ระบบเซ็นเซอร์ชื่อบุคคลออกจากเอกสารสแกนคำพิพากษาด้วยปัญญาประดิษฐ์

วิศรุต เหล่าดารา

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2565

SYSTEM FOR CENSORING PERSON'S NAME FROM COURT SENTENCES USING ARTIFICIAL INTELLIGENCE

WITSARUT LAODARA

A Thematic Paper Submitted in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering
Department of Big Data Engineering,
College of Innovative Technology and Engineering
Dhurakij Pundit University
Academic Year 2022

ใบรับรองสารนิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่

หัวข้อสารนิพนธ์ ระบบเซ็นเซอร์ชื่อบุคคลออกจากเอกสารสแกนคำพิพากษาด้วยปัญญาประดิษฐ์

เสนอโดย วิศรุต เหล่าตารา

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาสารนิพนธ์ ดร.ธนภัทร ฆังคะจิตร

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบสารนิพนธ์แล้ว

(ดร.สรรพฤทธิ์ มฤคทัต) ประธานกรรมการ

(ดร.ธนภัทร ฆังคะจิตร) กรรมการที่ปรึกษาสารนิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น) กรรมการ

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ รับรองแล้ว

(ดร.ชัยพร เขมะภาตะพันธ์) คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 31 เดือน กรกฎาคม พ.ศ. 2566

หัวข้อสารนิพนธ์ ระบบเซ็นเซอร์ชื่อบุคคลออกจากเอกสารสแกนคำพิพากษาด้วยปัญญาประดิษฐ์
ชื่อผู้เขียน วิศรุต เหล่าดารา
อาจารย์ที่ปรึกษา ดร. ธนภัทร ฆังคะจิตร
หลักสูตร วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา 2565

บทคัดย่อ

สำนักงานศาลยุติธรรมมีการจัดเก็บคำพิพากษามีการจัดเก็บคำพิพากษาฉบับเต็มในคดีรวบรวมจากการสแกนเอกสารผ่านสแกนเนอร์ในรูปแบบไฟล์ PDF ที่มีลักษณะเป็นไฟล์รูปภาพสำหรับใช้เป็นข้อมูลไฟล์อิเล็กทรอนิกส์และให้บริการคู่ความที่เกี่ยวข้องกับคดี แต่ทั้งนี้มีงานและความต้องการนำข้อมูลคำพิพากษาคดีเผยแพร่ให้ประชาชนทั่วไปสำหรับใช้เป็นแหล่งข้อมูลและการศึกษา เนื่องจากไม่สามารถเปิดเผยชื่อบุคคลในคดีสู่สาธารณะ กระบวนการนี้จึงต้องใช้เจ้าหน้าที่ในการย่อเนื้อหาของคำพิพากษาเพื่อตัดชื่อบุคคลหรืออ้างถึงบุคคลที่เกี่ยวข้องออกจึงนำออกเผยแพร่ได้ ซึ่งใช้ระยะเวลาและเจ้าหน้าที่ผู้ปฏิบัติในการดำเนินการ

การศึกษานี้จึงมีวัตถุประสงค์เพื่อจัดทำระบบกระบวนการที่นำเทคโนโลยีมาใช้ตรวจสอบชื่อบนเอกสารคำพิพากษานำมาช่วยผู้ปฏิบัติงานในการดำเนินการย่อคำพิพากษาและลดระยะเวลาและแบ่งเบาภาระในการดำเนินการของเจ้าหน้าที่ผู้ปฏิบัติงาน

ในงานวิจัยนี้ จึงนำไฟล์คำพิพากษาที่มีลักษระเป็นรูปภาพมาเข้ากระบวนการ Optical Character Recognition ทำการอ่านข้อความในเอกสารออกมาเป็นข้อความ ที่สามารถนำมาคัดแยกได้ด้วยรูปแบบของเอกสารและการจับคำนำหน้าชื่อและใช้ Named entity recognition เพื่อช่วยแยกชื่อบุคคลออกจากข้อความธรรมดา ซึ่งผลการวิจัยสามารถช่วยให้ตำแหน่งของชื่อบุคคลในคำพิพากษาและปกปิดข้อความส่วนใหญ่ได้

คำสำคัญ: คำพิพากษา, การรู้จำอักขระด้วยแสง, การระบุนิพจน์สำคัญ

(อาจารย์ที่ปรึกษา)

Thematic Paper Title	SYSTEM FOR CENSORING PERSON'S NAME FROM COURT SENTENCES USING ARTIFICIAL INTELLIGENCE
Author	WITSARUT LAODARA
Thematic Paper Advisor	Dr.Thanapat Kangkachit
Department	Big Data Engineering
Academic Year	2022

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 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 Optical Character Recognition process to get text in the document. That can be extracted person's name from single sentence with document format and split from 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

(Advisor)

กิตติกรรมประกาศ

ในงานวิจัยฉบับนี้ สำเร็จลุล่วงได้อย่างสมบูรณ์ด้วยความกรุณาอย่างยิ่งจาก ดร.ธนภัทร ฆังคะจิตร ที่ได้สละเวลาอันมีค่าแก่ผู้วิจัย เพื่อให้คำปรึกษาและแนะนำตลอดจนตรวจทางแกไขข้อบกพร่องต่างๆ ด้วยความเอาใจใส่เป็นอย่างยิ่ง จนงานวิจัยฉบับนี้สำเร็จสมบูรณ์ลุล่วงได้ด้วยดีผู้วิจัยขอกรอบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้ จากใจจริง

ขอขอบคุณ เหล่าคณะอาจารย์วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ ที่ได้กรุณาให้คำแนะนำช่วยเหลือ และปรับปรุงงานวิจัยจนเสร็จสมบูรณ์

สุดท้ายนี้ ขออุทิศความดีที่มีในการศึกษาวิจัยนี้แด่ บิดา มารดา ครอบครัวของผู้วิจัย ซึ่งสนับสนุนในทุกด้าน และกำลังใจจากมิตรแท้ทุกท่าน

วิศรุต เหล่าดารา

สารบัญ

สารบัญ (ต่อ)

บทที่ **หน้า**

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

สำนักงานศาลยุติธรรมมีการจัดเก็บคำพิพากษามีการจัดเก็บคำพิพากษาฉบับเต็มในคดีรวบรวมจากการสแกนเอกสารผ่านสแกนเนอร์ในรูปแบบไฟล์ PDF ที่มีลักษณะเป็นไฟล์รูปภาพสำหรับใช้เป็นข้อมูลไฟล์อิเล็กทรอนิกส์และให้บริการคู่ความที่เกี่ยวข้องกับคดี แต่ทั้งนี้มีงานและความต้องการนำข้อมูลคำพิพากษาคดีเผยแพร่ให้ประชาชนทั่วไปสำหรับใช้เป็นแหล่งข้อมูลและการศึกษา เนื่องจากไม่สามารถเปิดเผยชื่อบุคคลในคดีสู่สาธารณะ กระบวนการนี้จึงต้องใช้เจ้าหน้าที่ในการย่อเนื้อหาของคำพิพากษาเพื่อตัดชื่อบุคคลหรืออ้างถึงบุคคลที่เกี่ยวข้องออกจึงนำออกเผยแพร่ได้ ซึ่งใช้ระยะเวลาและเจ้าหน้าที่ผู้ปฏิบัติในการดำเนินการ ด้วยข้อมูลเดิมเป็นไฟล์คำพิพากษาที่มีการสแกนจัดเก็บไว้อยู่แล้ว นำเทคโนโลยีมาใช้เพื่อลดกระบวนการของเจ้าหน้าที่ลดขั้นตอนงานและนำข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์ จากข้อมูลตั้งต้นเป็นเอกสารสแกนมีลักษณะเป็นรูปภาพ ไม่ได้เป็นข้อความประมวลผล จึงต้องแยกข้อความออกมาด้วยการทำ Optical Character Recognition (OCR) เมื่อได้ข้อความมาแล้วจึงนำมาแยกค้นหาชื่อบุคคลที่เป็นชื่อเฉพาะด้วยเทคนิคต่างๆ ได้แก่ การค้นหาคำนำหน้า ค้นจากตำแหน่งบนเอกสาร รวมถึงการใช้ Named-entity recognition (NER) แยกชื่อออกจากข้อความทั่วไป ซึ่งกระบวนการดังกล่าวข้างต้นเป็นการทำงานด้วยคอมพิวเตอร์ช่วยแบ่งเบาภาระของเจ้าหน้าที่ในการดำเนินการจัดทำข้อมูล และนำข้อมูลไปใช้ประโยชน์ได้ต่อไป

1.2 วัตถุประสงค์ของงานวิจัย

การวิจัยนี้จึงมีวัตถุประสงค์เพื่อจัดทำระบบกระบวนการที่นำเทคโนโลยีมาใช้ตรวจสอบชื่อบนเอกสารคำพิพากษา นำมาช่วยผู้ปฏิบัติงานในการดำเนินการย่อคำพิพากษาและลดระยะเวลาและแบ่งเบาภาระในการดำเนินการของเจ้าหน้าที่ผู้ปฏิบัติงาน

1.3 ขอบเขตของงานวิจัย

ขอบเขตของการวิจัย การวิจัยครั้งนี้มีขอบเขตดังนี้

1. นำไฟล์คำพิพากษาที่จัดเก็บไว้มาเข้ากระบวนการ เพื่อได้ชื่อและตำแหน่งพิกัดของชื่อบุคคล บนรูปเอกสารคำพิพากษา และนำไปปกปิดข้อความต่อได้

2. คำพิพากษาที่จัดทำขึ้นมีหลากหลายรูปแบบ เช่นเป็นคำสั่งศาล หรือแบบฟอร์มที่มีการเขียนด้วยลายมือ ในงานวิจัยนี้จึงใช้คำพิพากษาทั่วไปอ้างอิงตามแบบพิมพ์ของสำนักงานศาลหมายเลข 31 ที่ได้จัดทำเมื่อคดีถึงที่สุด ซึ่งมีการพิมพ์ผ่านเครื่องคอมพิวเตอร์

สำหรับศาลใช้

○ (๓๑)
คำพิพากษา

คดีหมายเลขดำที่.....๒๕.....
คดีหมายเลขแดงที่.....๒๕.....

ในพระปรมาภิไธยพระมหากษัตริย์

ศาล.....
วันที่.....เดือน.....พุทธศักราช ๒๕.....

ความ.....

ระหว่าง { ..... โจทก์
..... จำเลย

เรื่อง.....

~เนื้อความ~

ภาพที่ 1.1 ตัวอย่างแบบฟอร์มคำพิพากษา (แบบฟอร์มศาลหมายเลข 31)

3. เป็นคำพิพากษาที่มีการตัดสินในปี 2562 และสแกนเป็นรูปภาพจัดเก็บในระบบเป็นไฟล์ PDF มีจัดเก็บแยกตามเลขคดี แต่ไม่มีข้อมูลคำพิพากษาฉบับเต็ม

1.4 ประโยชน์ที่ได้รับ

การใช้เทคโนโลยีเพื่อลดภาระงานหรือช่วยในการตัดสินใจแทนผู้ปฏิบัติงานมีความสำคัญประกอบกับเทคโนโลยีต่างๆที่ช่วยให้ระบบคอมพิวเตอร์สามารถแยกแยะเพื่อนำมาช่วยงานแทนคนทำงานหรือเจ้าหน้าที่มีการพัฒนามากขึ้นอาทิเช่น Optical Character Recognition(OCR) National Language processing Name entily reconnize และได้ระบบมาใช้ตรวจสอบชื่อบนเอกสารคำพิพากษา นำมาช่วยผู้ปฏิบัติงานในการดำเนินการย่อคำพิพากษาและลดระยะเวลาและแบ่งเบาภาระในการดำเนินการของเจ้าหน้าที่ผู้ปฏิบัติงาน

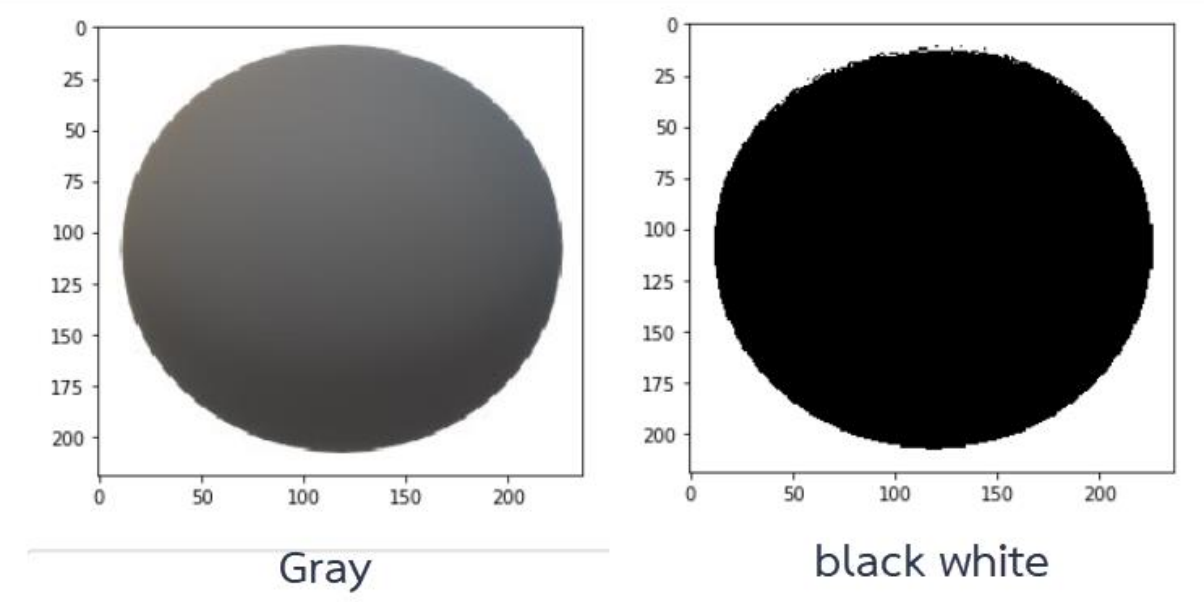
บทที่ 2
แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 Image processing

การประมวลผลภาพเป็นรูปแบบการนำเทคนิค ขั้นตอนแบบต่างๆ มาจัดการกับรูปภาพผ่านอัลกอริทึม (algorithm) ทั้งในลักษณะของภาพสองมิติหรือวีดีโอภาพเคลื่อนไหว ในงานวิจัยนี้ใช้ภาษา Python ในการพัฒนาจึงมีการใช้คลังโปรแกรม(Library)ที่เกี่ยวข้องกับการประมวลผลภาพ ดังนี้

1) OpenCV เป็นOpen-source library สำหรับการประมวลผลภาพด้วยคอมพิวเตอร์ Machine learning และ image processing ใช้ประมวลผลรูปภาพและวีดีโอ เพื่อตรวจจับวัตถุ ใบหน้ารวมถึงลายมือ

2) Pillow (Python imaging Library) ใช้สำหรับจัดการภาพรองรับไฟล์รูปภาพหลากหลายประเภท ถูกออกแบบมาเพื่อข้อมูลที่จัดเก็บในรูปแบบภาพพิกเซลได้อย่างดีและมีเครื่องมือพื้นฐานสำหรับจัดการรูปภาพ ตัวอย่างที่ใช้งานวิจัยนี้คือการปรับภาพเป็นภาพสีขาวดำเพื่อเพิ่มประสิทธิภาพในการทำ OCR

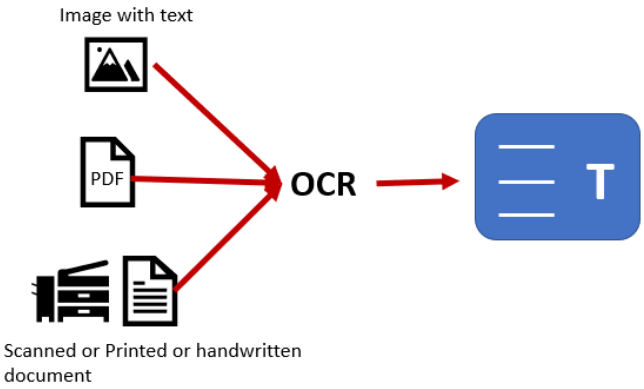


ภาพที่ 2.1 image processing แปลงสีของรูปภาพเป็นขาวดำ

3) Pdf2image เนื่องจากข้อมูลตั้งต้นอยู่ในรูปแบบไฟล์ PDF โมดูลนี้จึงนำเพื่อใช้แปลงไฟล์ในรูปแบบไฟล์รูปภาพ

2.2 Optical Character Recognition (OCR)

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



ภาพที่ 2.2 กระบวนการทำงานของ OCR

ในการวิจัยนี้ใช้ Library OCR ที่มีผู้พัฒนาแล้วมาใช้งาน โดยเลือกมา 2 ตัวที่มีความสามารถอ่านภาษาไทยได้จากการมีผู้พัฒนาโมเดลที่รองรับภาษาไทย ได้แก่

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

2) easyOCR เป็นหนึ่งใน OpenSource สำหรับ OCR ที่ดีที่สุดรองรับได้หลายภาษา รองรับการประมวลผลผ่านหน่วยประมวลผลภาพกราฟิก(GPU) โดยผู้พัฒนาระบุว่าถ้าทำงานผ่าน หน่วยประมวลผลภาพกราฟิก(GPU) จะทำงานได้เร็วเว่าเมื่อเทียบกับ หน่วยประมวลผลกลาง(CPU) การใช้งาน EasyOCR ด้วยภาษา Python [2]

2.3 National language processing

National Language processing เป็นกระบวนการประมวลผลข้อความโดยเน้นไปที่ภาษาที่ใช้ในแต่ละประเทศ และเป็นส่วนหนึ่งของระบบปัญญาประดิษฐ์ (AI) เพื่อให้คอมพิวเตอร์สามารถเข้าใจข้อความและภาษาพูดเช่นเดียวกับมนุษย์ [3]

2.4 Named-entity recognition

Named-entity recognition หรือ NER เป็นกระบวนการ natural language processing ประเภทหนึ่ง หัวข้อย่อยจากเรื่อง Information extraction เพื่อค้นระบุแยกแยะ เป็นเทคโนโลยีที่ช่วยแยกแยะกลุ่มประเภทของคำบนข้อความ

การทำงานของ NER มีพื้นฐานอยู่บน หลักไวยากรณ์ของภาษาและโมเดลที่สร้างขึ้นเป็นอัลกอริทึมจากการ Train กับข้อมูลของภาษานั้น ที่ระบุแยกแยะเป็น บุคคล องค์กร ค่าเงิน สถานที่ ช่วยให้โมเดลระบุรายละเอียดได้ถูกต้องยิ่งขึ้น

การตรวจจับเอกลักษณ์ของคำแต่ละคำและระบุจะแบ่งประเภทคำแต่ละคำเป็นกลุ่มด้วยเทคนิคที่เรียกว่า BIO กำหนดตำแหน่งของคำในเอกลักษณ์ที่ตรวจพบ โดย B หมายถึงคำอยู่ต้นข้อความ I หมายถึงคำอยู่ข้างในข้อความ และ O หมายถึงคำอยู่นอกข้อความ ซึ่งจะนำประเภทเอกลักษณ์ที่พบมาจัดทำเป็นข้อมูลที่มีรูปแบบหรือใช้ฝึกโมเดลให้มีความแม่นยำขึ้น [4]

ตารางที่ 2.1 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

2.5 งานวิจัยที่เกี่ยวข้อง

1) An Analysis of the Performance of Named Entity Recogntion over OCRed Documents [5]

งานวิจัยนี้กล่าวถึงการนำ named entity recognition มาใช้แยกหมวดหมู่ของเอกสารดิจิตัล ซึ่งพบปัญหาว่าเอกสารดิจิทัลที่ถูกจัดหมวดมาแล้วมีข้อมูลที่ผ่านการ OCR มาแล้วมีข้อผิดพลาดจากกการอ่านเกิดขึ้น เป้าหมายของงานจึงเป็นการระบุชื่อที่สำคัญและจัดหมวดหมู่ลงในกลุ่มที่เตรียมไว้ (บุคคล สถานที่ องค์กร) งานวิจัยนี้จึงมีการศึกษาความเกี่ยวข้องกันระหว่าความแม่นยำของ NER เกี่ยวพันกับอัตราข้อผิดพลาดจากการ OCR อย่างไร

2) Named entity extraction from images using natural language processing pipelines [6]

ผู้วิจัยมุ่งไปที่ Named entity extraction เป็นกระบวนการที่ใช้แยกเอกลักษณ์หรือชิ้นส่วนข้อมูลจากส่วนของข้อความหรือรูปภาพ ตัวอย่างเช่นการระบุถึงผู้สมัครจากเรซูเม่หรือการเลือกชื่อยี่ห้อจากกล่องขนมซีเรียล จึงใช้เทคโนโลยีอย่าง optical character recognition(OCR) และ named entity recognition มารวมกันสร้างเป็นกระบวนการทำงานเพื่อนำมาแยกคำที่เกี่ยวข้องด้วย named entity recognition ในงานวิจัยนี้ได้นำไปใช้กับแถบยาในส่วนของผู้บริโภคและของร้านค้าปลีก

3) Named Entity Recognition For Scanned Images [7]

ผู้วิจัยใช้ Optical Character Recognition(OCR) ในการแยกข้อมูลจากเอกสารที่สแกนหรือไฟล์รูปภาพโดยอัตโนมัติเพื่อเปลี่ยนเป็นข้อความที่คอมพิวเตอร์สามารถอ่านได้และนำมาใช้ประมาลผลข้อมูลต่ออย่างการแก้ไขหรือค้นหาข้อมูล ส่วน Named Entity Recognition (Ner) ใช้เป็นงานรองเพื่อดึงข้อมูลรายละเอียด และประเภทของชื่อออกจากข้อความที่ไม่มีโครงสร้างมาแบ่งเป็นกลุ่มประเภทที่เตรียมไว้แล้ว ได้แก่ ชื่อบุคคล องค์กร สถานที่ รหัสยา เวลา จำนวน ค่า เปอร์เซ็น ฯลฯ เป็นการเอาข้อมูลสำคัญออกจากรูปภาพ

4) Comparison of Named Entity Recognition tools for raw OCR text [8]

งานวิจัยนี้เป็นการวิเคราะห์ผลการทดลองเพื่อเปรียบเทียบผลของเครื่องมือ Named Entity Recognition (NER) หลายตัว โดยในกระบวนการจะใช้ข้อมูลที่ได้จากผลลัพท์การทำ Optical Character Recognition(OCR) ทางผู้วิจัยได้กำหนดแบ่งกลุ่มผลลัพท์เป็น บุคคล(PER) สถานที่(LOC) และองค์กร(ORG)ไว้แล้ว และได้ทดสอบโดยกับเครื่องมือ Named Entity Recognition (NER) ที่นำมาใช้ดังนี้ 1.OpenNLP 2.Stanford NER 3.Alchemy API 4.OpenCalais ผลลัพท์ที่ได้คือ Stanford NER ให้ผลลัพท์ดีที่สุด และให้ผลลัพท์ในกลุ่มบุคคล และกลุ่มสถานที่ได้เป็นอย่างดี Alchemy API ผลลัพท์ดีที่สุดในกลุ่มองค์กร

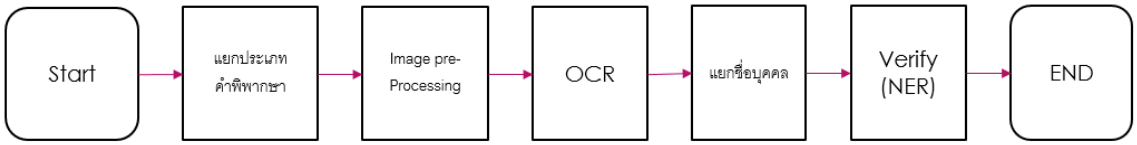
5) Extracting person names from diverse and noisy OCR text [9]

งานวิจัยนี้นำ named entity recognition มาใช้กับเอกสารสแกนที่เป็นเอกสารประวัติศาตร์เพื่อใช้ประกอบการค้นพบรายละเอียดทางประวัติศาสตร์ อย่างไรก็ตามปัญหาของชุดนี้คือผิดพลาดจากการอ่านคำยากที่จะคาดคะเนได้ ผู้วิจัยจึงแก้ปัญหาโดยใช้อัลกอริทึมในการแยกแยะ 3 รูปแบบแยกแยะชื่อบุคคลออกจากเอกสารและใช้ผลลัพท์ที่ผลโหวดสูงสุดซึ่งผลลัพท์ก็ช่วยเพิ่มประสิทธิภาพในการแยกแยะ

บทที่ 3
ระเบียบวิธีวิจัย

3.1 System Flow

ขั้นตอนการประมวลผลข้อความและดำเนินการสร้างระบบสำหรับตรวจสอบข้อความบนเอกสาร ได้ตามวัตถุประสงค์ของงานจึงได้กำหนดกระบวนการการดำเนินการเพื่อศึกษาข้อมูลมีขั้นตอนดังนี้



ภาพที่ 3.1 กระบวนการดำเนินการ

3.2 การสำรวจข้อมูล

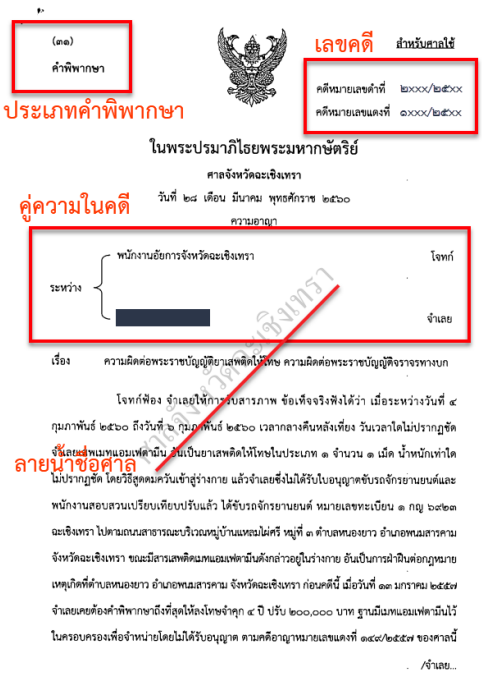
ศึกษาลักษณะของข้อมูลที่นำมาใช้โดยข้อมูลดังกล่าว คือคำพิพากษาที่ได้มีการจัดทำแล้วถูกสแกนผ่านเครื่องสแกนเนอร์ อัพโหลดเข้ามาจัดเก็บในระบบที่มีการจัดทำขึ้นเพื่อรวบรวมคำพิพากษาในระบบ ในรูปแบบไฟล์ PDF ที่มีลักษณะเป็นรูปภาพ (Image file) ดังนั้นกระบวนประมวลผลจากรูปภาพ(Image processing) การนำข้อมูลรูปภาพมาประมวลผลเพื่อแปลงข้อมูลให้ระบบคอมพิวเตอร์แยกแยะได้

จากการสำรวจข้อมูลไฟล์คำพิพากษาจึงสรุปได้ว่าไฟล์คำพิพากษาที่ถูกจัดเก็บโดยการสแกนจากต้นทางมีลักษณะดังนี้

3.2.1 เนื่องจากเป็นเอกสารที่สแกนเข้ามาจึงมีขนาดสัดส่วนรูปใกล้เคียงกัน

3.2.2 มีลายน้ำชื่อศาลวางเป็นแนวทแยงอยู่ทุกหน้าของเอกสาร

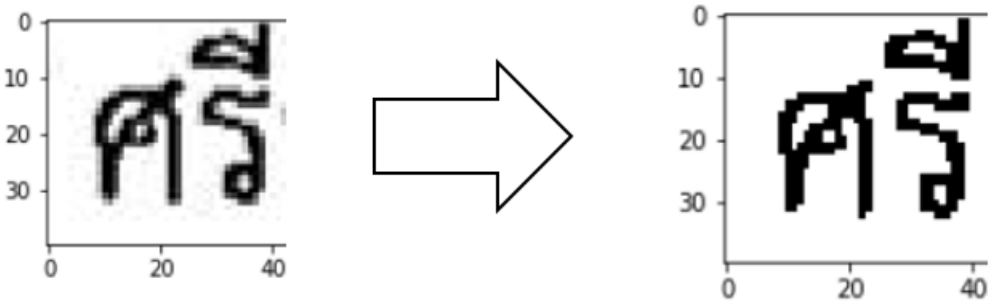
3.2.3 เอกสารมีรูปแบบฟอร์มในการจัดทำ แบ่งประเภทที่พบบ่อยได้แก่ คำพิพากษา คำสั่ง สัญญาประนีประนอม ซึ่งในรูปแบบฟอร์มศาล ส่วนหัวของเอกสารจะมีรูปแบบเฉพาะใกล้เคียงกัน ประกอบด้วยเลขคดี และคู่กรณีในคดีได้แก่ โจทก์ จำเลย หรือผู้ร้อง



ภาพที่ 3.2 ตัวอย่างรูปแบบของคำพิพากษา

3.3 Image Pre-Processing

ในงานวิจัยนี้ใช้เพื่อให้ผลของการ OCR ดีขึ้นจึงใช้มีการทำ Image Processing ก่อนเข้ากระบวนการ OCR โดยใช้การปรับสีภาพเป็นขาวดำ (Black White) ให้ภาพคอมชัดขึ้น



ภาพที่ 3.3 กระบวนการ Image Processing ปรับภาพเป็น Black White

3.4 OCR

ใช้กระบวนการ OCR เพื่อแปลงข้อมูลจากเอกสารสแกนที่มีลักษณะรูปภาพเป็นข้อความ โดยมีขั้นตอนดังนี้

3.4.1 แปลงภาพต้นฉบับ เป็นภาพขาวดำ เพื่อเพิ่มความแม่นยำในการทำ OCR

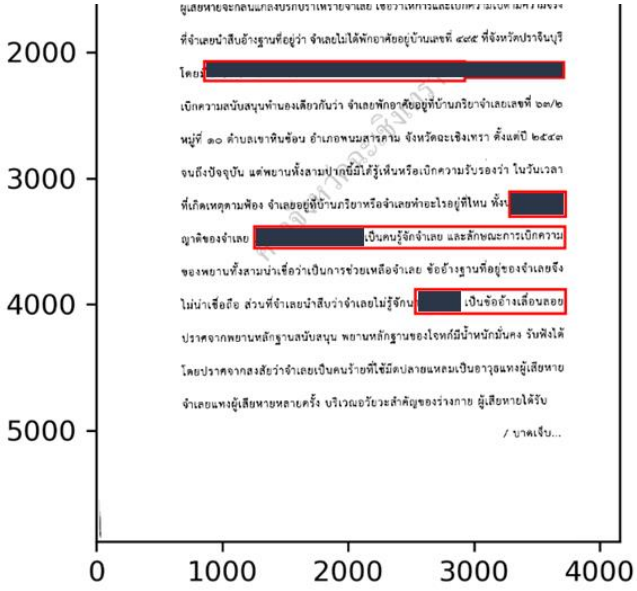
3.4.2 ได้ทดลองใช้งาน LIbrary OCR ได้แก่ tesserectOCR และ easyOCR พบว่า easyOCR นอกจากมีความแม่นยำในการระบุข้อความภาษาไทยได้ถูกต้องกว่าแล้ว สามารถระบุพิกัดตำแหน่งของประโยคในเอกสารได้

ภาพที่ 3.4 ตัวอย่างผลลัพท์ OCR

3.4.3 เมื่อได้ข้อความทั้งหมดบนเอกสารแล้ว จากนั้นคัดข้อความที่มีลักษณะเป็นชื่อเฉพาะโดยในการวิจัยนี้ใช้สมมุติฐาน 2 กรณี คือ

(1) ข้อความประกอบด้วยคำนำหน้าที่ใช้บ่อย ได้แก่ นาย นาง

(2) ในรูปแบบเอกสารของคำพิพากษาหน้าแรกจะระบุชื่อ โจทก์ จำเลย หรือ ผู้ร้อง และส่วนมาจากชื่อบุคคลที่ระบุในส่วนนี้ เนื้อหาจะระบุอ้างถึงด้วยชื่อความเกี่ยวข้องของคดี ตามด้วยลำดับที่ แต่ทั้งนี้กรณีมีบุคคลที่เกี่ยวข้องแต่ไม่ได้อ้างถึงข้างต้นก็จะใช้ชื่อเต็ม



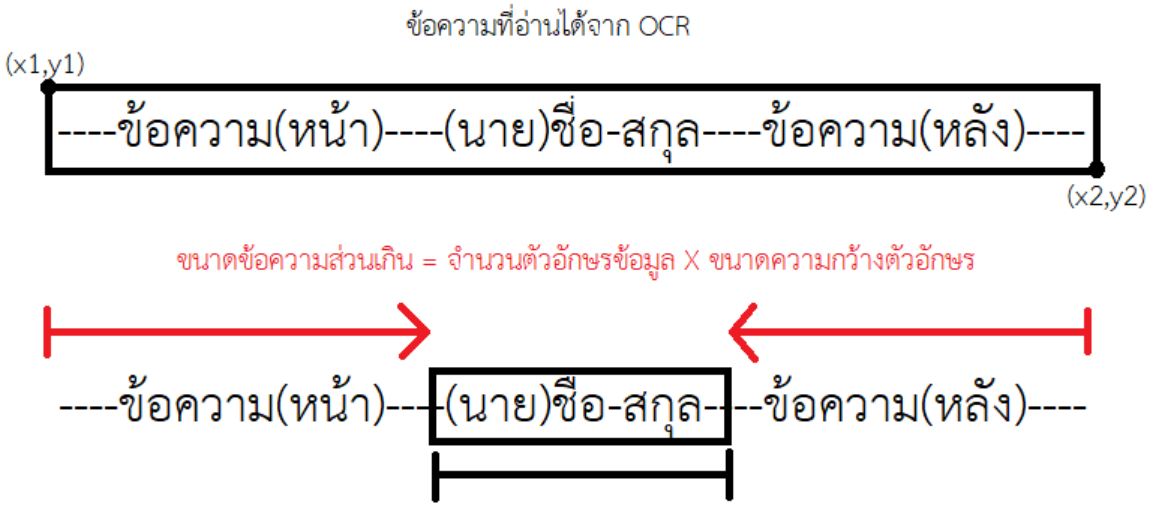
ภาพที่ 3.5 ตัวอย่างข้อความส่วนเกินจาก OCR

3.5 Named Entity Recognition

เพื่อยืนยันข้อมูลที่ได้โดย นำ NER(Named Entity Recognition) จาก Pythainlp มาใช้ตรวจสอบว่าคำดังกล่าวที่คัดมาได้เป็นชื่อบุคคลหรือไม่เป็นข้อมูลเพื่อนำมาเข้ากระบวนจำแนกต่อได้เพื่อใช้ยืนยันผลของข้อมูลที่ดึงแยกออกมา

3.6 การตัดข้อความ

เนื่องจากได้ตำแหน่งคำที่ต้องการ เซ็นเซอร์แต่ในการ OCR ได้ตำแหน่งของข้อความทั้งหมดในบรรทัด ซึ่งมีลักษณะเป็นข้อความยาวมีคำที่ไม่เกี่ยวข้องรวมอยู่ การ OCR ยังให้ข้อมูลพิกัดของกรอบข้อความดังกล่าวด้วย จึงต้องมีกระบวนการเพื่อตัดข้อความที่ไม่จำเป็นออกและคำนวณความกว้างของข้อความใหม่สำหรับนำไปใช้ต่อไป



ภาพที่ 3.6 การตัดข้อความส่วนเกิน

algorithm ขั้นตอนการตัดคำที่เกินมาจากข้อความที่อ่านได้จาก OCR

1. นำข้อความที่ได้จากการ OCR มาแบ่ง ใช้วิธีแบ่งได้ 2 กรณีดังนี้

    1) ข้อความด้านหน้า แบ่งข้อความออกด้วยคำนำหน้า (นาย,นาง)

    2) ข้อความด้านหลัง แบ่งข้อความด้าน NER แยกข้อความเฉพาะที่ระบุ Tag เป็น Person ส่วนข้อความอื่นๆ นำมารวมกัน

2. นำข้อความส่วนเกินมาตัด สระบน สระล่าง และวรรณยุกต์ในภาษาไทย ออกด้วยวิธี Replace

3. ใช้คำสั่ง length หาจำนวนตัวอักษรในข้อความ

4. ประมาณค่าความกว้างของข้อความโดยการนำ จำนวนตัวอักษร X ขนาดความกว้างของตัวอักษร

5. นำความกว้างไปคำนวณกับกรอบของข้อความเดิมออกเป็นกรอบใหม่

บทที่ 4
ผลการวิจัย

จากผลการดำเนินกระบวนการกับเอกสารคำพิพากษาจำนวน 30 เรื่อง เป็นหน้าเอกสารจำนวน 76 หน้า โดยผลลัพธ์ที่ได้จากระบบออกมาเป็นไฟล์ข้อมูล Jsonให้ระบบแบ่งผลลัพธ์สรุปได้ดังนี้

4.1 ผล OCR

ผลจากกระบวนการ OCR แล้วนำมาคัดแยกแล้วออกมาเป็นข้อมูลในรูปแบบ JSON ประกอบด้วย พิกัดของข้อความ ความกว้างความสูงของกรอบ และข้อความที่อ่านมา


```
{
    "filename": "encrypt_201700000254228_p_2.jpg",
    "result_ocr": [
        {
            "x": 1699,
            "y": 2157,
            "width": 340,
            "height": 137,
            "text": "[redacted]",
            "confident ": 0.4986003051242635
        },
        {
            "x": 3342,
            "y": 2174,
            "width": 365,
            "height": 144,
            "text": "[redacted]",
            "confident ": 0.597355790671238
        },
```


ภาพที่ 4.1 ตัวอย่างข้อมูล JSON

โดยในการวัดผลจะใช้วิธี Intersection over Union เพื่อวัดผลว่าข้อความที่พบตรงกับผลลัพธ์ที่ต้องการหรือจึงได้ผลลัพธ์เป็นความถูกต้องในการเซ็นเซอร์ชื่อบนเอกสาร **รายละเอียดตามภาคผนวก ก** โดยคิดจากขนาดของข้อความที่พบกับผลลัพธ์ที่ได้ดังนี้

$$\text{จำนวนข้อความที่พบ/ต้นฉบับ} = \frac{\text{จำนวนชื่อที่พบจาก Model}}{\text{จำนวนชื่อทั้งหมดบนเอกสาร}}$$

$$\text{ร้อยละความถูกต้องจาก OCR} = \frac{\text{ขนาดข้อความของชื่อที่ตรวจพบจาก OCR}}{\text{ขนาดข้อความของชื่อบนเอกสาร}}$$

4.2 ผลจาก NER

ผลจากกระบวนการ NER แล้วนำมาคัดแยกแล้วออกมาเป็นข้อมูลในรูปแบบ JSON ประกอบด้วยข้อความ และกลุ่มของประเภทคำที่พบ มีลักษณ์ของไฟล์และผลลัพธ์ดังนี้


```
],
"result_ner": [
    [
        "[redacted]",
        "B-PERSON"
    ],
    [
        "โอนสิทธิในการเข้าทำประโยชน์ในเขตปฏิรูปที่ดินดังกล่าวให้แก่",
        "O"
    ],
    [
```


ภาพที่ 4.2 ตัวอย่างผลจาก NER

ความถูกต้องในการเซ็นเซอร์ชื่อบนเอกสาร โดยคิดจากขนาดของข้อความที่พบกับผลลัพธ์ที่ได้ดังนี้

$$\text{ร้อยละความถูกต้องจาก NER} = \frac{\text{ขนาดข้อความของชื่อที่ตรวจพบจาก NER}}{\text{ขนาดข้อความของชื่อบนเอกสาร}}$$

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากผลการวิจัยการเซ็นเซอร์ชื่อบุคคลจากเอกสารสแกนคำพิพากษาด้วยเทคโนโลยีปัญญาประดิษฐ์ สรุปได้ดังนี้

5.1.1 ระบบใช้เรียกไฟล์คำพิพากษาเพื่อเซ็นเซอร์ชื่อบุคคลที่ปรากฏบนเอกสาร

5.1.2 จำนวนข้อความที่พบคิดเป็น 89 % จากข้อความทั้งหมด

5.1.3 การใช้ เทคนิค OCR ช่วยแยกคำจากเอกสารสแกนคำพิพากษาสามารถแยกชื่อบุคคลได้บางส่วนจากรูปแบบเอกสารหรือคำนำหน้าที่พบบ่อย

5.1.4 การนำ NER ตรวจสอบคำที่แยกมาแล้วเพื่อช่วยยืนยันการ โดยในการวิจัยจากตัวอย่างพบว่าตรวจสอบชื่อบุคคลถูกต้องเพิ่มขึ้น

5.1.5 การใช้ NER ช่วยตรวจสอบประโยคส่วนเกินที่ติดมาจากการขั้นตอนแยกหรือข้อความที่อยู่หลังชื่อได้

5.2 ข้อเสนอแนะ

5.2.1 เนื่องจากคำในภาษาไทยบางคำมีการเขียนใกล้เคียงกัน การค้นหาด้วยคำนำหน้าจึงอาจจะพบคำที่ไม่เกี่ยวข้องได้ เช่น พบคำว่า “ทนาย” แทนที่จะเป็นชื่อบุคคล

5.2.2 กระบวนการ OCR ในการประมวลผลนาน ในขั้นตอนติดตั้ง Library ของ easyOCR ควรเตรียมคอมพิวเตอร์มีประสิทธิภาพเพียงพอ และติดตั้ง Coda ให้ประมวลผลผ่านกราฟฟิกการ์ด ซึ่ง Library แนะนำว่าช่วยให้ประมวลผลได้เร็วขึ้น

บรรณานุกรม

บรรณานุกรม

[1] “tesseract-ocr/tesseract,” 23 พฤษภาคม 2023. [ออนไลน์]. Available: https://github.com/tesseract-ocr/tesseract.

[2] “JaidedAI/EasyOCR,” 23 พฤษภาคม 2023. [ออนไลน์]. Available: https://github.com/JaidedAI/EasyOCR.

[3] “What is natural language processing (NLP)?,” 23 พฤษภาคม 2023. [ออนไลน์]. Available: https://www.ibm.com/topics/natural-language-processing.

[4] “Nick Barney, "named entity recognition,” 23 พฤษภาคม 2023. [ออนไลน์]. Available: https://www.techtarget.com/whatis/definition/named-entity-recognition-NER.

[5] A. Hamdi, A. Jean-Caurant, N. Sidere, M. Coustaty และ A. Doucet, “An analysis of the performance of named entity recognition over ocred documents,” *In Proceedings of the 18th Joint Conference on Digital Libraries (JCDL '19). IEEE Press,* p. 333–334, 2020.

[6] A. Lodh, U. Saxena, A. Markhedkar, A. Khan, B. M. Votavat และ A. Mendon, “Named entity extraction from images using natural language processing pipelines,” *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India,* pp. 1777-1784, 2022.

[7] A. Goyal และ K. , “Named Entity Recognition For Scanned Images,” 2021.

[8] K. J. Rodriquez, M. Bryant, T. Blanke และ M. Luszczynska, “Comparison of Named Entity Recognition tools for raw OCR text,” *Proceedings of KONVENS 2012 (LThist 2012 workshop),* pp. 411-414, 21 กันยายน 2012.

[9] T. L. Packer, J. F. Lutes, A. P. Stewart, D. W. Embley, E. K. Ringger, K. D. Seppi และ L. S. Jensen, “Extracting person names from diverse and noisy OCR text,” *'10: Proceedings of the fourth workshop on Analytics for noisy unstructured text data,* pp. 19-26, ตุลาคม 2010.

ภาคผนวก

ภาคผนวก ก
ผลการเซ็นเซอร์ชื่อในไฟล์เอกสาร

ตารางที่ 1 ผลการเซ็นเซอร์ชื่อในไฟล์เอกสาร

ไฟล์	จำนวนข้อความที่พบ/ต้นฉบับ	ขนาดข้อความต้นฉบับ	ขนาดข้อความที่ได้จากOCR	ขนาดข้อความที่ได้หลัง NER
4228_p_1	2/2 (100%)	16	8 (50%)	8 (50%)
4228_p_1		16	15 (94%)	15 (94%)
4228_p_2	3/3 (100%)	16	8 (50%)	8 (50%)
4228_p_2		16	16 (100%)	16 (100%)
4228_p_2		19	19 (100%)	19 (100%)
4258_p_1	4/4 (100%)	25	25 (100%)	25 (100%)
4258_p_1		17	17 (100%)	17 (100%)
4258_p_1		17	17 (100%)	17 (100%)
4258_p_1		11	66 (17%)	11 (100%)
4258_p_2	4/4 (100%)	24	47 (51%)	26 (92%)
4258_p_2		17	7 (41%)	7 (41%)
4258_p_2		15	16 (93%)	16 (93%)
4258_p_2		18	19 (95%)	19 (95%)
5477_p_1	3/5 (60%)	17	18 (94%)	18 (94%)
5477_p_1		16	23 (70%)	8 (50%)
5477_p_1		7	44 (15%)	7 (100%)
5477_p_2	1/2 (50%)	7	64 (11%)	11 (63%)
5477_p_3	4/5 (80%)	16	33 (48%)	18 (89%)
5477_p_3		7	43 (16%)	14 (50%)
5477_p_3		7	39 (18%)	7 (100%)
5477_p_3		7	17 (41%)	7 (100%)
5477_p_4	1/1 (100%)	16	55 (29%)	18 (89%)
5477_p_5	3/4 (75%)	16	54 (30%)	23 (70%)
5477_p_5		7	35 (20%)	7 (100%)
5477_p_5		7	41 (17%)	7 (100%)
5477_p_6	0/0 (100%)			
5477_p_7	2/2 (100%)	17	19 (89%)	19 (89%)
5477_p_7		15	15 (100%)	15 (100%)
5496_p_1	5/5 (100%)	16	19 (84%)	19 (84%)
5496_p_1		16	30 (53%)	16 (100%)

5496_p_1		16	15 (94%)	8 (50%)
5496_p_1		17	19 (89%)	19 (89%)
5496_p_1		15	15 (100%)	15 (100%)
5524_p_1	6/6 (100%)	15	15 (100%)	15 (100%)
5524_p_1		16	17 (94%)	17 (94%)
5524_p_1		13	13 (100%)	13 (100%)
5524_p_1		16	21 (76%)	16 (100%)
5524_p_1		14	15 (93%)	15 (93%)
5524_p_1		15	7 (47%)	7 (47%)
5539_p_1	3/3 (100%)	13	15 (87%)	15 (87%)
5539_p_1		20	13 (65%)	13 (65%)
5539_p_1		20	49 (41%)	22 (91%)
5539_p_2	3/4 (75%)	13	51 (25%)	27 (48%)
5539_p_2		23	22 (96%)	22 (96%)
5539_p_2		15	15 (100%)	15 (100%)
5600_p_1	7/7 (100%)	15	21 (71%)	15 (100%)
5600_p_1		15	36 (42%)	14 (93%)
5600_p_1		15	68 (22%)	15 (100%)
5600_p_1		15	38 (39%)	14 (93%)
5600_p_1		15	15 (100%)	15 (100%)
5600_p_1		18	18 (100%)	18 (100%)
5600_p_1		21	21 (100%)	21 (100%)
5617_p_1	2/1 (50%)	18	17 (94%)	17 (94%)
5617_p_1		0	7	7
5617_p_2	1/1 (100%)	15	20 (75%)	15 (100%)
5617_p_3	1/1 (100%)	18	20 (90%)	20 (90%)
5767_p_1	2/2 (100%)	13	14 (93%)	14 (93%)
5767_p_1		18	18 (100%)	18 (100%)
5783_p_1	2/2 (100%)	16	10 (63%)	10 (63%)
5783_p_1		13	7 (54%)	7 (54%)
5783_p_2	1/2 (50%)	18	17 (94%)	17 (94%)
5783_p_3	2/1 (50%)	16	17 (94%)	15 (94%)
5783_p_3		13	18 (72%)	7 (54%)
5783_p_4	1/2 (50%)	17	15 (88%)	14 (82%)

5783_p_5	3/4 (75%)	16	15 (94%)	15 (94%)
5783_p_5		13	14 (93%)	14 (93%)
5783_p_5		18	18 (100%)	15 (83%)
5841_p_1	2/2 (100%)	15	18 (83%)	16 (94%)
5841_p_1		34	37 (92%)	33 (37%)
5841_p_2	0/0 (100%)			
5841_p_3	0/0 (100%)			
5841_p_4	0/0 (100%)			
5841_p_5	1/1 (100%)	17	17 (100%)	17 (100%)
5845_p_1	3/3 (100%)	15	15 (100%)	15 (100%)
5845_p_1		20	20 (100%)	20 (100%)
5845_p_1		14	4 (29%)	4 (29%)
5845_p_2	1/1 (100%)	14	35 (40%)	14 (100%)
5845_p_3	0/0 (100%)			
5845_p_4	2/1 (50%)			
5845_p_4		18	18 (100%)	18 (100%)
5851_p_1	2/2 (100%)	19	21 (90%)	18 (95%)
5851_p_1	2/2	18	23 (78%)	18 (100%)
5851_p_2	0/0 (100%)			
5851_p_3	0/0 (100%)			
5851_p_4	1/1 (100%)	17	17 (100%)	17 (100%)
5885_p_1	1/2 (50%)	38	18 (47%)	14 (37%)
5885_p_2	0/0 (100%)			
5885_p_3	1/1 (100%)	17	17 (100%)	17 (100%)
5912_p_1	3/3 (100%)	13	13 (100%)	13 (100%)
5912_p_1		20	23 (87%)	20 (100%)
5912_p_1		15	5 (33%)	5 (33%)
5912_p_2	1/1 (100%)	19	19 (100%)	19 (100%)
5927_p_1	1/1 (100%)	17	9 (53%)	9 (53%)
5927_p_2	0/0 (100%)			
5927_p_3	0/0 (100%)			
5927_p_4	1/1 (100%)	17	17 (100%)	17 (100%)
5931_p_1	2/2 (100%)	25	16 (64%)	16 (64%)
5931_p_1		26	16 (62%)	16 (62%)

5931_p_2	1/1 (100%)	23	22 (97%)	22 (97%)
5942_p_1	1/2 (50%)	18	19 (95%)	19 (95%)
5942_p_2	1/1 (100%)	23	23 (100%)	23 (100%)
5943_p_1	1/1 (100%)	16	16 (100%)	16 (100%)
5943_p_2	1/1 (100%)	23	22 (96%)	22 (96%)
5944_p_1	0/0 (100%)			
5944_p_2	0/0 (100%)			
5944_p_3	1/1 (100%)	18	20 (90%)	20 (90%)
5945_p_1	0/0 (100%)			
5945_p_2	1/1 (100%)	14	64 (22%)	14 (100%)
5945_p_3	1/1 (100%)	12	47 (26%)	12 (100%)
5945_p_4	1/1 (100%)	18	18 (100%)	18 (100%)
5946_p_1	1/1 (100%)	12	10 (83%)	10 (83%)
5946_p_2	1/1 (100%)	18	18 (100%)	18 (100%)
5947_p_1	5/3 (60%)	17	17 (100%)	17 (100%)
5947_p_1		22	17 (77%)	10 (45%)
5947_p_1		0	6 (0%)	6 (0%)
5947_p_1		18	18 (100%)	18 (100%)
5947_p_1		0	5 (0%)	5 (0%)
5948_p_1	3/3 (100%)	17	17 (100%)	17 (100%)
5948_p_1		16	16 (100%)	16 (100%)
5948_p_1		17	17 (100%)	17 (100%)
5949_p_1	0/1 (0%)			
5949_p_2	1/1 (100%)	18	18 (100%)	18 (100%)
5950_p_1	1/1 (100%)	21	12 (57%)	12 (57%)
5950_p_2	1/1 (100%)	18	18 (100%)	18 (100%)
5951_p_1	1/1 (100%)	15	16 (94%)	16 (94%)
5951_p_2	1/1 (100%)	18	18 (100%)	18 (100%)
5952_p_1	1/1 (100%)	29	11 (38%)	11 (38%)
5952_p_2	1/1 (100%)	18	18 (100%)	18 (100%)
5953_p_1	1/1 (100%)	14	14 (100%)	14 (100%)
5953_p_2	1/1 (100%)	18	18 (100%)	18 (100%)
5954_p_1	0/1 (0%)		0	0
5954_p_2	1/1 (100%)	18	18 (100%)	18 (100%)

ประวัติผู้เขียน

ชื่อ-นามสกุล วิศรุต เหล่าดารา

ประวัติการศึกษา

พ.ศ.2552 ปริญญาวิทยาศาสตร์บัณฑิต (วิทยาการคอมพิวเตอร์)
มหาวิทยาลัยเกษตรศาสตร์

ประวัติการทำงาน

พ.ศ.2554 สำนักเทคโนโลยีสารสนเทศ สำนักงานศาลยุติธรรม.