УДК 004.91

РАЗРАБОТКА ВЕБ-ПРИЛОЖЕНИЯ ДЛЯ АВТОМАТИЧЕСКОГО СКАНИРОВАНИЯ НАУЧНЫХ ПУБЛИКАЦИЙ НА НАЛИЧИЕ ЗАПРЕЩЁННЫХ МАТЕРИАЛОВ С ЛОКАЛИЗАЦИЕЙ ФРАГМЕНТОВ И ГЕНЕРАЦИЕЙ ОТЧЁТА

З.П. Рощин, И.А. Гергерт, Д.А. Витенко
бакалавр
E-mail:fseputem.zohan@gmail.com, gergert228@mail.ru, famaru544@gmail.com
Научный руководитель: **Г.С. Завалин**
Владивостокский государственный университет
Колледж информационных и креативных технологий, преподаватель
E-mail:Georgiy.Zavalin08@vvsu.ru

В работе описан процесс разработки веб-приложения для автоматического сканирования научных публикаций на предмет наличия в них запрещённых материалов: упоминаний иностранных агентов, нежелательных и запрещённых в Российской Федерации организаций. Реализован программный модуль, обеспечивающий загрузку документов в форматах PDF и DOCX, извлечение и нормализацию текста, поиск совпадений по справочнику запрещённых терминов, локализацию обнаруженных фрагментов и формирование структурированного PDF-отчёта. Приложение разработано с использованием фреймворка Flask и системы управления базами данных PostgreSQL.

Ключевые слова и словосочетания: веб-приложение, автоматизация, сканирование текста, обработка документов, Flask, PostgreSQL, Python, информационная безопасность, научные публикации.

Development of a web application for automatic scanning of scientific publications for prohibited materials with fragment localization and report generation

This paper describes the development of a web application for automatic scanning of scientific publications for the presence of prohibited materials: mentions of foreign agents, undesirable and banned organizations in the Russian Federation. A software module has been implemented that provides uploading of documents in PDF and DOCX formats, text extraction and normalization, searching for matches against a dictionary of prohibited terms, localization of detected fragments and generation of a structured PDF report. The application is developed using the Flask framework and the PostgreSQL database management system.

Keywords: web application, automation, text scanning, document processing, Flask, PostgreSQL, Python, information security, scientific publications.

В современных условиях усиления правового регулирования научной и образовательной деятельности в Российской Федерации перед авторами публикаций и редакциями научных изданий встают новые задачи, связанные с проверкой содержимого работ на соответствие действующему законодательству. В частности, согласно нормам, регулирующим деятельность средств массовой информации и распространение информации, упоминание физических и юридических лиц, признанных иностранными агентами, без сопроводительной маркировки, а также упоминание организаций, признанных нежелательными или запрещёнными на территории Российской Федерации, может повлечь административную ответственность с наложением штрафов в размере до 50 тысяч рублей.

В свою очередь, рост числа научных публикаций, ежегодно поступающих в редакции журналов, сборники конференций и системы антиплагиата, делает ручную проверку каждого документа на наличие подобных упоминаний крайне трудоёмкой задачей. Проверяющий вынужден последовательно просматривать весь текст, сверяясь с актуальными списками иностранных агентов и нежелательных организаций, что занимает значительное время и сопряжено с риском упустить отдельные фрагменты вследствие человеческого фактора.

Существующие на рынке программные решения, ориентированные преимущественно на проверку оригинальности текста, не обеспечивают автоматического выявления подобного рода контента и формирования отчётов о его местоположении в исходном документе.

Изложенные обстоятельства определяют актуальность разработки специализированного программного средства, способного автоматизировать процесс проверки научных публикаций. Целью настоящей работы является создание веб-приложения, обеспечивающего автоматическое сканирование загружаемых научных публикаций на наличие запрещённых материалов, локализацию выявленных фрагментов и генерацию структурированного отчёта по результатам проверки.

Для достижения поставленной цели в ходе работы решались следующие задачи: разработка механизма сканирования текста на запрещённые фрагменты, обеспечение визуальной локализации найденных фрагментов в исходном документе, а также реализация генерации структурированного отчёта по результатам проверки.

Архитектура разработанного приложения построена по классической трёхзвенной схеме «клиент — сервер — база данных». В качестве клиентской части использованы технологии HTML и CSS, обеспечивающие отображение пользовательского интерфейса и взаимодействие пользователя с системой. Серверная часть реализована на языке программирования Python с применением микрофреймворка Flask, отвечающего за обработку HTTP-запросов, управление логикой приложения и взаимодействие с хранилищем данных. В качестве системы управления базами данных выбрана PostgreSQL, обращение к которой осуществляется через объектно-реляционную библиотеку SQLAlchemy. Общая последовательность работы приложения представлена на рисунке 1:

«Рис.1. Общая схема работы приложения»

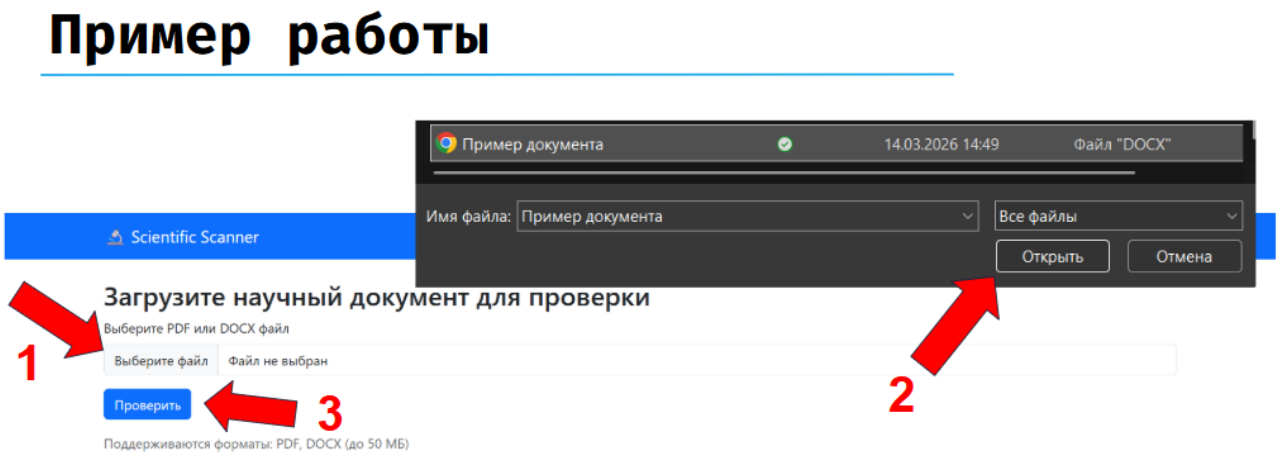
Ключевым элементом системы является поисковый движок, отвечающий за обнаружение запрещённых терминов в тексте загруженного документа. Алгоритм его работы организован следующим образом. На первом этапе из исходного файла, поступившего от пользователя в формате PDF или DOCX, извлекается текстовое содержимое. Затем выполняется токенизация: текст разбивается на отдельные слова, после чего каждое слово приводится к нижнему регистру и проходит процедуру нормализации. Полученные таким образом токены последовательно сравниваются со справочником запрещённых терминов, хранящимся в базе данных и содержащим перечень иностранных агентов, нежелательных и запрещённых организаций. При обнаружении совпадения фиксируется позиция фрагмента в исходном тексте и сохраняется его контекст — окружающее предложение, необходимое для последующего отображения в отчёте. Принципиальная схема функционирования поискового движка представлена на рисунке 2.

«Рис.2. Схема работы поискового движка»

По итогам сканирования возможны два сценария завершения работы. Если в проверяемом документе не выявлено ни одного совпадения со справочником, пользователю выдаётся сообщение об отсутствии нарушений. В противном случае система формирует структурированный PDF-отчёт, содержащий перечень обнаруженных запрещённых терминов с указанием их контекста в исходном документе. Такой формат вывода позволяет проверяющему быстро ознакомиться со списком потенциальных нарушений и при необходимости перейти к конкретным местам исходной публикации для принятия решения.

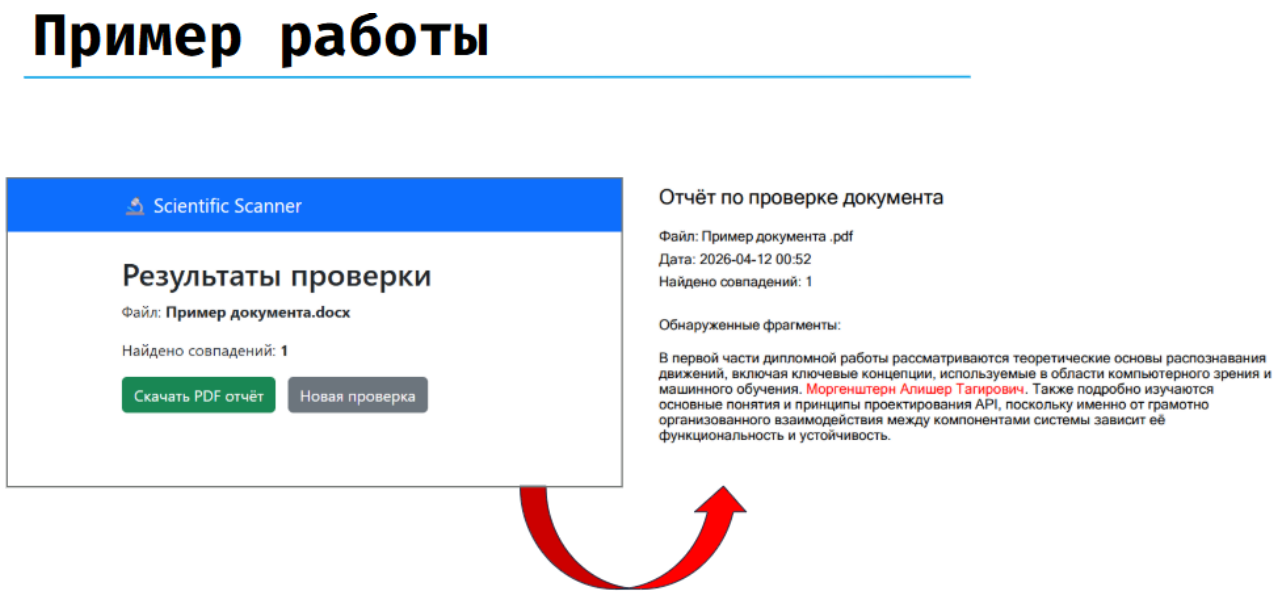
Пользовательский сценарий работы с приложением является максимально простым и состоит из трёх последовательных действий. На главной странице веб-приложения пользователь нажимает кнопку выбора файла, в открывшемся диалоговом окне операционной системы указывает документ для проверки в формате PDF или DOCX (с учётом ограничения на размер до 50 МБ), после чего нажимает кнопку «Проверить» для запуска сканирования. Внешний вид формы загрузки документа представлен на рисунке 3.

«Рис.3. Интерфейс загрузки документа»



После завершения процесса сканирования пользователь перенаправляется на страницу с результатами проверки, где отображается имя проверенного файла и общее количество обнаруженных совпадений. Пользователю предоставляется возможность скачать сформированный PDF-отчёт либо начать новую проверку. В сформированном отчёте указывается имя исходного файла, дата и время проведения проверки, общее количество найденных совпадений, а также перечень обнаруженных фрагментов с выделением запрещённого термина в тексте окружающего его предложения (рис 4):

«Рис.4. Страница результатов и пример отчёта»



В ходе выполнения работы было разработано полнофункциональное веб-приложение, обеспечивающее автоматизированную проверку научных публикаций на наличие запрещённых материалов. Поставленная цель достигнута: реализованы все три ключевые функции — сканирование текста, локализация выявленных фрагментов и генерация отчёта. Применение разработанной системы позволяет сократить трудозатраты на проверку публикаций и снизить влияние человеческого фактора на её результаты.

Дальнейшее развитие проекта предполагает доработку механизма поиска упоминаний с учётом контекста на основе методов обработки естественного языка, что позволит снизить количество ложноположительных срабатываний при совпадении нейтральных слов с элементами справочника. Также планируется добавление модуля авторизации с использованием учётных записей Владивостокского государственного университета через протокол OpenID Connect и реализация вывода отчёта непосредственно в интерфейсе приложения без необходимости его предварительного скачивания.

Список использованных источников:

1. Федеральный закон от 14.07.2022 № 255-ФЗ «О контроле за деятельностью лиц, находящихся под иностранным влиянием» [Электронный ресурс]. – Режим доступа: http://publication.pravo.gov.ru/Document/View/0001202207140014
2. Перечень иностранных агентов [Электронный ресурс] // Министерство юстиции Российской Федерации. – Режим доступа: https://minjust.gov.ru/ru/activity/directions/998/
3. Гринберг Дж. Flask. Разработка веб-приложений на Python / Дж. Гринберг. – М.: ДМК Пресс, 2019. – 320 с.
4. SQLAlchemy Documentation [Электронный ресурс]. – Режим доступа: https://docs.sqlalchemy.org/
5. PostgreSQL: The World's Most Advanced Open Source Relational Database [Электронный ресурс]. – Режим доступа: https://www.postgresql.org/