An Application of Non Negative Matrix Factorization in Text Mining

Nguyen Bao Tran(✉), Thanh Son Huynh, Ba Lam To, and Luong Anh Tuan Nguyen

Information Technology Department, Vietnam Aviation Academy, Ho Chi Minh City 72200, Vietnam

{baotn,sonth,lamtb,nlatuan}@vaa.edu.vn

Abstract. The field of text mining has increasingly relied on Non-negative matrix factorization (NMF) for its ability to perform high-dimensional data reduction and visualization. This paper aims to employ NMF in analyzing a dataset of 1,500 documents and 12,419 words in bags-of-words format, obtained from the UCI Machine Learning Repository. Our analysis demonstrates the utility of NMF in effectively classifying ambiguous and sparse textual data into distinct topics and extracting meaningful contents through the identification of relevant keywords. Further, we demonstrate the robustness of NMF in topic clustering by exploring the semantic relationship between extracted keywords and the topics to which they belonged. Our findings offered valuable insights into the application of NMF in text mining and suggested that universities in Vietnam could leverage this technique to analyze feedback and suggestions from students.

Keywords: NMF · text mining · topic classification · bags-of-words

1 Introduction

The growth of high-dimensional data has led to an increased need for advanced techniques to extract and derive valuable information from large volumes of data [1]. In data mining, the main objectives of such techniques are generally focused on reducing high-dimensionality while preserving the majority of original information and clustering and interpreting underlying features from the original data, which can enable the exploitation of valuable knowledge and information thereafter [2]. Accordingly, various techniques based on data decomposition, such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), Support Vector Machine (SVM), and Non-negative Matrix Factorization (NMF), have been utilized in data mining to extract pertinent information. Of these techniques, NMF has demonstrated robustness for non-negative data, which are naturally present in real-life datasets such as the number of pixels in an image, the number of occurrences of each word, and stock prices. The constraint of non-negative data in decomposed matrices leads to the part-based representation of NMF and improvement of interpretability [3, 4]. Therefore, NMF may be better suited for non-negative data and part-based representation compared to other techniques.

N.-S. Vo and H.-A. Tran (Eds.): INISCOM 2023, LNICST 531, pp. 287–295, 2023.
https://doi.org/10.1007/978-3-031-47359-3_21

NMF gained popularity after a seminal paper by Lee and Seung [5], which proposed an algorithm and highlighted the advantages of NMF, such as straightforward interpretability and potential ability of part-based representation. Indeed, NMF can effectively reduce the dimensionality of data with minimal information loss by approximately extracting high-dimensional data into low-rank matrices. Additionally, NMF automatically clusters the original data by sparse and meaningful features, which enables intuitive visualization of hidden correlations. As a result, NMF is a prominent tool to represent and factorize non-negative datasets in various fields, including signal processing, biomedical engineering, text mining, image processing, and more [6].

Due to the non-negative characteristic of word-document frequency of occurrence and NMF's automatic clustering abilities, NMF has proven effective and well-suited to analyze semantic and topic modeling in text mining [7]. Text mining can apply various applications, such as analyzing customer behaviors, conducting market research, and filtering malicious or spam content to enhance business performance [8]. Categorizing, interpreting, and discovering underlying features, word connections, and knowledge from a vast volume of text collections from document corpuses are considered as essential tasks in text mining. Topic clustering based on semantic might be the most prominent approach to text mining. To process data, text collections must undergo pre-processing, including tokenization and stop-word elimination under bags-of-words format. Then, NMF can automatically cluster meaningful topics by combining attributes in the two extracted matrices, resulting in low-dimensional data presentation. Given that the amount of textual data collections is growing larger and larger, such as on social network platforms (e.g., Facebook, Twitter, and Instagram) and email systems (Gmail and Yahoo), NMF applications in text mining are promising to uncover latent components and hidden topics of textual data by part-based representation.

In the context of university management, understanding students' viewpoints, reactions, and sentiments plays a vital role in enhancing educational quality and school facilities. The vast amount of feedback and comments sent from social media (e.g., Facebook and Twitter) and emails over the years should be made more compact, understandable, and intuitive to visualize, instead of becoming a burden for managers and staff. Accordingly, NMF can give the benefits of extracting the most concerned issues of students by topic clustering. In Vietnam, there are few studies in text mining that utilize NMF in text mining for education management in universities. Therefore, this paper aims to demonstrate the robustness of NMF in text mining by conducting an experiment from a data corpus. From the ambiguous and unstructured textual data, NMF might effectively cluster meaningful topics as well as gain insights into the underlying connections in terms of semantic between keywords within the clustered topics.

This study was structured into four main sections: Introduction, Methodology, Experiment, and Conclusions. The Introduction section presented an overview of the robustness of Non-negative Matrix Factorization (NMF) and its applications in text mining. In the Methodology section, a text mining process was proposed, which detailed the application of the NMF algorithm for topic clustering. The Experiment section provided a description of the dataset used, the experimental procedures employed for preprocessing the raw data, and the experimental results obtained. Finally, the Conclusions section outlined the significance of the study.

2 Methodology

2.1 Text Mining Process

In this study, a text mining process was proposed, which involved the application of the NMF algorithm (see Fig. 1). The process began with the pre-processing of raw data, which involved tokenization and stopword removal, and was referred to as the "Data Pre-processing" step in Fig. 1. The pre-processed data was then transformed into a bags-of-words format and transferred to a 2-dimensional matrix (n documents x m words), which served as the observed data, referred to as the "Text Transformation" step. The NMF algorithm was then applied to decompose the observed data into matrices, referred to as the "Applying NMF" step. Finally, the extracted topics accompanied by keywords within each topic were presented in the form of decomposed matrices, which constituted the "Text Mining" step.

Fig. 1. Proposed text mining process.

2.2 Non-negative Matrix Factorization Algorithm

From the algebraic perspective, NMF is relevant to matrix decomposition. Given a dataset consisting of non-negative elements, the NMF algorithm aims to approximate it as the product of two low-rank matrices, as shown in the following equation:

$$A \cong W \times H \tag{1}$$

where A is the observed data (n x m), which n dimensions and m data points of each dimension, W is the basic matrix (n x k), and H is the coefficient matrix (k x m). Each data point in the A matrix can be approximated by a linear combination of the rows of W and the columns of H. The columns of W can be interpreted as basis vectors or "building blocks," while the rows of H present the coordinates of data points of W, used to approximately reconstruct the observed data.

The matrices W and H are initially randomized, and then determined through iterative updates based on the convergence of local optimal matrix factorization. The value of k ranges from 1 to the minimum of n and m. In other words, the original n dimensions are reduced to k.

Since NMF approximates the A matrix, we could evaluate the fit-goodness between the reconstructed and observed data. The reconstructed matrix R = W x H is calculated as follows:

$$R = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1k} \\ w_{21} & w_{22} & \dots & w_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nk} \end{bmatrix} \times \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1m} \\ h_{21} & h_{22} & \dots & h_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ h_{k1} & h_{k2} & \dots & h_{km} \end{bmatrix} = W \times H \tag{2}$$

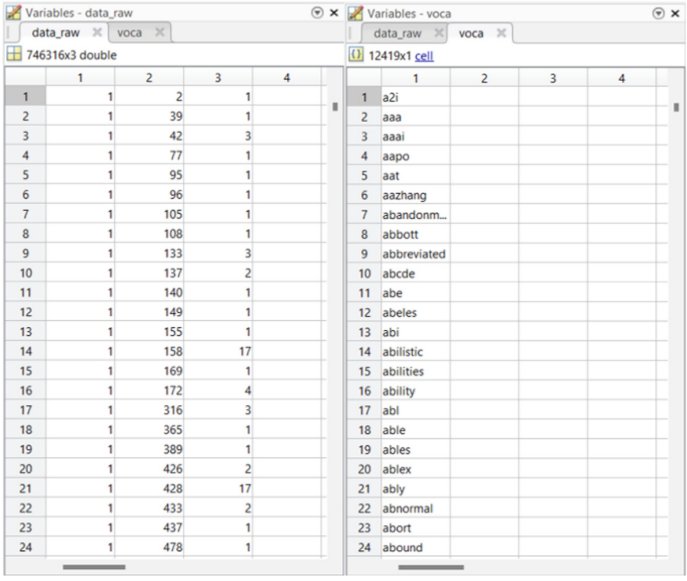
Values in W and H are updated and terminated based on the minimum least-squares error (E) optimization between the original (A) and reconstructed (R) matrices.

$$E = \min \|D - R\|_F^2 \tag{3}$$

3 Experiment

3.1 Data Set

For the experiment, we utilized a textual dataset obtained from the Neural Information Processing Systems (NIPS), a highly regarded machine learning conference. The dataset was obtained from the UCI Machine Learning Repository [9]. After initial processing steps such as tokenization and removal of stopwords, the dataset was transformed into bags-of-words format, and consisted of 12,419 words (n) in a vocabulary list and 1,500 documents (m), with approximately 2 million words in the colletion in total. Due to copyright restrictions, the name of each document was not provided, resulting in an untagged and ambiguous dataset, which was suitable for the purpose of topic clustering.



	1	2	3	4
1	1	2	1	
2	1	39	1	
3	1	42	3	
4	1	77	1	
5	1	95	1	
6	1	96	1	
7	1	105	1	
8	1	108	1	
9	1	133	3	
10	1	137	2	
11	1	140	1	
12	1	149	1	
13	1	155	1	
14	1	158	17	
15	1	169	1	
16	1	172	4	
17	1	316	3	
18	1	365	1	
19	1	389	1	
20	1	426	2	
21	1	428	17	
22	1	433	2	
23	1	437	1	
24	1	478	1	

	1	2	3	4
1	a2i			
2	aaa			
3	aaai			
4	aapo			
5	aat			
6	aazhang			
7	abandonm...			
8	abbott			
9	abbreviated			
10	abcde			
11	abe			
12	abeles			
13	abi			
14	abilistic			
15	abilities			
16	ability			
17	abl			
18	able			
19	ables			
20	ablex			
21	ably			
22	abnormal			
23	abort			
24	abound			

Fig. 2. The data obtained from NIPS and stored in Matlab. The left table presents the frequency of occurrence of each word in a single document. The right table presents the vocabulary list of 12,419 words.

3.2 Experimental Procedures

The raw data consisted of the vocabulary list of 12,419 words and the frequency of occurrence of each word in a single document. The latter was presented in the form

of [docID wordID count], as illustrated in Fig. 2. For instance, the first line [1 2 1] indicates that the second word (wordID = 2, "aaa" – obtained to the vocabulary list) occurs once in the first document (docID = 1). From the raw data, we constructed a word-document matrix A of dimensions 12,419 x 1,500 to store and process information using the TF-IDF (Term Frequency-Inverse Document Frequency) normalization method. For example, A[1, 2] = 1 means that the second word ID (wordID = 2) appears once in the first document (docID = 1).

Following the application of Non-negative Matrix Factorization (NMF), we obtained two decomposed matrices, W and H. W is the basic matrix (12,419 words x k topics), while H is the coefficient matrix (k topics x 1,500 documents). We normalized matrix W by the maximum value in each column, and matrix H by the maximum value in each row. W and H provided insights into the initial word-document matrix, which lacked any class labels and was seemingly chaotic.

Each column of matrix W represented a basic vector, where higher values (from 0 to 1) indicated greater significance of a word (or term) in a topic, as well as a high degree of semantic association with other words in the same topic. We selected the top 20 most important words, which were corresponding to the 20 largest values in each column of W as representatives, to investigate the content of each topic. Meanwhile, values in each row of matrix H (i x j) indicated the extent to which document j belonged to topic i. A high value (from 0 to 1) indicated that document j was strongly associated with topic i.

The number of topics (k) was determined based on the semantic relationships and combined attributes between words in matrix W. After matrix decomposition, we obtained the following information: (1) the number of clustered topics, (2) the most significant words in each topic (from matrix W), and (3) the topic to which a document belonged (from matrix H).

To evaluate the performance of NMF, we analyzed the relationship between the semantic of 30 words that appeared most frequently in a document and the semantic of the top 20 most important words in the topic to which the document belonged. We selected five documents that belonged to the five extracted topics as representatives to confirm the robustness of NMF in terms of semantic similarity between keywords in the documents and topics. A clear explanation of the relationship between the most frequent words in a document and the keywords in the topic to which the document belonged confirmed the efficacy of NMF.

We used Matlab (Mathworks Inc.) to perform all kinds of data processing.

3.3 Experimental Results

Due to the relatively large and extremely sparse nature of the data in bags-of-words format, the coefficient of determination between the original and reconstructed data was found to be low. For instance, when we chose the number of topics was 100 topics, which was too large and unrealistic, the r^2 value was still relatively low, i.e. 0.41. As such, the selection of the appropriate value of the number of topics (k) was based on semantic similarity, rather than r^2. Specifically, the value of k was reduced until the topics were distinguishable, while maintaining semantic coherence. Ultimately, a value of $k = 5$ was chosen, resulting in a mean coefficient of determination (r^2) of 0.27.

Upon completion of the matrix decomposition, the basic matrix W was obtained with dimensions of 12,419 words by 5 topics. From this matrix, the 20 most important words were selected to represent each of the topics. The resulting five topics (A to E) are presented in Fig. 3, along with their respective most important words. Topic A seemed to focus on neural network architectures, such as inputs, outputs, training set, and the number of hidden layers to enhance performance. Meanwhile, Topic B was highly related to functions and algorithms, which focused on parameters such as the number of classes, weights, and training data sets. Topic C was relevant to models using control systems for speech or image recognition, such as the Hidden Markov model (with the keyword "hmm"), which can be controlled or modified by parameters and methods related to Gaussian probability distribution. Topic D appeared to describe the processing to neural information of the biological brain, such as synaptic responses, spike signals, firing, circuits, or stimulus. Finally, Topic E appears to be related to reinforcement learning with specific keywords related to learning algorithm as policiy, rule, states, and action, which enables to obtain an optimal controller applied for robot.

	A	B	C	D	E
1	network	function	model	neuron	learning
2	unit	algorithm	data	cell	action
3	neural	set	parameter	input	system
4	input	data	system	visual	control
5	training	error	distribution	system	algorithm
6	output	vector	object	circuit	task
7	weight	problem	recognition	function	reinforcement
8	hidden	training	set	response	policy
9	system	method	likelihood	pattern	function
10	set	number	mixture	synaptic	model
11	layer	result	gaussian	signal	dynamic
12	error	weight	image	output	problem
13	pattern	distribution	control	firing	learn
14	net	point	point	field	optimal
15	recognition	linear	method	direction	step
16	performance	parameter	neural	spike	states
17	trained	learning	speech	neural	robot
18	number	bound	hmm	activity	controller
19	architecture	case	probability	current	learned
20	problem	class	word	stimulus	rule

Fig. 3. The data of 5 extracted topics (5 columns) and the most 20 important words (20 rows) in each topic.

The coefficient matrix H, consisting of 5 topics and 1500 documents, facilitated the determination of a document's topic by identifying the highest value within a given column. Figure 4 presents three subsets of 15 representative documents each, specifically

docID = 1 to 5, docID = 501 to 505, and docID = 1000 to 1005, which were used to represent the entire set of 1500 documents.

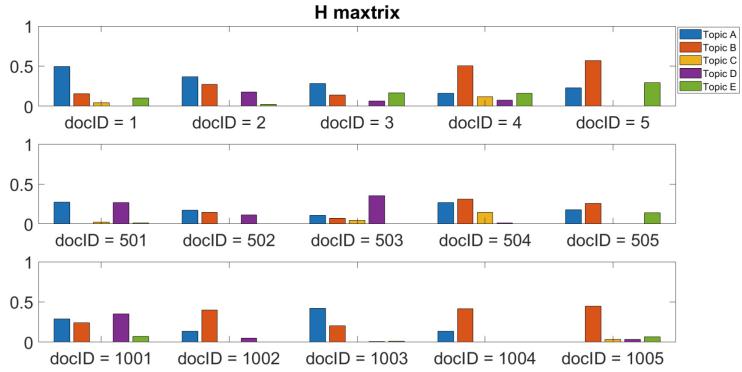


Fig. 4. The coefficient matrix H of 15 documents (docID from 1 to 5, 501 to 505, and 1000 to 1005) as representatives. The highest value of a bar means that document belonged to the topic.

As can be seen in Fig. 4, we could determine a document belonged to a certain topic based on the highest value of a column in the coefficient matrix H. For example, the first column of matrix H (docID = 1, see Fig. 4), the highest value was the first bar, which implied the first document (docID = 1) belonged to topic A.

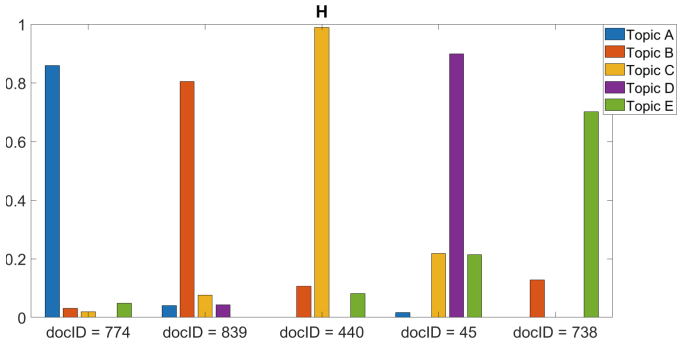


Fig. 5. Five selected documents which were apparently belonged to 5 topics.

To assess the efficacy of Non-negative Matrix Factorization (NMF) in topic clustering, we chose five documents that clearly pertained to the five topics (i.e., those with a high value in matrix H, approaching unity) as representatives. As depicted in Fig. 5, the documents with docIDs of 774, 839, 440, 45, and 738 belonged to topics A, B, C, D, and E, respectively. Subsequently, we examined the semantic association between the 30 most frequent words in a given document and the top 20 important words in the corresponding topic to which the document belonged, as can be seen in Fig. 3 and Fig. 6.

Firstly, with regards to the document with a docID of 774, which belonged to topic A (Fig. 5), it was apparent that the top 30 most frequent words and the top 20 most important

	docID = 774	docID = 839	docID = 440	docID = 45	docID = 738
1	**network**	**function**	**model**	**cell**	skill
2	order	**error**	merging	potential	**learning**
3	**output**	**bound**	**hmm**	membrane	**task**
4	recurrent	rate	transition	light	**action**
5	**neural**	**training**	states	**firing**	**reinforcement**
6	**number**	**problem**	**data**	**neuron**	loss
7	**problem**	cross	prior	learning	policies
8	**weight**	sample	**probability**	**model**	**function**
9	finite	validation	baum	ganglion	**algorithm**
10	**input**	plot	algorithm	**stimulus**	domain
11	**layer**	generalization	**likelihood**	network	pay
12	machine	target	samples	alkon	description
13	states	approximation	welch	associative	length
14	multilayer	complexity	**parameter**	effect	performance
15	tdnn	estimation	number	**input**	**policy**
16	comparison	behavior	sample	control	order
17	experiment	model	posterior	discharge	**problem**
18	high	input	markov	hair	environment
19	nodes	structure	path	mechanism	learner
20	gamma	log	initial	**response**	**states**
21	node	theory	learning	threshold	denoted
22	**training**	**data**	size	background	multiple
23	result	examples	bayesian	exposure	paper
24	**set**	interval	emission	**neural**	structure
25	**architecture**	**parameter**	hidden	pulse	finding
26	local	selection	structure	pulses	number
27	narendra	**number**	search	resistance	single
28	quadratic	size	string	hermissenda	defined
29	system	term	**distribution**	order	discover
30	architectures	wide	induction	shunting	grid

Fig. 6. Thirty words which has the most frequent in the 5 documents as representatives. Bold and red words indicate that these words also appeared in the topic that a document belonged to. (Color figure online)

words in topic A shared a significant degree of semantic similarity. Specifically, both the document and topic A contained keywords that pertain to neural network architectures, including terms such as “input”, “output”, “training”, “set”, “weight”, “number”, and “layer”.

Secondly, the document with a docID of 839, which was associated with topic B in Fig. 5, might be related to the generalization bound for neural networks. As such, this document shared certain keywords with topic B, such as “function”, “training data”, “parameter”, and “bound”.

Thirdly, the keywords of both the document with a docID of 440 and topic C were related to the Hidden Markov model (“HMM”), with various features such as “data”,

"parameter", "distribution", "probability", and "likelihood", leading to the document being grouped into topic C.

Moreover, both the document with a docID of 738 and topic D shared numerous keywords, including "cell", "firing", "neuron", "model", "stimulus", and "response". Therefore, it was suggested that the document may describe the nervous system in biology, and belonged to topic D.

Lastly, the document with a docID of 45 belonged to topic E (Fig. 5), and it was notable that the keywords in this document and topic E were related to reinforcement learning. Namely, certain terms such as "algorithm", "learning", "policiy", "function", "action", and "states" were employed to perform "task".

From the above-mentioned analyses of the five selected documents, it is apparent that the NMF technique is capable of reasonably handling topic clustering based on the semantic similarity of shared keywords.

4 Conclusions

This study seeks to examine the robustness of Non-negative Matrix Factorization (NMF) in topic clustering using a data corpus consisting of 1500 documents and 12419 words in the bags-of-words format. The study found that NMF was able to effectively cluster the relatively large and unstructured textual data into 5 meaningful topics and visually represent the topic context through keywords. Additionally, an investigation into the semantic relationship between the keywords of the documents and the corresponding extracted topics showed that NMF was an effective method for topic clustering. Thus, the study recommends the use of NMF as an efficient tool for education management in Vietnam.

References

1. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining, Pearson Addison Wesley (2006)
2. Xu, W., Liu, X., Gong, Y.: Document-clustering based on non-negative matrix factorization. In: Proceedings of SIGIR 2003, pp. 267–273. Toronto, CA (2003)
3. Wang, Y.-X., Zhang, Y.-J.: Nonnegative matrix factorization: a comprehensive review. IEEE Trans. Knowl. Data Eng. **25**(6), 1336–1353 (2013)
4. Ping, H., Xiaohua, X., Jie, D., Baichuan, F.: Low-rank nonnegative matrix factorization on Stiefel manifold. Inf. Sci. **514**, 131–148 (2020)
5. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
6. Andrzej, C., Rafal, Z., Huy, P., Shunichi, A.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation. John Wiley & Sons (2009)
7. Athukorala, S., Mohotti, W.: An effective short-text topic modelling with neighbourhood assistance-driven NMF in Twitter. Soc. Netw. Anal. Min. **12** (89) (2022)
8. Gensler, S., Völckner, F., Egger, M., Fischbach, K., Schoder, D.: Listen to your customers: insights into brand image using online consumer-generated product reviews. Int. J. Electron. Commer. (20), 112–141 (2015)
9. Dua, D., Graff, C.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2019). [http://archive.ics.uci.edu/ml]