Patient information extraction using optical character

Todsanai Chumwatana, Waramporn Rattana-amnuaychai, Patranuch Chauychu
International College, Rangsit University

Abstract

At present, digital transformation has played an important role for driving public health agencies to the age of digital transformation, and leading to information usage. However, some patient data still has stored in the format of hard copies, scanned document, images and PDFs which need to be transformed into digital form for future use. The paper aims to propose the technique for recognizing the text from a physical document into digital format, by using Optical Character Recognition, also called OCR that attempts to extract all the text from photocopies into structured data format. The experimental studies showed that the proposed technique makes the digitized documents completely searchable and editable with the average of accuracy performance around 74.62% for extracting attributes and 68.46% for extracting values from printed documents. Utilizing OCR helps the public health agencies to easily turn documents into digital form, provides more effective use of information, and also helps reduce amount of paper taking up space in the organization.

Keywords: *Information Extraction, Optical Character Recognition, OCR*

Received 10 January 2022, Revise: 25 March 2022, Accepted: 1 May 2022

Correspondence : Todsanai Chumwatana, International College, Rangsit University, 52/347 11th Building, Phaholyothin Road, Lak Hok, Muang, Pathum Thani 12000, E-mail: todsanai.c@rsu.ac.th

การสกัดข้อมูลผู้ป่วยด้วยเทคนิคตัวช่วยแปลงไฟล์เอกสาร

ทศนัย ชุ่มวัฒนะ, วรัมพร รัตนอำนวยชัย, ภัทรานุช ช่วยชู
วิทยาลัยนานาชาติ มหาวิทยาลัยรังสิต

บทคัดย่อ

ในปัจจุบัน เทคโนโลยีได้เข้ามามีบทบาทสำคัญในการขับเคลื่อนหน่วยงานด้านสาธารณสุขจากรูปแบบเก่าไปสู่รูปแบบใหม่ที่เหมาะสมในยุคดิจิทัล (Digital Transformation) เพื่อนำไปสู่การใช้ประโยชน์จากข้อมูลอย่างยั่งยืน อย่างไรก็ตาม ปัญหาหลักที่สำคัญคือการที่หน่วยงานเหล่านี้ ยังคงมีการจัดเก็บข้อมูลผู้ป่วยบางส่วนในรูปแบบของเอกสาร (hard copies) เอกสารสแกน รูปภาพ และไฟล์ PDF ซึ่งจำเป็นต้องแปลงเอกสารดังกล่าวให้อยู่ในรูปแบบดิจิทัลเพื่อการวิเคราะห์ และการใช้ประโยชน์ในด้านต่าง ๆ งานวิจัยฉบับนี้จึงมีวัตถุประสงค์เพื่อนำเสนอวิธีในการแปลงข้อความจากเอกสาร (physical document) ให้อยู่ในรูปแบบดิจิทัล โดยใช้ Optical Character Recognition หรือที่เรียกว่า OCR ซึ่งเป็นเทคโนโลยีที่จะช่วยสกัดข้อความทั้งหมดจากเอกสารลงสู่รูปแบบข้อมูลที่มีโครงสร้าง (Structure Data Format) ผลลัพธ์จากการวิจัยแสดงให้เห็นว่าการใช้เทคนิค OCR สามารถช่วยสกัดข้อความจากภาพหรือเอกสารให้อยู่ในรูปแบบดิจิทัลที่สามารถค้นหา แก้ไข และทำการวิเคราะห์ได้ โดยมีอัตราความแม่นยำในการประมวลผลประมาณ 74.62% โดยเฉลี่ย สำหรับการสกัดข้อมูลแอททริบิวต์ (Attributes) และ 68.46% สำหรับการสกัดข้อมูลที่เป็นค่า (Values) จากเอกสาร ดังนั้น ประโยชน์จากการใช้เทคโนโลยี OCR จะช่วยให้หน่วยงานด้านสาธารณสุขสามารถเปลี่ยนข้อมูลในรูปแบบเอกสารให้เป็นรูปแบบดิจิทัลเพื่อนำไปสู่การใช้ประโยชน์จากข้อมูลได้อย่างมีประสิทธิภาพ และช่วยลดปริมาณเอกสารที่มีการจัดเก็บภายในองค์กร

คำสำคัญ: *การสกัดข้อมูล, การอ่านอักขระด้วยแสง, การแปลงไฟล์เอกสาร*

วันที่รับต้นฉบับ 10 มกราคม 2565, วันที่แก้ไข: 25 มีนาคม 2565, วันที่ตอบรับ: 1 พฤษภาคม 2565

บทนำ

เนื่องด้วยจำนวนของเอกสารสิ่งพิมพ์ที่มีปริมาณมากขึ้นอย่างรวดเร็วในทศวรรษที่ผ่านมา และมีแนวโน้มที่จะเพิ่มขึ้นเรื่อย ๆ ซึ่งเอกสารเหล่านั้นไม่สามารถทำการแก้ไขได้โดยใช้เพียงโปรแกรมแก้ไขข้อความทั่วไป ทำให้การสกัดข้อมูลจากเอกสารเพื่อนำมาใช้ประโยชน์กลายเป็นงานที่ท้าทาย และเป็นงานที่ต้องใช้เวลามาก โดยเฉพาะหน่วยงานด้านสาธารณสุขที่ยังคงมีการจัดเก็บข้อมูลผู้ป่วยในแบบฟอร์มกระดาษ ทำให้ยากต่อการค้นหาและใช้ข้อมูล ปัญหาดังกล่าวนี้สามารถแก้ไขได้โดยการนำเทคโนโลยีการอ่านอักขระด้วยแสง หรือ OCR (ย่อมาจาก Optical Character Recognition) มาใช้เป็นเครื่องมือที่ช่วยในการแปลงเอกสาร ไฟล์ภาพ หรือไฟล์ PDF ให้เป็นไฟล์ที่สามารถแก้ไขได้ เพื่อใช้ในการวิเคราะห์ข้อมูล หรือเพื่อใช้งานในอนาคต แม้ว่า OCR จะได้รับการพัฒนาให้เป็นเทคโนโลยีที่ใช้กันอย่างกว้างขวาง แต่ส่วนใหญ่มักจะนำไปใช้ในการสกัดข้อความจากเอกสาร มากกว่าการแปลงเอกสารให้อยู่ในรูปแบบของข้อมูลที่มีโครงสร้าง

ความสามารถในการประมวลผลของ OCR จะขึ้นอยู่กับคุณภาพของเอกสารอินพุต โดยผลลัพธ์ที่ออกมาจะมีอัตราความถูกต้องมากขึ้นเมื่อประมวลผลกับภาพขาวดำ (Bitonal Image) ดังนั้น เพื่อให้การประมวลผลมีประสิทธิภาพมากขึ้น จึงจำเป็นต้องใช้กระบวนการแปลงภาพให้เป็นภาพขาวดำ (Binarization) [1],[2] ก่อนการสกัดข้อมูล

ด้วยเหตุนี้ งานวิจัยฉบับนี้ จึงมุ่งที่จะนำ Optical Character Recognition มาใช้สำหรับการแปลงไฟล์ดิจิทัล โดยใช้แบบฟอร์มลงทะเบียนผู้ป่วยเป็นเอกสารตัวอย่าง ตลอดจนนำเสนอกระบวนการสกัดข้อความ และจำแนกเขตข้อมูล (Fields) เช่น ชื่อ ที่อยู่ วันเกิดและอายุ เป็นต้น เพื่อเก็บข้อมูลสำคัญไว้ในลักษณะข้อมูลที่มีโครงสร้าง เช่น CSV หรือ Excel สำหรับการประมวลผลหรือใช้ประโยชน์ต่อไป

ผู้ประสานงาน: ทศนัย ชุ่มวัฒนะ, วิทยาลัยนานาชาติ มหาวิทยาลัยรังสิต 52/347 อาคาร 11 ถ.พหลโยธิน ต.หลักหก อ.เมือง จ.ปทุมธานี 12000 E-mail: todsanai.c@rsu.ac.th

งานวิจัยที่เกี่ยวข้อง

ปัจจุบัน เทคโนโลยี OCR เป็นที่ต้องการของหน่วยงานด้านสาธารณสุขและองค์กรต่าง ๆ มากขึ้น อีกทั้งเป็นเครื่องมือสำคัญที่ใช้กันอย่างแพร่หลายในองค์กรขนาดใหญ่ รวมไปถึงการพยายามใช้ OCR กับงานด้านต่าง ๆ ตัวอย่างเช่น ระบบจัดเก็บค่าโดยสารอัตโนมัติ (Automatic Fare Collection: AFC) ซึ่งประกอบด้วยเทคโนโลยีที่เกี่ยวข้องต่าง ๆ ทั้งการสื่อสารข้อมูลแบบไร้สาย (Near Field Communication: NFC), บัตรชิปการ์ด (Smart Card), แถบแม่เหล็ก (Magnetic Strip) รวมถึงเทคโนโลยี OCR ที่ใช้สำหรับการประมวลผลข้อมูล โดยมีรายงานจากเว็บไซต์ของ Global Market Insights ถึงขนาดตลาดของการใช้เทคโนโลยีเหล่านี้ในระบบดังกล่าว ตั้งแต่ปี พ.ศ.2555 และมีการคาดการณ์มูลค่าถึงปี พ.ศ.2566 อีกทั้งรายงานของเว็บไซต์ Grand View Research ระบุถึงมูลค่าตลาดทั่วโลกของ OCR อยู่ที่ 7.46 พันล้านดอลลาร์สหรัฐในปี พ.ศ.2563 และคาดการณ์ว่าจะมีอัตราการเติบโตปีละ 16.7% จากปี พ.ศ.2564 ถึง พ.ศ.2571 ซึ่งเป็นการยืนยันว่า OCR ได้เข้ามามีบทบาทกับงานด้านต่าง ๆ ทั่วโลก รวมไปถึงงานด้านสาธารณสุข

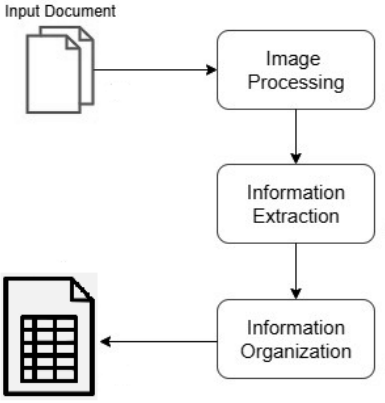
ในปี พ.ศ.2555 Chirag Patel, Atul Patel และ Dharmendra Patel ได้นำเสนอผลงานวิจัยชื่อว่า “Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study” [3] ซึ่งอธิบายถึงการใช้ OCR เพื่อสแกนและสกัดข้อความจากป้ายทะเบียนรถ โดยใช้เครื่องมือชื่อว่า Tesseract นอกจากนี้ G. Vamvakas, B. Gatos, N. Stamatopoulos และ S.J. Perantonis ได้นำเสนอผลงานวิจัยชื่อ “A Complete Optical Character Recognition Methodology for Historical Documents” ซึ่งนำเสนอในปี พ.ศ.2551 [1] งานวิจัยนี้อธิบายเทคนิค OCR โดยมีขั้นตอนที่กล่าวถึงในงานวิจัย ซึ่งสองขั้นตอนแรกเป็นการนำเสนอการใช้ชุดเอกสารเพื่อสร้างฐานข้อมูลสำหรับการ training และขั้นตอนที่สามคือการจดจำเอกสารใหม่

แม้ว่าจะมีงานวิจัยจำนวนไม่มากนักที่ใช้อัลกอริทึม OCR กับเอกสารที่ไม่ใช่ภาษาอังกฤษ เนื่องจากความซับซ้อนทางภาษาและความยากในการจดจำอักขระหรือข้อความ แต่ในความเป็นจริงแล้ว การใช้ OCR ส่วนมากยังต้องการการพัฒนาสำหรับอ่านภาษาอื่น ๆ เช่น ภาษาจีน ภาษาญี่ปุ่น ภาษาเกาหลี และภาษาไทย เป็นต้น ในปี พ.ศ.2542 คุณชาญฤทธิ์ สันตินานาเลิศ ได้นำเสนอวิทยานิพนธ์เรื่องการออกแบบและพัฒนาโปรแกรมโอซีอาร์ภาษาไทย [4] โดยอธิบายเกี่ยวกับ Thai-Optical Character Recognition (Thai-OCR) และงานพัฒนาระบบโดยการสร้างเอกสารในรูปแบบภาพ และแสดงความถูกต้องของการประมวลผล OCR กับภาษาไทย

ระเบียบวิธีวิจัย

สำหรับกระบวนการที่นำเสนอในวิจัยฉบับนี้ มีวัตถุประสงค์เพื่อให้จำแนกตัวอักษรภาษาไทยจากเอกสารอินพุต โดยประมวลผลด้วยโปรแกรม python และใช้เครื่องมือ OCR โอเพนซอร์ส (Open-source software) ที่รู้จักกันในชื่อ “Tesseract” [3]

การดำเนินการสกัดข้อมูลตามขั้นตอนหรือกระบวนการที่นำเสนอในงานวิจัยนี้ จะประกอบไปด้วย 5 ขั้นตอนหลัก ได้แก่ (ก) การสแกนเอกสารอินพุต (ข) การประมวลผลภาพ (ค) การสกัดข้อมูล (ง) การจัดระเบียบข้อมูล และ (จ) การจัดเก็บข้อมูลในรูปแบบที่มีโครงสร้าง เช่น CSV หรือ Excel



รูปที่ 2. ขั้นตอนการทำงาน

ก. เอกสารอินพุต (Input Document)

ขั้นตอนแรก คือการอัปโหลดเอกสารอินพุต (เช่น ไฟล์ PDF รูปภาพ หรือเอกสารประเภทอื่น ๆ) เข้าสู่ระบบ โดยเอกสารเหล่านั้น จะถูกเปลี่ยนให้เป็นรูปแบบดิจิทัล อย่างไรก็ตาม ข้อมูลที่ต้องการจากเอกสารนั้นยังไม่สามารถทำการแก้ไขหรือค้นหาได้ และไม่สามารถที่จะใช้ข้อมูลสำหรับการวิเคราะห์เพิ่มเติม


แบบฟอร์มลงทะเบียนผู้ป่วยใหม่

โปรดกรอกข้อมูลตามแบบฟอร์มให้ครบถ้วนชัดเจน

ข้อมูลผู้ป่วย

เลขที่บัตรประจำตัวประชาชน 3119900123455

คำนำหน้าชื่อ นาย ชื่อ สาริณ นามสกุล อิทธิไพบูล

วัน เดือน ปี เกิด 17 ส.ค. 2539 อายุ 25 ปี เชื้อชาติ ไทย สัญชาติ ไทย

ศาสนา พุทธ เพศ ชาย อาชีพ วิศวกร สถานภาพ โสด

ที่อยู่ปัจจุบันเลขที่ 1/1 หมู่ - ซอย สาริณ ถนน พหลโยธิน

ตำบล สามเสนใน อำเภอ พญาไท จังหวัด กรุงเทพ รหัสไปรษณีย์ 10400

โทรศัพท์(บ้าน) - โทรศัพท์(มือถือ) 0919957890 อีเมล์ sarinn@gmail

ชื่อ - นามสกุล บิดา นวริน อิทธิไพบูล ชื่อ – นามสกุล มารดา ดารัณ อิทธิไพบูล

ผู้ติดต่อได้กรณีฉุกเฉิน ดารัณ อิทธิไพบูล

ที่อยู่ผู้ติดต่อได้ 1/1 ซ.สาริณ ถ.พหลโยธิน แขวงสามเสนใน เขตพญาไท กรุงเทพมหานคร

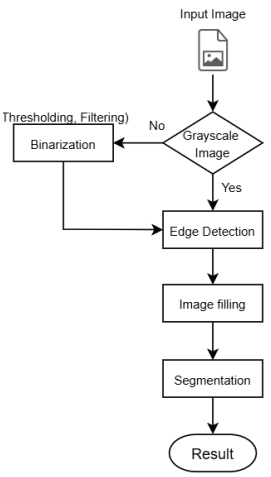
โทรศัพท์(บ้าน) - โทรศัพท์(มือถือ) 0635604544

รูปที่ 3. ตัวอย่างเอกสารอินพุต (แบบฟอร์มลงทะเบียนผู้ป่วย)

รูปที่ 3 แสดงตัวอย่างแบบฟอร์มลงทะเบียนผู้ป่วยที่ใช้เป็นเอกสารอินพุต โดยข้อมูลที่ปรากฏในเอกสารเป็นตัวพิมพ์ภาษาไทย และเป็นเอกสารที่มีโครงสร้าง (Structured Document) เนื่องจากอัลกอริทึมที่ใช้จะจดจำรูปแบบของคำ (Pattern of Words) เพื่อการสกัดข้อมูล จะเห็นว่ารายละเอียดของตัวอย่างเอกสารข้างต้น ประกอบไปด้วยข้อมูลต่าง ๆ ของผู้ป่วย เช่น ชื่อ-นามสกุล วันเกิด อายุ อาชีพ ที่อยู่ และหมายเลขโทรศัพท์ เป็นต้น เพื่อให้สามารถเข้าถึงข้อมูล (Data Accessibility) เหล่านี้ได้ จำเป็นต้องมีการประมวลผลภาพ ซึ่งจะอธิบายในขั้นตอนต่อไป

ข. การประมวลผลภาพ (Image Processing)

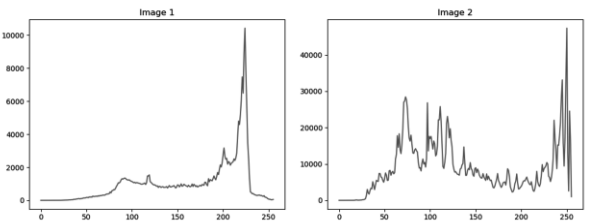
ขั้นตอนการประมวลผลภาพ หรือ Image Processing เป็นกระบวนการในการนำภาพมาประมวลผลเพื่อให้ได้ข้อมูลในรูปแบบดิจิทัล หรือข้อมูลที่ต้องการทั้งในเชิงคุณภาพและปริมาณ โดยจะทำให้ภาพมีความละเอียดมคมชัดมากขึ้น ขจัดสัญญาณรบกวนออกจากภาพ และปรับปรุงมาตรฐานของภาพ การดำเนินการในขั้นตอนนี้ ภาพที่เป็นเอกสารอินพุตจะถูกตรวจสอบก่อนว่าเป็นภาพสีหรือภาพขาวดำ ดังแสดงในรูปต่อไปนี้



รูปที่ 4. ผังงานการประมวลผลภาพ
(Image Processing Flowchart)

รูปที่ 4 แสดงขั้นตอนการประมวลผลภาพ จะเห็นว่าการแปลงภาพให้เป็นภาพขาวดำ หรือ Binarization สามารถแบ่งออกได้เป็น 2 กระบวนการหลัก คือ Thresholding ซึ่งเป็นวิธีที่ใช้สร้างภาพขาวดำ (Binary Image) [5] และ Filtering คือการกรองข้อมูลภาพ ซึ่งเป็นเทคนิคที่ใช้ในการปรับปรุงภาพให้ดีขึ้น ส่งผลให้คอมพิวเตอร์สามารถกำหนดขอบเขตของวัตถุในภาพ (Edge Detection) และเติมสีภาพ (Image Filling) โดยภาพขาวดำจะถูกกำหนดพื้นที่เฉพาะและเติมด้วยสีที่เลือก จากนั้นจะเป็นการใช้วิธีการตัดคำ (Segmentation) [2],[5,[6] เพื่อแบ่งข้อความภาษาไทยออกเป็นบรรทัด และแยกคำแต่ละคำ ซึ่งจะช่วยให้กระบวนการสกัดข้อมูลประมวลผลได้ง่ายมากขึ้น โดยเทคนิคการตัดคำภาษาไทย (Thai Text Segmentation) สามารถแบ่งออกได้เป็น 3 แนวทางหลัก ๆ ได้แก่ Rule Base, Dictionary Base และ Machine Learning Base [7] การใช้ Rule Base คือการสร้างกฎหรือเงื่อนไขเพื่อตัดคำตามโครงสร้างของภาษา ในขณะที่ Dictionary Base คือการตัดคำแบบอิงพจนานุกรมข้อมูล และการใช้ Machine Learning Base (หรือ Statistic Base) คือการนำอัลกอริทึมของ ML มาใช้ เช่น การทำ Part-of-speech tagging คือการกำกับโครงสร้างประโยค หรือการใช้ชุดคลังข้อมูลในการสร้างโมเดลทางสถิติ เพื่อให้โปรแกรมสามารถระบุขอบเขตของคำ (Word Boundaies) ในเอกสารได้

อย่างไรก็ตาม เพื่อให้อัตราความถูกต้องของผลลัพธ์จากการประมวลผลสูงขึ้น สามารถเพิ่มประสิทธิภาพของกระบวนการประมวลผลภาพ รวมถึงคุณภาพของเอกสารอินพุตให้สูงขึ้นได้ด้วยวิธีการต่าง ๆ เช่น การตัดขอบให้มีความคมชัดมากขึ้น (Unsharp Mask Filtering) การแปลงค่าความเข้มของภาพ (Intensity Transformation) การปรับค่าระดับความเข้มของภาพแบบเส้นตรง (Linear Contrast Stretch) การปรับค่าความเข้มแสงของภาพแบบ Non-Linear Contrast Stretch (หรือ Histogram Equalization) เป็นต้น รูปที่ 5 แสดงตัวอย่างกราฟฮีสโตแกรม (Histogram) หากภาพมีโทนสีเข้ม ความถี่ในกราฟทางด้านซ้ายจะสูงตามความเข้มของภาพ ภาพที่มีโทนสว่าง ความสูงของกราฟจะอยู่ทางด้านขวาตามความสว่างของภาพ หากภาพมีโทนสีเทามากกราฟจะมีความถี่อยู่ระหว่างกลาง และภาพที่หลากหลายสีกราฟจะมีความถี่แบบกระจาย



รูปที่ 5. ตัวอย่างกราฟฮีสโตแกรม

ฮีสโตแกรมบ่งบอกถึงปริมาณความสว่างของภาพ ซึ่งจะช่วยในการกำหนดช่วงความเข้มของภาพ เพื่อทำการปรับปรุงคุณภาพของอินพุตให้มีความเหมาะสมสำหรับการแสดงผลหรือการประมวลผลต่อไป

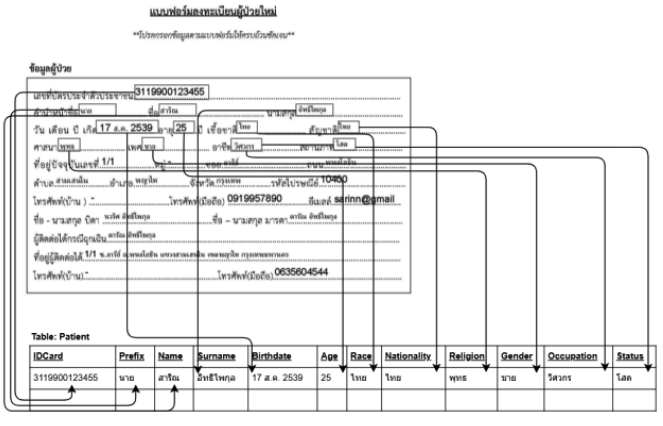
ค. การสกัดข้อมูล (Information Extraction)

กระบวนการสกัดข้อมูล หรือ Information Extraction (IE) [8] เป็นขั้นตอนสำคัญที่ใช้สำหรับงานหรือกรณีที่เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติ (Natural language processing: NLP) โดยกระบวนการนี้จะสกัดข้อมูลที่มีโครงสร้าง (Structured

Data) จากข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) โดยอัตโนมัติ เริ่มต้นจากการใช้ OCR เพื่อสกัดข้อความ และแบ่งออกเป็นคำแต่ละคำโดยใช้กระบวนการตัดคำ (Tokenization) [6] จากนั้นจะเป็นการกำกับคำ (Word Tagging) คือการแปลงข้อความเป็นรายการคำ (List of words) ซึ่งนำไปสู่ขั้นตอนการหาข้อมูลหรือชื่อเฉพาะ เพื่อแยกและกำหนดให้เป็นชื่อเอนทิตีหรือชื่อตาราง (Entity Detection) เช่น แพทย์ แผนก ผู้ป่วย หรือการรับยา เป็นต้น

ง. การจัดระเบียบข้อมูล (Information Organization)

Information Organization (IO) หรือ Knowledge organization (KO) เป็นหนึ่งในขั้นตอนสำคัญในการทำความเข้าใจข้อมูลเชิงลึกและประมวลผลชุดข้อมูลเพื่อช่วยในการตัดสินใจเชิงกลยุทธ์ (Strategic Decision) ข้อมูลที่ได้รับจากการประมวลผลตามกระบวนการก่อนหน้านี้จำเป็นต้องจัดระบบให้อยู่ในอยู่ในรูปแบบที่สามารถจัดการได้ โดยการทำ IO จะรวมถึงการทำดัชนีข้อมูล (Indexing) การเชื่อมโยงข้อมูล (Mapping) การจัดประเภท (Classification) และการจัดหมวดหมู่ (Categorization) สำหรับการจัดเก็บ และการดึงข้อมูล [8] รวมถึงการประเมินและการวิเคราะห์ในลักษณะที่สามารถเข้าใจได้ง่าย เมื่อพิจารณาจากเอกสารนำเข้า จะเห็นว่าเอนทิตีที่ปรากฏในเอกสาร คือ ข้อมูลผู้ป่วย (Patient) ซึ่งมีแอททริบิวต์ต่าง ๆ ตัวอย่างเช่น เลขบัตรประจำตัวประชาชน ชื่อ-นามสกุล ที่อยู่ หมายเลขโทรศัพท์ อาชีพ และชื่อผู้ติดต่อ เป็นต้น ในขณะที่ข้อความอื่น ๆ ที่อยู่ในกล่องข้อความหลังแอททริบิวต์เหล่านั้นถูกกำหนดให้เป็นค่าข้อมูลดังรูปที่ 6



รูปที่ 6. ตัวอย่างการสกัดแอททริบิวต์และข้อมูลจากเอกสาร

รูปที่ 6 แสดงตัวอย่างการสกัดข้อมูล และแอททริบิวต์เพื่อนำมาสร้างเป็นข้อมูลในรูปแบบที่มีโครงสร้าง เช่น CSV หรือ Excel โดยแอททริบิวต์จะเป็นส่วนของ Header และ ข้อมูลจะถูกนำเสนอในแต่ละแถวข้อมูล

เมื่อประมวลผลตามกระบวนการทั้งหมดข้างต้นแล้ว ผลลัพธ์ที่ออกมาจะอยู่ในรูปของข้อความที่สามารถค้นหาและแก้ไขได้ ซึ่งจะช่วยให้หน่วยงานด้านสาธารณสุขสามารถจัดระเบียบข้อมูลผู้ป่วยจำนวนมากภายในเอกสารที่มีอยู่ อีกทั้งสามารถจัดเก็บเข้าถึง ประมวลผล และวิเคราะห์ข้อมูลได้

ผลการวิจัย

ในงานวิจัยนี้ ผู้จัดทำได้มีการเก็บเอกสารตัวอย่างที่ใช้เป็นอินพุต จำนวน 20 ตัวอย่าง เพื่อนำมาประมวลผลโดยใช้เทคนิคช่วยแปลงไฟล์เอกสาร ซึ่งผ่านกระบวนการต่าง ๆ ที่ได้อธิบายในระเบียบวิธีการวิจัย ผลลัพธ์จากการประมวลผลแสดงให้เห็นว่า OCR สามารถสกัดข้อความได้ให้ระดับพอใช้ แม้ว่าจะยังมีคำแปลก ๆ อยู่บ้าง แต่อย่างไรก็ตาม อัตราความแม่นยำอาจเพิ่มขึ้นได้ขึ้นอยู่กับกระบวนการประมวลผลภาพ (Image Processing)

และเพื่อประเมินประสิทธิภาพของอัลกอริทึมในการสกัดข้อมูล ผู้วิจัยจึงได้ทำการวัดความถูกต้องของผลลัพธ์ที่ได้ดังแสดงในตารางที่ 1 โดยคำนวณผลลัพธ์ที่ถูกต้องทั้งหมด แสดงตัวเลขเป็นเปอร์เซ็นต์ ซึ่งคำนวณจากจำนวนแอททริบิวต์และจำนวนค่าทั้งหมดในเอกสารอินพุต (รวม 13 attributes และ 13 values)

ตารางที่ 1 การคำนวณค่าความถูกต้องผลลัพธ์

เอกสาร	ความถูกต้อง			
	จำนวน แอททริบิวต์ ที่สกัดได้ถูกต้อง	*อัตรา ความถูกต้อง (%)*	*จำนวนค่า ที่สกัดได้ถูกต้อง*	*อัตรา ความถูกต้อง (%)*
1	8	61.54	8	61.54
2	9	69.23	9	69.23
3	11	84.62	8	61.54
4	11	84.62	10	76.92
5	11	84.62	7	53.85
6	11	84.62	6	46.15
7	9	69.23	9	69.23
8	9	69.23	10	76.92
9	9	69.23	9	69.23
10	10	76.92	11	84.62
11	9	69.23	8	61.54
12	10	76.92	10	76.92
13	8	61.54	9	69.23
14	8	61.54	9	69.23
15	9	69.23	7	53.85
16	11	84.62	9	69.23
17	10	76.92	11	84.62
18	9	69.23	10	76.92
19	11	84.62	8	61.54
20	11	84.62	10	76.92
ค่าเฉลี่ย		**74.62**		**68.46**

จากตาราง สามารถสรุปได้ว่า OCR จะทำงานได้มีประสิทธิภาพที่ดีขึ้นเมื่อประมวลผลกับเอกสารที่มีคุณภาพสูง ซึ่งจะส่งผลให้อัตราความถูกต้องแม่นยำในการประมวลผลสูงขึ้นด้วย ขึ้นอยู่กับปัจจัยต่าง ๆ ที่ส่งผลต่อการทำงานของ OCR ในแง่ของความสว่าง (Brightness) คอนทราสต์ (Contrast) ความละเอียดภาพ (Resolution) จุดรบกวนในภาพ (Noise) ค่าความตรง (Straightness) และปัจจัยอื่น ๆ ในขณะเดียวกันอัตราความถูกต้องของโมเดลนี้สามารถลดลงได้ ในกรณีที่ข้อความยาวและอักขระมีความคลุมเครืออย่างมาก

บทสรุป

งานวิจัยฉบับนี้ได้อธิบายถึงกระบวนการสกัดข้อมูลผู้ป่วยด้วยเทคโนโลยีที่ช่วยแปลงไฟล์เอกสาร หรือ OCR ซึ่งเป็นเครื่องมือสำคัญที่จะช่วยรองรับและจัดการเอกสารเหล่านั้นให้สามารถเข้าถึงข้อมูลที่จะเป็นประโยชน์ได้ โดยเริ่มต้นจากการสแกนเอกสารอินพุต ปรับปรุงคุณภาพของเอกสารหรือภาพ ประมวลผลผ่านกระบวนการสกัดข้อความ จัดระเบียบข้อมูล และจัดเก็บข้อมูลในรูปแบบที่มีโครงสร้าง จากผลการวิจัยแสดงให้เห็นว่า OCR สามารถประมวลผลและแปลงเอกสารดิจิทัลเพื่อให้สามารถค้นหาและแก้ไขข้อความจากเอกสารได้ ผู้วิจัยได้ประเมินประสิทธิภาพของกระบวนการสกัดข้อมูล โดยคำนวณอัตราความถูกต้อง คือ 74.62% สำหรับการสกัดข้อมูลแอททริบิวต์ และ 68.46% สำหรับการสกัดข้อมูลที่เป็นค่า จากแบบฟอร์มลงทะเบียนผู้ป่วย ซึ่งความถูกต้องของผลลัพธ์จะขึ้นอยู่กับคุณภาพของเอกสาร เทคนิคการสกัดข้อความที่นำเสนอในผลงานวิจัยฉบับนี้จะเป็นประโยชน์อย่างมากสำหรับหน่วยงานด้านสาธารณสุขในการจัดระเบียบข้อมูลผู้ป่วยจำนวนมาก รวมทั้งการจัดเก็บ การเข้าถึง การประมวลผล และการวิเคราะห์ข้อมูลในมิติอื่น ๆ เพื่อนำไปใช้ประโยชน์ต่อไป

เอกสารอ้างอิง

[1] Vamvakas, G., Gatos, B., Stamatopoulos, N., & Perantonis, S. (2008). A Complete Optical Character Recognition Methodology for Historical Documents. 2008 The Eighth IAPR International Workshop on Document Analysis Systems.

[2] Pai, N., & Kolkure, V., S. (2015). Optical Character Recognition: An Encompassing Review. International Journal of Research in Engineering and Technology (IJRET), Volume-4(Issue-1), 407-409.

[3] Patel, C., Patel, A., & Patel, D. (2012). Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. International Journal of Computer Applications, 55(10), 50–56.

[4] Santinanalert, C. (1999). Design and development of a Thai-OCR Program. Bangkok, Thailand: Chulalongkorn University.

[5] Mithe, R., Indalkar, S., & Divekar, N. (2013). Optical Character Recognition. International Journal of Recent Technology and Engineering (IJRTE), Volume-2 (Issue-1), 72–75.

[6] Chumwatana, T. (2017). A Comparative Study of Clustering Techniques for Non-segmented Language Documents. RANGSIT JOURNAL OF ARTS AND SCIENCES (RJAS), Vol. 7, No. 1. (pp. 11-22).

[7] Todsanai Chumwatana “A Survey of Automatic Indexing Techniques for Thai Text Documents,” In Information Technology Journal, KING MONGKUT’S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK, Thailand Volume17,2013.

[8] Todsanai Chumwatana and Ichayaporn Chuaychoo “Automatic Filtering Non-English Complaint Emails in Tourism Industry Using N-gram Extraction and Classification Techniques”, The 2016 4th International Symposium on Computational and Business Intelligence (ISCBI 2016), pp. 216 – 220, Olten, Switzerland, September 5-7, 2016.