УДК 004.75

О.С. Сівак
О. С. Городецька

КОМП'ЮТЕРНА СИСТЕМА РОЗПОДІЛЕНОЇ ОБРОБКИ ТА ЗБЕРІГАННЯ ЕЛЕКТРОННИХ ДОКУМЕНТІВ

Вінницький національний технічний університет

Анотація

У роботі розроблено комп'ютерну систему обробки електронних документів, що забезпечує завантаження, OCR-розпізнавання, автоматичне виділення метаданих, класифікацію документів та пошук за змістом. Система підтримує обробку текстових PDF, сканованих PDF та зображень, використовує OCR для витягування тексту, а також AI-модуль для визначення типу документа, теми, ключових полів і побудови логічної структури зберігання. Реалізовано пакетне завантаження файлів, чергу обробки, повторну перевірку документів у різних режимах точності та аналітичний модуль для оцінювання продуктивності системи

Ключові слова: комп'ютерна система, розподілена обробка**,** електронні документи, OCR, штучний інтелект, класифікація документів, пошук.

Abstract

A computer system for processing electronic documents has been developed. It provides file upload, OCR recognition, automatic metadata extraction, document classification, and content-based search.The system supports processing of text-based PDFs, scanned PDFs, and images. It uses OCR to extract text, as well as an AI module to determine the document type, topic, key fields, and to build a logical storage structure.Batch file upload, a processing queue, re-validation of documents in different accuracy modes, and an analytical module for evaluating system performance have been implemented.

Keywords: computer system, distributed processing, electronic documents, OCR, artificial intelligence, document classification, search.

Вступ

У сучасних інформаційних системах значна частина даних надходить у вигляді неструктурованих або слабкоструктурованих документів: PDF-файлів, сканованих копій, зображень, виписок, заяв, лабораторних робіт та інших документів. Ручне сортування, аналіз і пошук таких матеріалів є повільним і помилковим процесом. Тому актуальним є створення комп'ютерної системи, що дозволяє автоматизувати етапи завантаження, розпізнавання, класифікації та подальшого пошуку документів.

Метою роботи є розробка інтелектуальної системи обробки електронних документів, яка забезпечує автоматичне витягування тексту з різних типів файлів, виділення метаданих, класифікацію, логічне групування та швидкий пошук документів.

Архітектура системи

Система побудована за клієнт-серверною архітектурою та включає три основні рівні: рівень взаємодії з користувачем, серверний рівень API та рівень фонової обробки документів. На рівні інтерфейсу користувач завантажує один або декілька файлів, переглядає результати обробки, виконує пошук та аналізує статистику. Серверний рівень приймає файли, перевіряє їх формат, зберігає оригінали, створює записи в базі даних та ставить документи у чергу. Фонові worker-процеси послідовно обробляють документи: виконують OCR за потреби, запускають AI-аналіз, формують метадані та повертають результат у сховище системи.

На відміну від простих систем зберігання файлів, запропоноване рішення підтримує обробку як текстових PDF, так і сканованих документів та зображень. Для текстових PDF використовується швидке пряме витягування тексту, а для сканів і зображень — OCR-розпізнавання. Після цього AI-модуль визначає тип документа, тему, назву, логічну папку, короткий опис та інші поля, необхідні для подальшого пошуку і сортування.

Функціональні ролі в системі

Система побудована на взаємодії кількох функціональних компонентів, кожен з яких виконує окрему роль у процесі обробки документів.

Користувач системи є джерелом документів та ініціатором процесу обробки. Користувач завантажує один або декілька файлів через вебінтерфейс системи, після чого має можливість переглядати результати автоматичного аналізу, виконувати пошук по документах та переглядати аналітичну інформацію щодо їх обробки.

Frontend-інтерфейс забезпечує взаємодію користувача із системою. Через нього здійснюється завантаження файлів, перегляд метаданих документів, запуск повторної перевірки документів та робота з системою пошуку.

Серверний API-модуль виконує перевірку форматів файлів, збереження оригіналів документів у файловому сховищі та створення відповідних записів у базі даних. Після цього документ отримує статус queued та передається до системи фонової обробки.

Worker-модуль обробки відповідає за виконання основної логіки аналізу документа. Worker отримує документ із черги та виконує визначення типу файлу. Для текстових PDF застосовується пряме витягування тексту, а для сканованих документів або зображень використовується OCR-розпізнавання.

AI-модуль аналізу документів виконує інтелектуальну обробку отриманого тексту. На цьому етапі визначаються тип документа, його тематика, формується короткий опис, логічна назва документа та структура папок для подальшого зберігання.

Аналітичний модуль відповідає за оновлення пошукових індексів, відображення результатів у dashboard системи та формування статистики обробки документів.

Алгоритм взаємодії та життєвий цикл заявки

Центральним елементом логіки системи є алгоритм обробки документа, який забезпечує автоматичний перехід документа від моменту завантаження користувачем до отримання структурованих метаданих та включення його до системи пошуку.

Процес обробки документа складається з таких етапів:

Завантаження документа.

Користувач завантажує один або декілька файлів через вебінтерфейс системи. Frontend передає файл на серверний API-модуль.

Перевірка та збереження файлу.

API перевіряє формат документа, після чого оригінальний файл зберігається у файловому сховищі. У базі даних створюється запис документа зі статусом queued.

Фонова обробка документа.

Worker-процес отримує документ із черги обробки та переводить його у стан processing.

Визначення типу документа.

Система аналізує тип файлу. Якщо документ є текстовим PDF, виконується пряме витягування тексту. Якщо документ є сканованим PDF або зображенням, запускається OCR-розпізнавання.

Інтелектуальний аналіз документа.

Після отримання тексту запускається AI-модуль, який виконує класифікацію документа, визначає його тематику, генерує короткий опис, формує назву документа та визначає логічну папку зберігання.

Збереження результатів.

Отримані метадані зберігаються у базі даних, документ отримує статус done, після чого результати стають доступними для пошуку та перегляду.

Схему процесу обробки документа наведено на рис. 1.

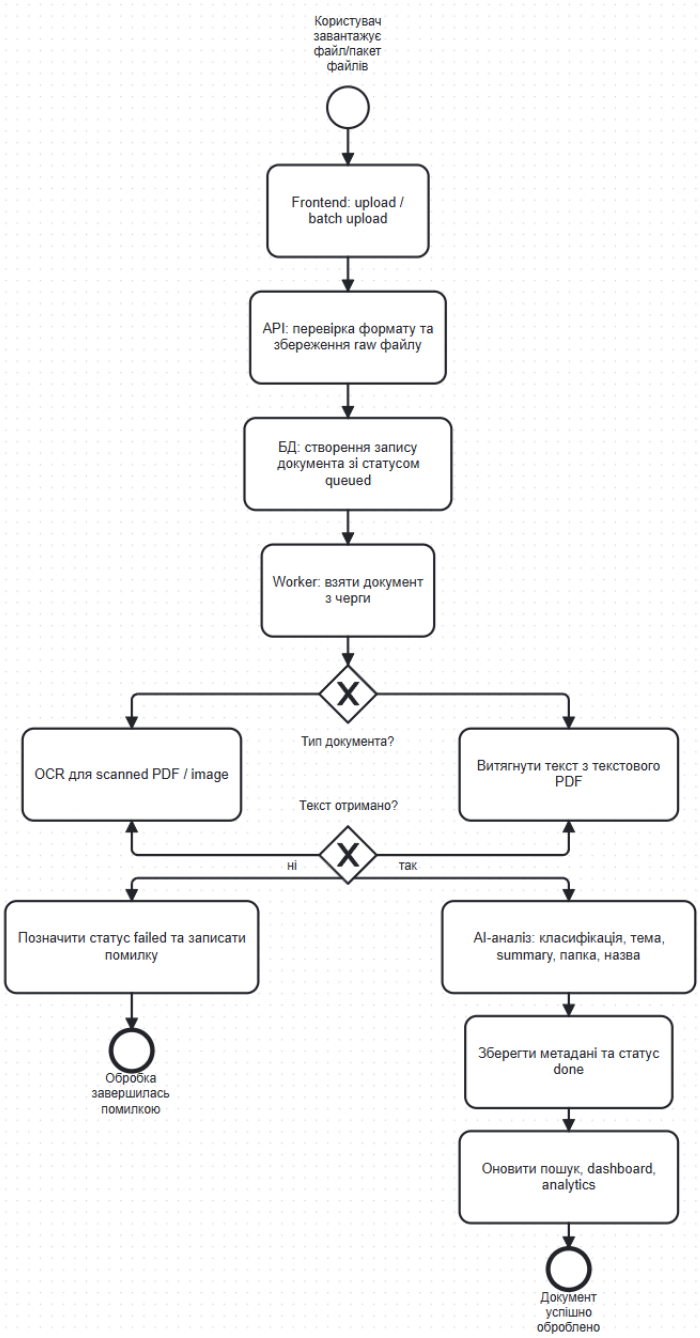

Рис. 1. BPMN-діаграми процесу обробки електронних документів

BPMN-діаграма відображає повний процес проходження документа через систему. Починаючи із завантаження файлу користувачем, документ проходить етап перевірки, потрапляє до черги обробки та обробляється worker-процесом. Залежно від типу документа виконується або пряме витягування тексту, або OCR-розпізнавання. Після отримання тексту запускається AI-аналіз, який визначає тип документа,

його тему та формує метадані. Результати обробки зберігаються у базі даних та стають доступними для пошуку і аналітики.

Висновки

У ході виконання роботи було розроблено інтелектуальну систему автоматичної обробки електронних документів, яка поєднує механізми завантаження файлів, OCR-розпізнавання, інтелектуальний аналіз тексту та систему пошуку.

На основі проведеного аналізу предметної області та реалізації програмного рішення можна зробити такі висновки.

Розроблена система реалізує клієнт-серверну архітектуру, що включає вебінтерфейс користувача, серверний API-модуль та систему фонової обробки документів. Така архітектура забезпечує масштабованість та можливість обробки великої кількості документів.

Запропонований алгоритм обробки документів дозволяє автоматизувати повний цикл роботи з документами: від моменту завантаження файлу до отримання структурованих метаданих та включення документа до системи пошуку.

Використання OCR-технологій забезпечує можливість роботи не лише з текстовими PDF-файлами, але і зі сканованими документами та зображеннями.

Інтеграція методів штучного інтелекту дозволяє автоматично класифікувати документи, визначати їх тематику, генерувати короткий опис та формувати логічну структуру зберігання. Це значно підвищує ефективність пошуку та організації електронних документів.

Результати роботи підтверджують доцільність використання інтелектуальних методів обробки документів для автоматизації систем електронного документообігу та інформаційного пошуку.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Tesseract OCR Engine Documentation [Електронний ресурс]. – Режим доступу: https://tesseract-ocr.github.io/ (дата звернення: 14.03.2026).
2. Google Cloud Vision API Documentation [Електронний ресурс]. – Режим доступу: https://cloud.google.com/vision/docs (дата звернення: 14.03.2026).
3. OpenAI API Documentation [Електронний ресурс]. – Режим доступу: https://platform.openai.com/docs (дата звернення: 14.03.2026).
4. Smith R. An Overview of the Tesseract OCR Engine // Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). – 2007. – P. 629–633.
5. Kyrylashchuk S., Horodetska O., Voitsekhovska O., Zakharchenko S. Order Forecasting System for Vehicles Based on Previous Statistics Requests // Lecture Notes in Data Engineering and Communications Technologies, vol. 219. – Springer, 2024. – P. 116–131.

Сівак Олександр Сергійович − студент групи 1СП-22б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: osivak472@gmail.com

Городецька Оксана Степанівна − кандидат технічних наук, доцент кафедри обчислювальної техніки Вінницького національного технічного університету, Вінниця.

Sivak Oleksandr – student of group 1SP-22b, faculty of information technologies and computer engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: osivak472@gmail.com

Horodetska Oksana – Candidate of Technical Sciences, Assistant Professor of the Computer Techniques Chair Vinnytsia National Technical University , Vinnytsia.