



Article Interpretable Topic Extraction and Word Embedding Learning Using Non-Negative Tensor DEDICOM

Lars Hillebrand ^{1,2,*,†}, David Biesner ^{1,2,*,†}, Christian Bauckhage ^{1,2} and Rafet Sifa ¹

- ¹ Fraunhofer IAIS, 53757 Sankt Augustin, Germany; christian.bauckhage@iais.fraunhofer.de (C.B.); rafet.sifa@iais.fraunhofer.de (R.S.)
- ² Department of Computer Science, University of Bonn, 53113 Bonn, Germany
- * Correspondence: lars.patrick.hillebrand@iais.fraunhofer.de (L.H.); david.biesner@iais.fraunhofer.de (D.B.)
- + These authors contributed equally to this work.

Abstract: Unsupervised topic extraction is a vital step in automatically extracting concise contentual information from large text corpora. Existing topic extraction methods lack the capability of linking relations between these topics which would further help text understanding. Therefore we propose utilizing the Decomposition into Directional Components (DEDICOM) algorithm which provides a uniquely interpretable matrix factorization for symmetric and asymmetric square matrices and tensors. We constrain DEDICOM to row-stochasticity and non-negativity in order to factorize pointwise mutual information matrices and tensors of text corpora. We identify latent topic clusters and their relations within the vocabulary and simultaneously learn interpretable word embeddings. Further, we introduce multiple methods based on alternating gradient descent to efficiently train constrained DEDICOM algorithms. We evaluate the qualitative topic modeling and word embedding performance of our proposed methods on several datasets, including a novel New York Times news dataset, and demonstrate how the DEDICOM algorithm provides deeper text analysis than competing matrix factorization approaches.

Keywords: matrix factorization; tensor factorization; word embeddings; topic modeling; NLP

1. Introduction

Matrix factorization methods have always been a staple in many natural language processing (NLP) tasks. Factorizing a matrix of word co-occurrences can create both lowdimensional representations of the vocabulary, so-called word embeddings [1,2], that carry semantic and topical meaning within them, as well as representations of meaning that go beyond single words to latent topics.

Decomposition into Directional Components (DEDICOM) is a matrix factorization technique that factorizes a square, possibly asymmetric, matrix of relationships between items into a loading matrix of low-dimensional representations of each item and an affinity matrix describing the relationships between the dimensions of the latent representation (see Figure 1 for an illustration).

We introduce a modified row-stochastic variation of DEDICOM, which allows for interpretable loading vectors and apply it to different matrices of word co-occurrence statistics created from Wikipedia based semi-artificial text documents. Our algorithm produces low-dimensional word embeddings, where one can interpret each latent factor as a topic that clusters words into meaningful categories. Hence, we show that row-stochastic DEDICOM successfully combines the task of learning interpretable word embeddings and extracting representative topics.

We further derive a similar model for factorization of three-dimensional data tensors, which represent word co-occurrence statistics for text corpora with intrinsic structure that allows for some separation of the corpus into subsets (e.g., a news corpus structered by time).



Citation: Hillebrand, L.; Biesner, D.; Bauckhage, C.; Sifa, R. Interpretable Topic Extraction and Word Embedding Learning Using Non-Negative Tensor DEDICOM. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 123–167. https://doi.org/10.3390 /make3010007

Received: 30 November 2020 Accepted: 13 January 2021 Published: 19 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/).



Figure 1. (a) The DEDICOM algorithm factorizes a square matrix $S \in \mathbb{R}^{n \times n}$ into a loading matrix $A \in \mathbb{R}^{n \times k}$ and an affinity matrix $R \in \mathbb{R}^{k \times k}$. (b) The tensor DEDICOM algorithm factorizes a three dimensional tensor $\underline{S} \in \mathbb{R}^{t \times n \times n}$ into a loading matrix $A \in \mathbb{R}^{n \times k}$ and a three dimensional affinity tensor $\underline{R} \in \mathbb{R}^{t \times k \times k}$.

An interesting aspect of this type of factorization is the interpretability of the affinity matrix. An entry in the matrix directly describes the relationship between the topics of the respective row and column and one can therefore use this tool to extract topics that a certain text corpus deals with and analyze how these topics are connected in the given text.

In this work we first describe the aforementioned DEDICOM algorithm and provide details on the modified row-stochasticity constraint and on optimization. We further expand our model to factorize three-dimensional tensors and introduce a multiplicative update rule that facilitates the training procedure. We then present results of various experiments on both semi-artificial text documents (combinations of Wikipedia articles) and real text documents (movie reviews and news articles) that show how our approach is able to capture hidden latent topics within text corpora, cluster words in a meaningful way and find relationships between these topics within the documents.

This paper is an extension of previous work [3]. In addition to the algorithms and experiments described there, we here add the extension of the DEDICOM algorithm to three-dimensional tensors, introduce a multiplicative update rule to increase training stability and present new experiments on two additional text corpora (Amazon reviews and New York Times news articles).

2. Related Work

Matrix factorization describes the task of compressing the most relevant information from a high-dimensional input matrix into multiple low-dimensional factor matrices, with either approximate or exact input reconstruction (see for example [4] for a theoretical overview of common methods and their applications). In this work we consider the DEDICOM algorithm, which has a long history of providing an interpretable matrix or tensor factorization, mostly for rather low-dimensional tasks.

First described in [5], it since has been applied to analysis of social networks [6], email correspondence [7] and video game player behavior [8,9]. DEDICOM also has successfully been employed to NLP tasks such as part of speech tagging [10], however to the best of our knowledge we provide the first implementation of DEDICOM for simultaneous word embedding learning and topic modeling.

Many works deal with the task of putting constraints on the factor matrices of the DEDICOM algorithm. In [7,8], the authors constrain the affinity matrix R to be non-negative, which aids interpretability and improves convergence behavior if the matrix to be factorized is non-negative. However, their approach relies on the Kronecker product between matrices in the update step, solving a linear system of $n^2 \times k^2$, where n denotes the number of items in the input matrix and k the number of latent factors. These dimensions make the application on text data, where n describes the number of words in the vocabulary, a computationally futile task. Constraints on the loading matrix, A, include non-negativity as well (see [7]) or column-orthogonality as in [8].

In contrast, we propose a new modified row-stochasticity constraint on *A*, which is tailored to generate interpretable word embeddings that carry semantic meaning and represent a probability distribution over latent topics.

The DEDICOM algorithm has previously been applied to tensor data as well, for example in [11], in which the authors apply the algorithm on general multirelational data by computing an exact solution for the affinity matrix. Both [6,7] explore a slight variation of our tensor DEDICOM approach to analyze relations in email data and [12] apply a similar model on non-square input tensors.

Previous matrix factorization based methods in the NLP context mostly dealt with either word embedding learning or topic modeling, but not with both tasks combined.

For word embeddings, the GloVe [2] model factorizes an adjusted co-occurrence matrix into two matrices of the same dimension. The work is based on a large text corpus with a vocabulary of $n \approx 400,000$ and produces word embeddings of dimension k = 300. In order to maximize performance on the word analogy task, the authors adjusted the co-occurrence matrix to the logarithmized co-occurrence matrix and added bias terms to the optimization objective.

A model conceived around the same time, word2vec [13], calculates word embeddings not from a co-occurrence matrix but directly from the text corpus using the skip-gram or continuous-bag-of-words approach. More recent work [1] has shown that this construction is equivalent to matrix factorization on the pointwise mutual information (PMI) matrix of the text corpus, which makes it very similar to the glove model described above.

Both models achieve impressive results on word embedding related tasks like word analogy, however the large dimensionality of the word embeddings makes interpreting the latent factors of the embeddings impossible.

On the topic modeling side, matrix factorization methods are routinely applied as well. Popular algorithms like non-negative matrix factorization (NMF) [14], singular value decomposition (SVD) [15,16] and principal component analysis (PCA) [17] compete against the probabilistic latent dirichlet allocation (LDA) [18] to cluster the vocabulary of a word co-occurrence or document-term matrix into latent topics. (More recent expansions of these methods can be found in [19,20].) Yet, we empirically show that the implicitly learned word embeddings of these methods lack semantic meaning in terms of the cosine similarity measure.

We benchmark our approach qualitatively against these methods in Section 4.3 and the Appendixes A and B.

3. Constrained DEDICOM Models

In this section we provide a detailed theoretical view at different constrained DEDICOM algorithms utilized for factorizing word co-occurrence based positive pointwise mutual information matrices and tensors.

We first consider the case of a two-dimensional input matrix S (see Figure 1a) in Section 3.1. We then present an extension of the algorithm for three-dimension input tensors <u>S</u> (see Figure 1b) in Section 3.2. Finally we derive a multiplicative update rule for non-negative tensor DEDICOM.

3.1. The Row-Stochastic DEDICOM Model for Matrices

For a given language corpus consisting of *n* unique words $X = x_1, ..., x_n$ we calculate a co-occurrence matrix $W \in \mathbb{R}^{n \times n}$ by iterating over the corpus on a word token level with a sliding context window of specified size. Then

$$W_{ij} =$$
#word *i* appears in context of word *j*. (1)

Note that the word context window can be applied symmetrically or asymmetrically around each word. We choose a symmetric context window, which implies a symmetric co-occurrence matrix, $W_{ij} = W_{ji}$.

We then transform the co-occurrence matrix into the pointwise mutual information matrix (PMI), which normalizes the counts in order to extract meaningful co-occurrences from the matrix. Co-occurrences of words that occur regularly in the corpus are decreased since their appearance together might be nothing more than a statistical phenomenon, the

co-occurrence of words that appear less often in the corpus give us meaningful information about the relations between words and topics. We define the PMI matrix as

$$\mathbf{PMI}_{ij} \coloneqq \log \mathbf{W}_{ij} + \log N - \log N_i - \log N_j \tag{2}$$

where $N := \sum_{ij=1}^{n} W_{ij}$ is the sum of all co-occurrence counts of W, $N_i := \sum_{j=1}^{n} W_{ij}$ the row sum and $N_j := \sum_{i=1}^{n} W_{ij}$ the column sum.

Since the co-occurrence matrix W is symmetrical, the transformed PMI matrix is symmetrical as well. Nevertheless, DEDICOM is able to factorize both symmetrical and non-symmetrical matrices. We expand details on symmetrical and non-symmetrical relationships in Section 3.4.

Additionally, we want all entries of the matrix to be non-negative, our final matrix to be factorized is therefore the positive PMI (PPMI)

$$S_{ij} = \mathbf{PPMI}_{ij} = \max\{0, \mathbf{PMI}_{ij}\}.$$
(3)

Our aim is to decompose this matrix using row-stochastic DEDICOM as

$$S \approx ARA^{T}$$
, with $S_{ij} \approx \sum_{b=1}^{k} \sum_{c=1}^{k} A_{ib}R_{bc}A_{jc}$, (4)

where $A \in \mathbb{R}^{n \times k}$, $R \in \mathbb{R}^{k \times k}$, A^T denotes the transpose of A and $k \ll n$. Literature often refers to A as the loading matrix and R as the affinity matrix. A gives us for each word i in the vocabulary a vector of size k, the number of latent topics we wish to extract. The square matrix R then provides possibility for interpretation of the relationships between these topics.

Empirical evidence has shown that the algorithm tends to favor columns unevenly, such that a single column receives a lot more weight in its entries than the other columns. We try to balance this behavior by applying a column-wise z-normalization on *A*, such that all columns have zero mean and unit variance.

In order to aid interpretability we wish each word embedding to be a distribution over all latent topics, i.e., entry A_{ib} in the word-embedding matrix provides information on how much topic *b* describes word *i*.

To implement these constraints we therefore apply a row-wise softmax operation over the column-wise z-normalized *A* matrix by defining $A' \in \mathbb{R}^{n \times k}$ as

$$A'_{ib} \coloneqq \frac{\exp(\bar{A}_{ib})}{\sum_{b'=1}^{k} \exp(\bar{A}_{ib'})}, \quad \bar{A}_{ib} \coloneqq \frac{A_{ib} - \mu_b}{\sigma_b},$$

$$\mu_b \coloneqq \frac{1}{n} \sum_{i=1}^{n} A_{ib}, \quad \sigma_b \coloneqq \sqrt{\frac{1}{n} \sum_{i=1}^{n} (A_{ib} - \mu_b)^2}$$
(5)

and optimizing A for the objective

$$S \approx A' R(A')^T. \tag{6}$$

Note that after applying the row-wise softmax operation all entries of A' are non-negative.

To judge the quality of the approximation (6) we apply the Frobenius norm, which measures the difference between *S* and $A'R(A')^T$. The final loss function we optimize our model for is therefore given by

$$\mathcal{L}(S, A, R) = \left\| S - A' R (A')^T \right\|_F^2$$
(7)

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\boldsymbol{S}_{ij}-\left(\boldsymbol{A}'\boldsymbol{R}(\boldsymbol{A}')^{T}\right)_{ij}\right)^{2}$$
(8)

with

$$\left(\boldsymbol{A}'\boldsymbol{R}(\boldsymbol{A}')^{T}\right)_{ij} = \sum_{b=1}^{k} \sum_{c=1}^{k} \boldsymbol{A}'_{ib} \boldsymbol{R}_{bc} \boldsymbol{A}'_{jc}$$
(9)

and A' defined in (5).

To optimize the loss function we train both matrices using alternating gradient descent similar to [8]. Within each optimization step we apply

$$A \leftarrow A - f_{\theta}(\nabla_A, \eta^A), \text{ where } \nabla_A = \frac{\partial \mathcal{L}(S, A, R)}{\partial A}$$
 (10)

$$R \leftarrow R - f_{\theta}(\nabla_R, \eta^R), \quad \text{where} \quad \nabla_R = \frac{\partial \mathcal{L}(S, A, R)}{\partial R}$$
 (11)

with η^A , $\eta^R > 0$ being individual learning rates for both matrices and $f_{\theta}(\cdot)$ representing an arbitrary gradient based update rule with additional hyperparameters θ . For our experiments we employ automatic differentiation methods. For details on the implementation of the algorithm above refer to Section 4.2.

3.2. The Constrained DEDICOM Model for Tensors

In this section we extend the model described above to three-dimensional tensors as input data. As above, the input describes the co-occurrences of vocabulary items in a text corpus. However, we consider additionally structured text: Instead of one matrix describing the entire corpus we unite multiple $n \times n$ matrices of co-occurrences into one tensor $\underline{S} \in \mathbb{R}^{t \times n \times n}$. Each of the *t* slices then consists of an adjusted PPMI matrix for a subset of the text corpus. This structure could originate for instance from different data (e.g., different Wikipedia articles), different topical subsets of the data source (e.g., reviews for different articles) or describe time-slices (e.g., news articles for certain time periods).

To construct the PPMI tensor we again take a vocabulary $X = x_1, ..., x_n$ over the entire corpus. For each subset l we then calculate a co-occurrence matrix $\underline{W}_l \in \mathbb{R}^{n \times n}$ as described above. Stacking these matrices yields the co-occurrence tensor $\underline{W} \in \mathbb{R}^{t \times n \times n}$.

When transforming slice \underline{W}_l into a PMI matrix we want to apply information from the entire corpus. We therefore do not only calculate the column, the row and the total sums on the corresponding subset but on the entire text corpus. Therefore

$$\underline{\mathbf{PMI}}_{lij} \coloneqq \log \underline{W}_{lij} + \log N - \log N_i - \log N_j, \tag{12}$$

where $N := \sum_{l=1}^{k} \sum_{ij=1}^{n} \underline{W}_{lij}$ is the sum of all co-occurrence counts of \underline{W} , $N_i := \sum_{l=1}^{k} \sum_{j=1}^{n} \underline{W}_{lij}$ the row sum and $N_j := \sum_{l=1}^{k} \sum_{i=1}^{n} \underline{W}_{lij}$ the column sum.

Finally we define the positive pointwise mutual information tensor as

$$\underline{S}_{lij} = \underline{\mathbf{PPMI}}_{lij} = \max\{0, \underline{\mathbf{PMI}}_{lij}\}.$$
(13)

We decompose this input tensor into a matrix $A \in \mathbb{R}^{n \times k}$ and a tensor $\underline{R} \in \mathbb{R}^{t \times k \times k}$, such that

$$\underline{S} \approx A\underline{R}A^T \tag{14}$$

where we multiply each slice of <u>**R**</u> with A and A^T to reconstruct the corresponding slice of <u>**S**</u>:

$$\underline{S}_{lij} \approx \sum_{b=1}^{k} \sum_{c=1}^{k} A_{ib} \underline{R}_{lbc} A_{jc}.$$
(15)

We keep our naming convention for A as the loading matrix and \underline{R} as the affinity tensor, since again A gives us for each word i in the vocabulary a vector of size k and for each slice l the square matrix $\underline{R}_l := (\underline{R}_{lij})_{i,j=1}^k$ provides information on the relationships between the topics in the l-th input slice.

Analogous to (7) we construct a loss function

$$\mathcal{L}(\underline{S}, A, \underline{R}) = \left\| \underline{S} - A\underline{R}(A)^T \right\|_F^2$$
(16)

$$=\sum_{l=1}^{I}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\underline{S}_{lij}-\left(\underline{A}\underline{R}_{l}(A)^{T}\right)_{ij}\right)^{2}$$
(17)

$$=\sum_{l=1}^{t} \mathcal{L}(\underline{S}_{l}, A, \underline{R}_{l})$$
(18)

with

$$\left(A\underline{R}_{l}(A)^{T}\right)_{ij} = \sum_{b=1}^{k} \sum_{c=1}^{k} A_{ib}\underline{R}_{lbc}A_{jc}.$$
(19)

Note that in this framework, the DEDICOM algorithm described in the previous section is equivalent to tensor DEDICOM with t = 1.

Update steps can then be taken via alternating gradient descent on A and \underline{R} . As in the previous section, one can now add additional constraints to A and \underline{R} and calculate the gradients as in (10), using automatic differentiation methods. Taking update steps of size η^A and $\eta^{\underline{R}}$ respectively leads to an eventual convergence to some local or global minimum of the loss (16) with respect to the original or constrained A and \underline{R} .

Alternatively, constraints can be added to A and \underline{R} by methods like projected gradient descent and the Frank-Wolfe algorithm [21] which either adjust the respective matrix or tensor to be constrained after the gradient step or apply the general gradient step such that the matrix or tensor never leaves the constrained area.

However, empirical results show that automatic differentiation methods lead to slow and unstable training convergence and worse qualitative results when applying the mentioned constraints on the factor matrices and tensors. We therefore derive an alternative method of applying alternating gradient descent to A and \underline{R} based on multiplicative update rules. This does not only improve training stability and convergence behavior but also lead to better qualitative results (see Section 4.3 and Figure 2).

We derive the gradients for A and \underline{R} analytically and set the learning rates η^A and $\eta^{\underline{R}}$ individually for each element (i, j) as η^A_{ij} for matrix A and for each element (l, i, j) as $\eta^{\underline{R}}_{lij}$ for tensor \underline{R} , such that the resulting update step is an element-wise multiplication of the respective matrix or tensor.

We derive the updates for the matrix algorithm first and later extend them for the tensor case. For detailed derivations refer to Appendix B. For *A* we derive the gradient analytically as

$$\frac{\partial \mathcal{L}(S, A, R)}{\partial A} = -2 \Big(SAR^T + S^T AR - A \Big(RA^T AR^T + R^T A^T AR \Big) \Big).$$
(20)

Therefore the update step is

$$\boldsymbol{A}_{ij} \leftarrow \boldsymbol{A}_{ij} + \eta_{ij}^{\boldsymbol{A}} 2 \left(\left[\boldsymbol{S} \boldsymbol{A} \boldsymbol{R}^{T} + \boldsymbol{S}^{T} \boldsymbol{A} \boldsymbol{R} \right]_{ij} - \left[\boldsymbol{A} \left(\boldsymbol{R} \boldsymbol{A}^{T} \boldsymbol{A} \boldsymbol{R}^{T} + \boldsymbol{R}^{T} \boldsymbol{A}^{T} \boldsymbol{A} \boldsymbol{R} \right) \right]_{ij} \right).$$
(21)



Figure 2. Reconstruction loss development during tensor factorization training. The *x*-axis plots the number of epochs on a logarithmic scale, the *y*-axis plots the corresponding reconstruction error for each method.

If we now chose η^A as

$$\eta_{ij}^{A} \coloneqq \frac{A_{ij}}{2[A(R^{T}A^{T}AR + RA^{T}AR^{T})]_{ij}}.$$
(22)

the update (21) becomes

$$A_{ij} \leftarrow A_{ij} \frac{[S^T A R + S A R^T]_{ij}}{[A(R^T A^T A R + R A^T A R^T)]_{ii}}.$$
(23)

For *R* we derive the gradient analytically as

$$\frac{\partial \mathcal{L}(\mathbf{S}, \mathbf{A}, \mathbf{R})}{\partial \mathbf{R}} = -2(\mathbf{A}^T \mathbf{S} \mathbf{A} - \mathbf{A}^T \mathbf{A} \mathbf{R} \mathbf{A}^T \mathbf{A}).$$
(24)

Therefore the update step is

$$\boldsymbol{R}_{ij} = \boldsymbol{R}_{ij} + \eta_{ij}^{\boldsymbol{R}} 2([\boldsymbol{A}^{T} \boldsymbol{S} \boldsymbol{A}]_{ij} - [\boldsymbol{A}^{T} \boldsymbol{A} \boldsymbol{R} \boldsymbol{A}^{T} \boldsymbol{A}]_{ij}).$$
(25)

Choose

$$\eta_{ij}^{\mathbf{R}} \coloneqq \frac{\mathbf{R}_{ij}}{2[\mathbf{A}^T \mathbf{A} \mathbf{R} \mathbf{A}^T \mathbf{A}]_{ij}},\tag{26}$$

and the update (25) becomes

$$\boldsymbol{R}_{ij} \leftarrow \boldsymbol{R}_{ij} \frac{[\boldsymbol{A}^T \boldsymbol{S} \boldsymbol{A}]_{ij}}{[\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{R} \boldsymbol{A}^T \boldsymbol{A}]_{ij}}.$$
(27)

Since $S_{ij} \ge 0$ for all *i*, *j*, in both (23) and (27) each element of the multiplier matrix is positive if both $A \ge 0$ and $R \ge 0$ in all entries. Therefore, initializing both matrices with positive values results in an update step that keeps the elements of A and R positive.

To extend this rule to tensor DEDICOM, consider that the analytical derivatives hold for \underline{R} and A by considering each slice \underline{S}_l and \underline{R}_l individually:

$$\frac{\partial \mathcal{L}(\underline{S}_l, A, \underline{R}_l)}{\partial \underline{R}_l} = -2(A^T \underline{S}_l A - A^T A \underline{R}_l A^T A),$$
(28)

$$\frac{\partial \mathcal{L}(\underline{S}_l, A, \underline{R}_l)}{\partial A} = -2(\underline{S}_l^T A \underline{R}_l + \underline{S}_l A \underline{R}_l^T - A(\underline{R}_l^T A^T A \underline{R}_l + \underline{R}_l A^T A \underline{R}_l^T)).$$
(29)

Since by (18) we have $\mathcal{L}(\underline{S}, A, \underline{R}) = \sum_{l=1}^{t} \mathcal{L}(\underline{S}_{l}, A, \underline{R}_{l})$ we can derive the full gradients as

$$\frac{\partial \mathcal{L}(\underline{S}, A, \underline{R})}{\partial \underline{R}} = (A^T \underline{S} A - A^T A \underline{R} A^T A), \tag{30}$$

$$\frac{\partial \mathcal{L}(\underline{S}, A, \underline{R})}{\partial A} = \sum_{l=1}^{t} -2(\underline{S}_{l}^{T} A \underline{R}_{l} + \underline{S}_{l} A \underline{R}_{l}^{T} - A(\underline{R}_{l}^{T} A^{T} A \underline{R}_{l} + \underline{R}_{l} A^{T} A \underline{R}_{l}^{T})).$$
(31)

For *A* we set η^A as

$$\eta_{ij}^{\boldsymbol{A}} \coloneqq \frac{A_{ij}}{2\sum_{l=1}^{t} [\boldsymbol{A}(\underline{\boldsymbol{R}}_{l}^{T} \boldsymbol{A}^{T} \boldsymbol{A} \underline{\boldsymbol{R}}_{l} + \underline{\boldsymbol{R}}_{l} \boldsymbol{A}^{T} \boldsymbol{A} \underline{\boldsymbol{R}}_{l}^{T})]_{ij}}.$$
(32)

Then the update step is

$$A_{ij} \leftarrow A_{ij} - \eta_{ij}^{A} \frac{\partial \mathcal{L}(\underline{S}, A, \underline{R})}{\partial A}$$
(33)

$$= A_{ij} \frac{\sum_{l=1}^{t} [\underline{S}_{l}^{T} A \underline{R}_{l} + \underline{S}_{l} A \underline{R}_{l}^{T}]}{\sum_{l=1}^{t} [A(\underline{R}_{l}^{T} A^{T} A \underline{R}_{l} + \underline{R}_{l} A^{T} A \underline{R}_{l}^{T})]_{ij}}.$$
(34)

For <u>**R**</u> we again set

$$\eta_{lij}^{\underline{R}} \coloneqq \frac{\underline{R}_{lij}}{2[\underline{A}^T A \underline{R}_l A^T A]_{ij}},$$
(35)

and the update (25) becomes

$$\underline{R}_{lij} \leftarrow \underline{R}_{lij} \frac{[A^T \underline{S}_l A]_{ij}}{[A^T A \underline{R}_l A^T A]_{ij}}.$$
(36)

Equations (23) and (27) provide multiplicative update rules that ensure the non-negativity of *A* and *R* without any additional constraints. Equations (33) and (36) provide the corresponding rules for matrix *A* and tensor <u>*R*</u> in tensor DEDICOM.

3.3. On Symmetry

The DEDICOM algorithm is able to factorize both symmetrical and asymmetrical matrices *S*. For a given matrix *A*, the symmetry of *R* dictates the symmetry of the product ARA^{T} , since

$$(ARA^{T})_{ij} = \sum_{b=1}^{k} \sum_{c=1}^{k} A_{ib}R_{bc}A_{jc} = \sum_{b=1}^{k} \sum_{c=1}^{k} A_{ib}R_{cb}A_{jc}$$
(37)

$$=\sum_{c=1}^{k}\sum_{b=1}^{k}A_{jc}\boldsymbol{R}_{cb}\boldsymbol{A}_{ib}=(\boldsymbol{A}\boldsymbol{R}\boldsymbol{A}^{T})_{ji}$$
(38)

if $R_{cb} = R_{bc}$ for all b, c. We therefore expect a symmetric matrix S to be decomposed into ARA^T with a symmetric R, which is confirmed by our experiments. Factorizing a non-symmetric matrix leads to a non-symmetric R, the asymmetric relation between items leads to asymmetric relations between the latent factors. The same relations hold for each slice \underline{S}_l and \underline{R}_l in tensor DEDICOM.

3.4. On Interpretability

We have

$$S_{ij} \approx \sum_{b=1}^{k} \sum_{c=1}^{k} A_{ib} R_{bc} A_{jc}, \qquad (39)$$

i.e., we can estimate the probability of co-occurrence of two words w_i and w_j from the word embeddings A_i and A_j and the matrix R, where A_i denotes the *i*-th row of A.

If we want to predict the co-occurrence between words w_i and w_j we consider the latent topics that make up the word embeddings A_i and A_j , and sum up each component from A_i with each component A_j with respect to the relationship weights given in R.

Two words are likely to have a high co-occurrence if their word embeddings have larger weights in topics that are positively connected by the R matrix. Likewise a negative entry $R_{b,c}$ makes it less likely for words with high weight in the topics b and c to occur in the same context. See Figure 3 for an illustrated example.



Figure 3. The affinity matrix R describes the relationships between the latent factors. Illustrated here are two word embeddings, corresponding to the words w_i and w_j . Darker shades represent larger values. In this example we predict a large co-occurrence at S_{ii} and S_{jj} because of the large weight on the diagonal of the R matrix. We predict a low co-occurrence at S_{ij} and S_{ji} since the large weights on A_{i1} and A_{j3} interact with low weights on R_{13} and R_{31} .

Having an interpretable embedding model provides value beyond analysis of the affinity matrix of a single document. The worth of word embeddings is generally measured in their usefulness for downstream tasks. Given a prediction model based on word embeddings as one of the inputs, further analysis of the model behavior is facilitated when latent input dimensions easily translate to semantic meaning.

In most word embedding models, the embedding vector of a single word is not particularly useful in itself. The information only lies in its relationship (i.e., closeness or cosine similarity) to other embedding vectors. For example, an analysis of the change of word embeddings and therefore the change of word meaning within a document corpus (for example a news article corpus) can only show how various words form different clusters or drift apart over time. Interpretability of latent dimensions would provide tools to also consider the development of single words within the given topics.

All considerations above hold for the three-dimensional tensor case, in which we analyze a slice \underline{R}_l together with the common word embedding matrix A to gain insight into the input data slice \underline{S}_l .

4. Experiments and Results

In the following section we describe our experimental setup in full detail (Our Python implementation to reproduce the results is available on https://github.com/LarsHill/text-dedicom-paper. Additionally, we provide a snapshot of our versions of the applied public datasets (Wikipedia articles and Amazon reviews).) and present our results on the simultaneous topic (relation) extraction and word embedding learning task. We compare these results against competing matrix and tensor factorization methods for topic modeling, namely NMF (including a Tucker-2 variation compatible with tensors), LDA and SVD.

4.1. Data

We conducted our experiments on three orthogonal text datasets which cover different text domains and allow for a thorough empirical analysis of our proposed methods.

The first corpus leveraged triplets of individual Wikipedia articles. The articles were retrieved as raw text via the official Wikipedia API using the wikipedia-api library. We differentiated between thematically similar (e.g., "Dolphin" and "Whale") and thematically different articles (e.g., "Soccer" and "Donald Trump"). Each article triplet was categorized

into one of three classes: All underlying Wikipedia articles were thematically different, two articles were thematically similar and one was different, and all articles were thematically similar. The previous paper [3] contained an extensive evaluation over 12 triples of articles in the supplementary material. In this work we focused on the three triples described in the previous main paper, namely

- 1. "Soccer", "Bee", "Johnny Depp",
- 2. "Dolphin", "Shark", "Whale", and
- 3. "Soccer", "Tennis", "Rugby".

Depending on whether the article triplets were represented as input matrix or tensor they were processed differently. In the case of a matrix input all three articles got concatenated to form a new artificially generated document. In the case of a tensor input the articles remained individual documents later which later represented slices in the tensor representation.

To analyze the topic extraction capability of constrained DEDICOM also on text which was rather prone to grammatical and syntactical errors, we utilized a subset of the Amazon review dataset [22]. In particular, we restricted ourselves to the "movie" product category and created a corpus consisting of six text documents holding the concatenated reviews from the following films respectively, "Toy Story 1", "Toy Story 3", "Frozen", "Monsters, Inc.", "Kung Fu Panda" and "Kung Fu Panda 2". Grouping the reviews by movie affiliation enabled us to generate a tensor representation of the corpus which we factorized via nonnegative tensor DEDICOM to analyze topic relations across movies. Table 1 lists the number of reviews per movie and shows that based on review count "Kung Fu Panda 1" was the most popular among the six films.

The third corpus represented a complete collection of New York Times news articles ranging from 1st September 2019 to 31st August 2020. The articles were taken from the New York Times website and covered a wide range of sections (see Table 2).

Instead of grouping the articles by section we binned and concatenated them by month yielding 12 news documents containing monthly information (see Table 3 for details on the article count per month). Thereby, the factorization of tensor DEDICOM allowed for an analysis of topic relations and their changes over time.

Movie	# Reviews
Toy Story 1	2491
Monsters, Inc.	3203
Kung Fu Panda 1	6708
Toy Story 3	1209
Kung Fu Panda 2	1208
Frozen	1292

Table 1. Amazon movie review corpus grouped by movie and number of reviews per slice of input tensor.

Before transforming the text documents to matrix or tensor representations we applied the following textual preprocessing steps. First, the whole text was made lower-cased. Second, we tokenized the text making use of the word-tokenizer from the nltk library and removed common English stop words, including contractions such as "you're" and "we'll". Lastly we cleared the text from all remaining punctuation and deleted digits, single characters and multi-spaces (see Table 4 for an overview of corpora statistics after preprocessing).

Section	# Articles
Politics	3204
U.S.	2610
Business	1624
New York	1528
Europe	988
Asia Pacific	839
Health	598
Technology	551
Middle East	443
Science	440
Economy	339
Elections	240
Climate	239
World	233
Africa	124
Australia	113
Canada	104

Table 2. New York Times news corpus composition by section and number of articles.

Table 3. New York Times news corpus grouped by month and number of articles. This corresponds to the number of articles per slice of input tensor.

Month	# Articles
September 2019	1586
October 2019	1788
November 2019	1623
December 2019	1461
January 2020	1725
Febuary 2020	1602
March 2020	1937
April 2020	1712
May 2020	1713
June 2020	1828
July 2020	1814
August 2020	1886

Table 4. Overview of word count statistics after preprocessing for all datasets. Columns represent from left to right the total number of words per corpus, total number of unique words per corpus, average number of total words per article, average number of unique words per article and the cutoff frequency of the 10,000th most common word. Wikipedia article combinations: Dolphin, Shark, Whale (DSW), Soccer, Bee, Johnny Depp (SBJ), Soccer, Tennis, Rugby (STR).

	Total	Unique	Avg. Total	Avg Unique	Cutoff
Amazon Reviews	252,400	15,560	15.2	13.4	1
Wikipedia DSW	14,500	4376	4833.3	2106.0	1
Wikipedia SBJ	10,435	4034	3478.3	1600.3	1
Wikipedia STR	11,501	3224	3833.7	1408.0	1
New York Times	12,043,205	141,591	582.5	366.5	118

Next, we utilized all preprocessed documents in a corpus to extract a fixed size vocabulary of n = 10,000 most frequent tokens. Since our dense input tensor was of dimensionality $t \times n \times n$, a larger vocabulary size led to a significant increase in memory consumption. Based on the total number of unique corpora words reported in Table 4, a maximum vocabulary size of n = 10,000 was reasonable for the three Wikipedia corpora and the Amazon reviews corpus. Only the New York Times dataset could potentially have benefited from a larger vocabulary size.

Based on this vocabulary a symmetric word co-occurrence matrix was calculated for each of the corpus documents. When generating the matrix we only considered context words within a symmetrical window around the base word. Analysis in [2,3] showed that the window size in the range of 6 to 10 had little impact on performance. Thus, following our implementation in [3], we chose a window size of 7, the default in the original glove implementation. Like in [2], each context word only contributed 1/d to the total word pair count, given it was d words apart from the base word. To avoid any bias or prior information from the structure and order of the concatenated Wikipedia articles, reviews or news articles, we randomly shuffled the vocabulary before creating the co-occurrence matrix. As described in Section 3 we then transformed the co-occurrence matrix to a positive PMI matrix. If the corpus consisted of just one document the generated PPMI matrix functions as input for the row-stochastic DEDICOM algorithm. If the corpus consisted of several documents (e.g., one news document per month) the individual PPMI matrices were stacked to a tensor which in turn represented the input for the non-negative tensor DEDICOM algorithm.

The next section sheds light upon the training process of row-stochastic DEDICOM, non-negative tensor DEDICOM and the above mentioned competing matrix and tensor factorization methods, which will be benchmarked against our results in Section 4.3 and in the Appendixes A and B.

4.2. Training

As thoroughly outlined in Section 3 we trained both the row-stochastic DEDICOM and non-negative tensor DEDICOM with the alternating gradient descent paradigm.

In the case of a matrix input and a row-stochasticity constraint on A we utilize automatic differentiation from the PyTorch library to perform update steps on A and R. First, we initialized the factor matrices $A \in \mathbb{R}^{n \times k}$ and $R \in \mathbb{R}^{k \times k}$, by randomly sampling all elements from a uniform distribution centered around 1, U(0,2). Note that after applying the softmax operation on A all rows of A were stochastic. Therefore, scaling R by

$$\bar{s} \coloneqq \frac{1}{n^2} \sum_{ij}^n S_{ij},\tag{40}$$

would result in the initial decomposition $A'R(A')^T$ yielding reconstructed elements in the range of \bar{s} , the element mean of the PPMI matrix S, and thus, speeding up convergence. Second, A and R were iteratively updated employing the Adam optimizer [23] with constant individual learning rates of $\eta^A = 0.001$ and $\eta^R = 0.01$ and hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$. Both learning rates were identified through an exhaustive grid search. We trained for num_epochs = 15,000 until convergence, where each epoch consisted of an alternating gradient update with respect to A and R. Algorithm 1 illustrates the just described training procedure.

In the case of a tensor input and an additional non-negativity constraint on \underline{R} we noticed an inferior training performance with automatic differentiation methods. Hence, due to faster and more stable training convergence and improved qualitative results, we updated A and \underline{R} iteratively via derived multiplicative update rules enforcing non-negativity. Again, we initialized $A \in \mathbb{R}^{n \times k}$ and $\underline{R} \in \mathbb{R}^{t \times k \times k}$, by randomly sampling all elements from a uniform distribution centered around 1, U(0, 2). In order to ensure that the initialized components yielded a reconstructed tensor whose elements were in the same range of the input, we calculated an appropriate scaling factor for each tensor slice \underline{S}_l as

$$\alpha_l \coloneqq \left(\frac{\bar{s}_l}{k^2}\right)^{\frac{1}{3}}, \quad \text{where} \quad \bar{s}_l \coloneqq \frac{1}{n^2} \sum_{ij}^n \underline{S}_{lij}.$$
(41)

Algorithm 1 The row-stochastic DEDICC	DM algorithm
1: initialize $A, R \leftarrow U(0, 2) \cdot \bar{s}$	\triangleright See Equation (40) for the definition of \bar{s}
2: initialize $\beta_1, \beta_2, \epsilon$	> Adam algorithm hyperparameters
3: initialize η^A , η^R	Individual learning rates
4: for <i>i</i> in 1,, num_epochs do	
5: Calculate loss $\mathcal{L} = \mathcal{L}(S, A, R)$	⊳ See Equation (7)
6: $A \leftarrow A - \operatorname{Adam}_{\beta_1,\beta_2,\epsilon}(\nabla_A,\eta^A),$	where $\nabla_A = rac{\partial \mathcal{L}}{\partial A}$
7: $R \leftarrow R - \operatorname{Adam}_{\beta_1,\beta_2,\epsilon}(\nabla_R,\eta^R),$	where $\nabla_{R} = \frac{\partial \mathcal{L}}{\partial R}$
8: return A' and R , where $A' = row_so$	oftmax(col_norm(A)) \triangleright See Equation (5)

Next, we scaled *A* by $\bar{\alpha} = \frac{1}{t} \sum_{l=1}^{t} \alpha_l$ and each slice \underline{R}_l by α_l before starting the alternating multiplicative update steps for num_epochs = 300. The detailed derivation of the update rules is found in Section 3.2 and their iterative application in the training process is described in Algorithm 2.

Algorithm 2 The non-negative tensor DEDICOM algorithm						
1: initialize $A, \underline{R} \leftarrow U(0, 2)$						
2: scale <i>A</i> by $\bar{\alpha}$ and <u><i>R</i></u> _{<i>l</i>} by α_l	▷ See Equation (41) for the definitions of $\bar{\alpha}$ and α_l					
3: for <i>i</i> in 1,, num_epochs do						
4: Calculate loss $\mathcal{L} = \mathcal{L}(\underline{S}, A, \underline{I})$	$\underline{\mathbf{R}}) \qquad \qquad \triangleright \text{ See Equation (17)}$					
5: $A_{ij} \leftarrow A_{ij} \frac{\left[\sum_{l=1}^{t} \left(\underline{S}_{l} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T} A \underline{R}_{l}^{T}\right)}}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T}\right)}}{\left[A \sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T} A \underline{R}_{l}^{T}\right) + A_{ij} \frac{\sum_{l=1}^{t} \left(\underline{R}_{l} A^{T} A \underline{R}_{l}^{T} A$	$\frac{\left[I + \underline{S}_{l}^{T} A \underline{R}_{l}\right]_{ij}}{\underline{\mathbf{Z}}_{l}^{T} + \underline{\mathbf{R}}_{l}^{T} A^{T} A \underline{\mathbf{R}}_{l})]_{ij}}$					
6: $\underline{R}_{lij} \leftarrow \underline{R}_{lij} \frac{[A^T \underline{S}_l A]_{ij}}{[A^T A \underline{R}_l A^T A]_{ij}}$						
7: return <i>A</i> and <u><i>R</i></u>						

We implemented NMF, LDA and SVD using the sklearn library. In all cases the learnable factor matrices were initialized randomly and default hyperparameters were applied during training. For NMF the multiplicative update rule from [14] was utilized.

Figure 4 shows the convergence behavior of the row-stochastic matrix DEDICOM training process and the final loss of NMF and SVD. Note that LDA optimized a different loss function, which is why the calculated loss was not comparable and therefore excluded. We see that the final loss of DEDICOM was located just above the other losses, which is reasonable when considering the row stochasticity contraint on A and the reduced parameter amount of $nk + k^2$ compared to NMF (2*nk*) and SVD (2*nk* + k^2).



Figure 4. Reconstruction loss development during matrix factorization training. The *x*-axis plots the number of epochs, the *y*-axis plots the corresponding reconstruction error for each method.

To also have a benchmark model for our constrained tensor DEDICOM methods to compare against, we implemented a Tucker-2 variation of NMF, named tensor NMF (TNMF), which factorized the input tensor \underline{S} as

$$\underline{S}_l \approx \underline{W} \underline{H}_l.$$
 (42)

Its training procedure closely followed the above described alternating gradient descent approach for non-negative tensor DEDICOM. However, due to the two-way factorization (three-way for DEDICOM) the scaling factor α_l to properly initialize \underline{W} and H had to be adapted to

$$\alpha_l := \left(\frac{\bar{s}_l}{k}\right)^{\frac{1}{2}}, \quad \text{where} \quad \bar{s}_l := \frac{1}{n^2} \sum_{ij}^n \underline{S}_{lij}.$$
(43)

Analogous to Figure 4, we compared the training stability and convergence speed of our implemented tensor factorization methods. In particular, Figure 2 visualizes the reconstruction loss development for non-negative tensor DEDICOM trained via multiplicative update rules, row-stochastic tensor DEDICOM trained with automatic differentiation and the Adam optimizer and tensor NMF. It could be clearly observed that row-stochastic tensor DEDICOM converged much slower than the other two models trained with multiplicative update rules (learning rates were implicit here and did not have to be tuned).

4.3. Results

In the following, we present our results of training the above mentioned constrained DEDICOM factorizations on different text corpora to simultaneously learn interpretable word embeddings and meaningful topic clusters and their relations.

First, we focused our analysis on row-stochastic matrix DEDICOM applied to the synthetic Wikipedia text documents described in Section 4.1. For compactness reasons we primarily considered the document "Soccer, Bee and Johnny Depp", set the number of topics to k = 6 and refer to Appendix A.1 for the other article combinations and competing matrix factorization results. Second, we extended our evaluation to the tensor representation of the Wikipedia documents (t = 3, one article per tensor slice) and compared the performance of non-negative (multiplicative updates) and row-stochastic (Adam updates) tensor DEDICOM. Lastly, we applied non-negative tensor DEDICOM to the binned Amazon movie and New York Times news corpora to investigate topic relations across movies and over time. We again point the interested reader to Appendix A for additional results and the comparison to tensor NMF.

In the first step, we evaluated the quality of the learned latent topics by assigning each word embedding $A'_i \in \mathbb{R}^{1 \times k}$ to the latent topic dimension that represents the maximum value in A'_i , e.g.,

 $A'_i = \begin{bmatrix} 0.05 & 0.03 & 0.02 & 0.14 & 0.70 & 0.06 \end{bmatrix}$, $\operatorname{argmax}(A'_i) = 5$,

and thus, A'_i was matched to Topic 5. Next, we decreasingly sorted the words within each topic based on their matched topic probability. Table 5 shows the overall number of allocated words and the resulting top 10 words per topic together with each matched probability.

Table 5. The top 10 representative words per dimension of the basis matrix A', trained on the Wikipedia data as input matrix using automatic gradient methods.

	Topic 1 #619	Topic 2 #1238	Topic 3 #628	Topic 4 #595	Topic 5 #612	Topic 6 #389
1	ball	film	salazar	cup	bees	heard
1	(0.77)	(0.857)	(0.201)	(0.792)	(0.851)	(0.738)
2	penalty	starred	geoffrey	football	species	court
Ζ	(0.708)	(0.613)	(0.2)	(0.745)	(0.771)	(0.512)
2	may	role	rush	fifa	bee	depp
5	(0.703)	(0.577)	(0.2)	(0.731)	(0.753)	(0.505)
4	referee	series	brenton	world	pollen	divorce
	(0.667)	(0.504)	(0.199)	(0.713)	(0.658)	(0.454)
5	goal	burton	hardwicke	national	honey	alcohol
5	(0.66)	(0.492)	(0.198)	(0.639)	(0.602)	(0.435)
6	team	character	thwaites	uefa	insects	paradis
0	(0.651)	(0.465)	(0.198)	(0.623)	(0.576)	(0.42)
7	players	played	catherine	continental	food	relationship
/	(0.643)	(0.451)	(0.198)	(0.582)	(0.536)	(0.419)
8	player	director	kaya	teams	nests	abuse
0	(0.639)	(0.45)	(0.198)	(0.576)	(0.529)	(0.41)
Q	play	success	melfi	european	solitary	stating
9	(0.606)	(0.438)	(0.198)	(0.57)	(0.513)	(0.408)
10	game	jack	raimi	association	eusocial	stated
10	(0.591)	(0.434)	(0.198)	(0.563)	(0.505)	(0.402)

Indicated by the high assignment probabilities, one can see that columns 1, 2, 4, 5 and 6 represent distinct topics, which can easily be interpreted. Topic 1 and 4 were related to soccer, where 1 focused on the game mechanics and 4 on the organizational and professional aspect of the game. Topic 2 and 6 clearly referred to Johnny Depp, where 2 focused on his acting career and 6 on his difficult relationship to Amber Heard. The fifth topic obviously related to the insect "bee". In contrast, Topic 3 did not allow for any interpretation and all assignment probabilities were significantly lower than for the other topics.

Further, we analyzed the relations between the topics by visualizing the trained R matrix as a heatmap (see Figure 5c).



Figure 5. (a) 2-dimensional representation of word embeddings A' colored by topic assignment. (b) 2-dimensional representation of word embeddings A' colored by original Wikipedia article assignment (words that occur in more than one article are excluded). (c) Colored heatmap of affinity matrix *R*.

One thing to note was the symmetry of R which was a first indicator of a successful reconstruction, $A'R(A')^T$, (see Section 3.3). In addition, the main diagonal elements were consistently blue (positive), which suggested a high distinction between the topics. Although not very strong one could still see a connection between Topic 2 and 6 indicated by the light blue entry $R_{26} = R_{62}$. While the suggested relation between Topic 1 and 4 was not clearly visible, element $R_{14} = R_{41}$ was the least negative one for Topic 1. In order to visualize the topic cluster quality we utilized Uniform Manifold Approximation ad Projection (UMAP) [24] to map the *k*-dimensional word embeddings to a 2-dimensional space. Figure 5a illustrates this low-dimensional representation of A', where each word is colored based on the above described word to topic assignment. In conjunction with Table 5 one could nicely see that Topic 2 and 6 (Johnny Depp) and Topic 1 and 4 (Soccer) were close to each other. Hence, Figure 5a implicitly shows the learned topic relations as well.

As an additional benchmark, Figure 5b plots the same 2-dimensional representation, but now each word is colored based on the original Wikipedia article it belonged to. Words that occurred in more than one article were not considered in this plot.

Directly comparing Figure 5a,b shows that row-stochastic DEDICOM did not only recover the original articles but also found entirely new topics, which in this case represented subtopics of the articles. Let us emphasize that for all thematically similar article combinations, the found topics were usually not subtopics of a single article, but rather novel topics that might span across multiple Wikipedia articles (see for example Table A2 in the Appendix A). As mentioned at the top of this section, we are not only interested in learning meaningful topic clusters, but also in training interpretable word embeddings that capture semantic meaning.

Hence, we selected within each topic the two most representative words and calculated the cosine similarity between their word embeddings and all other word embeddings stored in A'. Table 6 shows the four nearest neighbors based on cosine similarity for the top two words in each topic. We observed a high thematical similarity between words with large cosine similarity, indicating the usefulness of the rows of A' as word embeddings.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	ball	film	salazar	cup	bees	heard
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	penalty	starred	geoffrey	fifa	bee	court
1	(0.994)	(0.978)	(1.0)	(0.995)	(0.996)	(0.966)
C	referee	role	rush	national	species	divorce
Ζ	(0.992)	(0.964)	(1.0)	(0.991)	(0.995)	(0.944)
2	may	burton	bardem	world	pollen	alcohol
3	(0.989)	(0.937)	(1.0)	(0.988)	(0.986)	(0.933)
4	goal	series	brenton	uefa	honey	abuse
4	(0.986)	(0.935)	(1.0)	(0.987)	(0.971)	(0.914)
0	penalty	starred	geoffrey	football	species	court
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	referee	role	rush	fifa	bees	divorce
1	(0.999)	(0.994)	(1.0)	(0.994)	(0.995)	(0.995)
r	goal	series	salazar	national	bee	alcohol
2	(0.998)	(0.985)	(1.0)	(0.983)	(0.99)	(0.987)
2	player	burton	brenton	cup	pollen	abuse
3	(0.997)	(0.981)	(1.0)	(0.983)	(0.99)	(0.982)
4	ball	film	thwaites	world	insects	settlement
4	(0.994)	(0.978)	(1.0)	(0.982)	(0.977)	(0.978)

Table 6. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed. Matrix A' trained on the wikipedia data as input matrix using automatic gradient methods.

In comparison to DEDICOM, other matrix factorization methods also provided a useful clustering of words into topics, with varying degree of granularity and clarity. However, the application of these methods as word embedding algorithms mostly failed on the word similarity task, with words close in cosine similarity seldom sharing the same thematical similarity we have seen in DEDICOM. This can be seen in Table A1, which shows for each method, NMF, LDA and SVD, the resulting word to topic clustering and the cosine nearest neighbors of the top two word embeddings per topic. While the individual topics extracted by NMF looked very reasonable, its word embeddings did not seem to carry any semantic meaning based on cosine similarity; e.g., the four nearest neighbors of "ball" were "invoke", "replaced", "scores" and "subdivided". A similar nonsensical picture can be observed for the other main topic words. LDA and SVD performed slightly better on the similar word task, although not all similar words appeared to be sensible, e.g., "children", "detective", "crime", "magazine" and "barber". In addition, some topics could not be clearly defined due to mixed word assignments, e.g., Topic 4 for LDA and Topic 1 for SVD.

Before shifting our analysis to the Amazon movie review and the New York Times news corpus we investigated factorizing the tensor representation of the "Soccer, Bee and Johnny Depp" Wikipedia documents. In particular, we compared the qualitative factorization results of row-stochastic and non-negative tensor DEDICOM trained with automatic differentiation and multiplicative update rules, respectively. Tables 7 and 8 in conjunction with Figures 6 and 7 show the extracted topics and their relations for both methods.



Figure 6. Colored heatmap of affinity tensor <u>**R**</u>, trained on the Wikipedia data represented as input tensor using automatic gradient methods.

	Topic 1 #481	Topic 2 #661	Topic 3 #414	Topic 4 #457	Topic 5 #316	Topic 6 #1711
1	hind	game	film	heard	bees	disorder
	(0.646)	(0.83)	(0.941)	(0.844)	(0.922)	(0.291)
n	segments	football	starred	court	bee	collapse
Ζ	(0.572)	(0.828)	(0.684)	(0.566)	(0.868)	(0.29)
2	bacteria	players	role	divorce	honey	attrition
3	(0.563)	(0.782)	(0.624)	(0.51)	(0.756)	(0.285)
4	legs	ball	series	depp	insects	losses
	(0.562)	(0.777)	(0.562)	(0.508)	(0.68)	(0.284)
F	antennae	team	burton	sued	food	invertebrate
5	(0.555)	(0.771)	(0.547)	(0.48)	(0.634)	(0.283)
6	females	may	character	stating	species	rate
0	(0.549)	(0.696)	(0.499)	(0.45)	(0.599)	(0.283)
7	wings	play	success	alcohol	nests	businesses
/	(0.547)	(0.692)	(0.494)	(0.449)	(0.596)	(0.282)
0	small	competitions	played	paradis	flowers	virgil
0	(0.538)	(0.677)	(0.483)	(0.446)	(0.571)	(0.282)
0	groups	match	films	alleged	pollen	iridescent
9	(0.527)	(0.672)	(0.482)	(0.445)	(0.56)	(0.282)
10	males	penalty	box	stated	larvae	detail
10	(0.518)	(0.664)	(0.465)	(0.444)	(0.529)	(0.281)

Table 7. Top 10 representative words per dimension of the basis matrix A', trained on the wikipedia data as input tensor using automatic gradient methods.



Figure 7. Colored heatmap of affinity tensor <u>*R*</u>, trained on the Wikipedia data represented as input tensor using multiplicative update rules.

	Topic 1 #521	Topic 2 #249	Topic 3 #485	Topic 4 #871	Topic 5 #445	Topic 6 #1469
1	species	game	honey	allow	insects	depp
1	(3.105)	(3.26)	(2.946)	(0.668)	(2.794)	(2.419)
2	eusocial	football	bee	organised	pollen	film
2	(2.524)	(3.05)	(2.01)	(0.662)	(2.239)	(2.115)
2	solitary	players	beekeeping	winner	flowers	role
3	(2.279)	(2.699)	(1.933)	(0.632)	(2.019)	(1.32)
4	nest	ball	bees	officially	nectar	starred
4	(2.118)	(2.475)	(1.704)	(0.626)	(1.656)	(1.3)
F	females	may	increased	wins	wasps	actor
5	(1.993)	(2.447)	(1.589)	(0.617)	(1.602)	(1.155)
6	workers	team	humans	level	wings	series
0	(1.797)	(2.424)	(1.515)	(0.613)	(1.588)	(1.126)
7	nests	association	wild	free	many	burton
/	(1.787)	(1.92)	(1.415)	(0.6)	(1.588)	(1.112)
0	colonies	play	mites	constitute	hind	played
0	(1.722)	(1.834)	(1.4)	(0.596)	(1.577)	(1.068)
0	egg	referee	colony	regulation	hairs	heard
9	(1.692)	(1.809)	(1.35)	(0.595)	(1.484)	(1.005)
10	males	laws	beekeepers	prestigious	pollinating	success
10	(1.664)	(1.792)	(1.332)	(0.594)	(1.467)	(0.981)

Table 8. Top 10 representative words per dimension of the basis matrix *A*, trained on the wikipedia data as input tensor using multiplicative update rules.

It could be seen that non-negative tensor DEDICOM yielded a more interpretable affinity tensor \underline{R} (Figure 7) due to its enforced non-negativity. For example, it clearly highlighted the bee related Topics 1, 3 and 5 in the affinity tensor slice corresponding to the article "Bee". Moreover, all extracted topics in Table 8 were distinct and their relations were well represented in the individual slices of \underline{R} . In contrast, Topic 6 in Table 7 did not represent a meaningful topic, which was also indicated by the low probability scores of the ranked topic words. Although the results of the similar word evaluation were arguably better for row-stochastic tensor DEDICOM (see Tables 9 and 10) we prioritized topic extraction and relation quality. That is why in the further analysis of the Amazon review and New York Times news corpus we restriced tour evaluation to non-negative tensor DEDICOM.

As described in Section 4.1 our Amazon movie review corpus comprised human written reviews for six famous animation films. Factorizing its PPMI tensor representation with non-negative tensor DEDICOM and the number of topics set to k = 10 revealed not only movie-specific subtopics but also general topics that spanned over several movies. For example, Topics 1, 9 and 10 in Table 11 could uniquely be related to the films "Frozen", "Toy Story 1" and "Kung Fu Panda 1", respectively, whereas Topic 5 constituted bonus material on a DVD which held true for all films. The latter could also be seen in Figure 8 where Topic 5 was highlighted in each movie slice (strongly in the top and lightly in the bottom row). In the same sense one could observe that Topic 3 was present in both "Kung Fu Panda 1" and "Kung Fu Panda 2", which is reasonable considering the topic depicted the general notion of a fearsome warrior.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	hind	game	film	heard	bees	disorder
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	segments	football	starred	court	bee	collapse
1	(0.995)	(1.0)	(0.968)	(0.954)	(0.999)	(1.0)
r	legs	players	role	divorce	honey	losses
2	(0.994)	(0.999)	(0.954)	(0.925)	(0.99)	(1.0)
2	antennae	ball	series	sued	insects	attrition
3	(0.993)	(0.999)	(0.951)	(0.907)	(0.976)	(0.999)
4	wings	team	burton	alleged	food	businesses
4	(0.992)	(0.998)	(0.945)	(0.897)	(0.97)	(0.999)
0	segments	football	starred	court	bee	collapse
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	antennae	game	role	divorce	bees	disorder
T	(1.0)	(1.0)	(0.993)	(0.996)	(0.999)	(1.0)
2	wings	players	series	sued	honey	losses
2	(0.999)	(0.999)	(0.978)	(0.991)	(0.995)	(0.999)
3	bacteria	ball	burton	alleged	insects	pesticide
5	(0.999)	(0.999)	(0.975)	(0.981)	(0.984)	(0.998)
4	legs	team	film	alcohol	food	businesses
4	(0.998)	(0.999)	(0.968)	(0.981)	(0.976)	(0.998)

Table 9. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed. Matrix A', trained on the wikipedia data as input tensor using automatic gradient methods.

Table 10. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed. Matrix *A*, trained on the wikipedia data as input tensor using multiplicative update rules.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	species	game	honey	allow	insects	depp
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	easier	football	boatwrights	emancipation	ultraviolet	charlie
T	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
n	tiny	players	glade	broadly	mechanics	infiltrate
2	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	halictidae	association	tutankhamun	disabilities	exploit	thenwife
5	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	provision	team	oracle	total	swallows	tourist
т	(0.999)	(0.997)	(1.0)	(1.0)	(1.0)	(1.0)
0	eusocial	football	bee	organised	pollen	film
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	oligocene	players	subfamilies	comes	honeybees	starred
T	(1.0)	(1.0)	(0.995)	(1.0)	(1.0)	(1.0)
r	architecture	game	internal	shows	enlarged	smoking
2	(1.0)	(1.0)	(0.994)	(1.0)	(0.998)	(0.999)
2	uncommon	association	studied	attention	simple	dislocated
5	(1.0)	(1.0)	(0.994)	(1.0)	(0.998)	(0.999)
1	termed	team	cladogram	deductions	drove	injuries
4	(1.0)	(0.997)	(0.99)	(1.0)	(0.998)	(0.999)



Figure 8. Colored heatmap of affinity tensor <u>*R*</u>, trained on the Amazon review data represented as input tensor using multiplicative update rules.

Table 11. Top 10 representative words per dimension of the basis matrix *A*, trained on the Amazon review data as input tensor using multiplicative update rules.

	Topic 1 #528	Topic 2 #445	Topic 3 #1477	Topic 4 #1790	Topic 5 #1917	Topic 6 #597	Topic 7 #789	Topic 8 #670	Topic 9 #1599	Topic 10 #188
1	anna	shen	legendary	lasseter	disc	screams	code	mike	woody	ро
1	(4.215)	(4.21)	(1.459)	(2.887)	(1.367)	(3.779)	(3.292)	(4.325)	(6.12)	(5.737)
2	elsa	peacock	valley	director	birds	energy	email	crystal	buzz	master
2	(4.087)	(2.668)	(1.448)	(2.392)	(1.343)	(3.315)	(2.781)	(4.055)	(5.484)	(5.276)
2	olaf	oldman	temple	andrew	widescreen	monstropolis	promo	billy	andy	shifu
3	(2.315)	(2.627)	(1.31)	(2.158)	(1.327)	(3.13)	(2.645)	(3.911)	(4.355)	(4.707)
4	trolls	gary	kim	stanton	outtakes	world	free	goodman	toys	dragon
4	(2.241)	(2.423)	(1.307)	(2.119)	(1.238)	(3.109)	(2.343)	(3.812)	(4.119)	(4.344)
5	frozen	lord	fearsome	special	extras	monsters	promotion	sully	lightyear	warrior
5	(2.196)	(2.201)	(1.288)	(1.892)	(1.185)	(3.047)	(2.279)	(3.728)	(3.334)	(4.274)
6	kristoff	weapon	teacher	pete	dvd	city	promotional	wazowski	allen	tai
0	(2.155)	(1.469)	(1.288)	(1.612)	(1.142)	(2.994)	(2.266)	(3.513)	(2.752)	(4.082)
7	queen	wolf	battle	ranft	included	monster	amazon	randall	tim	lung
/	(2.055)	(1.405)	(1.264)	(1.564)	(1.13)	(2.978)	(2.129)	(3.404)	(2.609)	(3.993)
0	hans	inner	duk	joe	short	power	click	sulley	hanks	five
0	(2.054)	(1.38)	(1.257)	(1.564)	(1.101)	(2.919)	(2.024)	(3.396)	(2.471)	(2.952)
0	sister	yeoh	train	feature	games	scare	download	buscemi	cowboy	oogway
9	(1.904)	(1.359)	(1.221)	(1.555)	(1.1)	(2.614)	(1.889)	(3.266)	(2.407)	(2.918)
10	ice	michelle	warriors	ralph	tour	closet	instructions	james	space	furious
10	(1.839)	(1.354)	(1.22)	(1.518)	(1.031)	(2.555)	(1.877)	(3.264)	(2.375)	(2.806)

Figure 9 and Table 12 refer to our experimental results on the dataset of New York Times news articles. We saw a diverse array of topics extracted from the text corpus, ranging from US-politics (Topics 4, 6, 7) to natural disasters (Topic 8), Hollywood sexual assault allegations (Topic 10) and the COVID epidemic both from a medical view (Topic 3) and a view on resulting restrictions to businesses (Topic 9).

The corresponding heatmap allowed us to infer when certain topics were most relevant in the last year. While the entries relating to the COVID pandemic remain light blue for the first half of the heatmap we sawa the articles picking up on the topic around March 2020, when the effects of the Coronavirus started hitting the US. Even comparatively smaller events like the conviction of Harvey Weinstein and the death of George Floyd triggering the racism debate in the US could be recognized in the heatmap, with a large deviation of Topic 10 around February 2020 and Topic 4 around June 2020.



Figure 9. Colored heatmap of affinity tensor \underline{R} , trained on the New York Times news article data represented as input tensor using multiplicative update rules.

Further empirical results on the Amazon review and New York Times news corpora, such as two-dimensional UMAP representations of the embedding matrix A and extracted topics from tensor NMF, can be found in Appendixes A.3 and A.4, respectively. For example, Table A21 shows that the tensor NMF factorization also extracted high quality topics but lacked the interpretable affinity tensor \underline{R} which was crucial in order to properly comprehend a topic development over time.

Table 12. Top 10 representative words per dimension of the basis matrix *A*, trained on the New York Times news article data as input tensor using multiplicative update rules.

	Topic 1 #454	Topic 2 #5984	Topic 3 #567	Topic 4 #562	Topic 5 #424	Topic 6 #330	Topic 7 #515	Topic 8 #297	Topic 9 #431	Topic 10 #436
1	suleimani	loans	masks	floyd	contributed	confederate	ukraine	storm	restaurants	weinstein
1	(2.812)	(0.618)	(3.261)	(3.376)	(4.565)	(3.226)	(3.191)	(2.76)	(2.948)	(3.442)
2	iran	university	protective	minneapolis	reporting	statue	sondland	hurricane	bars	sexual
4	(2.593)	(0.551)	(2.823)	(2.551)	(2.788)	(2.649)	(2.881)	(2.622)	(2.021)	(2.71)
2	iraq	oil	gloves	police	michael	statues	zelensky	winds	reopen	rape
3	(2.453)	(0.549)	(2.516)	(2.255)	(2.707)	(2.416)	(2.133)	(1.715)	(1.684)	(2.102)
4	iranian	billion	ventilators	george	katie	monuments	ambassador	tropical	gyms	assault
4	(2.408)	(0.54)	(2.22)	(2.088)	(2.324)	(1.815)	(1.976)	(1.606)	(1.654)	(1.861)
E	iraqi	loan	surgical	protests	alan	monument	ukrainian	storms	stores	jury
5	(1.799)	(0.468)	(2.032)	(1.936)	(2.292)	(1.375)	(1.789)	(1.439)	(1.638)	(1.513)
6	baghdad	bonds	gowns	brutality	emily	flag	giuliani	coast	theaters	charges
0	(1.604)	(0.466)	(1.965)	(1.765)	(2.165)	(1.206)	(1.755)	(1.415)	(1.627)	(1.409)
7	qassim	payments	equipment	racism	nicholas	richmond	volker	laura	salons	predatory
/	(1.599)	(0.456)	(1.86)	(1.579)	(2.096)	(1.109)	(1.754)	(1.259)	(1.438)	(1.387)
0	strike	edited	supplies	knee	cochrane	symbols	investigations	isaias	closed	harvey
0	(1.597)	(0.452)	(1.816)	(1.435)	(1.934)	(1.089)	(1.602)	(1.217)	(1.424)	(1.35)
0	gen	trillion	gear	killing	rappeport	remove	testified	category	shops	guilty
2	(1.513)	(0.451)	(1.742)	(1.429)	(1.613)	(1.058)	(1.584)	(1.192)	(1.325)	(1.312)
10	maj	graduated	mask	officers	maggie	removal	testimony	landfall	indoor	sex
10	(1.504)	(0.449)	(1.502)	(1.405)	(1.529)	(1.003)	(1.558)	(1.106)	(1.247)	(1.3)

5. Conclusions and Outlook

We propose a constrained version of the DEDICOM algorithm that is able to factorize the pointwise mutual information matrices of text documents into meaningful topic clusters all the while providing interpretable word embeddings for each vocabulary item. Our study on semi-artificial data from Wikipedia articles has shown that this method recovers the underlying structure of the text corpus and provides topics with thematic granularity, meaning the extracted latent topics are more specific than a simple clustering of articles. A comparison to related matrix factorization methods has shown that the combination of relation aware topic modeling and interpretable word embedding learning given by our algorithm is unique in its class.

Extending this algorithm to factorize three-dimensional input tensors allows for the study of changes in the relations between topics across subsets of a structured text corpus, e.g., news articles grouped by time period. Algorithmically, this can be solved via alternating gradient descent by either automatic gradient methods or by applying multiplicative update rules which decrease training time drastically and enhance training stability.

Due to memory constraints from matrix multiplications of high dimensional dense tensors our proposed approach is currently limited in vocabulary size or time dimension.

In further work we aim for developing algorithms capable of leveraging sparse matrix multiplications to avoid the above mentioned memory constraints. In addition, we plan to expand on the possibilities of constraining the factor matrices and tensors when applying a multiplicative update rule and further analyze the behavior of the factor tensors, for example by utilizing time series analysis to discover temporal relations between extracted topics and to potentially identify trends. Finally, further analysis may include additional quantitative evaluations of our proposed methods' topic modeling performance with competing approaches.

Author Contributions: Conceptualization, D.B.; Methodology, L.H.; Project administration, L.H. and D.B.; Supervision, C.B. and R.S.; Writing—original draft, L.H. and D.B. All authors have read and agreed to the published version of the manuscript.

Funding: The authors of this work were supported by the Competence Center for Machine Learning Rhine Ruhr (ML2R) which is funded by the Federal Ministry of Education and Research of Germany (grant no. 01 | S18038C). We gratefully acknowledge this support.

Data Availability Statement: Publicly available datasets (Amazon reviews, Wikipedia articles) were analyzed in this study. This data can be found here: https://github.com/LarsHill/text-dedicom-paper. The NYT news article data presented in this study are available on request from the corresponding author. The data are not publicly available due to potential copyright concerns.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Additional Results

Appendix A.1. Additional Results on Wikipedia Data as Matrix Input

Articles: "Soccer", "Bee", "Johnny Depp".

		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
		#619	#1238	#628	#595	#612	#389
	1	ball	bees	film	football	heard	album
	2	may	species	starred	cup	depp	band
	3	penalty	bee	role	world	court	guitar
	4	referee	pollen	series	fifa	alcohol	vampires
	5	players	honey	burton	national	relationship	rock
	6	team	insects	character	association	stated	hollywood
	7	goal	food	films	international	divorce	song
	8	game	nests	box	women	abuse	released
	9	player	solitary	office	teams	paradis	perry
	10	player	eusocial	iack	uefa	stating	debut
NMF			1	GI	(d 1)	1 1	"
	0	ball	bees	film hourt and insisted	football	neard	album
	1	invoke	odors	burtondirected	athenaeus	crew	jones
	2	replaced		landar	paratympic	aneging	marilyn
	3	scores	colour	landau	governing	oped	roots
	4	subdivided	affected	brother	varieties	asserted	drums
	0	may	species	starred	cup	depp	band
	1	yd	niko	shared	inaugurated	refer	heroes
	2	ineffectiveness	commercially	whitaker	confederation	york	bowie
	3	tactical	microbiota	eccentric	gold	leaders	debut
	4	slower	strategies	befriends	headquarters	nonindian	solo
		#577	#728	#692	#607	#663	#814
	1	film	football	depp	penalty	bees	species
	2	series	women	children	heard	flowers	workers
	3	man	association	life	ball	bee	solitary
	4	played	fifa	role	direct	honey	players
	5	pirates	teams	starred	referee	pollen	colonies
	6	character	games	alongside	red	food	eusocial
	7	along	world	actor	time	increased	nest
	8	cast	cup	stated	goal	pollination	may
	9	also	game	burton	scored	times	size
LDA	10	hollow	international	playing	player	larvae	egg
	0	film	football	depp	penalty	bees	species
	1	charlie	cup	critical	extra	bee	social
	2	near	canada	february	kicks	insects	chosen
	3	thinking	zealand	script	inner	authors	females
	4	shadows	activities	song	moving	hives	subspecies
	0	series	women	children	heard	flowers	workers
	1	crybaby	fifa	detective	allison	always	carcases
	2	waters	opera	crime	serious	eusociality	lived
	3	sang	exceeding	magazine	allergic	varroa	provisioned
	4	cast	cuju	barber	cost	wing	cuckoo
		#1228	#797	#628	#369	#622	#437
	1	bees	depp	game	cup	heard	beekeeping
	2	also	film	ball	football	court	increased
	3	bee	starred	team	fifa	divorce	honey
	4	species	role	players	world	stating	described
SVD	5	played	series	penalty	european	alcohol	use
	6	time	burton	play	uefa	paradis	wild
	7	one	character	may	national	documents	varroa
	8	first	actor	referee	europe	abuse	mites
	9	two	released	competitions	continental	settlement	colony
	10	pollen	release	laws	confederation	sued	flowers
	0	bees	depp	game	cup	heard	beekeeping
	1	bee	iii	correct	continental	alleging	varroa
	2	develops	racism	abandoned	contested	attempting	animals
	3	studied	appropriation	maximum	contederations	finalized	mites
SVD	4	crops	march	clear	conmebol	submitted	plato
210	0	also	film	ball	football	court	increased
	1	although	waters	finely	er	declaration	usage
	2	told	robinson	poised	suffix	issued	farmers
	3	chosen	scott	worn	word	restraining	mentioned
	4	stars	costars	manner	appended	verbally	aeneid

Table A1. For each evaluated matrix factorization method we display the top 10 words for each topic and the five most similar words based on cosine similarity for the two top words from each topic.

Articles: "Dolphin", "Shark", "Whale".

	Topic 1 #460	Topic 2 #665	Topic 3 #801	Topic 4 #753	Topic 5 #854	Topic 6 #721
1	shark	calf	ship	conservation	water	dolphin
1	(0.665)	(0.428)	(0.459)	(0.334)	(0.416)	$(0.\hat{6}91)$
2	sharks	months	became	countries	similar	dolphins
-	(0.645)	(0.407)	(0.448)	(0.312)	(0.374)	(0.655)
3	fins	calves	poseidon	government	tissue	captivity
	(0.487) killod	(0.407)	(0.44) riding	(0.309)	(0.373)	(0.549)
4	(0.454)	(0.300)	(0.426)	(0.304)	(0.365)	(0.467)
_	million	blubber	dionysus	bycatch	swimming	behavior
5	(0.451)	(0.374)	(0.422)	(0.29)	(0.357)	(0.461)
6	fish	voung	ancient	cancelled	blood	bottlenose
6	(0.448)	(0.37)	(0.42)	(0.288)	(0.346)	(0.453)
7	international	sperm	deity	eastern	surface	sometimes
'	(0.442)	(0.356)	(0.412)	(0.287)	(0.344)	(0.449)
8	tin	born	ago	policy	oxygen	human
	(0.421) Galainan	(0.355)	(0.398)	(0.286)	(0.34)	(0.421)
9	(0.405)	(0.349)	(0.395)	(0.285)	(0.336)	(0.42)
	(0.405) teeth	(0.349) mysticetes	(0.393) greeks	imminent	(0.330) swim	various
10	(0.398)	(0.341)	(0.394)	(0.282)	(0.336)	(0.418)
0	shark	calf	ship	conservation	water	dolphin
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	sharks	calves	dionysus	south	prey	dolphins
-	(0.981)	(0.978)	(0.995)	(0.981)	(0.964)	(0.925)
3	fins	females	riding	states	swimming	sometimes
	(0.958) killod	(0.976)	(0.992) doity	(0.981)	(0.959)	(0.909)
4	(0.020)	(0.955)	(0.992)	(0.978)	(0.957)	(0.904)
_	fishing	voing	poseidon	endangered	(0.957) swim	bottlenose
5	(0.916)	(0.948)	(0.987)	(0.976)	(0.947)	(0.903)
0	sharks	months	became	countries	similar	dolphins
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	shark	born	old	eastern	surface	behavior
-	(0.981)	(0.992)	(0.953)	(0.991)	(0.992)	(0.956)
3	fins	young	later	united	brain	sometimes
	(0.936)	(0.992)	(0.946)	(0.969)	(0.97)	(0.945)
4	tiger	sperm	ago	caught	sound	various
	(0.894) killed	(0.985)	(0.939) modern	(U.907)	(0.968) object	(0.943)
5	(0.887)	(0.984)	(0.937)	(0.979)	(0.965)	(0.937)
	(0.007)	(0.201)	(0.207)	(0.777)	(0.200)	(0.707)

Table A2. Top half lists the top 10 representative words per dimension of the basis matrix *A*, bottom half lists the five most similar words based on cosine similarity for the two top words from each topic.

Articles: "Dolphin", "Shark", "Whale'.

		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
		#492	#907	#452	#854	#911	#638
	1	blood	international	evidence	sonar	ago	calf
	2	body	killed	selfawareness	may	teeth	young
	3	heart	states	ship	surface	million	females
	4	gills	conservation	dionysus	clicks	mysticetes	captivity
	5	bony	new	came	prey	whales	calves
	6	oxygen	united	another	use	years	months
	7	organs	shark	important	underwater	baleen	born
	8	tissue	world	poseidon	sounds	cetaceans	species
	9	water	endangered	mark	known	modern	male
NMF	10	via	islands	riding	similar	extinct	female
	0	blood	international	evidence	sonar	ago	calf
	1	travels	proposal	flaws	poisoned	consist	uninformed
	2	enters	lipotidae	methodological	signals	specialize	primary
	3	vibration	banned	nictating	≈- ,	legs	born
	4	tolerant	iniidae	wake	emitted	closest	leaner
	0	body	killed	selfawareness	may	teeth	young
	1	crystal	law	legendary	individuals	ruei	brood
	2	DIOCKS	consumers	numankina	neiping	lamp	lacking
	3	modified	pontoporndae	neipers	waste	Interfeeding	accurate
	4	Silts	org	performing	depression	Kfill	consistency
	1	#650	#785	#695	#815	#635	#6/4
	1	killed	teeth	head	species	meat	air
	2	system	baleen	fish dalahin	male	whale	using
	3	endangered	mysticetes	dolphin	remales	ft Gale ani an	causing
	4 5	olien	ago	IIII	whales	lisheries	currents
	6	close	family	eyes	sometimies	also	sounds
	7	countries	water	naw	voung	threate	groups
	8	since	includes	nopular	shark	children	research
	9	called	allow	tissue	female	nonulation	clicks
	10	vessels	oreater	tail	wild	bottom	burst
LDA		killed	teeth	head	species	meat	air
	1	nostures	dense	underside	along	norbeagle	australis
	2	dolphinariums	cetacea	grooves	another	source	submerged
	3	town	tourism	evesight	long	activities	melbourne
	4	onethird	planktonfeeders	osmoregulation	sleep	comparable	spear
	0	system	baleen	fish	male	whale	using
	1	dominate	mysticetes	mostly	females	live	communication
	2	close	distinguishing	swim	aorta	human	become
	3	controversy	unique	due	female	cold	associated
	4	agree	remove	whole	position	parts	mirror
		#1486	#544	#605	#469	#539	#611
	1	dolphins	water	shark	million	poseidon	dolphin
	2	species	body	sharks	years	became	meat
	3	whales	tail	fins	ago	ship	family
	4	fish	teeth	international	whale	riding	river
	5	also	flippers	killed	two	evidence	similar
	6	large	tissue	fishing	calf	melicertes	extinct
	7	may	allows	tin	mya	deity	called
	8	one	air	law	later	ino	used
	9	animals	feed	new	months	came	islands
SVD	10	use	bony	conservation	mysticetes	made	genus
	0	dolphins	water	shark	million	poseidon	dolphin
	1	various	vertical	corpse	approximately	games	aepictea
	2	militare	uniike	STOCKS	assigned	phaianthus	makara
	3 4	selfmade	lack	galea galeomorphii	appeared	statue isthmian	capensis goddess
		species	hody	charke	Voare	hecame	meat
	1	herd	heart	mostly	acanthodiane	pirates	contaminated
	2	reproduction	resistino	fda	spent	elder	harpoon
	3	afford	fit	lists	stretching	mistook	practitioner
	4	maturity	posterior	carcharias	informal	wealthy	pcbs

Table A3. For each evaluated matrix factorization method we display the top 10 words for each topic and the five most similar words based on cosine similarity for the two top words from each topic.

Articles: "Soccer", "Tennis", "Rugby".

	Topic 1 #539	Topic 2 #302	Topic 3 #563	Topic 4 #635	Topic 5 #650	Topic 6 #530
1	may	leads	tournaments	greatest	football	net
1	(0.599)	(0.212)	(0.588)	(0.572)	(0.553)	(0.644)
\mathbf{r}	penalty	sole	tournament	tennis	rugby	shot
Ζ	(0.576)	(0.205)	(0.517)	(0.497)	(0.542)	(0.629)
З	referee	competes	events	female	south	stance
5	(0.564)	(0.205)	(0.509)	(0.44)	(0.484)	(0.553)
4	team	extending	prize	ever	union	stroke
т	(0.517)	(0.204)	(0.501)	(0.433)	(0.47)	(0.543)
5	goal	fixing	tour	navratilova	wales	serve
0	(0.502)	(0.203)	(0.497)	(0.405)	(0.459)	(0.537)
6	kick	triggered	money	modern	national	rotation
0	(0.459)	(0.203)	(0.488)	(0.401)	(0.446)	(0.513)
7	play	bleeding	cup	best	england	backhand
,	(0.455)	(0.202)	(0.486)	(0.4)	(0.438)	(0.508)
8	ball	fraud	world	wingfield	new	hit
0	(0.452)	(0.202)	(0.467)	(0.394)	(0.416)	(0.507)
9	offence	inflammation	atp	sports	europe	forehand
/	(0.444)	(0.202)	(0.464)	(0.39)	(0.406)	(0.499)
10	foul	conditions	men	williams	states	torso
10	(0.443)	(0.201)	(0.463)	(0.389)	(0.404)	(0.487)
0	may	leads	tournaments	greatest	football	net
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
r	goal	tiredness	events	female	union	shot
2	(0.98)	(1.0)	(0.992)	(0.98)	(0.98)	(0.994)
2	play	ineffectiveness	tour	ever	rugby	serve
5	(0.959)	(1.0)	(0.989)	(0.971)	(0.979)	(0.987)
4	penalty	recommences	money	navratilova	association	hit
т	(0.954)	(1.0)	(0.986)	(0.967)	(0.96)	(0.984)
5	team	mandated	prize	tennis	england	stance
5	(0.953)	(1.0)	(0.985)	(0.962)	(0.958)	(0.955)
0	penalty	sole	tournament	tennis	rugby	shot
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
n	referee	discretion	events	greatest	football	net
Ζ	(0.985)	(1.0)	(0.98)	(0.962)	(0.979)	(0.994)
\mathbf{r}	kick	synonym	event	female	union	serve
3	(0.985)	(1.0)	(0.978)	(0.953)	(0.975)	(0.987)
4	offence	violated	atp	year	england	hit
4	(0.982)	(1.0)	(0.974)	(0.951)	(0.961)	(0.983)
Б	foul	layout	money	navratilova	wales	stance
9	(0.982)	(1.0)	(0.966)	(0.949)	(0.949)	(0.98)

Table A4. Top half lists the top 10 representative words per dimension of the basis matrix *A*, bottom half lists the five most similar words based on cosine similarity for the two top words from each topic.



Figure A1. (a) 2-dimensional representation of word embeddings A' colored by topic assignment. (b) 2-dimensional representation of word embeddings A' colored by original Wikipedia article assignment (words that occur in more than one article are excluded). (c) Colored heatmap of affinity matrix *R*.



Figure A2. (a) 2-dimensional representation of word embeddings A' colored by topic assignment. (b) 2-dimensional representation of word embeddings A' colored by original Wikipedia article assignment (words that occur in more than one article are excluded). (c) Colored heatmap of affinity matrix *R*.

Articles: "Soccer", "Tennis", "Rugby".

Topic 1 Topic 2 **Topic 3 Topic 4 Topic 5** Topic 6 #575 #402 #511 #453 #657 #621 net referee national tournaments rackets rules 12345678wingfield shot doubles balls penalty south serve may football singles made december hit size game kick cup events sports stance card europe tour must stroke listed strings lawn fifa prize backhand foul union money standard modern greek ball misconduct wales atp synthetic 9 leather server red africa men ta offence first 10 service width new grand NMF 0 rackets net referee national tournaments rules defensive retaken serbia bruno pressurisation collection 1 2 3 4 hourglass closer interference gold woodies become north eliminated unhappy somewhere dismissed equivalents center soares fully headquarters size originated 0 doubles shot balls wingfield penalty south 1 rotated prior asian combining express experimenting İlanelidan 2 3 4 execute vellow argentina becker ōz exclusively duration la bladder attended strive woodbridge antiphanes curve primary kong length #395 #413 #518 #616 #501 #776 1 penalty used net wimbledon world clubs ball episkyros 234567 forehand cup score rugby tournaments goal 11Se serve occurs schools large shot grass football team navratilova notable opponent roman fifa end forms players national also hit bc playing western lines occur international match sport , 8 9 twohanded ad goals greatest server europe doubles service island tournament time union 10 injury believed mav states scored war LDA 0 used net wimbledon world penalty clubs mistaken british 1 2 3 seconds result measure sees restrictions diagonal determined cancelled crossed papua although hollow exists combined requiring admittance 4 use perpendicular win wii teammate forces 0 episkyros forehand ball score rugby cup 1 2 multiple twohanded long roman penalty union deuce inline bar public grips bc 3 position facetiously island fifa fouled took 4 woodbridge allows believed published manufactured hour #1310 #371 #423 #293 #451 #371 players 123456789net tournaments stroke greatest balls player ball forehand singles ever rackets doubles female tennis shot stance size also serve tour wingfield square power play opponent slam backhand williams made football may prize torso navratilova leather game money team hit weight grip first service grand rotation said standard hitting events twohanded one serena width 10 line ranking used sports past SVD rugby players breaking greatest 0 tournaments stroke balls net 1 . lived pace masters rotates panels 2 3 one reach lowest achieve female sewn underhand entire running events face biggest 4 often tour adds potential leather air 0 player ball singles forehand ever rackets 1 autobiography utilize indian twohanded meanwhile keep 2 3 give hands doubles begins jack laminated pro rankings cońsistent converted pass backhand wood 4 gonzales either achieve touch strings

Table A5. For each evaluated matrix factorization method we display the top 10 words for each topic and the 5 most similar words based on cosine similarity for the 2 top words from each topic.

Appendix A.2. Additional Results on Wikipedia Data as Tensor Input

Wikipedia Articles "Soccer", "Bee", "Johnny Depp"—DEDICOM Automatic gradient method.



Figure A3. (a) 2-dimensional representation of word embeddings A' colored by topic assignment. (b) 2-dimensional representation of word embeddings A' colored by original Wikipedia article assignment (words that occur in more than one article are excluded).

Wikipedia Articles "Soccer", "Bee", "Johnny Depp"—DEDICOM Multiplicative Update Rules.



Figure A4. (a) 2-dimensional representation of word embeddings A' colored by topic assignment. (b) 2-dimensional representation of word embeddings A' colored by original Wikipedia article assignment (words that occur in more than one article are excluded).

	Topic 1 #226	Topic 2 #628	Topic 3 #1048	Topic 4 #571	Topic 5 #1267	Topic 6 #554
	cells	mysticetes	shark	bony	dolphin	whaling
1	(1.785)	(1.808)	(3.019)	(1.621)	(3.114)	(3.801)
0	brain	whales	sharks	blood	dolphins	iwc
2	(1.624)	(1.791)	(2.737)	(1.452)	(2.908)	(2.159)
2	light	feed	fins	fish	bottlenose	aboriginal
3	(1.561)	(1.427)	(1.442)	(1.438)	(1.629)	(2.098)
4	cone	baleen	killed	gills	meat	canada
4	(1.448)	(1.33)	(1.407)	(1.206)	(1.403)	(1.912)
5	allow	odontocetes	endangered	teeth	behavior	moratorium
5	(1.32)	(1.278)	(1.377)	(1.088)	(1.399)	(1.867)
6	greater	consist	hammerhead	body	captivity	industry
0	(1.292)	(1.162)	(1.269)	(1.043)	(1.298)	(1.855)
7	slightly	water	conservation	system	river	us
/	(1.269)	(1.096)	(1.227)	(1.027)	(1.281)	(1.838)
0	ear	krill	trade	skeleton	common	belugas
0	(1.219)	(1.05)	(1.226)	(1.008)	(1.275)	(1.585)
0	cornea	toothed	whitetip	called	selfawareness	whale
9	(1.158)	(1.003)	(1.203)	(0.99)	(1.248)	(1.542)
10	rod	sperm	finning	tissue	often	gb£
10	(1.128)	(0.991)	(1.184)	(0.875)	(1.218)	(1.528)

Wikipedia Articles "Dolphin", "Shark", "Whale"—DEDICOM Multiplicative Update Rules.

Table A6. Each column lists the top 10 representative words per dimension of the basis matrix *A*.

Table A7. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	cells	mysticetes	shark	bony	dolphin	whaling
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	sensitive	unborn	native	edges	hybrid	māori
T	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	cone	grind	tl	mirabile	hybridization	trips
2	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
3	rod	counterparts	predators—organisms	matches	yangtze	predominantly
5	(0.998)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
4	corneas	threechambered	cretaceous	turbulence	grampus	revenue
т	(0.998)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
0	brain	whales	sharks	blood	dolphins	iwc
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	receive	extended	reminiscent	hydrodynamic	superpod	distinction
1	(0.998)	(0.996)	(1.0)	(0.998)	(1.0)	(1.0)
r	equalizer	bryde	electrical	scattering	masturbation	billion
2	(0.998)	(0.996)	(1.0)	(0.998)	(1.0)	(1.0)
3	lobes	closes	induced	reminder	interaction	spain
5	(0.997)	(0.996)	(1.0)	(0.998)	(1.0)	(1.0)
4	clear	effects	coarsely	flows	stressful	competition
Ŧ	(0.997)	(0.996)	(1.0)	(0.998)	(1.0)	(1.0)



Figure A5. Colored heatmap of affinity tensor *R*.



Figure A6. (a) 2-dimensional representation of word embeddings *A* colored by topic assignment. (b) 2-dimensional representation of word embeddings *A* colored by original Wikipedia article assignment (words that occur in more than one article are excluded).



Wikipedia Articles "Soccer", "Tennis", "Rugby"—DEDICOM Multiplicative Update Rules.

Figure A7. Colored heatmap of affinity tensor *R*.

	Topic 1 #441	Topic 2 #861	Topic 3 #412	Topic 4 #482	Topic 5 #968	Topic 6 #57
1	rugby	titles	rackets	net	penalty	doubles
1	(2.55)	(1.236)	(2.176)	(2.767)	(1.721)	(2.335)
2	union	wta	wingfield	shot	football	singles
	(2.227)	(1.196)	(1.536)	(2.586)	(1.701)	(2.321)
3	wales	circuit	modern	serve	team	tournaments
5	(1.822)	(1.123)	(1.513)	(2.393)	(1.507)	(2.245)
4	georgia	futures	racket	hit	laws	tennis
4	(1.682)	(1.122)	(1.43)	(1.978)	(1.462)	(1.752)
F	fiji	earn	th	stance	referee	grand
5	(1.557)	(1.104)	(1.355)	(1.945)	(1.449)	(1.662)
6	samoa	offer	lawn	service	fifa	events
0	(1.474)	(1.096)	(1.316)	(1.83)	(1.439)	(1.648)
7	zealand	mixed	century	stroke	may	slam
/	(1.458)	(1.089)	(1.236)	(1.797)	(1.435)	(1.623)
8	new	draws	strings	server	goal	player
0	(1.414)	(1.085)	(1.179)	(1.761)	(1.353)	(1.344)
0	tonga	atp	yielded	backhand	competitions	professional
9	(1.374)	(1.072)	(1.121)	(1.692)	(1.345)	(1.328)
10	south	challenger	balls	forehand	associations	players
10	(1.369)	(1.07)	(1.101)	(1.554)	(1.288)	(1.316)

 Table A8. Each column lists the top 10 representative words per dimension of the basis matrix A.

Table A9. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	rugby	titles	rackets	net	penalty	doubles
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	ireland	hopman	proximal	hit	organisers	singles
T	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	union	dress	interlaced	formally	elapsed	tournaments
2	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.985)
3	backfired	tennischannel	harry	offensive	polite	grand
5	(1.0)	(0.998)	(1.0)	(1.0)	(1.0)	(0.975)
1	kilopascals	seoul	deserves	deeply	modest	slam
т	(1.0)	(0.998)	(1.0)	$(1.0)^{-1}$	(1.0)	(0.971)
0	union	wta	wingfield	shot	football	singles
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	rugby	helps	proximal	requires	circumference	doubles
1	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	ireland	hamilton	interlaced	backwards	touchline	tournaments
4	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.985)
3	backfired	weeks	harry	entail	sanctions	grand
5	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.975)
1	zealand	couple	deserves	torso	home	slam
+	(1.0)	(1.0)	(1.0)	(1.0)	(0.999)	(0.971)



Figure A8. (a) 2-dimensional representation of word embeddings *A* colored by topic assignment. (b) 2-dimensional representation of word embeddings *A* colored by original Wikipedia article assignment (words that occur in more than one article are excluded).

Wikipedia Articles "Soccer", "Tennis", "Rugby"-TNMF.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
	#275	#505	#607	#459	#816	#559
1	greatest	rackets	tournaments	net	football	penalty
	(39.36)	(33.707)	(29.126)	(29.789)	(27.534)	(27.793)
2	ever	modern	events	shot	rugby	referee
	(26.587)	(24.281)	(25.327)	(27.947)	(24.037)	(23.632)
3	female	balls	tour	serve	union	goal
	(25.52)	(22.016)	(23.488)	(25.722)	(21.397)	(23.072)
4	navratilova	wingfield	prize	hit	south	may
	(24.348)	(20.923)	(21.823)	(21.344)	(20.761)	(22.978)
5	best	tennis	atp	stance	national	team
	(24.114)	(19.863)	(21.124)	(20.75)	(19.586)	(21.258)
6	williams	strings	money	service	fifa	kick
	(22.207)	(18.602)	(20.667)	(19.7)	(19.331)	(21.052)
7	serena	racket	doubles	server	wales	foul
	(21.256)	(18.369)	(19.919)	(19.051)	(18.627)	(19.018)
8	said	made	ranking	stroke	league	listed
	(20.666)	(17.622)	(19.736)	(18.781)	(18.31)	(17.736)
9	martina	yielded	us	backhand	cup	free
	(20.153)	(17.284)	(19.431)	(17.809)	(17.015)	(17.702)
10	budge	th	masters	ball	association	goals
	(20.111)	(16.992)	(18.596)	(17.2)	(16.721)	(17.209)

Table A10. Each column lists the top 10 representative words per dimension of the basis matrix A'.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	greatest	rackets	tournaments	net	football	penalty
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	illustrated	garden	us	lob	midlothian	whole
1	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	johansson	construction	earned	receiving	alcock	corner
2	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
3	wilton	yielded	participating	rotates	capital	offender
5	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
4	jonathan	energy	receives	adds	representatives	stoke
4	(1.0)	$(1.0)^{-1}$	(1.0)	(1.0)	(1.0)	(1.0)
0	ever	modern	events	shot	rugby	referee
0	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	deserved	design	juniors	lobber	slang	dismissed
1	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)
2	stated	version	bowl	unable	colonists	showing
2	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)
З	female	shape	comprised	alter	sevenaside	stoppage
0	(1.0)	(0.998)	(1.0)	(1.0)	(1.0)	(1.0)
4	contemporaries	stitched	carlo	applying	seldom	layout
ч	(1.0)	(0.998)	(1.0)	(1.0)	(1.0)	(1.0)

Table A11. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.



Figure A9. (a) 2-dimensional representation of word embeddings *H* colored by topic assignment. (b) 2-dimensional representation of word embeddings *H* colored by original Wikipedia article assignment (words that occur in more than one article are excluded).

Wikipedia Articles "Dolphin", "Shark", "Whale"-TNMF.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
	#675	#996	#279	#491	#1190	#663
1	whaling	sharks	young	killed	dolphin	mysticetes
	(34.584)	(30.418)	(33.823)	(23.214)	(35.52)	(24.404)
2	whale	fish	born	shark	dolphins	flippers
	(25.891)	(23.648)	(27.62)	(22.6)	(31.881)	(22.059)
3	whales (21.653)	bony (19.689)	oviduct (23.706)	states (21.24)	bottlenose (18.198)	odontocetes (21.621)
4	belugas	prey	viviparity	endangered	behavior	water
	(20.933)	(18.785)	(23.694)	(20.976)	(18.003)	(21.087)
5	aboriginal	teeth	embryos	conservation	selfawareness	tail
	(19.44)	(18.242)	(22.966)	(20.398)	(16.48)	(18.268)
6	iwc	blood	continue	fins	meat	mya
	(19.226)	(16.521)	(21.752)	(18.641)	(16.02)	(17.79)
7	canada	gills	calves	new	often	baleen
	(18.691)	(13.34)	(21.25)	(18.445)	(15.687)	(17.189)
8	arctic	tissue	blubber	international	captivity	limbs
	(17.406)	(12.927)	(21.094)	(18.4)	(15.452)	(16.56)
9	industry	body	egg	drum	river	allow
	(16.837)	(12.691)	(20.735)	(17.587)	(14.68)	(16.552)
10	right	skeleton	fluids	finning	common	toothed
	(16.766)	(12.52)	(20.662)	(17.321)	(14.389)	(16.489)

 Table A12. Each column lists the top 10 representative words per dimension of the basis matrix *H*.

Table A13. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	whaling	sharks	young	killed	dolphin	mysticetes
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	antarctica	loan	getting	alzheimer	behaviors	digits
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	spain (1.0)	leopard (1.0)	insulation (1.0)	queensland (1.0)	familiar (1.0)	streamlined (1.0)
3	caro (1.0)	dogfish (1.0)	harsh (1.0)	als (1.0)	pantropical (1.0)	archaeocete (1.0)
4	excluded (1.0)	lifespans (0.999)	primary (1.0)	control (1.0)	test (1.0)	defines (0.999)
0	whale	fish	born	shark	dolphins	flippers
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	reason	like	getting	figure	levels	expel
	(1.0)	(0.997)	(1.0)	(1.0)	(1.0)	(1.0)
2	respected (1.0)	lifetime (0.992)	young (1.0)	sources (1.0)	moderate (0.999)	compress (1.0)
3	divinity	content	leaner	video	injuries	protocetus
	(0.999)	(0.992)	(1.0)	(0.998)	(0.999)	(1.0)
4	taken	hazardous	insulation	dògfishes	seems	nostrils
	(0.998)	(0.992)	(1.0)	(0.997)	(0.998)	(1.0)



Figure A10. (a) 2-dimensional representation of word embeddings *H* colored by topic assignment. (b) 2-dimensional representation of word embeddings *H* colored by original Wikipedia article assignment (words that occur in more than one article are excluded).

Wikipedia Articles "Soccer", "Bee", "Johnny Depp"-TNMF.

Table A14. Each column lists the top 10	0 representative words pe	er dimension of the basis m	atrix H .
---	---------------------------	-----------------------------	------------------

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
	#793	#554	#736	#601	#740	#616
1	film	ball	honey	football	heard	species
	(37.29)	(27.588)	(29.778)	(32.591)	(37.167)	(32.973)
2	starred	may	insects	fifa	depp	eusocial
	(23.821)	(25.768)	(27.784)	(25.493)	(30.275)	(25.001)
3	role	penalty	bees	world	court	females
	(23.006)	(25.04)	(27.679)	(25.414)	(20.771)	(24.24)
4	series	players	bee	cup	divorce	solitary
	(19.563)	(24.063)	(26.936)	(24.925)	(17.289)	(21.173)
5	burton (18.694)	referee (23.649)	food (23.44)	(22.331)	sued (16.105)	nest (20.198)
6	played	team	flowers	national	stated	males
	(17.583)	(22.9)	(22.374)	(20.958)	(15.984)	(18.3)
7	character	goal	pollination	women	alcohol	workers
	(16.646)	(22.859)	(18.09)	(20.668)	(15.238)	(17.16)
8	success (16.41)	player (22.054)	larvae (17.73)	international (20.16)	stating (15.199)	(16.886)
9	films	play	pollen	tournament	paradis	colonies
	(15.74)	(21.774)	(17.666)	(18.26)	(14.98)	(16.528)
10	box	game	predators	uefa	alleged	queens
	(15.024)	(20.471)	(17.634)	(18.029)	(14.971)	(16.427)

0

-5

0

(a)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	film	ball	honey	football	heard	species
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	avril	officials	triangulum	entered	obtained	progressive
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	office	invoke	consumption	most	countersued	halictidae
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
3	landau	heading	copper	excess	depths	temperate
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
4	chamberlain	twohalves	might	uk	mismanagement	spring
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
0	starred	may	insects	fifa	depp	eusocial
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	raimi	noninternational	blooms	oceania	city	unfertilized
	(1.0)	(0.992)	(1.0)	(1.0)	(1.0)	(1.0)
2	candidate	red	eats	sudamericana	tribute	females
	(1.0)	(0.991)	(1.0)	(1.0)	(1.0)	(1.0)
3	hardwicke	required	catching	widened	mick	paper
	(1.0)	(0.991)	(1.0)	(1.0)	(1.0)	(1.0)
4	peter	yd	disease	oversee	elvis	hibernate
	(1.0)	(0.989)	(1.0)	(1.0)	(1.0)	(1.0)

Table A15. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.

Appendix A.3. Additional Results on Amazon Review Data as Tensor Input



0

-5

0

(b)

10

Amazon Reviews—DEDICOM Mulitplicative Update Rules.

Figure A11. (a) 2-dimensional representation of word embeddings *H* colored by topic assignment. (b) 2-dimensional representation of word embeddings *H* colored by original Wikipedia article assignment (words that occur in more than one article are excluded).

10



Figure A12. (a) 2-dimensional representation of word embeddings *A* colored by topic assignment. (b) 2-dimensional representation of word embeddings *A* colored by original review article.



Figure A13. (a) 2-dimensional representation of word embeddings *H* colored by topic assignment. (b) 2-dimensional representation of word embeddings *H* colored by original review article.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	anna	shen	legendary	lasseter	disc	screams	code	mike	woody	ро
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	christoph (1.0)	canons (1.0)	sacred (0.999)	andrew (1.0)	thxcertified (1.0)	harvested (1.0)	confirm (1.0)	bogg (1.0)	rips (1.0)	panda (0.985)
2	readiness	yeoh	fulfill	stanton	presentation	speciallytrained	discount	chased	spy	fu
	(1.0)	(1.0)	(0.999)	(1.0)	(0.999)	(1.0)	(1.0)	(0.996)	(1.0)	(0.983)
3	carrots (1.0)	wolf (1.0)	roster (0.999)	eggleston (1.0)	upgrade (0.999)	screamprocessing (1.0)	browser (1.0)	flair (0.996)	josie (1.0)	black (0.983)
4	poverty	weapon	megafan	uncredited	featurettes	corporation	popup	slot	supurb	kung
	(1.0)	(1.0)	(0.999)	(1.0)	(0.998)	(1.0)	(1.0)	(0.995)	(1.0)	(0.981)
0	elsa	peacock	valley	director	birds	energy	email	crystal	buzz	master
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	shipwreck	shen	praying	producer	pressed	powered	confirm	oz	hist	shifu
	(1.0)	(1.0)	(0.999)	(0.998)	(1.0)	(0.998)	(1.0)	(1.0)	(1.0)	(0.999)
2	marriage	mcbride	kim	teaser	gadget	frightened	code	mae	wayne	warrior
	(1.0)	(1.0)	(0.997)	(0.997)	(1.0)	(0.997)	(1.0)	(1.0)	(1.0)	(0.999)
3	idena	yeoh	chorgum	globes	classically	screams	fwiw	celia	reunited	dragon
	(1.0)	(1.0)	(0.997)	(0.997)	(1.0)	(0.996)	(1.0)	(1.0)	(1.0)	(0.998)
4	prodding	michelle	preying	rousing	starz	scarry	android	cristal	hockey	martial
	(1.0)	(1.0)	(0.997)	(0.997)	(1.0)	(0.996)	(1.0)	(1.0)	(1.0)	(0.994)

Table A16. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.

Table A17. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	anna	shen	legendary	lasseter	disc	screams	code	mike	woody	po
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	christóph	canons	sacréd	andrew	thxcertified	harvested	confirm	bogg	rips	panda
	(1.0)	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.985)
2	readiness	yeoh	fulfill	stanton	presentation	speciallytrained	discount	chased	spy	fu
	(1.0)	(1.0)	(0.999)	(1.0)	(0.999)	(1.0)	(1.0)	(0.996)	(1.0)	(0.983)
3	carrots (1.0)	wolf (1.0)	roster (0.999)	eggleston (1.0)	upgrade (0.999)	screamprocessing (1.0)	browser (1.0)	flair (0.996)	josie (1.0)	black (0.983)
4	poverty	weapon	megafan	uncredited	featurettes	corporation	popup	slot	supurb	kung
	(1.0)	(1.0)	(0.999)	(1.0)	(0.998)	(1.0)	(1.0)	(0.995)	(1.0)	(0.981)
0	elsa	peacock	valley	director	birds	energy	email	crystal	buzz	master
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	shipwreck	shen	praying	producer	pressed	powered	confirm	oz	hist	shifu
	(1.0)	(1.0)	(0.999)	(0.998)	(1.0)	(0.998)	(1.0)	(1.0)	(1.0)	(0.999)
2	marriage	mcbride	kim	teaser	gadget	frightened	code	mae	wayne	warrior
	(1.0)	(1.0)	(0.997)	(0.997)	(1.0)	(0.997)	(1.0)	(1.0)	(1.0)	(0.999)
3	idena	yeoh	chorgum	globes	classically	screams	fwiw	celia	reunited	dragon
	(1.0)	(1.0)	(0.997)	(0.997)	(1.0)	(0.996)	(1.0)	(1.0)	(1.0)	(0.998)
4	prodding (1.0)	michelle (1.0)	(0.997)	rousing (0.997)	starz (1.0)	scarry (0.996)	android (1.0)	cristal (1.0)	hockey (1.0)	martial (0.994)

Amazon Reviews-TNMF.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
	#590	#1052	#456	#350	#4069	#733	#1140	#582	#423	#605
1	anna	woody	director	allen	widescreen	code	master	mike	film	screams
	(109.81)	(134.366)	(100.622)	(83.875)	(34.484)	(88.94)	(89.686)	(88.628)	(58.514)	(87.782)
2	elsa	buzz	lasseter	hanks	outtakes	email	ро	crystal	animation	energy
	(106.148)	(120.93)	(93.134)	(77.313)	(30.724)	(73.645)	(85.79)	(82.472)	(53.628)	(78.113)
3	olaf	andy	andrew	tim	disc	promo	shifu	billy	characters	world
	(59.353)	(105.728)	(81.452)	(75.511)	(30.688)	(67.483)	(82.465)	(79.244)	(46.861)	(73.888)
4	trolls	toys	stanton	rickles	extras	amazon	warrior	goodman	films	monstropolis
	(58.811)	(98.523)	(80.132)	(74.217)	(30.09)	(64.978)	(75.609)	(76.831)	(44.937)	(73.484)
5	kristoff	lightyear	john	tom	versions	promotion	dragon	sully	pixar	monsters
	(56.309)	(68.336)	(73.004)	(72.401)	(27.894)	(58.631)	(74.235)	(75.612)	(44.873)	(71.721)
6	hans	sid	pete	jim	included	free	tai	wazowski	even	city
	(55.628)	(52.34)	(70.556)	(69.776)	(27.455)	(58.207)	(71.721)	(71.588)	(44.176)	(71.352)
7	frozen	cowboy	docter	varney	material	promotional	lung	randall	animated	power
	(54.257)	(48.588)	(64.734)	(66.053)	(26.887)	(57.738)	(70.786)	(69.695)	(43.492)	(70.642)
8	queen	space	ralph	slinky	edition	click	furious	sulley	also	monster
	(53.956)	(47.88)	(54.884)	(62.326)	(26.546)	(55.373)	(63.232)	(69.604)	(43.484)	(70.197)
9	sister	room	joe	potato	contains	download	oogway	james	dvd	closet
	(52.749)	(42.655)	(53.7)	(62.237)	(25.386)	(50.788)	(60.879)	(68.574)	(42.736)	(61.451)
10	ice	toy	ranft	mr	extra	purchase	five	buscemi	well	scare
	(49.71)	(42.042)	(53.41)	(61.801)	(25.144)	(50.327)	(59.259)	(66.028)	(40.124)	(61.243)

Table A18. Each column lists the top 10 representative words per dimension of the basis matrix *H*.

Table A19. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	anna	woody	director	allen	widescreen	code	master	mike	film	screams
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	marriage	acciently	producer	trustworthy	benefactors	card	furious	longtime	films	electrical
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.995)	(1.0)	(1.0)
2	trolls (1.0)	limp (1.0)	jackson (1.0)	arguments (1.0)	pioneers (1.0)	confirmation (1.0)	dragon (1.0)	cyclops (0.995)	first (0.994)	screamprocessing (1.0)
3	flees	jealousey	rabson	hanks	keepcase	assuming	shifu	slot	animated	chlid
	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.995)	(0.993)	(1.0)
4	christian	swells	composer	knowitall	redone	android	warrior	humanlike	animation	shortage
	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.994)	(0.989)	(1.0)
0	elsa	buzz	lasseter	hanks	outtakes	email	ро	crystal	animation	energy
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	marriage	recive	nathan	trustworthy	storyboarding	promo	martial	billy	film	supply
	(1.0)	(0.999)	(1.0)	(1.0)	(0.993)	(1.0)	(0.998)	(1.0)	(0.989)	(1.0)
2	heals	zorg	officer	allen	informative	avail	fight	talkative	story	powered
	(1.0)	(0.999)	(1.0)	(1.0)	(0.991)	(1.0)	(0.997)	(0.999)	(0.987)	(1.0)
3	marrying	limp	cunningham	tom	contents	flixster	arts	competitor	scenes	collect
	(1.0)	(0.997)	(1.0)	(1.0)	(0.99)	(1.0)	(0.996)	(0.999)	(0.987)	(1.0)
4	feminist	acciently	derryberry	arguments	logo	contirming	adopted	devilishly	also	screams
	(1.0)	(0.997)	(1.0)	(1.0)	(0.99)	(1.0)	(0.995)	(0.999)	(0.984)	(0.999)

Appendix A.4. Additional Results on the New York Times News Article Data as Tensor Input New York Times News Articles – DEDICOM Mulitplicative Update Rules.

Table 1 20	Eastha most	cionificant two	words nor ton	is the form	nonwork noighborg	hazad on as	ain a ainailanita a	no listod
lable A20.	ror the most	significant two	words per top	ic, the four	nearest neighbors	based off co	sille silliality a	re insteu.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	suleimani	loans	masks	floyd	contributed	confederate	ukraine	storm	restaurants	weinstein
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	qassim	spend	sanitizer	brutality	alan	statue	lutsenko	storms	salons	raped
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	iran (1.0)	smallbusiness (1.0)	wipes (1.0)	police (1.0)	edmondson (1.0)	monuments (1.0)	ukrainians (1.0)	isaias (1.0)	cafes (1.0)	predatory (1.0)
3	iranian	rent	cloth	systemic	mervosh	statues	yovanovitch	landfall	pubs	mann
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
4	militias	incentives	homemade	knee	emily	honoring	burisma	forecasters	nightclubs	sciorra
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
0	iran	university	protective	minneapolis	reporting	statue	sondland	hurricane	bars	sexual
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	qassim (1.0)	jerome (0.999)	gowns (1.0)	breonna (1.0)	rabin (1.0)	monuments (1.0)	zelensky (1.0)	bahamas (1.0)	dining (1.0)	rape (1.0)
2	suleimani	oxford	ventilators	kueng	contributed	statues	volker	hurricanes	theaters	metoo
	(1.0)	(0.998)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
3	iranian	columbia	respirators	floyd	keith	confederate	giuliani	forecasters	venues	sexually
	(1.0)	(0.998)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
4	militias	economics	supplies	police	chokshi	honoring	quid	landfall	malls	mann
	(1.0)	(0.998)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)

New York Times News Articles – TNMF.

Table A21. Each column lists the top 10 representative words per dimension of the basis matrix *A*.

	Topic 1 #977	Topic 2 #360	Topic 3 #420	Topic 4 #4192	Topic 5 #489	Topic 6 #405	Topic 7 #1135	Topic 8 #748	Topic 9 #108	Topic 10 #1166
1	floyd (82 861)	contributed	iran (80,007)	masks	ship (87 178)	syria (94,686)	senator	restaurants	bloom (110 77)	ukraine
2	police (64 649)	reporting (84,889)	suleimani (78 78)	(40.403) patients (34.77)	(07.170) crew (70.694)	(94.000) syrian (82.565)	storm (43,439)	bars (64,889)	julie (103 282)	sondland
3	protesters (63 588)	michael (76 156)	(70.70) iranian (72.581)	ventilators	aboard (67 464)	kurdish (82.013)	hurricane (42 213)	reopen (57 541)	edited (100.159)	testimony (49 982)
4	minneapolis (63.216)	(70.130) katie (63.146)	(72.301) iraq (63.27)	protective (33 719)	passengers	(02.013) turkey (80.374)	iowa (41.985)	stores (55.435)	los (95 747)	testified
5	protests (61 585)	emily (60,696)	(00.27) gen (50.966)	(35.717) loans (28.178)	(03.695) cruise (63.535)	(00.374) turkish (75.912)	republican	gyms (50.487)	graduated	zelensky
6	george (53.378)	alan (59.499)	strike	supplies	princess (45 714)	kurds (63 639)	(40.775) gov (37.689)	theaters	angeles	ambassador
7	brutality (44.051)	nicholas	iraqi (46.027)	gloves	flight	fighters	buttigieg	closed	berkeley	weinstein
8	(44.051) officers (42.581)	cochrane	(46.027) qassim	equipment	(43.375) nasa (42.200)	forces	democrat	(40.344) indoor	(83.655) grew	ukrainian
9	(43.381) racism	(52.045) ben	(45.861) maj	(26.252) respiratory	(43.506) navy (40.20()	(57.659) troops	(37.087) representative	(44.541) salons	(84.145) today	(43.716) giuliani
10	(43.437) demonstrations (42.547)	(41.414) maggie (41.328)	(44.921) baghdad (44.867)	(25.447) testing (24.179)	(40.396) astronauts (37.073)	(54.045) isis (53.743)	(35.807) bernie (35.228)	(41.99) shops (40.311)	(41.818) california (38.807)	(42.674) sexual (39.966)
	. ,	. ,	. /	. /	. /	. /	. ,	. /	. /	. /

Table A22. For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.

-	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	floyd	contributed	iran	masks	ship	syria	senator	restaurants	bloom	ukraine
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	demonstrations	shear	suleimani	providers	aboard	isis	wyden	shops	graduated	volker
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
2	systemic	annie	retaliation	distressed	capsule	ceasefire	iowa	takeout	edited	inquiry
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)
3	protests	mazzei	qassim	tobacco	diamond	fighters	steyer	nightclubs	berkeley	transcript
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)
4	defund	kitty	revenge	selfemployed	dragon	syrian	klobuchar	pubs	grew	investigations
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)
0	police	reporting	suleimani	patients	crew	syrian	storm	bars	julie	sondland
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
1	systemic	luis	strike	treating	aboard	alassad	carolina	reopen	garcetti	testifying
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.993)	(1.0)
2	peaceful	beachy	maj	infection	capsule	recep	rubio	nonessential	graduated	mick
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.988)	(1.0)
3	peacefully	kaplan	iran	develop	princess	erdogan	hampshire	nail	edited	quid
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.988)	(1.0)
4	knee	glueck	retaliation	repay	cruise	kurds	landfall	takeout	berkeley	impeachment
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.988)	(1.0)



Figure A14. 2-dimensional representation of word embeddings A colored by topic assignment.



Figure A15. 2-dimensional representation of word embeddings A' colored by topic assignment).

Appendix B. Matrix Derivatives

In this section we derive the derivatives in (20) and (24) analytically. We write the loss in trace form by

$$\mathcal{L}(S, A, R) = \left\| S - ARA^T \right\|_F^2$$

= tr $\left[\left(S - ARA^T \right)^T \left(S - ARA^T \right) \right]$
= tr $\left[Q^T Q \right]$

Then

$$d\mathcal{L} = d \operatorname{tr} \left[Q^{T} Q \right]$$

= tr $\left[d \left(Q^{T} Q \right) \right]$
= tr $\left[(d Q)^{T} Q + Q^{T} d Q \right]$
= tr $\left[Q^{T} d Q + Q^{T} d Q \right]$
= tr $\left[\left(Q^{T} + Q^{T} \right) d Q \right]$
= 2 tr $\left[Q^{T} d Q \right]$
= 2 tr $\left[Q^{T} d \left(S - A'RA^{T} \right) \right]$
= 2 tr $\left[Q^{T} d S \right] - 2 \operatorname{tr} \left[Q^{T} d \left(ARA^{T} \right) \right]$
= -2 tr $\left[Q^{T} d \left(ARA^{T} \right) \right]$

Differential in terms of d *R*:

$$d\mathcal{L} = -2 \operatorname{tr} \left[Q^T d \left(A R A^T \right) \right]$$
$$= -2 \operatorname{tr} \left[Q^T A d R A^T \right]$$
$$= -2 \operatorname{tr} \left[A^T Q^T A d R \right]$$
$$\frac{\partial \mathcal{L}}{\partial R} = -2 \left(A^T Q A \right)$$
$$= -2 \left(A^T \left(S - A R A^T \right) A \right)$$
$$= -2 \left(A^T S A - A^T A R A^T A \right)$$

Differential in terms of d *A*:

$$d\mathcal{L} = -2 \operatorname{tr} \left[Q^T d \left(A R A^T \right) \right]$$

= $-2 \operatorname{tr} \left[Q^T d A R A^T + Q^T A R (d A)^T \right]$
= $-2 \operatorname{tr} \left[R A^T Q^T d A + R^T A^T Q d A \right]$
= $-2 \operatorname{tr} \left[\left(R A^T Q^T + R^T A^T Q \right) d A \right]$

$$egin{aligned} &rac{\partial \mathcal{L}}{\partial A} = -2 \Big(QAR^T + Q^TAR \Big) \ &= -2 \Big(\Big(S - ARA^T \Big) AR^T + \Big(S - ARA^T \Big)^TAR \Big) \ &= -2 \Big(SAR^T - ARA^TAR^T + S^TAR - AR^TA^TAR \Big) \ &= -2 \Big(SAR^T + S^TAR - A \Big(RA^TAR^T + R^TA^TAR \Big) \Big) \end{aligned}$$

References

- Levy, O.; Goldberg, Y. Neural Word Embedding as Implicit Matrix Factorization. In Advances in Neural Information Processing Systems; NIPS'14; MIT Press: Cambridge, MA, USA, 2014; Volume 2, pp. 2177–2185.
- Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
- 3. Hillebrand, L.P.; Biesner, D.; Bauckhage, C.; Sifa, R. Interpretable Topic Extraction and Word Embedding Learning Using Row-Stochastic DEDICOM. In *Machine Learning and Knowledge Extraction—4th International Cross-Domain Conference, CD-MAKE*; Lecture Notes in Computer Science; Springer: Dublin, Ireland, 2020; Volume 12279, pp. 401–422.
- 4. Symeonidis, P.; Zioupos, A. *Matrix and Tensor Factorization Techniques for Recommender Systems*; Springer International Publishing: New York, NY, USA, 2016. [CrossRef]
- Harshman, R.; Green, P.; Wind, Y.; Lundy, M. A Model for the Analysis of Asymmetric Data in Marketing Research. *Market. Sci.* 1982, 1, 205–242. [CrossRef]
- Andrzej, A.H.; Cichocki, A.; Dinh, T.V. Nonnegative DEDICOM Based on Tensor Decompositions for Social Networks Exploration. Aust. J. Intell. Inf. Process. Syst. 2010, 12, 10–15.
- 7. Bader, B.W.; Harshman, R.A.; Kolda, T.G. *Pattern Analysis of Directed Graphs Using DEDICOM: An Application to Enron Email;* Office of Scientific & Technical Information Technical Reports; Sandia National Laboratories: Albuquerque, NM, USA, 2006.
- 8. Sifa, R.; Ojeda, C.; Cvejoski, K.; Bauckhage, C. Interpretable Matrix Factorization with Stochasticity Constrained Nonnegative DEDICOM. In Proceedings of the KDML-LWDA, Rostock, Germany, 11–13 September 2017.
- Sifa, R.; Ojeda, C.; Bauckage, C. User Churn Migration Analysis with DEDICOM. In Proceedings of the 9th ACM Conference on Recommender Systems; RecSys '15; Association for Computing Machinery: New York, NY, USA, 2015; pp. 321–324.

- Chew, P.; Bader, B.; Rozovskaya, A. Using DEDICOM for Completely Unsupervised Part-of-Speech Tagging. In Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics, Boulder, CO, USA, 5 June 2009; Association for Computational Linguistics: Boulder, CO, USA, 2009; pp. 54–62.
- Nickel, M.; Tresp, V.; Kriegel, H.P. A Three-Way Model for Collective Learning on Multi-Relational Data. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2009; Association for Computing Machinery: Bellevue, WA, USA, 2011; pp. 809–816.
- 12. Sifa, R.; Yawar, R.; Ramamurthy, R.; Bauckhage, C.; Kersting, K. Matrix- and Tensor Factorization for Game Content Recommendation. *KI-Künstl. Intell.* 2019, 34, 57–67. [CrossRef]
- 13. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *Adv. Neural Inf. Proc. Syst.* 2013, *26*, 3111–3119.
- Lee, D.D.; Seung, H.S. Algorithms for Non-Negative Matrix Factorization. In Proceedings of the 13th International Conference on Neural Information Processing Systems; NIPS'00; MIT Press: Cambridge, MA, USA, 2000; pp. 535–541.
- Furnas, G.W.; Deerwester, S.; Dumais, S.T.; Landauer, T.K.; Harshman, R.A.; Streeter, L.A.; Lochbaum, K.E. Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure. In *ACM SIGIR Forum*; ACM: New York, NY, USA, 1988.
- Wang, Y.; Zhu, L. Research and implementation of SVD in machine learning. In Proceedings of the 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, China, 24–26 May 2017; pp. 471–475.
- 17. Jolliffe, I. Principal Component Analysis; John Wiley and Sons Ltd.: Hoboken, NJ, USA, 2005.
- 18. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. J. Mach. Learn. Res. 2003, 3, 993–1022.
- 19. Lebret, R.; Collobert, R. Word Embeddings through Hellinger PCA. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 482–490.
- Nguyen, D.Q.; Billingsley, R.; Du, L.; Johnson, M. Improving Topic Models with Latent Feature Word Representations. *Trans.* Assoc. Comput. Linguist. 2015, 3, 299–313. [CrossRef]
- 21. Frank, M.; Wolfe, P. An algorithm for quadratic programming. Naval Res. Logist. Q. 1956, 3, 95–110. [CrossRef]
- Ni, J.; Li, J.; McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 188–197.
- 23. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv 2014, arXiv:1412.6980.
- 24. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* 2018, arXiv:1802.03426.