4. Kingma, D.P., Ba, J. Adam: A Method for Stochastic Optimization // Proceedings of the 3rd International Conference on Learning Representations. 2015. doi: 10.48550/arXiv.1412.6980.

Мукомберо Хоуп (Зимбабве)

Томский политехнический университет, г. Томск

Научный руководитель: Цапко Ирина Валериевна, канд. техн. наук., доцент

ИСПОЛЬЗОВАНИЕ АЛГОРИТМ ШИНГЛОВ ДЛЯ ОПРЕДЕЛЕНИЯ СХОДСТВА ТЕКСТОВ

Введение

Плагиат является одной из самых серьезных этических проблем в образовании, в связи с чем возникает необходимость проверки уникальности содержания документов, представляемых студентами в рамках их учебной работы. Основной целью данного обзора является побуждение студентов к самостоятельному выполнению заданий, что, в свою очередь, повышает стандарты углубленной исследовательской работы и качество представляемых студентами результатов. Интеграция алгоритмов проверки на плагиат для обнаружения плагиата является обычным явлением для онлайн-проверки опубликованных материалов по глобальным академическим базам данных. Однако если речь идет о взаимной проверке студенческих работ на плагиат внутри вуза, то таких систем не так много, а существующие решения не обладают тем функционалом, который требуется. Настоящая работа направлена на решение этой проблемы и предлагает разработку приложения для сравнения уникальности документов, расположенных на локальном диске, с использованием алгоритма шингла.

Теоретическое понимание методов обнаружения плагиата

Плагиатом является использование чужих слов или идей без указания источника и представление их в качестве своих. Дословный плагиат, также известный как плагиат копирования и вставки, включает в себя прямое копирование и вставку текста из источника без указания автора. Создание совершенно нового текста путем копирования фраз и понятий из нескольких источников называется лоскутным или мозаичным плагиатом. Глобальный плагиат – это когда человек полностью берет чужую работу и выдает ее как свою [1]. Повторное использование ранее представленной работы или повторное использование идей, разработанных на основе предыдущих заданий, называется само плагиатом. Независимо от

того, что эта работа принадлежит человеку, повторная отправка ее как нового материала по-прежнему считается академической нечестностью, потому что признание за данную работу уже получено [2]. При сравнении студенческих отчётов, рефератов и других документов, расположенных на локальном диске, представляет интерес их взаимная проверка на мозаичное и глобальное сходство.

Алгоритм Шингла

Одним из распространённых алгоритмов при поиске заимствований в различных документах является алгоритм Шингла. Блок-схема программы для реализации алгоритма шингла представлена на рисунке 1.

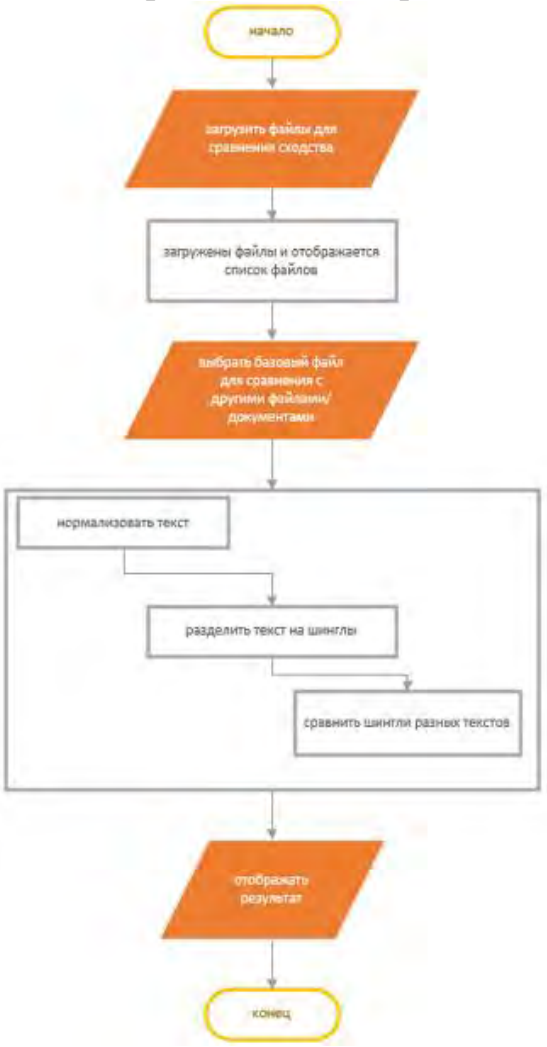


Рис. 1. Блок схема алгоритма программа «офлайн антиплагиата»

Шингла (от англ. scale, cell) – звено, из которого строится цепочка предложений, тем самым образуя текст. Шинглы помогают искать отдельные сочетания слов тем самым проверяя текстовые материалы на уникальность. Основные шаги этого алгоритма [3]:

Нормализация текста (обрезка ненужных слов и знаков препинания).

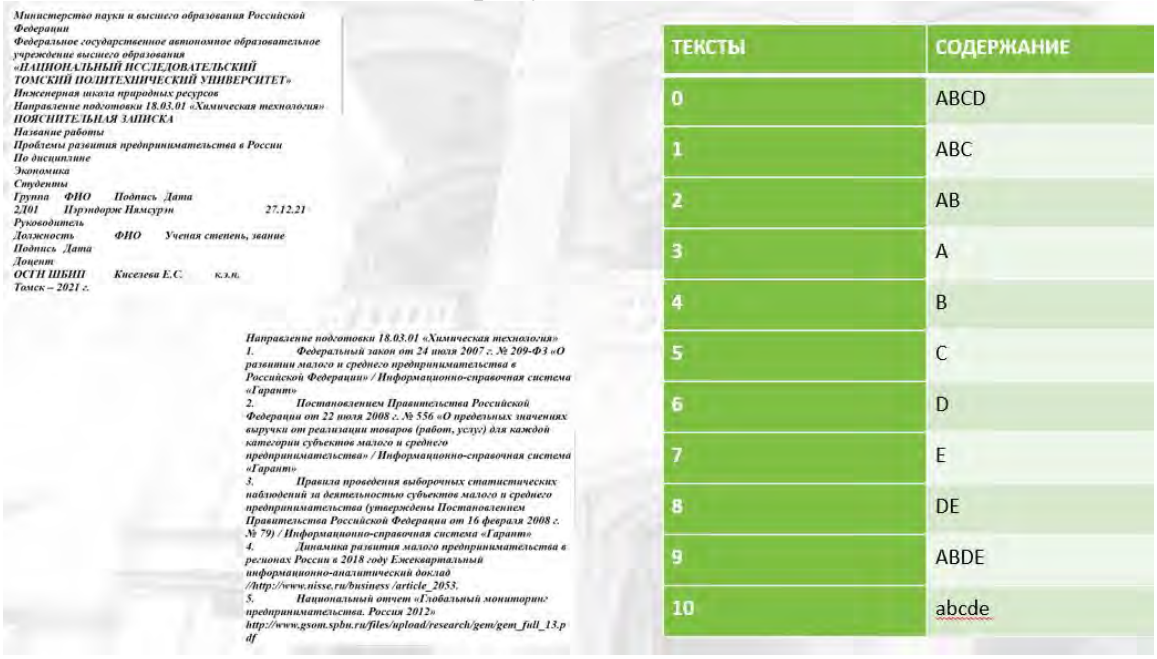
Разделение текста на звенья (чем меньше шингла, тем выше точность анализа).

Сравнение звеньев из разных текстов.

На основе приведенного алгоритма на языке C# была реализована программа проверки сходства/уникальности документов. Программа позволяет выбрать папку с документами для сравнения сходства и при сравнении их алгоритмом шингла выдает результаты в процентах.

Тестирование прототипа программы «офлайн антиплагиат»

Для удобства анализа функциональности разработанной программы в тестовую папку помещены 11 тестовых файлов. Они содержат систематизированное сочетание 5 разделов текста. Тексты имеют кодовые обозначения A, B, C, D и E соответственно. Например, полное сочетание текстов A, B и C представлено как текст ABC (текст 1), а частичное сочетание текстов A, B, C, D и E представлено как текст abcde (текст 10). Тексты соединяются, как показано на рисунок 2.


Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
Инженерная школа природных ресурсов
Направление подготовки 18.03.01 «Химическая технология»
ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
Название работы
Проблемы развития предпринимательства в России
По дисциплине
Экономика
Студенты
Группа ФИО Подпись Дата
2Д01 Пэрэндорж Нямсүрэн 27.12.21
Руководитель
Должность ФИО Ученая степень, звание
Подпись Дата
Доцент
ОСГН ШБИП Киселева Е.С. к.э.н.
Томск – 2021 г.

Направление подготовки 18.03.01 «Химическая технология»
1. Федеральный закон от 24 июля 2007 г. № 209-ФЗ «О развитии малого и среднего предпринимательства в Российской Федерации» / Информационно-справочная система «Гарант»
2. Постановлением Правительства Российской Федерации от 22 июля 2008 г. № 556 «О предельных значениях выручки от реализации товаров (работ, услуг) для каждой категории субъектов малого и среднего предпринимательства» / Информационно-справочная система «Гарант»
3. Правила проведения выборочных статистических наблюдений за деятельностью субъектов малого и среднего предпринимательства (утверждены Постановлением Правительства Российской Федерации от 16 февраля 2008 г. № 79) / Информационно-справочная система «Гарант»
4. Динамика развития малого предпринимательства в регионах России в 2018 году Ежеквартальный информационно-аналитический доклад //http://www.nisse.ru/business/article_2053.
5. Национальный отчет «Глобальный мониторинг предпринимательства. Россия 2012» http://www.gsom.spbu.ru/files/upload/research/gem/gem_full_13.pdf

ТЕКСТЫ	СОДЕРЖАНИЕ
0	ABCD
1	ABC
2	AB
3	A
4	B
5	C
6	D
7	E
8	DE
9	ABDE
10	abcde

Рис. 2. Сочетание тексты

При запуске программы графическое представление результата сравнения показано на графике (рисунок 3).

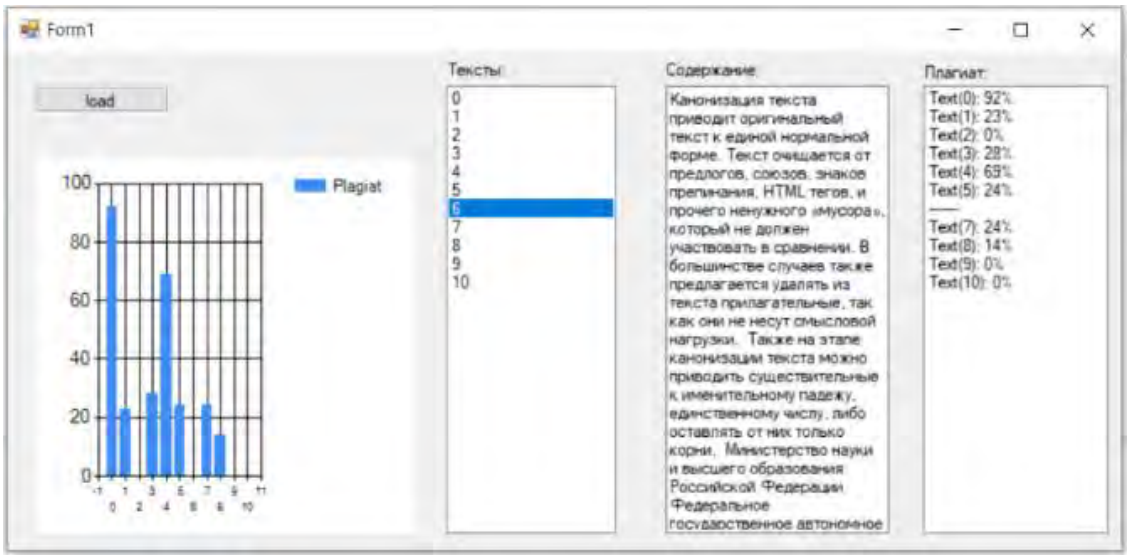


Рис. 3. Графический интерфейс прототипа «офлайн антиплагиата»

Интерпретация результатов сравнения текстов представлена на рисунке 4.

ТЕКСТЫ	СОДЕРЖАНИЕ / ТЕКСТЫ	0	1	2	3	4	5	6	7	8	9	10
0	ABCD		100	96	100	99	100	100	0	30	62	60
1	ABC	82		96	100	99	100	0	0	0	47	55
2	AB	42	51		100	99	0	0	0	0	45	25
3	A	14	17	32		0	0	0	0	0	15	3
4	B	26	32	60	0		0	0	0	0	28	22
5	C	36	44	0	0	0		0	0	0	0	30
6	D	13	0	0	0	0	0		0	28	14	4
7	E	0	0	0	0	0	0	0		66	33	23
8	DE	14	0	0	0	0	0	100	99		50	27
9	ABDE	57	53	96	100	99	0	100	99	100		53
10	abcde	31	35	30	11	44	43	15	39	31	30	

ключ:

очень непохожий текст	0-19
немного похожий текст	20-49
похожий текст	50-79
очень похожий текст	80-100

Для строки 1

В тексте 0 находим:

- 100% совпадение с текстом 1, 3, 5 и
- 99% совпадение с текстом 4;
- 96% совпадение с текстом 2;
- 62% совпадение с текстом 9;
- 60% совпадение с текстом 10;
- 30% совпадение с текстом 8 и
- 0% совпадение с текстом 7.

Для строки 8

В тексте 7 находим:

- 0% совпадение с текстом 0, 1, 2, 3, 4
- 66% совпадение с текстом 8;
- 33% совпадение с текстом 9 и
- 23% совпадение с текстом 10.

Рис. 4. Анализ результатов работы программы

Заключение

В результате была разработана программа, способная сравнивать содержимое документов и визуально показать их сходство. Алгоритм шингла, используемый в программе, не выявляет полностью все формы плагиата, но полезен при сравнении отчетов студентов. В будущем проект может быть расширен за счет, возможно, использования алгоритмов на основе машинного обучения, создания лучшего графического интерфейса и разрешения отображения плагиатных разделов.

СПИСОК ЛИТЕРАТУРЫ

1. What is Plagiarism? // URL: https://www.plagiarism.org/article/what-is-plagiarism (дата обращения: 4.01.2022).
2. Bretag T., Mahmud S. A model for determining student plagiarism: Electronic detection and academic judgement. Journal of University Teaching & Learning Practice, 6(1). – 2009 [Электронный ресурс]. – режим доступа: http://ro.uow.edu.au/jutlp/vol6/iss1/6 (дата обращения: 4.01.2022).
3. Stein, Benno; Lipka, Nedim; Prettenhofer, Peter (2011), "Intrinsic Plagiarism Analysis" (PDF), Language Resources and Evaluation, 45 (1): 63–82, doi:10.1007/s10579-010-9115-y, ISSN 1574-020X, S2CID 13426762, archived from the original (PDF) on 2 April 2012, (дата обращения: 15.12.2021).
4. Методы выявления плагиата [Электронный ресурс]. – // URL: https://commons.wikimedia.org/wiki/File:Методы выявления плагиата.png (дата обращения: 15.11.2020).

Нгикофа Фиел (Намибия),
Пономарев Сергей Викторович (Россия),
Волкова Татьяна Федоровна (Россия)

Томский политехнический университет, г. Томск

РАЗРАБОТКА УСТРОЙСТВА ДИАГНОСТИКИ COVID-19 И МОНИТОРИНГ СОСТОЯНИЯ ПАЦИЕНТОВ НА ОСНОВЕ ПЛАТФОРМЫ ESP32

Коронавирусное заболевание быстро распространяется по всему миру. Клинический спектр пневмонии SARS-CoV-2 варьируется от легких до критических случаев и требует раннего выявления и мониторинга в клинических условиях для критических случаев и удаленно для легких [1]. Важно знать симптомы COVID-19 и действовать соответствующим образом, если у вас есть эти симптомы [2]. Многие предпочитают экспресс-тесты как меру по снижению риска заболевания, поскольку это означает, что нет необходимости быть привязанным к расписанию врачей и терять время в очередях. «У каждого должно быть, по крайней мере, два домашних теста», – сказала медицинский аналитик CNN доктор Леана Вен [3]. Это наглядно показывает высокий спрос на устройства самоконтроля и экспресс-тесты, подобные разработанному устройству.