Научная статья

УДК 004.855.5
DOI 10.25205/1818-7900-2022-20-2-27-36

Классификация научных текстов по специальностям методами машинного обучения

Беруз Бурхонович Иномов[1], Марина Тропманн-Фрик[2]

[1]Политехнический институт Таджикского технического университета им. акад. М. С. Осими
Худжанд, Республика Таджикистан

[2]Гамбургский университет прикладных наук (HAW Hamburg)
Гамбург, Германия

[1]behruzinomov@gmail.com
[2]marina.tropmann-frick@haw-hamburg.de

Аннотация

Данная статья исследует экспериментальную задачу проблемы классификации научных текстовых материалов на основе методов машинного и глубокого обучения (Machine Learning & Deep Learning). Для решения задачи предложен метод классификации текстов, учитывающий предобработку и специфику научных текстовых материалов, позволяющий при использовании алгоритмов ML, повысить точность и быстродействие классификации текстов. Проведено исследование методов индексации и классификации по специальностям для базы научных текстовых материалов. Рассмотрены оценки качества алгоритмов ML и получены результаты сравнений классификации диссертационных работ по специальностям методами машинного обучения в рамках существующей обучающей выборки научных материалов.

Ключевые слова

классификация, диссертационная работа, Logistic Regression, SVM, SGD, MLP, Scikit-Learn

Для цитирования

Иномов Б. Б., Тропманн-Фрик М. Классификация научных текстов по специальностям методами машинного обучения // Вестник НГУ. Серия: Информационные технологии. 2022. Т. 20, № 2. С. 27–36. DOI 10.25205/1818-7900-2022-20-2-27-36

Scientific Texts Classification by Speciality with Machine Learning Methods

Beruz B. Inomov[1], Marina Tropmann-Frick[2]

[1]Polytechnic Institute of the Tajik Technical University named after academician M. S. Osimi
Khujand, Republic of Tajikistan

[2]Hamburg University of Applied Sciences (HAW Hamburg)
Hamburg, Germany

[1]behruzinomov@gmail.com
[2]marina.tropmann-frick@haw-hamburg.de

Abstract

This article investigates the problem of experimental study classification problem of scientific text materials by utilizing the methods of Machine Learning and Deep Learning. The experimental study based on text classification method which proposed preprocessing and specificity of scientific text materials by using the ML algorithms to improve accuracy and speed of text classification was conducted. The analysis of indexation and classification methods by specialties was conducted for a set of scientific text materials. The evaluation and comparison of ML algorithms' quality was considered, and the results of dissertational works' classification by machine learning methods within the framework of the existing training set of scientific materials were obtained.

Keywords
data classification, thesis work specialties, Logistic Regression, SVM, SGD, MLP, Scikit-Learn

For citation
Inomov1 B. B., Tropmann-Frick M. Scientific Texts Classification by Speciality with Machine Learning Methods. Vestnik NSU. Series: In-formation Technologies, 2022, vol. 20, no. 2, p. 27–36. (in Russ.) DOI 10.25205/1818-7900-2022-20-2-27-36

Введение

Правильная организация работы с документами в наши дни имеет большое значение, так как от эффективности реализации документооборота напрямую зависит эффективность работы любой организации. С увеличением количества информации в сфере образования и с развитием методов обучения у студентов и преподавателей расширилась возможность обмениваться информацией. Чаще всего эта информация относится к различным научным темам, предметным областям и специальностям. Такими типами информации могут быть электронные документы. Например: силлабусы, лекции, презентации, тесты, книги, статьи, диссертации и другие.

Выявление смысла является основной задачей анализа содержимого текстов. В данной статье в качестве обучающей и тестовой выборки текстов были отобраны документы: авторефераты и диссертации, которые относятся к определенным предметным областям и специальностям, чтобы автоматически классифицировать подобные документы на соответствующие группы.

Сопутствующие работы

Для модели классификации доступно много методов, таких как дерево решений, машина опорных векторов, k-ближайших соседей, наивный байесовский метод, случайный лес, логистическая регрессия, нейронные сеты и т. д. Однако эффективность этих методов зависит от конкретной решаемой задачи и обучающей выборки.

В статье [1] проведен сравнительный анализ методами k ближайших соседей (knn: k-nearest neighbor) и логистической регрессией для классификации научных текстовых материалов. В результате исследования метод логистической регрессии показал хорошую точность, чем k ближайших соседей.

Также проведено исследование [2] для оценки эффективности алгоритмов дерево решений и случайный лес для классификации научных текстовых материалов. В результате исследования метод Random Forest показал хорошую точность, чем Decision tree, так как для тренировки модели использовался больше, чем n_estimator> 1.

Некоторыми авторами [3] описана разработка системы классификации тестов по научным специальностям. Для формирования модели машинного обучения применена многоклассовая логистическая регрессия. Недостаток в том, что автор использовал только одну модель, но реализовал результат модели в виде веб-приложения.

Также можно встретить работы, описывающие исследования различных статистических методов классификации научных текстов [4]. В данной статье применяется метод опорных векторов. Будем сравнивать данный метод для одной и той же задачи так, как и были проведены исследования в статьях [1; 2].

Постановка задачи

Задана обучающая выборка: $\boldsymbol{X = \{(D_1, S_1), (D_2, S_2), \ldots (D_n, S_n)\}}$ и $\boldsymbol{X^* = \{(D_1, C_1), (D_2, C_2), \ldots (D_n, C_n)\}}$, где $\boldsymbol{D_i}$ – список диссертационных исследований, $\boldsymbol{S_i}$ – список наименований специальностей и $\boldsymbol{C_i}$ – список наименований сфер специальностей. Требуется построить алгоритм $\boldsymbol{a}$: $\boldsymbol{D \to S}$ и $\boldsymbol{D \to C}$, способный классифицировать произвольный объект $\boldsymbol{x} \in X$ и $\boldsymbol{x} \in X^*$.

В нашем случае, для новой входной диссертационной работы ***D***, необходимо определить точность и наиболее подходящую специальность ***S*** и сфер специальности ***C***.

Данные и методы исследований

В рамках данной работы рассматривается научная электронная библиотека диссертаций и авторефератов: dissercat.com.

Научная электронная библиотека диссертаций и авторефератов disserCat является одним из самых больших электронных каталогов научных работ в российском интернете. Фонд DisserCat.com составляет более 780 тысяч научно-исследовательских работ, свыше 410 тысяч диссертаций. Для большинства диссертационных исследований в качестве ознакомления доступны оглавление, введение и список литературы [5].

Все диссертации разделены по специальностям согласно ВАК РФ на 25 основных направлений и свыше 700 специальностей [15, стр. 124]. Каждая диссертация относится только к одной специальности. В качестве ознакомительного (бесплатного) материала disserCat предоставляет следующею информацию о диссертации: наименование темы диссертации, ФИО автора диссертации, степень автора диссертации, год защиты диссертации, город/субъект, где была проведена защита, специальность по ВАК РФ, количество страниц, оглавление диссертации, введение диссертации, заключение диссертации, список литературы диссертации.

Для реализации цели нашего исследования требуется четыре последних поля – оглавление, введение, заключение и список источников.

Классификация текстов диссертаций

Имеются отрывки (оглавление, введение, заключение и список источников) из диссертаций, где известна специальность каждого из них. На основе их нужно будет создавать классификаторы, которые будут предсказывать специальность тестовых текстов.

Компьютеры еще не научились понимать человеческий язык, так как слишком много контекстов и неологизмов. Поэтому обычно для классификации тексты переводятся в один определенный формат и будут учитываться на основе каких-нибудь критериев: количество частоты слов, их взаимосвязанные отношения, и другие критерии. Поэтому чаще всего используются методы n-грамм, векторизации и алгоритм tf-idf [16].

В данной работе будет использоваться следующий алгоритм классификации текста:

– очистка текста: удаление знаков пунктуации, пробелов, цифр, стоп-слов, а также перевод текста на нижний регистр;
– создание матрицы токенов через CountVectorizer [17];
– взвешивание текста через алгоритм TF-IDF;
– вскармливание результата в модель классификатора.

В первую очередь нужно очистить текст от элементов, которые не несут какой-либо смысловой нагрузки. Это знаки пунктуации, дополнительные пробелы, цифры и стоп-слова.

Затем нужно провести векторизации через матрицы токенов и взвешивание с помощью алгоритма TF-IDF.

Векторизация через CountVectorizer преобразовывает входной текст в матрицу, значениями которой являются количества вхождения данного ключа (слова) в текст. То есть этот векторизатор создает «словарь» из уникальных слов, и потом подсчитывает вхождение каждого слова в каждый из документов.

Подготовка данных к процессу классификации

В качестве начальных данных для тренировок было сгенерировано несколько наборов данных в формате CSV.

Из-за ограничения ресурсов компьютера (процессора, оперативной памяти) и времени, будет использоваться часть данных. При этом данные будут отбираться строго в равных количествах для более качественного анализа. Результаты приведены в табл. 1.

Таблица 1

Сгенерированные наборы данных*

Table 1

Generated datasets*

Набор	Тип данных	Количество записей	Размер (Мб)
introductions_total	Введения	38 700	985
heads_total	Оглавления	38 700	150
biblios_total	Списки источников литературы	38 700	1489
close_total	Заключения	38 700	486
all_total	Полнотекстовый набор из введения, оглавления, заключения и списка источников литературы	38 700	2 998
all_ten_cats	Полнотекстовый набор из 10 категорий и 2 000 диссертаций разных сфер	2 000	500
five_cats	Полнотекстовый набор из 5 категорий и 5000 диссертаций	2 500	300
all_two_cats	Полнотекстовый набор из 2 категорий и 3000 диссертаций	3000	85
two_cats	Полнотекстовый набор из 2 категорий и 3000 диссертаций одной сферы	3000	90

*Источник: исследование автора

Introductions_total – набор из введений 38 700 диссертаций, которые были отобраны с 387 специальностей, по 100 из каждой.

Heads_total – набор по параметрам схожий *introductions_total*, но вместо текстов введения диссертаций содержит тексты оглавления.

Biblios_total – набор по параметрам схожий *introductions_total*, но вместо текстов введения диссертаций содержит тексты списка источников литературы.

Close_total – набор по параметрам схожий *introductions_total*, но вместо текстов введения диссертаций содержит тексты заключений.

All_total – набор по параметрам схожий *introductions_total,* но содержит объединенные тексты содержаний, введения, заключения и библиографии.

All_ten_cats – полнотекстовый набор, который состоит из диссертаций 10 разных специальностей разных сфер.

Five_cats – полнотекстовый набор, который состоит из диссертаций 5 разных специальностей одной сферы.

All_two_cats – полнотекстовый набор, который состоит из диссертаций 2 разных специальностей разных сфер.

Two_cats – полнотекстовый набор, который состоит из диссертаций 2 разных специальностей одной сферы.

Каждый из наборов загружается в память и разделяется на две части: тренировочные данные и тестовые данные, в соотношении 90 % и 10 %. Для обучения моделей будут использованы тренировочные данные, для проверки и выявления результатов – тестовые данные.

Критерии тестирования

Тестирование результатов будет происходить по двум критериям – точность предсказывания специальности и точность предсказывания сферы. Для первого критерия результат будет положительным, в случае если тестовая категория будет совпадать с результатов предсказания. Для второго критерия результат будет положительным в случае, если родительская категория будет совпадать с родительской категорией предсказанного результата.

Точность для обоих критериев будет измеряться в процентах по формуле ниже:

$$\text{Точность} = \frac{\text{Количество правильных предсказаний}}{\text{Общее количество попыток}} \text{ x } 100$$

Точность зависит от количества правильных предсказаний и будет измеряться в процентах.

Алгоритмы и модели классификации

В качестве примера были выбраны пять различных алгоритмов и моделей. Их список и аргументы указаны в табл. 2.

Таблица 2

Методы и классификаторы, а также их аргументы

Table 2

Methods and classifiers, as well as their arguments

Классификатор	Параметр	Комментарий	Значение
RandomForest [1]	n_estimators	Кол. оценщиков	10
	n_jobs	Параллельных задач	8
LogisticRegression [2]	n_jobs	Параллельных задач	8
kNeighborsClassifier[2]	n_neighbors	Кол. соседей	5
	n_jobs	Параллельных задач	8
DecisionTree [1]	max_depth	Максимальная глубина дерева	none
SVC/SGD (SVM)	n_jobs	Параллельных задач	8
MLPClassifier	activation	Функция активации	relu

В целях проверки результатов помимо классификации научных текстов по специальностям также в качестве критерия были проведены исследования на точность родительской специальности.

В итоге обученные модели каждого из алгоритмов были сохранены на диск для дальнейшего использования. Исходный код классификатора доступен в [13, стр. 29–31].

Результаты работы

Результаты работы классификаторов и методов оказались весьма объективными. В табл. 3 указаны результаты для определенных наборов и сравнительные диаграммы классификаторов в [13].

Таблица 3

Итоговый результат точности предсказаний алгоритмов
для конкретных специальностей в общем наборе*

Table 3

The final result of the accuracy of predictions of algorithms for specific specialties
in the general set*

	Random Forest [1]	**kNN [2]**	**Logistic Regression [2]**	**Decision Tree [1]**	**SGD**	**MLP**
introductions_total	41	54,57	68	49,92	72,73	68,2
heads_total	36	41,52	58	32,61	62,5	61,4
biblio_total	42,4	54,13	71,4	47,4	70,98	69,7
close_total	30,39	48,22	65,98	33,58	65,86	65,33
all	44,53	55,7	76	53,47	74,21	75,27
all_ten_cats	89	94,5	99	85,5	98,5	98,2
five_cats	90,68	85,01	97,72	88,4	97,92	96,9
all_two_cats	100	99	100	99,66	100	99,75
two_cats	100	99,1	99,83	99,83	99,83	98,58

*Источник: исследование автора

В плане точности предсказаний лучшие результаты показали методы логистической регрессии и метод опорных векторов и нейросеть MLP. Результаты их почти сравнимы друг с другом, но в плане скорости метод опорных векторов оказался в разы быстрее. Однако, памяти для методов опорных векторов и нейронной сети потребовалось чуть больше.

Методы случайных лесов и дерева решений показали относительно неплохие результаты, но все же идеальными их назвать, увы, невозможно.

Если брать конкретно набор данных, состоящий из введений, то только три метода смогли пересечь черту в 50 % процентов точности. Все три метода смогли правильно предсказать более чем половину из тестовых наборов.

Самыми оптимальными типами данных для идентификации оказались тексты введения, библиографии и заключения. Идентификация по оглавлению оказалась не столь полезной и классификаторы чаще ошибались, чем угадывали. Единственным методом, который показывает относительно хорошие результаты, оказался метод опорных векторов, который при наборе heads_total [1–2] показал точность идентификации специальности в 62,5 %.

Многие алгоритмы смогли правильно предсказать родительскую категорию специальностей. Статистика работы показана в табл. 4.

Как видно по таблице выше многие методы смогли с вероятностью 60 % и более угадать сферу (родительскую категорию диссертации). Лучше всего себя показали методы логистической регрессии, опорных векторов и нейросеть MLP. Методы стабильно удерживали планку почты в 80 % процентов, а во многих случаях достигали точности более 90 % в общих наборах.

Результаты метода случайных лесов, дерева решений и k-ближайших соседей оставляют желать лучшего. Причем метод дерева решений является явным аутсайдером в плане точности и разница между ним и методом опорных векторов немаленькая.

Если судить по результатам в табл. 4, то можно определить четыре наиболее оптимальных алгоритма: kNN, LogisticRegression, SGD, MLP. При этом kNN является наиболее оптимальным в плане скорости и памяти, SGD – в плане скорости и точности, MLP – в плане точности, а Logistic Regression – в плане точности и памяти.

Таблица 4

Итоговый результат точности предсказаний алгоритмов
для родительских специальностей*

Table 4

The final result of the accuracy of prediction algorithms for parent specialties*

	Random Forest	**kNN**	**LogisticRegression**	**DecisionTree**	**SGD**	**MLP**
introductions_total	69,53	74,8	76	76,52	88,47	82,66
heads_total	65,4	63,86	67	61,84	82,07	77,3
biblio_total	70,82	75,95	84,4	74,1	86,88	73,41
close_total	58,1	70,27	84,64	60,9	83,46	82,67
all	72,85	78,35	90	77,57	89,53	86,3
all_ten_cats	89	94,5	99	85,5	98,5	97,63
five_cats	100	100	100	100	100	99,8
all_two_cats	100	99	100	99,66	100	99,96
two_cats	100	99,1	99,83	99,83	99,83	99,9

*Источник: исследование автора

Таблица 5

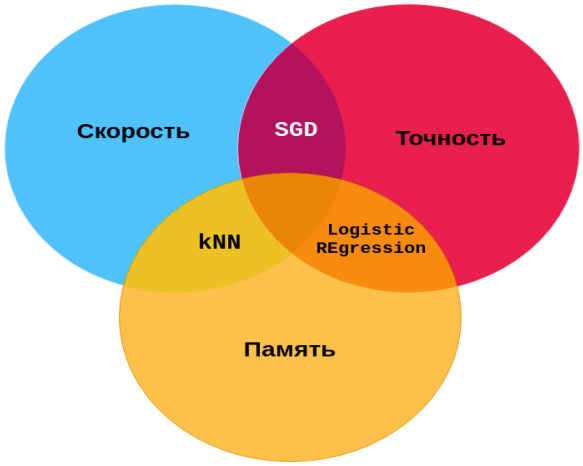
Скорость и память предсказаний алгоритмов*

Table 5

Speed and memory of prediction algorithms*

	Скорость (sec)	Память (MB)
kNN	~44	~23,83
SGD	~55	~42,86
LogReg	~140	~27,32
MLP	~220	~51,23

*Источник: исследование автора



Показатели оптимальности работы классификаторов для классификации диссертаций
Indicators of the optimality of the work of classifiers for the classification of dissertations

По результатам исследования было выявлено, что наиболее оптимальными методами классификации для задачи являются: логистическая регрессия, нейросеть и метод опорных векторов, которые показали хорошие результаты и высокую точность.

Заключение

В результате данной исследовательской работы был проведен анализ классификации научных текстов по специальностям и применены на практике алгоритмы машинного обучения. Была проведена классификация с использованием набора диссертаций из disserCat, основной целью которой является повышение объективности оценки принадлежности научной работы к определенной специальности.

Набор диссертаций был классифицирован и протестирован с помощью шести различных методов классификации. В результате данной работы выяснилось, что для классификации диссертационных исследований по специальностям нужны: оглавление, введение, заключение и список литературы, а не полный текст диссертации.

Данный классификатор текста может быть широко использован в научных целях, в том числе для классификации работ студентов, для оптимизации анализа анализаторов текста и антиплагиата, и во многих других целях. Все модели и результаты были сохранены, а исходные коды и другие данные предоставлены в материалах статьи [15].

Список литературы

1. **Максудов Х. Т., Иномов Б. Б., Муллоджанов Н. М.** Сравнительный анализ методов «дерево решений» и «случайный лес» – при определении специальности научных текстов // Вестник таджикского национального университета серия: естественных наук 2019. № 3. – Душанбе : ТНУ, 2019. С. 23–28.
2. **Максудов Х. Т., Иномов Б. Б.** Оценка эффективности методов k-ближайших соседей и логистической регрессии при определении специальности научных текстов // Политехнический Вестник серия: Интеллект. Инновации. Инвестиции. 4(48)2019. – Душанбе: ТТУ, 2019. С. 34–38.
3. **Гусев П. Ю.** Разработка системы классификации текстов по научным специальностям с применением методов машинного обучения // Вестник НГУ. Серия: Информационные технологии. 2021. Том 19, № 1
4. **Данилов Г. В. и др.** Сравнительный анализ статистических методов классификации научных публикаций в области медицины // Компьютерные исследования и моделирование. 2020. Т. 12, № 4. С. 921–933. DOI 10.20537/2076-7633-2020-12-4-921-933.
5. Научная электронная библиотека диссертаций и авторефератов: [Электронный ресурс]. URL: https://www.dissercat.com/ (дата обращения: 10-09-2018).
6. **Kepa M., Szymanski J.,** Two stage SVM and kNN text documents classifier, In: Pattern Recognition and Machine Intelligence, Kryszkiewicz M. (Ed.), Lecture Notes in Computer Science, Vol. 9124, pp. 279–289, 2015.
7. **Adeniyi D. A., Wei Z., Yongquan Y.** Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN)classification method // Applied Computing and Informatics. – 2016. – Т. 12. – № 1. С. 90–108.
8. **Baralis E., Cagliero L., Garza P.** EnBay: A novel pattern-based Bayesian classifier, Tkde, vol. 25, no. 12, pp. 2780–2795, 2013.
9. **Tang B. et al.** A Bayesian classification approach using class-specific features for text categorization // IEEE Transactions on Knowledge and Data Engineering. – 2016. – Т. 28. – № 6. – С. 1602–1606.

10. **Yoo J. Y., Yang D.** Classification scheme of unstructured text document using TF-IDF and naive bayes classifier // Advanced Scienceand Technology Letters. – 2015–Т. 3. – С. 263–266.
11. **Lilleberg J., Zhu Y., Zhang Y.** Support vector machines and word2vec for text classification with semantic features // Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on. – IEEE, 2015. – С. 136–140
12. **Barik R. C., Naik B.** A Novel Extraction and Classification Technique for Machine Learning using Time Series and Statistical Approach, Computational Intelligence in Data Mining, vol. 3, pp. 217–228, 2015.
13. **Liu Z., Lv X., Liu K., Shi S.** Study on SVM compared with the other text classification methods, 2nd Int. Work. Educ. Technol. Comput. Sci. ETCS 2010, vol. 1, pp. 219–222, 2010.
14. **Pliakos K., Geurts P., Vens C.** Global multi-output decision trees for interaction prediction // Machine Learning. – 2018. – С. 1–25.
15. **Иномов Б. Б.** Ресурсы, код, результаты работы. [Электронный ресурс]. URL: https://drive.google.com/open?id=13SaeBHidtCPpOdXTmtlGMWiT_WwbkujG (дата обращения: 06.04.2019).
16. TF-IDF — Википедия. [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/TF-IDF (дата обращения: 06.04.2019).
17. Sklearn.feature_extraction.text.CountVectorizer. [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html (дата обращения: 16-03-2019).

References

1. **Maqsudov Kh. T., Inomov B. B., Mullojonov N. M.** Comparative analysis of methods «Decision tree» and «random forest» in determining the specialty of scientific texts // Bulletin of the Tajik national university: series of natural sciences 2019. № 3. – Dushanbe: TNU, 2019. – pp. 23–28.
2. **Maqsudov Kh. T., Inomov B. B.** Evaluation of the effectiveness of k-nearest neighbors and logistic regression methods in determining the specialty of scientific texts // Polytechnic Bulletin series: intelligence. Innovation. Investments. 4 (48)2019. – Dushabe: TTU, 2019. – pp. 34–38.
3. **Gusev P. Yu.** Development of a Classification System for Texts by Scientific Specialties Using Machine Learning Methods // Vestnik NSU. Series: Information Technologies, 2021, vol. 19, no. 1
4. **Danilov G. V. et al.** Sravnitel'nyj analiz statisticheskih metodov klassifikacii nauchnyh publickacij v oblasti mediciny [Comparative analysis of statistical methods for the classification of scientific publications in the field of medicine]. Komp'juternye issledovanija i modelirovanie, 2020, vol. 12, no. 4, p. 921–933 (in Russ). DOI 10.20537/2076-7633-2020-12-4-921-933.
5. **Kepa M., Szymanski J.** Two stage SVM and kNN text documents classifier, In: Pattern Recognition and Machine Intelligence, Kryszkiewicz M. (Ed.), Lecture Notes in Computer Science, Vol. 9124, pp. 279–289, 2015
6. **Adeniyi D. A., Wei Z., Yongquan Y.** Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN)classification method // Applied Computing and Informatics. – 2016. – Vol. 12. – № 1. – С. 90–108.
7. **Baralis E., Cagliero L., Garza P.** EnBay: A novel pattern-based Bayesian classifier, Tkde, vol. 25, no. 12, pp. 2780–2795, 2013.
8. **Tang B. et al.** A Bayesian classification approach using class-specific features for text categorization // IEEE Transactions on Knowledge and Data Engineering. – 2016. – Vol. 28. – № 6. – С. 1602–1606.
9. **Yoo J. Y., Yang D.** Classification scheme of unstructured text document using TF-IDF and naive bayes classifier // Advanced Scienceand Technology Letters. – 2015. – Vol. 3. – С.263–266.

10. **Lilleberg J., Zhu Y., Zhang Y.** Support vector machines and word2vec for text classification with semantic features // Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on. – IEEE, 2015. – С. 136–140.
11. **Barik R. C., Naik B.** A Novel Extraction and Classification Technique for Machine Learning using Time Series and Statistical Approach, Computational Intelligence in Data Mining, vol. 3, pp. 217–228, 2015.
12. **Liu Z., Lv X., Liu K., Shi S.**, Study on SVM compared with the other text classification methods, 2nd Int. Work. Educ. Technol. Comput. Sci. ETCS 2010, vol. 1, pp. 219–222, 2010.
13. **Pliakos K., Geurts P., Vens C.** Global multi-output decision trees for interaction prediction // Machine Learning. – 2018. – С. 1–25.
14. **Inomov B. B.** Resources, source code, results. [Electronic resource]. URL: https://drive.google.com/open?id=13SaeBHidtCPpOdXTmtlGMWiT_WwbkujG (seen: 06.04.2019).
15. TF-IDF — Wikipedia. [Electronic resource]. URL: https://ru.wikipedia.org/wiki/TF-IDF (seen: 06.04.2019).
16. Sklearn.feature_extraction.text.CountVectorizer [Electronic resource]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html (seen: 16.03.2019).

Сведения об авторах

Иномов Бехруз Бурхонович, докторант (Ph.D.), старший преподаватель кафедры цифровой экономики, Политехнический институт Таджикского технического университета имени академика М. С. Осими (Худжанд, Республика Таджикистан)

Marina Tropmann-Frick, профессор науки данных кафедры компьютерных наук, Гамбургский университет прикладных наук (HAW Hamburg) (Гамбург, Федеративная республика Германия)

Information about the Authors

Behruz B. Inomov, Ph.D, Senior Lecturer of Digital Economy Department, Polytechnic Institute of the Tajik Technical University named after academician MS Osimi (Khujand, Republic of Tajikistan)

Marina Tropmann-Frick, Professor of Data Science, Department of Computer Science, University of Applied Sciences (HAW Hamburg) (Hamburg, Federal republic of Germany)

Статья поступила в редакцию 04.04.2022;
одобрена после рецензирования 30.05.2022; принята к публикации 30.05.2022
The article was submitted 04.04.2022;
approved after reviewing 30.05.2022; accepted for publication 30.05.2022