Методика автоматизированного классифицирования, нормоконтроля и проверки уникальности текстовых документов

В.В. Ветохин, email: daiolix@yandex.ru[1]
Д. А. Романов, email: videogameshateme@icloud.com[1]
К. С. Попова, email: popovaksenia128@gmail.com[1]
С. В. Лукьянец, email: lukyanec35@mail.ru[1]
К. А. Костюков, email: kostukovkonstantin70@gmail.com[1]

[1]ФГБОУ ВО «Воронежский государственный технический университет»

Аннотация. *В статье рассматривается проблема применения автоматизированных систем для проверки текстовых документов на наличие в них чужой интеллектуальной собственности и автоматизации нормоконтроля оформления текстовых документов. Был исследован процесс расширения функциональных возможностей инструмента автоматизированного поиска ошибок. В результате был разработан алгоритм проверки последовательности действий по формализации требований стандартов, регулирующих и регламентирующих оформление документации.*

Ключевые слова: *Антиплагиат, текстовый документ, автоматизация, нормоконтроль, цифровой документ, алгоритм, стандарты.*

Введение

В настоящее время, при стремительном развитии технологий всё больше и больше текстовой документации переходит в цифровой формат. Именно переход к цифровизации особенно четко требует выполнения правил, которые едины и прописаны в стандартах. Этот процесс сильно затронул технические учебные заведения, где большинство документов отныне используются в цифровом формате.

В документообороте университета содержится огромное количество данных в виде документов, поступающих от студентов на проверку преподавателям. Эти документы подлежат обязательной проверке на правильность оформления, однако ручная проверка работ, как правило, занимает немало времени проверяющего.

Существуют единые требования к оформлению текстовых документов, соблюдение которых необходимо. При этом подавляющая часть этого списка, а именно: пояснительные записки выпускных

квалификационных работ, курсовых работ и проектов, а также отчеты по лабораторным работам и практикам — подлежит как обязательной проверке преподавателем на содержание, так и обязательному нормоконтролю, поскольку каждый документ должен быть оформлен в соответствии с определенными требованиями и нормами.

При этом проблема осуществления качественного нормоконтроля при сокращении времени на рутинную работу стоит перед разработчиками достаточно давно. И ее актуальность со временем не уменьшается.

Таким образом, существование программного обеспечения, которое способно в кратчайшие сроки выявить все ошибки, могло бы очень помочь преподавателям.

Решением данного вопроса является интеллектуальная автоматизированная система, представляющая собой программно-аппаратный комплекс для проверки текстовых документов на правильность оформления и наличие заимствований из открытых источников сети Интернет и других источников, целью которой является сокращение времени на проверку и классификацию документом, за счет внедрения новой методики.

1. Анализ имеющихся систем

Существует три больших класса систем для поиска заимствований:

1. Поисковые системы сети Интернет. Не предназначены для поиска заимствований, но с их помощью можно искать заимствования вручную.

2. Метапоисковые системы и системы антиплагиата, не имеющие значимой собственной базы документов. Работают посредством формирования вызовов на основе проверяемого документа к популярным поисковым машинам сети Интернет, интерпретируют их результаты. Для ускорения работы оперируют стоп-словами и проводят непоследовательную проверку документа (метод выборок) и др.

3. Специализированные системы антиплагиата с собственными алгоритмами поиска совпадений и собственными базами документов.

На рисунке представлена схема работы сервисов, позволяющих проверить документ на заимствование их открытых источников.

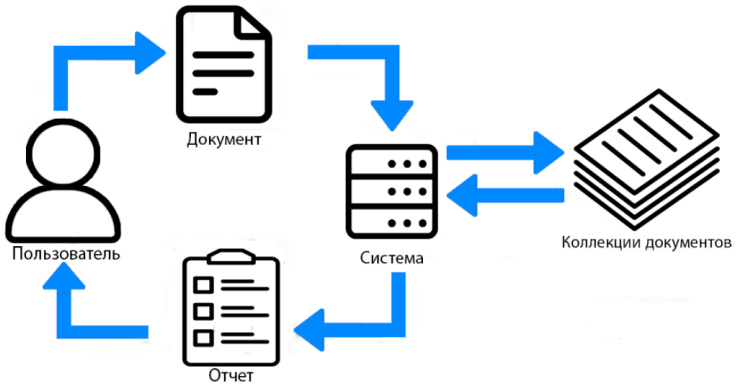

Рис. 1. Схема работы систем антиплагиат

Существует не так уж и много подобных сервисов, но и у них есть свои недостатки. Был проведён анализ и вычислены недостатки некоторых аналогов. Одним из таких является «Text.ru», данный антиплагиат имеет длительное время обработки информации, а также большую задержку в проверке текста. Так же не позволяет загружать файл в формате .txt, .doc, .pdf. Рассмотри другую систему – «Тектовод». Данный сервис также не позволяет загружать документы различных форматов, кроме того, одним из недостатков является непонятный и неудобный интерфейс.

По нормоконтролю и определению группы специальностей сервисы не известны.

2. Описание сервиса

Мы видим следующее решение проблем: мы создаём собственное отдельное веб-приложение, с понятным и доступным интерфейсом на русском языке, с возможностью вставки текстового документа любого формата и с бесплатной пробной подпиской. Данное решение упрощает и ускоряет обработку документов, проверку их оригинальности и нормоконтроля. В перспективе также рассматривается возможность добавления нового функционала, который также будет сверять УДК обрабатываемого документа и проверять его на соответствие с заданной темой.

Основные этапы работы выражены в следующем виде:

1. загрузка и сохранение нового документа на сервер, проверка его оформления;
2. сохранение документа на сервере;
3. получение свойств документа;

4. классификация;
5. проверка документа на соответствие выбранным;
6. сохранение последних результатов проверки для.

3. Работа сервиса

В основу разработки системы автоматизированной проверки соответствия требованиям оформления была положена клиент-серверная архитектура. Необходимо было предусмотреть реализацию архитектуры с учетом непрерывности процессов создания и проверки оформления документа. Основные идеи:

- реализация проверки оформления электронных документов единого наиболее распространенных форматов электронных документов — docx, pdf, txt;
- клиент отправляет файл в исходном виде на сервер;
- на сервере запускается ряд последовательных процессов по проверке документа на соответствие требованиям оформления соответствующего стандарта (включая модуль машинного обучения и модуль проверки);
- итогом работы серверной части является список абзацев и список ошибок к ним, а также процент заимствования.

4. Реализация проекта

В рамках проектирования сервиса, его работа была разделена на несколько модулей. Модуль для взаимодействия с документами в этой системе занимает одно из ключевых мест.

Один из таких модулей - python-docx, с помощью которого имеется возможность создавать и изменять документы с расширением docx. Данный модуль может помочь в анализе документов для создаваемого проекта. Изучение и реализация препроцессинга в python, подразумевающего собой очистку данных и токенизацию.

Для поиска сопоставлений между словами Мы используем модель Average Word Embeddings Model (модель среднего количества сопоставлений слов).

Для определения тематики текста, мы используем библиотеку NLTK для анализа текста и библиотеку wordcloud для построения облака слов.

Заключение

В статье изучена проблема избытка данных в документообороте университета и отсутствия средств для их автоматизированной проверки на соответствие правилам оформления. Также рассмотрены способы

работы с документами, на основе которых составлен алгоритм поиска ошибок оформления в отчетах и пояснительных.

Подводя итог, можно заключить, что подобное приложение - очень полезное решение для учебных учреждений в современном мире, оно способно в разы сократить и улучшить обработку цифровых документов, поэтому автоматизация данного продукта является необходимой задачей.

Список литературы

1. Зимина Е. В, Кайнова В. Н. Метрологическая экспертиза и нормоконтроль технической документации / В. Н. Кайнова. – М. : Лань, 2019. – 500 с.
2. Трофимов В. Б., Кулаков С. М. Интеллектуальные автоматизированные системы управления технологическими объектами / В.Б. Трофимов – М. : Инфра-Инженерия, 2020. – 256 с.
3. Бородин И.Ф. Автоматизация технологических процессов и системы автоматического управления / И.Ф. Бородин. – М. : КолосС, 2006. – 352 с.
4. Насыров Н.Ф., Кобец Е.А., Горлушкина Н.Н. Автоматизированная генерация учебных подзадач на основе методики тегов и критериев // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки – 2020. – № 3. — С. 102–107
5. Ивановский А.А. Объектная модель системы избирательного распространения информации. Научные и технические библиотеки. 2019;(4):61–75.
6. Байбаков В., Клименко Э. Опыт нормоконтроля. Техническое задание на разработку автоматизированной системы // Стандарты и качество – М., 2012– № 7 С: 42–47