. The same results were found for topic 2.

On the other hand, it can be seen in figure 3 that terms in topic 3 do not have a ratio of the overall term frequency to the estimated term frequency within the selected topic close to 1. This

ratio is close to or below 0.5 in most cases. This implies that terms in topic 3 are not confined to this topic and may occur as highly relevant terms in other topics. It can be seen in table 1 that this is the case since there is an overlap of terms in topic 3 and topic 4. The terms in these topics indicate that these topics can be defined as Covid-19 treatment related topics.

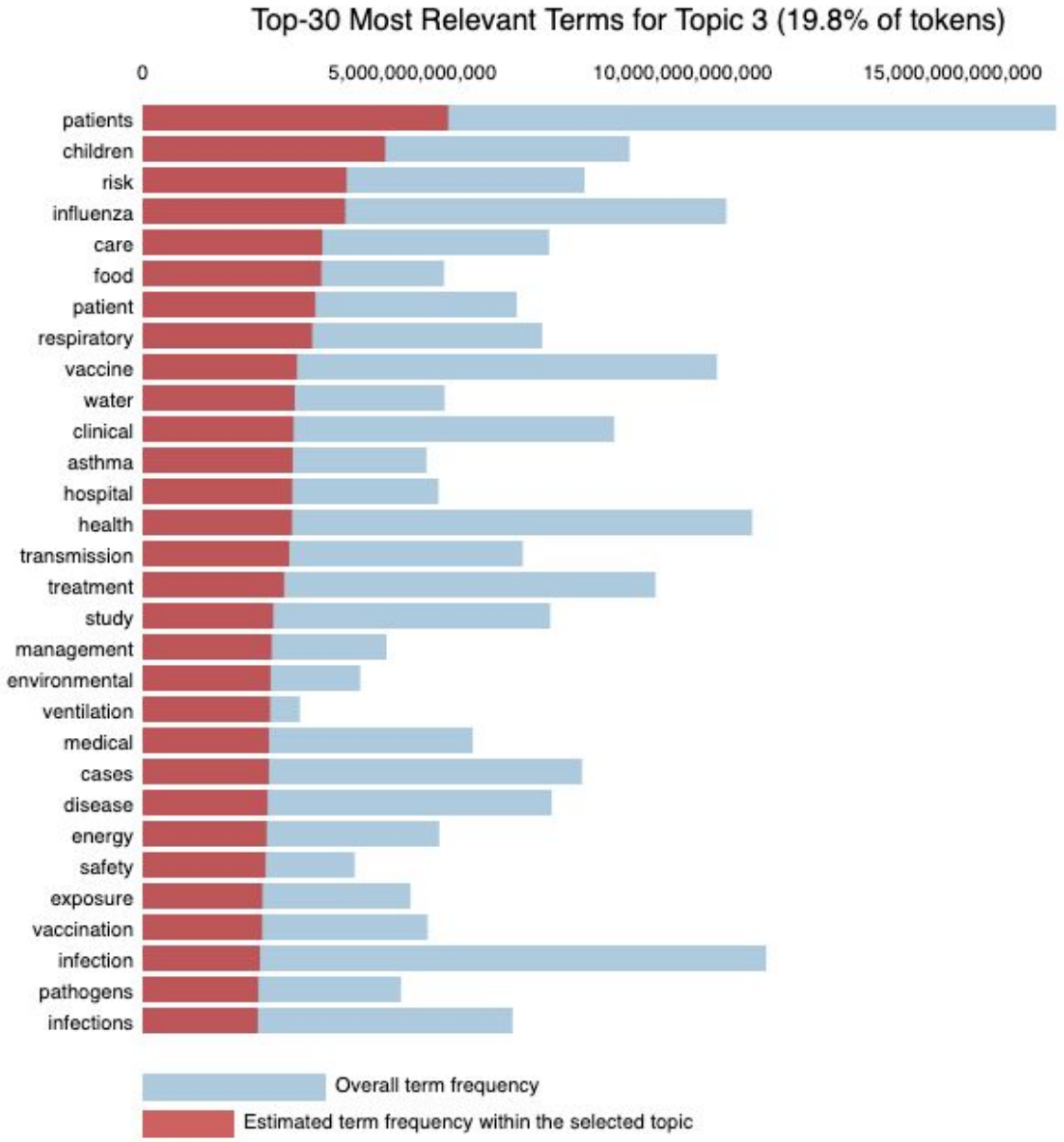

Fig 3: Overall term frequency and Estimated Term Frequency within topic 3

Interestingly, topic 5 is a list of common german words. These terms do not necessarily correlate with COVID-19 research. For example, eine translates to “a” and einer translates to “one”. This implies that there is a large number of documents that are written in german and our model was able to create a list of topics from common words in those documents. As expected,

this topic would have terms where the ratio of overall term frequency to the estimated term frequency with the topic is close to 1 with a few exceptions. This can be seen in figure 4.

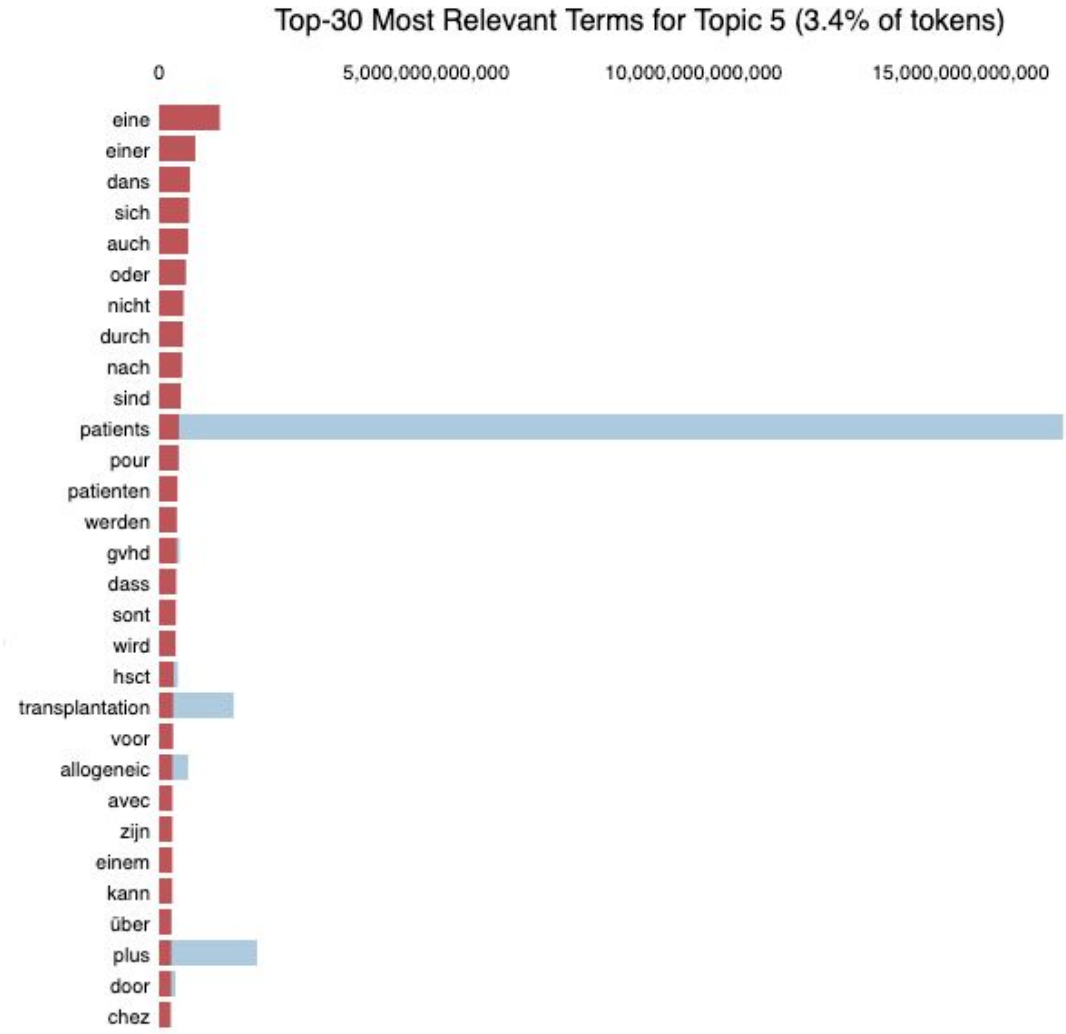

Fig 4: Overall term frequency and Estimated Term Frequency within topic 5

LDA 10 Topics

The ten most commonly occurring words in each topic is given below:

Topic	Words
1	Health, Social, Students, Participants, Public, Learning, Mental, People, Care, Training
2	Cells, Protein, Cells, Proteins, Mice, Expression, Binding, Ace2, Viral, Activity
3	Patients, Blood, Patient, Children, Clinical, Respiratory, Lung, Pulmonary, Treatment,

	Pneumonia
4	Sequences, Virus, Sequence, Model, Energy, Genome, Species, Market, Influenza, Figure
5	Cats, Dogs, Vaccinations, Animals, Vaccine, Transmission, Cases, Horses, Calves, Risk
6	Cells, Cirna, Genes, Viral, Expression, Virus, Sequence, Cell, Viruses, Protein
7	License, Preprint, Medrxiv, Holder, Copyright, Granted, Posted, Display, Version, Certified
8	Patients, Median, Cancer, Transplantation, Gvhd, Stem, Donor, Treatment, Allogenic, Hsct
9	Para, Asthma, Infl, Pacientes, Como, Copd, Renal, Lung, Hmpv, Children
10	Eine, Einer, Sich, Auch, Oder, Nicht, Durch, Nach, Sind, Calves, Patienten

When running the LDA model looking for 10 topics, the model was able to identify topics but this time with greater overlap between topics as can be seen in figure 5.

Fig 5: Ten Regions found by LDA model

Similar to the previous case, topic1 and topic 2 consist of words with a ratio of overall term frequency to the estimated term frequency within that topic close to 1 (Figure 6) thus showing that these topics consist of words that are mostly found only in those topics. This can also be seen in figure 5 above since topic 1 and topic 2 do not have any intersecting regions with any of the other topics.

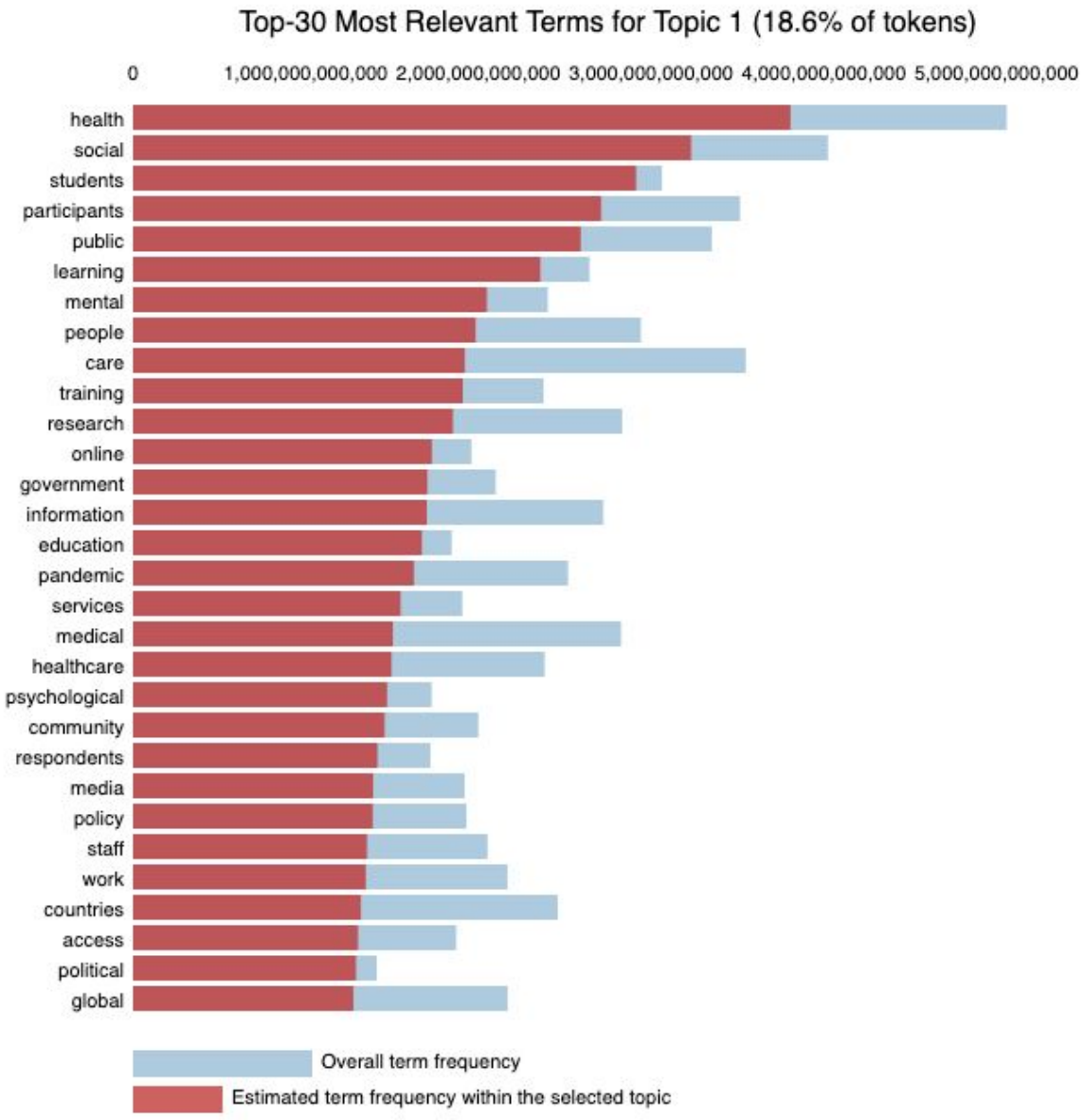

Fig 6: Overall term frequency and Estimated Term Frequency within topic 1

As can be seen in figure 5, topics 3, 4, 5, 6, and 8 cover similar regions in the graph and have overlapping areas. These topics consist of words relating to different aspects of Covid-19 experimental research papers. As expected, the ratio of overall term frequency to the estimated term frequency with that topic for these topics to be often close to 0.5 or below. Figure 7 shows these ratios for topic 5.

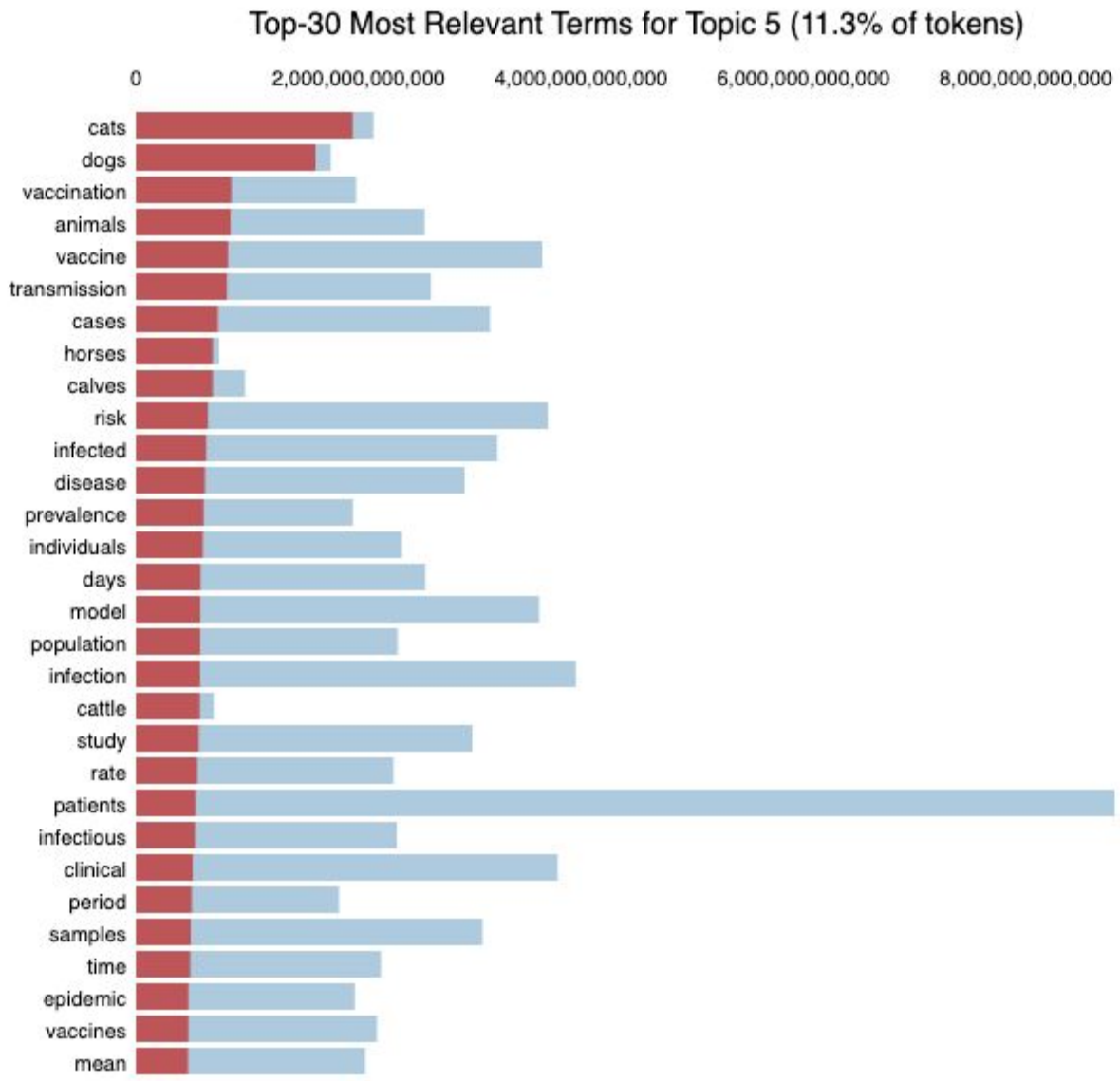

Fig 7: Overall term frequency and Estimated Term Frequency within topic 5

Once again, the language in which the paper was written played an important role in finding topics within the corpus. Like the previous case, topic 10 once again consisted of common German words which did not necessarily relate with COVID-19 research. Interestingly, topic 9 also displayed a similar trend, but with spanish words mixed in with english words. This indicated that the corpus includes several papers written in at least three different languages.

As can be seen in figure 8, spanish terms have a ratio of overall term frequency to the estimated term frequency with that topic close to one but the same ratio for english words is a lot lower. This is because those english words are present in several more documents (and topics) as well.

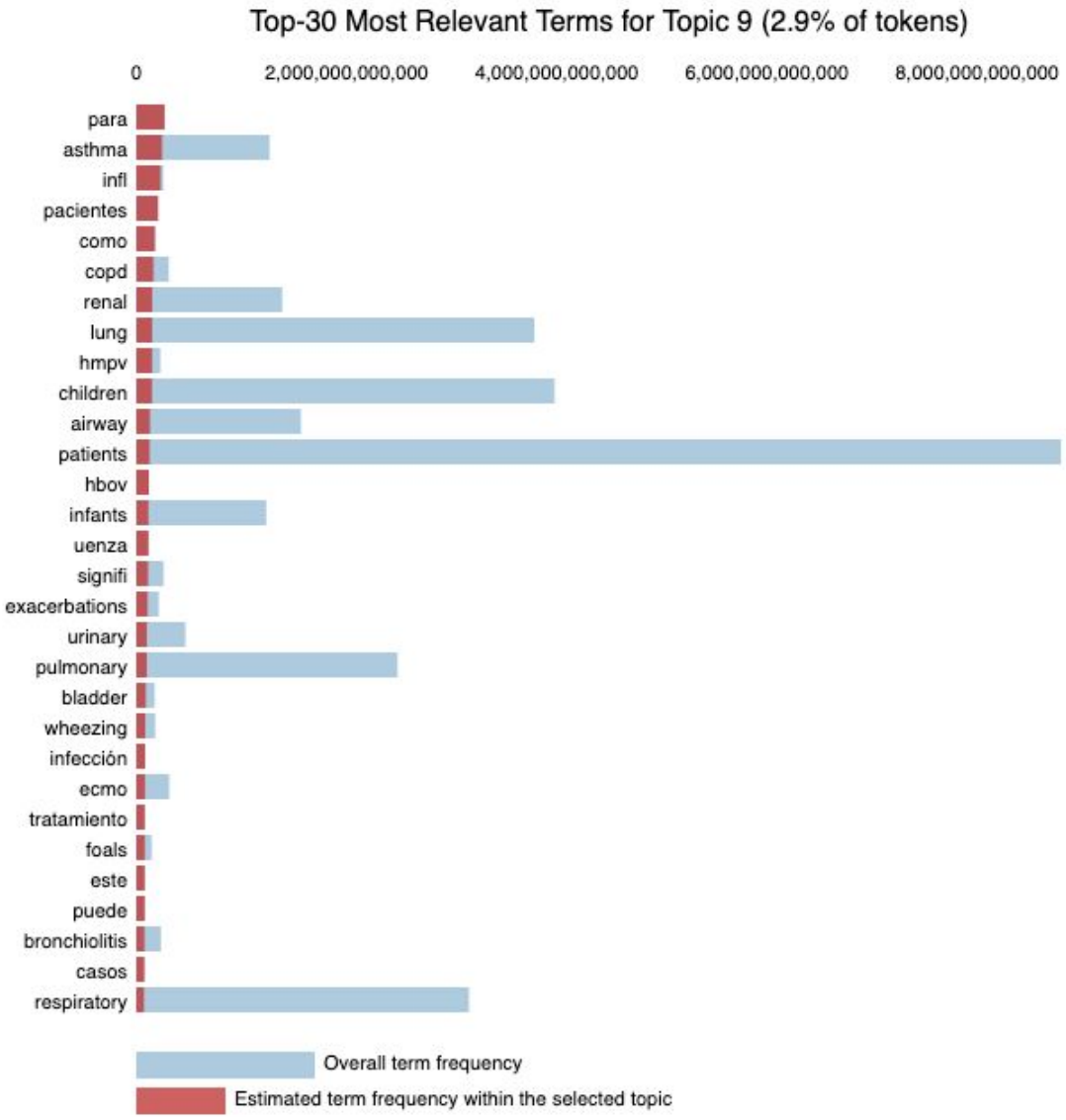

Fig 8: Overall term frequency and Estimated Term Frequency within topic 5

Conclusion

In conclusion, the LDA model in this report was able to concisely analyze the entire corpus of CORD-19 research papers and identify several terms associated with a given number of topics. Even though the model looking for 5 topics does provide more distinct topics, the model with outputting 10 topics performs better. This is because the latter model produces topics consisting of terms more closely related to each other. For ex, even though topics 3, 4, 5, and 6 can be classified as research related topics, it can be clearly seen that topic 3 is focused more on human research whereas topic 5 is more focused on animal research. This sort of distinction is not seen in the model where only five topics are output. Additionally, the recognition of more languages also provides valuable insight on the CORD-19 dataset.