ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse231942020.pdf
https://doi.org/10.30534/ijatcse/2020/231942020

Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus

Sara Mifrah[1], El Habib Benlahmar[2]
[1]Laboratory of Information Processing and Modeling, Hassan II University of Casablanca, Faculty of Sciences Ben M'sik, Casablanca, Morocco, mifrah.sara@gmail.com
[2]Laboratory of Information Processing and Modeling, Hassan II University of Casablanca, Faculty of Sciences Ben Msik Casablanca, Morocco, h.benlahmer@gmail.com

ABSTRACT

Topic modeling is a method for finding abstract topics in a large collection of documents. With it, it is possible to discover the mixture of hidden or "latent" topics that varies from document to document in a given corpus. As an unsupervised machine learning approach, topic models are not easy to evaluate since there is no labelled "ground truth" data to compare with. However, since topic modeling typically requires defining some parameters beforehand (first and foremost the number of topics k to be discovered), model evaluation is crucial in order to find an "optimal" set of parameters for the given data. Latent Dirichlet allocation (LDA) and Non-Negative Matrix Factorization (NMF) are the two most popular topic modeling techniques. LDA uses a probabilistic approach where as NMF uses matrix factorization approach. In this paper we want to assess which most relevant technique for topic coherence using c_v measure, we have chosen citations's Covid'19 Corpus for experimentations.

Key words: Topic Model, Topics Coherence, Machine Learning, LDA, NMF.

1. INTRODUCTION

Topic models learn topics (sets of words) automatically from unlabeled documents [14] [12] in an unsupervised way. This is an attractive method to bring structure to otherwise unstructured text data, but Topics are not guaranteed to be well interpretable, therefore, coherence measures have been proposed to distinguish between good and bad topics.

When we use a subject model (topic model), we are mainly concerned with the extent to which the subjects learned correspond to human judgments and help us to differentiate ideas. But until recently, the evaluation of these models has custom and application specific. Ratings ranged from substantially fully automated evaluations to manually designed external evaluations. Previous external evaluations used acquired subjects to represent small fixed vocabulary and to compare this distributive area with human judgments of similarity [1] [5] [13]. However, these reviews are hand-created and often costly to perform industry-specific topics. On the contrary, the intrinsic metrics evaluated the amount of information coded by topic, where confusion is a common example [2], however, [3] found that these essential metrics do not always relate to semantic explainable topics. Moreover, a few evaluations used the same metrics to compare distinct approaches like Latent Dirichlet Allocation (LDA) [15] [4] and Non-Negative Matrix Factors (NMF) [6]. This affected the knowledge of the most beneficial method for a given application, or in terms of extracting useful topics.

This paper is structured as follows: Section 2 presents a small description of the two models. Section 3 define different coherence measures, our experimentations and results presented in section 4 and 5 respectively, and finally a conclusions of this paper are provided in Section. 6.

2. TOPIC MODELS

2.1 LDA

The latent Dirichlet allocation [4][7] (LDA) is a thematic probabilistic modeling algorithm. It is based on the assumption that documents are composed of several themes (and not words), where a theme is a multinomial distribution on a fixed vocabulary W.

LDA is a mixture model that captures the exchangeability of both words and documents [4]. The assumption of exchangeability for words in a document means that the order of words in a document is not important, and likewise for the ordering of documents in a corpus.

The following are definitions of several terms:

- **Corpus**: a collection of M documents
- **Document**: a sequence of N words

- **Word**: the basic unit of discrete data represented by a unit-basis vector vocabulary indexed from 1 to V where one component is equal to 1 and all others are 0.

Recall that LDA is a generative model. Each document is assumed to be generated by the following generative process, where words are generated independently of other words, hence following a unigram bag-of-words model [8]:

To generate a document:

1. Randomly choose a distribution over topics (a distribution over a distribution).
2. For each word in a document:
   a. Randomly choose a topic from the distribution over topics.
   b. Randomly choose a word from the corresponding topic (distribution over the vocabulary).

More formally, the generative process finds the joint distribution of the hidden and observed variables [8]:

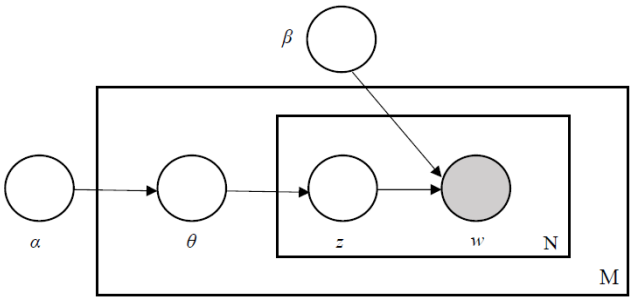

Figure 1 : LDA graphical representation

The variables θ and z are of major interest in that they prioritize the model and evaluate each component of the corpus (whole corpus, document, words). We distinguish in effect three distinct "levels" in (Fig 1):

- The "hyper parameters" **α** and **β** are global parameters defined for a entire corpus C; the other variables are generated from them
- **θ** is defined for a document; its probability density is expressed from properties of Dirichlet's law:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\dots\theta_K^{\alpha_K-1}|$$

- Finally **z** and **w** are specific to each word. We give the joint probability of a sequence of words w and a sequence of topic z:

$$p(w,z) = \int p(\theta)\left(\prod_{n=1}^{N} p(z_n|\theta)\,p(w_n|z_n)\right)d\theta$$

Using the Gibbs Sampling estimate, we obtain the probability matrix θ [document - subject] and the probability matrix Φ [subject - word]. For a new document of arbitrary length, we can deduce its latent subjects involved and in the meantime, we will assign a subject label for each word in the document.

2.2 NMF

Non-negative matrix factorization (NMF) is a linear-algebraic optimization algorithm [9] used for dimensionality reduction and data analysis [10] that solves the following problem (illustrated in Fig 2, which is taken from [11]):

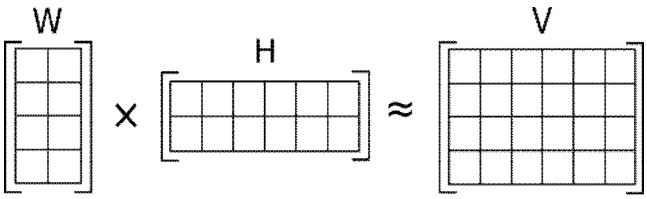

Figure 2: Non-negative matrix factorization diagram

Given a non-negative matrix **V**, find (usually) two non-negative matrices W and H such that:

$$V \approx WH$$

Thus, given a set of multivariate n dimensional data vectors, they are put into an n*m matrix V as its columns, where m is the number of examples in the data set. This matrix V is approximately factorized into an n*t matrix W and an t*m matrix H, where t is generally less than n or m. Hence, this results in a compression of the original data matrix.

In terms of topic modeling, the input document-term matrix V is factorized into a n*t document-topic matrix and a t*m topic-term matrix, where t is the number of topics produced.

Non-negativity is natural for many practical problems, such as color intensities, frequency counts, etc., and it also makes the resulting matrices easier to inspect [11]

3. COHERENCE MEASURES

A set of statements or facts is said to be coherent, if they support each other. Thus, a coherent fact set can be interpreted in a context that covers all or most of the facts. An example of a coherent fact set is "the game is a team sport",

"the game is played with a ball", "the game demands great physical efforts".

Let's take quick look at different coherence measures, and how they are calculated:

- C_v measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity
- C_p is based on a sliding window, one-preceding segmentation of the top words and the confirmation measure of Fitelson's coherence
- C_uci measure is based on a sliding window and the pointwise mutual information (PMI) of all word pairs of the given top words
- C_umass is based on document cooccurrence counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measure
- C_npmi is an enhanced version of the C_uci coherence using the normalized pointwise mutual information (NPMI)
- C_a is baseed on a context window, a pairwise comparison of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity

There is, of course, a lot more to the concept of topic model evaluation, and the coherence measure. However, keeping in mind the length, and purpose of this article, let's apply these concepts into developing a model that is at least better than with the default parameters.

4. EXPERIMENTATION

4.1 *Data Set*

In this study we have used a Corpus of Covid'19 Citations (2019-2020). This corpus consists of texts that were released as part of the COVID-19 Open Research Dataset (CORD-19) - from https://www.sketchengine.eu/covid19/ - . We are constructing our own corpus, we are selecting 13 000 Citations of Covid'19, structured them in a csv file which has the following form "Id; papers_source; url; citations".

4.2 *Pre-processing*

Preprocessing is an important task and critical step in Text mining, Natural Language Processing (NLP) and Information Retrieval (IR). Data preprocessing is used for extracting interesting and non-trivial and knowledge from unstructured text data.

In Preprocessing, a) tokenization is the procedure of splitting a text into words, phrases, or other meaningful parts, namely tokens. In other words, tokenization is a form of text segmentation. Typically, the segmentation is carried out considering only alphabetic or alphanumeric characters that are delimited by non-alphanumeric characters (e.g., punctuations, whitespace). b) Stop-words are the words that are commonly encountered in texts without dependency to a particular topic (e.g., conjunctions, prepositions, articles, etc.). c) Another widely used preprocessing step is lowercase conversion. Since uppercase or lowercase forms of words are assumed to have no difference, all uppercase characters are usually converted to their lowercase forms. d) The aim of stemming is to obtain stem, or root, forms of derived words. Since derived words are semantically similar to their root forms, word occurrences are usually computed after applying the stemming on a given text.

5. RESULTS

After the preprocessing phase we tried to extract 10 topics for each model, LDA and NMF, then calculated the consistency of the topics of the two models and finally analyzed the results obtained.

5.1 LDA Topics and Coherence of Topics

Table 1a: LDA Topics List

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
protein	transmission	use	cell	sequence
bind	use	test	protein	virus
structure	human	figure	receptor	coronavirus
may	virus	table	ace	bat
fig	evidence	epitope	bind	genome
affinity	model	sequence	entry	share
antibody	estimate	result	target	human
human	base	positive	show	identity
region	animal	detection	virus	identify
analysis	study	sample	also	high

Table 1b: LDA Topics List

Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
severe	treatment	level	case	patient
case	disease	gene	patient	infection
cause	virus	induce	confirm	study
report	infection	activity	test	infect
infection	control	design	hospital	clinical
spread	vaccine	significantly	suspect	symptom
acute	drug	peptide	positive	covid
respiratory	may	concentration	infection	day
human	effective	construct	virus	report
virus	cause	complex	local	respiratory

After extracting all topics we are calculating the coherences value of each topic using c_v measure and we have obtained the graph bellow (fig 3).

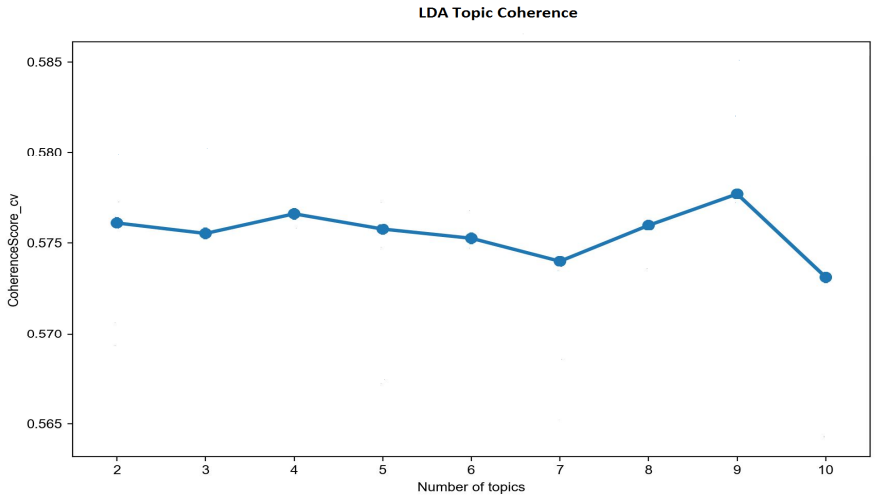

Figure 3: LDA Coherence Score with c_v mesure

5.2 NMF Topics and Coherence of Topics

Table 2a: NMF Topics List

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
2019	ace2	mers	patients	preprint
china	receptor	2012	infection	copyright
coronavirus	protein	saudi	sars	holder
wuhan	Sars	east	infected	1101
novel	Cells	middle	transmission	https
december	binding	arabia	cases	peer
outbreak	Rbd	first	clinical	reviewed
health	spike	respiratory	symptoms	doi
2020	human	cases	study	10
disease	Entry	reported	confirmed	2020

Table 2b: NMF Topics List

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
respiratory	Sars	pcr	19	al
syndrome	Bat	rt	covid	et
coronavirus	sequence	positive	disease	zaki
severe	genome	rna	caused	2012
acute	coronaviruses	time	pandemic	2020
disease	ratg13	real	sars	de
caused	identity	tested	virus	2003
middle	Virus	samples	outbreak	2005
east	Covs	detection	coronavirus	2013
novel	Sequences	swab	world	2004

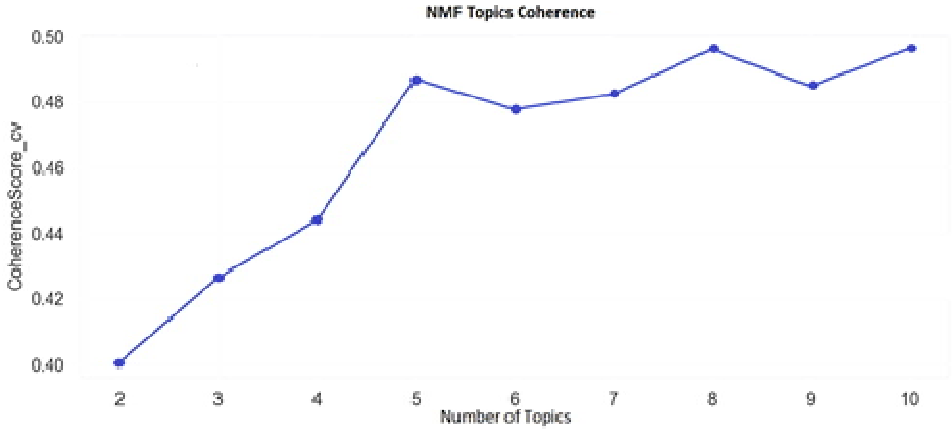

Figure 4: NMF Coherence Score with c_v mesure

5.2 Comparison

If we want to compare the two models, we can say that the coherence of topics generated by the LDA model (fig 3) is better than that generated by the NMF model (fig 4); moreover the words of each topic (Table 1a and Table 1b) for the LDA model is more significant than that of the NMF model (Table 2a and Table 2b) globally; but if we look at each topic we can see that: (the Topics 1,4,6,7,9) of NMF model are the best.

6. CONCLUSION

We can conclude that the LDA model is more relevant than the NMF model in the case of large corpus -13 000 citations - with long document or texts; in generation of coherence of topics.

Overall, we see that each topic model paradigm has its own strengths and weaknesses. Latent Dirichlet Allocation learn concise and coherent topics and achieved similar performance on our evaluations. However, NMF learns more incoherent topics than LDA. For applications in which a human end-user will interact with learned topics, the flexibility of LDA and the coherence advantages of LDA warrant strong consideration.

REFERENCES

1. Jurgens, D., & Stevens, K. "**The S-Space package: An open source package for word space models** ".proceedings of the ACL 2010 systems demonstrations, Uppsala, Sweden. 2010
2. Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno.." **Evaluation methods for topic models**". In Proceedings of the 26th International

Conference on Machine Learning (ICML). Omnipress. 2009

3. J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber and D. M. Blei "**Reading tea leaves: How humans interpret topic models**", Proc. Adv. Neural Inf. Process. Syst., pp. 288-296. 2009 https://doi.org/10.1145/1553374.1553515
4. D. M. Blei, A. Y. Ng and M. I. Jordan , **"Latent Dirichlet allocation**", Proc. J. Mach. Learn. Res., pp. 993-1022. 2003
5. Sara Mifrah and El Habib Benlahmar. **Semantic Relationship Study between Citing and Cited Scientific Articles Using Topic Modeling**. In Proceedings of the 4th International Conference on Big Data and Internet of Things (BDIoT'19), Rabat Morocco. 2019 doi: https://doi.org/10.1145/3372938.3372943
6. Daniel D. Lee and H. Sebastian Seung. "**Algorithms for non-negative matrix factorization**". In In NIPS, pages 556–562. MIT Press. 2000
7. T. L. Griffiths and M. Steyvers. "**Finding scientific topics**". Proceedings of the National Academy of Sciences, 101(Suppl. 1):5228–5235. 2004 https://doi.org/10.1073/pnas.0307752101
8. S. Clark . **Topic Modelling and Latent Dirichlet Allocation.** Available at "https://www.cl.cam.ac.uk/teaching/1213/L101/clark_lectures/lect7.pdf" 2013
9. P. Suri and N. R. Roy "**Comparison between LDA & NMF for event-detection from large text stream data**," 3rd International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, pp. 1-5, doi: 10.1109/CIACT.2017.7977281. 2017
10. .S. Dhillon and S. Sra **Generalized Nonnegative Matrix Approximations with Bregman Divergences**. Availibale at http://papers.nips.cc/paper/2757-generalized-nonnegative-matrix-approximations-with-bregman-divergences.pdf 2006
11. https://en.wikipedia.org/wiki/Non-negative_matrix_factorization
12. Mifrah, Sara., Benlahmar, El Habib. **Semantico-automatic Evaluation of Scientific Papers: State of the Art**. BDCA'17 Proceedings of the 2nd international Conference on Big Data, Cloud and Applications Tetouan, Morocco — March 29 – 30 2017, ISBN: 978-1-4503-4852-2 DOI: 10.1145/3090354.3090380
13. Hourrane, Oumaima, et al. "**Using Deep Learning Word Embeddings for Citations Similarity in Academic Papers**." International Conference on Big Data, Cloud and Applications. Springer, Cham 2018. DOI: 10.1007/978-3-319-96292-4_15
14. Sara Mifrah, El Habib Benlahmar, Youssef Mifrah and Mohamed Ezeouati. "**Toward a Semantic Graph of Scientific Publications: A Bibliometric Study**". International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May -June 2020, pp: 3323 -3330. DOI: 10.30534/ijatcse/2020/12993202
15. Deepti Sehrawat and Nasib Singh Gill "**Review and Comparative Analysis of Topic Identification Techniques**". International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.3, May - June 2019 DOI : 10.30534/ijatcse/2019/71832019