



# Interpretable Topic Extraction and Word Embedding Learning Using Row-Stochastic DEDICOM

Lars Hillebrand, David Biesner, Christian Bauckhage, Rafet Sifa

## ► To cite this version:

Lars Hillebrand, David Biesner, Christian Bauckhage, Rafet Sifa. Interpretable Topic Extraction and Word Embedding Learning Using Row-Stochastic DEDICOM. 4th International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE), Aug 2020, Dublin, Ireland. pp.401-422, 10.1007/978-3-030-57321-8\_22 . hal-03414746

**HAL Id: hal-03414746**

**<https://inria.hal.science/hal-03414746>**

Submitted on 4 Nov 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Interpretable Topic Extraction and Word Embedding Learning using row-stochastic DEDICOM

Lars Hillebrand<sup>\*1,2</sup>, David Biesner<sup>\*1,2</sup>, Christian Bauckhage<sup>1,2</sup>, and Rafet Sifa<sup>1</sup>

<sup>1</sup> Fraunhofer IAIS

<sup>2</sup> University of Bonn

**Abstract.** The DEDICOM algorithm provides a uniquely interpretable matrix factorization method for symmetric and asymmetric square matrices. We employ a new row-stochastic variation of DEDICOM on the pointwise mutual information matrices of text corpora to identify latent topic clusters within the vocabulary and simultaneously learn interpretable word embeddings. We introduce a method to efficiently train a constrained DEDICOM algorithm and a qualitative evaluation of its topic modeling and word embedding performance.

**Keywords:** Word Embeddings · Topic Analysis · Matrix Factorization · Natural Language Processing.

## 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 low-dimensional representations of the vocabulary, so-called word embeddings [11, 15], 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.

---

<sup>\*</sup> First authors, equal contribution.

Correspondence to [lars.patrick.hillebrand@iais.fraunhofer.de](mailto:lars.patrick.hillebrand@iais.fraunhofer.de)

Another 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 analyse 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 then present results of various experiments on semi-artificial text documents (combinations of Wikipedia 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.

## 2 Related Work

The DEDICOM algorithm has a long history of providing interpretable matrix factorization, mostly for rather low-dimensional tasks. First described in [6], it since has been applied to analysis of social networks [1], email correspondence [2] and video game player behaviour [16, 17]. DEDICOM also has successfully been employed to NLP tasks such as part of speech tagging [4], 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 [2, 17], the authors constrain the affinity matrix  $\mathbf{R}$  to be non-negative, which aids interpretability and improves convergence behaviour 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,  $\mathbf{A}$ , include non-negativity as well (see [2]) or column-orthogonality as in [17].

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

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 [15] 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 [11] 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.

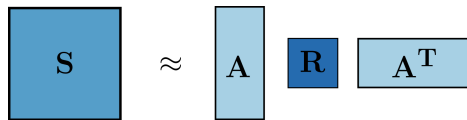
Both models achieve impressive results on word embedding related tasks like word analogy, however the large dimensionality of the word embeddings makes interpreting the latent factors of the embeddings impossible.

On the topic modeling side, matrix factorization methods are routinely applied as well. Popular algorithms like non-negative matrix factorization (NMF) [10], singular value decomposition (SVD) [5, 18] and principal component analysis (PCA) [7] compete against the probabilistic latent dirichlet allocation (LDA) [3] to cluster the vocabulary of a word co-occurrence or document-term matrix into latent topics.<sup>3</sup> 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 appendix.

### 3 The row-stochastic DEDICOM Model

In this section we provide a detailed theoretical view at the proposed row-stochastic DEDICOM algorithm for factorizing word co-occurrence based positive pointwise mutual information matrices.



$$\mathbf{S} \approx \mathbf{A} \mathbf{R} \mathbf{A}^T$$

Figure 1: The DEDICOM algorithm factorizes a square matrix  $\mathbf{S}$  into a loading matrix  $\mathbf{A}$  and an affinity matrix  $\mathbf{R}$ .

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

$$\mathbf{W}_{ij} = \# \text{word } i \text{ appears in context of word } j. \quad (1)$$

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,  $\mathbf{W}_{ij} = \mathbf{W}_{ji}$ .

We then transform the co-occurrence matrix into the pointwise mutual information matrix (PMI), which normalizes the counts in order to extract meaningful

<sup>3</sup> More recent expansions of these methods can be found in [9, 14].

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

$$\mathbf{PMI}_{ij} := \log \mathbf{W}_{ij} + \log N - \log N_i - \log N_j \quad (2)$$

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

Since the co-occurrence matrix  $\mathbf{W}$  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 3.2.

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

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

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

$$\mathbf{S} \approx \mathbf{A} \mathbf{R} \mathbf{A}^T, \quad \text{with} \quad \mathbf{S}_{ij} \approx \sum_{b=1}^k \sum_{c=1}^k \mathbf{A}_{ib} \mathbf{R}_{bc} \mathbf{A}_{jc}, \quad (4)$$

where  $\mathbf{A} \in \mathbb{R}^{n \times k}$ ,  $\mathbf{R} \in \mathbb{R}^{k \times k}$ ,  $\mathbf{A}^T$  denotes the transpose of  $\mathbf{A}$  and  $k \ll n$ . Literature often refers to  $\mathbf{A}$  as the loading matrix and  $\mathbf{R}$  as the affinity matrix.  $\mathbf{A}$  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  $\mathbf{R}$  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 behaviour by applying a column-wise z-normalization on  $\mathbf{A}$ , 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  $\mathbf{A}_{ib}$  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  $\mathbf{A}$  matrix by defining  $\mathbf{A}' \in \mathbb{R}^{n \times k}$  as

$$\begin{aligned} \mathbf{A}'_{ib} &:= \frac{\exp(\bar{\mathbf{A}}_{ib})}{\sum_{b'=1}^k \exp(\bar{\mathbf{A}}_{ib'})}, \quad \bar{\mathbf{A}}_{ib} := \frac{\mathbf{A}_{ib} - \mu_b}{\sigma_b}, \\ \mu_b &:= \frac{1}{n} \sum_{i=1}^n \mathbf{A}_{ib}, \quad \sigma_b := \sqrt{\frac{1}{n} \sum_{i=1}^n (\mathbf{A}_{ib} - \mu_b)^2} \end{aligned} \quad (5)$$

and optimizing  $\mathbf{A}$  for the objective

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

Note that after applying the row-wise softmax operation all entries of  $\mathbf{A}'$  are non-negative.

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

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

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

with

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

and  $\mathbf{A}'$  defined in (5).

To optimize the loss function we train both matrices using alternating gradient descent similar to [17]. Within each optimization step we apply

$$\mathbf{A} \leftarrow \mathbf{A} - f_\theta(\nabla_{\mathbf{A}}, \eta_{\mathbf{A}}), \quad \text{where} \quad \nabla_{\mathbf{A}} = \frac{\partial \mathcal{L}(\mathbf{S}, \mathbf{A}, \mathbf{R})}{\partial \mathbf{A}} \quad (10)$$

$$\mathbf{R} \leftarrow \mathbf{R} - f_\theta(\nabla_{\mathbf{R}}, \eta_{\mathbf{R}}), \quad \text{where} \quad \nabla_{\mathbf{R}} = \frac{\partial \mathcal{L}(\mathbf{S}, \mathbf{A}, \mathbf{R})}{\partial \mathbf{R}} \quad (11)$$

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

### 3.1 On Symmetry

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

$$(\mathbf{A}\mathbf{R}\mathbf{A}^T)_{ij} = \sum_{b=1}^k \sum_{c=1}^k \mathbf{A}_{ib} \mathbf{R}_{bc} \mathbf{A}_{jc} = \sum_{b=1}^k \sum_{c=1}^k \mathbf{A}_{ib} \mathbf{R}_{cb} \mathbf{A}_{jc} \quad (12)$$

$$= \sum_{c=1}^k \sum_{b=1}^k \mathbf{A}_{jc} \mathbf{R}_{cb} \mathbf{A}_{ib} = (\mathbf{A}\mathbf{R}\mathbf{A}^T)_{ji} \quad (13)$$

iff  $\mathbf{R}_{cb} = \mathbf{R}_{bc}$  for all  $b, c$ . We therefore expect a symmetric matrix  $\mathbf{S}$  to be decomposed into  $\mathbf{A}\mathbf{R}\mathbf{A}^T$  with a symmetric  $\mathbf{R}$ , which is confirmed by our experiments. Factorizing a non-symmetric matrix leads to a non-symmetric  $\mathbf{R}$ , the asymmetric relation between items leads to asymmetric relations between the latent factors.

### 3.2 On Interpretability

We have

$$\mathbf{S}_{ij} \approx \sum_{b=1}^k \sum_{c=1}^k \mathbf{A}_{ib} \mathbf{R}_{bc} \mathbf{A}_{jc}, \quad (14)$$

i.e. we can estimate the probability of co-occurrence of two words  $w_i$  and  $w_j$  from the word embeddings  $\mathbf{A}_i$  and  $\mathbf{A}_j$  and the matrix  $\mathbf{R}$ , where  $\mathbf{A}_i$  denotes the  $i$ -th row of  $\mathbf{A}$ .

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  $\mathbf{A}_i$  and  $\mathbf{A}_j$ , and sum up each component from  $\mathbf{A}_i$  with each component  $\mathbf{A}_j$  with respect to the relationship weights given in  $\mathbf{R}$ .

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  $\mathbf{R}$  matrix. Likewise a negative entry  $\mathbf{R}_{b,c}$  makes it less likely for words with high weight in the topics  $b$  and  $c$  to occur in the same context. See Figure 2 for an illustrated example.

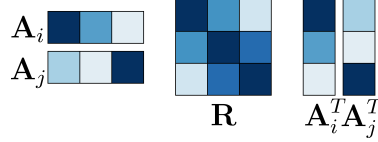


Figure 2: The affinity matrix  $\mathbf{R}$  describes the relationships between the latent factors. Illustrated here are two word embeddings, corresponding to the words  $w_i$  and  $w_j$ . Darker shades represent larger values. In this example we predict a large co-occurrence at  $\mathbf{S}_{ii}$  and  $\mathbf{S}_{jj}$  because of the large weight on the diagonal of the  $\mathbf{R}$  matrix. We predict a low co-occurrence at  $\mathbf{S}_{ij}$  and  $\mathbf{S}_{ji}$  since the large weights on  $\mathbf{A}_{i1}$  and  $\mathbf{A}_{j3}$  interact with low weights on  $\mathbf{R}_{13}$  and  $\mathbf{R}_{31}$ .

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 behaviour 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.

## 4 Experiments and Results

In the following section we describe our experimental setup in full detail<sup>4</sup> and present our results on the simultaneous topic (relation) extraction and word embedding learning task. We compare these results against competing matrix factorization methods for topic modeling, namely NMF, LDA and SVD.

### 4.1 Data

To conduct our experiments we leverage a synthetically created text corpus, whose documents consist of triplets of individual English Wikipedia articles. The articles are retrieved as raw text via the official Wikipedia API using the `wikipedia-api` library. Always three articles a time get concatenated to form a new artificially generated text document. We differentiate between thematically similar (e.g. “Dolphin” and “Whale”) and thematically different articles (e.g. “Soccer” and “Donald Trump”). Each synthetic document is categorized into one of three classes: All underlying Wikipedia articles are thematically different, two articles are thematically similar and one is different, and all articles are thematically similar. Table 3 in the appendix shows this categorization and the overall setup of our generated documents.

On each document we apply the following textual preprocessing steps. First, the whole document gets lower-cased. Second, we tokenize the text making use of the word-tokenizer from the `nltk` library and remove common English stop words, including contractions such as “you’re” and “we’ll”. Lastly we clear the text from all remaining punctuation and delete digits and single characters.

As described in Section 3 we utilize our preprocessed document text to calculate a symmetric word co-occurrence matrix, which, after being transformed to a positive PMI matrix, functions as input and target matrix for the row-stochastic DEDICOM algorithm. To avoid any bias or prior information from the structure and order of the Wikipedia articles, we randomly shuffle the vocabulary before creating the co-occurrence matrix. When generating the matrix we only consider context words within a symmetrical window of size 7 around the base word. Like in [15], each context word only contributes  $1/d$  to the total word pair count, given it is  $d$  words apart from the base word.

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

### 4.2 Training

As theoretically described in Section 3 we train row-stochastic DEDICOM with the alternating gradient descent paradigm utilizing automatic differentiation from the `PyTorch` library.

<sup>4</sup> All results are completely reproducible based on the information in this section. Our Python implementation to reproduce the results is available on <https://github.com/LarsHill/text-dedicom-paper>.



First, we initialize the factor matrices  $\mathbf{A} \in \mathbb{R}^{n \times k}$  and  $\mathbf{R} \in \mathbb{R}^{k \times k}$ , by randomly sampling all elements from a uniform distribution centered around 1,  $\mathcal{U}(0, 2)$ . Note that after applying the softmax operation on  $\mathbf{A}$  all rows of  $\mathbf{A}$  are stochastic. Therefore, scaling  $\mathbf{R}$  by

$$\bar{s} := \frac{1}{n^2} \sum_{ij}^n \mathbf{S}_{ij}, \quad (15)$$

will result in the initial decomposition  $\mathbf{A}'\mathbf{R}(\mathbf{A}')^T$  yielding reconstructed elements in the range of  $\bar{s}$ , the element mean of the PPMI matrix  $\mathbf{S}$ , and thus, speeding up convergence.

Second,  $\mathbf{A}$  and  $\mathbf{R}$  get iteratively updated employing the Adam optimizer [8] with constant individual learning rates of  $\eta_{\mathbf{A}} = 0.001$  and  $\eta_{\mathbf{R}} = 0.01$  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 train for `num_epochs` = 15,000 until convergence, where each epoch consists of an alternating gradient update with respect to  $\mathbf{A}$  and  $\mathbf{R}$ . Algorithm 1 illustrates the just described training procedure.

---

**Algorithm 1** The row-stochastic DEDICOM algorithm

---

- 1: initialize  $\mathbf{A}, \mathbf{R} \leftarrow U(0, 2) \cdot \bar{s}$  ▷ See Equation (15) for the definition of  $\bar{s}$
  - 2: initialize  $\beta_1, \beta_2, \epsilon$  ▷ Adam algorithm hyperparameters
  - 3: initialize  $\eta_{\mathbf{A}}, \eta_{\mathbf{R}}$  ▷ Individual learning rates
  - 4: **for**  $i$  in  $1, \dots, \text{num\_epochs}$  **do**
  - 5:   Calculate loss  $\mathcal{L} = \mathcal{L}(\mathbf{S}, \mathbf{A}, \mathbf{R})$  ▷ See Equation (8)
  - 6:    $\mathbf{A} \leftarrow \mathbf{A} - \text{Adam}_{\beta_1, \beta_2, \epsilon}(\nabla_{\mathbf{A}}, \eta_{\mathbf{A}})$ , where  $\nabla_{\mathbf{A}} = \frac{\partial \mathcal{L}}{\partial \mathbf{A}}$
  - 7:    $\mathbf{R} \leftarrow \mathbf{R} - \text{Adam}_{\beta_1, \beta_2, \epsilon}(\nabla_{\mathbf{R}}, \eta_{\mathbf{R}})$ , where  $\nabla_{\mathbf{R}} = \frac{\partial \mathcal{L}}{\partial \mathbf{R}}$
  - 8: **return**  $\mathbf{A}'$  and  $\mathbf{R}$ , where  $\mathbf{A}' = \text{row\_softmax}(\text{col\_norm}(\mathbf{A}))$  ▷ See Equation (5)
- 

We implement NMF, LDA and SVD using the `sklearn` library. In all cases the learnable factor matrices are initialized randomly and default hyperparameters are applied during training. For NMF the multiplicative update rule from [10] is utilized. Figure 3 shows the convergence behaviour of the row-stochastic DEDICOM training process and the final loss of NMF and SVD. Note that LDA optimizes a different loss function, which is why the calculated loss is not comparable and therefore excluded. We see that the final loss of DEDICOM locates just above the other losses, which is reasonable when considering the row stochasticity constraint on  $\mathbf{A}$  and the reduced parameter amount of  $nk + k^2$  compared to NMF ( $2nk$ ) and SVD ( $2nk + k^2$ ).

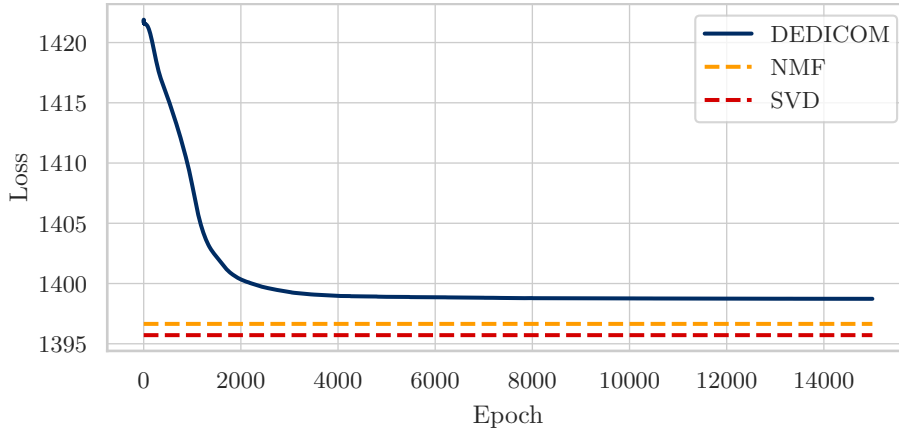


Figure 3: Reconstruction loss development during training. The  $x$ -axis plots the number of epochs, the  $y$ -axis plots the corresponding reconstruction error for each matrix factorization method.

### 4.3 Results

In the following, we present our results of training row-stochastic DEDICOM to simultaneously learn interpretable word embeddings and meaningful topic clusters and their relations. For compactness reasons we focus our main analysis on document id 3 in Table 3, “Soccer, Bee and Johnny Depp”, and set the number of latent topics to  $k = 6$ . We refer the interested reader to the appendix for results on other article combinations and comparison to other matrix factorization methods.<sup>5</sup>

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

$$\mathbf{A}'_i = [0.05 \ 0.03 \ 0.02 \ 0.14 \ 0.70 \ 0.06]$$

$$\operatorname{argmax}(\mathbf{A}'_i) = 5,$$

and thus,  $\mathbf{A}'_i$  gets matched to Topic 5. Next, we decreasingly sort the words within each topic based on their matched topic probability. Table 1 shows the overall number of allocated words and the resulting top 10 words per topic together with each matched probability.

Indicated by the high assignment probabilities, one can see that columns 1, 2, 4, 5 and 6 represent distinct topics, which easily can be interpreted. Topic 1 and 4 are related to soccer, where 1 focuses on the game mechanics and 4 on the organisational and professional aspect of the game. Topic 2 and 6 clearly

<sup>5</sup> We provide a large scale evaluation of all article combinations listed in Table 3, including different choices for  $k$ , as supplementary material at <https://bit.ly/3cBxsGI>.

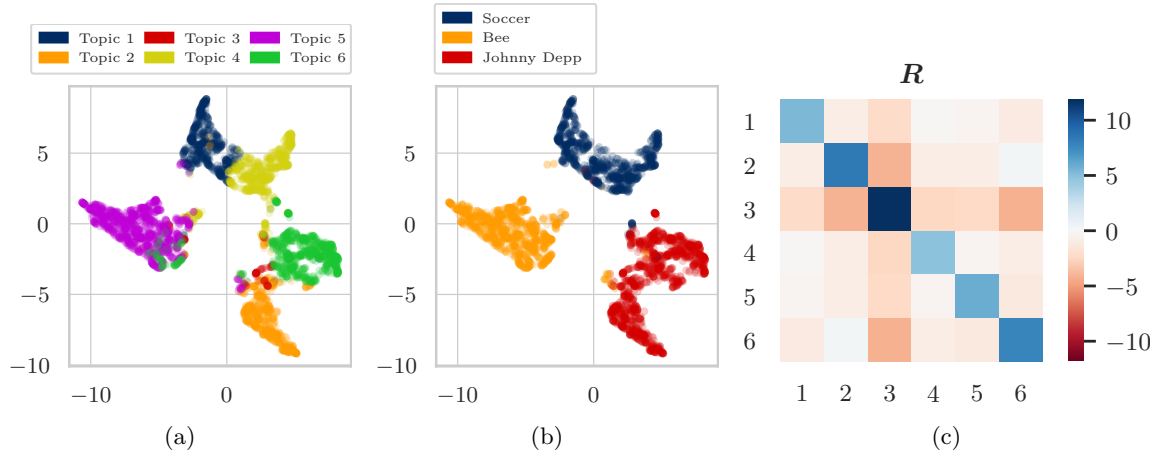


Figure 4: (a) 2-dimensional representation of word embeddings  $\mathbf{A}'$  colored by topic assignment. (b) 2-dimensional representation of word embeddings  $\mathbf{A}'$  colored by original Wikipedia article assignment (words that occur in more than one article are excluded). (c) Colored heatmap of affinity matrix  $\mathbf{R}$ .

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

Table 1: Each column lists the top 10 representative words per dimension of the basis matrix  $\mathbf{A}'$ .

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

Table 2: For the most significant two words per topic, the four nearest neighbors based on cosine similarity are listed.

refer to Johnny Depp, where 2 focuses on his acting career and 6 on his difficult relationship to Amber Heard. The fifth topic obviously relates to the insect “bee”. In contrast, Topic 3 does not allow for any interpretation and all assignment probabilities are significantly lower than for the other topics.

Further, we analyze the relations between the topics by visualizing the trained  $\mathbf{R}$  matrix as a heatmap (see Figure 4c).

One thing to note is the symmetry of  $\mathbf{R}$  which is a first indicator of a successful reconstruction,  $\mathbf{A}'\mathbf{R}(\mathbf{A}')^T$ , (see Section 3.1). Also, the main diagonal elements are consistently blue (positive), which suggests a high distinction between the topics. Although not very strong one can still see a connection between Topic 2 and 6 indicated by the light blue entry  $\mathbf{R}_{26} = \mathbf{R}_{62}$ . While the suggested relation between Topic 1 and 4 is not clearly visible, element  $\mathbf{R}_{14} = \mathbf{R}_{41}$  is the least negative one for Topic 1. In order to visualize the topic cluster quality we utilize UMAP (Uniform Manifold Approximation and Projection) [12] to map the  $k$ -dimensional word embeddings to a 2-dimensional space. Figure 4a illustrates this low-dimensional representation of  $\mathbf{A}'$ , where each word is colored based on the above described word to topic assignment. In conjunction with Table 1 one can nicely see that Topic 2 and 6 (Johnny Depp) and Topic 1 and 4 (Soccer) are

close to each other. Hence, Figure 4a implicitly shows the learned topic relations as well and arguably better than  $\mathbf{R}$ .

As an additional benchmark, Figure 4b plots the same 2-dimensional representation, but now each word is colored based on the original Wikipedia article it belonged to. Words that occur in more than one article are not considered in this plot.

Directly comparing Figure 4b and 4a shows that row-stochastic DEDICOM does not only recover the original articles but also finds entirely new topics, which in this case represent subtopics of the articles. Let us emphasize that for all thematically similar article combinations, the found topics are usually not subtopics of a single article, but rather novel topics that might span across multiple Wikipedia articles (see for example Table 5 in the appendix). 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 select within each topic the two most representative words and calculate the cosine similarity between their word embeddings and all other word embeddings stored in  $\mathbf{A}'$ . Table 2 shows the 4 nearest neighbors based on cosine similarity for the top 2 words in each topic. We observe a high thematic similarity between words with large cosine similarity, indicating the usefulness of the rows of  $\mathbf{A}'$  as word embeddings.

In comparison to DEDICOM, other matrix factorization methods also provide 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 fails on the word similarity task, with words close in cosine similarity seldom sharing the same thematic similarity we have seen in DEDICOM. This can be seen in Table 4, 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 look very reasonable, its word embeddings do not seem to carry any semantic meaning based on cosine similarity; e.g. the four nearest neighbors of “ball” are “invoke”, “replaced”, “scores” and “subdivided”. A similar nonsensical picture can be observed for the other main topic words. LDA and SVD perform slightly better on the similar word task, although not all similar words appear to be sensible, e.g. “children”, “detective”, “crime”, “magazine” and “barber”. Also, some topics cannot be clearly defined due to mixed word assignments, e.g. Topic 4 for LDA and Topic 1 for SVD.

For a comprehensive overview of our results for other article combinations, we refer to Tables 5, 6, 7, 8 and Figures 5, 6 in the Appendix.

## 5 Conclusion and Outlook

We propose a row-stochasticity 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 top modeling and interpretable word embedding learning given by our algorithm is unique in its class.

In future work we will expand on the idea of comparing topic relationships between multiple documents, possibly over time, with individual co-occurrence matrices resulting in stacked topic relationship matrices but shared word embeddings. Further extending this notion, we plan to utilize time series analysis to discover temporal relations between extracted topics and to potentially identify trends.

## 6 Acknowledgement

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.

## References

1. Andrzej, A.H., Cichocki, A., Dinh, T.V.: Nonnegative dedicom based on tensor decompositions for social networks exploration. *Aust. J. Intell. Inf. Process. Syst.* **12** (2010)
2. Bader, B.W., Harshman, R.A., Kolda, T.G.: Pattern analysis of directed graphs using dedicom: an application to enron email. (2006)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
4. 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*. pp. 54–62. Association for Computational Linguistics, Boulder, Colorado, USA (2009)
5. 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: *Proc. of ACM SIGIR* (1988)
6. Harshman, R., Green, P., Wind, Y., Lundy, M.: A model for the analysis of asymmetric data in marketing research. *Marketing Science* **1**, 205–242 (1982)
7. Jolliffe, I.: *Principal component analysis*. John Wiley and Sons Ltd (2005)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
9. Lebre, R., Collobert, R.: Word embeddings through hellinger PCA. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 482–490. Association for Computational Linguistics, Gothenburg, Sweden (2014)

10. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Proceedings of the 13th International Conference on Neural Information Processing Systems. pp. 535—541. NIPS'00, MIT Press, Cambridge, MA, USA (2000)
11. Levy, O., Goldberg, Y.: Neural word embedding as implicit matrix factorization. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. pp. 2177–2185. NIPS'14, MIT Press, Cambridge, MA, USA (2014)
12. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction (2018)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality (2013)
14. Nguyen, D.Q., Billingsley, R., Du, L., Johnson, M.: Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics* **3**(0) (2015)
15. 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). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014)
16. Sifa, R., Ojeda, C., Bauckhage, C.: User churn migration analysis with dedicom. In: Proceedings of the 9th ACM Conference on Recommender Systems. pp. 321—324. RecSys '15, Association for Computing Machinery, New York, NY, USA (2015)
17. Sifa, R., Ojeda, C., Cvejowski, K., Bauckhage, C.: Interpretable matrix factorization with stochasticity constrained nonnegative dedicom (2018)
18. YongchangWang, Zhu, L.: Research and implementation of svd in machine learning. In: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). pp. 471–475 (2017)

## Appendix

ID	Selection Type	Article 1	Article 2	Article 3
1	different	Donald Trump	New York City	Shark
2		Shark	Bee	Elephant
3		Soccer	Bee	Johnny Depp
4		Tennis	Dolphin	New York City
5	mixed	Donald Trump	New York City	Michael Bloomberg
6		Soccer	Tennis	Boxing
7		Brad Pitt	Leonardo Dicaprio	Rafael Nadal
8		Apple (company)	Google	Walmart
9	similar	Shark	Dolphin	Whale
10		Germany	Belgium	France
11		Soccer	Tennis	Rugby football
12		Apple (company)	Google	Amazon (company)

Table 3: Overview of our semi-artificial dataset. Each synthetic sample consists of the corresponding Wikipedia articles 1 – 3. We differentiate between *different* articles, i.e. articles that have little thematical overlap (for example a person and a city, a fish and an insect or a ball game and a combat sport), and *similar* articles, i.e. articles with large thematical overlap (for example European countries, tech companies or aquatic animals). We group our dataset into different samples (3 articles that are pairwise different), similar samples (3 articles that are all similar) and mixed samples (2 similar articles, 1 different).



## Articles: “Soccer”, “Bee”, “Johnny Depp”

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
NMF	#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 play	eusocial	jack	uefa	stating	debut
LDA	0 ball	bees	film	football	heard	album
	1 invoke	odors	burtondirected	athenaeus	crew	jones
	2 replaced	tufts	tone	paralympic	alleging	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
SVD	#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
	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	carcasses
	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
	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	confederations	finalized	mites
	4 crops	march	clear	conmebol	submitted	plato
	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 4: 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.

Articles: “Dolphin”, “Shark”, “Whale”

	Topic 1 #460	Topic 2 #665	Topic 3 #801	Topic 4 #753	Topic 5 #854	Topic 6 #721
1	shark (0.665)	calf (0.428)	ship (0.459)	conservation (0.334)	water (0.416)	dolphin (0.691)
2	sharks (0.645)	months (0.407)	became (0.448)	countries (0.312)	similar (0.374)	dolphins (0.655)
3	fins (0.487)	calves (0.407)	poseidon (0.44)	government (0.309)	tissue (0.373)	captivity (0.549)
4	killed (0.454)	females (0.399)	riding (0.426)	wales (0.304)	body (0.365)	wild (0.467)
5	million (0.451)	blubber (0.374)	dionysus (0.422)	bycatch (0.29)	swimming (0.357)	behavior (0.461)
6	fish (0.448)	young (0.37)	ancient (0.42)	cancelled (0.288)	blood (0.346)	bottlenose (0.453)
7	international (0.442)	sperm (0.356)	deity (0.412)	eastern (0.287)	surface (0.344)	sometimes (0.449)
8	fin (0.421)	born (0.355)	ago (0.398)	policy (0.286)	oxygen (0.34)	human (0.421)
9	fishing (0.405)	feed (0.349)	melicertes (0.395)	control (0.285)	system (0.336)	less (0.42)
10	teeth (0.398)	mysticetes (0.341)	greeks (0.394)	imminent (0.282)	swim (0.336)	various (0.418)
0	shark (1.0)	calf (1.0)	ship (1.0)	conservation (1.0)	water (1.0)	dolphin (1.0)
2	sharks (0.981)	calves (0.978)	dionysus (0.995)	south (0.981)	prey (0.964)	dolphins (0.925)
3	fins (0.958)	females (0.976)	riding (0.992)	states (0.981)	swimming (0.959)	sometimes (0.909)
4	killed (0.929)	months (0.955)	deity (0.992)	united (0.978)	allows (0.957)	another (0.904)
5	fishing (0.916)	young (0.948)	poseidon (0.987)	endangered (0.976)	swim (0.947)	bottlenose (0.903)
0	sharks (1.0)	months (1.0)	became (1.0)	countries (1.0)	similar (1.0)	dolphins (1.0)
2	shark (0.981)	born (0.992)	old (0.953)	eastern (0.991)	surface (0.992)	behavior (0.956)
3	fins (0.936)	young (0.992)	later (0.946)	united (0.989)	brain (0.97)	sometimes (0.945)
4	tiger (0.894)	sperm (0.985)	ago (0.939)	caught (0.987)	sound (0.968)	various (0.943)
5	killed (0.887)	calves (0.984)	modern (0.937)	south (0.979)	object (0.965)	less (0.937)

Table 5: Top half lists the top 10 representative words per dimension of the basis matrix  $A$ , bottom half lists the 5 most similar words based on cosine similarity for the 2 top words from each topic.

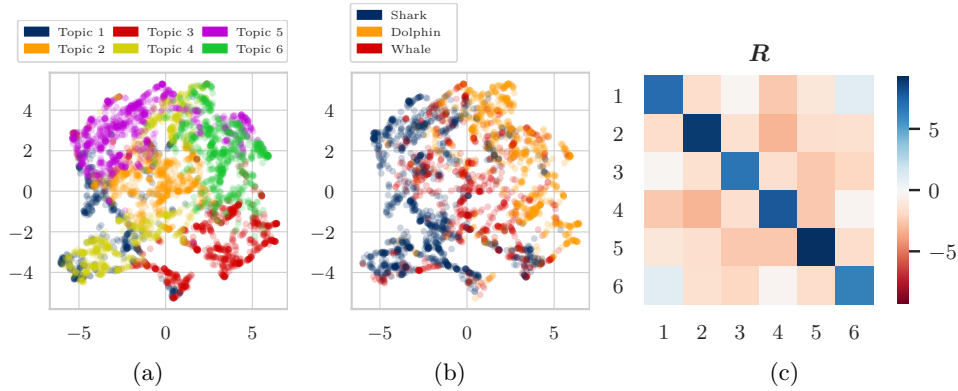


Figure 5: (a) 2-dimensional representation of word embeddings  $A'$  colored by topic assignment. (b) 2-dimensional representation of word embeddings  $A'$  colored by original Wikipedia article assignment (words that occur in more than one article are excluded). (c) Colored heatmap of affinity matrix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

## Articles: “Dolphin”, “Shark”, “Whale”

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
NMF	#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
	10 via	islands	riding	similar	extinct	female
LDA	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	fuel	brood
	2 blocks	consumers	humankind	helping	lamp	lacking
	3 modified	pontoporiidae	helpers	waste	filterfeeding	accurate
	4 slits	org	performing	depression	krill	consistency
SVD	#650	#785	#695	#815	#635	#674
	1 killed	teeth	head	species	meat	air
	2 system	baleen	fish	male	whale	using
	3 endangered	mysticetes	dolphin	females	ft	causing
	4 often	ago	fin	whales	fisheries	currents
	5 close	jaw	eyes	sometimes	also	sounds
	6 sharks	family	fat	captivity	ocean	groups
	7 countries	water	navy	young	threats	sound
	8 since	includes	popular	shark	children	research
	9 called	allow	tissue	female	population	clicks
SVD	10 vessels	greater	tail	wild	bottom	burst
	0 killed	teeth	head	species	meat	air
	1 postures	dense	underside	along	porbeagle	australis
	2 dolphinariums	cetacea	grooves	another	source	submerged
	3 town	tourism	eyesight	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
SVD	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	fin	mya	deity	called
	8 one	air	law	later	ino	used
SVD	9 animals	feed	new	months	came	islands
	10 use	bony	conservation	mysticetes	made	genus
	0 dolphins	water	shark	million	poseidon	dolphin
	1 various	vertical	corpse	approximately	games	depicted
	2 finding	unlike	stocks	assigned	phalanthus	makara
	3 military	chew	galea	hybodonts	statue	capensis
	4 selfmade	lack	galeomorphii	appeared	isthmian	goddess
	0 species	body	sharks	years	became	meat
	1 herd	heart	mostly	acanthodians	pirates	contaminated
	2 reproduction	resisting	fda	spent	elder	harpoon
SVD	3 afford	fit	lists	stretching	mistook	practitioner
	4 maturity	posterior	carcharias	informal	wealthy	pcbs

Table 6: 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.

Articles: “Soccer”, “Tennis”, “Rugby”

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

Table 7: Top half lists the top 10 representative words per dimension of the basis matrix  $A$ , bottom half lists the 5 most similar words based on cosine similarity for the 2 top words from each topic.

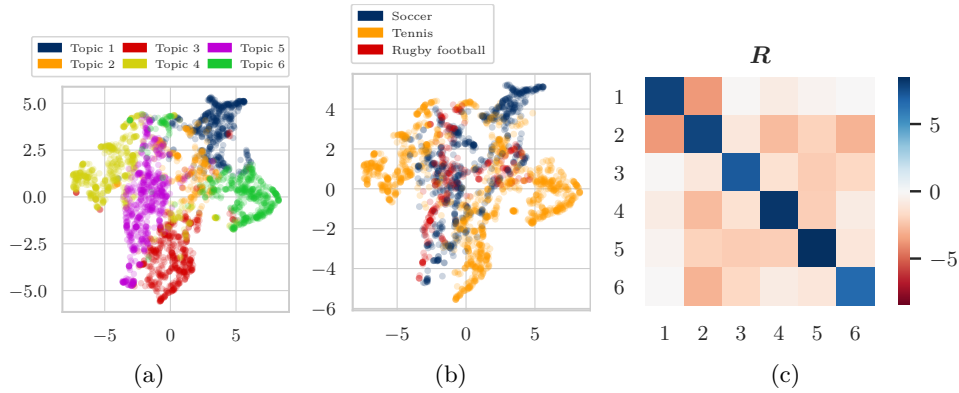


Figure 6: (a) 2-dimensional representation of word embeddings  $A'$  colored by topic assignment. (b) 2-dimensional representation of word embeddings  $A'$  colored by original Wikipedia article assignment (words that occur in more than one article are excluded). (c) Colored heatmap of affinity matrix  $R$ .

## Articles: “Soccer”, “Tennis”, “Rugby”

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

Table 8: 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.