УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2019 Том 19 № 6 ISSN 2226-1494 http://ntv.itmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTCS
November–December 2019 Vol. 19 No 6 ISSN 2226-1494 http://ntv.itmo.ru/en/
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

УДК 004.912 doi: 10.17586/2226-1494-2019-19-6-1058-1063

КЛАСТЕРИЗАЦИЯ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ СЕМАНТИКО-СИНТАКСИЧЕСКИХ СВЯЗЕЙ СЛОВ

С.В. Лапшин[a,b], И.С. Лебедев[a], А.И. Спивак[a]

[a] Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), Санкт-Петербург, 199178, Российская Федерация
[b] Санкт-Петербургский государственный университет, Санкт-Петербург, 191123, Российская Федерация
Адрес для переписки: sv.lapshin@gmail.com

Информация о статье
Поступила в редакцию 26.08.19, принята к печати 02.09.19
Язык статьи — русский

Ссылка для цитирования: Лапшин С.В., Лебедев И.С., Спивак А.И. Кластеризация текстов с использованием семантико-синтаксических связей слов // Научно-технический вестник информационных технологий, механики и оптики. 2019. Т. 19. № 6. С. 1058–1063. doi: 10.17586/2226-1494-2019-19-6-1058-1063

Аннотация

Предмет исследования. Выполнено исследование метода повышения показателей качества кластеризации текстов на естественном языке. Основное внимание уделено выделению признаков, составляющих математическую модель текстов. Для кластеризации полученного векторного представления текстов использовался метод k-means. **Метод.** Предложенный аналитический подход основан на использовании семантико-синтаксических признаков кластеризируемых текстов. Выделение признаков проведено с помощью Stanford CoreNLP Toolkit. Некоторые связи между словами текстов в представлении «Enhanced++ Dependencies» вместе со связывающими их словами кодируются. На основании частот их встречаемости в текстах вычисляются значения семантико-синтаксических признаков. **Основные результаты.** Эксперимент по сравнению показателей качества прототипа, разработанного на основе предложенного метода, и системы кластеризации на основе статистических признаков, показал, что использование предложенного метода позволяет сократить количество ошибок кластеризации в проведенном эксперименте более чем на 15 %. **Практическая значимость.** Для получения семантико-синтаксических признаков текстов не требуется предобучение. Рассматриваемый подход может быть использован для повышения показателей качества кластеризации в условиях отсутствия больших корпусов текстов, которые необходимы для предобучения статистических моделей языка на основе «word embeddings».

Ключевые слова
кластеризация текстов, семантико-синтаксические признаки, контекст слов, k-means

Благодарности
Работа выполнена по программе фундаментальных исследований РАН по приоритетным направлениям, определяемым президиумом РАН № 2 «Механизмы обеспечения отказоустойчивости современных высокопроизводительных и высоконадежных вычислений».

doi: 10.17586/2226-1494-2019-19-6-1058-1063

TEXT CLUSTERING POWERED BY SEMANTICO-SYNTACTIC FEATURES

S.V. Lapshin[a,b], I.S. Lebedev[a], A.I. Spivak[a]

[a] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation
[b] Saint Petersburg State University, Saint Petersburg, 191123, Russian Federation
Corresponding author: sv.lapshin@gmail.com

Article info
Received 26.08.19, accepted 02.09.19
Article in Russian

For citation: Lapshin S.V., Lebedev I.S., Spivak A.I. Text clustering powered by semantico-syntactic features. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2019, vol. 19, no. 6, pp. 1058–1063 (in Russian). doi: 10.17586/2226-1494-2019-19-6-1058-1063

Abstract

Subject of Research. The performed study is devoted to improvement of the text clustering quality indicators. The main attention is paid to the feature extraction that describes the mathematical model of the texts. The k-means method is used for clustering of the resulting vector representation of the texts. **Method.** An analytical approach was proposed based on the use of semantico-syntactic features of the clustered texts. Feature extraction was performed using the Stanford CoreNLP Toolkit. Some links

between the words of the texts in “Enhanced ++ Dependencies” representation were encoded together with the words connecting them. The values of semantico-syntactic features were calculated based on the frequencies of encoded links in the texts. **Main Results.** An experiment has shown that by comparison of the quality indicators of a prototype developed on the basis of the proposed method and a clustering system based on statistical features, the proposed method application provides for decrease in the number of clustering errors by more than 15 %. **Practical Relevance.** Pre-training is not required to obtain semantico-syntactic features of the texts. Therefore, the proposed approach can be used to improve clustering quality indicators in the absence of large text corpuses, which are necessary for pre-training of statistical language models based on word embeddings.

Keywords
text clustering, semantico-syntactic features, word context, k-means

Acknowledgements
This work has been performed according to the program of fundamental research of the Russian Academy of Sciences in priority areas determined by the Presidium of the Russian Academy of Sciences No. 2 “Mechanisms for ensuring fault tolerance of modern high-performance and highly reliable computing”.

Введение

Кластеризация текстов является важным этапом решения многих прикладных задач в области обработки естественного языка. В частности, кластеризация находит широкое применение при создании рекомендательных систем, определении профилей пользователей [1] и разработке интеллектуальных ассистентов.

Процесс кластеризации текстов, как правило, разделяется на два этапа. На первом этапе происходит выделение признаков из текстов и формирование множества векторов, составляющих математическую модель кластеризируемых текстов. На втором этапе на основании расстояний между векторами происходит выделение кластеров текстов с помощью одного из методов кластеризации: k-means, метод иерархической кластеризации, DBSCAN и других [2].

Методы выделения признаков из текстов в свою очередь разделяются на несколько групп:

1) методы на основе «мешка слов» (например, «bag-of-words», BM25 [3]),

2) методы на основе Latent Semantic Analysis (например, LSA [4], pLSA [5], LDA [6]),

3) методы на основе «word embeddings» (word2vec [7], PV-DM, STC2-LPI [1], BERT [8]),

4) методы на основе семантических признаков [9, 10].

Показатели качества перечисленных методов, очевидно, разнятся в зависимости от характера текстов и объема данных. Однако на основании ряда исследований [1, 2, 11] можно выделить методы кластеризации на основе «word embeddings».

Важной особенностью этих методов является то, что для их применения требуется предобучение. Предобучение осуществляется на текстах, содержащих термины, которые будут использоваться в кластеризируемых текстах для выявления «семантической близости» слов или n-грамм, используемых для построения векторной модели. Так, например, модели для BERT, предоставленные Google, были получены в результате предобучения на текстах Wikipedia совместно с BookCorpus[1].

В некоторых прикладных задачах обработки естественного языка получение таких и даже небольших корпусов не всегда возможно или экономически целесообразно. Например, в задаче кластеризации аннотаций научных публикаций по некоторым узким областям науки нет достаточного количества текстов в принципе. В задаче кластеризации текстовых сообщений пользователей (в рамках разработки чат-ботов) по некоторым тематикам сложности вызывает использование специфических сленговых выражений и сокращений, встречающихся только в разговорной речи и к тому же изменяющихся с течением времени [12]. В этих случаях использование методов на основе «word embeddings» не всегда позволяет решить задачу кластеризации текстов с приемлемыми показателями качества. Поэтому разработка новых методов кластеризации текстов, подходящих для использования в описанных условиях, является актуальной задачей.

Формальная постановка задачи кластеризации текстов

Рассмотрим множество текстов $T = \{t_1 ... t_n\}$.

Векторное представление текстов T задается матрицей $\mathbf{X}$ размерности $n \times m$, такой, что каждая строка матрицы $\boldsymbol{x}_\mathbf{i} = \{x_{i1} ... x_{im}\}$ содержит m признаков i-того текста.

Функция F_1:$F_1(T) = \mathbf{X}$ реализует один из методов выделения признаков из текстов.

Задана функция расстояния между векторами признаков текстов $\rho(x_i, x_j)$.

Пусть $C = \{c_1 ... c_k\}$ — множество кластеров, каждый из которых содержит как минимум один вектор $\mathbf{x}$:$c = \{\mathbf{x}\}$.

Функция F_2:$F_2(\mathbf{X}) = C$ реализует один из методов кластеризации на основании расстояния между векторами признаков текстов $\rho(\mathbf{x}_i, \mathbf{x}_j)$.

Требуется найти такие F_1 и F_2, чтобы для текстов T выполнялось (1), (2).

[1] https://github.com/google-research/bert

$$\sum_{y=1}^{k} \sum_{\mathbf{x_i}, \mathbf{x_j} \in c_y} \rho(\mathbf{x_i}, \mathbf{x_j}) \to \min, \quad (1)$$

$$\sum_{y=1}^{k} \sum_{\mathbf{x_i} \in c_y,\ \mathbf{x_i} \notin c_y} \rho(\mathbf{x_i}, \mathbf{x_j}) \to \max. \quad (2)$$

Иными словами, нужно найти такие функции получения признаков F_1 и кластеризации F_2, чтобы каждый кластер состоял из близких по метрике векторов, а векторы из разных кластеров существенно отличались.

В данном исследовании основное внимание уделено поиску подходящей функции F_1 для кластеризации текстов в условиях, когда методы на основе «word embeddings» не позволяют получить достаточно высокие показатели качества.

Предлагаемый метод

Методы на основе «мешка слов» не требуют предобучения, но их применение на лексически близких текстах не позволяет получить достаточно высокие показатели качества кластеризации. Проблема в том, что статистические признаки, сформированные на основании частот встречаемости терминов или n-грамм, не учитывают контекст применения этих терминов и их семантическую роль в тексте.

Рассмотрим пример из двух названий статей:

Developing convolutional neural networks for computer vision systems;

Development of computer vision systems powered by convolutional neural networks.

Эти предложения лексически очень близки, поэтому при использовании статистической модели на основе «bag of words» они будут отнесены к одному кластеру. Но в рамках задачи кластеризации научных статей ожидается другой результат – первая может быть отнесена к кластеру «Алгоритмы», а вторая — к кластеру «Компьютерное зрение».

Методы на основе «word embeddings» частично решают эту проблему за счет предобучения, в ходе которого формируются векторы, позволяющие учитывать семантическую роль слов, встречающихся в текстах. Но в некоторых прикладных задачах получить такие «embeddings» невозможно в силу отсутствия достаточной предобучающей выборки, поэтому для качественного решения поставленной задачи нужен другой способ получения семантической информации.

Предлагаемый метод основан на том, что признаки, необходимые для различения лексически близких текстов, можно получить, используя семантико-синтаксические связи между словами. Для пояснения этой идеи удобно использовать деревья разбора предложений, приведенных в примере выше.

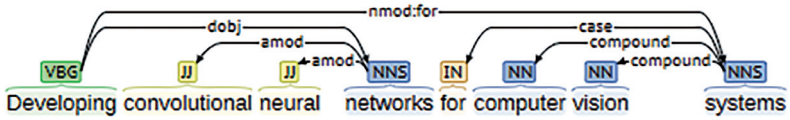

Рис. 1. Дерево семантико-синтаксического разбора первого предложения: dobj, amod, nmod:for, case, compound — типы связей в представлении «Universal Dependencies»[1]; VBG, JJ, NN, NNS, IN — части речи в нотации «The Penn Treebank Project»[2]

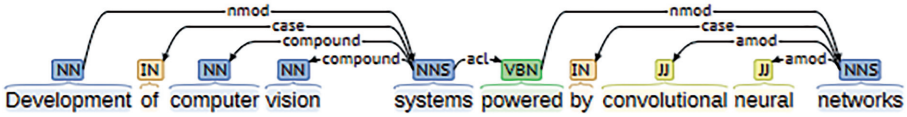

Рис. 2. Дерево семантико-синтаксического разбора второго предложения: dobj, amod, nmod, case, compound, acl — типы связей в представлении «Universal Dependencies»; IN, NN, VBN, JJ, NNS, IN — части речи в нотации «The Penn Treebank Project»

Деревья, изображенные на рис. 1, 2, являются результатом семантико-синтаксического анализа предложений в формате «Universal Dependencies» [13]. В первом предложении фразы «Developing» и «convolutional neural networks» объединены связью типа «dobj» (direct object). Во втором они напрямую вообще не связаны, и слово «Development» сопряжено с «computer vision systems» связью типа «nmod» (nominal modifier). Связь в первом предложении более типична для текстов публикаций, относящихся к кластеру «Алгоритмы», в то время как вторая — для текстов кластера «Компьютерное зрение». Это отличие позволит алгоритму кластеризации разделить такие тексты по разным кластерам.

[1] Расшифровка всех обозначений представлена в документе https://nlp.stanford.edu/software/dependencies_manual.pdf

[2] Расшифровка всех обозначений представлена на сайте https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Для формирования численных признаков текстов, описывающих семантико-синтаксические связи между словами, необходимо ввести следующие понятия.

L — множество семантико-синтаксических связей языка кластеризируемых текстов.

$G = \{V, E\}$ — граф, описывающий результат семантико-синтаксического анализа текста, где V — вершины, E — ребра графа. При этом вершины графа описывают слова и знаки препинания, а ребра описывают тип связи между словами, которые они соединяют.

Таким образом, результатом анализа текста будет множество кортежей (v_i, v_j, l_k), где v_i и v_j — это два связанных слова, а l_k — связь k-того типа между ними.

Для вычисления семантико-синтаксических признаков предлагается произвести следующие действия:

1) каждому кортежу (v_i, v_j, l_k) поставить в соответствие некоторое числовое значение w, уникальное для каждого кортежа;

2) для каждого текста $t_1 \in T$ получить вектор семантико-синтаксических признаков $x_i' = \{x_{i1}' \ldots, x_{iq}'\}$, где q — количество уникальных кортежей;

3) из векторов $\{\mathbf{x_i}\}$ получить матрицу $\mathbf{X'}$ и применить к ней функцию TF-IDF.

В результате будет получена матрица, содержащая семантико-синтаксические признаки кластеризируемых текстов. Эти признаки можно использовать для кластеризации отдельно, но поскольку пары одинаковых слов, объединенных связью одного типа, встречаются в текстах относительно редко, то это подходит только для длинных текстов. Во многих случаях целесообразно использовать семантико-синтаксические признаки *совместно* с признаками, полученными другими методами. Например, с помощью методов на основе «мешка слов». В этом случае вычисляется матрица $\mathbf{X''}$:

$$\mathbf{X''} = [\mathbf{X} \quad \mathbf{X'}] = \begin{bmatrix} x_{11} \cdots x_{1m} x'_{11} \cdots x'_{1q} \\ \vdots \ddots \vdots \quad \vdots \ddots \vdots \\ x_{n1} \cdots x_{nm} x'_{n1} \cdots x'_{nq} \end{bmatrix}, \tag{3}$$

где $\mathbf{X}$ — матрица статистических признаков.

Это позволяет «не проваливаться» на текстах, в которых не удалось выделить достаточное количество семантико-синтаксических признаков для кластеризации.

Экспериментальная реализация предложенного метода

На основе предложенного метода был реализован программный прототип системы кластеризации текстов. Для семантико-синтаксического анализа использовался Stanford CoreNLP Parser [14], позволяющий получать деревья разбора в формате «Universal Dependencies».

Статистические признаки текстов вычислялись с помощью «мешка слов» и объединялись с семантико-синтаксическими в матрицу $\mathbf{X''}$ (3). Для выделения кластеров текстов к полученной матрице применялся метод k-means.

Эксперименты проводились на датасете «20 newsgroups». Классы текстов, которые использовались в каждом эксперименте, описаны в табл. 1.

Таблица 1. Классы текстов, использовавшихся в экспериментах

Номер эксперимента	Классы текстов	Количество текстов	
		в классе	в эксперименте
1	comp.graphics	584	1753
	comp.os.ms-windows.misc	591	
	comp.sys.mac.hardware	578	
2	comp.graphics	584	2343
	comp.os.ms-windows.misc	591	
	comp.sys.mac.hardware	578	
	comp.sys.ibm.pc.hardware	590	
3	comp.graphics	584	2936
	comp.os.ms-windows.misc	591	
	comp.sys.mac.hardware	578	
	comp.sys.ibm.pc.hardware	590	
	comp.windows.x	593	

Для оценки качества разработанного прототипа использовался показатель «точность» (accuracy) отнесения текстов одного класса к одному кластеру. Сравнение разработанного прототипа производилось

с референтной системой кластеризации на основе «bag of words», в которой также применялся k-means. В результате были получены следующие результаты (табл. 2).

Таблица 2. Результаты оценки точности кластеризации текстов

Номер эксперимента	Метод выделения признаков текстов	Количество верно кластеризованных текстов	Количество ошибочно кластеризованных текстов	accuracy
1	BOW[1]	1346	407	0,7679
	BOW + SEM[2]	1370	383	0,7816
2	BOW	1063	1280	0,4537
	BOW + SEM	1178	1165	0,5028
3	BOW	1157	1779	0,3941
	BOW + SEM	1438	1498	0,4898

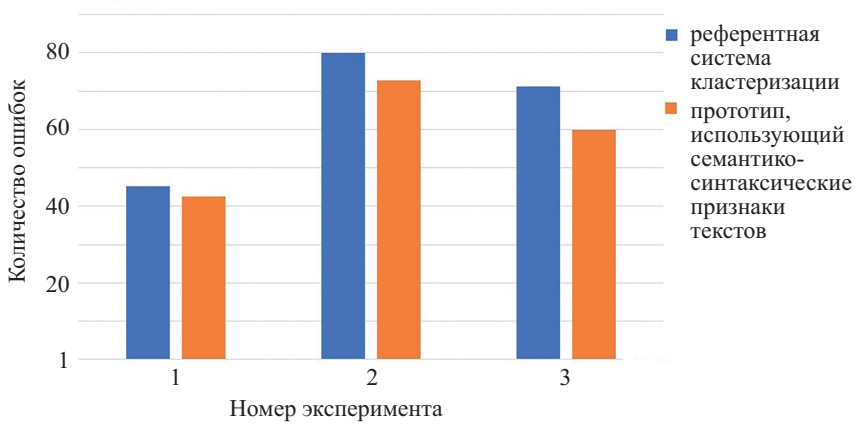

Рис. 3. Количество ошибочно кластеризованных текстов, нормированных по количеству классов

По гистограмме (рис. 3) видно, что применение прототипа позволяет сократить количество ошибок кластеризации текстов во всех экспериментах. При этом в третьем эксперименте он позволил повысить точность более чем на 15 %. Таким образом, можно сделать вывод, что предложенный метод решает поставленную задачу повышения качества кластеризации текстов.

Заключение

В работе предложен метод повышения показателей качества кластеризации текстов за счет использования семантико-синтаксических признаков. Важным отличием от методов на основе «word embeddings» является то, что для его работы не требуется предобучение на данных. Это позволяет использовать его для кластеризации семантически близких текстов при отсутствии предобучающих выборок.

Разработан прототип системы кластеризации текстов, реализующий описанный подход. Проведенные на прототипе эксперименты показали, что использование семантико-синтаксических признаков текстов позволяет поднять точность кластеризации. При этом данные признаки могут использоваться совместно с другими, полученными любым методом, что делает предложенный метод универсальным.

[1] BOW — «мешок слов».

[2] BOW + SEM — объединение признаков, полученных методом «мешка слов» и семантико-синтаксических признаков.

Литература

1. Xu J., Xu B., Wang P., Zheng S., Tian G., Zhao J., Xu B. Self-taught convolutional neural networks for short text clustering // Neural Networks. 2017. V. 88. P. 22–31. doi: 10.1016/j.neunet.2016.12.008
2. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов // Труды ИСП РАН. 2017. Т. 29. № 2. С. 161–200. doi: 10.15514/ISPRAS-2017-29(2)-6
3. Whissell J.S., Clarke C.L.A. Improving document clustering using Okapi BM25 feature weighting // Information Retrieval. 2011. V. 14. N 5. P. 466–487. doi: 10.1007/s10791-011-9163-y
4. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. Indexing by latent semantic analysis // Journal of the American Society for Information Science. 1990. V. 41. N 6. P. 391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
5. Hofmann T. Probabilistic latent semantic indexing // Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999). 1999. P. 50–57. doi: 10.1145/312624.312649
6. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. V. 3. N 4-5. P. 993–1022.
7. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // Proc. 1st International Conference on Learning Representations (ICLR 2013). 2013.
8. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv:1810.04805. 2018.
9. Staab S., Hotho A. Ontology-based text document clustering // Proc. International Intelligent Information Systems/ Intelligent Information Processing and Web Mining Conference (IIS: IIPWM'03). 2003. P. 451–452.
10. Choudhary B., Bhattacharyya P. Text clustering using semantics [Электронный ресурс]. URL: http://vima01220.ethz.ch/CDstore/www2002/poster/79.pdf (дата обращения: 23.10.2019)
11. Liang S., Yilmaz E., Kanoulas E. Collaboratively tracking interests for user clustering in streams of short texts // IEEE Transactions on Knowledge and Data Engineering. 2019. V. 31. N 2. P. 257–272. doi: 10.1109/TKDE.2018.2832211
12. Попова С.В., Данилова В.В. Представление документов в задаче кластеризации аннотаций научных текстов // Научно-технический вестник информационных технологий, механики и оптики. 2014. Т. 19. № 1(89). С. 99–107.
13. Schuster S., Manning C.D. Enhanced english universal dependencies: an improved representation for natural language understanding tasks // Proc. 10th International Conference on Language Resources and Evaluation (LREC 2016). 2016. P. 2371–2378.
14. Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S.J., McClosky D. The Stanford CoreNLP natural language processing toolkit // Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014. P. 55–60. doi: 10.3115/v1/P14-5010

References

1. Xu J., Xu B., Wang P., Zheng S., Tian G., Zhao J., Xu B. Self-taught convolutional neural networks for short text clustering. *Neural Networks*. 2017, vol. 88, pp. 22–31. doi: 10.1016/j.neunet.2016.12.008
2. Parhomenko P.A., Grigorev A.A., Astrakhantsev N.A. A survey and an experimental comparison of methods for text clustering: application to scientific articles. *Proceedings of ISP RAS*, 2017, vol. 29, no. 2, pp. 161–200. (in Russian). doi: 10.15514/ISPRAS-2017-29(2)-6
3. Whissell J.S., Clarke C.L.A. Improving document clustering using Okapi BM25 feature weighting. *Information Retrieval*, 2011, vol. 14, no. 5, pp. 466–487. doi: 10.1007/s10791-011-9163-y
4. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990, vol. 41, no. 6, pp. 391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
5. Hofmann T. Probabilistic latent semantic indexing. *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, 1999, pp. 50–57. doi: 10.1145/312624.312649
6. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, vol. 3, no. 4-5, pp. 993–1022.
7. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. *Proc. 1st International Conference on Learning Representations (ICLR 2013)*, 2013.
8. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*. 2018.
9. Staab S., Hotho A. Ontology-based text document clustering. *Proc. International Intelligent Information Systems/ Intelligent Information Processing and Web Mining Conference (IIS: IIPWM'03)*, 2003, pp. 451–452.
10. Choudhary B., Bhattacharyya P. *Text clustering using semantics*. Available at: http://vima01220.ethz.ch/CDstore/www2002/poster/79.pdf (accessed: 23.10.2019)
11. Liang S., Yilmaz E., Kanoulas E. Collaboratively tracking interests for user clustering in streams of short texts. *IEEE Transactions on Knowledge and Data Engineering*, 2019, vol. 31, no. 2, pp. 257–272. doi: 10.1109/TKDE.2018.2832211
12. Popova S., Danilova V. Document representation for clustering of scientific abstracts. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2014, vol. 19, no. 1(89), pp. 99–107. (in Russian)
13. Schuster S., Manning C.D. Enhanced english universal dependencies: an improved representation for natural language understanding tasks. *Proc. 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 2371–2378.
14. Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S.J., McClosky D. The Stanford CoreNLP natural language processing toolkit. *Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60. doi: 10.3115/v1/P14-5010

Авторы

Лапшин Сергей Владимирович — кандидат технических наук, научный сотрудник, Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), Санкт-Петербург, 199178, Российская Федерация, доцент, Санкт-Петербургский государственный университет, Санкт-Петербург, 191123, Российская Федерация, ORCID ID: 0000-0001-7102-4702, sv.lapshin@gmail.com

Лебедев Илья Сергеевич — доктор технических наук, профессор, заведующий лабораторией, Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), Санкт-Петербург, 199178, Российская Федерация, Scopus ID: 56321781100, ORCID ID: 0000-0001-6753-2181, lebedev@cit.ifmo.ru

Антон Игоревич Спивак — кандидат технических наук, заведующий лабораторией, Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН), Санкт-Петербург, 199178, Российская Федерация, ORCID ID: 0000-0002-6981-8754, anton.spivak@gmail.com

Authors

Sergei V. Lapshin — PhD, Scientific Researcher, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation; Associate Professor, Saint Petersburg State University, Saint Petersburg, 191123, Russian Federation, ORCID ID: 0000-0001-7102-4702, sv.lapshin@gmail.com

Ilya S. Lebedev — D.Sc., Professor, Laboratory Head, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation, Scopus ID: 56321781100, ORCID ID: 0000-0001-6753-2181, lebedev@cit.ifmo.ru

Anton I. Spivak — PhD, Laboratory Head, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation, ORCID ID: 0000-0002-6981-8754, anton.spivak@gmail.com