Deep Mining Covid-19 Literature

Joshgun Sirajzade(✉), Pascal Bouvry, and Christoph Schommer

Department of Computer Science, University of Luxembourg, Belval, 6 avenue de la
Fonte, 4264 Esch-sur-Alzette, Luxembourg
{joshgun.sirajzade,pascal.bouvry,christoph.schommer}@uni.lu
https://wwwen.uni.lu/recherche/fstm/dcs

Abstract. In this paper we investigate how scientific and medical papers about Covid-19 can be effectively mined. For this purpose we use the CORD19 dataset which is a huge collection of all papers published about and around the SARS-CoV2 virus and the pandemic it caused. We discuss how classical text mining algorithms like Latent Semantic Analysis (LSA) or its modern version Latent Drichlet Allocation (LDA) can be used for this purpose and also touch more modern variant of these algorithms like word2vec which came with deep learning wave and show their advantages and disadvantages each. We finish the paper with showing some topic examples from the corpus and answer questions such as which topics are the most prominent for the corpus or how many percentage of the corpus is dedicated to them. We also give a discussion of how topics around RNA research in connection with Covid-19 can be examined.

Keywords: CORD19 · SARS-CoV-2 · Covid-19 · Pandemic · Topic modeling · Latent Drichlet Allocation · word2vec

1 Introduction

The Covid-19 virus and the pandemic it caused hit the world very hard in many aspects, including the economy, political and social life as well as health care systems of different countries. As answer to that, in many countries around the world a huge effort has been put in order to fight against the pandemic; the virus was studied profoundly, vaccines were developed and many measures against the spread of the virus were taken. Despite that, the pandemic is still going on as of September 2022 and many aspects of it, especially the best ways of fighting it, are still not very well understood, yet. At the same time, the nations around the world gained a great experience in fighting a pandemic and are prepared like never before. The huge portion of this experience and knowledge is hidden in the scientific and medical publications around the world. The number of these publications is meanwhile increasing to a six digit figure and it is very hard or sometimes impossible to keep the overview. Time needed for reading or reviewing that much publications has grown out of a life span of an human being long time ago. This is where the digital organisation and techniques of information retrieval

H. Florez and H. Gomez (Eds.): ICAI 2022, CCIS 1643, pp. 121–133, 2022.
https://doi.org/10.1007/978-3-031-19647-8_9

and text mining come in handy. We believe that with the help of intelligent text mining techniques one can create a good overview of the existing papers about Covid-19. In this paper we discus how the algorithms like LDA or Word2Vec can be used in order to extract topics from Covid-19 literature. We shed light into the inner workings of the algorithms and into the history and the future of extracting topics. Then we investigate how it can be best adapted for Covid-19 literature in the example of the CORD19 dataset.

2 Dataset

Cord19 (COVID-19 Open Research Dataset) is a big and open source dataset consisting of scientific publications about and around Covid-19 pandemic gathered by the Allen Institute for AI in collaboration with the White House Office of Science and Technology Policy (OSTP), the National Library of Medicine (NLM) and many other institutions and research organisations[1]. Its aim is to centralize all the research papers about SARS-CoV-2 virus, the disease it causes, the pandemic and its prevention as well as its impact from social, political and economical view [16]. This dataset is constantly growing as new papers are published. In the time of this paper (September 2022) the data comprises of slightly under half a million papers over 30 GB as pure JSON files.

3 Related Work

3.1 Search Engines

From the moment the data was published Allen Institute announced competitions in different tasks in Kaggle[2]. Shortly after several information retrieval systems were built for the dataset. As one of the projects worth mentioning is a neural search engine which was developed by the Amazon Web Services AI team[3]. This search engine is publicly available online and uses traditional scalable information retrieval methods combined with natural language querying possibilities [1]. It is based on Amazon Kendra which utilises deep learning techniques for search engines. The search engine performs document ranking, passage ranking, question answering and FAQ matching and leverages knowledge graphs and topic modeling for better structuring the search. Although also a classical topic modeling technique like Z-label LDA was applied, the developers reduced the topics generated by the algorithm to the following ten after the consultation of medical professionals: Vaccines/immunology, Genomics, Public health Policies, Epidemiology, Clinical Treatment, Virology, Influenza, Healthcare, Industry, Pulmonary Infections and (human) Lab Trials [1]. After that the whole corpus was multi-label classified with the mentioned topics.

[1] https://allenai.org/data/cord-19.

[2] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

[3] https://cord19.aws/.

Another search engine was developed by the joint effort of researchers from the Universities of Waterloo, Delaware, New York and Canadian Institute for Advanced Research. The creators of this engine called it Covidex[4]. It also uses the classical keyword search in the core of their platform which was supplemented with the sequence to sequence transformer models for reranking and feedback [18]. The algorithm used for the latest is the so called doc2query [12] which, in its turn, is an extension to the well known BERT model. Also the search engine built by google ai labs[5] is worth mentioning.

3.2 Mining of Cord19

Besides the efforts of building a search engine there has been also some attempts to analyse, study and mine the cord19 dataset. These are mostly based on finding topics in the dataset or applying other machine and deep learning technologies to the dataset. Otmakhova et al. [13] have an interesting approach by applying Latent Drichlet Allocation (LDA) to the documents which were transformed into an unordered set of Unified Medical Language System (UMLS) concepts. Worth mentioning is also the work of Karami et al. [7] who applied descriptive statistics and topic modeling to a small corpus of 9298 articles about Covid-19. Here the terms were pre-annotated as “chemical” or “disease” and were incorporated into the analysis.

4 Text Mining

Text Mining is a fast growing sub discipline, historically inspired from Data Mining as being its special form. Its aim is to discover and extract information from large text data why it is also called Text Data Mining. This information is mostly either hidden and can not be spotted by the reader immediately or the text data at hand is so huge that it would take a lot of time to read and be analyzed by a human. The spilt of Text Mining from Data Mining is mostly due to the fact that text data has a different structure (also known as unstructured data) compared to well structured data that Data Mining is using. In this regard Text Mining is closely related to Natural Language Processing (NLP). Usually, apart from the traditional data mining techniques, the topics in this field are Word Association Mining, Text Clustering, Text Categorisation, Text Summerization, Topic Analysis, Opinion Mining and Sentiment Analysis [17].

4.1 Classical Topic Modelling with Bag of Words: From LSA to LDA

Generating Term Document Frequencies and Vector Space Model dominated the IR and Text Mining research for a long time. Compressing the text documents

[4] https://github.com/castorini/covidex.

[5] https://covid19-research-explorer.appspot.com/.

to the topics with the help of Singular Value Decomposition (SVG) opened a new research field that is also called Latent Semantic Analysis (LSA) [5]. In its basic form it is usually formulated as $A_{mn} = U_{mk}\Sigma_{kk}V_{kn}^{\top}$ where A stands for the Document Term Matrix, U and V are orthogonal matrices of left and right singular vectors (columns) respectively, and Σ is a diagonal matrix of the corresponding singular values. The expressing of a (usually large and sparse) document term matrix in 3 components can be interpreted as U and V expressing document to term or term to document relationship respectively and Σ being a reduced form of Document Term Matrix to topics. Note that, the shape of Σ or the number of topics can be freely defined. Usually, it depends on the size of the text collection and is a number somewhere between 40 up to 300 in the real life LSA applications.

In the late 90s this process was formulated probabilistically which was called Probabilistic Latent Semantic Analysis (PLSA) [6]. It formulates the topics, thus the latent variable as $z \in Z = \{z_1, ..., z_k\}$ assuming that the documents $d \in D = \{d_1, ..., d_N\}$ are composed of topics Z, which is an unobserved variable, and the topics are composed of terms $w \in W = \{w_1, ..., w_M\}$. This way, the whole model is defined as mixture of

$$P(d, w) = P(d)P(w|d), P(w|d) = \sum_{z \in Z} P(w|z)P(z|d). \tag{1}$$

Here z is designed as bottleneck since its cardinality is smaller than the number of documents or words. The whole model can also be rewritten as

$$P(d, w) = \sum_{z \in Z} P(z)P(d|z)P(w|z), \tag{2}$$

where our diagonal matrix Σ will be equivalent to $diag(P(z_k))_k$, so PLSA is very similar to LSA. Model Fitting here happens with the help of EM Algorithm where the posterior probabilities for the words and documents given the topic are calculated

$$P(w|z) \propto \sum_{d \in D} n(d, w)P(z|d, w), \tag{3}$$

$$P(d|z) \propto \sum_{w \in W} n(d, w)P(z|d, w), \tag{4}$$

$$P(z) \propto \sum_{d \in D} \sum_{w \in W} n(d, w)P(z|d, w). \tag{5}$$

With the development of PLSA the research was paying more attention to the fact that all these approaches did not consider the order of words in a text. This is known as the so called Bag-Of-Words approach which has a great advantage and flexibility in terms of mapping of all documents and queries to one fix number of dimension, usually the total number of words in all documents. However, this approach has also some shortcomings. These were more apparent when PLSA was further developed to Latent Drichlét Allocation (LDA). First, in PLSA the number of parameters in the model grows linearly with the corpus size. Secondly, it was not yet clear how to assign probability to a document outside of the training set [2]. This problem was solved by adding the uncertainty of the topic coverage distribution for a document and entire collection in the form of an integral. It assumes that each document in a collection is generated by a random mixture of latent topics which is chosen from a Dirichlet distribution $\theta \sim Dir(\alpha)$. By plugging in all the variables we get the marginal distribution of a document:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n,\beta) \right) d\theta. \tag{6}$$

The probability of a corpus is then obtained as product of the marginal probabilities of single documents:

$$p(\mathbf{D}|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_d n} p(z_d n|\theta_d) p(w_d n|z_d n,\beta) \right) d\theta_d. \tag{7}$$

One of the main advantages of LDA, which is basically a Bayesian formulation of PLSA, is that by addressing and adding the topic mixture the parameters of the model are reduced to $k + kV$ which avoids the theoretical overfitting problem of PLSA. However, the most important detail here is the dilemma of using the power of bag-of-words approach – indeed LDA takes advantage of the *exchangebility* of words, topics and the documents – on one hand and on the other hand the reflection on the shortcoming of it by adding the notion of n-grams to topic modeling. In fact, Blei et al. [2] suggests that LDA can be used as mixture model of larger structural units such as n-grams or even paragraphs. In probabilistic language modeling the usage of n-grams has a rich tradition and LDA makes an attempt to combine these two.

4.2 Beyond Bag of Words: Text Classification and Clustering with Word Embeddings

Traditionally, regarding the word order the probabilistic language modeling was the opposite of bag-of-words approach. It considers the order of words (or characters), usually in form of *bi*- or *tri*-grams, but initially they do not capture and ignore the whole context or document. Language models were successfully applied in various tasks and are especially handy for example when dealing with unknown words in pos-tagging or in guessing/building the next word in automatic speech recognition. One of the famous algorithms used in this context was the so called Hidden Markov Models. With the time, the usage of artificial neural networks (ANN) would play a central role in NLP and Text Mining. Many ANN architectures were inspired by the traditional probabilistic language modelling. However, with the development and success of special architectures like Convolutional Neural Networks (CNN) or Long Term Short Term Memory (LSTM) a new approach emerged. Using many layers in (more or less) complex architectures with an embedding layer, delivered good results in solving many tasks. That is why, using some additional semantic information or word embeddings made sense and became popular. This again was an attempt to combine the context of documents and the order of words or other units. Despite the fact that this idea was around for some time, the real brake through came with the development of the word2vec algorithm by Mikolov [9,10]. All word2vec does is to generate vectors for words from a window with the help of a logistic regression (or a so called shallow network because it has only one hidden layer). Usually, this window is a small number like 5 words to the left and right of the word for which the vectors needs to be generated. The window in a way imitates a sentence and the created vectors capture the semantic representation of words. The success of wordvec was not only due to the fact that it could capture semantically related words well, but due to fact that the handy embedding vectors could be used for other tasks like text classification.

4.3 Topic Analysis with Deep Architectures

In our opinion the developments in deep learning had several impacts on topic modelling. First, using topics – especially probabilistic topic modelling – for search engines became less popular although it was the very thing that made topics popular. Instead, modern deep learning architectures seem to deliver better results in information retrieval tasks. The reason for that lies probably in the nature of the topic phenomenon. Term distributions as topics usually create a more general linguistic description of text collections, especially when word order does not matter. This might be great for text mining purposes in order to see what a text collection is about, however in a search one might want to receive more concrete results. In fact, the recent developments in DL based NLP widened the application and retrieval possibilities which are getting more sophisticated. They are meanwhile far beyond of search and find and resemble more human like communication. These are tasks like sentence completion, textual entailment, question answering etc. to name a few [15]. New models like BERT support these all functionalities [4]. The intuition behind it is that the searching of documents will be and is more like asking questions or beginning a sentence and hoping the search engine can complete it for you. And getting a much precise, concrete and narrowed down answer is priority.

While these developments are astonishing, they result in the fact that the questions asked also need to be concrete. The elegance of the classical topic modeling lies in the fact that one could understand a text collection even without having a question. It is the possibility to brake down many documents into few topics. However, there is also some debate on how interpretive word distributions are as topics, especially the ones which are generated by the bag-of-words approach [3]. Despite this fact, the expressive power of topics should be investigated and similar approaches further developed.

Reviewing the recent research one can observe two main approaches; the first one tries to combine algorithms and techniques from topic modeling with language technologies from ANN/DL research. Moody [11] tries for example to bring LDA and word verctors together. The second group of research tries to formulate the whole Bayesian inference process of topic modeling (like LDA) with the help of deep neural networks, also using of additional information or embedding layers [19].

5 Experiments

5.1 Preparing the Data

Text files in the dataset are delivered in two folders – `pdf_json` and `pmc_json`. Every paper resides inside one file with some metadata. First, every file was read, cleaned and appended into one file. Metadata about authors, date etc. was removed and only the content of json `text` element was kept. Also, some non unicode characters were removed as far as they would obviously confuse the analysis, especially the ones contained in the formulas or in the name of chemical elements. After cleaning the data the size of pure text dropped to under 10GB. Of course the link to the `metadata.csv` file which is additional table with the information such as author names, title, publishing data etc. was kept in order to be able to identify every paper.

In the experiments two prominent libraries – `gensim` [14] and `mallet` [8] – were used, both have implementions of above mentioned algorithms like LSA, PLSA and LDA. `Gensim` was implemented in python and has support of multi-core and even distributed computing for some of the algorithms. However, the data pre-processing pipeline supports only one core running which means it takes a lot of time for datasets bigger than 10GB. During our experiments we needed go back, in order to remove some unwanted stop words like *et al.* which appears a lot in the corpus. `Mallet` was written in Java. It is unfortunately not very well scalable, nonetheless it is very easy to use and produces well dependable results. It is also very well utilizing the type safety in the programming language Java. In dealing with big datasets both libraries require huge amount of memory, we recommend to use more than 50 GB.

5.2 Finding the Most Prominent Topics

We run LDA with the help of `mallet` and `gensim` libraries. Since the dataset is relatively huge, the number of topics was set to 400. It is wort mentioning that the output of the algorithm is twofold; $P(w_i|z_i)$ the probability distribution of each word in each topic and $P(z_i|d_i)$ the probability of each topic in each document:

$$\begin{array}{cc} & Topics \\ Words & \begin{bmatrix} P(w_1|z_1) & \cdots & P(w_1|z_n) \\ \vdots & \ddots & \vdots \\ P(w_n|z_1) & \cdots & P(w_n|z_n) \end{bmatrix} \end{array} \tag{8}$$

$$\begin{array}{cc} & \textit{Documents} \\ \textit{Topics} & \begin{bmatrix} P(z_1|d_1) & \cdots & P(z_1|d_n) \\ \vdots & \ddots & \vdots \\ P(z_n|d_1) & \cdots & P(z_n|d_n) \end{bmatrix} \end{array} \tag{9}$$

In both cases the result is a matrix where the values of rows in a column always add up to the number 1 because these are probabilities. This means the sum of the probabilities of all topics for every document is 1, however some topics for a document have higher probability than others. The first 15 topics and their respective keywords are shown in Table 1. We see from the topics that some are really related to infections like number 9 or studies to active cases like in 2 or restrictions put on schools like in 7.

Table 1. The first 15 topics from mallet

0	stroke cerebral brain ischemic eeg neurological epilepsy cognitive ich seizures seizure hemorrhage icp neurolo gy tbi mri intracranial acute scale outcome
1	cns demyelination disease myelin eae astrocytes spinal cord lesions brain oligodendrocytes demyelinating autoimmune microglia mbp sclerosis multiple encephalomyelitis matter tmev
2	patients results methods study years age conclusion months treatment patient cases clinical data therapy performed aim analysis background disease conclusions
3	food products consumption foods production meat safety consumers milk agricultural nutrition foodborne eating diet vegetables dietary agriculture fresh produce farmers
4	animal animals farms veterinary livestock farm production cattle sheep meat fmdv farmers disease control risk goats btv poultry veterinarians zoonotic
5	fig data number table values analysis observed shown results high average similar total calculated time based set higher compared range
6	article rights protected copyright reserved accepted reservedthe elsevier reserved.the reserved.accepted the this hrcs edx nicorandil gie reserved.in andthis b.v cecs
7	school household schools households closure closures students childcare teachers closed attack members home secondary reopening absenteeism classrooms children elementary classroom
8	opioid drug gambling overdose substance cocaine opioids vcp illicit cannabis injection addiction drugs buprenorphine methadone heroin reward abuse taar harm
9	fever infection infections cases measles illness hepatitis risk symptoms transmission infected blood transmitted days united exposure endemic children contact skin
10	infection hand hygiene mrsa infections control ipc nosocomial prevention compliance healthcare hospital hai ha is hospitals rates catheter practices vre patient
11	cells xbe patients expression response mbl human ifn-g cell immune results bacteria iga protein responses blood levels complement increased production
12	lasv lcmv arenavirus arenaviruses tfr junv gpc lassa macv fever stt world mopv thermometer pcn-dosed dbs hemor rhagic forehead mixing candid
13	n/a airbnb leprosy dot ulcerans bms bmp donkeys leprae globin nans-p ahr ecn hookworm ccdab hipab rental epz besnoitia rfhgst-s
14	health public care services system population medical national insurance coverage healthcare systems people private access prevention community service diseases social

In the Table 2 the 20 topics containing the term “rna” can be seen. These can be interpreted as topics from the papers which do a research on a rna vaccine or the sequencing of the genome of the virus. Whether all these topics can be accumulated to one, needs to be further investigated.

Table 2. The 20 topics containing the search phrase `"rna"`

41	viral virus protein proteins host replication viruses cell cells infection cellular rna genome infected membrane entry cycle interaction antiviral virions
66	hcv hbv hepatitis patients chronic liver genotype hbsag hcc therapy viral infection core svr dna rna treatment huh ribavirin weeks
107	frameshifting structure trna sequence frameshift pseudoknot stem base rna codon prf ribosome mrna ribosomal structures loop site frame sequences translation
128	sars-cov sars spike viral coronavirus anti-sars-cov sars-cov-infected coronaviruses epi_isl nucleocapsid vero pandemic mers-cov gisaid rna respiratory syndrome orf patient global
152	samples viral positive detection detected rna swabs sample specimens virus collected negative swab rt-pcr pcr load tested results sampling clinical
162	cells expression cell human growth protein gene proliferation mrna results increased role induced levels endothelial tissue mice receptor differentiation factor
177	rna replication genome rnas synthesis sequence viral transcription end genomic fig polymerase mrna strand region helicase nucleotides structure template subgenomic
187	tlr activation rig-i signaling innate immune irf mda mavs response dsrna trim rna tlrs ifn traf sting nlrp antiviral dna
206	ifn-l nmd upf tcv ifnlr sst tudor-sn pemv smg heo hou ifn-a/b cob elephants balb/cv sgrna pvx-gfp mmtv prokunina-olsson gfp-l
221	hpv cas editing cervical crispr ifi crispr-cas crispr/cas sgrna crrna crrnas target sgrnas hts types vrti eri lvs grna acrs
227	sirna mir sirnas rnai mirnas mirna target rna silencing gene expression mrna targeting antisense rnas dsrna sequence shrna dicer genes
233	india indian lncrnas lncrna delhi kerala states state till pradesh maharashtra bengal neat mumbai nrav west gujarat tamil nadu districts
239	mrna translation eif rna mrnas initiation rnase ires pkr splicing sgs stress utr cap transcripts cleavage ribosome ribosomal synthesis translational
251	plant mosaic tmv coat dsrna plants baculovirus tbsv insect yeast movement cpmv pvx leaves ctv benthamiana protoplasts baculoviruses orf symptoms
310	virus viruses viral rna family human dna genome genus species host acid nucleic capsid families hepatitis members group related infect
317	nsp orf activity rna plp nsps exon cap domain conserved replication mtase replicase nidoviruses nidovirus eav capping mhv cov complex
335	sars-cov coronavirus coronaviruses cov sars human virus respiratory mers-cov covs protein viruses spike viral ncov host humans rna receptor syndrome
337	pcr rna samples primers min primer kit rt-pcr dna performed reverse table reaction gene cdna xce/xbcl extracted usa positive study
347	rdrp polymerase nucleotide template rdrps incorporation importin-a ntp remdesivir motif polymerases thumb fidelity rna palm active fingers atp triphosphate motifs
366	bortezomib cml imatinib chl wolters kluwer hoct kir lymphoma survivors unauthorized abl lines mutation reproduction gltscr qol mrna ppl picts

One of the most interesting results of the mining process can be found in the Table 3. Here, the 15 most prominent topics in the entire collection are shown. We use the sum of the topics in the documents and we see here that the corpus is indeed representative for doing Covid-19 research. While the most prominent topic, 311 is about clinical cases the second most prominent topic, 117 focuses on the social measures done by the government such as lockdown. In the figure1

Table 3. The 15 most used topics

311	covid patients sars-cov severe disease infection respiratory acute risk clinical syndrome viral coronavirus reported mortality injury data ards higher severity
117	covid pandemic health people social measures lockdown public virus distancing spread coronavirus march countries due government outbreak world impact home
320	time case important make long number made fact part large point form means problem system general result good process small
300	results based study case considered number analysis order due important present studies specific method high table main data level information
5	fig data number table values analysis observed shown results high average similar total calculated time based set higher compared range
183	covid care pandemic staff patients time health patient clinical team services providers healthcare resources provide support access including crisis virtual
344	covid symptoms sars-cov cases patients infection disease respiratory coronavirus confirmed reported asymptomatic clinical fever positive infected severe virus case china
38	patients study patient hospital data clinical days table reported performed admission median included years time admitted group disease medical received
89	cells immune inflammatory response cytokines inflammation production cytokine activation macrophages levels role increased expression responses cell effects receptor shown disease
340	risk studies study reported found increased evidence factors data higher compared high increase recent significant potential disease important including exposure
336	e.g. potential provide including important studies multiple include critical approach impact systems current ability significant future understanding additional strategies specific
399	infection patients patient ppe transmission risk room contact equipment staff care protective isolation respiratory medical control procedures protection precautions workers
329	study age table higher found reported significant prevalence studies years data number compared analysis cases total group groups significantly differences
37	disease clinical diagnosis cases patients common acute treatment infection chronic include present severe syndrome therapy symptoms reported diagnostic including infections
114	global health development countries public national international community support policy local government systems capacity response including resources world research approach

the proportion of the most prominent topics are shown. The y-axis stands for the percentage cover of the topic in the collection. The most prominent topic 311 occupies 2% of the collection and with every other topic this number drops.

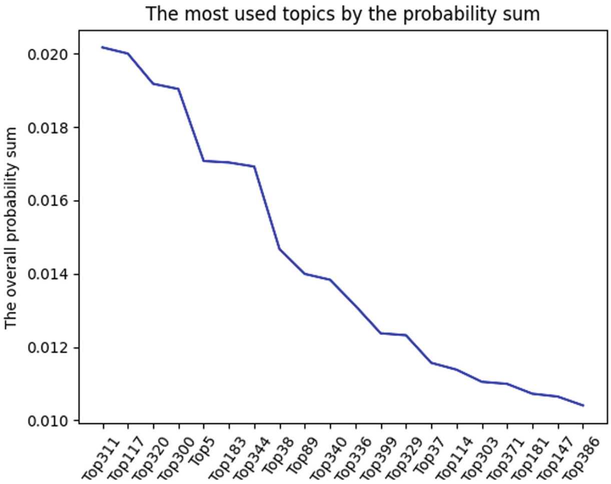

Fig. 1. The proportion of the most prominent topics in the dataset.

6 Conclusion

Almost three years passed since the Covid-19 pandemic hit the planet. Still, we fight against it and try to understand how our society can respond to it. In this paper we show that topic modeling can be very helpful in achieving this goal.

Topic modeling was gaining a lot of attention since LDA was first published in 2004. Despite the fact that the recent research on language technologies introduced many kinds of new algorithms, with BERT being the culmination, it is not clear, yet how these technologies can be optimally and efficiently used for topic extraction. BERT indeed outperforms many other existing algorithms in tasks like sentiment analysis, question answering or just semantic search querying, because it creates a very detailed and contextualized language representation. However, when we do topic modeling, we do not ask a specific question to our corpus, but we try to extract clear and easy understandable information about its content. LDA delivers more or less understandable results where topics created from the dataset are abstract enough for a quick distant reading and at the same time concrete enough, in order to capture the slight semantic differences in the topics or documents. In the future, we want to investigate how topics can be extracted and presented in a similar manner with the help of newer algorithms.

References

1. Bhatia, P., et al.: AWS CORD-19 search: a neural search engine for COVID-19 literature. http://arxiv.org/abs/2007.09186
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

3. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Proceedings of the 22nd International Conference on Neural Information Processing Systems, pp. 288–296. NIPS 2009, Curran Associates Inc., Red Hook, NY, USA (2009)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2018). https://doi.org/10.48550/ARXIV.1810.04805, https://arxiv.org/abs/1810.04805
5. Dumais, S.T.: Latent semantic analysis. Ann. Rev. Inf. Sci. Technol. **38**, 188–230 (2005)
6. Hofmann, T.: Probabilistic latent semantic analysis. CoRR abs/1301.6705 (2013). http://arxiv.org/abs/1301.6705
7. Karami, A., Bookstaver, B., Nolan, M.S., Bozorgi, P.: Investigating diseases and chemicals in Covid-19 literature with text mining. Int. J. Inf. Manag. Data Insights **1**, 100016–100016 (2021)
8. McCallum, A.K.: Mallet: a machine learning for language toolkit (2002). http://mallet.cs.umass.edu
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, 2–4 May 2013, Workshop Track Proceedings (2013). http://arxiv.org/abs/1301.3781
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. CoRR abs/1310.4546 (2013). http://arxiv.org/abs/1310.4546
11. Moody, C.E.: Mixing Dirichlet topic models and word embeddings to make LDA2vec. CoRR abs/1605.02019 (2016). http://arxiv.org/abs/1605.02019
12. Nogueira, R., Yang, W., Lin, J., Cho, K.: Document expansion by query prediction. arXiv preprint arXiv:1904.08375 (2019)
13. Otmakhova, Y., Verspoor, K., Baldwin, T., Šuster, S.: Improved topic representations of medical documents to assist COVID-19 literature exploration. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.nlpcovid19-2.12, https://www.aclweb.org/anthology/2020.nlpcovid19-2.12
14. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, Malta, May 2010. http://is.muni.cz/publication/884893/en
15. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: a multi-task benchmark and analysis platform for natural language understanding (2018). https://doi.org/10.48550/ARXIV.1804.07461, https://arxiv.org/abs/1804.07461
16. Wang, L.L., et al.: Cord-19: The COVID-19 open research dataset. ArXiv (2020)
17. Zhai, C., Massung, S.: Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, 1st edn. ACM Books, San Rafael (2016). OCLC: ocn957355971

18. Zhang, E., et al.: Covidex: neural ranking models and keyword search infrastructure for the COVID-19 open research dataset. In: Proceedings of the First Workshop on Scholarly Document Processing, pp. 31–41. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.sdp-1.5, https://www.aclweb.org/anthology/2020.sdp-1.5
19. Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., Buntine, W.: Topic modelling meets deep neural networks: a survey (2021). https://doi.org/10.48550/ARXIV.2103.00498, https://arxiv.org/abs/2103.00498