Рецензия на выпускную квалификационную работу бакалавра
Абубакирова Азата Радисовича
на тему
«Автоматизация сверки и устранения дубликатов в персональных данных»

Выпускная квалификационная работа бакалавра, представленная Абубакировым А.Р., посвящена исследованию и реализации методов автоматизированного поиска и фильтрации дубликатов в персональных данных с целью повышения качества этих данных. Для проведения работы используются данные из Медицинского информационно-аналитического центра (МИАЦ), содержащие записи с ФИО и датой рождения пациентов. Данная тема является актуальной, так как персональные данные зачастую вносятся в информационные системы вручную, что может привести к опечаткам и ошибкам. Автоматизированное обнаружение и исправление ошибок может значительно облегчить и улучшить работу с подобными данными.

Выпускная квалификационная работа содержит 30 страниц, состоит из пяти глав, а также разделов с введением, постановкой задачи и заключением. В ходе работы для достижения цели автор решает следующие задачи: анализ существующих подходов и решений, анализ и выявление типов ошибок в данных, разработка методов для устранения выявленных ошибок, разработка алгоритма нахождения дубликатов, применение алгоритма на реальных данных и анализ результатов, оформление разработанного программного кода в виде библиотеки.

В Главе 1 автор проводит подобный анализ литературы и существующих решений, на основе чего выявляет проблемы и обозначает способы для их решения. Глава 2 описывает обрабатываемые данные, их характеристики и свойства тестового набора данных. Глава 3 посвящена предобработке данных, подробно описывает виды возможных ошибок и способы их устранения. Здесь же представлены результаты анализа тестового набора данных по видам встречающихся ошибок. Глава 4 представляет алгоритм поиска дубликатов, предложенный автором на основе анализа предыдущего опыта. Автор рассматривает и сравнивает разные варианты предложенного алгоритма, в частности, использующие либо рекурсивный, либо линейный поиск, подбирает оптимальные параметры конфигурации алгоритма на тестовом наборе данных и применяет итоговый алгоритм на полном наборе исходных данных. Глава 5 рассказывает о разработке библиотеки на языке Python, раскрывает назначение ее модулей.

К недостаткам работы следует отнести отсутствие оценки времени выполнения различных вариантов алгоритмов и отсутствие замеров времени при тестовых запусках программ для оценки их производительности. Требование к производительности может оказаться существенным при обработке больших объемов исходных данных.

Данная выпускная работа представляет собой полноценное исследование с анализом существующих подходов к созданию систем выявления дубликатов в данных, с выбором и разработкой требуемых алгоритмов, с практической реализацией программной системы и анализом ее работы.

Считаю, что поставленная задача выполнена полностью, выпускная квалификационная работа Абубакирова А.Р. удовлетворяет требованиям к бакалаврским выпускным квалификационным работам и заслуживает оценки “отлично”, а автор – присвоения степени бакалавра.

Рецензент,
доцент кафедры КМиМС, PhD  В.В. Корхов