УДК 004.588

ОПТИМИЗАЦИЯ СИСТЕМЫ ПРОВЕРКИ ПЛАГИАТА В УЧЕБНЫХ ОТЧЕТАХ

Чжо Зейя Наинг (ИТМО)

Научный руководитель – кандидат технических наук, доцент Горлушкина Н.Н.
(ИТМО)

Введение. Для повышения точности систем проверки плагиата предлагается использовать гибридный подход, объединяющий TF-IDF, BERT. Этот подход позволит учитывать как синтаксическое сходство текстов, так и их семантическое содержание.

Цель исследования. Оптимизация систем проверки плагиата путем использования современных методов семантического анализа и алгоритмов машинного обучения [1].

Задачи исследования

1) провести анализ существующих систем проверки плагиата и выявить их основные недостатки,
2) исследовать методы семантического анализа (BERT) и их применение в задачах сравнения текстов,
3) оптимизация алгоритмы TF-IDF и косинусного сходства для базового сравнения текстов,
4) интегрировать семантический анализ для учета контекста и смыслового содержания текстов,
5) оптимизировать хранение и обработку данных в MySQL для повышения производительности системы,
6) провести тестирование системы и оценить её точность и эффективность.

Основная часть.

В работе используются Python, Django, MySQL, NLTK, Scikit-learn, BERT. TF-IDF обеспечивает базовую векторизацию текста, а семантический анализ позволяет находить сходства между текстами, даже если они перефразированы [2], [3].

Основные этапы реализации

1) предварительная обработка текстов: удаление стоп-слов, лемматизация и нормализация данных,
2) векторизация текстов: применение TF-IDF для преобразования текста в числовые векторы,
3) семантический анализ: использование моделей BERT для выявления семантических сходств,
4) алгоритм сравнения: вычисление косинусного сходства между векторами текстов,
5) оптимизация хранения данных: настройка MySQL для быстрого доступа и обработки больших объемов текстовых данных,
6) отчетность: создание детализированных отчетов с выделением заимствованных фрагментов и процентом сходства.

Ожидаемые результаты

− повышение точности проверки за счет учета семантического анализа.
− уменьшение количества ложноположительных и ложноотрицательных результатов.
− оптимизация времени обработки больших объемов текстов.
− удобный и адаптивный пользовательский интерфейс.

Выводы. Использование методов семантического анализа и машинного обучения позволяет значительно повысить точность систем проверки плагиата. Оптимизированная система

будет полезна для образовательных учреждений, обеспечивая надежный инструмент для поддержки академической честности и оригинальности учебных работ.

Список использованных источников:

1. Проверка на плагиат научных работ: методы и значимость https://treeofbonsai.ru/wiki/proverka-na-plagiat-naucnyx-rabot-metody-i-znacimost/
2. Как работает антиплагиат: алгоритмы, этапы и технологии выявления заимствований https://antiplagiat.live/blog/vsyo-ob-antiplagiate/kak-proveryaetsya-antiplagiat
3. ТОП-15 нейросетей для проверки на плагиат в 2025 году — ТОП рейтинг на DTF https://dtf.ru/top-smm/3380651-top-15-neirosetei-dlya-proverki-na-plagiat-v-2025-godu