МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ
Кафедра интеллектуальных систем

Аннотация к дипломной работе

РАСПОЗНАВАНИЕ И ПРЕОБРАЗОВАНИЕ ТАБЛИЧНЫХ ДАННЫХ

Басацкая Анастасия Олеговна

Научный руководитель: кандидат физико-математических наук, доцент Е.И. Козлова

Минск, 2025

РЕФЕРАТ

Дипломная работа: 45 страниц, 13 рисунков, 1 приложение, 14 использованных литературных источников.

РАСПОЗНАВАНИЕ ТАБЛИЦ, ПРЕДОБРАБОТКА ДАННЫХ, СВЕРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ, КОМПЬЮТЕРНОЕ ЗРЕНИЕ, ОПТИЧЕСКОЕ РАСПОЗНАВАНИЕ СИМВОЛОВ, DETECTRON2.

Объект исследования: табличные данные в документах формата PDF и изображениях, включая таблицы с явными и неявными границами, а также их автоматическое детектирование и преобразование в редактируемые форматы (CSV, XML).

Цель дипломной работы - разработка системы для автоматического детектирования и извлечения таблиц из документов в форматах PDF и изображений с сохранением их структуры и преобразованием данных в форматы, пригодные для дальнейшего анализа и редактирования.

В результате исследования изучены существующие методы распознавания и извлечения данных из таблиц, различные подходы, технологии и инструменты, которые используются для автоматического распознавания таблиц, разработана система на языке Python, способная детектировать таблицы в PDF-документах и изображениях, извлекать текстовые данные из таблиц с использованием OCR (Tesseract), сохранять извлеченные данные в структурированных форматах (CSV, XML), проведено сравнение двух подходов: метода компьютерного зрения (OpenCV), нейросетевого подхода (Detectron2).

Методы исследования: теоретические: сравнительный анализ, качественный анализ, функциональный анализ; практические: разработка и реализация собственного алгоритма.

Область применения: финансовый сектор, страховые компании, документооборот, научные исследования.

ABSTRACT

Diploma thesis: 45 pages, 13 figures, 1 appendix, 14 literature sources used.

TABLE DETECTION, DATA PREDOCTRATION, CONVOLUTIONAL NEURAL NETWORKS, COMPUTER VISION, OPTICAL CHARACTER RECOGNITION, DETECTRON2.

The object of the research is tabular data in PDF documents and images, including tables with explicit and implicit boundaries, as well as their automatic detection and conversion into editable formats (CSV, XML).

The aim of the thesis is to develop a system for automatic detection and extraction of tables from documents in PDF and image formats, preserving their structure and converting the data into formats suitable for further analysis and editing.

As a result of the research the existing methods of recognising and extracting data from tables, different approaches, technologies and tools that are used for automatic table recognition were studied, a system in Python language was developed, capable of detecting tables in PDF documents and images, extracting text data from tables using OCR (Tesseract), saving the extracted data in structured formats (CSV, XML), comparison of two approaches: computer vision method (OpenCV), neural network approach (Detectron2).

Research methods: theoretical: comparative analysis, qualitative analysis, functional analysis; practical: development and implementation of a custom algorithm.

Field of application: financial sector, insurance companies, document management, and scientific research.

РЭФЕРАТ

Дыпломная праца: 45 старонак, 13 малюнкаў, 1 дадатак, 14 выкарыстаных крыніц.

РАСПАЗНАВАННЕ ТАБЛІЦ, ПРАДАПРАЦОЎКА ДАНЫХ, ЗВЯРТОЧНЫЯ НЕЙРОННЫЯ СЕТКІ, КАМП'ЮТЭРНЫ ЗРОК, АПТЫЧНАЕ РАСПАЗНАВАННЕ СІМВАЛАЎ, DETECTRON2.

Аб'ект даследавання: Таблічныя даныя ў дакументах фармату PDF і выявах, уключаючы табліцы з відавочнымі і нявідавочнымі межамі, а таксама іх аўтаматычнае выяўленне і пераўтварэнне ў рэдагавальныя фарматы (CSV, XML).

Мэта дыпломнай працы: распрацоўка сістэмы для аўтаматычнага выяўлення і здабывання табліц з дакументаў у фарматах PDF і выяваў з захаваннем іх структуры і пераўтварэннем даных у фарматы, прыдатныя для далейшага аналізу і рэдагавання.

Вынікі даследавання: вывучаны існуючыя метады распазнавання і здабывання даных з табліц, розныя падыходы, тэхналогіі і інструменты, якія выкарыстоўваюцца для аўтаматычнага распазнавання табліц. Распрацавана сістэма на мове Python, здольная выяўляць табліцы ў PDF-дакументах і выявах, здабываць тэкставыя даныя з табліц з выкарыстаннем OCR (Tesseract), захоўваць здабытыя даныя ў структураваных фарматах (CSV, XML). Праведзена параўнанне двух падыходаў: метаду камп'ютарнага зроку (OpenCV) і нейрасецевага падыходу (Detectron2).

Метады даследавання: тэарэтычныя: параўнальны аналіз, якасны аналіз, функцыянальны аналіз; практычныя: распрацоўка і рэалізацыя ўласнага алгарытму.

Вобласць прымянення: фінансавы сектар, страхавыя кампаніі, дакументаабарот, навуковыя даследаванні.