Using Quantum Probability for Word Embedding Problem

Aleksandr Trukhanov[0000−0002−5054−3439],
Aleksei Platonov[0000−0002−8485−1296], and Igor Bessmertny[0000−0001−6711−6399]

ITMO University, Kronverksky Ave. 49, St.Petersburg, 197101, Russian Federation
{astrukhanov, avplatonov, bessmertny}@itmo.ru

Abstract. Over the past years, there has been being a contradiction between the growth rate of data that is available to humanity and the possibilities of their intellectual processing. Most of the knowledge that mankind operates is stored in the form of text documents in natural languages, which are not accompanied by additional markup tools for automated text processing tools. Thus, the exponential increase of the amount of information in the baggage of knowledge of mankind is faced with the inability to process it effectively. To resolve this contradiction, there are systems for the automatic processing of natural language data. Most intelligent data processing algorithms operate with numerical data, so the basic task of any process of working with natural language texts is to represent text units in numerical form. In our research we propose to use framework of of the quantum theory of probabilities. In this case we can operating correctly with as clean as entangled states of words. For implementation of calculation of the matrix for generalized context we using the machine learning technique, named gradient descent, and apply some of restrictions ensuring for elimination extra degrees of freedom. Our approach provides a probabilistic interpretation of research results. And it allows easy to find a probability that word's context similar to another one. The proposed model of word will can be to integrate to different text data analysis processes. This paper presents the results of a comparison of our proposed method with other similar algorithms.

Keywords: Word Embedding · Natural Language Processing · Quantum Probability Theory· Machine Learning

1 Introduction

The possibility of a qualitative analysis of the context of words is one of the most important tasks of our time. It is very important to find a mathematical description that would make it possible to make predictions about the meanings of words with high accuracy. Today, the most promising approach is to describe

the context of a word through its surround. We can imagine a word as a vector, which is the sum of all its contexts. And the purpose of this study is to find a formal description of the vector space of various contexts of words on with help set of texts. In this article, to solve the problem of finding context, we use the mathematical framework of the quantum theory of probabilities. And to calculate the context in vector space, we use large sparse matrices is defined by the power of the set of words.

2 Related Works

Today, there are a number of basic approaches to modeling such spaces. In the general case, all approaches can be divided into two large groups: distribution [1] and structural [2]. In the structural approach, parsing of text data is performed and a parsing tree is constructed, which is used to identify semantic relations between words. A set of such relations for a word defines its semantic representation.The structural approach is most typical for knowledge base systems, for example, ontologies, in which the relations between concepts are specified explicitly, we can immediately determine the severity of these relations, which ultimately allows we to build some kind of analogue of the space in which the concepts are defined and a metric for determining their similarities. However, in most cases, building knowledge bases is done manually or by automated means, which certainly affects the speed of developing tools using such representations. In the case of the distributive approach, the Harris distribution hypothesis is used, which says that words with close semantic meanings will be more often found in texts that have similar sets of words. According to this hypothesis, the concept of a word context is introduced, which includes, in its simplest form, the set of surrounding words in a fixed-size window taken in the original text sample. In this paper, we use the distribution hypothesis, and all further algorithms considered, including the approach we propose, are based on this hypothesis.
All algorithms based on the distribution approach require large sets of texts. But also in most algorithms no manual markup is required. And we can build vector representations of words without a teacher. The simplest model that allows vectorization of words is the Bag-of-Word algorithm [3]. In order to obtain a vector representation of a word in this algorithm, we need to select a fixed-size scanning window, and calculate the frequency of words, in such a context for a given word for the entire set of texts. The calculated word frequencies are used as values for the vector components corresponding to the word of their context. Finally, the third is an approach that uses neural networks, and which is now the most popular. In particular, the Word2Vec algorithm and its derivatives [4]. In this algorithm, a set of vectors is supplied to the input of a single-layer perceptron, each of which is a vector representing a word from the context, modeled for example by the Bag-of-Words model. At its output, we get the same vector, but for the central word in the window in question. This neural network in a hidden layer will have a vector of small dimension, about several hundred components, than the size of the entire lexicon. In a model trained in predicting contexts, a

hidden layer is used to represent a word vector. Thus, the Word2Vec algortime constructs dense small-size vectors in contrast to the previous algorithms that generate sparse representations [7].

3 Quantum probability theory in problem of vector representation of words

3.1 Passing word meanings through vector representation

With the work [8] begins the presentation of the possibilities of using the apparatus of quantum theory for information retrieval models. This paper uses quantum formalism to model documents, users, and ranking. Further, this approach was developed and a whole direction in information retrieval, called Quantum Information Retrieval, was born, which is engaged in modeling information retrieval processes using the framework of quantum mathematics and applies these models for the needs of information retrieval, for example ranking.
One of the most important areas in information retrieval is the solution of the problem of the vector representation of objects involved in information retrieval, such as words or documents. This direction is important because vector representations allow we to enter comparison operations using mathematical tools (for example, using a cosine metric), which certainly increases the interpretability of the model and the possibility of its use in other algorithms as opposed to algorithms based on classical machine learning.
Consider the problem of constructing a vector of word, that will to represent the meaning this word. To solve this problem, it is necessary to develop an algorithm, that takes a word and returns L-dimensional vector $V_{word} = [v_1 v_2 ... v_L]^T$ in the space of real numbers as the output. This operation can be written as:

$$a(W) : W \to R^L \tag{1}$$

where W is a set of all words, and $w \in W$ is a target word. In this paper, it is assumed that the meaning of the word in question is determined through its context, which corresponds to the Harris distribution hypothesis. If the context of the word is represented by a vector, then in this representation for the i-th position of the vector the number of repetitions of the i-th word in the dictionary W will be assigned in a window around the word w. So for the sentence "he want eat" and a dictionary consisting of these three words, the context of the word "want" will be determined by the vector $v = [101]^T$.

3.2 Tomography of quantum states

Quantum physics allows us to describe the behavior of quantum particles that are not amenable to direct observation. So we cannot track, for example, the momentum and coordinate of an electron without changing its trajectory. Therefore,

in quantum physics there is no way to describe the behavior of individual particles, and physicists tend to describe the state of ensembles of particles, resorting naturally to statistical methods.
In quantum physics, the described system can be represented by some vector $|\Psi\rangle$ in a Hilbert space. The distributions of all possible states that the system can take are described by the density matrix (operator), which was proposed in the works of L.D. Landau and J. von Neumann. The state of a quantum system is described in some basis of the chosen Hilbert space and the density operator is a function of the density distribution in the phase space of this system.
In the general case, for a selected basis $|\Psi_1\rangle \cdots |\Psi_n\rangle$, the state of a quantum system can be described by a density matrix according to the following formula:

$$\rho = \sum_{i=1}^{n} p_i \left|\Psi_i\right\rangle \left\langle\Psi_i\right|, \tag{2}$$

where $\langle\Psi_i|$ means the conjugate vector for the vector $|\Psi_i\rangle$, and p_i is the probability of finding the system under study in the state $|\Psi_i\rangle$. If the system is in a pure state, then it can be uniquely described by only one vector in the selected basis and will have a density matrix $\rho = |\Psi\rangle \langle\Psi|$. In accordance with the classical theory of probability, a pure state corresponds to an elementary event in probability space. If the system cannot be described by a single vector, then it is described by expression (2) and this state is called mixed.
Thus the quantum probability theory uses vectors and matrices to describe quantum systems [5]. Thus, it is a convenient framework to get a vector representation of a word. Quantum probability theory, in isolation from quantum physics, can be interpreted by a geometric generalization of classical probability theory. There is an equation in this theory, which binds the state of the system under study, the expected state of the system, and the probability of observing the expected state: [6]:

$$\langle A\rangle = Tr(A\rho), \tag{3}$$

where ρ is the density matrix describing the distribution of the Ψ system states, A is the projector on subspace corresponding to the expected state of the system, $\langle A\rangle$ is the probability observation of the state. Matrices ρ and A have the equal dimensions. Then, the probability that system Ψ is in state A is equal to calculating the trace of the product between density matrix D_Ψ, and the projector to the subspace of the state.

3.3 Quantum probability theory analogy

In this work, we use the hypothesis that the framework of density matrices is applicable to modeling vector or matrix representations of natural language words. This hypothesis is based on two facts. First, there are experiments showing that the vectors obtained by modeling using word contexts have a quantum-like statistical structure. This gives some reason to assume that the quantum theory

framework can be used to model vector representations of words. Secondly, if we represent the contexts in which words can appear (for example, bag-of-words vectors, thematic modeling vectors and others) as observable, that is, as some basic vectors, and the average for such expected is the probability of this word appearing in this context, when constructing projectors on context vectors, we can restore the density matrix for a given word. Consider the following approach. Let A_k be the matrix-projector on the subspace corresponding to some context k. This context is in set of N contexts.Consider context k, which is one of the known contexts N.To obtain such a projector, it is necessary to present the context of the word as a vector. Since the matrix (3) obtained from such vector must have the properties of a projector, the normalization condition of such vector is necessary.

$$A_k = \frac{v_k \cdot v_k^T}{|v_k|^2}, \tag{4}$$

We know the value of P_{A_k}, since we know the corpus of texts on which training is conducted. Then it is sufficient to group the context vectors according to their exact coincidence and express the probability of their occurrence in terms of the frequency, and we get a probability distribution on N-contexts for the word under study. In the particular case, the probability of any such context will be equal to $\frac{1}{N}$, if all contexts are unique.

3.4 Projector-Matrix Search Task

Continuing the discussion of the preceding sections, the density matrix ρ is a description of the state of the "system" corresponding to the word under study. It is this matrix that needs to be restored from expression 3, and using this to get the matrix representation of the generalized context of the word. Unlike quantum theory, we will further consider all matrices as objects over a field of real numbers. Extending the model to the field of complex numbers is the next step in our study. Thus, the problem of obtaining the representation of a word is reduced to the problem of finding a matrix satisfying equation 3. More detailed representation of the matrices is presented below:

$$\rho = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1L} \\ \vdots & \ddots & \vdots \\ \rho_{L1} & \cdots & \rho_{LL} \end{bmatrix}, A_k = \begin{bmatrix} a_{11} & \cdots & a_{1L} \\ \vdots & \ddots & \vdots \\ a_{L1} & \cdots & a_{LL} \end{bmatrix}. \tag{5}$$

Thus, the following expression shows how can be to calculate the trace of the product between these matrices.

$$Tr(A_k \rho) = \sum_{i=1}^{L} \sum_{j=1}^{L} a_{ij} \cdot \rho_{ji} = P_k \tag{6}$$

where P_{A_k} is the probability of a specific context.

3.5 Density matrix recovery

In the work of Pivovarsky [9], we can trace the approach to obtaining density matrices based on obtaining a weighted sum of the matrices of projectors on the contexts of a word:

$$\rho = \sum_{k=1}^{L} \nu_k A_k, \ \sum_{k=1}^{L} \nu = 1, \ \nu_k \geq 0 \tag{7}$$

In such a sum, the weights correspond to the frequencies of occurrence of the word contexts. This approach is simple from a computational point of view and it retains all the properties of the density matrix from formula:

$$\begin{cases} \rho_{i,i} \geq 0 \\ Tr\rho = 1 \\ Tr(\rho^2) \leq 1 \\ \rho_{ij} = \overline{\rho_{ji}}, i \neq j \end{cases} \tag{8}$$

However, it is not suitable for non-orthogonal contexts of words, as well as for systems that are mixed states. Such an algorithm does not restore the target distribution for contexts.
In our experiments, we determined that it is necessary to restore the density matrix based on a metric called Kullback–Leibler divergence:

$$D_{KL}(\rho||P) = -\sum_{k=1\ldots L} Tr(\rho \cdot A_k) \cdot ln \frac{\nu_k}{Tr(\rho \cdot A_k)}, \tag{9}$$

This metric allows us to approximate the initial distribution in contexts to that which is actually in the training data. In this paper, we propose to use the approach often used in machine learning, based on solving the optimization problem by the gradient descent method. If we take metric 9 as the main objective function, and $\lambda_i(\rho)$ is considered as a set of regularizers for preserving the properties of density matrix 3, then we obtain the following optimization problem:

$$\min_{\rho} Q(\rho, P) = D_{KL}(\rho||P) + \sum_{i=1}^{R} \lambda_i(\rho) \rightarrow \min_{\rho}, \tag{10}$$

Here R is the number of regularizers λ. And the expression of the gradient in terms of the density matrix parameters for formula 10 will have the following form:

$$\nabla Q(\rho, P) = \sum_{k=1}^{K}(ln\frac{\nu_k}{Tr(\rho \cdot A_k)} - 1)A_k^T + \sum_{i=1}^{R}\nabla\lambda_i(\rho). \tag{11}$$

Once the expression for the gradient of the objective function is obtained, gradient descent methods can be applied to optimize and obtain the density matrix for the word.

4 Introducing a phase in computing

A feature of most algorithms that reconstruct in one form or another the density matrix for processing natural language texts is that the developers of these algorithms strive to circumvent the use of the complex number field, thus working in the field of real numbers. However, as you know, if we take quantum probability theory as a probability theory with a quadratic measure, then degeneracy can be avoided when compiling the density matrix only when using the field of complex numbers. An approximate solution in the field of real numbers can be obtained according to the algorithm described above, but from the point of view of increasing the degrees of freedom in training and the accuracy of restoring the initial distribution, the use of complex numbers is a justified step. In this section, we will try to convert the original expressions used in solving the optimization problem to expressions working in the field of complex numbers.
During the transition to the field of complex numbers, the initial vectors representing the basis by which the density matrix is reconstructed can be represented as follows:

$$|\psi_k\rangle = \left[\overline{\psi_{k1}} \cdot e^{i\phi_{k1}} \dots \overline{\psi_{kn}} \cdot e^{i\phi_{kn}}\right] = \left[\overline{\psi_{kn}} \cdot e^{i\phi_{kn}}\right]_{n=1\dots N}, \tag{12}$$

where N is the dimension of the vector.
When receiving a projector on such a vector, we get a matrix of the following form:

$$A_k = |\psi_k\rangle\langle\psi_k| = \begin{bmatrix} \overline{\psi_{k1}}^2 & \overline{\psi_{k1}} \cdot \overline{\psi_{k2}}e^{i(\phi_{k1}-\phi_{k2})} & \dots & \overline{\psi_{k1}} \cdot \overline{\psi_{kn}}e^{i(\phi_{k1}-\phi_{kn})} \\ \overline{\psi_{k1}} \cdot \overline{\psi_{k2}}e^{i(\phi_{k2}-\phi_{k1})} & \overline{\psi_{k2}}^2 & \dots & \overline{\psi_{k2}} \cdot \overline{\psi_{kn}}e^{i(\phi_{k2}-\phi_{kn})} \\ \vdots & \dots & \ddots & \vdots \\ \overline{\psi_{kn}} \cdot \overline{\psi_{k1}}e^{i(\phi_{kn}-\phi_{k1})} & \overline{\psi_{k2}} \cdot \overline{\psi_{kn}}e^{i(\phi_{kn}-\phi_{k2})} & \dots & \overline{\psi_{kn}}^2 \end{bmatrix} \tag{13}$$

And if we shorten the record:

$$A_k = \left[\overline{\psi_{kn}} \cdot \overline{\psi_{km}}e^{i(\phi_{kn}-\phi_{km}}\right], n, m \in [1 \dots N]. \tag{14}$$

After receiving this projector expression, we can substitute it into the main expression to calculate the probability of the observed:

$$Tr(\rho \cdot A_k) = \sum_{m=1}^{N}\sum_{n=1}^{N}\rho_{mn} \cdot A_{knm} = \sum_{m=1}^{N}\sum_{n=1}^{N}\rho_{mn}\overline{\psi_{km}} \cdot \overline{\psi_{kn}}e^{i(\phi_{km}-\phi_{kn})} \tag{15}$$

From this expression, substituting it into the objective function (10), one can obtain expressions of partial gradients for performing gradient decent. Note that with such a statement of the problem, we do not know anything about what phase values have the basis vectors. We will also find them using the gradient decent method, thus fulfilling some semblance of automatic phase adjustment for the basis in the learning process. Therefore, we need to know two expressions for gradient decent. The first expression is the partial derivative of the objective function with respect to the values of the destiny matrix, which we reconstruct. This expression is the first part of the gradient of objective function (11):

$$\begin{aligned} \frac{\partial D_{KL}(\rho||P)}{\partial Tr} &\approx \sum_{k=1}^{L} ln(\frac{\nu_k}{Tr(\rho \cdot A_k)}) \\ \frac{\partial D_{kl}(\rho||P)}{\partial \rho_{nm}} &\approx \sum_{k=1}^{L} A_{kmn} \cdot ln(\frac{\nu_k}{Tr(\rho \cdot A_k)}) \end{aligned} \tag{16}$$

It should be noted that, in fact, the original expression has not changed, only now the values of A_{kmn} are complex numbers. We write the process of deriving the partial derivative by phase:

$$\begin{aligned} \frac{\partial D_{KL}}{\partial \phi_{kn}} = \sum_{m=1}^{N} i \cdot [\rho_{mn} \cdot \overline{\psi_{kn}} \cdot \overline{\psi_{kn}} \cdot e^{i(\phi_{kn}-\phi_{km})} - \\ -\rho_{nm} \cdot \overline{\psi_{km}} \cdot \overline{\psi_{kn}} \cdot e^{i(\phi_{km}-\phi_{kn})}] \end{aligned} . \tag{17}$$

The index m appears due to the fact that the phase ϕ_{kn} appears several times in all expressions for the remaining phases. Next, we can perform the following replacement in order to simplify the recording and knowing the properties of the density matrix:

$$\rho_{mn} = \overline{\rho_{mn} \cdot e^{i\alpha_{nm}}}, \rho_{nm} = \rho_{mn}^{*}, \Delta\phi_{knm} = \phi_{kn} - \phi_{km} \tag{18}$$

After performing this replacement, we can get the following expression:

$$\frac{\partial D_{KL}}{\partial \phi_{kn}} = \sum_{m=1}^{N} i \cdot \overline{\rho_{mn}} \cdot \overline{\psi_{kn}} \cdot \overline{\psi_{km}} \cdot \left[e^{i(\alpha_{mn}+\Delta\phi_{nm})} - e^{-i(\alpha_{mn}+\Delta\phi_{nm})} \right] \tag{19}$$

If we transforming the notation of complex exponentials in parentheses using the Euler formula, we can obtain the final expression for the partial derivative in phase:

$$\frac{\partial D_{KL}}{\partial \phi_{kn}} = -2 \sum_{m=1}^{N} \overline{\rho_{mn}} \cdot \overline{\psi_{kn}} \cdot \overline{\psi_{km}} \cdot sin(\alpha_{mn} + \Delta\phi_{nm}) \tag{20}$$

Thus, using expressions (16) and (20), we can go to the space of complex numbers to search for the density matrix with automatic phase adjustment.

5 Testing of algorithm

To evaluate the algorithm, the well-known WordSim353 package was chosen. This corpus consists of two parts with a total number of example strings equal to 353. Each row of this data set is a pair of English words and a set of 16 ratings of people reflecting the degree of similarity of this pair of words with each other. A couple of words can get a rating from 1 to 10, where 10 means the maximum semantic similarity between words in the opinion of a person. For evaluation needs developed of the algorithm, we used the average value of people's ratings for word pairs as a reference.
To give a comparative evaluation of the developed algorithm, two text data vectorization algorithms were chosen: an algorithm based on the idea of a word bag and tf-idf statistics and the word2vec algorithm (CBoW architecture was used) as the values of the coordinates of the context vectors. For both algorithms, the cosine distance was used as a measure of the proximity of the resulting vector representations:

$$d(W_1, W_2) = \frac{W_1 \cdot W_2}{|W_1|\,|W_2|}, \tag{21}$$

The Tensorflow library was used to build the architecture and teach the Word2Vec model. The degree of closeness of the density matrices was estimated using expression 6. Formally, this expression is not suitable for obtaining the probability value that the words in question belong to the same semantic context, but the expression can still be used as an indicator of the similarity of the two density matrices. In addition, in the current implementation of the semantic tomography algorithm, phase auto-adjustment was not used due to which all the work on constructing density matrices was carried out in the field of real numbers, which could potentially affect the quality of the restoration of distributions over the Kullback-Leibler divergence.
As a training sample for algorithms, a slice of English Wikipedia was chosen. All texts were subjected to the following processing: paragraphs were glued together into one text, punctuation marks were deleted, and all words were first reduced to lowercase and then stamped. All non-alphabetic characters (i.e. numbers, dashes, colons, etc.) were removed from the sample as well as words corresponding to conjunctions and prepositions. For punching, we used Porter [11] from the NLTK library [10] for the Python programming language, which is currently the standard tool for working with texts in natural languages. As a measure for comparing the algorithms with each other, the Pearson correlation coefficient was selected by evaluating the proximity of words by a person and the metric for this model. The algorithm for constructing the matrix representation is given below in the pseudo-code:

Data: Ds - set of documents, C_{max} - maximum number of context clusters, ϵ - threshold for stopping gradient descent, I_{max} - maximum number of iterations of gradient descent
Result: $[\rho_1, \rho_2, \dots, \rho_L]$ - density matrices for size lexicon words L

$Dict \leftarrow build\_dict(Ds)$// building vocabulary of dimension L
$BoWs \leftarrow make\_bag\_of\_words(Ds, Dict)$// for each word generation of multiple word bags
$result \leftarrow []$
// i - word index, ctxs - set bags of words
for $(i, word, ctxs) \leftarrow BoWs$ **do**
  $[ctxs', P] \leftarrow clusterize(ctxs, K_{max})$ // context clustering, P - probability distribution of clusters
  $N \leftarrow |ctxs'|$ // total number of context clusters
  $A \leftarrow [\frac{v_k \cdot v_k^T}{|v_k|} : v_k \in ctxs']$ // projectors on word contexts, v_k - word context
  $\rho_i \leftarrow \frac{1}{N} \sum_{k=1}^{N} A_k$ // initialization of the density matrix for the i-th word
  $loss \leftarrow +\infty$, $\nabla_0 \leftarrow \mathbf{0}$
  for $iter \leftarrow 1 \dots I_{max}$ **do**
    $current\_loss \leftarrow \sum_{k=1}^{N} Tr(\rho_i \cdot A_k) \cdot log\frac{P_k}{Tr(\rho_i \cdot A_k)}$ // loss function calculation $Q(\rho, P)$
    if $|current\_loss - loss| < \epsilon$ **then**
      // if changes $Q(\rho, P)$ are minor, then stop the gradient descent
      break
    end
    $\nabla_{iter} \leftarrow \sum_{k=1}^{N} (log\frac{P_k}{Tr(\rho_i \cdot A_k)} - 1)A_k^T + \sum_{j=1}^{R} \nabla\lambda_j(\rho_i)$ // gradient value calculation $\nabla Q(\rho, P)$ at current point
    $\rho_i \leftarrow adam\_grad(\rho_i, \nabla_{iter-1}, \nabla_{iter})$ // performing adaptive gradient descent step
    $\rho_i \leftarrow norm(\rho_i)$ // matrix trace normalization if necessary
  end
  result.add(ρ_i)
end
return result

The comparison results of the algorithms are shown in table 1:

Table 1: The result of comparing algorithms for a sample of WordSim353

	tf-idf	word2vec (CBow)	QST
Measure of similarity	$cos(w_1, w_2)$	$cos(w_1, w_2)$	$Tr(A\rho)$
Part 1	0.0831	0.1461	0.2213
Part 2	0.1023	0.1644	0.2107

6 Conclusion

The result of comparing algorithms for a sample of WordSim353. This article discusses the analogy between quantum tomography and the process of constructing vector representations of words for text documents. cops. Such an analogy allows we to adapt the mathematical apparatus used to describe the process of quantum tomography, which is used to describe the statistical properties of objects with quantum-like properties. From the point of view of modeling semantics, such an atematic apparatus allows one to take into account the properties of superposition and entanglement of word contexts that occur during vectorization of the analyzed word. In this case, the superposition is considered from the point of view of the dictionary - each individual word in the context is a semantic concept from the dictionary, and the analyzed word is, respectively, in a state of superposition of all words in its context, i.e. in a state of uncertainty about its meaning, expressed through the words of context. As for the analogy with a mixed state, the analyzed word is found in a number of contexts and, from this point of view, the representation of the word in the form of a density matrix, one should take into account the many encountered contexts as a mixture of density matrices. The analogy with entanglement can be carried out through the determination of the presence of correlation of the words encountered in the analyzed contexts. As for further research aimed at improving the presented vectorization algorithm, here we can distinguish a number of improvements, namely:

1. This model does not take into account the frequency characteristics of the words encountered in contexts. For example, there is no automatic filtering of the most garbage words, which can be performed using the tf-idf algorithm. The use of such normalization of context words seems to be a simple and at the same time useful step for cleaning the contexts of words;
2. Density matrices use the word bag model, learning on the contexts of the dimension of the entire lexicon, thus generating matrices of a very large dimension. Further research should be aimed at reducing the dimensionality of these objects and, as a consequence, reducing computational costs.

References

1. Z. Harris. Distributional structure. Word, 10(23):146–162, 1954.
2. N. Chomsky. Three models for the description of language. IRE Transactions on Information Theory, 2:113–124, 1956.

3. Zhang, Y., Rong J., Zhi-Hua Z.: Understanding bag-of-words model. A statistical frame-work. International Journal of Machine Learning and Cybernetics. 1. 43-52. (2010).
4. Jeong, Y., Song, M.: Applying content-based similarity measure to author co-citation analysis. In: Proceedings of Conference 2016. (2016).
5. Haven, E., Khrennikov, A.: Quantum probability and the mathematical modelling of deci-sion-making. Philosophical Transactions. Series A. Mathematical, physical, and engineering science. vol. 374. (2016).
6. Gleason, Andrew M.: Measures on the closed subspaces of a Hilbert space. Indiana Uni-versity Mathematics Journal. 6 885 (1957).
7. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013.
8. C. J. v. Rijsbergen. The Geometry of Information Retrieval. Cambridge University Press, New York, NY, USA, 2004.
9. B. Piwowarski and M. Lalmas. A quantum-based model for interactive information retrieval. In L. Azzopardi, G. Kazai, S. Robertson, S. Rűger, M. Shokouhi, D. Song, and E. Yilmaz, editors, Advances in Information Retrieval Theory, pages 224–231, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
10. E. Loper and S. Bird. Nltk: The natural language toolkit. In In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguis.
11. P. M.F. An algorithm for suffx stripping. 14(3):130–137, Jan 1980.