

ระบบเซ็นเซอร์ชื่อบุคคลออกจากเอกสารสแกนคำพิพากษาด้วยปัญญาประดิษฐ์

AI-BASED SYSTEM FOR CENCORING PERSONAL NAMES FROM SCANNED DOCUMENTS OF COURT SENTENCES

วิสูตร เหล่าดารา¹

ดร. ธนภัทร มั่งคะจิตร²

บทคัดย่อ

สำนักงานศาลยุติธรรมมีการจัดเก็บคำพิพากษามีการจัดเก็บคำพิพากษานับเต็มในคดีรวบรวมจากการสแกนเอกสารผ่านสแกนเนอร์ในรูปแบบไฟล์ PDF ที่มีลักษณะเป็นไฟล์รูปภาพสำหรับใช้เป็นข้อมูลไฟล์อิเล็กทรอนิกส์และให้บริการข้อมูลที่เกี่ยวข้องกับคดี แต่ทั้งนี้มีความจำเป็นและความต้องการนำข้อมูลคำพิพากษาคัดเลือกเผยแพร่ให้ประชาชนทั่วไปสามารถใช้เป็นแหล่งข้อมูลและการศึกษา เนื่องจากไม่สามารถเปิดเผยชื่อบุคคลในคดีสู่สาธารณะ กระบวนการนี้จึงต้องใช้เจ้าหน้าที่ในการย่อเนื้อหาของคำพิพากษาเพื่อตัดชื่อบุคคลหรืออ้างถึงบุคคลที่เกี่ยวข้องออกจึงนำออกเผยแพร่ได้ ซึ่งใช้ระยะเวลาและเจ้าหน้าที่ผู้ปฏิบัติในการดำเนินการ การศึกษานี้จึงมีวัตถุประสงค์เพื่อจัดทำระบบกระบวนการที่นำเทคโนโลยีมาใช้ตรวจสอบชื่อบนเอกสารคำพิพากษา นำมาช่วยผู้ปฏิบัติงานในการดำเนินการย่อคำพิพากษาและลดระยะเวลาและแบ่งเบาภาระในการดำเนินการของเจ้าหน้าที่ผู้ปฏิบัติงาน ในงานวิจัยนี้ จึงนำไฟล์คำพิพากษาที่มีลักษณะเป็นรูปภาพมาเข้ากระบวนการ Optical Character Recognition ทำการอ่านข้อความในเอกสารออกมาเป็นข้อความ ที่สามารถนำมาคัดแยกได้ด้วยรูปแบบของเอกสารและการจับคำนำหน้าชื่อ และใช้ Named entity recognition เพื่อช่วยแยกชื่อบุคคลออกจากข้อความธรรมดา ซึ่งผลการวิจัยสามารถช่วยให้ตำแหน่งของชื่อบุคคลในคำพิพากษาและปกปิดข้อความส่วนใหญ่ได้

คำสำคัญ: คำพิพากษา, Optical Character Recognition, Named entity recognition

¹ นักศึกษาหลักสูตรวิศวกรรมมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่

² ที่ปรึกษาสารนิพนธ์

ABSTRACT

The Office of the Court of Justice stores judgments, stores full judgments in cases compiled from scanned documents via scanners in the form of PDF files that look like image files for use as electronic file data, and provides services for parties related to the case. However, there are tasks and demands to disseminate judgment information to the general public for use as a source of information and education. Because the name of the person in the case cannot be disclosed to the public this process therefore requires officials to abbreviate the content of the judgment in order to remove the name of the person or refer to the person involved, so that it can be published. Which takes time and staff to perform the process? This study aims to establish a process system that uses technology to verify names on judgment documents. Brought to help operators in executing judgments and reducing time and lighten the burden of the operators. In this research, use the judgment file image file to the Optical Character Recognition process to get text in the document. That can be extracted person's name from a single sentence in document format and split from the title's person, and named entity recognition is used to help identify people's names from plain text. It can help the placement of a person's name in the judgment and conceal most of the text.

Keywords: Judgment, Optical Character Recognition, Named entity recognition

บทนำ

สำนักงานศาลยุติธรรมมีการจัดเก็บคำพิพากษามีการจัดเก็บคำพิพากษาฉบับเต็มในคดีรวบรวมจากการสแกนเอกสารผ่านสแกนเนอร์ในรูปแบบไฟล์ PDF ที่มีลักษณะเป็นไฟล์รูปภาพสำหรับใช้เป็นข้อมูลไฟล์อิเล็กทรอนิกส์และให้บริการสู่ความที่เกี่ยวกับคดี แต่ทั้งนี้มีความและความต้องการนำข้อมูลคำพิพากษาคัดเผยแพร่ให้ประชาชนทั่วไปสำหรับใช้เป็นแหล่งข้อมูลและการศึกษา เนื่องจากไม่สามารถเปิดเผยชื่อบุคคลในคดีสู่สาธารณะ กระบวนการนี้จึงต้องใช้เจ้าหน้าที่ในการย่อเนื้อหาของคำพิพากษาเพื่อตัดชื่อบุคคลหรืออ้างถึงบุคคลที่เกี่ยวข้องออกจึงนำออกเผยแพร่ได้ ซึ่งใช้ระยะเวลาและเจ้าหน้าที่ผู้ปฏิบัติงานในการดำเนินการ ด้วยข้อมูลเดิมเป็นไฟล์คำพิพากษาที่มีการสแกนจัดเก็บไว้อยู่แล้ว นำเทคโนโลยีมาเพื่อลดกระบวนการของเจ้าหน้าที่ ลดขั้นตอนงานและนำข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์ จากข้อมูลตั้งต้นเป็นเอกสารสแกนมีลักษณะเป็นรูปภาพไม่ได้เป็นข้อความประมวลผล จึงต้องแยกข้อความออกมาด้วยการทำ Optical Character Recognition (OCR) เมื่อได้ข้อความมาแล้วจึงนำมาแยกค้นหาชื่อบุคคลที่เป็นชื่อเฉพาะด้วยเทคนิคต่างๆ ได้แก่ การค้นหาคำนำหน้า ค้นจากตำแหน่งบนเอกสาร รวมถึงการใช้ Named-entity recognition (NER) แยกชื่อออกจากข้อความทั่วไป ซึ่ง

กระบวนการดังกล่าวข้างต้นเป็นการทำงานด้วยคอมพิวเตอร์ช่วยแบ่งเบาภาระของเจ้าหน้าที่ในการดำเนินการจัดทำข้อมูล และนำข้อมูลไปใช้ประโยชน์ได้ต่อไป

วัตถุประสงค์ของงานวิจัย

การวิจัยนี้จึงมีวัตถุประสงค์เพื่อจัดทำระบบกระบวนการที่นำเทคโนโลยีมาใช้ตรวจสอบข้อบกพร่องเอกสารคำพิพากษา นำมาช่วยผู้ปฏิบัติงานในการดำเนินการย่อคำพิพากษาและลดระยะเวลาและแบ่งเบาภาระในการดำเนินการของเจ้าหน้าที่ผู้ปฏิบัติงานต่อไป

ขอบเขตงานวิจัย

ขอบเขตของการวิจัย การวิจัยครั้งนี้มีขอบเขตดังนี้

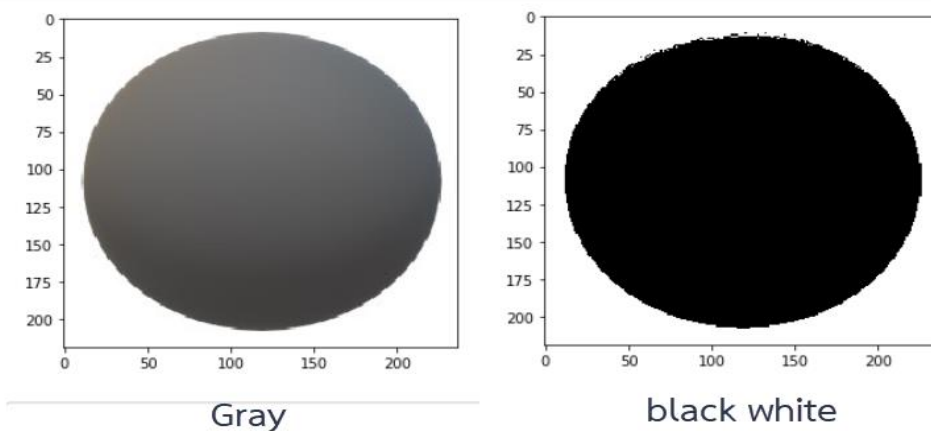
1. นำไฟล์คำพิพากษาที่จัดเก็บไว้มาเข้ากระบวนการ เพื่อได้ชื่อและตำแหน่งพิกัดของข้อมูล บนรูปเอกสารคำพิพากษา และนำไปปกปิดข้อความต่อไป
2. คำพิพากษาที่จัดทำขึ้นมีหลากหลายรูปแบบ เช่นเป็นคำสั่งศาล หรือแบบฟอร์มที่มีการเขียนด้วยลายมือ ในงานวิจัยนี้จึงใช้คำพิพากษาทั่วไปอ้างอิงตามแบบพิมพ์ของสำนักงานศาลหมายเลข31 ที่ได้จัดทำเมื่อคดีถึงที่สุด ซึ่งมีการพิมพ์ผ่านเครื่องคอมพิวเตอร์
3. เป็นคำพิพากษาที่มีการตัดสินในปี 2562 และสแกนเป็นรูปภาพจัดเก็บในระบบเป็นไฟล์ PDF มีจัดเก็บแยกตามเลขคดี แต่ไม่มีข้อมูลคำพิพากษาลบเบียด

ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

Image processing

การประมวลผลภาพเป็นรูปแบบการนำเทคนิค ขั้นตอนแบบต่างๆจัดการกับรูปภาพผ่านอัลกอริทึม (algorithm) ทั้งในลักษณะของภาพสองมิติหรือวิดีโอภาพเคลื่อนไหว ในงานวิจัยนี้ใช้ภาษา Python ในการพัฒนาจึงมีการใช้คลังโปรแกรม(Library)ที่เกี่ยวข้องกับการประมวลผลภาพ ดังนี้

- 1) OpenCV เป็น Open-source library สำหรับการประมวลผลภาพด้วยคอมพิวเตอร์ Machine learning และ image processing ใช้ประมวลผลรูปภาพและวิดีโอ เพื่อตรวจจับวัตถุ ใบหน้ารวมถึงลายมือ
- 2) Pillow (Python imaging Library) ใช้สำหรับจัดการภาพรองรับไฟล์รูปภาพหลากหลายประเภท ถูกออกแบบมาเพื่อข้อมูลที่จัดเก็บในรูปแบบภาพพิกเซลได้อย่างดีและมีเครื่องมือพื้นฐานสำหรับการจัดการรูปภาพ ตัวอย่างที่ใช้งานวิจัยนี้คือการปรับภาพเป็นภาพสีขาวดำเพื่อเพิ่มประสิทธิภาพในการทำ OCR

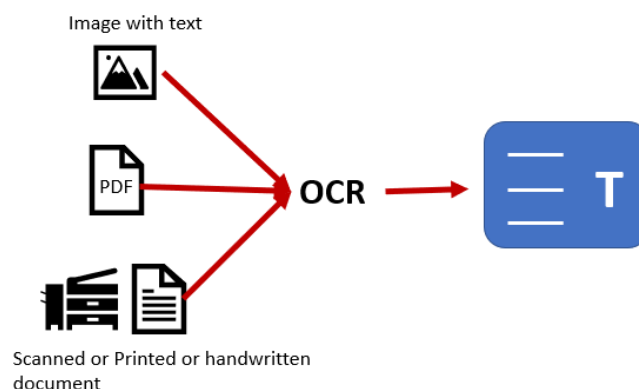


ภาพที่ 1 image processing แปลงสีของรูปภาพเป็นขาวดำ

3) pdf2image เนื่องจากข้อมูลตั้งต้นอยู่ในรูปแบบไฟล์ PDF โมดูลนี้จึงนำเพื่อใช้แปลงไฟล์ในรูปแบบไฟล์รูปภาพ

OCR (optical character recognition)

OCR ย่อมาจาก Optical Character Recognition เป็นเทคโนโลยีที่ช่วยให้ระบบคอมพิวเตอร์อ่านข้อความจากรูปภาพ ออกมาอยู่ในรูปแบบของข้อความ ชุดของตัวอักษรที่สามารถนำไปจัดเก็บเพื่อสืบค้นเรียกข้อมูล ทั้งนี้ข้อมูลที่อ่านจากเอกสารอาจจะไม่สามารถอ่านได้ 100 % เนื่องจากโมเดลที่ใช้ สภาพของไฟล์เอกสาร ความคมชัด รายละเอียดซึ่งมีผลต่อการอ่านข้อความ



ภาพที่ 2 กระบวนการทำงานของ OCR

ในการวิจัยนี้ใช้ Library OCR ที่มีผู้พัฒนาแล้วมาใช้งาน โดยเลือกมา 2 ตัวที่มีความสามารถอ่านภาษาไทยได้จากการมีผู้พัฒนาโมเดลที่รองรับภาษาไทย ได้แก่

1. Tesseract OCR เป็น OpenSource ที่เดิมพัฒนาโดยบริษัท Hewlett-Packard เป็น OCR Engine มุ่งเป้าไปที่การอ่านเป็นบรรทัดปัจจุบันพัฒนามาถึงเวอร์ชัน 5 รองรับยูนิโค้ด (UTF-8) และภาษาต่างๆมากกว่า 100 ภาษา เขียนโปรแกรมด้วยภาษา Python รองรับไฟล์หลายประเภทได้แก่ jpeg, png, gif, BMP ฯลฯ

2. easyOCR เป็นหนึ่งใน OpenSource สำหรับ OCR ที่ดีที่สุดรองรับได้หลายภาษา รองรับการประมวลผลผ่านหน่วยประมวลผลภาพกราฟิก(GPU) โดยผู้พัฒนาระบุว่าถ้าทำงานผ่าน หน่วยประมวลผลภาพกราฟิก(GPU) จะทำงานได้เร็วกว่าเมื่อเทียบกับ หน่วยประมวลผลกลาง(CPU) การใช้งาน EasyOCR ด้วยภาษา Python

NER (named-entity recognition)

Named-entity recognition หรือ NER เป็นกระบวนการ natural language processing ประเภทหนึ่ง หัวข้อย่อยจากเรื่อง Information extraction เพื่อกั้นระบุแยกแยะ เป็นเทคโนโลยีที่ช่วยแยกแยะกลุ่มประเภทของคำบนข้อความ [3]

การทำงานของ NER มีพื้นฐานอยู่บนหลักไวยากรณ์ของภาษาและโมเดลที่สร้างขึ้นเป็นอัลกอริทึมจากการ Train กับข้อมูลของภาษานั้น ที่ระบุแยกแยะเป็น บุคคล องค์กร ค่าเงิน สถานที่ ช่วยให้โมเดลระบุรายละเอียดได้ถูกต้องยิ่งขึ้น

การตรวจจับเอกลักษณ์ของคำแต่ละคำและระบุจะแบ่งประเภทคำแต่ละคำเป็นกลุ่มด้วยเทคนิคที่เรียกว่า BIO กำหนดตำแหน่งของคำในเอกลักษณ์ที่ตรวจพบ โดย B หมายถึงคำอยู่ต้นข้อความ I หมายถึงคำอยู่ข้างในข้อความ และ O หมายถึงคำอยู่นอกข้อความ ซึ่งจะนำประเภทเอกลักษณ์ที่พบมาจัดทำเป็นข้อมูลที่มีรูปแบบหรือใช้ฝึกโมเดลให้มีความแม่นยำขึ้น

ตารางที่ 1 เอกลักษณ์ที่พบจากการตรวจจับด้วย Named-entity recognition

Name	B-NAME, I-NAME, O-NAME
DES (designation)	B-DES, I-DES, O-DES
PHONE	B-PHONE, I-PHONE, O--PHONE

ORG(Organisation)	B-ORG, I-ORG, O-ORG
EMAIL	EMAIL
WEB	WEB
O	ไม่พบเอกลักษณ์ คำนี้ไม่พบคำสำคัญ

งานวิจัยที่เกี่ยวข้อง

An Analysis of the Performance of Named Entity Recognition over OCR'd Documents งานวิจัยนี้กล่าวถึงการนำ named entity recognition มาใช้แยกหมวดหมู่ของเอกสารดิจิทัล ซึ่งพบปัญหาว่าเอกสารดิจิทัลที่ถูกจัดหมวดหมู่แล้วมีข้อมูลที่ผ่านการ OCR มาแล้วมีข้อผิดพลาดจากการอ่านเกิดขึ้น เป้าหมายของงานจึงเป็นการระบุชื่อที่สำคัญและจัดหมวดหมู่ลงในกลุ่มที่เตรียมไว้ (บุคคล สถานที่ องค์กร) งานวิจัยนี้จึงมีการศึกษาความเกี่ยวข้องกันระหว่างความแม่นยำของ NER เกี่ยวพันกับอัตราข้อผิดพลาดจากการ OCR อย่างไร

Named entity extraction from images using natural language processing pipelines ผู้วิจัยมุ่งไปที่ Named entity extraction เป็นกระบวนการที่ใช้แยกเอกลักษณ์หรือชิ้นส่วนข้อมูลจากส่วนของข้อความหรือรูปภาพ ตัวอย่างเช่นการระบุถึงผู้สมัครจากเรซูเม่หรือการเลือกชื่อยี่ห้อจากกล่องขนมชีเรียล จึงใช้เทคโนโลยีอย่าง optical character recognition(OCR) และ named entity recognition มารวมกันสร้างเป็นกระบวนการทำงานเพื่อนำมาแยกคำที่เกี่ยวข้องด้วย named entity recognition ในงานวิจัยนี้ได้นำไปใช้กับแถบยาในส่วนของผู้บริโภคและของร้านค้าปลีก

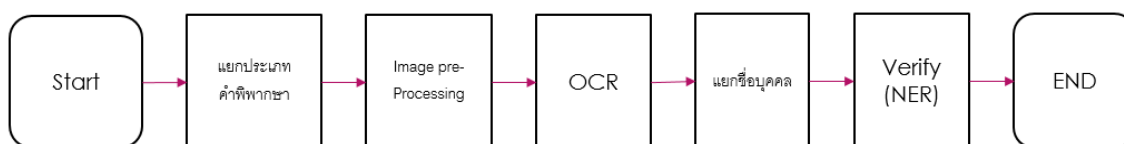
Named Entity Recognition For Scanned Images ผู้วิจัยใช้ Optical Character Recognition(OCR) ในการแยกข้อมูลจากเอกสารที่สแกนหรือไฟล์รูปภาพโดยอัตโนมัติเพื่อเปลี่ยนเป็นข้อความที่คอมพิวเตอร์สามารถอ่านได้และนำมาใช้ประมวลผลข้อมูลต่ออย่างการแก้ไขหรือค้นหาข้อมูล ส่วน Named Entity Recognition (Ner) ใช้เป็นงานรองเพื่อดึงข้อมูลรายละเอียด และประเภทของชื่อออกจากข้อความที่ไม่มีโครงสร้างมาแบ่งเป็นกลุ่มประเภทที่เตรียมไว้แล้วได้แก่ ชื่อบุคคล องค์กร สถานที่ รหัสยา เวลา จำนวน ค่า เปอร์เซ็นต์ ฯลฯ เป็นการเอาข้อมูลสำคัญออกจากรูปภาพ

Comparison of Named Entity Recognition tools for raw OCR text งานวิจัยนี้เป็นการวิเคราะห์ผลการทดลองเพื่อเปรียบเทียบผลของเครื่องมือ Named Entity Recognition (NER) หลายตัว โดยในกระบวนการจะใช้ข้อมูลที่ได้จากผลลัพธ์การทำ Optical Character Recognition(OCR) ทางผู้วิจัยได้กำหนดแบ่งกลุ่มผลลัพธ์เป็น บุคคล(PER) สถานที่(LOC) และองค์กร(ORG)ไว้แล้ว และได้ทดสอบโดยกับเครื่องมือ Named Entity

Recognition (NER) ที่นำมาใช้ดังนี้ 1.OpenNLP 2.Stanford NER 3.Achemy API 4.OpenCalais ผลลัพธ์ที่ได้คือ Stanford NER ให้ผลลัพธ์ที่ดีที่สุด และให้ผลลัพธ์ในกลุ่มบุคคล และกลุ่มสถานที่ได้เป็นอย่างดี Alchemy API ผลลัพธ์ที่ดีที่สุดในกลุ่มองค์กร

Extracting person names from diverse and noisy OCR text งานวิจัยนี้ นำ named entity recognition มาใช้กับเอกสารสแกนที่เป็นเอกสารประวัติศาสตร์เพื่อใช้ประกอบการค้นพบรายละเอียดทางประวัติศาสตร์ อย่างไรก็ตามปัญหาของชุดนี้คือผิดพลาดจากการอ่านคำ ขากที่จะคาดคะเนได้ ผู้วิจัยจึงแก้ปัญหาโดยใช้ อัลกอริทึมในการแยกแยะ 3 รูปแบบแยกแยะชื่อบุคคลออกจากเอกสารและใช้ผลลัพธ์ที่ผลโหวตสูงสุดซึ่งผลลัพธ์ที่ช่วยเพิ่มประสิทธิภาพในการแยกแยะ

ระเบียบวิธีวิจัย



ภาพที่ 3 กระบวนการดำเนินการ

ขั้นตอนการประมวลผลข้อความและดำเนินการสร้างระบบสำหรับตรวจสอบข้อความบนเอกสารได้ตามวัตถุประสงค์ของงานจึงได้กำหนดกระบวนการดำเนินการเพื่อศึกษาข้อมูลมีขั้นตอนดังนี้

การสำรวจข้อมูล

ศึกษาลักษณะของข้อมูลที่นำมาใช้โดยข้อมูลดังกล่าว คือคำพิพากษาที่ได้มีการจัดทำแล้วถูกสแกนผ่านเครื่องสแกนเนอร์ อพโหลดเข้ามาจัดเก็บในระบบที่มีการจัดทำขึ้นเพื่อรวบรวมคำพิพากษาในระบบ ในรูปแบบไฟล์ PDF ที่มีลักษณะเป็นรูปภาพ (Image file) ดังนั้นกระบวนการประมวลผลจากรูปภาพ(Image processing) การนำข้อมูลรูปภาพมาประมวลผลเพื่อแปลงข้อมูลให้ระบบคอมพิวเตอร์แยกแยะได้

Image processing จากการสำรวจข้อมูลไฟล์คำพิพากษาจึงสรุปได้ว่าไฟล์คำพิพากษาที่ถูกจัดเก็บโดยการสแกนจากต้นทางมีลักษณะดังนี้

- 1) เนื่องจากเป็นเอกสารที่สแกนเข้ามาจึงมีขนาดสัดส่วนรูปใกล้เคียงกัน
- 2) มีลายน้ำชื่อศาลวางเป็นแนวทแยงอยู่ทุกหน้าของเอกสาร

3) เอกสารมีรูปแบบฟอร์มในการจัดทำ แบ่งประเภทที่พบบ่อยได้แก่ คำพิพากษา คำสั่ง สัญญา ประณิษระยอม ซึ่งในรูปแบบฟอร์มศาล ส่วนหัวของเอกสารจะมีรูปแบบเฉพาะใกล้เคียงกัน ประกอบด้วยเลขคดี และคู่กรณีในคดีได้แก่ โจทก์ จำเลย หรือผู้ร้อง

(๓๑)
คำพิพากษา

เลขคดี สำหรับศาลใช้
คดีหมายเลขคำที่ ๒๐๐๐/๒๕๕๐
คดีหมายเลขแดงที่ ๑๐๐๐/๒๕๕๐

ประเภหคำพิพากษา

ในพระปรมาภิไธยพระมหากษัตริย์

ศาลจังหวัดฉะเชิงเทรา

วันที่ ๒๘ เดือน มีนาคม พุทธศักราช ๒๕๖๐

คู่ความในคดี ความอาญา

ระหว่าง

พนักงานอัยการจังหวัดฉะเชิงเทรา โจทก์

จำเลย

เรื่อง ความผิดต่อพระราชบัญญัติยาเสพติดให้โทษ ความผิดต่อพระราชบัญญัติจราจรทางบก

โจทก์ฟ้อง จำเลยให้การรับสารภาพ ขอให้จึงจริงฟังได้ว่า เมื่อระหว่างวันที่ ๔ กุมภาพันธ์ ๒๕๖๐ ถึงวันที่ ๖ กุมภาพันธ์ ๒๕๖๐ เวลาากลางคืนหลังเที่ยง วันเวลาใดไม่ปรากฏชัด

จำเลยเสพและจำหน่ายยาเสพติดให้โทษในประเภท ๑ จำนวน ๑ เม็ด น้ำหนักเท่าใด ไม่ปรากฏชัด โดยวิธีดูดดมครั้นเข้าสู่ร่างกาย แล้วจำเลยซึ่งไม่ได้รับใบอนุญาตขับรถจักรยานยนต์และ

พนักงานสอบสวนเปรียบเทียบปรับแล้ว ได้จับรถจักรยานยนต์ หมายเลขทะเบียน ๑ กญ ๖๙๒๓ ฉะเชิงเทรา ไปตามถนนสาธารณะบริเวณหมู่บ้านแหลมไม้ศรี หมู่ที่ ๓ ตำบลหนองยาว อำเภอพนมสารคาม จังหวัดฉะเชิงเทรา ขณะมีสารเสพติดเมกแอมเฟตามีนดังกล่าวอยู่ในร่างกาย อันเป็นการฝ่าฝืนต่อกฎหมาย

เหตุเกิดที่ตำบลหนองยาว อำเภอพนมสารคาม จังหวัดฉะเชิงเทรา ก่อนคดีนี้ เมื่อวันที่ ๑๓ มกราคม ๒๕๕๗

จำเลยเคยต้องคำพิพากษาถึงที่สุดให้ลงโทษจำคุก ๔ ปี ปรับ ๒๐๐,๐๐๐ บาทฐานมีเมกแอมเฟตามีนไว้ในครอบครองเพื่อจำหน่ายโดยมิได้รับอนุญาต ตามคดีอาญานหมายเลขแดงที่ ๑๔๔/๒๕๕๗ ของศาลนี้

/จำเลย...

ภาพที่ 4 รูปแบบคำพิพากษา

OCR ใช้กระบวนการ OCR เพื่อแปลงข้อมูลจากเอกสารสแกนที่มีลักษณะรูปภาพเป็นข้อความ โดยมีขั้นตอนดังนี้

- 1) แปลงภาพต้นฉบับ เป็นภาพขาวดำ เพื่อเพิ่มความแม่นยำในการทำ OCR
- 2) ได้ทดลองใช้งาน LLibrary OCR ได้แก่ tesseractOCR และ easyOCR พบว่า easyOCR นอกจากมีความแม่นยำในการระบุข้อความภาษาไทยได้ถูกต้องกว่าแล้ว สามารถระบุพิกัดตำแหน่งของประโยคในเอกสารได้

(๓๓) คำพิพากษา

ส่วนรับศาลใช้

คดีหมายเลขคดีที่

คดีหมายเลขแดงที่

ในพระปรมาภิไธยพระมหากษัตริย์

ศาลจังหวัดขอนแก่น

วันที่ ๒๕ เดือน ธันวาคม พุทธศักราช ๒๕๖๐

ความว่า

พนักงานอัยการจังหวัดขอนแก่น โจทก์

ระหว่าง

จำเลย

เรื่อง ความผิดของพระราชบัญญัติอาชญากรรมคดีอาญา ความผิดของพระราชบัญญัติอาชญากรรม

โจทก์ฟ้อง จำเลยได้กระทำความผิด ดังนี้จึงจึงรับฟ้องได้ว่า จำเลยกระทำความผิด

ต่อกฎหมายอาชญากรรมคดีอาญา ดังนี้ เมื่อระหว่างวันที่ ๘ ตุลาคม ๒๕๖๐ เวลาประมาณ

วันที่ ๘ ตุลาคม ๒๕๖๐ เวลาประมาณ ๑๖.๐๐ น. จำเลยได้กระทำความผิดโดยนำ

เงินเป็นอาชญากรรมคดีอาญาประเภท ๓ จำนวน ๒๐,๐๐๐ บาท ไปฝากธนาคารออมสิน

ด้วยวิธีนำเงินฝากออมสินมาฝากในบัญชีออมสินที่ชื่อ จำเลย และนำเงินฝากออมสิน

แล้วสุ่มรับเอาเงินนั้นไปใช้จ่ายค่า ค่าใช้จ่ายในการดำเนินคดีอาชญากรรม จำเลยและจำเลย

ประกอบอยู่ อันเป็นอาชญากรรมคดีอาญาประเภท ๒ ไม่ทราบจำนวนและนำเงินคืน

นำเงินคืนมาฝากในบัญชีออมสินที่ชื่อ จำเลย และนำเงินฝากออมสินไปฝาก

เพื่อรักษาไว้ตามคำสั่งของเจ้าพนักงานอัยการหรือผู้ประกอบคดีอาชญากรรมปัจจุบัน

ขึ้นหนังสือคำสั่งของเจ้าพนักงานอัยการที่ได้นำมาฝากในบัญชีออมสินที่ชื่อ จำเลย

๒๕๖๐ เวลาประมาณ ๑๖.๐๐ น. จำเลยได้กระทำความผิดโดยนำเงินฝากออมสิน

๒๕๖๐ นี้อยู่ที่ธนาคารออมสิน จำนวนเงินฝากออมสินที่ชื่อ จำเลยได้

โจทก์

ภาพที่ 5 ตัวอย่างผลลัพธ์จาก OCR

3) เมื่อได้ข้อความทั้งหมดบนเอกสารแล้ว จากนั้นคัดข้อความที่มีลักษณะเป็นชื่อเฉพาะ โดยในการวิจัยนี้ใช้สมมติฐาน 2 กรณี คือ (1) ข้อความประกอบด้วยคำนำหน้าที่ใช้บ่อย ได้แก่ นาย นาง และ (2) ในรูปแบบเอกสารของคำพิพากษานี้แรกจะระบุชื่อ โจทก์ จำเลย หรือ ผู้ร้อง และส่วนมาจากชื่อบุคคลที่ระบุในส่วนนี้ เนื้อหาจะระบุอ้างอิงด้วยข้อความเกี่ยวข้องของคดี ตามด้วยลำดับที่ แต่ทั้งนี้กรณีมีบุคคลที่เกี่ยวข้องแต่ไม่ได้อ้างถึงข้างต้นก็จะใช้ชื่อเต็มคัดข้อความที่เกิน

ข้อเสนอแนะ

- 1 คำในภาษาไทยบางคำมีการเขียนใกล้เคียงกัน การค้นหาด้วยคำนำหน้าจึงอาจจะพบคำที่ไม่เกี่ยวข้องได้ เช่น พบคำว่า “ทนาย” แทนที่จะเป็นชื่อบุคคล
- 2 กระบวนการ OCR ในการประมวลผลนาน ในขั้นตอนติดตั้ง Library ของ easyOCR ควรเตรียมคอมพิวเตอร์มีประสิทธิภาพเพียงพอ และติดตั้ง Coda ให้ประมวลผลผ่านกราฟฟิการ์ด์ ซึ่ง Library แนะนำว่าช่วยให้ประมวลผลได้เร็วขึ้น

บรรณานุกรม

- Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickael Coustaty, Antoine Doucet (2020). An Analysis of the Performance of Named Entity Recognition over OCR'd Documents. Retrieved June 2023, from ACM DL Digital Library website
<https://dl.acm.org/doi/10.1109/JCDL.2019.00057>
- Anirudh Lodh, Utkarsh Saxena, Akshay Markhedkar, Ajmal Khan, Bhavya Manoj Votavat, Akshay Mendon (2022). Named entity extraction from images using natural language processing pipelines. Retrieved June 2023, from IEEE explore website
<https://ieeexplore.ieee.org/document/9917902>
- Anuva Goyal, Kritika (2021). Named Entity Recognition For Scanned Images. Retrieved June 2023, from ResearchGate website
https://www.researchgate.net/publication/357420807_NAMED_ENTITY_RECOGNITION_FOR_SCANNED_IMAGES_PROJECT_REPORT_2021
- JaidedAI/EasyOCR" (Retrieved 23 May 2023) <https://github.com/JaidedAI/EasyOCR>
- Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, Magdalena Luszczynska (2012). Comparison of Named Entity Recognition tools for raw OCR text. Retrieved June 2023, from Österreichische Gesellschaft für Artificial Intelligence website
http://www.oegai.at/konvens2012/proceedings/60_rodriguez12w/60_rodriguez12w.pdf
- Nick Barney, "named entity recognition" (Retrieved 23 May 2023)
<https://www.techtarget.com/whatis/definition/named-entity-recognition-NER> "tesseract-ocr/tesseract" (Retrieved 23 May 2023) <https://github.com/tesseract-ocr/tesseract>

Thomas L. Packer, Joshua F. Lutes, Aaron P. Stewart, David W. Embley, Eric K. Ringger, Kevin D. Seppi,
Lee S. Jensen (2010). Extracting person names from diverse and noisy OCR text. Retrieved June
2023, from ACM DL Digital Livrary website <https://dl.acm.org/doi/10.1145/1871840.1871845>