# Principal Component Analysis in Topic Modelling of Short Text Document Collections

Hennadii Dobrovolskyi, Nataliya Keberle

Department of Computer Science, Zaporizhzhya National University,
Zhukovskogo st. 66, 69600, Zaporizhzhya, Ukraine,
gen.dobr@gmail.com, nkeberle@gmail.com

**Abstract.** This paper presents the motivation for and the preliminary theoretical investigations of the PhD project by the first author. The objective of the research is to propose and to experimentally verify the approach of application of eigendecomposition in principal component analysis for topic modelling of short text document collections. The main hypothesis examined in this project, is that principal component analysis applied to word co-occurrence statistics turns topic modelling into well-defined problem having unique solution with natural fitting parameters. The project is performed at the Dept. of Computer Science of Zaporizhzhya National University.

**Keywords:** text mining, short text document, topic modelling, principal component analysis, eigendecomposition, clusterization.
**KeyTerms:** MathematicalModel, MachineIntelligence, DescriptiveModel, KnowledgeRepresentation

## 1 Introduction

This paper presents a PhD project aimed at developing the method for probabilistic topic modelling of collections of short text documents. It is assumed that documents are literate texts or have another known structure that allows to discover in a natural way all the links between terms. For instance collection of scientific paper abstracts and titles contains well-formed sentences that can be parsed with NLP tools. The project concentrates on the analysis scientific abstracts covering one vague domain of knowledge. That means, documents have one principal topic and some extra ones associated with related domains of knowledge.

It is well known that number of scientific publication grows faster than average scientist can analyze. Thus it is important to have a tool that can maintain actual state of documents collection ensuring that it completely covers a domain of interest. Therefore developing a well-grounded method to determine topics an unknown document belongs to is useful.

The main hypothesis examined in this project, is that principal component analysis[1] applied to word co-occurrence statistics turns topic modelling into well- defined problem having unique solution with natural fitting parameters.

The discovered topics can be used to represent short documents as vector of real numbers appropriate for retrieval and clustering. Moreover the terms associated with topics can be used to search for new documents and to extend the collection.

The known common document topic modelling is the ill-posed problem that does not have unique solution. Therefore different additional conditions are added and combined to get comprehensible topic models. Often many restrictions are poorly grounded heuristics that require diverse tricks to combine them[2]. Applying common document topic modelling to short texts leads to the lower quality of discovered topics if compared to ones derived from long texts.

The objective of the presented project is to develop and evaluate a method to determine a set of topics of short text document collection and assign topic weights to each document in the collection.

As a theoretical background the project uses the natural language processing (NLP) methods like part-of-speech tagging [3], stemming [4], sentence splitting and parsing [5]. Information retrieval approaches [4] are used to exclude insignificant information during preprocessing. Following the mainstream of probabilistic topic model the Principal Component Analysis [1] is applied to derive the most significant collection topics from word co-occurrence frequencies. The HEP collection [6] provides a sample data to verify the suggested method.

The rest of the paper is organized as follows. Section 2 contains a short review of the related work. This is followed by short description of the suggested method in Section 3. Experiment goals and workflow is illustrated and explained in Section 4. Finally, the conclusive remarks are presented briefly in section 5 and several possible directions in future are also pointed out.

## 2 Related Works and Motivation

Probabilistic topic models [2] are the set of algorithms providing a statistical solution to the problem of handling large collection of documents. The basic idea behind the topic modelling is to construct a low-dimensional document representation using few groups of tightly connected significant terms instead of separate words. The most known method of topic modelling is Latent Dirichlet Allocations (LDA) [7] which overcomes deficiencies of earlier approaches and is successful and simple enough. A general introduction and survey of the topic modelling can be found in [2] along with a novel approach, called Additive Regularization of Topic Models. However the primary direction of topic model enhancement still is a regularization, i.e. incorporating different restrictions into basic algorithm. Origins of the restrictions are not limited and sometimes (as in LDA) the additional condition is applied because it is manageable and it works.

Another drawback of common topic model is the shorter are documents in a collection the less accurate is the result. It is overpassed with approaches utilizing word co-occurrence statistics [8, 9] instead of counting document-word pairs.

Similar results can be reached in quite a different way, with a combination of NLP and clustering algorithms [10]. However the clustering in a high-dimensional discrete space is a time demanding problem. So mapping documents to low-dimensional space $R^n$ can accelerate the clustering and subsequent analysis.

Therefore, the method of topic modelling that replaces the magic restrictions with comprehensible ones will be valuable. The project presented in this paper aims at the development, evaluation and application of the method based on the PCA approach for word co-occurrence probabilities.

## 3 Method Description

Let $D$ be a collection of documents, $W$ - a dictionary containing all terms used in $D$. Each document $d \in D$ is a sequence of $n_d$ terms $(w_1, \ldots, w_{n_d})$. The term can occur many times in the document. "Term" may be a word or group of words.

The suggested method shares with the common document topic model [2] the following assumptions:

Assumption 1: Each term $w$ in the document $d$ is related to a topic $t$ from a set of topics $T$. Collection of documents is formed as set of triples $(d, w, t)$, independently selected in a random way from discrete probability $p\left(d, w, t\right)$ defined over set $D \times W \times T$. The document $d$ and the term $w$ are observable and the topic $t$ is the hidden parameter.

Assumption 2. Order of terms in a document doesn't matter.

Assumption 3: Order of documents in the collection doesn't matter.

Assumption 4: Conditional probability $p\left(w|d, t\right)$ is independent on the document $d$, i.e. $p\left(w|d, t\right) = p\left(w|t\right)$.

As well as Word Network Topic Model [8] and Biterm Topic Model [9] the suggested method utilizes probability $p\left(w_i, w_k\right)$ that both word $w_i$ and word $w_k$ occur in the same document or document fragment

$$p\left(w_i, w_k\right) = \sum_{t=1}^{T} p\left(w_i|t\right) p\left(t\right) p\left(w_k|t\right) \tag{1}$$

where $p\left(w_i, w_k\right)$ is a joint probability and $t$ is a topic identifier. In the presented project the probability is estimated as relative number of pairs $(w_i, w_k)$.

Term pairs $(w_i, w_k)$ are collected in two steps. First, each document $d_k$ in the collection is mapped to set of short term sequences $S\left(d_k\right) = (s_{k1}, s_{k2}, \ldots)$, where $s_{kq} = (w_{kq1}, \ldots, w_{kqr})$. Second, each sentence $s_{kq}$ is mapped to pairs $(w_i, w_k)$, $w_i \in s_{kq}$ , $w_k \in s_{kq}$ , $w_i \neq w_k$ .

Topic model creation is an estimation of probabilities $p\left(w_i|t\right)$ and $p\left(t|d_k\right)$. It is assumed that a number of significant topics is far smaller than the number of words and the number of documents that simplifies the further manipulations like search, comparison, clustering etc.

In our document generation model, the document $d_k$ is represented with the set of conditional probabilities $p\left(t|d_k\right)$. $d_k$ is a bag of terms and we apply the Gibbs sampling to create such a bag of terms. First, the document covariance matrix $p\left(w_i, w_k\right)$ is calculated using Eq.(1) where topic probabilities $p(t)$ are replaced with $p\left(t|d_k\right)$. Second, a random topic $t$ is selected according to the conditional probability $p(t|d_k)$ and the initial set of $N$ terms for the document is randomly selected based on the term probabilities $p(w_i|t)$. Third, the repetitive sampling is used to replace each term in the document. One iteration of the sampling is a three-step process:

1. choose the term position $j$ that will be updated;

2. calculate naïve Bayes probability for each term $w$ in the dictionary $W$

$$p\left(w|w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_N\right) = \frac{1}{Z}\, p\left(w\right) \prod_{i \neq j} \frac{p(w, w_i)}{p(w)} \tag{2}$$

where $Z$ is a normalizing denominator and $w_i$ is a term placed at $i$-th position in the document;

3. get new random term $w$ according to the probability (2) and place it at the position $j$.

In the presented work, dimensionality of covariance matrix is decreased through stemming and omitting words which are not nouns or adjectives, stop-words, and rare words.

Words which are not nouns or adjectives are readily detected with part-of-speech tagger [3]. They are proved to make small contribution to document topic assignment [10].

Stop-words are the terms that do not affect topic detection. There are two groups of stop-words: collection-specific and common stop-words. Various lists of common words are available online[1] but the collection-specific ones have to be constructed.

To extract a set of collection-specific stop-words the covariance $p\left(w_i, w_j\right)$ is employed. The hypothesis is that the stop-word has a large value of the Shannon information entropy

$$H(w_i) = -\sum_{j=1}^{|W|} p\left(w_i, w_j\right)\, log\left[p\left(w_i, w_j\right)\right] \tag{3}$$

The value of $H(w_i)$ exceeding some threshold value $H_{max}$ signals that $w_i$ can accompany any other word. Therefore it is not effective when detecting topics and should be dropped out. $H_{max}$ may be considered as additional parameter to adjust the algorithm.

Rare words are detected with comparison of the single word probability $p(w_i)$ and a threshold value $P_{max}$ where

$$p(w_i) = \sum_{j=1}^{|W|} p\left(w_i, w_j\right) \tag{4}$$

[1] For instance list of English stop words is available at Snowball stemmer site http://snowball.tartarus.org/algorithms/english/stop.txt

One of the ways to define $P_{max}$ is to require that cumulative distribution function equals to some parameter $\alpha$

$$\alpha = \sum_{p(w_i)\geq P_{max}} p(w_i) \tag{5}$$

That means, the kept terms cover predefined percentage $\alpha$ of occurrences.

After all the excessive words are dropped out the joint probability matrix becomes much smaller and should be decomposed into product of three matrixes according to Eq.(1). The main point of the presented method is setting number of topics $T$ to dimensionality of the square covariance matrix $P_{ij}$ . Then Eq.(1) becomes eigendecomposition problem such that its solution produces conditional word probabilities $p\left(w_j|t\right)$ as eigenvectors and topic probabilities $p(t)$ as eigenvalues.

Next step is to reduce the number of topics. Method of Principal Component Analysis [1] states that the matrix $P_{ij}$ can be approximated by setting the smallest values of topic probabilities $p(t)$ to zero. Also PCA suggests a way to select the most significant topics relying on calculated topic probabilities. After values of $p(w_j|t)$ are calculated, the topic detection of the document is performed using the expression

$$p(t|d) = \sum_{i=1}^{|W|} p\left(t|w_i\right) p\left(w_i|d\right) \tag{6}$$

where $p(t|w_i)$ is found from the Bayes equation

$$p(t|w_i) = \frac{p(w_i|t)p(t)}{p(w_i)} \tag{7}$$

$p(w_i)$ ( see Eq.(4) ) is the probability of word $w_i$ occurs in the collection, and $p(w_i|d)$ is the relative frequency of word $w_i$ in document d.

The comprehensive and automated evaluation measure of topic quality is Topic Coherence [11]

$$C(z, M) = \sum_{t_1=1}^{T} \sum_{t_2=1}^{t_1-1} log\left[\frac{D(m_{t_1}, m_{t_2}) + \epsilon}{D(m_{t_1})D(m_{t_2})}\right] \tag{8}$$

where $M = (m_1, \ldots, m_T)$ is the list of the $T$ most probable terms in a topic $z$, $D(m)$ counts the number of documents containing the term $m$, $D(m_1, m_2)$ counts the number of documents containing both $m_1$ and $m_2$, and $\epsilon$ is used to avoid $log(0)$. The evaluation metric of the entire topic model is the average coherence score of all topics. Topic coherence is directly related to probability the top topic terms can be found in the same document. So the higher topic coherence indicates better topic quality.

## 4 Experiment planning

Experiments aim to check if the method presented above does provide the valid and high-quality topic definitions. To answer the main question experiments should explore the impact of factors enlisted in the Section 3 on the quality of topic definitions, namely:

1. How does the topic model quality depend on the threshold value of information entropy?
2. How does the topic model quality depend on the threshold value of rare word frequency?
3. Which of the available PCA decompositions fits the explored method?
4. How does the lower limit of topic probability influence the quality of results?
5. Which method of word pairs extraction is better:
   (a) combination of consecutive words (as in [10]);
   (b) combination of adjacent terms in a grammar tree;
   (c) combination of all possible words in separate sentence;
   (d) all possible word pairs in sliding window of size $r$ [8, 9].

General experiment workflow contains the following steps:

1. Extract title and abstract from each document of the collection;
2. Extract word pairs from titles and abstracts using one of the word pairs extraction methods;
3. Apply stemming, omit words which are not nouns or adjectives, stop-words, and rare words setting $H_{max}$, $\alpha$;
4. Apply one of PCA alternatives to extract word probabilities in topics, setting minimal value of topic probability;
5. Calculate average topic coherence to measure quality of topic set.

The experiment will use the HEP data collection [6] which is oriented to the study of multi-label classifiers text. It consists of scientific papers in the field of High Energy Physics (HEP) obtained from the document server of European Nuclear Physics Laboratory (CERN).

The experiments should show the dependencies of average topic coherence on $H_{max}$, $\alpha$, minimal value of topic probability, type of PCA and extraction method.

## 5 Conclusive remarks and future works

The project has been started in December 2016 and is in stage of detailed planning of experiments and exploration of background technologies and theories. The future plans include (a) implementation of all the necessary software components; (b) evaluating quality of proposed basic algorithms; (c) component integration and running all the workflow; (d) application of developed method to practical tasks.

## References

1. Jolliffe, I. Principal Component Analysis (2ed.). Springer, 2002.
2. Vorontsov, K.V., Potapenko, A.A.: Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. In: Ignatov, D.I. et al. (Eds.) Proc. AIST2014, CCIS 436, pp. 29-46 (2014).
3. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Hearst, M. and Ostendorf, M. (Eds.) Proc. HLT-NAACL2003, pp. 252-259 (2003)
4. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, 2008.
5. Chen, D., Manning, C.D.: A Fast and Accurate Dependency Parser using Neural Networks. In: Moschitti, A. et al. (Eds.) Proc. EMNLP 2014, pp.740-750 (2014)
6. Montejo-Rez, A., Steinberger, R., Urea-Lpez, L.A.: Adaptive Selection of Base Classifiers in One-Against-All Learning for Large Multi-Labeled Collections. In: Vicedo J. L. et al. (Eds.) Proc. 4th Intl Conf. Advances in Natural Language Processing (EsTAL2004), LNAI 3230, pp.1-12 (2004)
7. Blei, D.M., Ng, A., Jordan, M.I.: Latent Dirichlet Allocation. In: J. Machine Learning Research, Vol. 3, pp.993-1022 (2003)
8. Zuo, Yu., Zhao, Ji., Xu, K.: Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts. The Computer Research Repository (CoRR), http://arxiv.org/abs/1412.5404, December 2014.
9. Yan, X., Guo, Ji., Lan, Ya., Cheng, Xu.,A Biterm Topic Model For Short Texts. In: Schwabe, D. et al.(Eds.) Proc. 22nd Intl Conf. World Wide Web, ACM, pp.1445-1456 (2013)
10. Popova, S., Khodyrev, I., Egorov, A., Logvin, S., Gulayev, S., Karpova, M., Mouromtsev, D.: Sci-Search: Academic Search and Analysis System Based on Keyphrases. In: Klinov, P., Mouromtsev, D. (Eds.) Proc. 4th Intl Conf. Knowledge Engineering and the Semantic Web, CCIS 394, pp. 281-288 (2013)
11. Aletras, N., Stevenson, M.: Evaluating Topic Coherence Using Distributional Semantics. In: Proc. 10th Intl Workshop on Computational Semantics (IWCS2013), pp.13-22 (2013)