# Latent Topic Refinement based on Distance Metric Learning and Semantics-assisted Non-negative Matrix Factorization

**Tran-Binh Dang, Ha-Thanh Nguyen, Le-Minh Nguyen**
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
{binhdang, nguyenhathanh, nguyenml}@jaist.ac.jp

## Abstract

SeaNMF stands for the Semantics-assisted non-negative factorization. This approach is the current state-of-the-art method for topic modeling. In this study, we propose a new method *(i.e., DML-SeaNMF)* for improving the latent topic by utilizing the distance metric learning. The main idea is to iteratively learn the appropriate term-document and term-term relations based on extracted topics in the previous step. Our experiments show that the DML-SeaNMF outperforms the SeaNMF in evaluating based on the topic coherence and topic-based document classification accuracy on several datasets.

## 1 Introduction

With the development of the social network, the textual data repository is being enriched by a huge amount of informative posts, comments, and questions from the internet, by which, we can extract latent information and knowledge by using various text mining methods. Among them, topic modeling is a well-known problem. For short text data (posts, comments, and questions), there are some methods like biterm topic model (BTM)(Yan et al., 2013), LeadLDA (Li et al., 2016), etc. These methods used topic modeling variants to reduce the effects of sparsity issues on topic modeling. Miao (2017) proposed an approach that used deep neural network architecture for topic modeling. In another way, Non-negative matrix factorization(NMF) is a solution for topic modeling. Choo (2013) successfully applied this approach. Yan (2013) used factorize a symmetric term correlation matrix for topic model. In WWW 2018, the SeaNMF (Tian Shi et al., 2018) was proposed to learn topics from short texts. The model combines document-word relation and word-context relation as inputs. SeaNMF outperforms state-of-the-art methods for topic modeling such as LDA (Blei et al., 2003), NMF, PTM (Zuo et al., 2016), and GPUDMM (Li et al., 2016).

In this paper, we propose a novel method that incorporates the distance metric learning(DML) for refining latent topics extracted by the SeaNMF, which is called the DML-SeaNMF.

A proper topic is a cluster of words that share a common semantic. We suppose that in the topics extracted by the SeaNMF, there exists a subset of proper topics. Hence, we aim to refine non-proper topics by changing the input matrices of the SeaNMF that is based on the proper ones. To this end, we consider the most likely topic of each word $w$ as the "soft label" of such a word. Besides, each row in the term-document matrix or the term-term matrix (semantic matrix) is a vector representation of each word, denoted by $\overrightarrow{v}_w$. Hence, we aim to learn the new representation of word $\overrightarrow{v'}_w$ based on $\overrightarrow{v}_w$, which satisfies that vectors representing words in the same topic are close in the vector space (they have the small Euclidean distance). That means the distance $d_{euc}(\overrightarrow{v'}_{w_i}, \overrightarrow{v'}_{w_j})$ is small if $w_i$ and $w_j$ have the same topic. We learn $\overrightarrow{v'}_w$ by using the Large Margin Nearest Neighbor. New vectors $\overrightarrow{v'}_w$ form the new term-document and term-term matrices that are input to the SeaNMF to revise current latent topics.

We compare the performance of our approach

with SeaNMF. The experimental results show that our proposed method significantly outperforms the SeaNMF regarding the following points: (i) the coherence of the topics; and (ii) the effectiveness of topic-based representation for document classification.

The outline of the paper is organized as follows. Section 2 presents the basic knowledge about SeaNMF and Distance metric learning. Section 3 presents an approach named DML-SeaNMF. Section 4 shows experiment results which we did. Finally, we conclude the content of the paper in section 5.

## 2 Background

### 2.1 Non-negative matrix factorization - NMF

Non-negative matrix factorization(Tian Shi et al., 2018) is a method that divides origin matrix to two sub-matrix, with the property that all three matrices have no negative elements. It is useful when analyzing objects which are high-dimensional data. In topic modeling, NMF is no less than the generative probabilistic model. With a corpus has N documents and the number of work/keyword in vocabulary is M, we will have a word-document matrix $A$. The column of $A$ represents bag of words a document on vocabulary. Using NMF for this matrix $A$ was created two lower-dim matrices $W, H$. Multiple of two matrices approximates by matrix $A$.

$$min_{W,H\geq 0}||A - WH^T||_F^2 \tag{1}$$

More detail, matrix $A$ has size $M\ \ words \times N\ documents$. After factorizing, with K topic, we have two matrices. The matrix $W$ represents the distribution of words in the topic. Each column is the presence of a topic in vocabulary. Size of $W$ is $M\ \ words \times K\ \ topic$. The matrix $H$ shows the distribution topic in documents. Each row is the latent topic space of documents. Size of $H$ is $N\ documents \times K\ topic$.

### 2.2 Semantics-assisted NMF - SeaNMF

SeaNMF (Tian Shi et al., 2018) is the model based on Non-negative matrix factorization to discover topics from short texts. SeaNMF uses semantics information to implement the information in its learning process. The representation of semantic information explains pointwise mutual information(PMI) (Levy and Goldberg, 2014). SeaNMF uses two matrices as the input of the model: term-document matrix $A$ and semantic correlation matrix $S$. The matrix $S$ shows the relationship between keyword and their contexts (term - term relation). The objective function of SeaNMF is calculated as follow:

$$min_{W,W_c,H\geq 0}\left|\left|\begin{pmatrix} A^T \\ \sqrt{\alpha}S^T \end{pmatrix} - \begin{pmatrix} H \\ \sqrt{\alpha}W_c \end{pmatrix} W^T\right|\right|_F^2 \tag{2}$$

With input matrices and the number of topics $K$, the SeaNMF model has factorized to three output matrices $W, W_c$ and $H$. The matrix $W$ represents the distribution of words in the topic. Each column is the presence of a topic in vocabulary. The matrix $H$ shows the distribution topic in documents. Each row is the latent topic space of documents. In SeaNMF, there is a new output matrix - $W_c$. The matrix $W_c$ represents the word in context.

### 2.3 Distance metric learning

The distance metric uses distance function which provides a relationship metric between each element in the dataset. Traditionally, practitioners would choose a standard distance metric (Euclidean, City-Block, Cosine, etc.) using a priori knowledge of the domain. Distance metric learning (or simply, metric learning) is the sub-field of machine learning dedicated to automatically constructing optimal distance metrics.

Weinberger and Saul (2009) proposed Large margin nearest neighbor - LMNN as one of the most widely-used Mahalanobis distance learning methods. The method was designed to work with the nearest neighbor classifiers. It can help to improve the performance of the nearest neighbor classifier. LMNN works based on the proposed: label of samples will be more believed if nearest neighbors have the same labels.

Give a dataset: $X = \{x_1, x_2, x_3, ..., x_n\}$ and their labels: $Y = \{y_1, y_2, y_3, ..., y_n\}$. Consider three samples $x_i, x_j, x_k$: $x_j$ is target neighbor of $x_i$, $x_k$ is impostor.

$S = \{(x_i, x_j) : y_i = y_j;\ x_j\ is\ neighbor\ of\ x_i\}$

$R = \{(x_i, x_k) : y_k \neq y_i;\ x_k\ is\ neighbor\ of\ x_i\}$

With a sample $x_i$, sample $x_j$ is a *target neighbor* of $x_i$ if label of $x_j$ is the same with label of $x_i$

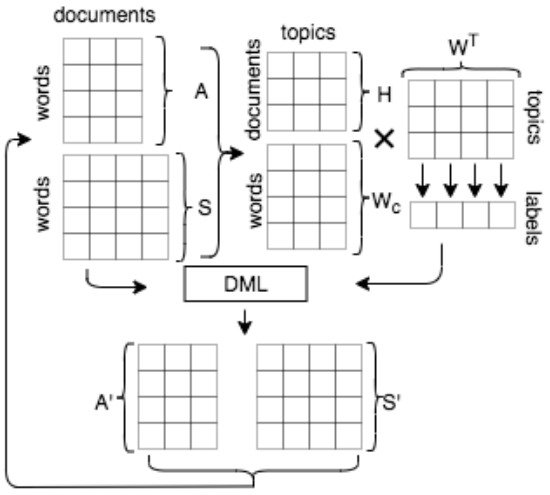


Figure 1: The proposed approach DML-SeaNMF (Example with a corpus: 3 documents, 4 words, 3 topics)

($y_j = y_i$) and $x_j$ is k-nearest neighbor of $x_i$. After establishing the target neighbor, a perimeter is created by the distance of each sample in the dataset $X$. LMNN tried to learn a distance that no sample difference label in this perimeter. So, a margin is built based on the radius of the perimeter. Any sample of a different class that invades this margin will be called an *impostor*. Now, LMNN brings the target neighbor closer and try to keep impostors as far away as possible.

LMNN uses two penalties in learning process. The first one penalized distant target neighbors ($\varepsilon_{pull}$) and the second one penalized nearby impostors($\varepsilon_{push}$). Combine two penalties above with parameter $t$ controls the "pull/push" trade-off will create object function of LMNN:

$$\min \{(1-t)\varepsilon_{pull} + t\varepsilon_{push}\} \quad t \in [0,1] \qquad (3)$$

## 3 The propose method: Distance metric learning for SeaNMF

As mentioned above, the SeaNMF employs an unsupervised approach to effectively learn latent topics in short texts. To improve this method, we aim to iteratively refine topics learned by the SeaNMF as follow.

SeaNMF uses two matrices term-document $A$ and semantic $S$ are input. Term-document matrix $A$ used bag-of-words to show the relationship between word and document. Semantic matrix $S$ was built by the calculation of PMI - a measure of association. The initialization of two matrices is based on the corpus. And, the performance of SeaNMF model depends on the quality of the input matrices. So, we aim to learn better non-negative matrices $A$ and $S$ by a linear transformation $f$.

To this end, we propose a method that incorporates the SeaNMF with distance metric learning (DML) for topic refinement. The idea behind this method is that: (i) assign the most likely topic, obtained by the SeaNMF, for each word which is called a soft label for such a word; (ii) with the soft label and DML, learns parameter of Mahalanobis distance which is a transformation $f$. Thus, after refining, new input inherited the essence of latent topics from the previous step. And, new latent topics which learned in the next step is more clear and better.

With a corpus, matrix $A$ and matrix $S$ were built as the input of the SeaNMF model. After the process, three lower-rank matrices were born - $W$, $W_c$ and $H$. Based on the result of SeaNMF, we can use the learned topic as a soft label for each word. They depend on SeaNMF's results. Let $W$ be a matrix in which, each row in $W$ represents the probability of a word with $K$ topics in latent topic space. Thus, the soft label of each word was determined that the topic has the max probability in each row of $W$. Each word in vocabulary has a corresponding label.

Based on these soft labels, "Distance metric learning"(DML) process tries to learn transformation $f$ by Large margin nearest neighbor. LMNN is an approach using nearest neighbor to improve the performance of clustering. In some cases, topic modeling is also considered as a clustering problem. So, LMNN can support the topic modeling method to increase quality. In our approach, dataset $X$ of LMNN is the vocabulary of the corpus, each word is treated as a sample. However, each sample in the dataset has two representations corresponding with two matrices $A$ and $S$: (i) a vector $N$-dimensions ($N$ documents) with matrix $A$; (ii) a vector $M$-dimensions (size of vocabulary) with matrix $S$. And their labels $Y$ are soft labels of each word. Through the distance metric learning process, there is a transformation matrix $L$ for each representation. The matrix $A'$ and matrix $S'$ are transformed into metric space by:

$$A' = A \times L_1^T \; and \; S' = S \times L_2^T \qquad (4)$$

After transformation complete, $A'$ and $S'$ are checked non-negative condition and used as input of SeaNMF. The negative values in the matrix were replaced by $0$.

This is the end of a time-step in a loop. The process will run with $T$ time-steps. We expect to take three output matrices with their best state. With DML-SeaNMF, we designed a condition to solve the issue as follows: use measure evaluation of topic models - Topic coherence (David et al., 2010). This metric is calculated after the SeaNMF process to check. With time-step $t$ and time-step $t-1$, if topic coherence of $t$ is less than topic coherence of $t-1$, the learning process is stopped. At the time, three lower-rank matrices in time-step $t-1$ is the final output.

## 4 Experiments

### 4.1 Datasets

Experiments are conducted with the benchmark short text dataset. SeaNMF is a model fit with short text data. We used three datasets include:

- **TagNews**[1] The dataset is a part of TagMyNews dataset. It is news extracted from RSS feeds of popular newspaper websites. Categories are: Sport, Business, Entertainment, US, World, Health, Sci_tech.
- **Question 2002**[2] This dataset was used in learning question classification experiments of Xin Li, Dan Roth(2002). Data is public dataset.
- **StackOverflow**[3] The dataset used by Jiaming Xu et al.(2015) in VSM-NLP workshop NAACL 2015. It is questioned in StackOverflow through July 31st to August 14 2012.

Table 1 shows the static information of the three datasets, which we used in our experiments.

### 4.2 Evaluation metrics

Topic coherence (David et al., 2010) is typically an evaluation method for evaluating topic models. With topic $k$: After the models generates topic consisting of words, this metric is applied on the top $n$ words of the topic. Given a topic $k$, PMI value is computed on this topic as described in (Tian Shi et al., 2018). Topic coherence is the average value of PMI on all of the topics.

$$TC_k = \frac{2}{n(n-1)} \sum_{1 \le i \le j \le n} log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (5)$$

where $n$ is top-$n$ words in topic $k$, $p(w_i, w_j)$ is the probability of word $w_i, w_j$ co-occurring. $p(w_i)$ , $p(w_j)$ is marginal probability of $w_i$, $w_j$.

$$Topic\ Coherence = \frac{\sum_{k=1}^{K} TC_k}{K} \quad (6)$$

The higher topic coherence, the better model. In our experiments, the number of the top words $n$ is set to 10, the number of topics $K$ set to 50.

The model runs with the stop condition described in Section 3. Then, the result is compared with the output of the SeaNMF model, which runs separately. We evaluate with two metrics: (i) max topic coherence: the largest value of topic coherence in time-steps and (ii) average topic coherence: the ratio between the sum of topic coherence values and number of time-steps. This experiment compares when we use or not use distance metrics learning.

Besides, we also use document classification performance to measure topic model effectiveness. Latent topics extracted from the models are used as features for a single fully connected layer neural network to perform classification. Training and testing data are randomly split with a ratio of 4:1. The quality is measured by three measures: precision, recall, and F-score.

| Dataset | Docs | Terms | Avg doc-len | Labels |
|---|---|---|---|---|
| TagNews | 1000 | 3505 | 7.77 | 7 |
| Question 2002 | 1000 | 2837 | 8.61 | 6 |
| StackOverflow | 1000 | 2502 | 8.69 | 20 |

Table 1: Basic statistics of datasets in our experiments

| | | SeaNMF | DML-SeaNMF |
|---|---|---|---|
| TagNews | Max | 2.671 | 3.374 |
| | Avg | 2.589 | 3.073 |
| Question 2002 | Max | 2.445 | 3.183 |
| | Avg | 2.228 | 2.782 |
| StackOverflow | Max | 2.233 | 2.545 |
| | Avg | 2.187 | 2.389 |

Table 2: Topic coherence results with three datasets

[1] https://github.com/isthegeek/News-Classification

[2] https://cogcomp.org/Data/QA/QC/

[3] https://github.com/jacoxu/StackOverflow?

| | TagNews | StackOverflow |
|---|---|---|
| LDA | 2.023 | 0.675 |
| NMF | 2.426 | 1.035 |
| SeaNMF | 2.671 | 1.919 |
| DML-NMF | 2.819 | 2.009 |
| DML-SeaNMF | 3.374 | 3.11 |

Table 3: Topic coherence results on 5 methods : LDA, NMF, SeaNMF, DML-NMF, DML-SeaNMF In this experiments, we use StackOverflow dataset with 4000 samples

| | | SeaNMF | DML-SeaNMF |
|---|---|---|---|
| TagNews | Recall | 0.36 | 0.39 |
| | Precision | 0.35 | 0.36 |
| | F-score | 0.35 | 0.37 |
| Question 2002 | Recall | 0.55 | 0.56 |
| | Precision | 0.57 | 0.59 |
| | F-score | 0.55 | 0.56 |
| StackOverflow | Recall | 0.42 | 0.55 |
| | Precision | 0.43 | 0.56 |
| | F-score | 0.42 | 0.54 |

Table 4: Performance of SeaNMF and DML-SeaNMF in documents classification

### 4.3 Results and Discussion

In our experiments, we compared the performance of our method with SeaNMF. The topic coherence value is shown in table 2. With two metrics: max topic coherence and average topic coherence, DML-SeaNMF is better than SeaNMF on all of the three datasets. After 2-3 time-steps, DML-SeaNMF could find a better state of input matrices to return higher topic coherence value. The difference between the average performance of the two models is significant.

To analysis overview, we do more an experiment with datasets: TagNews and StackOverflow on 5 methods: LDA, NMF, DML-NMF, SeaNMF, and DML-SeaNMF. The result shown in table 3.

The document classification result is shown in table 4. On the three datasets, DML - SeaNMF all outperformed SeaNMF. The difference between the two methods is about 2-3% on TagNews and Question 2002. This number is the largest on StackOverflow (13%). That happened because StackOverflow samples often contain name entities that are identical with the class labels. Our model extracted this kind of features better than SeaNMF. The refinement of the latent topic helps the topic feature become more descriptive.

After learning latent topics on TagNews and StackOverflow datasets, we find similar topics obtained from DML-SeaNMF and SeaNMF based on the top-10 keywords. The list of the top-10 keywords in the selected topics obtained is shown in Table 5. As we can see, two topics for TagNews are about Sport and Japan news. The topic selected from StackOverflow related to Visual Studio.

In this paper's scope, we conduct experiments with short text datasets. However, our proposal is not limited to short text. Investigating and optimizing the method for long documents is also one of our possible future directions.

## 5 Conclusion

This paper presents a method to refine latent topics. Our method proposes the combination of Distance metric learning and SeaNMF. Large margin nearest neighbor(LMNN) is chosen to use in the learning distance process. LMNN takes latent topics as labels and sample is the word. This learning process creates a transformation matrix to update the input matrices of the topic model. We compared DML-SeaNMF with one of the state-of-the-art methods(SeaNMF) on three datasets. Experimental results showed that our model is effective when testing the benchmark data. In future works, we want to improve and extend this method especially on long documents.

## References

Tian Shi, Kyeongpil Kang, Jaegul Choo and Chandan K. Reddy 2018, *Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations*, In Proceedings of the International Conference on World Wide Web (WWW) Lyon, France

Bac Nguyen, BernardDe Baets 2018, *An approach to supervised distance metric learning based on difference of convex functions programming*, Pattern Recognition Volume 81, pp. 562-574

Weinberger, Kilian Q and John Blitzer and Lawrence K. Saul, 2006, *Distance Metric Learning for Large Mar-*

Table 5: Discovered topic by DML-SeaNMF and SeaNMF

| | TagNews | | | | StackOverflow | |
|---|---|---|---|---|---|---|
| Category | Sport | | Japan | | Visual Studio | |
| | DML-SeaNMF 6 | SeaNMF 34 | DML-SeaNMF 42 | SeaNMF 10 | DML-SeaNMF 2 | SeaNMF 2 |
| Top 10 keywords | league | keeps | japan | japan | Visual | Visual |
| | basketball | winning | nuclear | street | Studio | Studio |
| | play | semi-finals | trust | wall | Window | project |
| | global | champions | crisis | nuclear | FreezingTFS | Code |
| | champions | roundup | government | worries | Might | Using |
| | semi-finals | nbc | shut | deals | screen | projects |
| | soccer | basketball | rescue | stocks | IFEnd | Can |
| | uconn | drought | reactors | dow | Refactoring | Keyboard |
| | winning | play-off | radioactivity | rescue | Structured | build |
| | fans | share | quake | quake | IntelliSense | Setup |


*gin Nearest Neighbor Classification*, NIPS 2005, pp. 1473–1480, MIT Press

Shiming Xiang, Feiping Nie, Changshui Zhang 2008, *Learning a Mahalanobis distance metric for data clustering and classification*, Pattern Recognition Volume 41, Issue 12, pp. 3600-3612

Xin Li, Dan Roth 2002, *Learning Question Classifiers* COLING'02

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, Hongwei Hao, 2015, *Short Text Clustering via Convolutional Neural Networks* VSM-NLP workshop, NAACL

David M. Blei, Andrew Y. Ng, and Michael I. Jordan 2003, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, 3:993–1022

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng 2013, *A Biterm Topic Model for Short texts*, In 22nd International World Wide Web Conference, WWW 2013 Rio de Janeiro, Brazil, pp. 1445-1456

Jing Li, Ming Liao, Wei Gao, Yulan He, and KamFai Wong, 2016, *Topic Extraction from Microblog Posts Using Conversation Structures*, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume1: Long Papers Berlin, Germany

Yishu Miao, Edward Grefenstette, and Phil Blunsom, 2017, *Discovering Discrete Latent Topics with Neural Variational Inference*, In Proceedings of the 34th International Conference on Machine Learning, ICML 2017 Sydney, NSW, Australia, pp. 2410–2419

Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park, 2013, *Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization*, IEEE transactions on visualization and computer graphics 19

Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park, 2015, *Weakly supervised nonnegative matrix factorization for user-driven clustering* Data Mining and Knowledge Discovery 29,pp. 1598–1621

Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang 2013, *Learning topics in short texts by non-negative matrix factorization on term correlation matrix*, In Proceedings of the 2013 SIAM International Conference on Data Mining. SIAM, 749–757

Kilian Q. Weinberger and Lawrence K. Saul, 2009, *Distance Metric Learning for Large Margin Nearest Neighbor Classification*, Journal of Machine Learning Research 10 (2009) pp. 207-244

Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong, 2016, *Topic Modeling of Short Texts: A Pseudo-Document View*, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2105–2114

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma, 2016, *Topic Modeling for Short Texts with Auxiliary Word Embeddings*, In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 165–174

Newman David, Lau Jey Han, Grieser Karl and Baldwin Timothy, 2010, *Automatic evaluation of topic coherence*, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100-108

Jacob Goldberger, Geoffrey Hinton, Sam Roweis, Ruslan Salakhutdinov, 2005, *Neighbourhood Components Analysis*, Advances in Neural Information Processing Systems 17, pp. 513–520, MIT Press

Levy, Omer and Goldberg, Yoav, 2014, *Neural Word Embedding As Implicit Matrix Factorization*, Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14