УДК 004.42

А. Р. Аргинбаев[1], П. Ю. Бугаков[1]*

Разработка и программная реализация алгоритма определения оригинальности текстовых электронных работ обучающихся

[1] Сибирский государственный университет геосистем и технологий, г. Новосибирск, Российская Федерация
* e-mail: arthur.arginbaev@gmail.com

Аннотация. Статья посвящена актуальной проблеме оригинальности самостоятельных работ обучающихся в образовательных учреждениях. В настоящий момент проверке подвергаются только наиболее значимые виды работ (выпускные квалификационные и курсовые работы). Остальные же (рефераты, домашние и лабораторные работы, отчеты по практике) не проверяются на заимствования, что обусловлено стремлением образовательных учреждений лимитировать финансовые расходы. Наиболее распространенные программные системы выявления заимствований работают с крупными базами текстовых документов, которые формируются большим сообществом пользователей. При этом ручная корректировка списка файлов, используемых для проверки оригинальности, существенно ограничена. Однако результат такой проверки внутри небольшого коллектива, например, группы или потока, мог бы стать дополнительным критерием оценки качества. В связи с этим был разработан и протестирован прототип программного обеспечения, позволяющий проверять текстовые документы на предмет заимствования, используя для этого внутреннюю гибко конфигурируемую коллекцию письменных работ учебного заведения. В дальнейшем данное программное обеспечение планируется использовать в учебном процессе на кафедре прикладной информатики и информационных систем СГУГиТ.

Ключевые слова: программное обеспечение, плагиат, антиплагиат, учебный процесс, оценка оригинальности, opencorpora

A. R. Arginbaev[1], P. Yu. Bugakov[1]*

Development and Software Implementation Algorithm for Determining Originality Text Electronic Works of Students

[1] Siberian State University of Geosystems and Technologies, Novosibirsk, Russian Federation
* e-mail: arthur.arginbaev@gmail.com

Abstract. The article is devoted to the actual problem of originality of independent works of students in educational institutions. At the moment, only the most significant types of work (final qualification and term papers) are subject to verification. The rest (abstracts, home and laboratory work, practice reports) are not checked for borrowing, which is due to the desire of educational institutions to limit financial costs. The most common software systems for identifying borrowings work with large databases of text documents that are formed by a large community of users. At the same time, manual adjustment of the list of files used to check originality is significantly limited. However, the result of such a check within a small team, for example, a group or a stream, could become an additional criterion for assessing quality. In this regard, a software prototype was developed and tested that allows you to check text documents for borrowing, using the internal flexibly configurable collection of written works of an educational institution. In the future, this software is planned to be used in the educational process at the Department of Applied Informatics and Information Systems of SSUGT.

Keywords: software, plagiarism, anti-plagiarism, educational process, originality assessment, opencorpora

Введение

Письменные работы являются одним из лучших средств оценки академических достижений обучающихся [1-3]. В настоящее время основным информационным источником является интернет [4- 6]. С одной стороны, свободный доступ ко всему многообразию цифровых научных изданий способствует более качественному и всестороннему раскрытию изучаемой темы, но с другой, соблазн быстрого копирования необходимой информации подталкивает обучающихся к бездумному заимствованию [7- 9]. В итоге письменные работы стали демонстрировать не степень развитости профессиональных навыков, а умение найти и скомпилировать приемлемый текст с минимальным количеством усилий [10-12].

В настоящий момент проверке подвергаются только наиболее значимые виды работ (выпускные квалификационные и курсовые работы) [13-15]. Остальные же (рефераты, домашние и лабораторные работы, отчеты по практике) не проверяются на заимствования, что обусловлено стремлением образовательных учреждений лимитировать финансовые расходы. Наиболее распространенные программные системы для выявления заимствований работают с крупными базами текстовых документов, которые формируются большим сообществом пользователей [16-18]. При этом ручная корректировка списка файлов, используемых для проверки оригинальности, существенно ограничена. Однако результат такой проверки внутри небольшого коллектива, например, группы или потока, мог бы стать дополнительным критерием оценки качества.

В связи с этим возникает необходимость разработки программного обеспечения, позволяющего определять оригинальность текстовых электронных работ обучающихся, используя для этого внутреннюю гибко конфигурируемую коллекцию письменных работ учебного заведения. Для достижения поставленной цели необходимо выполнить следующие задачи: составить список требований к программе; разработать алгоритм и программный прототип; выполнить его тестирование и проанализировать результаты.

Методы и технологии

Программа для определения оригинальности текста должна обеспечивать:

– выбор типов анализируемых документов;

– выбор типов анализируемых значений;

– выбор типов анализируемых частей речи;

– установку минимально допустимого уровня оригинальности;

– выбор режима анализа; выявление плагиата после перестановки фраз, предложений и абзацев;

– игнорирование изменения времен, падежей, и других грамматических категорий слов;

– выявление плагиата после незначительного добавления слов в исходное предложение;

– определение точных и частичных совпадений фрагментов текстов;

– сохранение результатов проверки.

Запрос на выполнение анализа текстовых работ формируется из набора таких параметров, как пути к проверяемым и сравниваемым файлам, путь для сохранения результатов анализа, список поддерживаемых форматов документов, списки анализируемых типов слов и анализируемых частей речи, минимально допустимый уровень оригинальности и т.д.

В процессе обработки запроса формируются списки проверяемых и сравниваемых документов. Для каждого документа создается объект, который содержит список абзацев. Каждый абзац, в свою очередь, хранит список предложений, а каждое предложение – список слов (рис. 1).

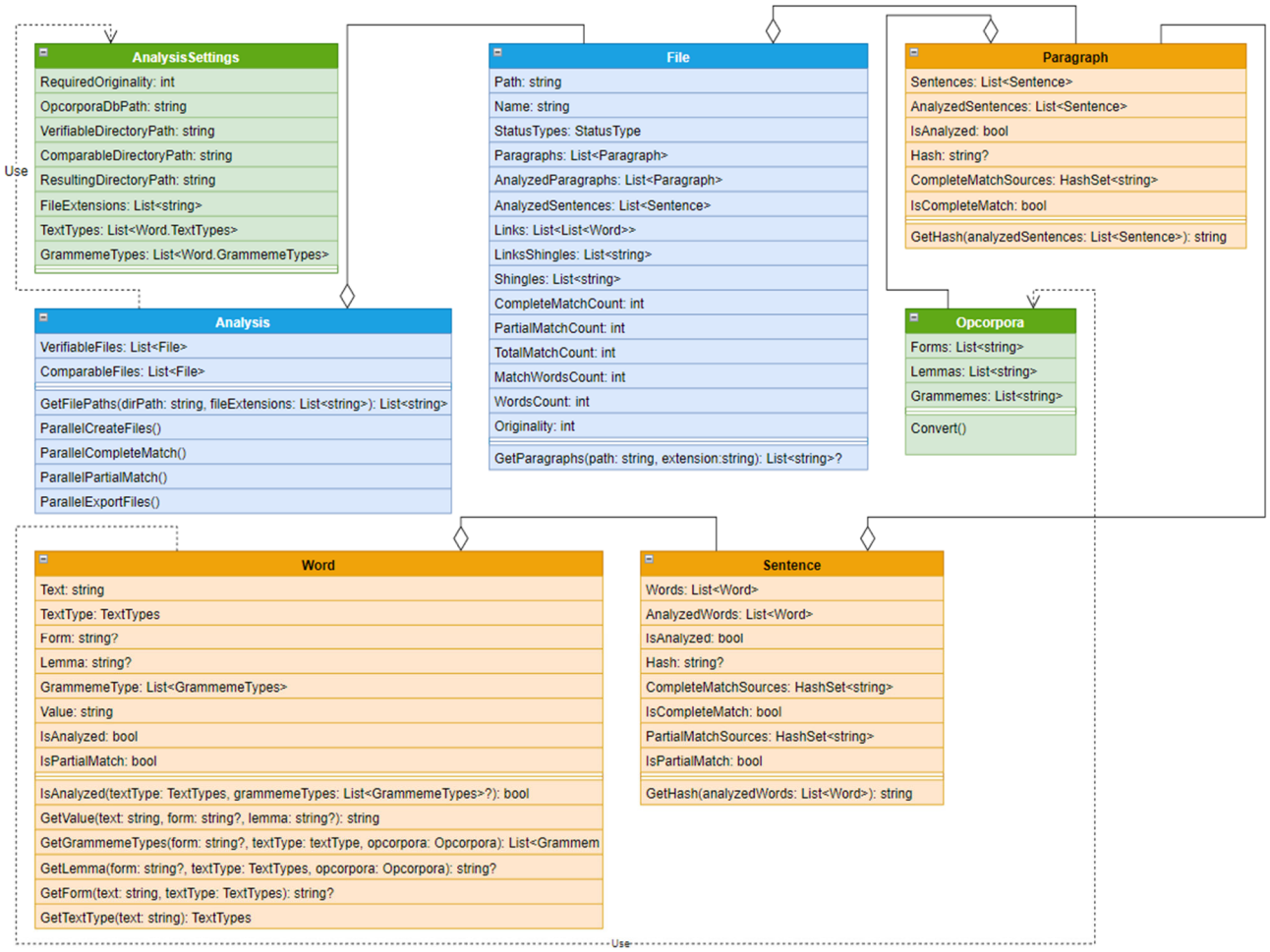

Рис. 1. Структура классов

Для каждого слова определяется его тип: латиница, кириллица, число или символ. Если слово русскоязычное, то для него определяется нормальная форма и вероятные части речи с помощью словаря корпуса русского языка OpenCorpora. В зависимости от полученных результатов, для каждого слова формируется его значение. Если тип слова и его вероятные части речи подходят под параметры анализа, то оно помечается, как анализируемое.

Если предложение содержит хоть одно анализируемое слово, то для него генерируется хеш с помощью хеш-функции SHA256. При этом входным значением является строка, образованная списком значений анализируемых слов, отсортированных по возрастанию. Предложение с хешем помечается как анализируемое. После структурного анализа документов и формирования информационных объектов начинается двухэтапный поиск совпадений.

На первом этапе происходит поиск точных совпадений. Сначала выполняются пересечения множеств хэшей абзацев проверяемых и сравниваемых документов [19]. В случае совпадения хэша в список источников проверяемого абзаца добавляется название сравниваемого документа. Затем выполняются пересечения множеств хэшей предложений проверяемых и сравниваемых документов. В случае совпадения хэша проверяемого предложения в список его первоисточников добавляется название документа, с которым происходило сравнение (рис. 2).

Рис. 2. Алгоритм определения точных совпадений

На втором этапе происходит поиск частичных совпадений. Каждое предложение, не имеющие совпадений на прошлом этапе, разбивается на фрагменты, называемые шинглами (от английского «shingle»). Они формируется из сортированных по возрастанию анализируемых слов предложения. Каждый шингл состоит из двух слов. Выборка слов происходит внахлест, а не встык, то есть каждое слово, кроме первого и последнего, попадает в два соседних шингла. Про-

верка проходит путем пересечения объединенного множества шинглов, проверяемых и сравниваемых документов. В случае совпадения шингл слова и его образующие помечаются как совпавшие (рис. 3). Предложениям, в которых соотношение совпавших слов к анализируемым превышает коэффициент 0.75, в список источников добавляется название сравниваемого документа.

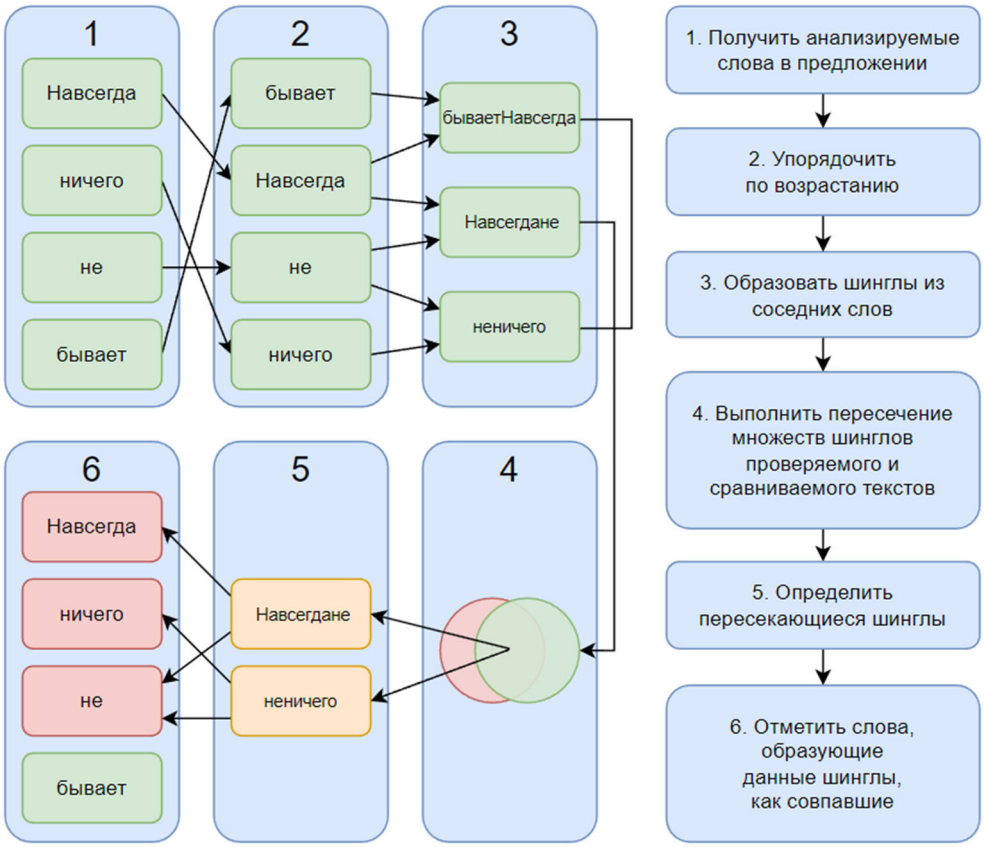

Рис. 3. Алгоритм определения частичных совпадений

Результаты

Для апробации алгоритма был создан прототип программы с графическим интерфейсом (рис. 4).

Для написания кода использовался язык программирования C#. В качестве тестовых данных были взяты 60 курсовых работ, более 100 страниц в каждой. Анализировались только русскоязычные слова и информативные части речи (исключены местоимения, предикативы, предлоги, союзы, частицы и междометия). Тестирование проводилось на аппаратной платформе со следующими характеристиками: процессор AMD Ryzen 7 2700X с тактовой частотой 4 ГГц, 8 ядер, 16 потоков; оперативная память 16 Гб DDR4, с тактовой частотой 3200 МГц, в двухканальном режиме; SSD M.2 накопитель.

При использовании такой конфигурации обработка запроса заняла 30 секунд, максимальная загрузка процессора достигала 60%, объем занятой оперативной памяти – 7 Гб, время, затраченное на проверку приведено в табл. 1.

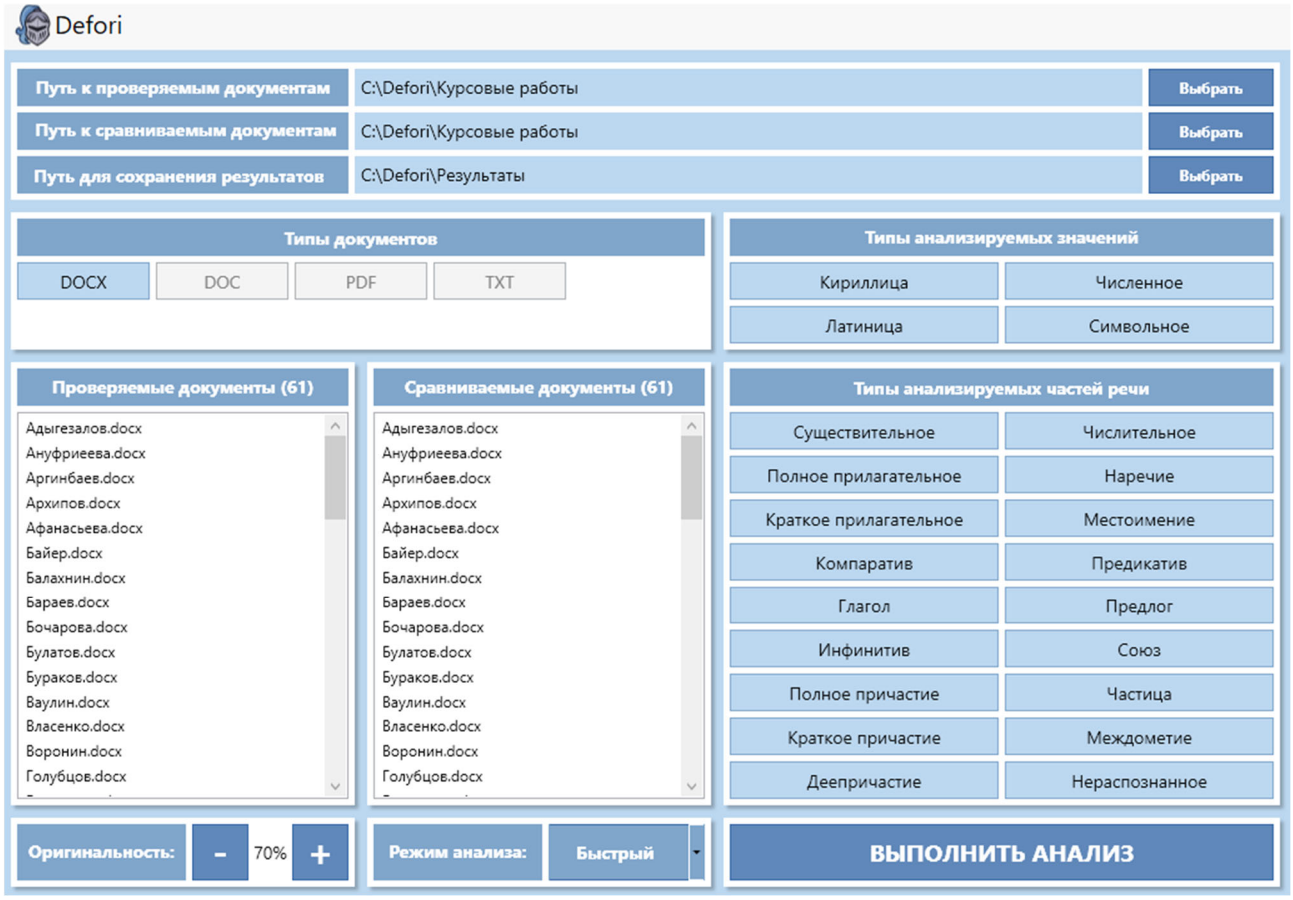

Рис. 4. Интерфейс формирования запроса на выполнение анализа

Таблица 1

Время, затраченное на анализ

Этап анализа	Затраченное время, сек
Открытие и подготовка файлов для сравнения	20,94
Поиск полных совпадений	1,06
Поиск частичных совпадений	0,61
Сохранение результатов поиска	6,71
Общее время выполнения	29,32

Результаты анализа сохраняются в виде документов в формате .docx. В начале отображается краткая сводка, содержащая информацию о количестве совпавших предложений, оригинальности документа и десяти самых объемных источниках заимствований. Процент оригинальности определяется через отношение количества слов из совпавших фрагментов документа к их общему количеству. Затем располагается текст проверяемого документа. В конце абзацев и предложений отображается список источников. Если абзац и предложение имеют общий источник, то в списке источников предложения он не отображается (рис. 5).

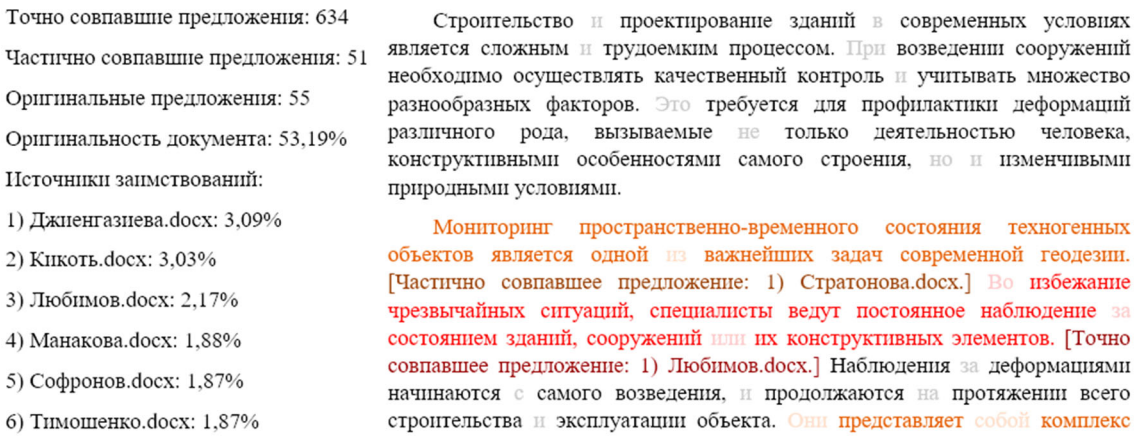

Рис. 5. Результат анализа документа

Обсуждение

Широко используемая в высших образовательных учреждениях система Антиплагиат.ВУЗ позволяет выполнять проверку курсовых и лабораторных работ, однако база текстовых документов в ней формируется коллективно – всеми уполномоченными пользователями в вузе. Они могут самостоятельно добавлять новые или удалять ранее добавленные ими файлы, но не могут управлять файлами других пользователей. Таким образом, ручное формирование списка файлов, используемых для проверки оригинальности письменных работ, становится существенно ограниченным.

Разработанный алгоритм и созданный на его основе программный прототип устраняет этот недостаток, позволяет пользователю самостоятельно выбирать наборы файлов для сравнения, указывать глубину проводимого анализа и получать подробные отчеты о проведенной проверке (рис. 6).

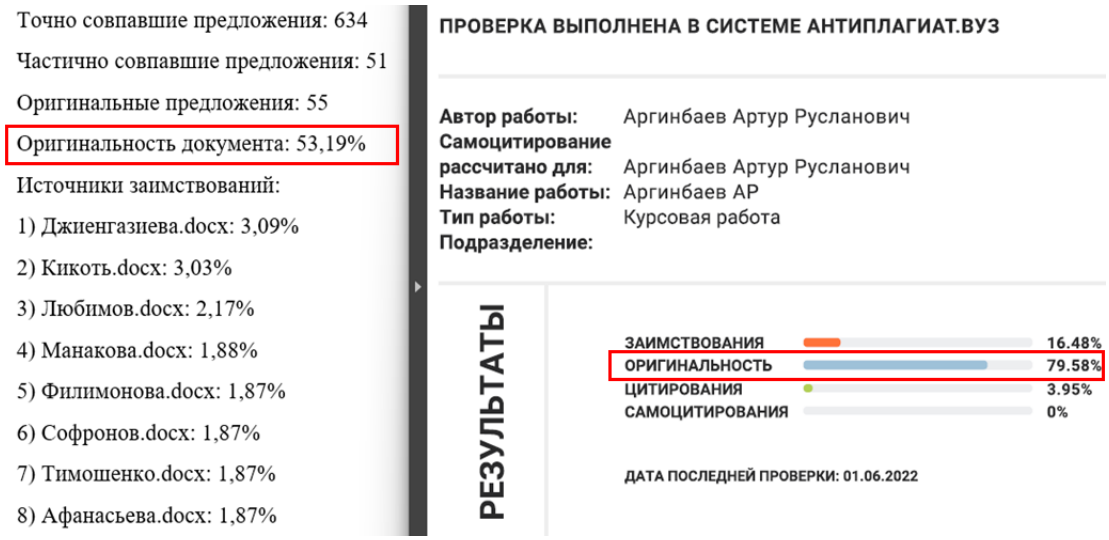

Рис. 6. Сравнение результатов

Заключение

Был разработан и протестирован прототип, позволяющий проверять текстовые документы на предмет заимствования [20]. Он позволяет: определять точные и частичные совпадения; выявлять плагиат после перестановки слов, фраз и предложений, смены формы слов, при незначительном добавлении новых слов в исходное предложение; игнорировать изменения времен, падежей, и других грамматических категорий слова. В дальнейшем данное программное обеспечение планируется использовать в учебном процессе на кафедре прикладной информатики и информационных систем СГУГиТ для проверки оригинальности отчетов лабораторных и курсовых работ.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Васильева В.А., Шабаева А.А. Плагиат глазами студентов: мошенничество или норма // Социально-гуманитарные знания, № 3, 2023, С. 20-29.
2. Болкунов И.А. Списывание в школе: причины, параметры, способы снижения масштабов // Проблемы современного педагогического образования, 2022, С. 42-44.
3. Кацко С.Ю., Кокорина И.П. Проверка ВКР: корректные заимствования, плагиат и оригинальность текста // Актуальные вопросы образования, 2021, № 1, С. 142-145.
4. Плещенко В.И. О плагиате в научных публикациях и выпускных работах // Высшее образование в России. 2018, №8-9, С. 62-70.
5. Рыжко Е. Н. Проблемы текстовых заимствований в учебных и научных работах // Журналистский ежегодник, 2016, №5, С. 41-44.
6. Яркова И.С. Соотношение норм права и морали при разрешении споров, связанных с академическим плагиатом // Социальные нормы и практики, 2021, №1, С. 50-56.
7. Литвинов В.А. Некоторые вопросы оценки качества дистанционного обучения // Вестник УЮИ. 2021. №3 (93), С 159-164.
8. Анисимов А.П., Козлова М. Ю. Плагиат как феномен современной действительности // Имущественные отношения в РФ. 2013. №9 (144). С. 6-13.
9. Левин В.И. Плагиат, его сущность и борьба с ним // Высшее образование в России. 2018. №1. С. 143-150.
10. Витко В. С. О содержании понятия "самоплагиат" // Вестн. Том. гос. ун-та. 2021. №467. С. 235-243.
11. Никитов А.В., Орчаков О.А, Чехович Ю.В. Плагиат в работах студентов и аспирантов: проблема и методы противодействия / Университетское управление: практика и анализ – Екатеринбург, 2012. – С. 61-68
12. Островская А. С. Плагиат в XXI веке: кому это нужно? // ВСП. 2016. №2. С. 148-153.
13. Игнатова И.В. Плагиат как угроза инновационному развитию общества // Вестник евразийской науки. 2012. №4 (13). С. 27.
14. Севостьянов Д.А. Плагиат в современном образовании: беда или симптом? // Высшее образование в России. 2017. №3. С. 17-25.
15. Куликова Е.Ю. Краденая Наука: почему плагиат и самоплагиат неприемлемы // Вестник РГМУ. 2016. №6. С. 50-53.
16. Радаев В.В., Чириков И.С. Отношение студентов и преподавателей к наказаниям за плагиат и списывание // Университетское управление: практика и анализ. 2006. №4. С. 77-82.
17. Бажанов В.А., Козина О.А. Феномен плагиата и его восприятие в академической среде // Вестн. Том. гос. ун-та. Философия. Социология. Политология. 2019. №48. С. 225-235.
18. Кичерова М.Н., Кыров Д.Н., Смыкова П.Н., Пилипушко С.А. Плагиат в студенческих работах: анализ сущности проблемы // Вестник евразийской науки. 2013. №4 (17). С. 82.

19. Петровский А. Б., Теория измеримых множеств и мультимножеств – Наука, 2018. – 244 с.

20. Свидетельство о государственной регистрации программы для ЭВМ 2022682095 Российская Федерация. Программа для анализа оригинальности лабораторных и курсовых работ / П. Ю. Бугаков, А. Р. Аргинбаев; заявитель и правообладатель Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный университет геосистем и технологий». – № 2022681876; заявл. 18.11.2022; опубл 18.11.2022.