УНИВЕРСИТЕТ ИТМО
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2022 Vol. 22 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2022 Том 22 Номер 1

doi: 10.17586/2226-1494-2022-22-1-114-119
УДК 004.912

Исследование способов векторизации неструктурируемых текстовых документов на естественном языке по степени их влияния на качество работы различных классификаторов

Виктор Викторович Шадский[1]✉, Александр Борисович Сизоненко[2], Максим Алексеевич Чекмарев[3], Андрей Васильевич Шишков[4], Даниил Андреевич Исакин[5]

[1,2,3,4] Краснодарское высшее военное училище им. С.М. Штеменко, Краснодар, 350063, Российская Федерация
[5] Новосибирский государственный технический университет, Новосибирск, 630087, Российская Федерация

[1] vdvryazan57@yandex.ru✉, https://orcid.org/0000-0002-9221-2283
[2] siz_al@mail.ru, https://orcid.org/0000-0001-8201-9159
[3] max.chek13@gmail.com, https://orcid.org/0000-0002-6832-9991
[4] shishkov-andrey00@mail.ru, https://orcid.org/0000-0002-1841-8750
[5] pm11.isakin@gmail.com, https://orcid.org/0000-0001-7307-6258

Аннотация

Предмет исследования. Повсеместное увеличение объемов обрабатываемой информации на объектах критической информационной инфраструктуры, представленной в текстовой форме на естественном языке, создает проблему ее классификации по степени конфиденциальности. Успех решения данной задачи зависит как от самой модели-классификатора, так и от выбранного способа извлечения признаков (векторизации). Требуется максимально полно передать модели-классификатору свойства исходного текста, содержащие всю совокупность демаркационных признаков. В работе представлена эмпирическая оценка эффективности алгоритмов линейной классификации, основанная на выбранном способе векторизации, а также значении количества настраиваемых параметров в случае применения векторизатора хеширования (Hash Vectorizer). **Метод.** В качестве датасета для обучения и тестирования алгоритмов классификации использованы государственные текстовые документы, условно выступающие в роли конфиденциальных. Выбор подобного текстового массива обусловлен наличием специфической терминологии, повсеместно встречающейся в рассекреченных документах. Терминированность, являясь примитивной демаркационной границей и выступая в роли классификационного признака, облегчает работу алгоритмов классификации, что в свою очередь позволяет сконцентрировать внимание на той доли вклада, которую вносит выбранный способ векторизации. Метрикой оценки качества работы алгоритмов выступает величина ошибки классификации. За величину ошибки принята величина, обратная доле правильных ответов алгоритма (accuracy). Проведена оценка алгоритмов по времени обучения. **Основные результаты.** Полученные гистограммы отражают величину ошибки алгоритмов и время обучения. Выделены наиболее и наименее эффективные алгоритмы для конкретно заданного способа векторизации. **Практическая значимость.** Результаты работы позволяют повысить эффективность решения реальных практических классификационных задач текстовых документов небольшого объема со свойственной специфической терминологией.

Ключевые слова

способ векторизации, TF-IDF, Hash Vectorizer, алгоритм классификации, accuracy

Благодарности

Работа выполнена в Краснодарском высшем военном училище им. С.М. Штеменко в рамках диссертационного исследования в области обработки естественного языка.

Ссылка для цитирования: Шадский В.В., Сизоненко А.Б., Чекмарев М.А., Шишков А.В., Исакин Д.А. Исследование способов векторизации неструктурируемых текстовых документов на естественном языке по степени их влияния на качество работы различных классификаторов // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22, № 1. С. 114–119. doi: 10.17586/2226-1494-2022-22-1-114-119

A study of vectorization methods for unstructured text documents in natural language according to their influence on the quality of work of various classifiers

Viktor V. Shadsky[1]✉, Alexander B. Sizonenko[2], Maxim A. Chekmarev[3], Andrey V. Shishkov[4], Daniil A. Isakin[5]

[1,2,3,4] Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation
[5] Novosibirsk State Technical University, Novosibirsk, 630087, Russian Federation

[1] vdvryazan57@yandex.ru✉, https://orcid.org/0000-0002-9221-2283
[2] siz_al@mail.ru, https://orcid.org/0000-0001-8201-9159
[3] max.chek13@gmail.com, https://orcid.org/0000-0002-6832-9991
[4] shishkov-andrey00@mail.ru, https://orcid.org/0000-0002-1841-8750
[5] pm11.isakin@gmail.com, https://orcid.org/0000-0001-7307-6258

Abstract
The widespread increase in the volume of processed information at the objects of critical information infrastructure, presented in text form in natural language, causes a problem of its classification by the degree of confidentiality. The success of solving this problem depends both on the classifier model itself and on the chosen method of feature extraction (vectorization). It is required to transfer to the classifier model the properties of the source text containing the entire set of demarcation features as fully as possible. The paper presents an empirical assessment of the effectiveness of linear classification algorithms based on the chosen method of vectorization, as well as the number of configurable parameters in the case of the Hash Vectorizer. State text documents are used as a dataset for training and testing classification algorithms, conditionally acting as confidential. The choice of such a text array is due to the presence of specific terminology found everywhere in declassified documents. Termination, being a primitive demarcation boundary and acting as a classification feature, facilitates the work of classification algorithms, which in turn allows one to focus on the share of the contribution that the chosen method of vectorization makes. The metric for evaluating the quality of algorithms is the magnitude of the classification error. The magnitude of the error is the inverse of the proportion of correct answers of the algorithm (accuracy). The algorithms were evaluated according to the training time. The resulting histograms reflect the magnitude of the error of the algorithms and the training time. The most and least effective algorithms for a given vectorization method are identified. The results of the work make it possible to increase the efficiency of solving real practical classification problems of small-volume text documents characterized by their specific terminology.

Keywords
vectorization method, TF-IDF, Hash Vectorizer, classification algorithm, accuracy

Acknowledgements
The work was carried out at the Krasnodar Higher Military School as part of a dissertation in the field of natural language processing.

For citation: Shadsky V.V., Sizonenko A.B., Chekmarev M.A., Shishkov A.V., Isakin D.A. A study of vectorization methods for unstructured text documents in natural language according to their influence on the quality of work of various classifiers. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 1, pp. 114–119 (in Russian). doi: 10.17586/2226-1494-2022-22-1-114-119

Введение

Увеличение объемов обрабатываемой информации на объектах критической информационной инфраструктуры, в том числе представленной в текстовой форме на естественном языке, непременно влечет за собой целый ряд трудностей, связанных, в первую очередь, с ее обработкой. Одна из таких задач обработки – задача классификации документов [1].

Электронные текстовые документы на естественном языке, независимо от степени конфиденциальности, имеют достаточно широкий спектр классификационных признаков, начиная от тематической составляющей, степени формализации и шаблонизации, и заканчивая тональностью и стилем изложения [2, 3]. Из данного факта следует вывод: для каждой отдельно взятой задачи классификации и соответствующего ей датасета существует детерминированный и не всегда очевидный метод ее решения. Таким образом, перед решением классификационной задачи необходимо провести ее детальный анализ с целью выделения наиболее важных демаркационных признаков и максимального сужения «диапазона» возможных методов ее решения [4].

Например, выберем для составления датасета материалы рассекреченных документов о положении дел в стране[1], выступающие в роли конфиденциальных, и открытых нормативно-правовых актов[2]. Документы представлены в текстовой форме на естественном языке. Отметим, что для более качественной их демаркации необходимо обратить внимание на такой классификационный признак как уровень терминированности [5], который у рассекреченных документов носит ярко выраженный и специфический характер [6, 7].

Для успешного решения задачи классификации указанных документов достаточно использовать классификатор, способный разделить пространство признаков обычной разделяющей гиперплоскостью [8]. В связи с этим в настоящей работе приведены результаты работы простейших алгоритмов классификации [9].

[1] Рассекреченные источники // Исторические материалы [Электронный ресурс]. URL: https://istmat.info/node (дата обращения: 20.12.2021).

[2] Указы и распоряжения Президента Российской Федерации // Кодификация РФ [Электронный ресурс]. URL: https://rulaws.ru/president (дата обращения: 20.12.2021).

Один из ключевых факторов, определяющих успешность решения классификационной задачи — выбор наилучшего способа представления знаний [10], содержащихся в текстовых документах, при котором максимально сохраняются свойства исходного текста и содержащиеся в нем закономерности [9], что в свою очередь отражается на результатах работы моделей-классификаторов.

Выбор способа векторизации

Самые применяемые в настоящее время способы векторизации текста: TF-IDF (Term Frequency-Inverse Document Frequency) и Hash Vectorizer [11]. Выполним сравнение данных способов по величине ошибки, а также времени обучения алгоритмов классификации.

Векторизатор TF-IDF часто используется для создания векторов слов из набора документов. Он масштабирует частоту термов, придавая меньшую значимость наиболее часто встречаемым из них [12]. Однако в таком случае заранее требуется наличие полного словаря, что для большинства практических задач невозможно ввиду периодического его пополнения. Вследствие чего постоянный пересчет векторов TF-IDF крайне неэффективен в вычислительном плане, учитывая при этом широкую применимость векторизатора на больших объемах корпусов[1].

Для словаря с динамически изменяющимся количеством термов более эффективен векторизатор Hash

[1] [Электронный ресурс]. https://habr.com/ru/post/515036/, https://habr.com/ru/post/515084 (дата обращения: 20.12.2021).

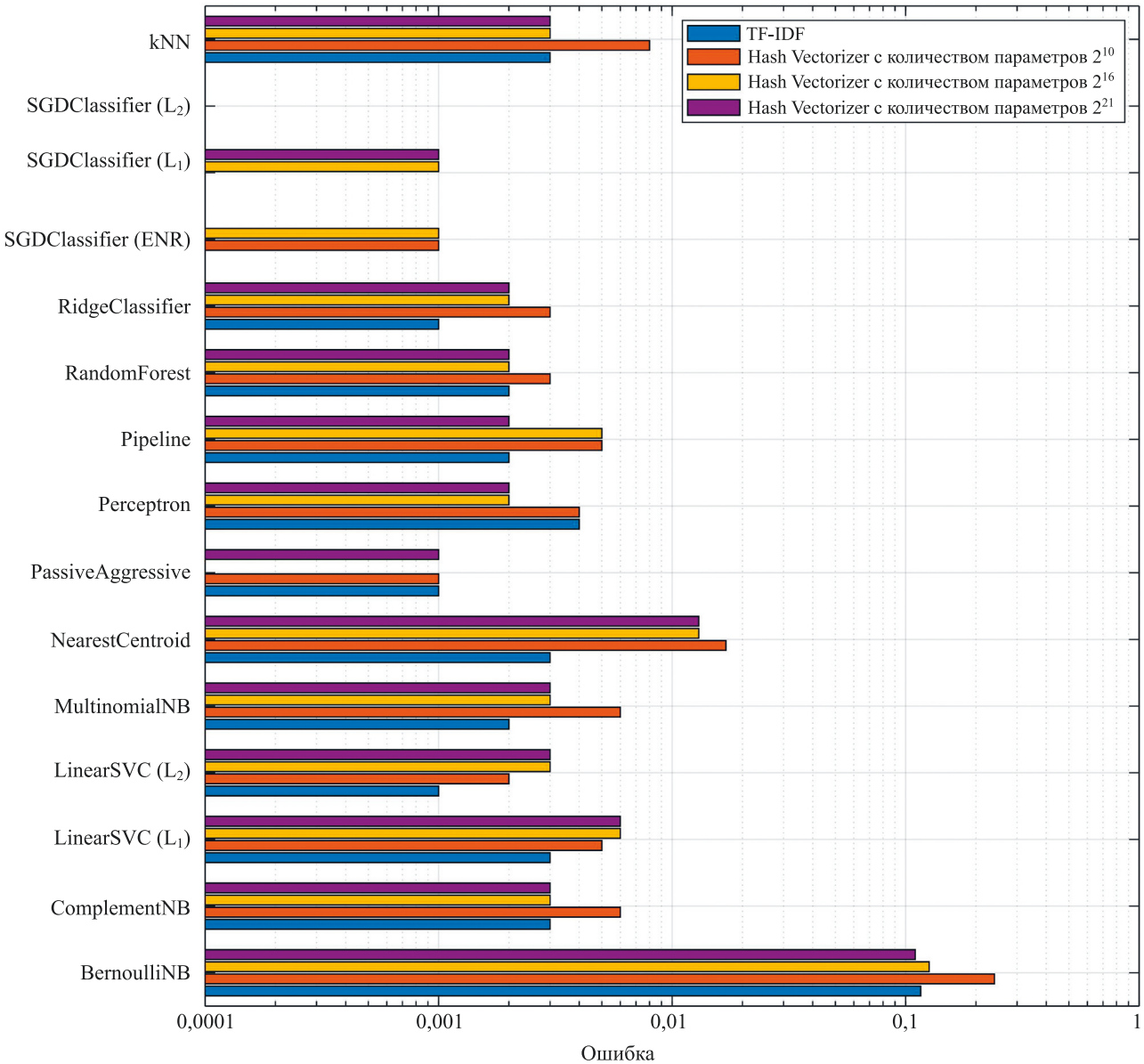

Рис. 1. Величина ошибки алгоритмов классификации при использовании векторизаторов TF-IDF и Hash Vectorizer с различным числом параметров

Fig. 1. The magnitude of the error of classification algorithms when using TF-IDF and Hash Vectorizer with different number of parameters

Vectorizer, популярность которого для большинства практических задач все чаще возрастает. Используя 32-битную функцию MurmurHash3[1], разработанную Остином Эпплби, этот векторизатор позволяет проводить непрерывное обучение по мере доступности данных, не уточняя заранее объем словаря [13].

Результаты работы классификаторов

Ввиду того, что задача классификации в данной работе носит достаточно примитивный характер, большинство алгоритмов классификации показывают практически идеальные результаты по показателю accuracy. По этой причине, для более наглядного представления результатов работы и качественной их оценки, перейдем к рассмотрению величины ошибки, являющейся обратной величиной для доли правильных ответов алгоритма (accuracy). Гистограммы, представленные на рис. 1 в логарифмическом масштабе, показывают, что для векторизатора Hash Vectorizer при увеличении количества параметров с 2^{10} до 2^{21} прослеживается тенденция к уменьшению величины ошибки.

Переходя к рассмотрению такого показателя, как время обучения, подчеркнем, что в зависимости от выбранного способа векторизации данная величина варьируется в широких пределах и увеличивается с ростом количества параметров Hash Vectorizer (рис. 2). Для более наглядного представления шкала, отображающая показатель времени, представлена в логарифмическом масштабе.

[1] MurmurHash [Электронный ресурс]. URL: https://www.sites.google.com/site/murmurhash (дата обращения: 20.12.2021).

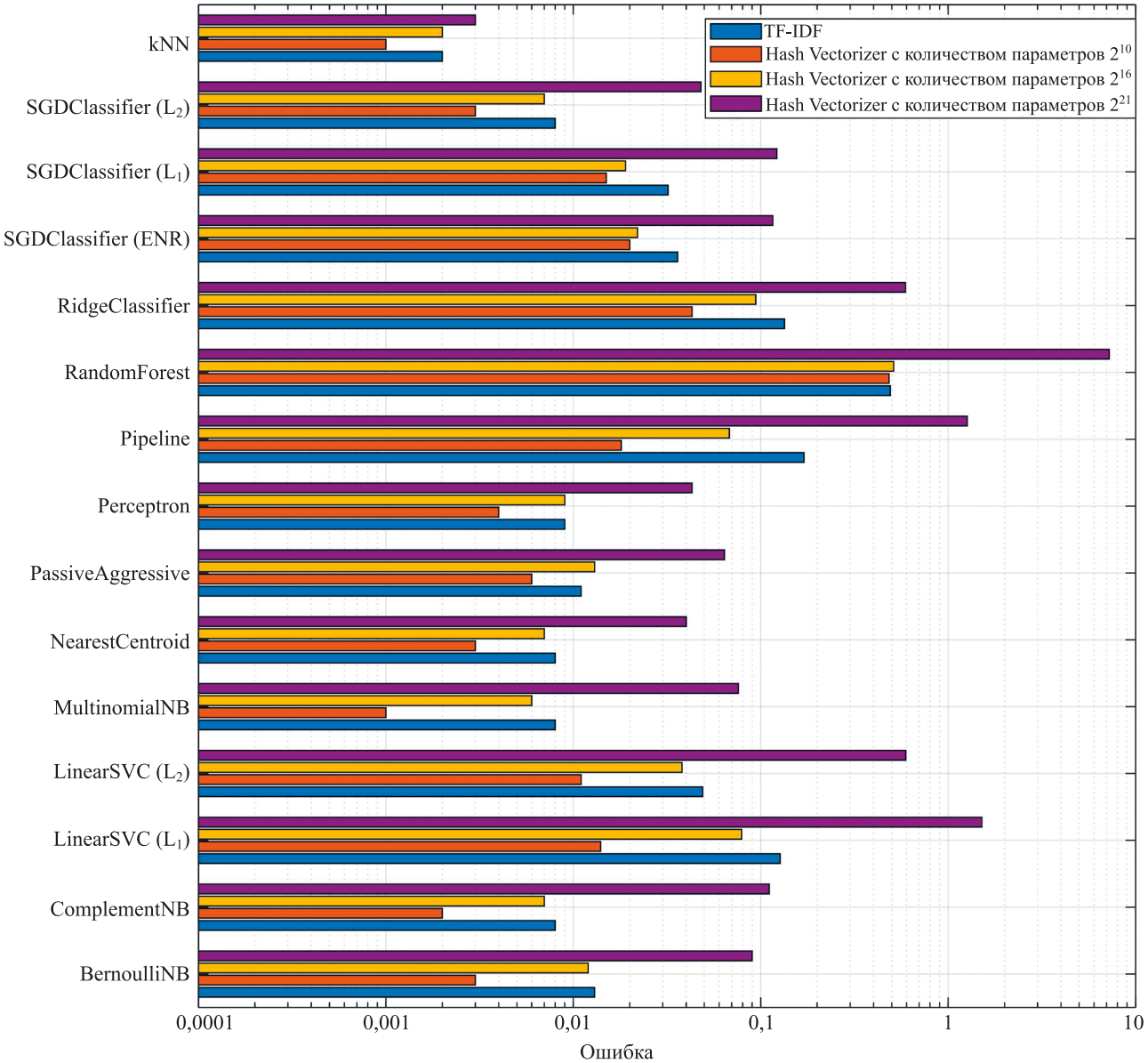

Рис. 2. Время обучения алгоритмов классификации при использовании векторизаторов TF-IDF и Hash Vectorizer с различным числом параметров

Fig. 2. Training time of classification algorithms when using TF-IDF and Hash Vectorizer with different number of parameters

Более детальное сравнение рассмотренных алгоритмов классификации с результатами их работы, отражающими время обучения, тестирования и долю правильных ответов алгоритма размещено в документе[1].

Отметим, что при увеличении числа настраиваемых параметров векторизатора Hash Vectorizer с 2^{10} до 2^{21} алгоритмы классификации показывают лучшие результаты, чем при использовании TF-IDF векторизатора. При этом значительно увеличиваются временные затраты на обучение модели, что связано с увеличением размерности целочисленных индексов, присваиваемых соответствующим токенам словаря. Это приводит к уменьшению числа коллизий [14, 15] и значительному увеличению времени обучения.

Оценивая эффективность алгоритмов путем отношения доли правильных ответов к времени обучения (accuracy/train-time) видно, что для всех исследуемых способов векторизации самый лучший алгоритм классификации — kNN, показавший лучшие результаты по доли правильных ответов и времени обучения. Наиболее высокие результаты (0,992/0,001) данный алгоритм показал при использовании векторизатора Hash Vectorizer с количеством параметров 2^{10}.

Наихудший результат показал алгоритм классификации Random Forest Classifier с результатами (0,998/7,239) для случая применения векторизатора Hash Vectorizer с количеством параметров 2^{21}.

Заключение

В работе выполнена сравнительная оценка качества работы алгоритмов линейной классификации в зависимости от выбранного способа векторизации (представления) исходного текста.

Несмотря на очевидное преимущество векторизатора Hash Vectorizer по сравнению с TF-IDF применительно к документам небольшого объема, проявляющееся в увеличении доли правильных ответов алгоритмов и, соответственно, уменьшении величины ошибки, показатель времени обучения в общей тенденции также возрастает.

Полученный эмпирический опыт по выявлению соответствия величин показателей ошибки и времени обучения алгоритмов классификации количеству настраиваемых параметров способа векторизации Hash Vectorizer позволит существенно повысить качество решения подобных практических классификационных задач путем сужения диапазона выбора настраиваемых параметров моделей-классификаторов.

[1] [Электронный ресурс]. https://docs.google.com/document/d/1pSdWQ8E7H-CBeuTZPxoQPf-yFsLWiXxy/edit (дата обращения: 20.12.2021).

Литература

1. Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы. 2017. № 1. С. 85–99. https://doi.org/10.15827/0236-235X.030.1.085-099
2. Бортников В.И., Михайлова Ю.Н. Документная лингвистика: учебно-методическое пособие / Министерство образования и науки Российской Федерации, Уральский государственный юридический университет. Екатеринбург: Изд-во Уральского университета, 2017. 132 с.
3. Роготнева Е.Н. Документная лингвистика: сборник учебно-методических материалов. Томск: Изд-во Томского политехнического университета, 2011. 784 с.
4. Орлов А.И. Математические методы теории классификации // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 95. С. 23–45.
5. Косова М.В., Шарипова Р.Р. Терминированность как основа классификации документных текстов // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2016. Т. 15. № 4. С. 245–252. https://doi.org/10.15688/jvolsu2.2016.4.26
6. Терских Н.В. Термин как единица специального знания // Система ценностей современного общества. 2008. № 3. С. 97–104.
7. Розенталь Д.Э., Голуб И.Б., Теленкова М.А. Современный русский язык. 13-е изд. М.: АЙРИС-пресс, 2014. 448 с.
8. Крашенинников А.М., Гданский Н.И., Рысин М.Л. Линейная классификация объектов с использованием нормальных гиперплоскостей // Инженерный вестник Дона. 2012. № 4-1 (22). С. 94–99.
9. Dan Nelson. Overview of Classification Methods in Python with Scikit-Learn // Stack Abuse [Электронный ресурс]. URL: https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/ (дата обращения: 20.12.2021).
10. Woods W. Important issues in knowledge representation // Proceedings of the IEEE. 1986. V. 74. N 10. P. 1322–1334. https://doi.org/10.1109/PROC.1986.13634
11. Рашка С., Мирджалили В. Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и

References

1. Batura T.V. Automatic text classification methods. *Software & Systems*, 2017, no. 1, pp. 85–99. (in Russian). https://doi.org/10.15827/0236-235X.030.1.085-099
2. Bortnikov V.I., Mikhailova Iu.N. *Documentary Linguistics*. Ekaterinburg, Izdatel'stvo Ural'skogo universiteta Publ., 2017, 132 c. (in Russian)
3. Rogotneva E.N. *Documentary Linguistics*. Teaching materials. Tomsk, Tomsk Polytechnic University Publ., 2011, 784 c. (in Russian)
4. Orlov A.I. Mathematical methods of classification theory. *Polythematic online scientific journal of Kuban State Agrarian University*, 2014, no. 95, pp. 23–45. (in Russian)
5. Kosova M.V., Sharipova R.R. Termination as the basis for classification of document texts. *Science Journal of Volgograd State University. Linguistics*, 2016, vol. 15, no. 4, pp. 245–252. (in Russian). https://doi.org/10.15688/jvolsu2.2016.4.26
6. Terskikh N.V. Term as a unit of specialized knowledge . *Sistema cennostej sovremennogo obshhestva*, 2008, no. 3, pp. 97–104. (in Russian)
7. Rozental D.E., Golub I.B., Telenkova M.A. *Contemporary Russian Language*. Moscow, AJRIS-press Publ., 2014, 448 p. (in Russian)
8. Krasheninnikov A.M., Gdanskiy N.I., Rysin M.L. Linear classification of objects using normal hyperplanes. *Engineering journal of Don*, 2012, no. 4-1 (22), pp. 94–99. (in Russian)
9. Dan Nelson. Overview of Classification Methods in Python with Scikit-Learn. *Stack Abuse*. Available at: https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/ (accessed: 20.12.2021).
10. Woods W. Important issues in knowledge representation. *Proceedings of the IEEE*, 1986, vol. 74, no. 10, pp. 1322–1334. https://doi.org/10.1109/PROC.1986.13634
11. Raschka S., Mirjalili V. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019, 770 p.
12. Qaiser S., Ali R. Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 2018, vol. 181, no. 1, pp. 25–29. https://doi.org/10.5120/ijca2018917395

TensorFlow 2 / пер. с англ. – 3-е изд. СПб.: ООО «Диалектика», 2020. 848 с.
12. Qaiser S., Ali R. Text mining: Use of TF-IDF to examine the relevance of words to documents // International Journal of Computer Applications. 2018. V. 181. N 1. P. 25–29. https://doi.org/10.5120/ijca2018917395
13. Kavita Ganesan. HashingVectorizer vs. CountVectorizer [Электронный ресурс]. URL: https://kavita-ganesan.com/hashingvectorizer-vs-countvectorizer/#.YcGOyavP2Ul (дата обращения: 20.12.2021).
14. Jason Brownlee. How to Encode Text Data for Machine Learning with scikit-learn // Machine learning mastery [Электронный ресурс]. URL: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/ (дата обращения: 20.12.2021).
15. Max Pagels. Introducing One of the Best Hacks in Machine Learning: the Hashing Trick // Medium [Электронный ресурс]. URL: https://medium.com/value-stream-design/introducing-one-of-the-best-hacks-in-machine-learning-the-hashing-trick-bf6a9c8af18f (дата обращения: 20.12.2021).

13. Kavita Ganesan. *HashingVectorizer vs. CountVectorizer*. Available at: https://kavita-ganesan.com/hashingvectorizer-vs-countvectorizer/#.YcGOyavP2Ul (accessed: 20.12.2021).
14. Jason Brownlee. How to Encode Text Data for Machine Learning with scikit-learn. *Machine learning mastery*. Available at: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/ (accessed: 20.12.2021).
15. Max Pagels. Introducing One of the Best Hacks in Machine Learning: the Hashing Trick. *Medium*. Available at: https://medium.com/value-stream-design/introducing-one-of-the-best-hacks-in-machine-learning-the-hashing-trick-bf6a9c8af18f (accessed: 20.12.2021).

Авторы

Шадский Виктор Викторович — адъюнкт, Краснодарское высшее военное училище им. С.М. Штеменко, Краснодар, 350063, Российская Федерация, https://orcid.org/0000-0002-9221-2283, vdvryazan57@yandex.ru

Сизоненко Александр Борисович — доктор технических наук, доцент, начальник кафедры, Краснодарское высшее военное училище им. С.М. Штеменко, https://orcid.org/0000-0001-8201-9159, siz_al@mail.ru

Чекмарев Максим Алексеевич — адъюнкт, Краснодарское высшее военное училище им. С.М. Штеменко, Краснодар, 350063, Российская Федерация, https://orcid.org/0000-0002-6832-9991, max.chek13@gmail.com

Шишков Андрей Васильевич — студент, Краснодарское высшее военное училище им. С.М. Штеменко, Краснодар, 350063, Российская Федерация, https://orcid.org/0000-0002-1841-8750, shishkov-andrey00@mail.ru

Исакин Даниил Андреевич — студент, Новосибирский государственный технический университет, Новосибирск, 630087, Российская Федерация, https://orcid.org/0000-0001-7307-6258, pm11.isakin@gmail.com

Authors

Viktor V. Shadsky — PhD Student, Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation, https://orcid.org/0000-0002-9221-2283, vdvryazan57@yandex.ru

Alexander B. Sizonenko — D.Sc., Associate Professor, Head of Department, Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation, https://orcid.org/0000-0001-8201-9159, siz_al@mail.ru

Maxim A. Chekmarev — PhD Student, Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation, https://orcid.org/0000-0002-6832-9991, max.chek13@gmail.com

Andrey V. Shishkov — Student, Krasnodar Higher Military School, Krasnodar, 350063, Russian Federation, https://orcid.org/0000-0002-1841-8750, shishkov-andrey00@mail.ru

Daniil A. Isakin — Student, Novosibirsk State Technical University, Novosibirsk, 630087, Russian Federation, https://orcid.org/0000-0001-7307-6258, pm11.isakin@gmail.com

Статья поступила в редакцию 06.12.2021
Одобрена после рецензирования 29.12.2021
Принята к печати 29.01.2022

Received 06.12.2021
Approved after reviewing 29.12.2021
Accepted 29.01.2022