บทความวิจัย (Research Article)

การพัฒนาระบบกระจายเอกสารพีดีเอฟอัตโนมัติด้วยการบูรณาการโอซีอาร์และการเทียบเคียงฐานข้อมูลพนักงานเพื่อการแจ้งรายบุคคล

Development of an Automated PDF Document Routing System Integrating OCR and Employee Database Matching for Targeted Personnel Notification

ทรงฤทธิ์ กิติศรีวรพันธุ์* และ ณัฐวัตร แพงแสน

สาขาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์
มหาวิทยาลัยนครพนม
* ผู้รับผิดชอบบทความ
songriti@npu.ac.th

Received: 9 Dec 2025
Revised: 20 Dec 2025
Accepted: 22 Dec 2025

บทคัดย่อ

ในยุคปัจจุบันการบริหารจัดการเอกสารอิเล็กทรอนิกส์มีบทบาทสำคัญในองค์กร โดยเฉพาะไฟล์รูปแบบพีดีเอฟที่ได้รับความนิยมอย่างแพร่หลาย อย่างไรก็ตามเอกสารจำนวนมากในระบบงานสารบรรณเป็นเอกสารที่เกิดจากการสแกน ซึ่งมีลักษณะเป็นไฟล์รูปภาพ ทำให้ไม่สามารถสืบค้นข้อมูลหรือคัดลอกข้อความได้โดยตรง ส่งผลให้กระบวนการคัดกรองและส่งต่อเอกสารไปยังผู้เกี่ยวข้องต้องใช้ทรัพยากรบุคคลและเวลาเป็นจำนวนมาก งานวิจัยนี้ได้เสนอการพัฒนาระบบอัตโนมัติเพื่อลดกระบวนการคัดกรองและส่งต่อเอกสารไปยังผู้เกี่ยวข้อง โดยประยุกต์ใช้เทคโนโลยีโอซีอาร์ ด้วยซอฟต์แวร์ Tesseract ร่วมกับการใช้นิพจน์ปกติ ในการสกัดรายชื่อภาษาไทยออกจากเอกสาร และใช้อัลกอริทึมการเปรียบเทียบความคล้ายคลึงของสตริง ร่วมกับฐานข้อมูลพนักงานในระบบบริการไดเร็กทอรี่ เพื่อระบุชื่อบุคคลและส่งไฟล์ให้รายบุคคลผ่านทางอีเมล ระบบถูกพัฒนาบนสถาปัตยกรรมไมโครเซอร์วิส เพื่อความยืดหยุ่นในการขยายระบบ ผลการทดลองแสดงให้เห็นว่าระบบสามารถทำงานได้อย่างมีประสิทธิภาพ โดยการกำหนดค่าเกณฑ์ความคล้ายคลึง ที่ร้อยละ 80 ให้ความแม่นยำในการระบุตัวบุคคลสูงสุดถึงร้อยละ 98.1 ซึ่งช่วยลดความผิดพลาดและระยะเวลาในการจัดการเอกสารได้อย่างมีนัยสำคัญ

คำสำคัญ: โอซีอาร์ พีดีเอฟ เร็กกูลาร์เอ็กซ์เพรสชัน ฐานข้อมูลแอบแดป ระบบอัตโนมัติ การแจ้งรายบุคคล

Abstract

In the digital era, electronic document management plays a pivotal role in organizations, particularly with the widespread adoption of the Portable Document Format (PDF). However, a significant portion of administrative documents consists of scanned files (image-based PDFs), rendering their text content unsearchable and uncopiable directly. This limitation necessitates manual processing for screening and routing documents to relevant personnel, consuming substantial human resources and time. This research proposes the development of an automated system to streamline the document screening and routing process. The system integrates Optical Character Recognition (OCR) technology using Tesseract software combined with Regular Expressions to extract Thai names from documents. Furthermore, String Similarity Matching algorithms are employed in conjunction with an employee database within a Directory Service system to accurately identify individuals and dispatch specific files via email. The system

is developed based on a Microservices architecture to ensure scalability. Experimental results demonstrate the system's effectiveness, with a similarity threshold setting of 80% yielding a maximum identification accuracy of 98.1%. This significantly minimizes errors and reduces the time required for document management workflows.

Keywords: Optical Character Recognition, PDF, Regular Expression, Directory Service, Automated Document Routing

1. บทนำ

การบริหารจัดการเอกสารเป็นหัวใจสำคัญของการดำเนินงานในหน่วยงานภาครัฐและภาคธุรกิจ ในอดีตกระบวนการเหล่านี้พึ่งพากระดาษเป็นหลัก แต่ด้วยความก้าวหน้าทางเทคโนโลยีสารสนเทศ การสื่อสารผ่านอีเมล (E-mail) และการใช้ไฟล์เอกสารดิจิทัลจึงเข้ามาแทนที่ โดยรูปแบบไฟล์มาตรฐานที่นิยมใช้ในการแลกเปลี่ยนข้อมูลคือไฟล์พีดีเอฟ (Portable Document Format: PDF) [1] เนื่องจากมีความสามารถในการคงรูปแบบของเอกสารต้นฉบับไว้ได้ไม่ว่าจะเปิดบนแพลตฟอร์มใด

อย่างไรก็ตาม กระบวนการทำงานจริงในปัจจุบันยังคงมีความซับซ้อน กล่าวคือ เอกสารต้นฉบับมักถูกพิมพ์ลงกระดาษเพื่อนำเสนอผู้มีอำนาจลงนามรับรอง จากนั้นจึงถูกนำกลับเข้าสู่ระบบดิจิทัลด้วยวิธีการสแกนหรือถ่ายเอกสารเพื่อบันทึกเป็นไฟล์พีดีเอฟ ก่อนจะแจกจ่ายทางอีเมล กระบวนการนี้ทำให้ไฟล์พีดีเอฟ ที่ได้มีคุณสมบัติเป็นไฟล์รูปภาพ ซึ่งคอมพิวเตอร์ไม่สามารถอ่านค่าตัวอักษรภายในได้โดยตรง การค้นหาชื่อบุคคลหรือเนื้อหาสำคัญในเอกสารจึงเป็นเรื่องยากและต้องอาศัยมนุษย์ในการเปิดอ่านทีละหน้า ซึ่งมีความเสี่ยงต่อความผิดพลาด (Human Error) และทำให้เกิดความล่าช้าในการปฏิบัติงาน โดยเฉพาะในองค์กรที่มีปริมาณเอกสารหมุนเวียนจำนวนมาก

เพื่อแก้ไขปัญหาดังกล่าว เทคโนโลยีโอซีอาร์ (Optical Character Recognition: OCR) [2] จึงถูกนำมาประยุกต์ใช้ในการแปลงไฟล์ภาพเอกสารให้เป็นข้อมูลข้อความที่คอมพิวเตอร์สามารถประมวลผลได้ งานวิจัยนี้จึงมุ่งเน้นการพัฒนาระบบที่สามารถ "อ่าน" เอกสารสแกนแทนมนุษย์ โดยมีวัตถุประสงค์หลัก 3 ประการ คือ 1) เพื่อประยุกต์ใช้เทคโนโลยีโอซีอาร์ ร่วมกับนิพจน์ปกติ (Regular Expression) ในการค้นหารายชื่อบุคคลภาษาไทยในเอกสาร 2) เพื่อพัฒนาระบบค้นหาและจับคู่ชื่อบุคคลโดยเปรียบเทียบข้อมูลที่สกัดได้กับฐานข้อมูลองค์กร และ 3) เพื่อสร้างระบบอัตโนมัติในการส่งต่อเอกสารไปยังผู้ที่เกี่ยวข้องผ่านทางอีเมล ซึ่งจะช่วยลดภาระงานและเพิ่มประสิทธิภาพในการบริหารจัดการเอกสารขององค์กร

แม้แนวโน้มการแก้ปัญหาในปัจจุบันจะมุ่งเน้นไปที่การสร้างโมเดล Deep Learning ขนาดใหญ่ แต่การนำมาประยุกต์ใช้ในกระบวนการทำงานสำนักงานทั่วไปอาจก่อให้เกิดความซับซ้อนในการคำนวณ (Computational Complexity) ที่เกินความจำเป็น งานวิจัยนี้จึงนำเสนอแนวทางที่เน้นประสิทธิภาพการใช้ทรัพยากร (Resource-efficient Approach) โดยประยุกต์ใช้เทคโนโลยีโอซีอาร์ ร่วมกับการใช้นิพจน์ปกติโดยพัฒนาให้มีการทำงานร่วมกันของ ซอฟต์แวร์ Tesseract OCR ร่วมกับการใช้นิพจน์ทั่วไป (Regular Expression) จะสามารถจัดการกับสัญญาณรบกวนจาก OCR ได้อย่างมีประสิทธิภาพ โดยไม่จำเป็นต้องพึ่งพาทรัพยากรประมวลผลระดับสูง

ขอบเขตของงานวิจัยนี้ครอบคลุมการรองรับรายชื่อบุคคลภาษาไทย และสามารถประมวลผลไฟล์เอกสารนามสกุล PDF PNG JPG และ JPEG โดยมุ่งเน้นการทดสอบกับเอกสารทางราชการหรือเอกสารบันทึกข้อความที่มีรูปแบบกึ่งโครงสร้าง

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 โอซีอาร์ (Optical Character Recognition: OCR)

OCR คือกระบวนการแปลงภาพของตัวอักษร ไม่ว่าจะเป็นจากเอกสารสแกน หรือภาพถ่าย ให้กลายเป็นข้อความที่เครื่องคอมพิวเตอร์สามารถเข้าใจและแก้ไขได้ กระบวนการทำงานประกอบด้วยขั้นตอนสำคัญ ได้แก่ การนำเข้าภาพ (Scanning) การประมวลผลภาพเบื้องต้น (Pre-processing) เช่น การปรับระนาบ การกำจัดจุดรบกวน การตรวจจับข้อความ (Text Detection) การแปลงเป็นตัวอักษร (Recognition) และการประมวลผลหลังการแปลง (Post-processing) ในงานวิจัยนี้เลือกใช้ Tesseract OCR ซึ่งเป็นโอเพนซอร์สที่มีประสิทธิภาพสูง รองรับภาษาไทย และมีการพัฒนาอย่างต่อเนื่องโดย Google [3] ซึ่ง Chaimooltan [4] ได้เปรียบเทียบและพบว่า Tesseract ทำงานได้ดีกับเอกสารภาษาไทย

2.2 นิพจน์ปกติ (Regular Expression: Regex)

นิพจน์ปกติหรือ Regex เป็นรูปแบบการเขียนคำสั่งเพื่อค้นหาชุดตัวอักษรที่มีรูปแบบเฉพาะเจาะจง งานวิจัยของ Cox [5] ระบุว่าอัลกอริทึมของ Regex ที่ไม่ซับซ้อนจะให้ประสิทธิภาพการทำงานที่รวดเร็ว ในงานวิจัยนี้ Regex ถูกนำมาใช้ในการกรอง (Filter) ข้อมูลขยะที่ได้จากกระบวนการ OCR และดึงเฉพาะส่วนที่เป็นรายชื่อบุคคลออกมา โดยอาศัยคำนำหน้าชื่อ (เช่น นาย นาง นางสาว) เป็นจุดสังเกตหลัก

2.3 การเปรียบเทียบความคล้ายคลึง (String Similarity)

เนื่องจากการทำงานของโอซีอาร์อาจมีความผิดพลาดในการอ่านอักขระภาษาไทย (เช่น วรรณยุกต์ หรือสระลอย) การเปรียบเทียบชื่อที่สกัดได้กับฐานข้อมูลจึงต้องใช้วิธีการหาความคล้ายคลึง (Similarity Matching) แทนการเปรียบเทียบแบบตรงตัว เพื่อเพิ่มความแม่นยำในการระบุตัวบุคคล

2.4 โพรโทคอลการรับส่งอีเมล์ (Mail Server)

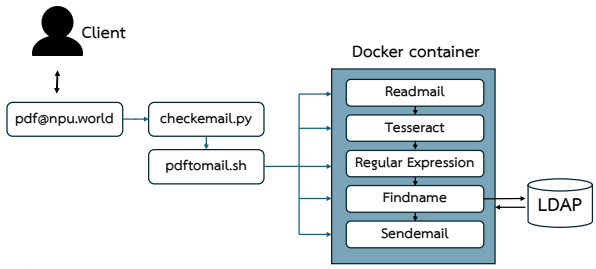
ระบบจำเป็นต้องมีการรับและส่งข้อมูลผ่านเครือข่ายอินเทอร์เน็ต โดยใช้โพรโทคอลมาตรฐาน ได้แก่ SMTP (Simple Mail Transfer Protocol) สำหรับการส่งอีเมลผ่านพอร์ต 25 หรือ 587 และ IMAP (Internet Message Access Protocol) สำหรับการดึงข้อมูลอีเมลจากเซิร์ฟเวอร์ผ่านพอร์ต 143 ซึ่งช่วยให้ระบบสามารถตรวจสอบสถานะอีเมล (เช่น Unread) และดาวน์โหลดไฟล์แนบมาประมวลผลได้

2.5 ระบบบริการไดเร็กทอรี่และโพรโตคอล LDAP

LDAP เป็นโพรโทคอลสำหรับเข้าถึงและจัดการข้อมูลแบบ Directory Service ซึ่งนิยมใช้เก็บข้อมูลบุคลากรในองค์กร โครงสร้างข้อมูลจัดเก็บแบบต้นไม้ (Tree Structure) โดยมี Attribute สำคัญที่งานวิจัยนำมาใช้ คือ cn (Common Name) สำหรับชื่อ-นามสกุล และ mail สำหรับที่อยู่อีเมล

3. วิธีดำเนินการวิจัย

ระบบถูกออกแบบภายใต้แนวคิดสถาปัตยกรรมซอฟต์แวร์แบบ Microservices โดยแบ่งการทำงานออกเป็นโมดูลย่อยที่ทำงานบน Docker Container เพื่อความสะดวกในการจัดการและขยายผล โดยมีขั้นตอนการทำงานรวมดังแสดงในรูปที่ 1 โดยระบบประกอบด้วย 5 ส่วนการทำงานหลัก รูปกล่องสี่เหลี่ยมใช้อธิบายขั้นตอนของการทำงานในแต่ละส่วนงาน เมื่อทำงานเสร็จจะส่งต่อไปยังส่วนงานต่อไปตามลำดับลูกศร ซึ่งเขียนควบคุมด้วยสคริปต์ภาษาไพทอน (Python) และ Shell Script



รูปที่ 1 ภาพรวมสถาปัตยกรรมและการทำงานของระบบ

จากรูปที่ 1 การทำงานของระบบการค้นหารายชื่อบุคคลในเอกสารพีดีเอฟ เริ่มต้นจากระบบได้จัดเตรียมดอกเกอร์คอนเทนเนอร์ทั้ง 5 ส่วนได้แก่ “Readmail” “Tesseract” “Regular Expression” “Findname” และ “Sendemail” หลังจากผู้ใช้งานส่งอีเมลไปยัง pdf@npu.world โค้ด checkemail.py จะทำการตรวจสอบอีเมล เมื่อตรวจพบอีเมล์ที่ต้องการ จึงสั่งใช้งานสคริปต์ pdfmail.sh เพื่อทำการใช้งานด๊อกเกอร์คอนเทนเนอร์ เมื่อแต่ละคอนเทนเนอร์จะทำหน้าเสร็จจึงส่งผลลัพธ์ไปยังคอนเทนเนอร์ถัดไปจนกระทั่งจบการทำงาน

3.1 ขั้นตอนการรับเอกสาร (Readmail Container)

ทำหน้าที่เชื่อมต่อกับ Mail Server ขององค์กร (ในที่นี้ได้แก่ mail.npu.world) ผ่านโพรโทคอล IMAP สคริปต์ checkemail.py จะทำงานวนลูปตรวจสอบกล่องจดหมายเข้า (INBOX) ทุกๆ 5 วินาที หากพบอีเมลที่มีสถานะ "ยังไม่อ่าน" (UNSEEN) ระบบจะทำการดาวน์โหลดไฟล์แนบ (Attachment) ที่เป็นพีดีเอฟ มาเก็บไว้ในโฟลเดอร์ชั่วคราว (/output) และเรียกสคริปต์หลัก pdftomail.sh ให้เริ่มกระบวนการถัดไป

3.2 การแปลงภาพเป็นข้อความ (Tesseract Container)

เนื่องจาก Tesseract ทำงานกับไฟล์รูปภาพ ไฟล์พีดีเอฟ ที่ได้รับมาจึงต้องถูกแปลงเป็นไฟล์ภาพเสียก่อน โดยใช้ไลบรารี pdf2image แปลงแต่ละหน้าของเอกสารให้เป็นไฟล์ภาพความละเอียดสูง (300-500 DPI) จากนั้นจึงส่งเอกสารไฟล์รูปภาพเข้าสู่กระบวนการโอซีอาร์ โดยใช้ด้วยคำสั่ง pytesseract.image_to_string

โดยกำหนดภาษาเป็นภาษาไทย (lang='tha') ผลลัพธ์ที่ได้คือไฟล์ข้อความ (.txt) ที่บรรจุเนื้อหาทั้งหมดของเอกสาร รวมถึงขยะอักขระที่เกิดจากความผิดพลาดในการอ่าน

```
((|ดร\.?|ศ\.?|รศ\.?|ผศ\.?|ผศ\.ดร\.?|รศ\.ดร\.?|ศ\.ดร\.?|
หม่อม|หม่อมราชวงศ์|หม่อมหลวง|พระองค์เจ้า|สมเด็จพระ|
พระเจ้าวรวงศ์เธอ|พระยา|เจ้าคุณ|คุณหญิง|คุณชาย|ท่าน
ผู้หญิง|พลเอก|พลโท|พลตรี|พันเอก|พันโท|พันตรี|ร้อยเอก|ร้อย
โท|ร้อยตรี|สิบเอก|สิบโท|สิบตรี|จ่าสิบเอก|จ่าสิบโท|จ่าสิบตรี|
ว่าที่ร้อยตรี|ว่าที่ร้อยโท|ว่าที่ร้อยเอก|พลตำรวจเอก|พลตำรวจ
โท|พลตำรวจตรี|พันตำรวจเอก|พันตำรวจโท|พันตำรวจตรี|
ร้อยตำรวจเอก|ร้อยตำรวจโท|ร้อยตำรวจตรี|สิบตำรวจเอก|สิบ
ตำรวจโท|สิบตำรวจตรี|จ่าสิบตำรวจ|ดาบตำรวจ|ส.ต.|ส.ต.ท.|
ส.ต.อ.|ด.ต.|ร.ต.|ร.ต.ท.|ร.ต.อ.|พ.ต.|พ.ต.ท.|พ.ต.อ.|พล.ต.|
พล.ต.ท.|พล.ต.อ.ข) [ก-๙]+\s+[ก-๙]+)
```

รูปที่ 2 ตัวอย่าง Pattern Matching ชื่อบุคคล

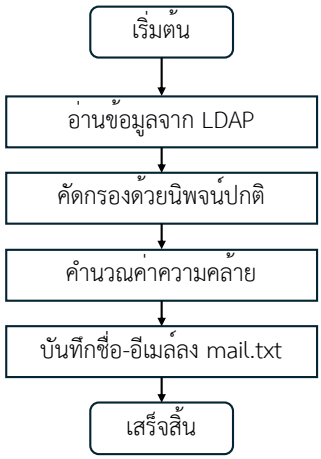
3.3 การคัดกรองรายชื่อด้วยนิพจน์ปกติ (Regular Expression Container)

ข้อความดิบจากกระบวนการโอซีอาร์ จะถูกส่งต่อมายังส่วนการคัดกรองรายชื่อเพื่อสกัดเฉพาะรายชื่อบุคคล โค้ด regex.py เป็นการปรับรูปแบบข้อความของไฟล์ที่ได้จากการทำงานของดอกเกอร์คอนเทนเนอร์ Tesseract เพื่อความสะดวกในการใช้งานในส่วนของการค้นหาบุคคล บันทึกเป็นไฟล์ชื่อ person.txt โดยไฟล์ regex.py มีฟังก์ชันการทำงานเพื่อจัดรูปแบบข้อความ สรุปการทำงานของนิพจน์ปกติมีดังนี้

- Preprocessing ทำการลบบรรทัดว่างและอักขระเดี่ยวที่ไม่มีความหมาย
- Pattern Matching ใช้นิพจน์ปกติเพื่อค้นหาคำนำหน้าชื่อบุคคล ตามด้วยสตริงภาษาไทย (ชื่อจริง) เว้นวรรค และสตริงภาษาไทย (นามสกุล) มีรูปแบบดังรูปที่ 2 ซึ่งสามารถเพิ่มคำนำหน้าชื่อด้วยการกำหนดรูปแบบนิพจน์ปกติ ได้ในอนาคตเพื่อใช้เป็นเครื่องมือยืนยันชื่อบุคคลได้
- Refinement จัดรูปแบบข้อความให้เหลือเพียงรายชื่อบุคคลบรรทัดละ 1 ชื่อ เพื่อเตรียมส่งให้ขั้นตอนถัดไป

3.4 ส่วนระบุชื่อบุคคลและค้นหาอีเมล์

รายชื่อที่สกัดได้อาจมีความคลาดเคลื่อนจากต้นฉบับเนื่องจากข้อจำกัดของโอซีอาร์ (เช่น สระลอย หรือวรรณยุกต์ผิดเพี้ยน) ระบบจึงไม่สามารถเปรียบเทียบแบบตรงตัว (Exact Match) กับฐานข้อมูลได้ จึงต้องใช้อัลกอริทึม Sequence Matcher จากไลบรารี difflib เพื่อหาค่าอัตราส่วนความคล้ายคลึง (Similarity Ratio) ลำดับการทำงานดังแสดงในรูปที่ 3



รูปที่ 3 ลำดับขั้นการระบุชื่อและค้นหาอีเมล์

3.5 ส่วนส่งเอกสาร (Sendemail Container)

ส่วนการทำงานนี้ทำหน้าที่เป็นขั้นตอนสุดท้ายของระบบ โดยรับข้อมูลอินพุตเป็นไฟล์รายการอีเมล (mail.txt) ที่ได้จากการค้นหาชื่อบุคคลในขั้นตอนก่อนหน้า และไฟล์พีดีเอฟต้นฉบับที่ถูกจัดเก็บไว้ในโฟลเดอร์ชั่วคราว การทำงานของโมดูลนี้เขียนด้วยภาษาไพทอนโดยใช้ไลบรารี smtplib สำหรับการสื่อสารผ่านโพรโทคอล SMTP และ email.mime สำหรับการจัดการเนื้อหาอีเมลและไฟล์แนบ

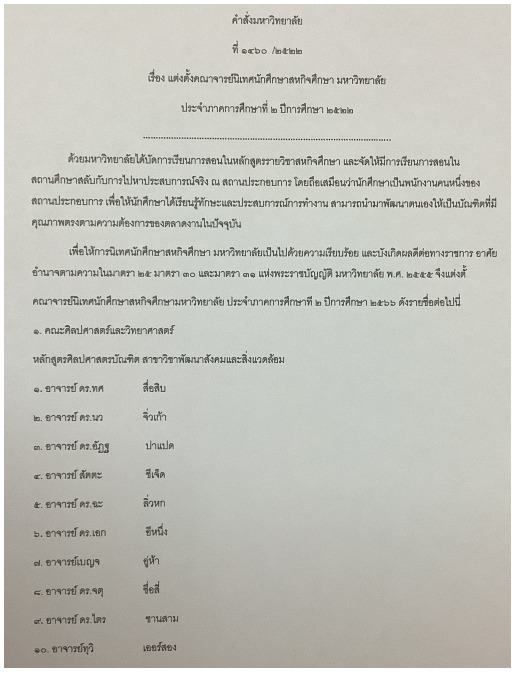
กระบวนการทำงานเริ่มจากการเชื่อมต่อกับเครื่องแม่ข่ายไปรษณีย์อิเล็กทรอนิกส์ (mail.npu.world) ผ่านพอร์ต 587 เพื่อความปลอดภัยในการส่งข้อมูล จากนั้นทำการระบุชื่อบุคคลด้วยบัญชีอีเมลของระบบ (pdf@npu.world) ระบบจะอ่านรายชื่ออีเมลจากไฟล์ข้อมูลและดำเนินการวนซ้ำ เพื่อสร้างอีเมลรายฉบับ โดยกำหนดหัวข้อเรื่องและเนื้อหาจดหมายในรูปแบบ HTML พร้อมแนบไฟล์พีดีเอฟ ที่เกี่ยวข้องให้กับผู้รับแต่ละราย

หลังจากดำเนินการส่งอีเมลให้แก่ผู้เกี่ยวข้องครบถ้วนแล้ว ระบบจะสร้างอีเมลรายงานผล ซึ่งระบุรายละเอียดวันเวลาและรายชื่อผู้รับทั้งหมดที่ระบบได้ดำเนินการส่งเอกสารให้ ส่งไปยังอีเมลของผู้ดูแลระบบ (report-pdf@npu.world) เพื่อเป็นการบันทึกประวัติการทำงาน และแจ้งเตือนให้ทราบว่ากระบวนการทั้งหมดเสร็จสิ้นสมบูรณ์

4. ผลการทดลองและการวิเคราะห์

การทดลองดำเนินการบนเครื่องแม่ข่ายจำลองระบบปฏิบัติการ Linux Ubuntu 22.04 โดยแบ่งการทดสอบออกเป็น 2 ส่วนหลัก คือ การทดสอบประสิทธิภาพด้านเวลา (Time Performance) และการทดสอบความแม่นยำ (Accuracy)

ในการทดลองนี้ได้ทำการทดสอบกับชุดเอกสารจำลอง จำนวน 30 ฉบับ ที่ผ่านกระบวนการพิมพ์และสแกนเพื่อให้เกิดสัญญาณรบกวนจริงดังรูปที่ 4 โดยมีรายชื่อบุคคลเป้าหมายรวม 900 รายชื่อ



คำสั่งมหาวิทยาลัย

ที่ ๑๔๖๐ /๒๕๖๒

เรื่อง แต่งตั้งคณาจารย์นิเทศนักศึกษาสหกิจศึกษา มหาวิทยาลัย

ประจำภาคการศึกษาที่ ๒ ปีการศึกษา ๒๕๖๒

ด้วยมหาวิทยาลัยได้บัดการเรียนการสอนในหลักสูตรรายวิชาสหกิจศึกษา และจัดให้มีการเรียนการสอนในสถานศึกษาสลับกับการไปหาประสบการณ์จริง ณ สถานประกอบการ โดยถือเสมือนว่านักศึกษาเป็นพนักงานคนหนึ่งของสถานประกอบการ เพื่อให้นักศึกษาได้เรียนรู้ทักษะและประสบการณ์การทำงาน ตามวรถนำมาพัฒนาตนเองให้เป็นบัณฑิตที่มีคุณภาพตรงตามความต้องการของตลาดงานในปัจจุบัน

เพื่อให้การนิเทศนักศึกษาสหกิจศึกษา มหาวิทยาลัยเป็นไปด้วยความเรียบร้อย และบังเกิดผลดีต่อทางราชการ อาศัยอำนาจตามความในมาตรา ๒๕ มาตรา ๓๐ และมาตรา ๓๑ แห่งพระราชบัญญัติ มหาวิทยาลัย พ.ศ. ๒๕๕๕ จึงแต่งตั้ง

คณาจารย์นิเทศนักศึกษาสหกิจศึกษามหาวิทยาลัย ประจำภาคการศึกษาที่ ๒ ปีการศึกษา ๒๕๖๖ ดังรายชื่อต่อไปนี้

๑. คณะศิลปศาสตร์และวิทยาศาสตร์

หลักสูตรศิลปศาสตรบัณฑิต สาขาวิชาพัฒนาสังคมและสิ่งแวดล้อม

๑. อาจารย์ ดร.ทศ สื่อสิบ

๒. อาจารย์ ดร.นว จิ๋วเก้า

๓. อาจารย์ ดร.ธัญ ปาแปด

๔. อาจารย์ ลัดดะ ชีเจ็ด

๕. อาจารย์ ดร.ฉะ สิ่วหก

๖. อาจารย์ ดร.เอก ชีหนึ่ง

๗. อาจารย์เบญจ ชู้ห้า

๘. อาจารย์ ดร.จตุ ชี่อสี่

๙. อาจารย์ ดร.ไตร ชานสาม

๑๐. อาจารย์ทุวิ เอยร์สอง

รูปที่ 4 ตัวอย่างเอกสารที่จำลอง

การนำเอกสารเข้าสู่ระบบคอมพิวเตอร์โดยสแกนเอกสารด้วยเครื่องพิมพ์ ในรูปที่ 5 รุ่น Brother DCP-T310 และใช้งานโปรแกรม Brother&iPrintScan โดยตั้งค่าความละเอียดที่ 150x150 DPI

รูปที่ 5 การนำเอกสารจำลองเข้าคอมพิวเตอร์ด้วยเครื่องสแกนเนอร์

ตารางที่ 1 ความสัมพันธ์ระหว่างขนาดไฟล์และจำนวนหน้ากระดาษของชุดข้อมูลทดสอบ

ชื่อไฟล์	ขนาดไฟล์(ไบต์)	จำนวนหน้า
100k.pdf	102127	2
200k.pdf	222072	5
300k.pdf	320373	6
400k.pdf	441155	7
500k.pdf	512034	5
600k.pdf	625420	8
700k.pdf	754769	8
800k.pdf	849976	16
900k.pdf	1004970	15
1M.pdf	1044444	18
2M.pdf	2885808	58
3M.pdf	3123252	9
4M.pdf	4235610	8
5M.pdf	5169252	9
6M.pdf	6228179	10
7M.pdf	7240593	6

4.1 การทดสอบประสิทธิภาพด้านเวลา

ในการประเมินประสิทธิภาพของระบบ ผู้วิจัยได้ดำเนินการทดสอบเพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อระยะเวลาในการประมวลผล โดย

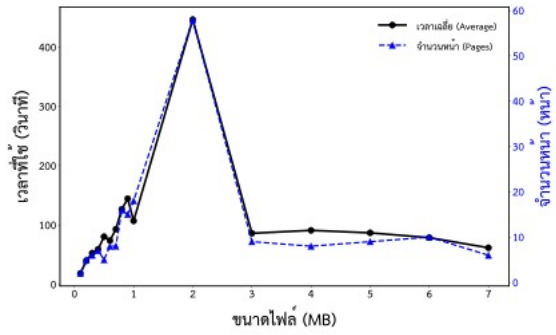
กำหนดตัวแปรอิสระ 3 ประการ ได้แก่ ขนาดของไฟล์เอกสาร จำนวนหน้ากระดาษ และจำนวนรายชื่อบุคคลในเอกสาร เพื่อจำแนกผลกระทบที่มีต่อเวลาการทำงานรวมและเวลาของแต่ละโมดูลย่อย (Microservices) ดังรายละเอียดต่อไปนี้

4.1.1 ผลกระทบจากขนาดไฟล์และจำนวนหน้ากระดาษ

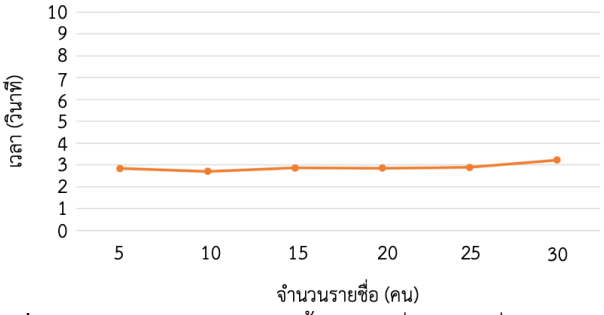
ผู้วิจัยได้กำหนดชุดข้อมูลทดสอบ โดยแปรผันปัจจัยนำเข้า 2 ประการ ได้แก่ ขนาดของไฟล์และจำนวนหน้ากระดาษ ความสัมพันธ์ระหว่างขนาดไฟล์และจำนวนหน้ากระดาษของชุดข้อมูลทดสอบปรากฏดัง ตารางที่ 1 ในการทดสอบประสิทธิภาพการประมวลผลโดยกำหนดตัวแปรด้านขนาดไฟล์และจำนวนหน้า พบว่าในช่วงขนาดไฟล์ 0.1 MB ถึง 2.0 MB ระยะเวลาที่ใช้ในการประมวลผลมีแนวโน้มเพิ่มขึ้นตามจำนวนหน้ากระดาษ โดยมีจุดสูงสุดที่ขนาดไฟล์ 2.0 MB มีจำนวน 58 หน้า ทั้งนี้เนื่องจากโมดูล Tesseract ต้องเพิ่มจำนวนรอบการทำงาน (Iterations) ในการประมวลผลภาพและสกัดค่าอักขระตามจำนวนหน้าที่มีปริมาณมาก ในทางตรงกันข้าม เมื่อพิจารณาในช่วงขนาดไฟล์ 3.0 MB ถึง 7.0 MB ซึ่งเป็นการทดสอบโดยกำหนดให้จำนวนหน้าลดลงเหลือ 6–10 หน้า แต่ไฟล์ได้เพิ่มความละเอียดของภาพให้สูงขึ้นแทน พบว่าระยะเวลาการประมวลผลรวมในรูปที่ 6 กลับลดต่ำลงอย่างชัดเจนเมื่อเปรียบเทียบกับกลุ่มไฟล์ 2.0 MB ผลการทดสอบนี้ชี้ให้เห็นว่า จำนวนหน้า มีอิทธิพลต่อระยะเวลาในการประมวลผลของโมดูล Tesseract มากกว่า ความหนาแน่นของข้อมูลภาพที่ส่งผลให้ขนาดไฟล์ใหญ่ขึ้น

4.1.2 ผลกระทบจากจำนวนรายชื่อในเอกสาร

ในการทดสอบปัจจัยด้านจำนวนรายชื่อ ผู้วิจัยได้กำหนดช่วงการทดสอบไว้ระหว่าง 5 ถึง 30 รายชื่อ เพื่อให้สอดคล้องกับบริบทการใช้งานจริงของเอกสารทางราชการทั่วไป จากผลการทดลองใน รูปที่ 7 พบว่าจำนวนรายชื่อเป็นปัจจัยหลักที่ส่งผลกระทบต่อระยะเวลาการทำงานของโมดูล Sendemail เนื่องจากระบบจำเป็นต้องใช้เวลาในการวนซ้ำ เพื่อสร้างและจัดส่งอีเมลไปยังผู้รับแต่ละรายตามจำนวนที่ระบบตรวจพบ ยิ่งจำนวนรายชื่อมาก เวลาที่ใช้ในการส่งข้อมูลผ่าน SMTP Server ก็จะเพิ่มขึ้นตามลำดับ ส่วนขั้นตอนการประมวลผลก่อนหน้า ไม่ได้รับผลกระทบจากปัจจัยด้านจำนวนรายชื่อแต่อย่างใด



รูปที่ 6 ผลทดลองระยะเวลาการทำงานของ Tesseract แปรผันตามขนาดไฟล์และจำนวนหน้ากระดาษ



รูปที่ 7 ผลทดลองระยะเวลาการค้นหารายชื่อบุคคลเมื่อแปรผันตามจำนวนรายชื่อบุคคล

4.2 การทดสอบความแม่นยำในการระบุชื่อบุคคล

จากการทดลองเพื่อวัดประสิทธิภาพของระบบในการค้นหาและระบุรายชื่อบุคคลในเอกสารพีดีเอฟจำลองจำนวน 30 ชุด รวมรายชื่อเป้าหมายทั้งสิ้น 900 รายชื่อ ภายใต้การกำหนดค่าเกณฑ์ความคล้ายคลึง (Similarity Threshold) 3 ระดับ ได้แก่ 100%, 90% และ 80% ผลการทดลองเชิงประจักษ์ปรากฏดัง ตารางที่ 2 ข้อมูลผลการทดลองจากตารางนำมาคำนวณหา อัตราการระบุตัวตนถูกต้อง (Correct Identification Rate) หรือความระลึก (Recall) เพื่อประเมินความสามารถของระบบในการสืบค้นข้อมูลที่ต้องการ โดยใช้สมการที่ (1)

$$\text{Recal}(\%) = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \qquad (1)$$

โดยที่ TP (True Positive) คือจำนวนรายชื่อที่ระบบตรวจพบและระบุตัวตนถูกต้อง และ FN (False Negative) คือจำนวนรายชื่อที่ระบบไม่สามารถตรวจพบ สรุปได้ดังนี้

ตารางที่ 2 ผลการทดสอบประสิทธิภาพการระบุรายชื่อบุคคล (N=900)

ครั้งที่	จำนวนรายชื่อทั้งหมด (P)	TP 100%	TP 90%	TP 80%
1	30	26	27	30
2	30	26	28	30
3	30	26	28	29
4	30	23	26	30
5	30	23	27	30
6	30	24	26	30
7	30	24	27	30
8	30	18	21	23
9	30	22	23	23
10	30	23	26	30
11	30	24	27	29
12	30	24	27	30
13	30	26	27	30
14	30	26	27	30
15	30	27	29	30
16	30	26	27	30
17	30	27	27	30
18	30	26	27	30
19	30	28	29	30
20	30	28	29	30
21	30	26	27	30
22	30	25	28	30
23	30	26	28	30
24	30	26	27	30
25	30	25	27	29
26	30	27	29	30
27	30	26	27	30
28	30	26	28	30
29	30	27	28	30
30	30	25	27	30
Total	900	755	811	883
Recal(%)l	100.00	83.89	90.11	98.11

- ที่ระดับความคล้ายคลึง 100% ระบบสามารถระบุรายชื่อได้ถูกต้องจำนวน 755 รายชื่อ คิดเป็นค่า Recall ร้อยละ 83.89 ความผิดพลาดส่วนใหญ่เกิดจากความคลาดเคลื่อนเล็กน้อยในกระบวนการโอซีอาร์ ทำให้ข้อความไม่ตรงกับฐานข้อมูลอย่างสมบูรณ์
- ที่ระดับความคล้ายคลึง 90% ประสิทธิภาพเพิ่มขึ้นโดยระบุได้ถูกต้อง 811 รายชื่อ คิดเป็นค่า Recall ร้อยละ 90.11
- ที่ระดับความคล้ายคลึง 80% เป็นจุดที่ระบบทำงานได้ดีที่สุด โดยสามารถระบุรายชื่อได้ถึง 883 รายชื่อ จากทั้งหมด 900 รายชื่อ คิดเป็นค่า Recall สูงถึงร้อยละ 98.11 ซึ่งถือเป็นระดับนัยสำคัญที่ยอมรับได้สำหรับการใช้งานจริง

ค่าประสิทธิภาพแสดงให้เห็นว่า การกำหนดค่า จับคู่ความคล้ายคลึงที่ร้อยละ 80 เป็นจุดสมดุลที่ดีที่สุด (Optimal Point) สำหรับการจัดการเอกสารภาษาไทยที่มีสัญญาณรบกวนจากการสแกน อัตราการตกหล่นของข้อมูล (Miss Rate) ที่ร้อยละ 1.89 ส่วนใหญ่เกิดจากคุณภาพของต้นฉบับที่ต่ำมากจนส่งผลให้รูปร่างอักษรเพี้ยนไปจนเกินขอบเขตที่อัลกอริทึมจะยอมรับได้ อย่างไรก็ตาม ในการทดสอบ ไม่สามารถจับคู่ข้อมูลได้ถูก (False Positive)

5. สรุปผลการวิจัยและข้อเสนอแนะ

การใช้งาน LLM (Large Language Mode) หรือโมเดล NLP ขั้นสูงจำเป็นต้องใช้ทรัพยากรฮาร์ดแวร์ประสิทธิภาพสูง โดยเฉพาะหน่วยประมวลผลกราฟิก และหน่วยความจำขนาดใหญ่เพื่อรองรับการประมวลผล ซึ่งสวนทางกับข้อจำกัดด้านงบประมาณและโครงสร้างพื้นฐานทางไอทีของหน่วยงานราชการส่วนใหญ่ ในทางตรงกันข้าม ข้อเสนอในงานวิจัยนี้ สามารถทำงานได้อย่างลื่นไหลบนเซิร์ฟเวอร์พื้นฐานหรือแม้แต่เครื่องคอมพิวเตอร์สำนักงานทั่วไป ซึ่งช่วยลดต้นทุนการดำเนินการ ได้อย่างมหาศาลและเอื้อต่อการนำไปใช้งานจริงในวงกว้าง อีกทั้งปัญญาประดิษฐ์เชิงทำงานบนพื้นฐานของความน่าจะเป็น ซึ่งมีความเสี่ยงที่จะเกิดปรากฏการณ์ 'การสร้างข้อมูลเท็จ' หรือ Hallucinations โดยเฉพาะการเติมแต่งคำหรือการเดาบริบทที่อาจทำให้ชื่อบุคคลผิดเพี้ยนไปจากเอกสารต้นฉบับ ซึ่งเป็นสิ่งที่ยอมรับไม่ได้ในระบบงานสารบรรณที่ต้องการความถูกต้องแม่นยำสูงสุด การใช้ Regex ซึ่งเป็นอัลกอริทึมเชิงกำหนด (Deterministic Algorithm) จึงเป็นทางเลือกที่ปลอดภัยกว่าในการรับประกันว่าข้อมูลที่ถูกดึงออกมานั้นปรากฏอยู่จริงในเอกสาร 100% โดยปราศจากการปรุงแต่งโดยปัญญาประดิษฐ์

งานวิจัยนี้ให้ผลสำเร็จตามวัตถุประสงค์ในการพัฒนาระบบอัตโนมัติสำหรับค้นหาและแจกจ่ายเอกสารพีดีเอฟ โดยได้บูรณาการเทคโนโลยีโอซีอาร์ เข้ากับอัลกอริทึมตรวจสอบความคล้ายคลึงของข้อความ ซึ่งช่วยแก้ปัญหาความไม่สมบูรณ์ของข้อมูลจากการแปลงไฟล์ภาพได้เป็นอย่างดี ผลการทดสอบยืนยันว่าที่เกณฑ์ความคล้ายคลึงร้อยละ 80 ระบบมีความแม่นยำสูงถึงร้อยละ 98.11 และสามารถลดเวลาการทำงานของมนุษย์ได้อย่างชัดเจน โดยเอกสารส่วนใหญ่ใช้เวลาประมวลผลเฉลี่ยไม่เกิน 3 นาทีต่อไฟล์

แม้งานวิจัยนี้ได้บรรลุตามเป้าหมายที่ผู้วิจัยได้วางไว้ เพื่อให้เกิดความก้าวหน้า ผู้วิจัยพบข้อจำกัดซึ่งสามารถพัฒนาต่อยอดได้ในอนาคต ได้แก่ ปัญหาด้านประสิทธิภาพเมื่อประมวลผลเอกสารที่มีตารางหรือเส้นขอบจำนวนมาก ซึ่งควรแก้ไขด้วยการพัฒนาขั้นตอนการประมวลผลภาพเบื้องต้น เพื่อลบเส้นตารางก่อนเข้าสู่กระบวนการโอซีอาร์

สำหรับแนวทางการพัฒนาต่อยอดในอนาคต ควรขยายขีดความสามารถให้รองรับการค้นหารายชื่อภาษาอังกฤษ และรองรับเอกสารที่มีโครงสร้างซับซ้อนยิ่งขึ้น เช่น เอกสารแบบหลายคอลัมน์หรือเอกสารลายมือเขียน

6. เอกสารอ้างอิง

[1] D. Mailserver. (23 October 2025). *Docker mailserver*. [Online] Available : https://github.com/docker-mailserver/docker-mailserver

[2] K. Chaimooltan. (26 July 2021). *TH-National-Document-OCR-Part-II: First releases of THND OCR Part II*. [Online] Available : https://zenodo.org/records/5136432 doi: 10.5281/zenodo.5136432

[3] R. Hengprasert, "An end-to-end trainable Thai OCR system using deep recurrent neural network," *Journal of Science Innovation for Sustainable Development*, Vol. 2 (1), pp. 78–83, 2020.

[4] T. Khumphakdee, S. Waijanya, and N. Promrit. "Natural language processing to improve the errors caused by the optical character recognition," *KKU Science Journal*, Vol. 51 (2), pp. 126–141, 2023.

[5] R. Cox. (23 October 2025). *Regular expression matching can be simple and fast*. [Online] Available : http://swtch.com/~rsc/regexp/regexp1.html