INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

TOPNMF: Topic based Document Clustering using Non-negative Matrix Factorization

R Parimala[1]*, **K Gomathi**[2]

1 Assistant Professor, Department of Computer Science, Periyar EVR College, Tiruchirapalli, 620023, Tamil Nadu, India
2 Research Scholar, Department of Computer Science, Periyar EVR College, Tiruchirapalli, 620023, Tamil Nadu, India

OPEN ACCESS

Received: 13.07.2021

Accepted: 26.08.2021

Published: 27.09.2021

Citation: Parimala R, Gomathi K (2021) TOPNMF: Topic based Document Clustering using Non-negative Matrix Factorization. Indian Journal of Science and Technology 14(31): 2590-2595. https://doi.org/10.17485/IJST/v14i31.1293

* **Corresponding author**.

rajamohanparimala@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This work focuses on creating targeted content-specific topic-based clusters. They can help users to discover the topics in a set of documents information more efficiently. **Methods/Statistical analysis:** The Non-negative Matrix Factorization (NMF) based models learn topics by directly decomposing the term-document matrix, which is a bag-of-word matrix representation of a text corpus, into two low-rank factor matrices namely Word-Topic feature Matrix(WTOM) and Document-Topic feature Matrix(DTOM). Topic clusters and Document clusters are extracted from obtained features matrices. This method does not require any statistical distribution and probability. Experiments were carried out on a subset of BBC sport Corpus. **Findings:** The experimental results indicate that the accuracy of TONMF clusters was observed as 100 percent. **Novelty/Applications:** NMF often fails to improve the given clustering result as the number of parameters increases linearly with the size of the corpus. The computational complexity of the TOPNMF is better than exact decomposition like Singular Value Decomposition (SVD).

Keywords: Topic cluster; Document cluster; Non-negative matrix factorization; K-means clustering; Word cloud

1 Introduction

Online activities are generating an outsized volume of unstructured text within emails, blog spot, social media posts, on-line reviews, news articles etc. Manually grouping massive amounts of text results in creating mistakes and inconsistencies and also is a time overwhelming aspect. Topic clustering is an unsupervised learning problem which finds unknown groups of similar data. Topic modeling attempts to discover and annotate thematic structure in the collection of documents[1]. There are a variety of commonly used topic modeling algorithms including SVD, NMF, Latent Dirichlet Allocation (LDA), and Structural Topic Model (STM). SVD is a method of decomposing a structured format of unstructured text into the orthogonal left singular matrix, which represents the relationship between word and latent topics; a diagonal matrix which describes the strength of each latent topic, and right singular matrix, which indicates the similarity between documents and latent topics.

LDA is one of the probabilistic topic models for discovering the latent topics that occur in text corpus[2]. STM performs topic modeling with document-level covariate information on latent topics with uncertainty. Nonnegative Matrix Factorization (NMF) was introduced as a dimension reduction method for pattern analysis[3]. NMF has found applications in the areas of face detection, speech recognition, text and video/audio document processing, and genetics.

Numerically, NMF decomposes the structured matrix into two lower rank matrices namely WTOM and DTOM whose elements are non-negative, since it gives semantically meaningful hidden word topic features and document topic features.

The proposed TOPNMF approach identifies related topics rather than exact factorization like SVD. SVD is a linear model it might not do well with non-linear. LDA model is a probabilistic model determines the group of terms with different probabilities. LDA cannot learn the infrequent word correct semantics accurately in the case of the co-occurring topic.

In[4] proposed NMF for document clustering. In[5] employed NMF to the text corpus which utilizes a word embedding model. In[6] conducted a study on Topic modeling in short-text using non-negative matrix factorization based on deep reinforcement machine learning. The author concluded with a graph-based word embedding technique which improves the performance of topics modeling of short text document. In[7] made an attempt to find Nonnegative Matrix Factorization for Signal and Data Analytics. In[8] revealed that a deep NMF model is used for unsupervised topic modeling. In[9] analyzed Short-Text Topic Modeling via Non-negative Matrix Factorization method enriched Local Word-Context Correlations and concluded that the proposed SeaNMF as an effective topic model for short texts. In[10] experimented with topic extraction using NMF and LDA on Swedish news articles and segment articles. In[11] found that the LDA model is more relevant than the NMF model in the case of large corpus. In[12] claimed that symmetric NMF is equivalent to kernel K-means. The computational cost of kernel K-means is higher than the standard one. The traditional text mining method is also hard to identify topics in a short text corpus. The evaluating quality of topic modeling is a challenging one. Topic cluster depends upon three parameters namely the number of Topics and list of all related words under particular topic. The proposed TOPNMF algorithm finds broad topics and dominant words that categorize the corpus. The number of the topic to model is the pre-defined parameter This study makes it easier, faster, investigate unstructured text corpus and extracts the foremost vital words used concerning every of the topics.

2 Methodology

This section presents the design and methodology of the proposed model: Topic based Document Clustering using Non-negative Matrix Factorization (TOPNMF). Topic modeling is the process of finding groups of co-occurring words in texts. These groups are called topics[12]. A text corpus is a collection of unstructured noisy nature of all documents. The TOPNMF framework consists of reading corpus, preprocessing, Document Term Matrix (DTM), Word by TOPic Matrix (WTOPM), and Document by TOPic Matrix (DTOPM). The design of the model is given in Figure 1.

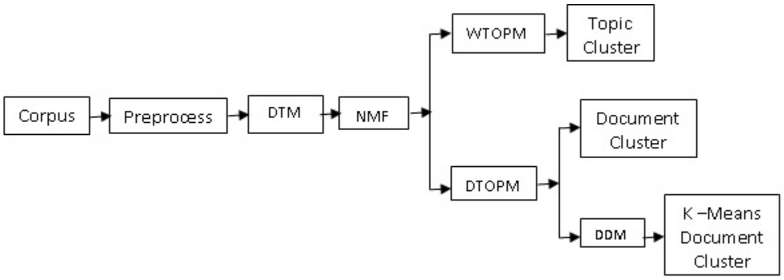

Fig 1. Framework of TOPNME

Preprocessing transforms unstructured text corpus into the structured form[13]. The input passed to NMF algorithm is in a structured form. Tokenization is used to split the documents into a sequence of individual units. Preprocessing involves remove punctuations, remove numbers and special characters and strip white spaces. The undesirable English stopwords are removed from a corpus. Transform case transforms all characters to lower case to remove case sensitivity. TF-IDF vectorizer converts a collection of raw documents to a matrix of **TF-IDF** features. It aims to find the importance of the words relative to other words in the document and corpus[14]. A structured corpus DTM of size mxn is generated using the TF-IDF score where m is the number of documents, and n is the number of features in a corpus. NMF decomposes DTM into WTOPM and DTOPM of size mxr and rxn, where r is the predefined number of topics. The hidden topics present in the DTM can be represented as a distribution over the words, and every document is a blend of topics. WTOPM extracts topic clusters and DTOPM extracts Document clusters. Document – Document Matrix (DDM) generated from DTOPM. The DDM is passed into the input parameters of the K-means clustering algorithm, where k is the number of topics which results in Document clusters.

NMF method

Given Document Term matrix A, find WTOPM(W) and DTOPM(D) such that A mxn≈$W_{mxr}H_{rxn}$, where all elements of A, W, and D are strictly non-negative. Mathematically,

$$\min_{W\geq 0, D\geq 0} ||A-WD||_F^2$$

W and H are updated using the multiplication update rule[15] areas

$$W = W \odot \frac{AD^T}{WDD^T}$$

$$D = D \odot \frac{W^TA}{W^TWD}$$

The extracted word-topic component (W) is clustered into word-topic-cluster and the document-topic component (D) is clustered into the document-topic cluster. Document-Document matrix (DDM) is formed by-product of document-topic-matrix and its transpose. DDM is clustered into k-groups using the K-means clustering algorithm. The results are presented. The various measures are used to evaluate the clusters. The algorithm for the TOPNMF model is as follows.

Algorithm TOPNMF

1. Read Corpus.
2. Perform preprocessing.
3. Create DTM.
4. Apply NMF on DTM.
5. Factorize DTM into W and D
6. Return cluster of size K and top-5 words of each topic.
7. Evaluate the cluster.
8. Create DDM from D
9. Apply K-means clustering on DDM
10. Record the cluster.

Corpus used

The BBCSport dataset includes 737 documents about articles on five topical areas as Athletics, Cricket, Football, Rugby and Tennis from BBC sports web site between the years 2004 to 2005[15]. First 25 news articles from each Cricket and Tennis (CT) news article are considered as BBC- CT news Corpus.

3 Results and Discussion

NMF is a multi-variate matrix factorization method, and it can be applied to any matrix-like data. The proposed algorithm is implemented in R programming, tested on BBC-CT news corpus text content of each document is examined. The first corpus is loaded into the R environment. The Corpus consists of 50 documents and 2120 words. Before fitting the NMF model to the BBC-CT news corpus, we perform common NLP pre-processing procedures on the corpus such as tokenization, remove numbers, remove punctuation, transforms to lowercase, strip white spaces, and remove English stop words. The user-defined stop words (”also”, ”didn't”, ”may”, ”said”, ”can”, ”sinc”, ”will”) are eliminated from the corpus.

DTM is constructed with a TF-IDF score. A DTM, where those words from DTM are removed from the corpus whose sparsity is greater than 80%. The reduced size of DTM is 50 documents and 98 features. The proposed NMF topic model is employed on the corpus and it estimates word-by-topic matrix (W) and Document-by-topic matrix (D). The number of the topic K is set to 2. The top-15 most likely words in each topic are estimated. The word-by-topic matrix with each entry denotes the weight of the corresponding word in the corresponding topic. The document-by-topic matrix with each entry denotes the weight of the corresponding document in the corresponding topic. The top-15 words and their scores of BBC-CT news corpus topics are presented in Tables 1 and 2. Shows the existing study of finding Top 10 frequent words for CT topic of LDA model for BBC Sport[16]. The same word can appear in a different topic due to its ambiguity and the word with the same meaning may appear on the same topic. For example, "first" and "plai" are shared by both the topics in existing study but not in TOPNMF model.

Table 1. Top-15 words of topic of TOPNMF model for BBCSport.

Topic-I(Cricket)		Topic-II(Tennis)	
Word	TF-IDF Score	Word	TF-IDF Score
england	0.10426458	seed	0.11609316
test	0.10048576	dubai	0.08174925
india	0.08150266	champion	0.07171886
south	0.08123415	beat	0.07105532
cricket	0.08056095	top	0.06897649
new	0.07569391	set	0.06658506
tour	0.07329428	number	0.06633464
seri	0.07020662	tenni	0.06487354
intern	0.06931591	open	0.06286429
australia	0.06850402	round	0.06206863
bat	0.06354931	point	0.05992050
captain	0.05872574	win	0.05761346
wicket	0.05765118	world	0.05575125
run	0.05749031	face	0.05455274
oneday	0.05695719	year	0.05019818

Table 2. Top-10 frequent words of topic of LDA model for BBCSport.

Topic-I (Cricket)	test	cricket	plai	england	first	seri	south	match	run	Australia
Topic-II (Tennis)	plai	open	win	match	first	set	year	final	game	roddick

Word cloud displaying top-30 words for corpus and top-15 words of each topic represented in Figures 2, 3 and 4.

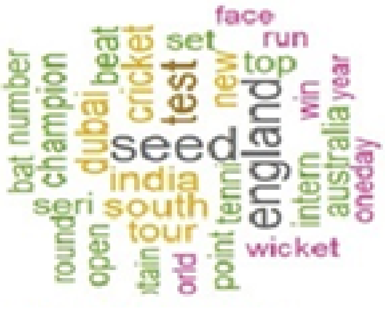

Fig 2. Word cloud of BBC-CT news corpus

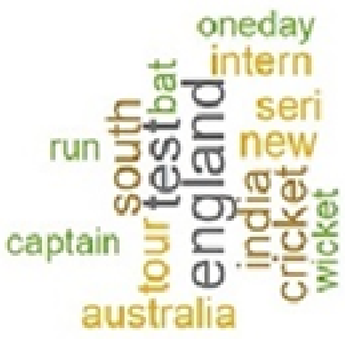

Fig 3. Word cloud (Topic-I)

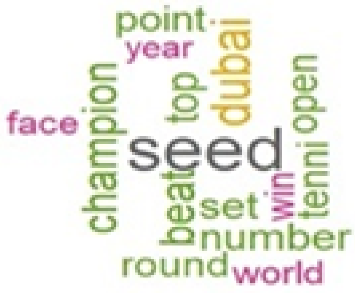

Fig 4. Word cloud (Topic-II)

The algorithm returns the cluster membership of each sample and/ or each feature. The DD matrix is clustered using the k-means cluster. It is observed that both the results of the NMF cluster and k-means cluster are the same. Random initialization of centroids in the k-means method produces different clusters in each run. Table 2 represents the summary of the results of the TOPNMF model. The consensus map of the NMF clustering results is shown in Figure 5. The clustering method attempts to maximize the between-cluster variation (and thus to minimize the within-cluster variation), but how much it succeeds in it depends on the data. From Figure 6. It is observed that there is high similarity within each group and low similarity between each group.

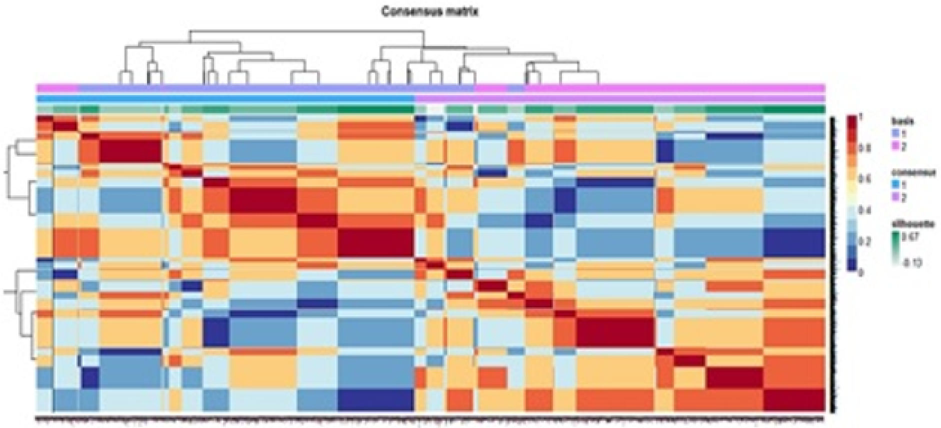

Fig 5. A consensus map of NMF Cluster

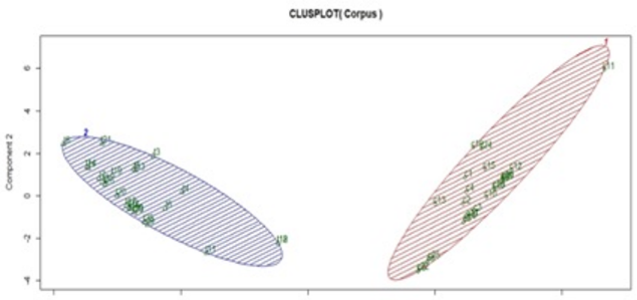

Fig 6. A cluster plot of BBC-CT news corpus

4 Conclusion

The proposed topic modeling automatically identifies topics present in a corpus and to drive hidden pattern exhibited by a text corpus. The results of the TOPNMF model have better semantic representation and the clustering result. The TOPNMF model runs for different topic number, and choose the best cluster according to the evaluation measure. The prediction of the new news article is not implemented in this study. Techniques for efficiently appending features and documents in the corpus will be investigated in the future.

References

1) Blei DM. Probabilistic topic models. *Commun ACM*. 2012;55(4):77–84. Available from: https://doi.org/10.1145/2133806.2133826.
2) Korshunova I, Xiong H, Fedoryszak M, Theis L. Discriminative topic modeling with logistic LDA. *arXiv*. 2019. Available from: http://arxiv.org/abs/1909.01436.
3) Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*. 2007;23(12):1495–1502.
4) Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR ’03. ACM Press. 2003.
5) Ailem M, Salah A, Nadif M. Non-negative matrix factorization meets word embedding. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’17. ACM Press. 2017.
6) Shahbazi Z, Byun YC. Topic modeling in short-text using non-negative matrix factorization based on deep reinforcement learning. *J Intell Fuzzy Syst*. 2020;39(1):753–770.
7) Fu X, Huang X, Sidiropoulos ND, Ma WK. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *arXiv* . 2018. Available from: http://arxiv.org/abs/1803.01257.
8) Wang J, Zhang XL. Deep NMF topic modeling. *arXiv* . 2021. Available from: http://arxiv.org/abs/2102.12998.
9) Shi T, Kang K, Choo J, Reddy CK. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW ’18. ACM Press. 2018.
10) Divaportal.org. 2021. Available from: http://uu.divaportal.org/smash/record.jsf?pid=diva2%3A1512130&dswid=4056.
11) Mifrah S. Topic modeling coherence: A comparative study between LDA and NMF models using COVID’19 corpus. *International Journal of Advanced Trends in Computer Science and Engineering*. 2020;9(4):5756–5761.
12) Dieng AB, Ruiz F, Blei DM. Topic modeling in embedding spaces. *Trans Assoc Comput Linguist*. 2020;8:439–453.
13) Zhang F, Wang C, Trapp A, Flaherty P. A global optimization algorithm for sparse mixed membership matrix factorization. *arXiv* . 2016. Available from: http://arxiv.org/abs/1610.06145.
14) Feinerer I, Hornik K, Meyer D. Text Mining Infrastructure in R. *J Stat Software*. 2008;25(5):1–54.
15) BBC Datasets. 2021. Available from: http://mlg.ucd.ie/datasets/bbc.html.
16) Chen S, Wang Y. 2021. Available from: https://acsweb.ucsd.edu/~yuw176/report/lda.pdf.