Министерство образования Республики Беларусь

Белорусский государственный университет

Механико-математический факультет

Кафедра веб-технологий и компьютерного моделирования

Минск, 2025

БАГРЕЙ Иван Викторович

Аннотация к дипломной работе

**Разработка веб-сервиса по распознаванию таблиц в печатных документах**

Научный руководитель:

доцент, кандидат физико-математических наук

Б. М. Дубров

АННОТАЦИЯ

Дипломная работа содержит 54 страницы, 4 картинки и 11 наименований источников в списке использованной литературы.

**Ключевые слова:** PDF-ДОКУМЕНТЫ, ТАБЛИЧНЫЕ ДАННЫЕ, FLASK, TABLE-TRANSFORMER, TESSERACT OCR, ИЗВЛЕЧЕНИЕ ТАБЛИЦ, КОНВЕРТАЦИЯ СТРАНИЦ, ВЕБ-СЕРВИС, ИНФЕРЕНС, POPPLER.

**Объект исследования** — программный инструмент, предназначенный для автоматического выделения таблиц из документов в формате PDF, преимущественно полученных путем сканирования. Этот инструмент ориентирован на обработку изображений страниц, содержащих табличные структуры, с целью извлечения информации, пригодной для дальнейшего анализа и хранения.

**Цель дипломной работы** — разработка веб-сервиса, способного без участия пользователя распознавать табличные блоки в визуальном содержимом PDFдокументов и предоставлять результат в виде структурированной информации.

**Методы исследования:** в работе применялись методы анализа существующих решений в области распознавания таблиц, практическая реализация модели Microsoft Table-Transformer, а также экспериментальное тестирование функциональности сервиса. В качестве инструментов использовались библиотеки PyTorch, pdf2image, Pillow, Tesseract OCR и Flask, обеспечившие комплексную реализацию как серверной, так и клиентской части приложения.

**Результатом** дипломной работы стал полнофункциональный веб-сервис, успешно прошедший тестирование и продемонстрировавший устойчивую работу при обработке документов различного формата и сложности. Сервис корректно извлекает таблицы из PDF-файлов, минимизируя количество ложных срабатываний и обеспечивая высокую точность при типичных сценариях использования.

**Область применения:** разработанный веб-сервис может применяться в сфере документооборота, бухгалтерского учета, архивной обработки, подготовки отчетности, научных публикаций и других областях, где требуется автоматическое извлечение табличных данных из PDF-документов.

Дипломная работа выполнена автором самостоятельно.

# АНАТАЦЫЯ

Дыпломная работа змяшчае 54 старонкі, 4 ілюстрацыі і 11 найменняў крыніц у спісе выкарыстанай літаратуры.

**Ключавыя словы:** PDF-ДАКУМЕНТЫ, ТАБЛІЧНЫЯ ДАДЗЕНЫЯ, FLASK, TABLE-TRANSFORMER, TESSERACT OCR, ВЫДЗЯЛЕННЕ ТАБЛІЦ, КАНВЕРТАЦЫЯ СТАРОНОК, ВЭБ-СЭРВІС, ІНФЕРЭНС, POPPLER.

**Аб’ект даследавання** — праграмны інструмент, прызначаны для аўтаматычнага выдзялення табліц з дакументаў у фармаце PDF, пераважна атрыманых шляхам сканавання. Гэты інструмент арыентаваны на апрацоўку выяваў старонак, што ўтрымліваюць таблічныя структуры, з мэтай вылучэння інфармацыі, прыдатнай для далейшага аналізу і захавання.

**Мэта дыпломнай работы** — распрацоўка вэб-сэрвісу, здольнага без удзелу карыстальніка распазнаваць таблічныя блокі ў візуальным змесце PDF-дакументаў і прадастаўляць вынік у структураваным выглядзе.

**Метады даследавання:** у працы выкарыстоўваліся метады аналізу існуючых рашэнняў у галіне распазнавання табліц, практычная рэалізацыя мадэлі Microsoft Table-Transformer, а таксама эксперыментальнае тэставанне функцыянальнасці сэрвісу. У якасці інструментаў прымяняліся бібліятэкі PyTorch, pdf2image, Pillow, Tesseract OCR і Flask, што забяспечылі комплексную рэалізацыю як сервернай, так і кліенцкай часткі прыкладання.

**Вынікам** дыпломнай працы стаў поўнафункцыянальны вэб-сэрвіс, паспяхова пратэставаны і які паказаў устойлівую працу пры апрацоўцы дакументаў рознага фармату і складанасці. Сэрвіс карэктна выдзяляе табліцы з PDF-файлаў, мінімізуючы колькасць ілжывых спрацоўванняў і забяспечваючы высокую дакладнасць у тыпавых сцэнарах выкарыстання.

**Вобласць ужывання:** распрацаваны вэб-сэрвіс можа быць выкарыстаны ў сферы дакументаабароту, бухгалтарскага ўліку, архіўнай апрацоўкі, падрыхтоўкі справаздачнасці, навуковых публікацый і іншых галінах, дзе патрабуецца аўтаматызаванае выдзяленне таблічных дадзеных з PDF-дакументаў. Дыпломная работа выканана аўтарам самастойна.

# ANNOTATION

The thesis comprises 54 pages, includes 4 figures, and references 11 sources in the bibliography.

**Keywords**: PDF DOCUMENTS, TABULAR DATA, FLASK, TABLETRANSFORMER, TESSERACT OCR, TABLE EXTRACTION, PAGE

CONVERSION, WEB SERVICE, INFERENCE, POPPLER.

The **object of the research** is a software tool designed for the automatic extraction of tables from PDF documents, primarily those obtained through scanning. This tool focuses on processing page images that contain tabular structures in order to extract information suitable for further analysis and storage.

**The goal of the thesis** is to develop a web service capable of recognizing tabular blocks within the visual content of PDF documents without user involvement and providing the output in a structured format.

**Research methods** include the analysis of existing solutions in the field of table recognition, practical implementation of the Microsoft Table-Transformer model, and experimental testing of the service’s functionality. The development employed libraries such as PyTorch, pdf2image, Pillow, Tesseract OCR, and Flask to implement both the server-side and client-side components.

As a **result** of the thesis, a fully functional web service was developed and successfully tested, demonstrating stable performance when processing documents of various formats and complexities. The service accurately extracts tables from PDF files, minimizing false positives and ensuring high precision under typical usage scenarios.

**Application arrea**: the developed web service can be applied in document workflow systems, accounting, archival processing, report generation, scientific publishing, and other fields requiring automated extraction of tabular data from PDF documents.

The thesis was completed independently by the author.