**Natural Language Processing**

Indiana University

Team Project

Spring 2020 (#32961)

David Kurniadi

Rene Lizarraga

Xin Gu

**COVID-19 Open Research Dataset Challenge (CORD-19)**

**Introduction**

Corona Virus Disease 2019 (COVID-19) is a contagious respiratory illness caused by a novel coronavirus. Although we are still learning about the origins of the virus, we know the first cases were reported after a cluster of severe pneumonia cases on New Year’s Eve 2019 in Wuhan, China. The virus has spread rapidly with the first confirmed case in the United States reported on January 20, 2020 (Holshue, 2019). According to the World Health Organization (WHO, 2020), as of April 19, 2020, there are over 2.2 million confirmed cases worldwide and 152,707 deaths worldwide, with 695,353 and 32,427 of those being in the United States, respectively. As infections and the death toll continue to rise worldwide, scientists and medical researchers race to both contain the spread of the virus and develop a vaccine.

Artificial Intelligence (AI) can help medical researchers fight the virus by scouring tens of thousands of relevant research papers at an unprecedented pace (Etzioni, 2020). In fact, Kaggle, an online community of data scientists and machine learning practitioners published an open competition (“COVID-19 Open Research Dataset Challenge (CORD-19)”) to analyze over 52,000 scholarly articles, including 41,000 with full-text, about COVID-19, SARS-CoV-2, and related coronaviruses. The competition is an attempt to support the ongoing COVID-19 response efforts worldwide through the implementation of Natural Language Processing (NLP) to quickly mine the literature to answer key scientific questions (Kaggle, 2020). There are 10 tasks as part of the challenge. We have selected to focus on one the tasks; “What is known about transmission, incubation, and environmental stability?” This task has 14 task questions to answer as listed below.

1. Range of incubation periods for the disease in humans (and how this varies across age and health status) and how long individuals are contagious, even after recovery.
2. Prevalence of asymptomatic shedding and transmission (e.g., particularly children).
3. Seasonality of transmission.
4. Physical science of the coronavirus (e.g., charge distribution, adhesion to hydrophilic/phobic surfaces, environmental survival to inform decontamination efforts for affected areas and provide information about viral shedding).
5. Persistence and stability on a multitude of substrates and sources (e.g., nasal discharge, sputum, urine, fecal matter, blood).
6. Persistence of virus on surfaces of different materials (e,g., copper, stainless steel, plastic).
7. Natural history of the virus and shedding of it from an infected person
8. Implementation of diagnostics and products to improve clinical processes
9. Disease models, including animal models for infection, disease and transmission
10. Tools and studies to monitor phenotypic change and potential adaptation of the virus
11. Immune response and immunity
12. Effectiveness of movement control strategies to prevent secondary transmission in health care and community settings
13. Effectiveness of personal protective equipment (PPE) and its usefulness to reduce risk of transmission in health care and community settings
14. Role of the environment in transmission

Our contribution is to apply Natural Language Processing (NLP) techniques to the journal corpus to help scientists and researchers get quick answers to these questions about COVID-19 to expedite their goals.

We compare 3 approaches to answer the task questions, evaluate the results, and assess the pros and cons of each approach. The first approach applies Latent Dirichlet Allocation (LDA) to retrieve relevant medical articles, leaving it to the researcher to read the article in order to find the task answers within the article. The second approach applies Global Vectors of Work Representation (GloVe) and Cosine Similarity together to evaluate the corpus to retrieve relevant text that answers the questions. The third approach combines the first two approaches; Latent Dirichlet Allocation (LDA) retrieves the relevant medical articles and then runs GloVe with Cosine Similarity on those articles to provide the text to answer the questions.

We then evaluate the results by reading the retrieved article and text from each approach and evaluate if the question is answered.

**Literature Review**

The volume of medical literature related to corona viruses is growing rapidly as scientists and medical communities are working tirelessly to quickly curb the spread of COVID-19 and discover a cure. To keep up and gain the most benefit from the explosion in literature, researchers need a tool to sort, monitor, and prioritize this wealth of valuable information. Text mining tools can help achieve that goal. Machine learning algorithms and Natural Language Processing (NLP) models have been created to help review and classify medical literatures (Harmston, Filsell, & Stumpf, 2010), such models include Naïve Bayes (Frunza, Inkpen, Matwin, Klement, & O’blenis, 2011), RNN (Hsieh, Chang, Chang, & Hsu, 2017), SVM, CNN (Bao et al., 2019). Training a model from scratch can be a time-consuming process that requires large data sets and manual labeling of classes. However, COVID-19 literature shares a similar domain vocabulary and pertinent information can be extracted by word embedding. GloVe (Pennington, Socher, & Manning, 2014) is an unsupervised model, meaning classes are not labeled, to learn word vectors from global log-bilinear regression. The goal of finding word embedding vector is to represent the probability of appearance of surrounding context. It has both advantages in efficiently use of corpus statistical information, as in matrix factorization methods and in finding word analogy, as in skip-gram. In addition, GloVe model was trained with non-zero entries of a global word-word matrix which contains information about the frequency of two words appearing together which is computationally efficient.  By investigating word-word co-appearance we may capture underlying meanings.

Cosine Similarity is a commonly used metric in NLP that measures the similarity of two documents. Cosine similarity metric measures the cosine of the angle between two vectors projected in a multi-dimensional space. The Cosine similarity of two documents will range from 0 to 1. If the Cosine similarity score is 1, it means two vectors have the same orientation. (Kanani, 2019). Cosine similarity has been used for many use cases including building a classifier of exam questions using Bloom’s Taxonomy. (Jayakodi, 2016)

Bei and others stated that Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete data such as text corpora. LDA is based on a simple exchangeability assumption for the words and topics in a document. (Bei et al. 2002)

We use LDA to measure similarity and show the confidence in the recommendations then plot topic-distribution of an articles. The LDA model understands topics, words and context. Select a time range to limit the articles that are considered, or you can decide if you want to find the latest publications. You also can choose the option to only suggest COVID-19-papers which contain COVID-19, SARS-CoV-2, 2019-nCov, SARS Coronavirus 2 or 2019 Novel Coronavirus in the text body. (Wolffram 2020)

**Methodology**

**Data Description**

The data for analysis are open-source and available for download online by Kaggle. The data are stored in CSV and JSON format files and organized into 18 columns including the journal abstract and full journal text. The metadata are well documented on the Kaggle website.

As of April 19, 2020, there are over 52,000 articles included in the data set with new articles being added regularly. It is important to note that at the time our models were run, the dataset included approximately 29,000 articles. The articles are published in English and range in year of publication between 2010 and 2020.

**Data pre-processing**

First, we downloaded the dataset (CORD-19-research-challenge) in zip file format. This is a dataset based on this date (2020-03-13) for this NLP project. It contains metadata sources in CSV format and all articles are in JSON format.

Second, we created a blank data frame to hold medical papers. We retrieved all JSON files from the 4 different folders ("biorxiv\_medrxiv", "common\_use\_sub", "non\_common\_use", "pmc\_custom\_license") and used the glob function to iterate over the files and populate them into data frame. As the last step, we merged the data frame with the all sources metadata (CSV file) and saved to a new CSV file, named as “Kaggle\_Covid19\_All\_Sources.csv”.

For data preparation we replaced null with -1 for “publish year” and we converted title to string because there are some numeric values.

For preprocessing we used [scispaCy](https://allenai.github.io/scispacy/), which is a Python package containing [spaCy](https://spacy.io/) models for processing biomedical, scientific or clinical text. We imported en\_core\_sci\_md which is a full spaCy pipeline for biomedical data with a “medium” vocabulary and 50k work vectors.

For data cleansing, we used spaCy to remove punctuations, space, and default stopwords. We also created customized stopwords and removed them from the text.

## Generate models

We used the same preprocessed dataset to run 3 competing approaches. The first approach applies Latent Dirichlet allocation (LDA) to retrieve relevant medical articles. The second approach applies Global Vectors of Work Representation (GloVe) and Cosine Similarity together to the entirety of the cleaned corpus to retrieve the relevant text. The third approach applies Latent Dirichlet allocation (LDA) to retrieve relevant medical articles and then runs those articles through Glove/Cosine Similarity model to provide the text that answers the task questions. Figure 1 shows our workflow for the 3 approaches.

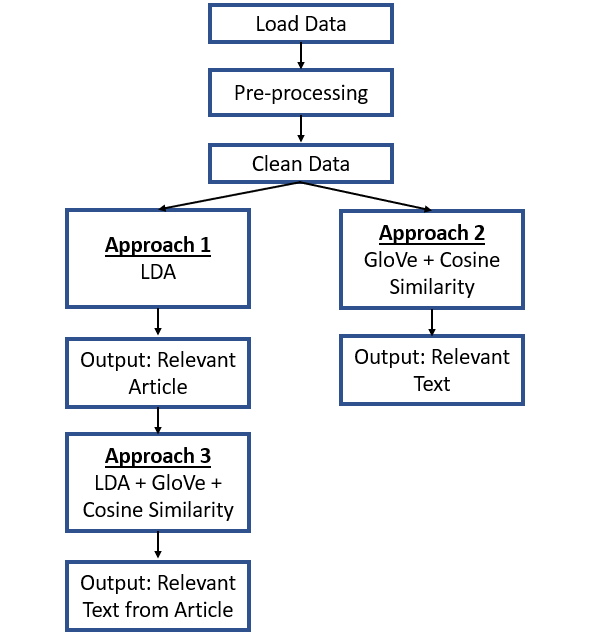


Figure 1

In the GloVe-Cosine model, we used [Wikipedia 2014](http://dumps.wikimedia.org/enwiki/20140102/) + [Gigaword 5](https://catalog.ldc.upenn.edu/LDC2011T07) word embeddings which is 100 dimensional and was trained with 6B tokens, 400K uncased vocab[2]. After collecting sentence representations with GloVe word vectors for both text body sentences and question sentences, cosine similarity was calculated. In the dataframe, we got a similarity score of every sentence for each question, the top correlated sentences can be found by sorting the highest similarity score.

The LDA model has an “understanding” of topics, words and context, most importantly combining them all together then put it to the test. The model takes the article reads in the text as a bag of words, it only understands the words that we used to train the model. The text corpus and pre-processing of words are a critical step for the model so the model can try to produce such topic distribution that would best explain how the article could have been generated. A document is a probability distribution over topics. And a topic is a probability distribution over words. (Knispelis 2016)

We can use LDA space simplex to assign a meaningful identifier so that we can compare the topics. We visualized three topics in a space. We can measure distance using Jensen Shannon distance, a standard way to measure similarity between two probability distributions giving the values from 0 to 1 (Figure 2). Zero being close and 1 being far away so the LDA model gave us a topic distribution for this article. (Knispelis 2016)

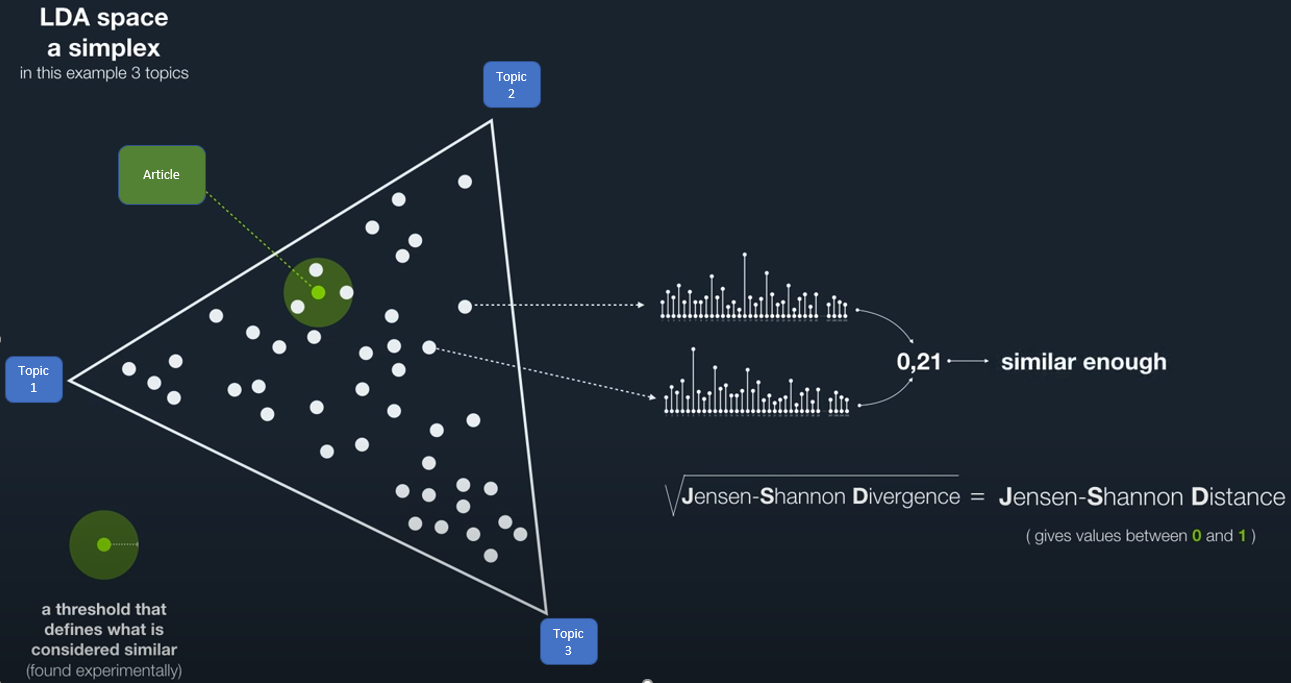
****

Figure 2

We continued with tokenization using spacy tokenizer, fit and transform text. Continue fit the result with LDA (LatentDirichletAllocation(n\_components=50, random\_state=0, n\_jobs=-1) and saved it as a csv file, named as “lda.csv”. The goal was to find the index of the k nearest articles as by Jensen–Shannon divergence in topic space. As a similarity measure, we use 1 - Jensen-Shannon distance.

**Results and Evaluation**

We evaluated the results of Approach 1 (LDA) by reading the recommended articles to determine if the question was answered in the articles. Approaches 2 (GloVe-Cosine Similarity) and 3 (LDA-GloVe-Cosine Similarity) were evaluated by determining whether that retrieve answers the question. We than assigned a binary result of Y (Yes) or N (No) to tabulate the outcome and calculated the percent of the 14 questions that were answered. The results are displayed in Figure 3. Please refer to “Appendix A (Supplemental Material)” document for sample article and text results that answer the 14 topic questions.

It is important to note that we are not medial professional or researchers. The standard for determining where the question was answered is from the perspective of the 3 graduate students in data science writing this paper. A gold standard would be a scientific researcher who can evaluate the robustness of the article or text in answering the question as well as a secondary reviewer to validate the results and assessing the level of agreement, such as by a Cohen’s kappa statistic.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Question #** | **Question** | **LDA** | **GloVe-Cosine** | **LDA-GloVe-Cosine** |
| 1 | Range of incubation periods for the disease in humans (and how this varies across age and health status) and how long individuals are contagious, even after recovery. | Y | Y | Y |
| 2 | Prevalence of asymptomatic shedding and transmission (e.g., particularly children). | Y | Y | Y |
| 3 | Seasonality of transmission. spring, winter, summer, autumn, fall, cold, hot, warm | Y | N | Y |
| 4 | Physical science of the coronavirus (e.g., charge distribution, adhesion to hydrophilic/phobic surfaces, environmental survival to inform decontamination efforts for affected areas and provide information about viral shedding). | N | Y | N |
| 5 | Persistence and stability on a multitude of substrates and sources (e.g., nasal discharge, sputum, urine, fecal matter, blood). | Y | Y | Y |
| 6 | Persistence of virus on surfaces of different materials (e,g., copper, stainless steel, plastic). | Y | Y | Y |
| 7 | Natural history of the virus and shedding of it from an infected person | Y | Y | Y |
| 8 | Implementation of diagnostics and products to improve clinical processes | Y | Y | Y |
| 9 | Disease models, including animal models for infection, disease and transmission | Y | Y | Y |
| 10 | Tools and studies to monitor phenotypic change and potential adaptation of the virus | Y | N | Y |
| 11 | Immune response and immunity | Y | N | Y |
| 12 | Effectiveness of movement control strategies to prevent secondary transmission in health care and community settings | Y | N | Y |
| 13 | Effectiveness of personal protective equipment (PPE) and its usefulness to reduce risk of transmission in health care and community settings, mask, google, gloves | N | N | N |
| 14 | Role of the environment in transmission | Y | N | Y |
| **Results** | **Percent of Questions that were Answered** | **85.7%** | **57.1%** | **85.7%** |

Figure 3

Approach 1 (LDA) successful retrieve articles that answered 12 of 14 questions resulting in an 85.7% success rate. Approach 2 (GloVe-Cosine Similarity) successfully retrieved text that answered 8 of the 14 questions. Approach 3 (LDA- GloVe-Cosine Similarity) successfully retrieved text to answer 12 of the 14 questions. Although the success rates for Approach 1 and 3 are equal, Approach 1 required the researcher to read through the article to get an answer which Approach 3 returns the actual text that answers the question. Depending on the intention of the researcher, he or she may want a quick answer versus evaluating an entire article. This is analogous to doing a Google search on the population of China. Google may retrieve a link to an article about Chinese population growth over time or it may retrieve the number and directly display in at the top of the webpage. This is a consideration before deciding which approach is more beneficial to the researcher.

Performing LDA prior to performing GloVe-Cosine Similarity appears to add a significant performance lift over GloVe-Cosine alone. Our interpretation is that once the most relevant articles are identified by LDA, then GloVe-Cosine can cleanly evaluate similarities between the question and the words in the article alone. This can cut the noise from evaluating all corpus together.

In our analysis of results, it appears that Approach 3 efficiently retrieved target literatures or sentences. Take question number one as example (“Range of incubation periods for the disease in humans (and how this varies across age and health status) and how long individuals are contagious, even after recovery.”), in 30 result sentences from Glove-Cosine model, there are 7 sentences closely related with the question and 1 sentence can directly answer the question with incubation time period. A sample result is: “Usually, COVID-19 has an incubation period of 2-7 days 2 with no obvious 41 symptoms, during which time the virus can spread from infected to uninfected individuals.” There are also 4 sentences talking about COVID-19 and incubation with ambiguous numbers, such as “few”, “some”, or “longer.” Five sentences are related to SARS which is related to COVID-19. Other sentences have the keywords in question sentences, but do not give any useful information about the question. According to the question sentence one, the result sentence only gave one direct answer about incubation period, only a touch about children are at low risk. The coverage of question aspects is low. For LDA-Glove-Cosine model, 13 sentences are closely related with the question and can give direct answers with numbers about incubation period, age difference and severity. There are 2 sentences talking about COVID-19 and incubation but with ambiguous numbers and other sentences are relevant to COVID-19, but without useful information about answering the question. According to the results from both models, model LDA-Glove-Cosine has high coverage of question aspects, result sentences are more focused on certain topic, and higher probability to get direct answers. For example, “A detailed analysis of one of the early COVID-19 clusters by Chan and colleagues 19 revealed symptomatic infections in five adult members of the same household, while a child in the same household aged 10 years was infected but remained asymptomatic, potentially indicating biological differences in the risk of clinical disease driven by age.”

The differences are even more obvious if we do not look for specific numbers to answer questions in that way the hits of keywords may lead to mismatch of question and answers. For example, in question number two: “Prevalence of asymptomatic shedding and transmission (e.g., particularly children).” Glove-Cosine model shows 13 sentences directing to other disease, such as HIV and Hepatitis B. Seven sentences gives closely related answers, and 3 are relevant with COVID-19, but not answering the question. Others are not related with the question. For LDA-Glove-Cosine model, there are 20 closely related sentences that can be used to answer the question and 9 with related knowledge and only 1 is talking about other disease.

Question number 4 is about the physical science of the coronavirus (e.g., charge distribution, adhesion to hydrophilic/phobic surfaces, environmental survival to inform decontamination efforts for affected areas and provide information about viral shedding). Neither Glove-Cosine nor LDA-Glove-Cosine give satisfying result sentences. Only one sentence from Glove-Cosine talked about disinfecting surfaces could help limit the transmission. It may due to the lack of the research in this area at the time we downloaded the data from the Kaggle website. This may be improved with updated database or additional keywords.

**Conclusion**

We believe that our models improve the goal of accelerating the identification of literature and text from articles to quickly answer relevant questions asked by researchers battling COVID-19. Based on our evaluation, if the researcher wants the be directed to an article that is likely to have an answer to his or her question, Approach 1 (LDA) is recommended for it’s computational efficiency. If the researcher prefers to have to answer to his or her question answered directly, Approach 3 (LDA-GloVe-Cosine) is recommended as it gives better, more nuanced results than Approach 2 (GloVe-Cosine) while getting directly to the answer text, which is lacking from Approach 1.

For future projects, we recommend adding robustness to the evaluation of results by implementing a gold standard by a scientific or medical researcher along with a secondary reviewer.

**References**

Kaggle COVID-19 Open Research Dataset Challenge (CORD-19) (2020). Retrieved from

<https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>

Holshue, M., DeBolt, C., Lindquist, S., et al. (January 31, 2020). First Case of 2019 Novel

Coronavirus in the United States. *The New England Journal of Medicine.* <https://www.nejm.org/doi/full/10.1056/NEJMoa2001191>

World Health Organization COVID-19 data dashboard (2020) Retrieved from

<https://covid19.who.int/>

Etzioni, O., Decario, N., (March 28, 2020). AI Can Help Scientists Find a Covid-19 Vaccine.

*Wired.* <https://www.wired.com/story/opinion-ai-can-help-find-scientists-find-a-covid-19-vaccine/>

Kanani, B., (September 29, 2019). Cosine Similarity – Text Similarity Metric in NLP. Retrieved

from <https://studymachinelearning.com/cosine-similarity-text-similarity-metric-in-nlp/>

Jayakodi, K., Bandara, M. (2016). Wordnet and Cosine Similarity Based Classifier of

Exam Questions using Bloom’s Taxonomy. *iJET,* 11(4), 142-149. http://dx.doi.org/10.3991/ijet.v11i04.5654

Bao, Y., Deng, Z., Wang, Y., et al (2019). Using machine learning and natural language processing to review and classify the medical literature on cancer susceptibility genes. *JCO clinical cancer informatics, 1*, 1-9.

Frunza, O., Inkpen, D., Matwin, S., Klement, W., & O’blenis, P. (2011). Exploiting the systematic review protocol for classification of medical abstracts. *Artificial intelligence in medicine, 51*(1), 17-25.

Harmston, N., Filsell, W., & Stumpf, M. P. (2010). What the papers say: Text mining for genomics and systems biology. *Human genomics, 5*(1), 17.

Hsieh, Y.-L., Chang, Y.-C., Chang, N.-W., & Hsu, W.-L. (2017). *Identifying protein-protein interactions in biomedical literature using recurrent neural networks with long short-term memory.* Paper presented at the Proceedings of the eighth international joint conference on natural language processing (volume 2: short papers).

Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation.* Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Blei, D., Ng, A., Jordan, M. (2020). Latent Dirichlet Allocation. *Journal of Machine Learning*

*Research* 3 (2003) 993-1022. <https://ai.stanford.edu/~ang/papers/jair03-lda.pdf>

Wolffram, D. (2020). Topic Modeling: Finding Related Articles.

<https://www.kaggle.com/danielwolffram/topic-modeling-finding-related-articles/notebook>

Knispelis, A. (2016). LDA Topic Models.

<https://www.youtube.com/watch?v=3mHy4OSyRf0&t=515s>