Topic Modeling with Non-negative Matrix Factorization

Saurav Kumar
M.Tech(CSE)-2002038
IIIT Guwahati
saurav.kumar@iiitg.ac.in

Dipankar Mitra
M.Tech (CSE)-2002013
IIIT Guwahati
dipankar.mitra@iiitg.ac.in

***Abstract*—Topic modeling is an unsupervised natural language processing technique to discover the topic or set of topics that best describes a given text document. Each topic can be thought as a word or a set of words.Topic modeling allows us to cut through the noise (deal with the high dimensionality of text data) and identify the signal (the main topics) of our text data. One of the mostly used algorithm for Topic modelling is Non-negative Matrix Factorization(NMF). Non-negative Matrix Factorization is applied with two different objective functions: the Frobenius norm, and the generalized Kullback-Leibler divergence. The NMF technique examines documents and discovers topics in a mathematical framework through probability distributions. NMF is basically a linear algebra technique where a matrix is factorized into two matrices, with the property that all three matrices have no negative elements.**

***Keywords*- NMF, PLSA, DF-IDF vectorizer.**

I. Introduction

A topic model is a kind of a probabilistic generative model that has been used widely in the field of computer science with a specific focus on text mining and information retrieval in recent years. Since this model was first proposed, it has received a lot of attention and gained widespread interest among researchers in many research fields. So far, besides text mining, there also have been successful applications in the fields of computer vision, population genetics, and social networks.

The origin of a topic model is latent semantic indexing (LSI) ; it has served as the basis for the development of a topic model. Nevertheless, LSI is not a probabilistic model; therefore, it is not an authentic topic model. Based on LSI, probabilistic latent semantic analysis (PLSA) was proposed by Hofmann and is a genuine topic model. Published after PLSA, latent Dirichlet allocation (LDA) proposed by Blei et al. (2003) is an even more complete probabilistic generative model and is the extension of PLSA. Nowadays, there is a growing number of probabilistic models that are based on LDA via combination with particular tasks. Nonetheless, all the above-mentioned topic models have initially been introduced in the text analysis community for unsupervised topic discovery in a corpus of documents.

Non-negative Matrix Factorization(NMF) is a linear algebra technique that can be used for Topic Modeling. NMF has two main advantages when compared to LDA. The first is that there are completely deterministic algorithms for its resolution. Second, NMF allows for an easier tuning and manipulation of its parameters. NMF is an analog of SVD(Singular Value Decomposition) where all vectors are nonnegative.

II. PROBLEM STATEMENT

The problem statement is to discover hidden semantic structure in a corpus of text. Suppose a nonnegative matrix $A \in R^{m \times n}$ is given. When the desired lower dimension is k, the goal of NMF is to find the two matrices $W \in R^{m \times n}$ and $H \in R^{k \times n}$ having only nonnegative entries such that

$$A \approx WH \tag{1}$$

According to (1), each data point, which is represented as the column of A, can be approximated by an additive combination of the non negative basis vectors, which are represented as the columns of W. As the goal of dimension reduction is to discover compact representation in the form of (1), k is assumed to satisfy that $k < \min\{m, n\}$. The matrices W and H are found by solving an optimization problem defined by

$$min_{W \geq 0, H \geq 0} f(W, H) = ||A - WH||_F^2 \tag{2}$$

The constraints in (2) indicate that all the entries of W and H are nonnegative. Because of the nonnegativity constraints in NMF, the result of NMF can be viewed as document clustering and topic modeling results directly. The goal of this project is to provide an overview of NMF used as a topic modeling method for document data.

III. Non-negative Matrix Factorization

NMF (Nonnegative Matrix Factorization) is a matrix factorization method[3] where we constrain the matrices to be nonnegative. In order to understand NMF, we should clarify the underlying intuition between matrix factorization. Suppose we factorize a matrix X into two matrices W and H so that

$$X \approx WH$$

There is no guarantee that we can recover the original matrix, so we will approximate it as best as we can.

Now, suppose that X is composed of m rows $x_1, x_2, ...x_m$, W is composed of k rows $w_1, w_2, ...w_k$, H is composed of m rows $h_1, h_2, ...h_m$. Each row in X can be considered a data point. For instance, in the case of decomposing images, each row in X is a single image, and each column represents some feature.

The meaning of this equation becomes clearer when we visualize it.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_k \end{bmatrix}, W = \begin{bmatrix} w_1 \\ w_2 \\ ... \\ w_k \end{bmatrix}, H = \begin{bmatrix} h_1 \\ h_2 \\ ... \\ h_k \end{bmatrix}$$

Figure 1: NMF(vectors)

$$x_i = \begin{bmatrix} w_{i1} & w_{i2} & ... & w_{ik} \end{bmatrix} \times \begin{bmatrix} h_1 \\ h_2 \\ ... \\ h_k \end{bmatrix} = \sum_{j=1}^{k} w_{ij} \times h_i$$

Figure 2: NMF components

Basically, we can interpret x_i to be a weighted sum of some components (or bases), where each row in H is a component, and each row in W contains the weights of each component.

IV. Topic modeling as a use case of NMF

Suppose we wanted to decompose a term-document matrix, where each column represented a document, and each element in the document represented the weight of a certain word (the weight might be the raw count or the tf-idf weighted count or some other encoding scheme).

We can see what happens when we decompose this into two matrices. Suppose if the documents came from news articles. The word "eat" would be likely to appear in food-related articles, and therefore co-occur with words like "tasty" and "food". Therefore, these words would probably be grouped together into a "food" component vector, and each article would have a certain weight of the "food" topic.

Therefore, an NMF decomposition of the term-document matrix would yield com- ponents that could be considered "topics", and decompose each document into a weighted sum of topics. This is called topic modeling and is an important applica- tion of NMF.

Note that this interpretation would not be possible with other decomposition meth- ods. We cannot interpret what it means to have a "negative" weight of the food topic. This is another example where the underlying components (topics) and their weights should be non-negative.

V. How do we conduct NMF?

In order to conduct NMF we formalize an objective function and iteratively optimize it. NMF is an NP-hard problem in general, so we will aim for a good local minima. The objective function that is used is :
$minimise\ ||X-WH||_F^2$ w.r.t. W, H s.t. $W, H \geq 0$

A Multiplicative Update rule is used to update W and H iteratively :

$$H \longleftarrow H \bigodot \frac{W^T X}{W^T W H}$$
$$W \longleftarrow H \bigodot \frac{X H^T}{W H H^T}$$

where the matrix not being updated is kept constant.

VI. Vectorizer

Text data requires special preparation before we can start using it for modeling.The text must be parsed to remove words, called tokenization. Then the words need to be encoded as integers or floating point values for use as input to a machine learning algorithm, called feature extraction (or vectorization).

A. Word Counts with CountVectorizer

The CountVectorizer provides a simple way to both tok- enize a collection of text doc- uments and build a vocabulary of known words, but also to encode new documents using that vocabulary. An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document. Because these vectors will contain a lot of zeros, we call them sparse.

B. Word Frequencies with Tf-idfVectorizer

Word counts are a good starting point, but are very basic. One issue with simple counts is that some words like "the" will appear many times and their large counts will not be very meaningful in the encoded vectors. An alternative is to calculate word frequencies, and by far the most popular method is called TF-IDF. This is an acronym than stands for "Term Frequency – Inverse Document" Frequency which are the components of the resulting scores assigned to each word.

- **Term Frequency:** This summarizes how often a given word appears within a document.
- **Inverse Document Frequency:** This downscales words that appear a lot across documents.

TF-IDF are basically word frequency scores that try to highlight words that are more interesting, e.g. frequent in a document but not across documents. The TfidfVectorizer will tokenize documents, learn the vocabulary and inverse document frequency weightings, and allows to encode new documents.

VII. Data Pre-processing

The data are initially in raw form. So we need to preprocess the data so that we can use it accordingly. The NLP syntactic tasks - tokenization, lemmatization are applied. Text may contain stop words like 'the', 'is', 'are'. Stop words can be filtered from the text to be processed. We would not want these words taking up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to be stop words. NLTK(Natural Language Toolkit) in python has a list of stopwords. Using this list, we first filter out the stop words so that only the words which contribute semantically to the sentence are retained.

VIII. Implementing NMF

For NMF, we need to obtain a design matrix. To obtain a Counts design matrix, SKLearn's CountVectorizer module has been used. The transformation will return a matrix of size (Documents × Features), where the value of a cell is going to be the number of times the feature (word) appears in that document.

IX. RESULTS

A. ABC News Headlines dataset

The generated topics and top words can be viewed in figure-5.1

	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10
0	weather	nightmare	egyptsaikal	police	interview	indigenous	court	new	fire	warning
1	abc	bridge	schools	car	anthony	quarter	plane	highway	found	economy
2	wednesday	grafton	teacher	charges	griffin	ep4	crash	old	house	winter
3	tuesday	dud	children	hunt	kemp	australian	dead	push	toddler	sparks
4	news	bridgewater	underperforming	fatal	corey	service	australian	wa	two	national
5	business	closer	testing	bay	brett	public	wa	pm	guilty	spark
6	higgins	am1	grants	muswellbrook	ricky	staff	sentenced	govt	lab	burglaries
7	presenter	historic	monaro	wide	martin	bridge	lebanese	highlights	pool	spike
8	paul	urged	sporting	crash	ponting	myskina	three	australia	set	fires
9	reports	public	faces	rammed	parker	doubt	explosion	shipping	attacked	call
10	analysis	say	assault	victim	white	open	springs	flood	safe	close
11	market	worker	stalking	disappearance	chris	grafton	alice	tasmania	fly	bushfire
12	far	entitlements	calls	charged	mccaw	historic	high	service	woman	nz
13	counter	project	charges	faces	gayle	robber	murder	bans	missing	emergency
14	guidelines	scallop	rewarded	escaped	richie	shipping	investigation	farmers	bushwalker	australian
15	claims	production	parade	death	mukhlas	children	charge	centres	wandering	park
16	lashes	mill	christmas	investigate	downer	say	adamson	family	streets	upgrades
17	north	sugar	efforts	prisoners	questions	sentenced	supreme	smoking	pleads	travel
18	bias	closes	concerns	chief	call	lebanese	nsw	deal	toddlers	philippines
19	wild	cancer	rise	station	media	urged	appointed	work	geelong	news

Figure 3: Topics and top wordsin each discovered by NMF

X. CONCLUSION

In this project, we have presented non-negative matrix factorization(NMF) for topic modeling. We have first introduced the NMF formulation and its applications to topic modeling. We have applied NMF along with Natural Language Process-ing techniques on two datasets - ABC News Headlines dataset and 20Newsgroups dataset. Experimental results on these datasets show the advantage of NMF algo- rithm in terms of topic quality and consistency.

The excellence of NMF in clustering and topic modeling poses numerous excit- ing research directions. One important direction is to improve the scalability of NMF. Parallel distributed algorithms are essential for this purpose, but at the same time,the real-time interaction capability can also be considered from the perspective of a human perception. Another direction is to allow users to better understand clustering and topic modeling outputs.

References

[1] D. Kuang, J. Choo, H. Park *Nonnegative matrix factorization for interactive topic modeling and document clustering, in: Partitional Clustering Algorithms, Springer, 2015*

[2] J. Choo, C. Lee, C. K. Reddy, H. Park, Utopian: *User-driven topic modeling based on interactive nonnegative matrix factorization,* IEEE transactions on visualization and computer graphics 19 (12) (2013) 1992–2001.

[3] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: *T. K. Leen, T. G. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems 13*, MIT Press, 2001, pp. 556–562.

[4] A. Purpura, *Non-negative matrix factorization for topic modeling.*

[5] S. Arora, R. Ge, A. Moitra, *Learning topic models–going beyond svd, in*: 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, IEEE, 2012, pp. 1–10.

[6] Steven Bird, Ewan Klein and Edward Loper, *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit* url: http://www.nltk.org/book