Topic modeling for CORD-19 dataset*

*Final Project Report

Feyza Nur KILICASLAN
Computer Engineering
Hacettepe University
Ankara, Turkey
feyzanur@cs.hacettepe.edu.tr

I. Abstract

Coronavirus disease (COVID-19) emerged in the last days of 2019 in the city of Wuhan, Hubei Province of China. As this disease started to spread among the continents very quickly, it was declared as a pandemic by the World Health Organization (WHO) as of March 11. Many scientists are researching for the treatment of this disease and trying to establish relationships between existing data and catch meaningful patterns. Considering this tragic situation, two research questions are raised in this study. The first research question is how to split literature resources about coronavirus disease to help researchers who need to study articles according to their topics and the second one is how to use obtained topics for recommendation system. Thus the obtained topics can be used for the information retrieval process. In this project,topic modeling methods, Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), are implemented using CORD-19 data set [1], which is a compilation of publications in the literature.

II. Introduction

Many people suffer from coronavirus disease and some of them, mostly elderly people, die for the same reason since the medical community could not find any validated effective treatment or vaccine yet. Since the time coronavirus disease emerged, many scientists contributed the literature by producing studies that present some drugs and treatments that may be a cure for this disease [2], [3], relations coronavirus disease with other chronic diseases [4], etc. All these studies related to this topic is very valuable. Therefore these literature resources need to be split according to their topics in order to increase the utilization of reaching them. Separating the articles by topics will make it easier to research about this disease since it is a really hard task to examine all articles in the literature. Analyzing literature articles in the CORD-19 data set by applying topic modeling techniques will facilitate the process of reaching an appropriate article.

Topic modeling is one of the trend methods in machine learning, text mining, and Natural Language Processing (NLP) fields. It is used in many areas such as in social media analytics, real-time event detection, recommendation systems, etc. In this study, topic modeling techniques are used to meet the above-mentioned need in the CORD-19 data set. Obtained topics are used to build a search engine by calculating the cosine Similarity between query terms and topic terms. As a result, the search engine presents the most relevant topics as recommendations according to the query terms.

III. Methods and Dataset

In this study, the topic modeling tool which is known as 'unsupervised' machine learning is used in CORD-19 data set. Topic modeling enables an effective way to analyze large volumes of unlabeled text. In order to perform topic modeling, it is utilized from Latent Semantic Analysis (LSA) also known as Latent Semantic Index (LSI), which is a natural language processing technique and Latent Dirichlet Allocation (LDA), which is a type of probabilistic topic modeling.

These algorithms use different mechanisms to identify topics.The aim of LDA is to break down a conditional term by document probability distribution into two distributions: topic distribution over keywords and document distribution over topics [5]. In [6] LSA described as it delegates the meaning of a word as a kind of average of the meaning of all text in which it exists and the meaning of a text as a kind of average of the meaning of all the words it embodies.

Graphical model representation of LDA is demonstrated in Figure 1. As it is seen from Figure 1, there are three levels in LDA demonstration. α and β are corpus-level parameters that represent document-topic density and topic-word density, respectively. θ is the topic distribution for document, z and w are word-level variables. z is the topic for the word in document and w is the specific word. M represents number of documents and N represents total number of words in all documents.

The data CORD-19, placed at the Kaggle platform, is chosen as a dataset. It is stated that CORD-19 is a resource of over 45,000 scholarly articles, including over 33,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses in [1]. But in this study the whole dataset was not used, instead, articles were selected only from the bioRxiv archival services folder in the dataset in order to implement LDA and LSA algorithms. In the final dataset there are 1053 articles and 7 variables.

IV. Experiment

In this study, experiments are applied by using R Language. In the final dataset, abstract and paper id variables are used.

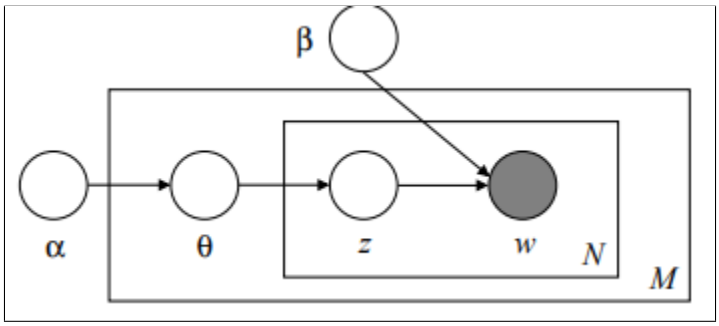


Fig. 1: Graphical model representation of LDA

Before starting implementing topic modeling algorithms, some text cleaning processes are applied. The cleaning process consists of following steps:

- Firstly missing and undefined values (NA and NULL) are found and removed.
- Tokenization consist of splitting strings such as text documents into smaller units. These units can be individual word or term. Each of these smaller units are called tokens.The corpus, text documents are tokenized as words. This process enables to find the frequency of the word.
- Case folding is a kind of normalization technique. The lower casing is a very common pre processing way in terms of case folding. If we do not apply this step, our model might assume that a word which is at the beginning of the sentence with uppercase is different from the same word which appears later in the sentence with lowercase. In order to prevent this situation and improve the accuracy of the model, all words were converted into lowercase.
- The stopword removal step helps to focus on the important words instead of commonly used words that have low information about text.“a”, “the”, “is”, “are” and etc. are an example of stop words for the English language. Also custom stopwords like author,https, preprint, etc are defined and removed from text.
- Punctuation and numbers are removed.
- Leading and/or trailing whitespaces are removed from text.

After preprocessing and removal of stop words, createDTM function is used to create a document term matrix. It is a sparse matrix containing terms and documents as dimensions. When building the document term matrix, sentences are broke up into a window size of 1–2 words. It means number of minimum n gram is 1 and maximum *n* gram is 2. The term frequency matrix are explored, which shows the number of times the word/phrase is occurring in the entire corpus of text. If the term is less than 2 times, it is eliminated, because it does not add any value to the algorithm, and it will help to reduce computation time as well.

A. Latent Dirichlet Allocation

The LDA algorithm is run for topic modelling with document term matrix. Firstly a number of topics (k) is assigned manually to 30. Then the coherence score is calculated in order to choose optimal a number of topics from 2 to 30. Coherence score is a score that calculates if the words in the same topic make sense when they are put together. This gives degree of the quality of the topics being produced. The higher the score for the specific number of *k*, it means for each topic, there will be more related words together and the topic will make more sense. Figure 2 shows that k equals 18 gives the highest coherence score. Furthermore, probabilistic coherence of obtained best model is shown in 3a.Probabilistic coherence measures how related words are in a topic, controlling for statistical independence.

After finding the optimal number of topics, the different words within the topic is analyzed. Table I shows the top 10 terms describe what the topic is about for 5 topics as a demonstration purpose. The words are in ascending order of phi-value. Here phi value means that probability of word *w* occurring in topic *k*. The first word implies a higher phi value.

TABLE I: Top 10 words in topics from LDA

Topic Labels				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
patients	cells	medrxiv	fitness	transcription
clinical	cell	display	selection	dna
severe	protein	display perpetuity	evolution	termination
disease	frameshifting	granted medrxiv	viral	jbp
pneumonia	sequence	medrxiv display	genetic	protein
age	membrane	perpetuity	bees	pp
symptoms	ccr	granted	foragers	proteins
hospital	mm	allowed	immunity	resistance
study	previously	reuse	mutation	complex
ct	human	health	strains	dux

Topic, word and frequencies are examined as it is shown in

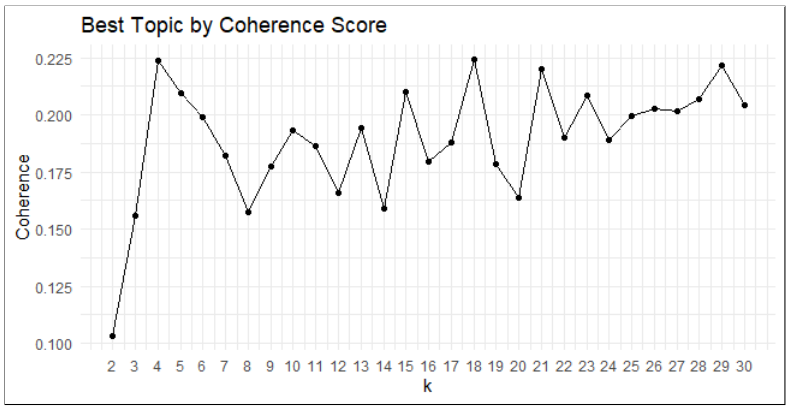
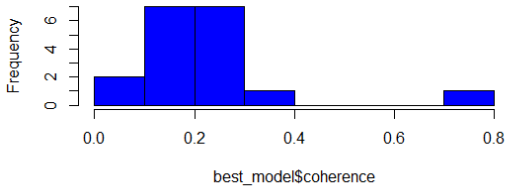
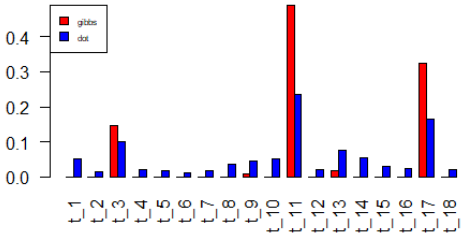


Fig. 2: Coherence score of LDA



(a) Histogram of probabilistic coherence from LDA



(b) The prediction comparison between Gibbs and Dot methods in LDA

Fig. 3: The results of LDA

Table II. For instance term frequency of word ”sars” in topic 13 is 993 and document frequency is 267.

Hellinger distance is used to analyze which topics are closely related. Figure 4 illustrates the dendrogram that shows that there is a great similarity between topic 8 and 14.

In LDA, predictions are implemented by using Gibbs and Dot. The comparison between Gibbs and Dot methods is represented in Figure 3b. When we compare the two methods from Figure 3b, it is seen that Gibbs method is has a much less noisy result although it is slower than Dot method.

TABLE II: Sample topics, word, frequencies from LDA

Topic-Label	Word	Value	Term-Frequency	Document-Frequency
13	sars	0.0592	993	267
13	cov	0.0543	1084	249
13	sars cov	0.0519	846	226
1	patients	0.05032	861	213
18	cov	0.03331	1084	249
16	rna	0.02705	508	137
8	china	0.0258	431	239
7	viral	0.0218	633	238
7	host	0.02090	452	150
18	sars	0.0194	993	267

B. *Latent Semantic Analysis*

Latent semantic analysis (LSA) uses vectors to represent every statement in a given document. It uses a single value decomposition on a set of matrices such as a document term matrix, TF-IDF matrix. Analyzes done in LDA are also done in LSA algorithm. Table III shows top 10 words from 5 topics produced from LSA. Figure 5a shows the probabilistic coherence in order to evaluate individual topics. In LSA, again predictions are implemented by using Gibbs and Dot. The comparison between Gibbs and Dot methods is represented in Figure 5b.

C. *Recommendation System*

Cosine similarity metric is used to obtain relevant topics with user query. Cosine similarity metric measures closeness between two vectors of an inner product space in terms of similarity. To be able to calculate cosine similarity, each document is need to be represented by a term-frequency vector. Therefore, term document matrices of obtained topics

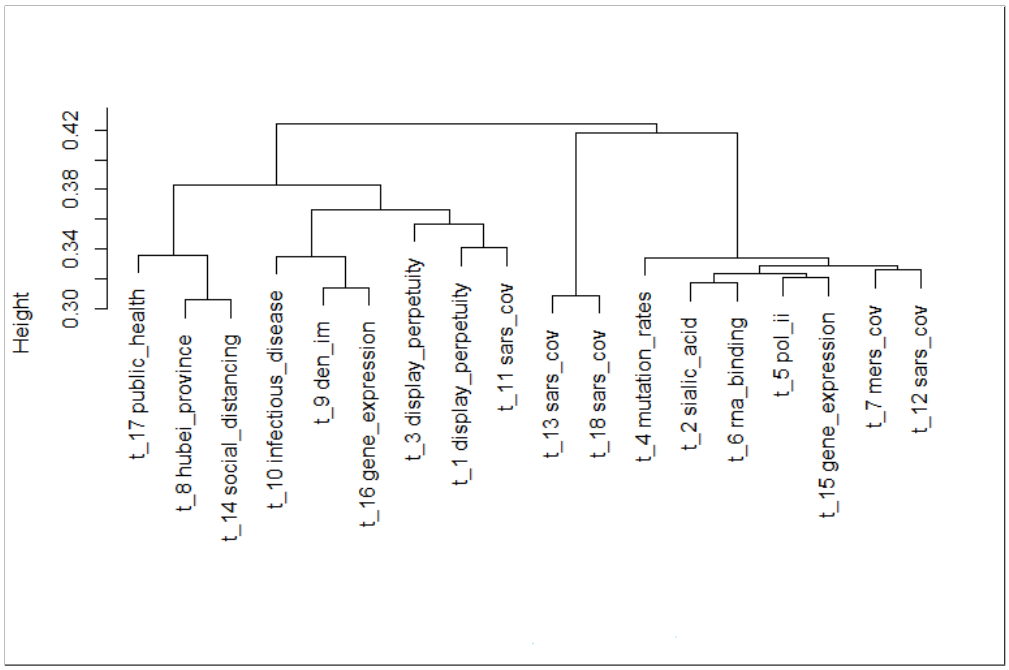
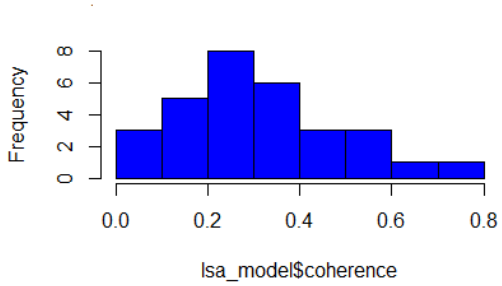
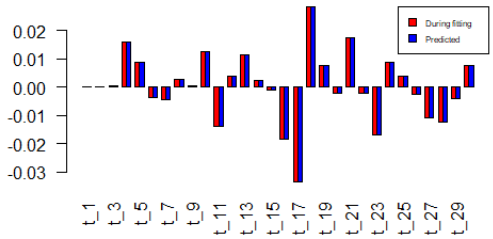


Fig. 4: Topic similarities



(a) Histogram of probabilistic coherence from LSA



(b) The prediction comparison between Gibbs and Dot methods in LSA

Fig. 5: The results of LSA

TABLE III: Top 10 words in topics from LSA

Topic Labels				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
teams	development	cov	sars cov	sars
background	background march	background outbreak	patients diagnosed	diagnosed
words	main words	words main	main	words reuse
forecasted	status	analyzed	cov	sars
cov	sars	fmw	sars cov	sleep
fmw	sleep	sleep disturbances	disturbances	prevalence sleep
background march	patients diagnosed	diagnosed	march	patients
pathogenic bacteria	host proteins	limiting	resource	bacteria
background outbreak	outbreak spreading	outbreak	ncp	risk infection
background suspected	cluster coronavirus	cluster	suspected	cov

and query terms are created. As a result of recommendation system, it is shown most relevant topic terms, topic labels and similarity score to user. For instance ”patients clinical china covid” query is resulted as shown in Table IV for LDA and Table V for LSA.

TABLE IV: Recommendation system results using LDA

Topic terms	Score	Topic Label
patients clinical severe disease pneumonia age symptoms hospital study ct	0.2498	topic 1
china epidemic wuhan coronavirus confirmed outbreak model ncov spread hubei	0.1330	topic 8

TABLE V: Recommendation system results using LSA

Topic terms	Score	Topic Label
transcription dna termination jbp protein pp proteins resistance complex dux	0.1670	topic 5
transmission human population disease health pathogen infection risk influenza diseases	0.1569	topic 17
fitness selection evolution viral genetic bees foragers immunity mutation strains	0.1419	topic 4
viral host replication rna species exon mutations activity coronavirus bat	0.0476	topic 7

V. Discussion

Evaluation of topic modeling is commonly divided into two categories as intrinsic and extrinsic. The intrinsic evaluation consists of holdout-log likelihood/perplexity and with human-in-the-loop (word intrusion) while extrinsic evaluation is conducted through classification test data and task-specific metrics.

In the literature, several methods are presented for evaluating topic models such as perplexity [7], coherence methods [8], the log probability of a held-out set of document [9]. In this work, the evaluations of topic modeling techniques were proceeded by probabilistic coherence and "eyeballing" approach. According to the results of implementation of two algorithm, LDA gives more meaningful topic terms than LSA. Also in recommendation system, similarity score between topic terms and query terms is higher in LDA results.

VI. Conclusion

In this study, CORD-19 dataset is used to implement topic modeling techniques. As the main goal of this study, topic modeling techniques that are Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are used to create a recommendation system. The results of applied algorithms are analyzed using some statistical methods. The cosine similarity metric is used for the purpose of building a recommendation system. In the recommendation system, it is found that LDA topic terms are closer to query terms in terms of similarity.

References

[1] https://pages.semanticscholar.org/coronavirus-research

[2] Sanders JM, Monogue ML, Jodlowski TZ, Cutrell JB. Pharmacologic Treatments for Coronavirus Disease 2019 (COVID-19): A Review. JAMA. 2020;323(18):1824–1836. doi:10.1001/jama.2020.6019 Wen Zhang, Yan Zhao, Fengchun Zhang, Qian Wang, Taisheng Li, Zhengyin Liu, Jinglan Wang, Yan Qin, Xuan Zhang, Xiaowei Yan, Xiaofeng Zeng, Shuyang Zhang,

[3] The use of anti-inflammatory drugs in the treatment of people with severe coronavirus disease 2019 (COVID-19): The Perspectives of clinical immunologists from China,Clinical Immunology,Volume 214,2020,108393,ISSN 1521-6616, https://doi.org/10.1016/j.clim.2020.108393.

[4] Mengmeng Zhao, Menglong Wang, Jishou Zhang, Jing Ye, Yao Xu, Zhen Wang, Di Ye, Jianfang Liu, Jun Wan,Advances in the relationship between coronavirus infection and cardiovascular diseases,Biomedicine and Pharmacotherapy,Volume 127,2020, 110230,ISSN0753-3322, https://doi.org/10.1016/j.biopha.2020.110230.

[5] Bergamaschi, Sonia, Laura Po, and Serena Sorrentino. "Comparing Topic Models for a Movie Recommendation System." WEBIST (2). 2014.

[6] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." Discourse processes 25.2-3 (1998): 259-284.

[7] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09). AUAI Press, Arlington, Virginia, USA, 27–34.

[8] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12). Association for Computational Linguistics, USA, 952–961.

[9] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09). Association for Computing Machinery, New York, NY, USA, 1105–1112. DOI:https://doi.org/10.1145/1553374.1553515