Система автоматического обнаружения плагиата в научных работах

1. Цель системы

Система предназначена для автоматизированного анализа текстов научных работ с целью выявления заимствований из открытых источников и публикаций. Основные задачи системы: обеспечение академической честности, повышение качества научных публикаций и сокращение времени проверки на плагиат.

2. Пользователи системы

1. **Администратор системы** — управляет правами доступа, настраивает параметры проверки, обновляет базы данных и алгоритмы, формирует сводные отчёты.
2. **Преподаватель/Рецензент** — загружает тексты для проверки, анализирует результаты, принимает решения о допуске работ, формирует отчёты о проверке.
3. **Студент/Автор** — загружает собственную работу для предварительной проверки, получает отчёт о возможных заимствованиях и рекомендации по исправлению.

3. Структура системы и её подсистем

1. **Подсистема загрузки и предобработки текстов** — принимает документы в распространенных форматах (.docx, .doc, .pdf, .txt, .rtf), извлекает текст, нормализует его (приводит к единому формату, удаляет стоп-слова).
2. **Подсистема сравнения и анализа** — сравнивает текст с базами данных источников, использует алгоритмы семантического и синтаксического анализа для выявления заимствований, рассчитывает процент уникальности.

3. **Подсистема хранения данных** — содержит базы данных научных работ, публикаций, веб-источников.
4. **Подсистема формирования отчётов** — генерирует отчёты с указанием источников заимствований, визуализирует результаты (графики, таблицы, выделение текста).
5. **Интерфейсная подсистема** — предоставляет web-интерфейс для взаимодействия с пользователями: загрузка документов, настройка параметров, просмотр отчётов.

4. Связи и взаимодействия между подсистемами

1. Пользователь загружает документ через интерфейсную подсистему в подсистему загрузки и предобработки.
2. Обработанный текст передаётся в подсистему сравнения и анализа, которая взаимодействует с подсистемой хранения данных для поиска совпадений.
3. Результаты анализа направляются в подсистему формирования отчётов.
4. Готовый отчёт выводится через интерфейсную подсистему пользователю.
5. Действия пользователя (например, настройки проверки) влияют на работу всех подсистем в последующих циклах обработки.

5. Основные сценарии работы пользователей с системой

1. **Сценарий проверки работы на плагиат.**
   Преподаватель загружает документ, система проводит анализ и выдает отчёт с указанием процента заимствований и источников. Преподаватель принимает решение о допуске работы.
2. **Сценарий самопроверки автором.**
   Студент загружает свою работу, получает предварительный отчёт, вносит правки и повторно проверяет текст.
3. **Сценарий администрирования и формирования сводной отчётности.**
   Администратор системы настраивает алгоритмы проверки,

обновляет базы источников, формирует отчёты по активности пользователей и результатам проверок.

6. Три направления развития системы

Интеграция с внешними источниками

Подключение API eLibrary и CrossRef для проверки публикаций на русском и английском языках. Настройка регулярного обновления базы источников каждые 24 часа. Реализация модуля кэширования результатов поиска, чтобы ускорить повторные проверки.

Машинное обучение и обработка текста (NLP)

Внедрение модели распознавания перефразированного текста (на основе BERT/RuBERT). Разработка алгоритма автоматической классификации типа заимствования (прямое цитирование, перевод, перефразирование). Тестирование модели на корпусе из 10 000 проверенных работ для настройки точности.

Аудит и контроль действий пользователей

Добавление журнала активности (кто, когда и какой документ загрузил/проверил). Реализация автоматической генерации отчётов для администратора (ежемесячная статистика проверок, распределение по факультетам). Настройка системы уведомлений при аномальной активности (например, массовая загрузка документов с одного аккаунта).