

СИСТЕМА РОЗПІЗНАВАННЯ І КЛАСИФІКАЦІЇ ФІНАНСОВИХ ДОКУМЕНТІВ

Анотація. Дослідження присвячене автоматизації обробки документів, зокрема товарно-транспортних накладних (ТТН), для цифровізації підприємств. Основною метою є створення гнучкої програмної системи, яка автоматично витягує реквізити з документів різних форматів із високою точністю. Система використовує класичні методи обробки зображень, оптичне розпізнавання символів (OCR) через бібліотеку *Nicomsoft OCR*, а також шаблонний підхід до семантичного аналізу тексту. Дані проходять нормалізацію, забезпечуючи стандартизоване зберігання інформації. Реалізація виконана мовою *Java*, із використанням *Swing* для UI, *Apache POI* для роботи з документами та *Firebird SQL* для збереження даних. Завдяки модульній архітектурі система легко інтегрується в корпоративні інформаційні платформи, підвищує ефективність обробки документів та може бути масштабована для інших типів документації, що сприяє цифровій трансформації підприємств.

Ключові слова: автоматизація обробки документів, оптичне розпізнавання символів (OCR), семантичний аналіз тексту, цифрова трансформація підприємств, модульна архітектура системи.

Цифрова трансформація підприємств значно змінює методи ведення документації, особливо в сферах торгівлі та логістики. Незважаючи на активне впровадження електронного документообігу, значна частина процесів все ще пов'язана із паперовими накладними, що вимагає ручного внесення даних до комп'ютерних систем. Це призводить до значних витрат часу, підвищеного ризику помилок та зростання витрат на адміністрування.

Автоматизація обробки товарно-транспортних накладних (ТТН) потребує інтеграції сучасних методів комп'ютерного бачення та оптичного розпізнавання символів (OCR). Враховуючи різноманітність форматів ТТН, ключовим завданням є розробка адаптивної системи, здатної коригувати викривлення та нерівномірність освітлення під час сканування або фотографування документів. Застосування методів попередньої обробки, таких як коригування контрасту, видалення артефактів та покращення роздільної здатності, дозволяє підвищити точність подальшого текстового аналізу.

Традиційні OCR-системи ефективні у розпізнаванні стандартних шрифтів і чітких документів, проте в умовах змінних форматів ТТН можуть виникати труднощі з інтерпретацією нестандартних символів, розпізнаванням числових полів або коректним відокремленням розділових знаків. Для усунення таких проблем доцільним є застосування додаткових алгоритмів семантичного аналізу тексту та валідації числових реквізитів.

Запропонована система інтегрує технологію OCR з шаблонним підходом до обробки даних, що дозволяє швидко адаптувати розпізнавання під різні формати документів. Реалізація на базі *Java* та *Firebird SQL* забезпечує високу продуктивність, гнучкість інтеграції з корпоративними інформаційними системами та масштабованість рішення. Впровадження автоматизованої обробки ТТН сприяє оптимізації процесів управління поставаннями, скороченню витрат та підвищенню точності бухгалтерських розрахунків, що є важливим аспектом цифрової трансформації підприємств.

Автоматизація обробки документів є важливим завданням для цифровізації сучасних підприємств. Особливої актуальності це набуває в тих галузях, в яких присутній великий обсяг паперової документації, зокрема це галузі логістики, торгівлі і виробництва. Ручне введення даних із накладних займає багато часу, супроводжується помилками та вимагає залучення додаткового персоналу. Основною метою цього стартапу є створення гнучкої програмної системи, здатної автоматично витягувати реквізити з товарно-транспортних накладних (ТТН) різних форматів з високою точністю і зберігати результати у структурованому вигляді для подальшої обробки.

Основною частиною проекту є класичні етапи обробки вхідного зображення: вирівнювання, покращення контрастності, фільтрація шумів, видалення фону та зайвих артефактів. Данні операції покращують якість зображення для підвищення точності подальшого розпізнавання тексту. Система повинна бути орієнтованою на обробку зображень знятих на мобільні пристрої. Враховуючи велику різноманітність форматів надходження документів, система повинна бути здатна самостійно адаптувати зображення до прийнятного формату для аналізу. Саме тому було реалізовано механізми автоматичного визначення орієнтації документа, а також коригування спотворень зображень, знятих під кутом, які часто виникають при ручній зйомці.

Після етапу попередньої обробки зображення застосовується технологія оптичного розпізнавання символів (OCR). Обраною бібліотекою для даної задачі стала Nicomsoft OCR, яка забезпечує стабільне розпізнавання кирилиці, зокрема української мови. Важливою перевагою є можливість локального використання, що дозволяє забезпечити збереження конфіденційності оброблених даних без потреби у зовнішніх сервісах. Бібліотека має зручний інтерфейс для інтеграції в Java-застосунки та підтримує налаштування зон розпізнавання, що дозволяє визначати, які саме ділянки документа потрібно аналізувати. Інструменти Nicomsoft OCR дозволяють точно контролювати процес розпізнавання: задавати мови, формати символів, зональні обмеження, що критично важливо при роботі з багатосторінковими та неструктурованими документами. Висока гнучкість API цієї бібліотеки дозволяє зменшити час розробки та підвищити адаптивність до нестандартних форматів документів. На Рис.1. показано концептуальну архітектуру програми Nicomsoft, яка складається з трьох основних компонентів: тексту, зображення та OCR.

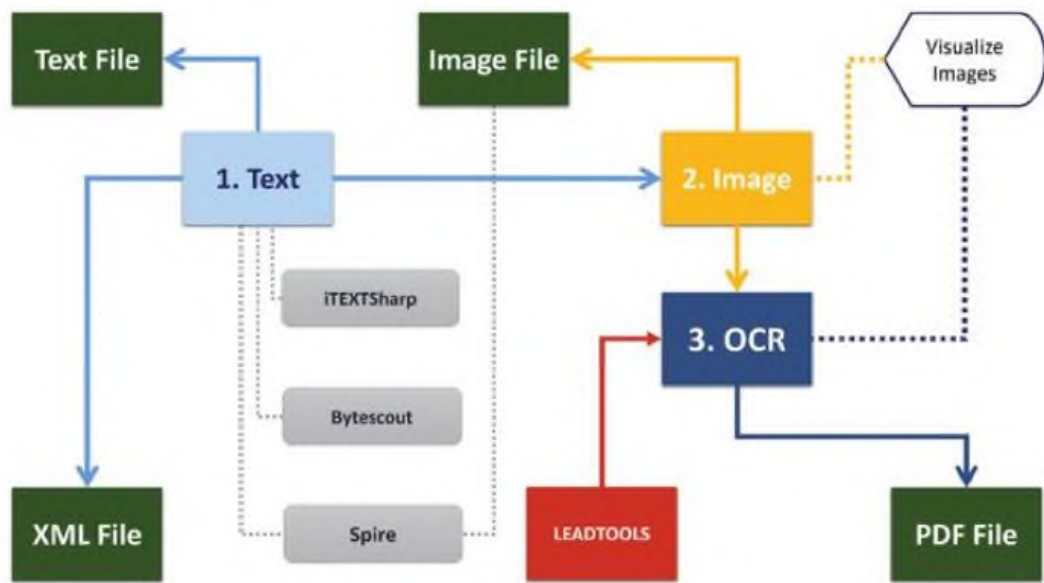


Рис. 1. Концептуальна архітектура OCR

Наступним етапом є семантичний аналіз розпізнаного тексту. Для реалізації цієї частини було запроваджено шаблонний підхід: для кожного нового постачальника створюється шаблон із вказівкою координат ключових полів документа. Завдяки такому рішення при повторному надходженні схожого документа система автоматично застосовує відповідний шаблон і витягує необхідні дані. Це дозволяє значно підвищити швидкість і точність обробки. У майбутньому така система може бути розширена у бік машинного навчання, тобто дозволити їй самостійно створювати шаблони на основі попереднього досвіду без участі користувача. Враховуючи обмеження поточного етапу реалізації, шаблонний метод виступає перехідним етапом між повністю ручною і повністю автоматизованою системою.

Дані, які отримали внаслідок розпізнавання, проходять додаткову нормалізацію. Тут мається на увазі приведення числових значень, дат, одиниць виміру до єдиного формату. Наприклад, ціна «1 000,00 грн» та «1000.00 UAH» повинні зберігатися у базі даних у єдиному форматі. Також проводиться обробка розділових знаків, пробілів, символів валюти, форматів дати. У разі помилкового розпізнавання форматів система надає оператору можливість перевірити відскановані дані та виправити їх вручну. Але для подальшого вдосконалення система повинна мати змогу попереджати користувача, або автоматично виправляти типові неточності спираючись на вбудовані словники та шаблони даних. Стандартизація запису ключових реквізитів відіграє вирішальну роль при подальшому аналітичному аналізі, звітуванні та контролі бухгалтерського обліку. Також у нормалізації використовуються правила перетворення регістрів тексту, стандартизація позначень одиниць виміру (наприклад, шт. → штуки) та валют (грн, UAH), що дозволяє уникати неоднозначності в подальшому аналізі.

1 прим. – вантажовідправнику
2 прим. – вантажоодержувачу
3 і 4 прим. – перевізнику

ТОВАРНО-ТРАНСПОРТНА НАКЛАДНА

Формат № 1-ТН затверджено наказом Міністерства
України 05.12.2013 р. №983

№ 241901 від 20 р.

Автомобіля / Пратін/навантажувача / Вид перевезення /
(марка, модель, рік реєстрації номер) (марка, модель, рік, реєстраційний номер)

Автомобільний перевізник Водій
(назва підприємства П.І.Б.) (П.І.Б., місце проживання водія)

Вантажовідправник /
(назва підприємства, місце реєстрації П.І.Б., місце проживання)

Вантажоодержувач /
(назва підприємства, місце реєстрації П.І.Б., місце проживання)

Пункт навантаження / Пункт розвантаження /
(назва підприємства) (назва підприємства)

Перезадресування вантажу /
(назва підприємства, місце реєстрації П.І.Б., місце проживання вантажоодержувача, П.І.Б., місце та місце відправлення водія)

Відпуск за довіреністю вантажоодержувача: серія № 04 від 20 р. (назва підприємства)

Вантаж наданий для перевезення у стані, що / вантажам перевезення відомостей вантажів, номер габаритів (за наявності)
кількість місць / масою бруто, т / отримав водій/експедитор / (П.І.Б., місце, підпис)

Бухгалтер (відповідає особа вантажовідправника) / Відпуск дозволено / (П.І.Б., місце, підпис, печатка)

Усього відпущено на загальну суму / у т.ч. ПДВ /
(сума, у гривнях і копійках) (сума, у гривнях і копійках)

Супровідні документи на вантаж /
Транспортні послуги, які надаються автомобільним перевізником /

ВІДОМОСТІ ПРО ВАНТАЖ

№ з/п	Найменування вантажу (номер автотранспорту, у разі перевезення побутових вантажів, клас кобеселених рішень, до якого віднесено вантаж)	Одиниця виміру	Кількість вантажів	Ціна без ПДВ (за відомостями) грн.	Загальна сума з ПДВ, грн.	Відпуск вантажів	Документи з вантажів	Маса бруто, т
1	2	3	4	5	6	7	8	9
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								
31								
32								
33								
34								
35								
36								
37								
38								
39								
40								
41								
42								
43								
44								
45								
46								
47								
48								
49								
50								
51								
52								
53								
54								
55								
56								
57								
58								
59								
60								
61								
62								
63								
64								
65								
66								
67								
68								
69								
70								
71								
72								
73								
74								
75								
76								
77								
78								
79								
80								
81								
82								
83								
84								
85								
86								
87								
88								
89								
90								
91								
92								
93								
94								
95								
96								
97								
98								
99								
100								

Дані (назва підприємства, місце реєстрації П.І.Б., місце, підпис, печатка)

Примітка (назва підприємства, місце реєстрації П.І.Б., місце, підпис)

Зам. вантажів експедитор (П.І.Б., місце, підпис)

Примітка (назва підприємства, місце реєстрації П.І.Б., місце, підпис, печатка)

ВАНТАЖНО-РОВАНТАЖУВАЛЬНІ ОПЕРАЦІЇ

Операція	Маса бруто, т	Час (гол., хв.)		Підпис відповідального особи	
		прибуття	відбуття		
10	11	12	13	14	15
Навантаження					
Розвантаження					

Рис.2. Приклад ТТН

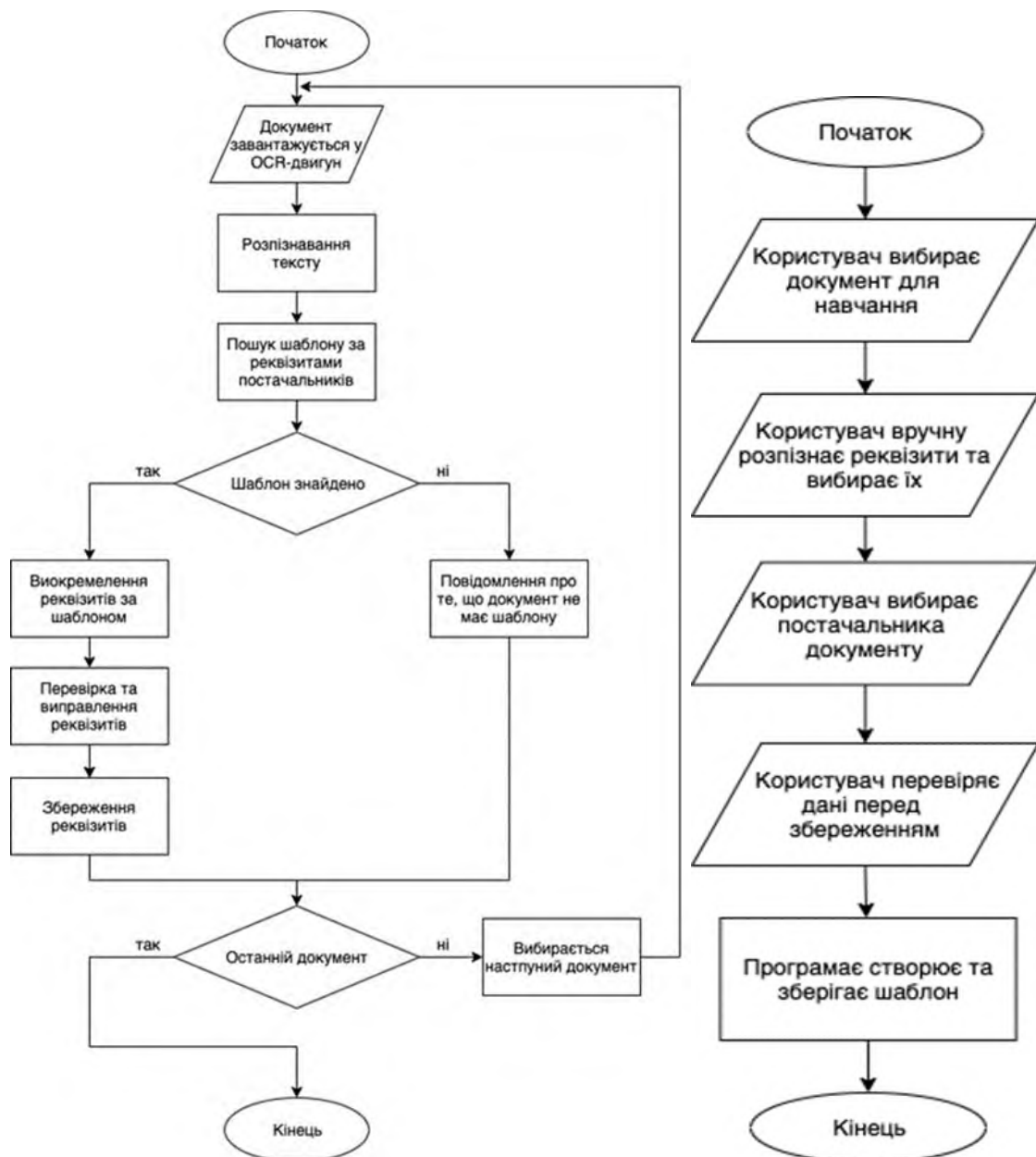
Програмна реалізація системи виконана мовою Java. Такий вибір пояснюється її платформеною незалежністю, підтримкою багатопотоковості, високою продуктивністю та великою кількістю бібліотек. Бібліотека Swing використовується для створення графічного інтерфейсу користувача. Вона дозволяє будувати зручні вікна з формами, кнопками, полями введення та іншим функціоналом. Для обробки файлів є бібліотека Apache Commons IO, вона спрощує роботу з каталогами, потоками та файлами. Apache POI використовується для формування таблиць та звітів, вони можуть бути експортовані у формат Excel для подальшого аналізу. Log4j забезпечує системне логування, що необхідне для відстеження помилок і подій у програмі. Завдяки такому пакету бібліотек, кожен етап обробки, від отримання документа до виводу обробленої інформації, реалізується прозоро та стабільно.

Оброблені та нормалізовані дані зберігаються у базі даних Firebird SQL. Ця СУБД з відкритим кодом, з низькими системними вимогами, швидкою роботою на малих об'ємах і простотою адміністрування. Вона не потребує окремого сервера, що робить її придатною для вбудованих або локальних рішень. База підтримує тригери, збережені процедури, механізми забезпечення цілісності даних. Також дозволяє реалізовувати складні логічні залежності між таблицями за допомогою SQL-запитів. Такий підхід дозволяє спростити логіку роботи з даними, покращити контроль за змінами у БД та забезпечити надійність довготривалого зберігання інформації. Система підтримує транзакції – будь-яка група операцій виконується як єдине ціле, що особливо важливо при збереженні змін з декількох джерел одночасно або під час аварійного завершення роботи програми.

Проаналізувавши вимоги та функціонал майбутньої системи було прийнято рішення впровадити роботу системи у двох режимах: інтерфейсний - використовується для навчання та створення шаблонів бланків, та автоматичний - створений для автоматичного розпізнавання реквізитів у документах на основі існуючих шаблонів бланків для них. Створений шаблон зберігається у базі даних. Розпізнавання інших документів, його типу бланку вже буде доступно у автоматичному режимі. Щодо прецеденту «перенавчання» бланку - воно проходить абсолютно так само як і навчання. Програма сама оновить дані щодо шаблону у базі даних. Таку процедуру навчання потрібно провести для усіх типів бланків, які будуть проходити розпізнавання у автоматичному режимі.

Алгоритм роботи у інтерфейсному режимі представлений у Рис. 3.

Алгоритм роботи у автоматичному режимі представлений на Рис. 4.



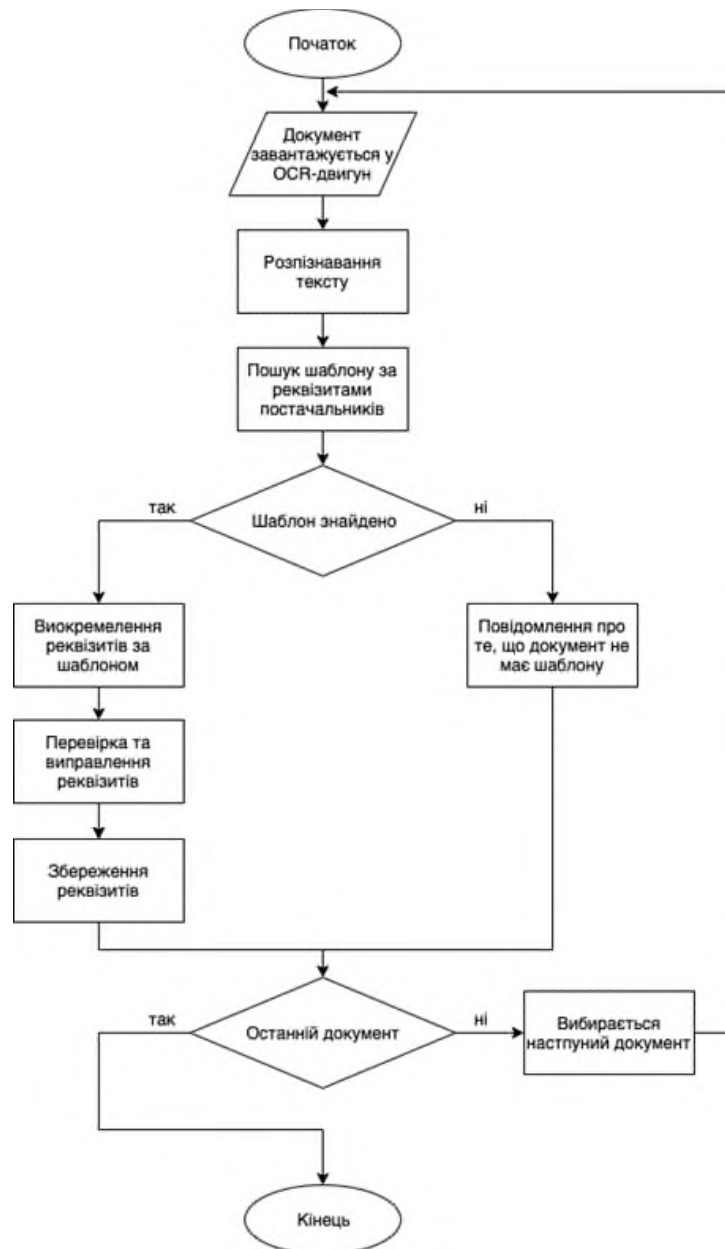


Рис.3. Алгоритм роботи у інтерфейсному режимі

Рис.4. Алгоритм роботи у автоматичному режимі

Під час роботи у автоматичному режимі, на початку програма сама сканує базу даних на предмет нерозпізнаних документів. Або відразу завантажує потрібний документ, якщо є параметр `doc_pi_import_id`. Після цього починає їх розпізнавати. Після розпізнавання тексту, потрібно знайти шаблон документу. Для цього у базі даних є дані усіх постачальників. У розпізаному документі шукають атрибути одного із збережених постачальників. Такими атрибутами є електронна пошта, номер телефону та назва. Після того як шаблон знайдений, за його допомогою виокремлюються потрібні реквізити. Перед збереженням, реквізити перевіряються. Наприклад, що реквізит суми не містить в собі літери. А всі дати переводяться до одного формату. Всі перевірені реквізити програма відразу зберігає у базі даних. При помилках чи не вдалому розпізнаванні, наприклад не розізнався певний реквізит чи не було знайдено шаблону для документа, програма позначає це в Log-файлі.

У роботі в автоматичному режимі відсутній інтерфейс, програма лише записує результати в Log-файл та відображає їх в консолі. Саме можливість запускати програму та задавати їй усі параметри із консолі дозволяють їй бути незалежною та легко-вбудованою у інші системи.

Таким чином, система, побудована на основі технологій Java, Nicosoft OCR, Swing, Firebird SQL, Apache Commons IO та Apache POI, створює гнучке та ефективне рішення для автоматизації обробки ТТН. Вона легко адаптується до різних форматів документів, забезпечує точність розпізнавання і високу продуктивність роботи, а її модульна структура відкриває можливості для подальшого розвитку та масштабування. Перевагою цієї архітектури є те, що вона дозволяє створити не просто застосунок, а платформу, яка може бути адаптована для інших типів документів, наприклад, актів виконаних робіт,

рахунків-фактур, сертифікатів та чеків. Завдяки цьому система набуває універсальності і може бути впроваджена у підприємствах з різними формами документообігу. Крім цього, програмний код, побудований на відкритих технологіях, дозволяє зменшити вартість впровадження та спрощує модифікацію у разі потреби.

Висновок. Автоматизація обробки документів є важливим етапом цифровізації підприємств, спрямованим на підвищення точності обробки даних та зниження витрат часу на ручне введення інформації. Запропонована система базується на сучасних методах комп'ютерного бачення та оптичного розпізнавання символів (OCR) з використанням бібліотеки Nicomsoft OCR, що забезпечує ефективне витягування реквізитів з товарно-транспортних накладних. Семантичний аналіз тексту із застосуванням шаблонного підходу сприяє адаптації алгоритмів до різних форматів документів. Використання мов програмування Java та бази даних Firebird SQL забезпечує високу продуктивність, надійність зберігання даних та гнучкість інтеграції з корпоративними інформаційними системами. Модульна архітектура дозволяє розширювати функціонал системи, інтегрувати її з зовнішніми сервісами та застосовувати для обробки інших типів документів. Ці технологічні рішення сприяють підвищенню ефективності управління документацією та впровадженню сучасних підходів до цифрової трансформації підприємств.

References

1. Shahar Markovitch and Paul Willmott/McKinsey – Accelerating the digitalization of business processes[Електронний ресурс] - Режим доступу: <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/accelerating-the-digitization-of-business-processes>
2. Sayallar S., Sayar A., Babalik N.: An OCR Engine for Printed Receipt Images using Deep Learning Techniques. International Journal of Advanced Computer Science and Applications, Vol. 14, No. 2, 2023. URL: <https://avesis.kocaeli.edu.tr/yayin/aa68b0f4-568a-46da-b680-98d60f7e561c/an-ocr-engine-for-printed-receipt-images-using-deep-learning-techniques> (дата звернення: 22.04.2025)
3. Yue A. Automated Receipt Image Identification, Cropping, and Parsing, 2018. URL: https://www.cs.princeton.edu/courses/archive/spring18/cos598B/public/projects/System/COS598B_spr2018_ReceiptParsing.pdf (дата звернення: 23.04.2025)
4. Drobac S., Linden K.: Optical character recognition with neural networks and post-correction with finite state methods. International Journal on Document Analysis and Recognition (IJ DAR) (2020) 23:279–295 URL: <https://link.springer.com/article/10.1007/s10032-020-00359-9> (дата звернення: 23.04.2025)

DOI 10.34132/mspc2025.01.14.13

Oleksandr Obushko
Inessa Kulakovska

SYSTEM FOR RECOGNIZING AND CLASSIFYING FINANCIAL DOCUMENTS

Annotation: This study focuses on the automation of document processing, specifically the extraction of requisites from waybills (TTN) to facilitate digital transformation in enterprises. The main objective is to develop a flexible software system capable of accurately extracting data from documents of various formats. The system incorporates image preprocessing techniques, optical character recognition (OCR) using Nicomsoft OCR, and a template-based semantic text analysis approach. Extracted data undergoes normalization to ensure standardized storage. The implementation is based on Java, utilizing Swing for the user interface, Apache POI for document processing, and Firebird SQL for data storage. The modular architecture enables seamless integration with corporate platforms and scalability for processing different document types. By enhancing document management efficiency, this solution supports enterprise digitalization, optimizing data processing speed and accuracy while reducing operational costs. The adaptability of the system makes it a valuable tool for modernizing administrative workflows across various industries.

Key words: document processing automation, optical character recognition (OCR), semantic text analysis, digital transformation of enterprises, modular system architecture.

Відомості про автора / Information about the Author

Олександр Обушко, здобувач ступеню бакалавра, факультет комп'ютерних наук, Чорноморського національного університету імені Петра Могили, м. Миколаїв, Україна, 54003. E-mail: vebvongolla@gmail.com, orcid: <https://orcid.org/0009-0000-7421-5993>

Oleksandr Obushko, Bachelor's Degree Candidate, Department of Intelligent Information Systems, Faculty of Computer Science, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine, 54003, E-mail: vebvongolla@gmail.com, ORCID: <https://orcid.org/0009-0000-7421-5993>

Інесса Кулаковська, кандидат фізико-математичних наук, доцент кафедри Інтелектуальних інформаційних систем, факультет комп'ютерних наук, Чорноморського національного університету імені

Петра Могили, м. Миколаїв, Україна. E-mail: inessa.kulakovska@chmmu.edu.ua, orcid: <http://orcid.org/0000-0002-8432-1850>

Inessa Kulakovska, PhD of Physical and Mathematical Sciences, Associate Professor, Department of Intelligent Information Systems, Faculty of Computer Science, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine, 54003, E-mail: Inessa.Kulakovska@chmmu.edu.ua, ORCID: <http://orcid.org/0000-0002-8432-1850>

© О. Обушко, І. Кулаковська
19.05.2025.

Стаття отримана редакцією

