

Разработка системы классификации текстов по научным специальностям с применением методов машинного обучения

П. Ю. Гусев

*Воронежский государственный технический университет
Воронеж, Россия*

Аннотация

Описана разработка системы классификации текстов по научным специальностям. В качестве научных текстов использована выборка авторефератов по группе специальностей научных работников 05.13.00 «Информатика, вычислительная техника и управление». Приведены основные сведения по методике сбора и фильтрации исходных данных. Описаны способы подготовки исходных данных к моделированию. Для векторизации исходных текстов использована модель TF-IDF, которая позволила при ограниченных технических возможностях загрузить весь объем данных. Для разработки модели машинного обучения применена многоклассовая логистическая регрессия. В работе также описывается процесс и результат разработки веб-приложения для классификации текстов авторефератов. Приведены сведения о возможных путях развития системы классификации.

Ключевые слова

классификация, машинное обучение, NLP, научная специальность, интеллектуальные системы, веб-приложение

Для цитирования

Гусев П. Ю. Разработка системы классификации текстов по научным специальностям с применением методов машинного обучения // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19, № 1. С. 39–47. DOI 10.25205/1818-7900-2021-19-1-39-47

Development of a Classification System for Texts by Scientific Specialties Using Machine Learning Methods

P. Yu. Gusev

*Voronezh State Technical University
Voronezh, Russian Federation*

Abstract

In the process of preparing a dissertation, the researcher is faced with the problems of determining the research topic, as well as the problems of writing a text for a particular specialty. To increase the objectivity of the classification of scientific texts by specialties, a system has been developed using machine learning algorithms. In the process of building the classification system, the following tasks were solved: development tools were identified, the collection and processing of initial data was carried out, machine learning models were built, and a web application was developed.

The initial dataset is a sample of texts for the group of Russian scientific specialties “Informatics and Computer Engineering”. For the objectivity of the study, the initial data was filtered - the least representative classes were removed. Conversion of texts for vectorization was also carried out.

For the vectorization of the source texts, the TF-IDF model was used, which allowed loading the entire data volume with limited technical capabilities. Multiclass logistic regression is chosen as a machine learning model used to classify a scientific specialty. For training, the initial data divided into 2 parts - training and test in a ratio of 80 to 20. Accu-

racy is used as a metric of the quality of the machine learning model. The choice of the metric is due to the sufficient balance of the classes. The model trained on the training data made it possible to achieve the accuracy of determining the scientific specialty on test data equal to 0.87.

To use a ready-made machine learning model for classification of a scientific specialty, a web application has been developed using Flask. The web app is currently located at <http://predict-spec.herokuapp.com/>

The most urgent tasks for finalizing the system at the moment are: transferring a web application to a more powerful server, finalizing machine learning models, displaying visual information on the analyzed work.

Keywords

classification, machine learning, NLP, scientific specialty, intelligence systems, web-application

For citation

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, p. 39–47. (in Russ.) DOI 10.25205/1818-7900-2021-19-1-39-47

Введение

Формирование тематики исследования для молодого ученого всегда является трудоемкой задачей. Основную помощь в решении подобной задачи может оказать опытный научный руководитель, который является экспертом в определенной области знаний. Ориентируясь на свой опыт профессиональной деятельности, научный руководитель в состоянии определить актуальность тематики исследования и принадлежность к определенной научной специальности. Однако, учитывая объем и доступность информации в современном мире, а также личный взгляд на научные проблемы, мнение одного эксперта не всегда может совпадать с мнением других экспертов или членов диссертационного совета.

Повторно трудности с определением научной специальности у молодого ученого могут возникнуть уже с готовой диссертационной работой. Эти трудности связаны с требованиями различных диссертационных советов. Особенно остро задача определения научной специальности возникает при выполнении работы на стыке специальностей. В этом случае типичной является ситуация, когда молодой ученый подаёт в диссертационный совет научную работу и экспертная группа совета выдает рекомендации по приведению работы к специальности. Но ввиду незнания специфики специальности молодой ученый не в состоянии привести работу в надлежащий вид. А учитывая, что советы по доработке диссертации основываются на экспертном мнении, их невозможно выразить в точном формализованном виде. Как следствие, может возникнуть ситуация, когда соискатель отказывается от защиты своей работы.

В журналах можно встретить ряд работ, посвященных статистическому исследованию тематики диссертаций [1–3]. Подобные исследования позволяют оценить актуальность тематик исследований, но без привязки к работе, для которой требуется определение специальности. Также можно встретить работы, описывающие исследования различных статистических методов классификации научных текстов [4]. Однако подобные работы имеют теоретическое значение и не реализованы в виде программных продуктов.

Сотрудниками лаборатории 404 МГУ разработана система «ИСТИНА», предназначенная для перманентного сбора, систематизации, хранения и анализа наукометрической информации в вузах и научных организациях с целью подготовки и принятия управленческих решений [5]. Но средства интеллектуального анализа сведений о научной деятельности находятся еще в разработке.

Для решения проблем определения научной специальности предлагается система классификации текстов авторефератов на основе алгоритмов машинного обучения, основной целью которой является повышение объективности оценки принадлежности научной работы к определенной специальности. При построении системы классификации текстов авторефератов решены следующие задачи:

- определены инструменты разработки;

- проведены сбор и обработка исходных данных;
- построены модели машинного обучения;
- разработано web-приложение.

Исходные данные

В настоящее время полные тексты диссертаций и авторефераты публикуются в открытом доступе. На сайте ВАК также имеется архив с текстами диссертаций за те годы, когда требование о публикации материалов диссертаций не являлось обязательным. Таким образом, сбор исходных данных представлял собой исключительно техническую задачу.

В рамках данной работы рассматривались научные специальности, относящиеся к группе 05.13.00 «Информатика, вычислительная техника и управление». Такое ограничение продиктовано необходимостью отработки алгоритмов работы системы для дальнейшего проецирования на другие группы специальностей, а также ограниченностью доступных вычислительных мощностей.

Обработка предложений, описываемых в статье, производилась с использованием только авторефератов диссертационных работ. Связано это с тем фактом, что автореферат является краткой выдержкой из диссертационной работы и должен полностью соответствовать выбранной специальности. Также для построения моделей машинного обучения отдельно от текста работы использованы названия диссертаций.

Исходные данные хранятся в файлах формата csv отдельно для каждой специальности из группы 05.13.00. Каждый файл содержит 15 столбцов-признаков, среди которых: тип работы, название работы, научная специальность, дата публикации, дата защиты, текст автореферата. Исходно текст автореферата содержался в формате pdf. Но для разработки моделей машинного обучения текст из формата pdf записан в отдельный столбец таблицы.

На рис. 1 представлен фрагмент таблицы, содержащей информацию о диссертациях по научной специальности 05.13.01.

topic	code	direction	diss_comm	place	date
Алгоритмы повышения точности системы навигации...	05.13.01 - Системный анализ, управление и обра...	Технические науки	NaN	NaN	2020
Интеллектуальная система выбора методов прогно...	05.13.01 - Системный анализ, управление и обра...	Технические науки	NaN	NaN	2020
Информационная система контроля качества произ...	05.13.01 - Системный анализ, управление и обра...	Технические науки	NaN	NaN	2020

Рис. 1. Фрагмент таблицы с информацией

Fig. 1. Fragment of a table with information

Как можно увидеть из рис. 1, не все поля таблицы заполнены. В значительной части работ присутствуют пропуски в указании диссертационного совета и места защиты. Эти столбцы созданы для перспективного исследования работы диссертационных советов и будут заполнены данными из авторефератов.

Для адекватной работы системы классификации научной специальности требуется использовать максимально сбалансированный набор данных. Сбалансированность набора данных в первую очередь определяется количеством авторефератов по каждой анализируемой специальности. Вторым важным показателем сбалансированности – год первой защищенной работы. Однако второй показатель в большей степени влияет на актуальность исследования. Далее в таблице приведены сведения по собранной информации об авторефератах с указанием специальности, количества и года публикации самого раннего автореферата.

Сведения об исходных данных
Information about initial data

Специальность	Количество авторефератов	Год публикации первого автореферата
05.13.01	2200	2008
05.13.05	873	1982
05.13.06	1499	2006
05.13.10	1684	1982
05.13.11	1498	1983
05.13.12	1083	1983
05.13.15	162	2001
05.13.17	896	1997
05.13.18	2600	2008
05.13.19	474	1997

Наибольшее количество диссертаций защищается по специальностям 05.13.01, 05.13.06, 05.13.18, поэтому решено ограничить количество авторефератов этих специальностей по году самой ранней защиты: 2008 год для специальностей 05.13.01 и 05.13.18, 2006 год для специальности 05.13.06. Наименьшее количество работ представлено по специальности 05.13.15 (6,2 % от специальности 05.13.18) и по специальности 05.13.19 (18,2 % от специальности 05.13.18). Выборка по данным специальностям не репрезентативна и будет оказывать негативное влияние на результаты работы моделей машинного обучения. При разработке системы классификации научной специальности от анализа данных специальностей решено отказаться. Таким образом, система позволяет проводить анализ научных работ на соответствие следующим специальностям: 05.13.01, 05.13.05, 05.13.06, 05.13.10, 05.13.11, 05.13.12, 05.13.17, 05.13.18.

Дальнейшая подготовка исходных данных заключалась в обработке текстов авторефератов и названий диссертаций. Применены следующие преобразования:

- все буквы переведены в нижний регистр;
- удалены слова длиной 3 и меньше букв;
- удалены все символы, за исключением букв и цифр (знаки пунктуации и т. д.).

Проведенная подготовка данных позволила привести тексты к единому формату. Но при развитии системы для построения более точных моделей требуется углубленная подготовка. Также дополнительно из текстов авторефератов удалены названия специальностей.

После объединения всех фрагментов данных по всем анализируемым специальностям получен набор данных, содержащий 10685 строк.

Построение моделей машинного обучения

Обязательный предварительный этап построения моделей машинного обучения для обработки естественного языка – векторизация текстовых данных. Существуют разные способы векторизации данных, например: TF-IDF, Word2vec, Glove и т. д. [6–8]. Модели Word2vec и Glove являются сильными word embeddings моделями и их применение планируется в перспективе. Но для первой реализации системы классификации научной специальности решено использовать модель TF-IDF, так как она позволяет запустить систему даже на слабом сервере.

Однако для обучения TF-IDF всё же требуется вычислительная машина с большим объемом оперативной памяти. На объем требуемой при вычислениях оперативной памяти в значительной степени влияет используемое количество подряд идущих слов, которые будут преобразованы в признаки. При доступных для вычислений 16 Гб оперативной памяти, удалось провести векторизацию только каждого слова в отдельности.

В качестве модели машинного обучения, используемой для классификации научной специальности, выбрана многоклассовая логистическая регрессия [9; 10]. Выбор модели обусловлен устойчивостью работы на разреженных данных, а также быстротой обучения и использования. Дополнительным преимуществом логистической регрессии является результат в виде вероятности принадлежности к классу, что позволяет сделать расширенный вывод о принадлежности текста к определенной научной специальности.

Для обучения исходные данные разделены на 2 части – тренировочную и тестовую в соотношении 80 к 20. Таким образом, 8548 авторефератов и тем диссертаций использованы для обучения, а 2137 – для тестирования. В качестве метрики качества работы модели машинного обучения использована точность (ассигасу). В дальнейшем в дополнение к ассигасу для наиболее сложных для классификации классов могут быть использованы метрики precision и recall.

Модель, обученная на тренировочных данных авторефератов, позволила достичь точности классификации научной специальности на тестовых данных, равной 0,87. При этом неверно было распознано 275 авторефератов. Для общей оценки функционирования модели удобно использовать матрицу ошибок (рис. 2).

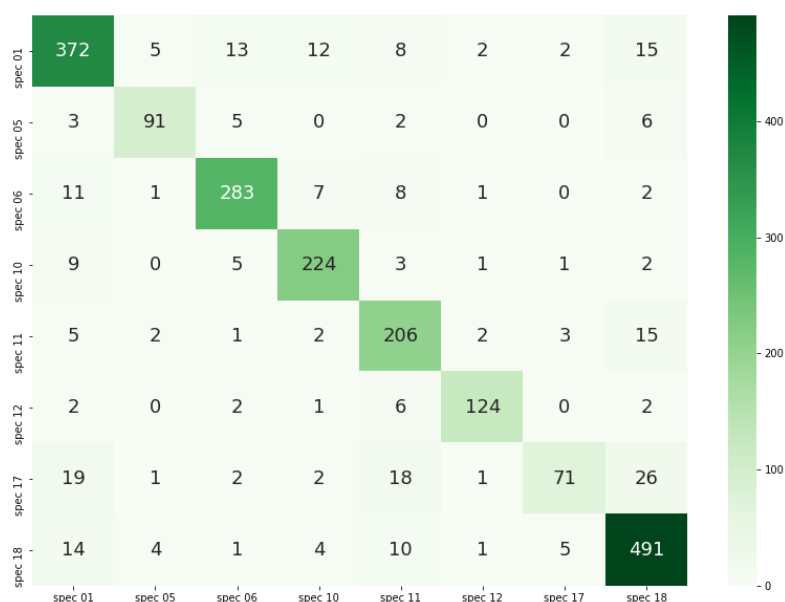


Рис. 2. Матрица ошибок

Fig. 2. Matrix of errors

Как видно из рис. 2, наиболее сложными для классификации научных специальностей являются 05.13.01, 05.13.17, 05.13.18. Сложность определения принадлежности к научным специальностям 05.13.01, 05.13.18 объясняется наибольшим количеством авторефератов в тестовой выборке. Сложность классификации специальности 05.13.17 возникает вследствие малого количества авторефератов данной специальности в обучающей выборке. В дальнейшем планируется построение дополнительных моделей отдельно для специальности 05.13.17 и создание ансамбля моделей на их основе.

Разработка web-приложения

Для использования готовой модели машинного обучения классификации научной специальности разработано веб-приложение. Без приложения запуск модели возможен только напрямую из программного кода, что представляется сложным для массового использования. Определяющими факторами при выборе приложения в виде веб-страницы выступили возможности массового доступа к приложению и выбора конфигурации сервера в зависимости от нагрузки на приложение.

Разработка веб-приложения осуществлена с использованием фреймворка flask, который обеспечивает создание веб-приложений на языке программирования Python с набором инструментов Werkzeug. Flask предоставляет возможности удобного масштабирования приложения в зависимости от возникающих потребностей. Так, например, в первой реализации приложения не предусмотрена авторизация для постоянных пользователей. В дальнейшем, при добавлении авторизации, не потребуется разработка нового приложения, а потребуется только подключение и настройка модуля авторизации.

Для взаимодействия с пользователем предусмотрено 6 шаблонов отображения веб-страниц. В первую очередь разработан базовый шаблон, обеспечивающий единообразие отображения информации на всех страницах. При создании базового шаблона использован набор инструментов bootstrap. Остальные шаблоны приложения наследуются от базового шаблона. Шаблон отображения главной страницы содержит описание функционала приложения. Также предусмотрены шаблоны для проверки названия и текста работы, каждый из которых разделен на 2 шаблона: шаблон загрузки данных и шаблон вывода результата.

Алгоритм проверки научной специальности, реализованный в приложении, состоит из 4 пунктов:

- 1) загрузка названия или текста работы;
- 2) обработка текста с использованием модели векторизации;
- 3) классификация научной специальности с использованием модели машинного обучения;
- 4) вывод результатов.

Первый пункт алгоритма осуществлен с использованием модуля FlaskForm из расширения Flask-WTF. Форма ввода названия отличается наличием окна для ввода текста, а форма загрузки автореферата предусматривает возможность загрузки файла в формате pdf.

Обработка текста и классификация научной специальности осуществляется с использованием уже обученных моделей, которые импортируются в приложение через формат pickle. Учитывая, что фреймворк flask использует язык программирования Python, не возникло трудностей с использованием моделей.

Загрузка названия или текста работы в модель приводит к получению результата, который отображает принадлежность материала к каждой из специальностей. На рис. 3 представлен результат работы веб-приложения.

Как видно из рис. 3, проверенный автореферат находится на стыке специальностей 05.13.11 и 05.13.12. Полученный результат подтвержден экспертами диссертационного совета по специальности 05.13.11.

Подходит для специальности: 05.13.12 С долей
достоверности **45.1 %**

Специальность:	% вероятности:
05.13.12	45.1 %
05.13.11	44.5 %
05.13.18	9.8 %
05.13.17	0.3 %
05.13.01	0.2 %
05.13.05	0.1 %
05.13.06	0.0 %
05.13.10	0.0 %

Рис. 3. Результат работы приложения
Fig. 3. Results of web-application working

В дальнейшем планируется добавить возможность вывода информации, обеспечивающей объяснение полученной принадлежности [11; 12].

Заключение

Проверка актуальности работы и её принадлежности к научной специальности всегда была и остаётся прерогативой опытных экспертов. Однако уже оцененный экспертами материал является ценной базой данных для построения консультационных систем для молодых ученых.

В результате разработки системы классификации научной специальности получены модели машинного обучения, обеспечивающие проверку названий диссертаций и текстов авторефератов. Полученные модели использованы при создании веб-приложения, позволяющего любому пользователю проверить свою работу. В настоящее время веб-приложение расположено по адресу <http://predict-spec.herokuapp.com/>.

Наиболее актуальными задачи по доработке системы на данный момент являются: перенос веб-приложения на более мощный сервер, доработка моделей машинного обучения, вывод визуальной информации по анализируемой работе.

Список литературы

1. **Суслова С. И.** Специальность 12.00.03: статистическое исследование тематики и количества диссертационных работ, представленных к защите // Пролог: журнал о праве. 2018. № 2. С. 48–54. DOI 10.21639/2313-6715.2018.2.8
2. **Девецкий Э. И.** Специальность 12.00. 05: статистическое исследование тематики и количества диссертационных работ, представленных к защите // Пролог: журнал о праве. 2020. № 1 (25). С. 50–56. DOI 10.21639/2313-6715.2020.1.6
3. **Качурова Е. С., Суслова С. И.** Специальность 12.00.08: статистическое исследование тематики и количества диссертационных работ, представленных к защите // Пролог: журнал о праве. 2018. № 3. С. 54–60. DOI 10.21639/2313-6715.2018.3.8
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. **Васенин В. А. и др.** Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА) // Информационное общество. 2013. Т. 1, № 03. С. 21–30.

6. **Бородин А. И., Вейнберг Р. Р., Литвишко О. В.** Методы обработки текста при создании чат-ботов // Хуманитарни Балкански изследвания. 2019. Т. 3, № 3 (5). DOI 10.34671/SCH.HBR.2019.0303.0026
7. **Artama M., Sukajaya I. N., Indrawan G.** Classification of official letters using TF-IDF method. *Journal of Physics: Conference Series*, 2020, vol. 1516, no. 1, p. 012001–012001. DOI 10.1088/1742-6596/1516/1/012001
8. **Grohe M.** Word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In: *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2020, p. 1–16. DOI 10.1145/3375395.3387641
9. **Жилина Е. В.** Использование бинарной логистической регрессии для оценки качества адаптивного теста // Вестник Томского государственного университета. 2010. № 334. С. 106–109.
10. **Попова Е. П., Леоненко В. Н.** Прогнозирование реакции пользователей в социальных сетях методами машинного обучения // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20, № 1. С. 118–124. DOI 10.17586/2226-1494-2020-20-1-118-124
11. **Yang F. et al.** How do visual explanations foster end users' appropriate trust in machine learning? In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020, p. 189–201. DOI 10.1145/3377325.3377480
12. **Collaris D., Wijk J. J. van.** ExplainExplore: Visual Exploration of Machine Learning Explanations. In: *IEEE Pacific Visualization Symposium (PacificVis)*, 2020, p. 26–35. DOI 10.1109/PacificVis48177.2020.7090

References

1. **Suslova S. I.** Special'nost' 12.00.03: statisticheskoe issledovanie tematiki i kolichestva dissertacionnyh rabot, predstavlenykh k zashhite [Specialty 12.00.03: statistical research of the topic and the number of dissertations submitted for defense]. *Prolog: zhurnal o prave*, 2018, no. 2, p. 48–54. (In Russ.). DOI:10.21639/2313-6715.2018.2.8
2. **Devitsky E. I.** Special'nost' 12.00.05: statisticheskoe issledovanie tematiki i kolichestva dissertacionnyh rabot, predstavlenykh k zashhite [Specialty 12.00.05: statistical study of the topic and the number of dissertations submitted for defense]. *Prolog: zhurnal o prave*, 2020, no. 1 (25), p. 50–56. (in Russ.). DOI 10.21639/2313-6715.2020.1.6
3. **Kachurova E. S., Suslova S. I.** Special'nost' 12.00.08: statisticheskoe issledovanie tematiki i kolichestva dissertacionnyh rabot, predstavlenykh k zashhite [Specialty 12.00.08: statistical research of the topic and the number of dissertations submitted for defense]. *Prolog: zhurnal o prave*, 2018, no. 3, p. 54–60. (in Russ.). DOI 10.21639/2313-6715.2018.3.8
4. **Danilov G. V. et al.** Sravnitel'nyy analiz statisticheskikh metodov klassifikatsii nauchnykh publikatsiy v oblasti mediciny [Comparative analysis of statistical methods for the classification of scientific publications in the field of medicine]. *Komp'yuternye issledovaniya i modelirovanie*, 2020, vol. 12, no. 4, p. 921–933. (in Russ.). DOI 10.20537/2076-7633-2020-12-4-921-933
5. **Vasenin V. A. et al.** Intellektual'naya sistema temacheskogo issledovaniya nauchno-tehnicheskoy informatsii (ISTINA) [Intelligent Science and Technology Information Case Study System (ISTINA)]. *Informatsionnoe obshchestvo*, 2013, vol. 1, no. 3, p. 21–30. (in Russ.).
6. **Borodin A. I., Vejnberg R. R., Litvishko O. V.** Metody obrabotki teksta pri sozdanii chat-botov [Text processing techniques for creating chat bots]. *Humanitarni Balkanski izsledvanija*, 2019, vol. 3, no. 3 (5). (in Russ.). DOI 10.34671/SCH.HBR.2019.0303.0026
7. **Artama M., Sukajaya I. N., Indrawan G.** Classification of official letters using TF-IDF method. *Journal of Physics: Conference Series*, 2020, vol. 1516, no. 1, p. 012001–012001. DOI 10.1088/1742-6596/1516/1/012001

8. **Grohe M.** Word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In: Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, 2020, p. 1–16. DOI 10.1145/3375395.3387641
9. **Zhilina E. V.** Ispol'zovanie binarnoj logisticheskoy regressii dlja ocenki kachestva adaptivnogo testa [Using binary logistic regression to assess the quality of an adaptive test]. *Vestnik tomskogo gosudarstvennogo universiteta*, 2010, no. 334, p. 106–109. (in Russ)
10. **Popova E. P., Leonenko V. N.** Prognozirovanie reakcii pol'zovatelej v social'nyh setjah metodami mashinnogo obucheniya [Predicting user reactions on social networks using machine learning methods]. *Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki*, 2020, vol. 20, no. 1, p. 118–124. DOI 10.17586/2226-1494-2020-20-1-118-124
11. **Yang F. et al.** How do visual explanations foster end users' appropriate trust in machine learning? In: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, p. 189–201. DOI 10.1145/3377325.3377480
12. **Collaris D., Wijk J. J. van.** ExplainExplore: Visual Exploration of Machine Learning Explanations. In: IEEE Pacific Visualization Symposium (PacificVis), 2020, p. 26–35. DOI 10.1109/PacificVis48177.2020.7090

Материал поступил в редколлегию

*Received
15.01.2021*

Сведения об авторе

Гусев Павел Юрьевич, кандидат технических наук, доцент кафедры компьютерных интеллектуальных технологий проектирования Воронежского государственного технического университета (Воронеж, Россия)
gusevpvl@gmail.com
ORCID 0000-0002-3752-0152

Information about the Author

Pavel Yu. Gusev, candidate of technical sciences, associate professor of Computer Intelligent Design Technology department, Voronezh State Technical University (Voronezh, Russian Federation)
gusevpvl@gmail.com
ORCID 0000-0002-3752-0152