Topic Modeling of COVID-19 Research Papers

Brandon Le, Yves Wienecke, Angie McGraw, Tu Vu, Keaton Kraiger

1. Project Description

The COVID-19 pandemic is a major world event, having an extensive impact on the ebb and flow of society. The pandemic has significantly disrupted the manner in which people communicate and learn. One of the leading voices in the pandemic is the medical field, providing us a little light in these uncertain times. Given the credibility and influence of medical research papers, our group is interested in applying machine learning techniques to identify patterns and make sense of the madness. We hope to utilize natural language processing (NLP) to learn about the shift in research topics leading up to and during the pandemic. Moreover, we believe this project will help elucidate the shifts in scholarly conversation and how these topics change over time and possibly, by geographic location. Our motivation for this project stems primarily from the significant effect that COVID-19 has had on each of us and our communities. In addition, some of us have a connection to the medical field and see this situation as an opportunity to interweave experience in the fields of medicine and computer science. Although we are isolated from the community and living in our new realities, we are nevertheless interested in understanding the new realities of others. For the final project, our group is proposing to perform topic modeling on research papers obtained from the Allen Institute Database, by date of publication. Two of the algorithms that we will be looking at are word2vec and Latent Dirichlet Allocation (LDA). Once the algorithms are conducted, we plan to analyze and visualize our results.

Topic modeling is a useful form of analysis for large amounts of unlabeled text[1]. Given some number of 'topics', topic modelling algorithms learn to partition a dataset (corpus) into sets that best describe these topics. In our project, we hope to employ topic modelling to extract the salient features of COVID-19 articles and discover the salient topics present in the corpus. Our corpus will be broken down into subcorpora defined by time and geographic location. By training models on these subcorpora, we can perform a qualitative analysis of the topical drift. We can also perform a quantitative analysis of the convergence or divergence of these topics.

COVID-19 topic modelling is relevant to machine learning in many aspects, and requires careful consideration of the dataset and models that we choose to use. For this project, we are learning about the challenges associated with finding and working with unstructured text and unlabelled data. Notably, some of the research papers from the dataset may not have an abstract. We will have to carefully filter and preprocess our corpus to account for discrepancies. In addition to filtering, our focus is on saliency, which is defined in our project as the importance or significance of a given keyphrase or topic. Here, we are

[1] "Topic Modeling - Mallet." http://mallet.cs.umass.edu/topics.php. Accessed 18 May. 2020.

analyzing topics extracted from COVID-19 research articles from different publications. Topic modeling involves the usage of machine learning for automatic keyphrase extraction and associating these keyphrases with topics. For our purposes, we define keyphrases as notable words or sequences of words that stand out due to a variety of factors. These factors may include word frequency, uniqueness, and grammatical complexity. We hope to use several popular machine learning approaches in order to perform topic modeling, from purely statistical approaches, such as TF-IDF, to approaches involving probability and word embeddings, such as LDA and word2vec. Evaluation of the performance of each model will be done qualitatively and quantitatively, potentially through the usage of several metrics for comparing topic distributions, including perplexity, KLD, Jaccard, and Hellinger measurements.

Our goals with this project are to gain experience working with novel datasets and to employ a few of the preprocessing and machine learning modeling techniques that were briefly discussed in class. In addition, we seek to share our project and findings to provide insight into the shifts in research during the pandemic. We expect to see shifts in topics related to health, employment, misinformation, and motivation. These topics may vary in importance by geographic area in accordance with the unique circumstances of the region.

2. Methods

We will be using the COVID-19 Open Research Dataset Challenge (CORD-19)[2], a dataset composed of over 134,000 scholarly articles about the COVID-19, SAR-CoV-2, and related coronaviruses. Next, topic modelling will involve the usage of pretrained models and analyzation techniques implemented by the Gensim[3] python library. The Gensim library exposes functions for Latent Dirichlet Allocation (LDA), Term Frequency Inverse Document Frequency (TF-IDF), and word2vec, Doc2Vec, and other machine learning algorithms. We aim to use at least LDA and word2vec for topic modeling, but we hope to use the other algorithms for comparison among the models; if the time constraints and potential complexity of the Gensim model permits. Lastly, we intend on using Matplotlib[4] and pyLDAvis[5] to visualize our results. This project will be implemented using the Python 3 programming language.

[2] "COVID-19 Open Research Dataset Challenge (CORD-19)." https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. Accessed 29 May. 2020.
[3] "gensim: About - Radim Řehůřek." 1 Nov. 2019, https://radimrehurek.com/gensim/about.html. Accessed 18 May. 2020.
[4] "Matplotlib." https://matplotlib.org/. Accessed 18 May. 2020.
[5] "pyLDAvis's documentation! - Read the Docs." https://pyldavis.readthedocs.io/en/latest/. Accessed 18 May. 2020.

3. References

"COVID-19 Open Research Dataset Challenge (CORD-19)." https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. Accessed 29 May. 2020.

DocNow/twarc. (2020, May 11). Retrieved from https://github.com/DocNow/twarc

Gensim: Topic Modelling for Humans. (n.d.). Retrieved May 18, 2020, from https://radimrehurek.com/gensim/about.html

Latent Dirichlet allocation. (2020, April 28). Retrieved May 18, 2020, from https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Tf–idf. (2020, May 3). Retrieved May 18, 2020, from https://en.wikipedia.org/wiki/Tf–idf

Topic Modeling. (n.d.). Retrieved May 18, 2020, from http://mallet.cs.umass.edu/topics.php

Visualization with Python. (n.d.). Retrieved May 18, 2020, from https://matplotlib.org/

Welcome to pyLDAvis's documentation! (n.d.). Retrieved May 18, 2020, from https://pyldavis.readthedocs.io/en/latest/

word2vec. (2020, May 6). Retrieved May 18, 2020, from https://en.wikipedia.org/wiki/word2vec