

Project No. 67

ระบบจัดเก็บและจัดการเอกสารภายในห้องบรรณสารสนเทศ

Mr.Akarapon Boonsermsakul

Ms.Thanaporn Pitianusorn

Mr.Annop Kongsombatcharoen

A Project Submitted in Partial Fulfillment

of the Requirements for

the Degree of Bachelor of Engineering (Computer Engineering)

Faculty of Engineering

King Mongkut's University of Technology Thonburi

2020

Project Committee

.....

Project Advisor

(Asst.Prof. Suthathip Manee, Ph.D.)

.....

Committee Member

(Dr.Prapong Prechaprapranyawong, Ph.D.)

.....

Committee Member

(Asst.Prof.Sanan Srakaew)

.....

Committee Member

(Asst.Prof.Surapont Toomnark)

Project Title	Project No. 67 ระบบจัดเก็บและจัดการเอกสารภายในห้องสมุดสารสนเทศ
Credits	3
Member(s)	Mr.Akarapon Boonsermsakul Ms.Thanaporn Pitianusorn Mr.Annop Kongsombatcharoen
Project Advisor	Asst.Prof. Suthathip Manee, Ph.D.
Program	Bachelor of Engineering
Field of Study	Computer Engineering
Department	Computer Engineering
Faculty	Engineering
Academic Year	2020

### **Abstract**

KMUTT's library have collected the archive of valued documents. Because these document have not transformed into digital form, there is vital problem in searching for information in these document for librarian and patrons. In this project, we developed web platform to digitize these document into digital format and implement the search function that facilitate the librarian and patron to search for information. The platform consists of 2 components. The first part is importing documents and digitization. In this step, we applied image processing techniques such as Morphology Transformation to preprocess the images of documents and transform the images to full text data by using Tesseract. After getting the text files, we tokenize the text into words by using the Deepcut library and find the significant words of the document by using the TF-IDF algorithm. In the second part, we start by getting the input from the user and use the word2Vec model to find a similar word. And take input and similar words to get the TF-IDF score that we generate at first to find the best document for the input word.

**Keywords:** Natural language processing / RESTful Service / Optical character recognition / Image Processing / Information retrieval / Term Frequency-Inverse Document Frequency / Word2Vec / Word Embedded

หัวข้อปริญญาอินพนธ์	ระบบจัดเก็บและจัดการเอกสารภายในห้องบรรณสารสนเทศ KMUTT Archives Management Platform
หน่วยกิต	3
ผู้เขียน	นายอัครพล บุญเสริมศักดิ์กุล นางสาวอนพร ปิติอนุสรณ์ นายอรรถพ กองสมบัติเจริญ
อาจารย์ที่ปรึกษา	ผศ.ดร.สุราทิพย์ มณีวงศ์วัฒนา
หลักสูตร	วิศวกรรมศาสตรบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
คณะ	วิศวกรรมศาสตร์
ปีการศึกษา	2563

### บทคัดย่อ

การจะสืบค้นข้อมูลจากเอกสารหรือขั้นหนังสือที่มีการรวบรวมข้อมูลไว้ตั้งแต่อดีตจนเป็นปัจจุบันอย่างหนึ่งของเจ้าหน้าที่บรรณาธิการ ที่ต้องทำการค้นหาเอกสารและหนังสือแล้วนั้น เนื่องจาก การที่ยังไม่มีการเก็บหนังสือและเอกสารให้อยู่ในรูปแบบของข้อมูลดิจิตอลทำให้ต้อง สืบค้น โดยการค้นหาเอกสารและหนังสือแต่ละเล่มโดยการดูจากเนื้อหาสารบัญเพื่อให้ได้หนังสือที่ตรงกับข้อมูลที่ต้องการมากที่สุด ซึ่งการที่ค้นหาจากหน้าสารบัญของ หนังสือแต่ละเล่ม ก็จะทำให้การค้นหาเป็นไปอย่างล่าช้า และบางครั้งการคูเพียงแค่สารบัญก็อาจจะทำให้ได้หนังสือที่ไม่ตรงกับความต้องการของผู้ที่เข้ามายืนหนังสือ ในโครงการนี้เราได้ทำการพัฒนาการระบบจัดเก็บและค้นหาเอกสารอิเล็กทรอนิกส์ โดยแบ่งออกเป็น 2 ขั้นตอนคือ การนำเข้าข้อมูล และการสร้างระบบค้นหา โดยขั้นตอนการนำเข้าข้อมูล เราจะเริ่มจากการ ทำ image processing เพื่อเตรียมข้อมูลรูปภาพที่ได้มา ก่อนจะนำไปผ่านกระบวนการ OCR เพื่อแปลงรูปภาพเหล่านี้ให้อยู่ในรูปของข้อมูลดิจิตอล โดยการเก็บข้อมูลในรูปแบบของ Information Retrieval เพื่อช่วยให้ความเร็วการค้นหาใหม่ประสิทธิภาพมากยิ่งขึ้น และนำข้อมูลนั้นมาทำการตัดคำ และเช็คคำพิเศษ จากนั้นจะนำมาคำคำคัญของหนังสือหรือเอกสารนั้น ๆ โดยการใช้การหาค่าคะแนนแบบ TF-IDF ส่วนการสร้างระบบการค้นหาจะเริ่มจากรับคำค้นหาจากผู้ใช้และทำการคำนวณที่ได้ไปข้างมานี้แล้วword2Vec เพื่อหาคำที่ใกล้เคียง จำนวนคำใกล้เคียงและคำค้นหาไปดึงคะแนน TF-IDF ที่เก็บไว้เพื่อค้นหาว่า มีเอกสารหรือหนังสือเล่มไหนที่มีคะแนนที่ตรงและใกล้เคียงกับคำค้นหามากที่สุด

**คำสำคัญ:** Natural language processing / RESTful Service / Optical character recognition / Image Processing / Information retrieval / Term Frequency-Inverse Document Frequency / Word2Vec / Word Embedded

## กิตติกรรมประกาศ

ขอขอบคุณนางสาวอาจารยา ศรีบัวบาน เจ้าหน้าที่หอบรรณสารสนเทศและ พศ.ดร.สุราทิพย์ มณีวงศ์วัฒนา อาจารย์ที่ปรึกษาร่วมทั้งเจ้าหน้าที่ภายในหอสมุดมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรีที่เสียเวลาให้ความรู้ความเข้าใจ ทั้งในเรื่องการเก็บข้อมูลและคอยแนะนำวิธีการจัดการกับปัญหาต่างๆที่เกิดขึ้น นำมาสู่การทำทั่วช้อปวิถุญาณพินธ์ฉบับนี้ให้สำเร็จตามที่ต้องการ

<b>สารบัญ</b>	
<b>หน้า</b>	
<b>ABSTRACT</b>	ii
<b>บทคัดย่อ</b>	iii
<b>กิตติกรรมประกาศ</b>	iv
<b>สารบัญ</b>	viii
<b>สารบัญตาราง</b>	ix
<b>สารบัญรูปภาพ</b>	x
<b>สารบัญสัญลักษณ์</b>	xii
<b>สารบัญคำศัพท์ทางเทคนิคและคำย่อ</b>	xiii
<b>บทที่ 1 บทนำ</b>	1
1.1      คำสำคัญ	1
1.2      ความสำคัญของปัญหา	1
1.3      ประเภทของโครงงาน	1
1.4      วิธีการที่นำเสนอ	1
1.5      วัตถุประสงค์	2
1.6      ขอบเขตของงานวิจัย	2
1.7      เนื้อหาทางวิชวกรรมที่เป็นต้นฉบับ	2
1.8      การแยกย่อยงาน และร่างแผนการดำเนินงาน	3
1.9      ตารางการดำเนินงาน	4
1.9.1      ผลการดำเนินงานในภาคการศึกษาที่ 1	5
1.9.2      ผลการดำเนินงานในภาคการศึกษาที่ 2	5
<b>บทที่ 2 ที่มา ทฤษฎีและงานวิจัยที่เกี่ยวข้อง</b>	6
2.1      บทนำ	6
2.2      แนวความคิดทางทฤษฎี	6
2.2.1      Image Processing	6

2.2.1.1	Contour	6
2.2.1.2	Morphology Transformation	7
2.2.2	Optical character recognition (OCR)	7
2.2.3	Natural language processing	8
2.2.3.1	Information retrieval	8
2.2.3.2	TF-IDF	9
2.2.3.3	Cosine Similarity	9
2.2.3.4	Minimum Edit Distance	10
2.2.4	RESTful Service	11
2.2.5	Word Embedding	12
2.3	ภาษาคอมพิวเตอร์และเทคโนโลยี	12
2.3.1	Open source Computer Vision (OpenCV)	12
2.3.2	Tesseract OCR	12
2.3.3	DeepCut	12
2.3.4	ReactJS	13
2.3.5	Python	13
2.3.5.1	Django	13
2.3.6	NodeJS	13
<b>บทที่ 3</b>	<b>การออกแบบและระบบวิธีจัดการข้อมูล</b>	<b>14</b>
3.1	System Overview	14
3.2	Feature lists	14
3.2.1	การแปลงเอกสารเป็นรูปภาพ	14
3.2.2	Image preparation	14
3.2.3	Image to word	15
3.2.4	Text preprocessing	15
3.2.5	Tag generated	15

3.2.6	Search	15
3.2.7	Manage Book	15
3.2.8	Login	16
3.3	System requirements	16
3.4	Database Design	18
3.4.1	Database Structure	21
3.4.2	Database Dictionary	22
3.5	UML Design	28
3.5.1	Use case diagram	28
3.5.2	Sequence diagram	28
3.5.2.1	Use case Add Document	28
3.5.2.2	Use case Manage word in document	29
3.5.2.3	Use case Verify Document to Generate Keyword	30
3.5.2.4	Use case Edit Document	31
3.5.2.5	Use case Delete Document	32
3.5.2.6	Use case View Document & Search Document	33
3.5.2.7	Use case Login	34
3.6	GUI Design	36
3.6.1	Homepage	36
3.6.2	Homepage2	37
3.6.3	Login	38
3.6.4	Insert Book(1)	39
3.6.5	Insert Book (2)	40
3.6.6	Insert Book (3)	41
3.6.7	Insert Book (4)	42
3.6.8	Insert Book (5)	43
3.6.9	Insert Book (6)	44

3.6.10	Search	45
3.6.11	Document View	46
3.6.12	Manage book	47
3.6.13	Edit Book	48
3.6.14	Upload Status Page	51
3.6.15	Evaluate Process Design	51
	บรรณานุกรม	54

## สารบัญตาราง

ตารางที่	หน้า
1.1 ตารางการดำเนินงาน ภาคการศึกษาที่ 1/2563	4
1.2 ตารางการดำเนินงาน ภาคการศึกษาที่ 2/2563	4
2.1 Information retrieval ในลักษณะ Boolean Retrieval	8
3.1 ตารางอธิบายความหมายตาราง term_word	22
3.2 ตารางอธิบายความหมายตาราง user	23
3.3 ตารางอธิบายความหมายตาราง score	23
3.4 ตารางอธิบายความหมายตาราง pre_term_in_page	23
3.5 ตารางอธิบายความหมายตาราง page_in_document	24
3.6 ตารางอธิบายความหมายตาราง nodejs_log	24
3.7 ตารางอธิบายความหมายตาราง knex_migrations_lock	24
3.8 ตารางอธิบายความหมายตาราง knex_migrations	24
3.9 ตารางอธิบายความหมายตาราง indexing_publisher_document	25
3.10 ตารางอธิบายความหมายตาราง indexing_issued_date_document	25
3.11 ตารางอธิบายความหมายตาราง indexing_creator_orgname_document	25
3.12 ตารางอธิบายความหมายตาราง indexing_creator_document	25
3.13 ตารางอธิบายความหมายตาราง indexing_contributor_document	25
3.14 ตารางอธิบายความหมายตาราง document	26
3.16 ตารางอธิบายความหมายตาราง django_log	27
3.17 ตารางอธิบายความหมายตาราง dc_type	27
3.18 ตารางอธิบายความหมายตาราง dc_relation	27
3.19 ตารางอธิบายความหมายตาราง dc_keyword	27
3.20 ตารางประเมินการทำ OCR	51
3.21 ตารางประเมินระบบการค้นหา	52
3.22 ตารางประเมิน Design	52
3.23 ตารางประเมิน test	53

## สารบัญรูปภาพ

รูปที่	หน้า
2.1 แสดงการหาค่าโครงร่างภายในรูป	6
2.2 แสดงการทำ dilation เพื่อเพิ่มพื้นที่สีขาว	7
2.3 แสดงการทำ erosion เพื่อกร่อนพื้นที่สีขาว	7
2.4 Information retrieval ในลักษณะ Index Retrieval	9
2.5 หลักการการเข้า edit distance [10]	10
2.6 ตัวอย่างตารางการทำ minimum edit distance [10]	10
2.7 แสดงถึงโครงสร้างของ HTTP Request [12]	11
2.8 แสดงถึงโครงสร้างของ HTTP Response [12]	12
3.1 System Overview	14
3.2 แสดง ER Diagram ของฐานข้อมูล	18
3.3 แสดง ER Diagram ส่วนของการเก็บความผิดพลาดในการสร้างคีย์เวิร์ดจากเอกสาร	18
3.4 แสดง ER Diagram ส่วนของคีย์เวิร์ดและคะแนนความสำคัญในระบบ	18
3.5 แสดง ER Diagram ส่วนของการเก็บคำจากแต่ละหน้าที่แปลงมาจากการ	19
3.6 แสดง ER Diagram ส่วนของประวัติของผู้ใช้งานมีการสร้างหรือแก้ไขเอกสาร	19
3.7 แสดง ER Diagram ส่วนของการเก็บข้อมูล keyword, relation, type ของเอกสาร	19
3.8 แสดง ER Diagram ส่วนของ Creator มีความเกี่ยวข้องกับเอกสารใหม่บ้าง	19
3.9 แสดง ER Diagram ส่วนของ Creator Organized Name มีความเกี่ยวข้องกับเอกสารใหม่บ้าง	20
3.10 แสดง ER Diagram ส่วนของ Publisher มีความเกี่ยวข้องกับเอกสารใหม่บ้าง	20
3.11 แสดง ER Diagram ส่วนของ Contributor มีความเกี่ยวข้องกับเอกสารใหม่บ้าง	20
3.12 แสดง ER Diagram ส่วนของ Issued Date มีความเกี่ยวข้องกับเอกสารใหม่บ้าง	20
3.13 แสดง ER Diagram ส่วนของ Knex module ที่ใช้สำหรับ Migration ฐานข้อมูล	21
3.14 แสดง ER Diagram ส่วนของการเก็บประวัติการ HTTP Request NodeJS ไปยัง Django	21
3.15 Use case diagram	28
3.16 แสดง Scenario 1 เพิ่มเอกสารเข้าระบบ	29

3.17 แสดง Scenario 2 การจัดการคำที่ถูกเก็บได้จากเอกสารในระบบ	30
3.18 แสดง Scenario 3 ยืนเอกสารว่าพร้อมสำหรับการถูกนำไปสร้างคีย์เวิร์ด	31
3.19 แสดง Scenario 4 แก้ไขข้อมูลเอกสาร	32
3.20 แสดง Scenario 5 ลบเอกสาร	33
3.21 แสดง Scenario 6 ดูข้อมูลเอกสาร และการค้นหาเอกสาร	34
3.22 แสดง Scenario 7 ระบบล็อกอิน	35
3.23 ภาพแสดงหน้าหลักของเว็บไซต์	36
3.24 ภาพแสดงหน้าหลักของเว็บไซต์หลังจากการกดเปิดเมนู	37
3.25 ภาพแสดงหน้าเข้าสู่ระบบ	38
3.26 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าสู่ระบบขั้นเลือกไฟล์	39
3.27 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าสู่ระบบขั้นกรอกข้อมูลขั้นที่ 1	40
3.28 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าสู่ระบบขั้นกรอกข้อมูลขั้นที่ 2	41
3.29 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าขั้นโหลดข้อมูลเข้าสู่ระบบ	42
3.30 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าสู่ระบบขั้นแก้ไขคำผิด	43
3.31 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าสู่ระบบขั้นแก้ไขและเพิ่มคำสำคัญ	44
3.32 ภาพแสดงหน้าค้นหาข้อมูล	45
3.33 ภาพแสดงหน้าดูหนังสือ	46
3.34 ภาพแสดงหน้าการจัดการหนังสือที่เพิ่มเข้าสู่ระบบ	47
3.35 ภาพแสดงขั้นตอนการแก้ไขหนังสือขั้นที่ 1	48
3.36 ภาพแสดงขั้นตอนการแก้ไขหนังสือขั้นที่ 2	49
3.37 ภาพแสดงขั้นตอนการแก้ไขหนังสือขั้นที่ 3	50
3.38 ภาพแสดงหน้าการโหลดข้อมูล	51

## สารบัญสัญลักษณ์

SYMBOL		UNIT
$\alpha$	Test variable	$m^2$
$\lambda$	Interarrival rate	jobs/ second
$\mu$	Service rate	jobs/ second

## สารบัญคำศัพท์ทางเทคนิคและคำย่อ

ABC	=	Adaptive Bandwidth Control
MANET	=	Mobile Ad Hoc Network

# บทที่ 1 บทนำ

## 1.1 คำสำคัญ

Natural language processing, RESTful Service, Optical character recognition, Image Processing, Information retrieval, Term Frequency-Inverse Document Frequency, Word2Vec, Word Embedded

## 1.2 ความสำคัญของปัญหา

นับตั้งแต่การก่อตั้งหอสมุดมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรีได้มีการเก็บรวบรวมองค์ความรู้จากประสบการณ์การทำงานของคณะอาจารย์ผู้เชี่ยวชาญในทางด้านศาสตร์ต่าง ๆ ในรูปแบบลายมือและสื่อสิ่งพิมพ์ไม่ว่าจะเป็น หนังสือ เอกสาร รวมถึงบันทึกเหตุการณ์ในอดีต ในรูปของจดหมายเหตุเพื่อส่งต่อประวัติศาสตร์ความรู้ไปยังคนรุ่นหลังโดยมีการจัดเก็บอยู่ภายใต้หมวดหมู่ที่มีเจ้าหน้าที่บรรณารักษ์เป็นผู้ดูแล และเนื่องจากการที่ เอกสาร หนังสือยังไม่ได้มีการจัดเก็บในรูปแบบดิจิตอลทำให้เมื่อมีบุคคลภายนอกที่ต้องการข้อมูลเพื่อนำไปทำกิจกรรมต่าง ๆ ไม่ว่าจะเป็นการทำวิจัย รายงาน หรือหาข้อมูลเพื่อประกอบการประชุมก็ตามแต่ ก็จำเป็นที่จะต้องมาติดต่อเจ้าหน้าที่บรรณารักษ์ผู้ดูแลเพื่อที่จะให้เจ้าหน้าที่บรรณารักษ์ทำการค้นหาหนังสือที่มีเนื้อหาตามที่เราต้องการ ซึ่งการค้นหาข้อมูลที่ต้องการนั้นเจ้าหน้าที่จะต้องทำการค้นหาด้วยระบบมือทำให้การค้นหาข้อมูลดำเนินการไปอย่างล่าช้า นอกจากนั้นวิธีการหาข้อมูลของเจ้าหน้าที่บรรณารักษ์จะเลือกตรวจสอบข้อมูลของหนังสือจากการดูสารบัญทำให้ข้อมูลที่ได้รับมาอาจจะแตกคลื่นจากข้อมูลเดิมอื่นได้

เพื่ออำนวยความสะดวกให้กับบรรณารักษ์ในการสืบค้นข้อมูลและทำให้การบริการในการสืบค้นเอกสารต่าง ๆ และให้บุคคลภายนอกสามารถทำการค้นหาข้อมูลได้ด้วยตนเองครบถ้วนทางคณิตศาสตร์จัดทำโครงการจึงได้พัฒนาระบบการจัดเก็บเอกสารและระบบการค้นหาโดยการใช้เครื่องมือในการทำ OCR เพื่อแปลงเอกสารให้อยู่ในรูปแบบของเอกสาร digital และหาคำสำคัญในการสร้าง tag ด้วยวิธี Term Frequency - Inverse Document Frequency เพื่อเพิ่มประสิทธิภาพให้กับการค้นหา

## 1.3 ประเภทของโครงงาน

นำเสนอความต้องการของผู้มีส่วนได้ส่วนเสียเฉพาะกลุ่ม

## 1.4 วิธีการที่นำเสนอ

ระบบการค้นหาเอกสาร มีขั้นตอนการทำงานดังนี้

1. นำเอกสารมาแปลงเป็นรูปภาพในรูปแบบสแกน
2. นำรูปภาพเข้าสู่ระบบโดยใช้การรับส่งข้อมูลแบบ RESTful API ในระบุประเภทของการใช้งาน
3. นำรูปภาพผ่านกระบวนการ Image Processing โดยใช้ OpenCV ในการลบส่วนอื่น ๆ ที่ไม่ใช่ข้อความออกและตัดเฉพาะข้อความเพื่อนำไปใช้ในขั้นตอนต่อไป
4. นำรูปที่ผ่านการทำ Image Processing มาเข้าสู่ระบบ OCR เพื่อแปลงข้อมูลจากรูปภาพมาเป็นข้อความในระบบดิจิตอล
5. นำข้อมูลที่เก็บไวามาทำการตัดแบ่งคำภาษาไทยและแก้คำผิด
6. ค้นหาคำสำคัญโดยใช้วิธี TF-IDF เพื่อนำมาใช้ในการสร้าง Tag

7. นำข้อมูลที่ถูกแปลงเก็บและข้อมูลเกี่ยวกับ Tag ลงในดาต้าเบส
8. ทำระบบค้นหาในรูปแบบ Cosine Similarity
9. ทำระบบหาคำใกล้เคียงโดยใช้วิธี Word2Vec
10. ทำแพลตฟอร์มเว็บไซต์เพื่อเป็น User Interface ให้กับผู้ใช้งานได้ใช้งานสำหรับการใช้งานในการค้นหาข้อมูลและเพิ่มข้อมูลหนังสือลงในฐานข้อมูลเพิ่ม

## 1.5 วัตถุประสงค์

1. สร้างระบบแปลงข้อมูลเอกสารให้อยู่ในรูปแบบดิจิตอล
2. สร้าง web platform เพื่อทำการค้นหาเอกสารจากคำค้น และพัฒนาเครื่องมือสนับสนุนการทำงานของบรรณาธิการประจำห้องบรรณสารสนเทศ
3. สร้างระบบการค้นหาโดยการใช้วิธีการ อินฟอร์เมชันรีทรีฟวอล ซึ่งวัดความใกล้เคียงกันระหว่างคำค้นและข้อมูลในฐานข้อมูลโดยวิธี cosine ซึ่งมีผลิติ
4. เพิ่มประสิทธิภาพในการเข้าถึงข้อมูลในรูปแบบดิจิตอล
5. เรียนรู้เรื่องการทำ Image processing

## 1.6 ขอบเขตของงานวิจัย

1. ระบบแปลงข้อมูลจากเอกสารและหนังสือเก่า รองรับเฉพาะเอกสารที่เป็นตัวอักษรแบบพิมพ์ และรองรับไฟล์เอกสารเฉพาะ PDF เท่านั้น
2. ระบบตัดคำ Stop word ภาษาไทยโดยอ้างอิงมาจาก pythainlp และภาษาอังกฤษจาก nltk
3. ทำระบบค้นหาแบบ Cosine Similarity ในระบบ Information retrieval
4. ข้อมูลหนังสือที่นำมาใช้คือหนังสือจำพวก งานแสดงกตเวทิตาจิต เอกสารรายงานประจำปี ตั้งแต่ปี พ.ศ. 2527 ถึง 2560 รวมประมาณ 44 เล่ม จากหอดหมายเหตุมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
5. ทำ platform เว็บไซต์ในรูปแบบ responsive แต่ไม่รองรับขนาดมือถือ รองรับเฉพาะคอมพิวเตอร์หรือโน้ตบุ๊ค
6. การแปลงสิ่งพิมพ์เป็นดิจิตอลใช้ Tesseract ใน การแปลงเอกสารและหนังสือเป็นรูปแบบดิจิตอล
7. การตัดคำภาษาไทยทางคณผู้จัดทำ จะใช้ freeware เช่น DeepCut มาใช้ในส่วนของการตัดคำภาษาไทย

## 1.7 เนื้อหาทางวิศวกรรมที่เป็นต้นฉบับ

- การทำ Image processing สำหรับการเตรียมภาพก่อนนำไปทำ OCR

โครงการของเราทำเกี่ยวกับการทำ OCR เพื่ออ่านภาพให้กลายเป็น text แต่ถึงแม้ว่าภาพที่ได้มาจะจากการสแกนหรือการถ่ายรูป แต่ถึงอย่างนั้น OCR ที่ใช้ก็ยังคงมีข้อจำกัดในเรื่องของคุณภาพของภาพที่ใช้ ถ้าเกิดว่าภาพที่ใช้มี noise จะทำให้การอ่านมีประสิทธิภาพน้อยลง นอกจากนี้การตัดภาพแยกย่อหน้าแต่ละย่อหน้าทำให้การอ่านมีความถูกต้องมากยิ่งขึ้น

- การพัฒนาเว็บไซต์สำหรับการค้นหาหนังสือในหอดหมายเหตุ

เว็บไซต์ของเราระยะใช้ ReactJS, NodeJs, python ใน การพัฒนาเว็บไซต์เป็น Interface ให้กับ user สำหรับการใช้งานระบบการค้นหาหนังสือ รวมถึงการอัปโหลดเอกสารเพื่อแปลงเอกสารเข้าสู่ระบบดิจิตอลและ API ต่าง ๆ

- คัดเลือกคำสำคัญอกรมาเพื่อสร้าง tag  
สำหรับแบ่งแยกหมวดหมู่ของหนังสือโดยใช้หลักการของ TF-IDF ในการค้นหาคำสำคัญของหนังสือเพื่อนำมาสร้าง tag และใช้สำหรับการค้นหาข้อมูล
- ระบบค้นหาโดยใช้คำที่มีความหมายใกล้เคียง  
สำหรับการค้นหาเราจะนำคำแนะนำ TF-IDF มาใช้เป็นเครื่องแอบนเพื่อใช้ในการค้นหาแบบ Cosine similarity และค้นหาคำใกล้เคียง (Query Expansion) เพื่อทำให้การค้นหาเจอผลลัพธ์ที่ต้องการเพิ่มมากขึ้น

## 1.8 การแยกย่อ้งาน และร่างแผนการดำเนินงาน

- ศึกษาและค้นคว้าปัญหาของโครงการ
- เสนอหัวข้อโปรเจค
- ค้นหาข้อมูลเกี่ยวกับเทคโนโลยีที่ใช้ในโปรเจค
- ประเมินความเป็นไปได้และกำหนดขอบเขตของโปรเจค
- จัดเก็บ requirement จากกลุ่มผู้ใช้งาน
  - ติดต่อเจ้าหน้าที่ของหอสมุด
  - เก็บข้อมูลที่ต้องการแปลงเข้าสู่ระบบดิจิตอล
- นำเสนอโครงการครั้งที่ 1
- ออกแบบ UX/UI
- แปลงรูปภาพเป็น Full-text
  - นำเอกสารมาแปลงเป็นรูปภาพในรูปแบบสแกน
  - ศึกษาการใช้งาน OpenCV
  - สร้างระบบ Image processing เพื่อทำการปรับแต่งรูปภาพและทำการปรับแต่งจนได้ระบบที่รองรับกับ Data ที่มี
  - นำรูปที่ผ่านการทำ Image Processing มาเข้าสู่ระบบ OCR เพื่อแปลงข้อมูลจากรูปภาพมาเป็นข้อความในระบบดิจิตอล
- นำข้อมูลที่เก็บไว้มาทำการตัดแบ่งคำภาษาไทยและหาคำสำคัญโดยใช้ TF-IDF
  - ทำการตัดแบ่งคำ (Tokenization)
  - ลบ stop word ออกจากข้อมูล
- ระบบค้นหา
  - ระบบค้นหาโดยใช้หลักการ Cosine Similarity
  - ทำการค้นหาด้วยคำใกล้เคียงโดยใช้ Word2Vec
- จัดทำเว็บไซต์แพลตฟอร์ม
- ทดสอบระบบ
- ปรับปรุงแก้ไข
- นำเสนอโปรเจค

## 1.9 ตารางการดำเนินงาน

ตารางที่ 1.1 ตารางการดำเนินงาน ภาคการศึกษาที่ 1/2563

ตารางที่ 1.2 ตารางการดำเนินงาน ภาคการศึกษาที่ 2/2563

### 1.9.1 ผลการดำเนินงานในภาคการศึกษาที่ 1

- ทำระบบ Image processing สำหรับการเตรียมรูปภาพสำหรับการแปลงข้อมูลเป็นดิจิตอล
- ทำ API ในการตัดคำและจัดการ stop word สำหรับการเตรียมการ text processing
- ทำระบบ Term Frequency-Inverse Document Frequency สำหรับการค้นหาคำสำคัญเพื่อสร้าง tag
- ทำส่วนของการทำการค้นหาข้อมูลเบื้องต้น

### 1.9.2 ผลการดำเนินงานในภาคการศึกษาที่ 2

- ทำระบบค้นหาให้เสร็จสิ้น
- ปรับปรุงระบบค้นหาให้ตอบโจทย์มากยิ่งขึ้น
- ทำเว็บไซต์ platform ทั้งฝั่ง frontend และ backend

## บทที่ 2 ที่มา ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

### 2.1 บทนำ

โดยทฤษฎีที่เกี่ยวข้องกับโปรเจคนี้มีหลากหลายสาขาด้วยกันโดยจะแบ่งเป็นส่วนของ Image Processing โดยการใช้ Open source Computer Vision (OpenCV) เพื่อนำไปใช้กับส่วนของการทำ Optical character recognition (OCR), Tesseract OCR และส่วนของการทำ Natural language processing (NLP) โดยการใช้ Team Frequency Inverse Document Frequency (TF-IDF), Minimum Edit Distance, Deep Cut ส่วนต่อไป Search Engine ได้ใช้ Cosine Similarity และในส่วนการสร้าง Web application โดยใช้ RESTful API และส่วนสุดท้ายการทำ Word Embedding

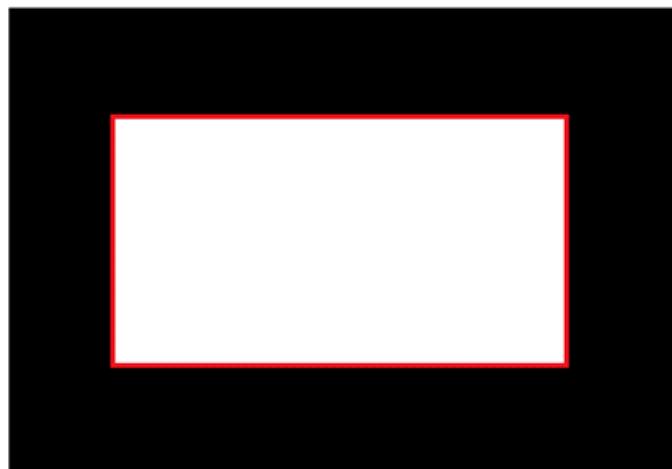
### 2.2 แนวความคิดทางทฤษฎี

#### 2.2.1 Image Processing

เป็นการประมวลผลรูปภาพที่แปลงภาพให้เป็นข้อมูลทางดิจิตอลเพื่อใช้สำหรับปรับคุณภาพของภาพให้ตรงตามความต้องการ อย่างการตัดสิ่งรบกวน การลบกรอบ การมนุรูป หรือการปรับให้ภาพมีความคมชัดมากยิ่งขึ้น ในโปรเจคของเรานั้นเรามาใช้ในการปรับคุณภาพของรูปภาพเพื่อช่วยให้การทำ OCR แม่นยำมากยิ่งขึ้น

##### 2.2.1.1 Contour

Contour [1] คือเส้นเค้าโครงของรูปภาพ ที่เว้าหักขอบเขตพื้นที่ที่มีค่าสีต่อเนื่องกัน หรือค่าเดียวกัน โดยใช้การเปลี่ยนให้รูปภาพอยู่ในรูปของ matrix และเช็คดูว่าค่าสีที่มีความแตกต่างอย่างชัดเจนเริ่มที่ตรงไหนและสร้างเป็นเส้นเค้าโครงขึ้นมาดังรูป 2.1 ซึ่งการทำเส้นเค้าโครงจะทำงานได้ถูกต้องเมื่อเป็นรูปภาพแบบ Binary

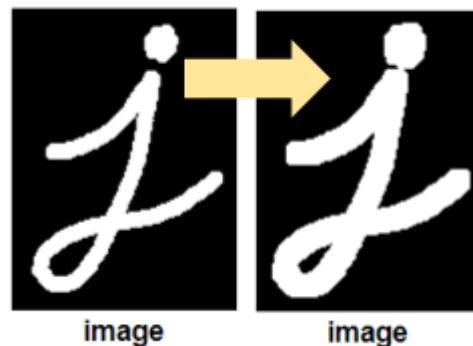


รูปที่ 2.1 แสดงการทำเค้าโครงภายในรูป

### 2.2.1.2 Morphology Transformation

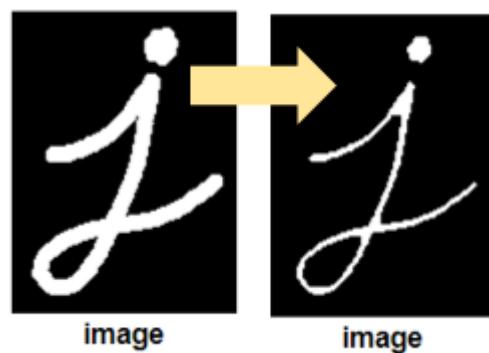
เป็นกระบวนการทาง Image Processing ที่จะทำการนำรูปภาพมาทำการเปลี่ยนแปลงลักษณะ รูปร่างของวัตถุภายในภาพ ปกติแล้วจะใช้ภาพที่เป็น Binary ซึ่งส่วนใหญ่จะใช้สำหรับการกำจัด noise การซ้อมแซมรูปร่างของภาพ หรือการเพิ่มขนาดให้กับวัตถุนั้นๆ โดยการทำ Morphology Transformation นั้นจะมีวิธีการดำเนินการพื้นฐานอยู่ 2 วิธีคือ Dilation และ Erosion

Dilation คือการเพิ่มพื้นที่สีขาวของรูปเพิ่มพื้นที่สีขาวตามขอบพื้นที่สีขาวและจะเปลี่ยนพื้นที่สีดำให้กลายเป็นสีขาวทำให้พื้นที่สีขาวมีความหนามากขึ้นดังรูป



รูปที่ 2.2 แสดงการทำ dilation เพื่อเพิ่มพื้นที่สีขาว

Erosion คือการกร่อนภาพ หรือก็คือจะลดพื้นที่สีขาวของภาพออกไปซึ่งวิธีการนี้ส่วนใหญ่จะใช้สำหรับการแยกสิ่งที่องที่อยู่ด้านหลัง หรือลบ pepper noise ที่เป็น noise เล็กๆได้ โดยจะใช้หลักการเดียวกับ Dilation เพียงแต่จะเปลี่ยนจากพื้นที่สีขาวให้กลายเป็นพื้นที่สีดำดังรูป



รูปที่ 2.3 แสดงการทำ erosion เพื่ogrอนพื้นที่สีขาว

### 2.2.2 Optical character recognition (OCR)

OCR เป็นกระบวนการของการแปลงอักษรบนสื่อสิ่งพิมพ์ให้เป็นข้อความที่สามารถค้นหา แปลงแบบและแก้ไขได้โดยที่ไม่ต้องพิมพ์ขึ้นมาใหม่ ด้วยการทำ Deep learning ใน การเรียนรู้ภาพเพื่อแปลงออกมานเป็นตัวอักษร ซึ่งในปัจจุบันทำได้ด้วยระบบเกี่ยวกับค้นหาที่จะต้องคัดคำอ่านจากสื่อพิมพ์เหล่านั้น จึงจำเป็นที่จะต้องใช้ OCR ในการแปลงภาพดันแบบออกมานเป็นตัวอักษรก่อนที่จะนำไปใช้งานต่อ

จากการศึกษาพบว่าการทำ OCR ภาษาไทยนั้นมีอยู่มากหลายในปัจจุบัน หนึ่งในนั้นมี T - OCR ซึ่งเป็น library ของ AI For Thai [6] และ Tesseract ของ Google [5] ที่ใช้สำหรับแปลงภาพเป็น text ซึ่งโดยกลุ่มของเราเลือกที่จะใช้ Tesseract ในการทำ OCR เนื่องจากไม่เสียค่าใช้จ่ายเมื่อเทียบกับการใช้ OCR ของ AI For Thai นอกจากนั้นเรื่องของการเรียกใช้งานอย่างต่อเนื่อง Tesseract สามารถทำได้ดีกว่าเนื่องจากไม่จำเป็นต้องเรียกใช้งาน AI For Thai จากภายนอก

### 2.2.3 Natural language processing

Natural language processing คือกระบวนการที่ใช้ในทางปัญญาประดิษฐ์ซึ่ง เป็นกระบวนการที่ทำการวิเคราะห์ทางด้านภาษาซึ่งเอาไปประยุกต์ทำให้ปัญญาประดิษฐ์ (AI) สามารถทำให้คอมพิวเตอร์เข้าใจภาษาและตอบกลับได้ใกล้เคียงกับมนุษย์มากขึ้น โดยในปัจจุบันนี้จะใช้มาช่วยในการหาคำสำคัญของหนังสือ และบทความต่าง ๆ เพื่อช่วยให้การค้นหาบทความมีประสิทธิภาพมากขึ้น

#### 2.2.3.1 Information retrieval

Information retrieval คือ เทคโนโลยีการเก็บข้อมูลอย่างนึงโดยจะมีทั้งหมด 2 ลักษณะ ลักษณะที่ 1 คือ Boolean Retrieval เป็นการสร้างโครงสร้างข้อมูลในรูปแบบ Matrix ที่มีค่าเพียงแค่ 0, 1 โดยที่ 0 คือไม่มีคำ (Term) ในเอกสารนั้น และ 1 คือมีคำ (Term) อยู่ภายในเอกสารนั้นหรือเรียกได้ว่าเป็น Term-Document Incidence Matrix ดัง ตารางที่ 2.1 โดยที่ถ้าเราพิจารณาในรูปแบบแຄรายการจะได้ Vector ของ Term นั้นที่ปรากฏอยู่ในเอกสาร ไหนบ้าง แต่การเก็บในรูปแบบ Boolean Retrieval เมื่อมีเอกสาร เเยอยังจะทำให้เกิดค่า 0 ที่ไม่มีประโยชน์มากขึ้นจึงมีลักษณะที่ 2 คือโครงสร้างแบบ Inverted index เป็นการเก็บเพียง Term นั้นอยู่ภายในเอกสาร ไหนบ้างเพื่อจะเก็บแต่เพียงข้อมูลสำคัญเอาไว้ดัง ตารางที่ 2.2 โดย คำ (Term) จะผ่านกระบวนการ Text Processing ประกอบไปด้วย Tokenization (การตัดคำจากประโยค), Normalization (การจัดการคำย่อ), Stemming (การแปลงคำให้อยู่รูปแบบเดียวกัน), Stop words (จัดการคำที่ไม่มีความหมาย) เพื่อเป็นการจัดรูปของคำให้อยู่ในรูปแบบเดียวกันก่อนที่จะนำไปใช้งาน ซึ่งการเก็บข้อมูลแบบ Information retrieval (IR) จะทำให้การค้นหาข้อมูลภาษาในฐานข้อมูลได้อย่างรวดเร็วและมีประสิทธิภาพ

ตารางที่ 2.1 Information retrieval ในลักษณะ Boolean Retrieval

	Antony & Cleopatra	Julius Ceasar	The Tempest	Hamlet	Othello
Antony	1	1	0	0	0
Brutus	1	1	0	1	0
Ceasar	1	1	0	1	1
Calpurnia	0	0	1	0	0
Cleopatra	1	0	1	1	1
Mercy	1	0	1	1	1

Term	Frequency	ID : Document				
		1	Antony & Cleopatra	2	Julius Caesar	
Antony	2	1	Antony & Cleopatra	2	Julius Caesar	
Brutus	3	1	Antony & Cleopatra	2	Julius Caesar	4 Hamlet
Caesar	4	1	Antony & Cleopatra	2	Julius Caesar	4 Hamlet 5 Othello
Calpurnia	1	3	The Tempest	3	The Tempest	4 Hamlet 5 Othello
Cleopatra	4	1	Antony & Cleopatra	3	The Tempest	4 Hamlet 5 Othello
Mercy	4	1	Antony & Cleopatra	3	The Tempest	4 Hamlet 5 Othello

รูปที่ 2.4 Information retrieval ในลักษณะ Index Retrieval

### 2.2.3.2 TF-IDF

เป็นเทคนิคในการคัดแยกความสำคัญผ่านการให้น้ำหนักในแต่ละคำ โดยแบ่งเป็นสองส่วนนั้นคือ TF (Term Frequency) เป็นการดูว่าคำนี้ หรือว่า Term นี้ปรากฏขึ้นภายใน document มา กันอย่างเพียงไหน และ IDF (Inverse Document Frequency) คือการหาความผกผันในความถี่ของเอกสารโดยจะแทนความผกผันที่ทำให้รู้ว่าคำนี้เป็นคำที่มีความสำคัญและพารากราฟในเอกสารนี้ แต่เนื่องจากการดูคุณภาพ IDF เพียงอย่างเดียวไม่สามารถบอกได้ว่า Term นั้นเป็นคำสำคัญ จึงจำเป็นต้องนำค่า TF มาคูณกับ IDF เป็นค่า TF-IDF เพื่อคุณภาพสำคัญของ Term นั้น ในส่วนการคำนวณนี้เพื่อนำไปใช้ในการค้นหาแบบ Cosine Similarity ต่อไป โดยที่ TF จะใช้เป็น Log normalization โดยคำนวนได้จากสมการ 2.1 ซึ่ง  $f_{t,d}$  คือความถี่ของคำ (Term) ที่ปรากฏขึ้นภายใน Document ส่วน IDF จะคำนวนจากสมการ 2.2 ซึ่ง N คือจำนวน Document ที่มีภายในระบบ และ  $n_t$  คือ จำนวนของ document ที่มีคำ (term) น้อยๆ เมื่อหาค่าทั้ง TF และ IDF ได้แล้วก็จะหาค่าของ TF-IDF ได้จากสมการ 2.3

$$tf = \log(1 + f_{t,d}) \quad (2.1)$$

$$idf = \log \frac{N}{n_t} \quad (2.2)$$

$$TF - IDF = tf * idf \quad (2.3)$$

### 2.2.3.3 Cosine Similarity

เป็นหน่วยวัดความคล้ายคลึงกันระหว่างข้อมูลสอง Vector โดยวัดจากมุม cosine ของจาก Vector ทั้งสองโดยคำนวนได้จากสมการ 2.4 โดยที่  $\|x\|, \|y\|$  คือ สมการของ Euclidean norm ของ Verter x, y ดังสมการ 2.5 โดยในโปรเจคนี้เราได้นำค่าน้ำหนักของ TF-IDF มาเป็นน้ำหนักในการคิดค่า Cosine Similarity โดยนำประโยชน์ที่จะค้นหามาผ่านกระบวนการ Text processing ก่อนที่จะนำมาค้นหาว่า document ไหนมีค่า relevance score (คะแนนความสัมพันธ์) เพื่อนำมาเรียงค่าคะแนนสูงสุดแสดงเป็นผลลัพธ์การค้นหา

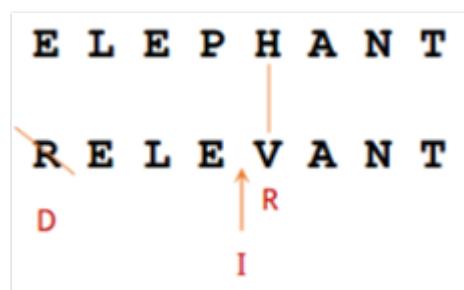
$$\sin(x, y) = \frac{x * y}{\|x\| \|y\|} \quad (2.4)$$

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \quad (2.5)$$

#### 2.2.3.4 Minimum Edit Distance

เป็นหลักการที่หาระยะห่างที่สั้นที่สุดจากคำนึงไปสู่อีกคำนึงจะมีความแตกต่างกันเท่าไหร่ซึ่งจะหลักการเช็คความห่างของคำทั้งหมดสามรูปแบบ

- รูปแบบ Insert(I) จะเป็นการเพิ่มตัวอักษรลงไประบในคำนั้น เพื่อคำดังเดิมของเราจะเปลี่ยนแปลงเป็นคำที่เราต้องการ
- รูปแบบ Delete(D) จะเป็นการลบตัวอักษรออกไประบในคำนั้น เพื่อคำดังเดิมของเราจะเปลี่ยนแปลงเป็นคำที่เราต้องการ
- รูปแบบ Replace(R) จะเป็นการเปลี่ยนตัวอักษรนั้นให้เป็นตัวอักษรใหม่ เพื่อคำดังเดิมของเราจะเปลี่ยนแปลงเป็นคำที่เราต้องการ



รูปที่ 2.5 หลักการการเช็ค edit distance [10]

หลังจากมีรูปแบบการวัดระยะห่างของคำดังรูปภาพที่ 2.5 เล้า จะต้องทำการหาคำที่สั้นที่สุดผ่านรูปแบบของตารางดังรูปภาพที่ 2.6 ซึ่งการคำนวนผ่านตารางจะเป็นการนำกระทำก่ออันหน้ามาคำนวนเรื่อยๆ จนได้รูปการเปลี่ยนเป็นคำใหม่ที่ใช้การเปลี่ยนน้อยที่สุด

	E	L	E	P	H	A	N	T	
R	0	1	2	3	4	5	6	7	8
E	1	1	2	3	4	5	6	7	8
L	2	1	2	2	3	4	5	6	7
E	3	2	1	2	3	4	5	6	7
V	4	3	2	1	2	3	4	5	6
A	5	4	3	2	2	3	4	5	6
N	6	5	4	3	3	3	3	4	5
T	7	6	5	4	4	4	4	3	4
	8	7	6	5	5	5	5	4	3

รูปที่ 2.6 ตัวอย่างตารางการทำ minimum edit distance [10]

ซึ่งในโครงการของเราได้ตั้งหลักการ Minimum edit distance มาใช้ในการตรวจสอบหากคำที่ส่องไม่ถูกต้องโดยมีเกณฑ์ตั้งไว้ว่าถ้าเกินที่กำหนดไว้จะถือว่าคำ ๆ นั้นส่องไม่ถูกต้องแล้วถูกแก้ให้เป็นคำที่ส่องถูกต้อง

#### 2.2.4 RESTful Service

เป็นการสร้าง web service โดยเรียกใช้ผ่านทาง HTTP Method ทั้ง 4 ประเภท GET/POST/PUT/DELETE ส่งข้อมูลออกมายังรูปของ XML ทำให้ปริมาณข้อมูลที่ส่งมากน้อยกว่าการใช้ Protocol SOAP โดยโครงสร้างของ HTTP Request ดังรูปภาพที่ 2.7 ประกอบด้วย

1. VERB: แสดง method ของ HTTP
2. URI: ตำแหน่งของข้อมูลที่ต้องการ
3. HTTP Version: เวอร์ชันของ HTTP
4. Request Header: Metadata ที่เก็บข้อมูลในรูปแบบ Key-Value ของ header
5. Request Body: ส่วนเก็บข้อมูลของเนื้อหา

### HTTP Request



รูปที่ 2.7 แสดงถึงโครงสร้างของ HTTP Request [12]

HTTP Response ดังรูปภาพที่ 2.8 ประกอบด้วย

1. HTTP Version: เวอร์ชันของ HTTP
2. Response Code: รหัสผลลัพธ์ของการทำงานในระดับ HTTP เป็นเลข 3 หลัก
3. Response Header: Metadata ที่เก็บข้อมูลในรูปแบบ Key-Value ของ header
4. Response Body: ส่วนเก็บข้อมูลของเนื้อหา

## HTTP Response



รูปที่ 2.8 แสดงถึงโครงสร้างของ HTTP Response [12]

### 2.2.5 Word Embedding

เป็นวิธีการที่จะเปลี่ยนคำปกติเป็น vector ที่อยู่ในหลากหลายมิติและขนาดเพื่อให้สามารถเปรียบเทียบคำต่าง ๆ ว่ามีความสัมพันธ์ใกล้เคียงกับคำไหนบ้างในระบบเพื่อที่ใช้สำหรับการทำคำที่มีความหมายใกล้เคียงกันโดยมีการทำ word embedding มากมายไม่ว่าจะเป็น Word2Vec [11] [4] ที่ถูกสร้างโดยทีมวิจัยของ Google FastText [3] เป็น word embedding อีกหนึ่งตัวที่สร้างขึ้นจากทีมวิจัยของ facebook หรือจะเป็น ELMo [9] ที่เป็นรูปแบบการ word embedding ที่ครุ่นคำโดยรอบเป็นต้น

## 2.3 ภาษาคอมพิวเตอร์และเทคโนโลยี

### 2.3.1 Open source Computer Vision (OpenCV)

เป็นซอฟต์แวร์ที่เกี่ยวกับการประมวลผลภาพที่มีการสนับสนุนการพัฒนาจาก Intel Corporation โดยที่ตัว OpenCV นั้นเป็น Library Open Source โดยมีจุดประสงค์เพื่อให้นำไปต่อยอดการพัฒนาโปรแกรมในด้าน การรับรู้ของเห็นของคอมพิวเตอร์ (Computer Vision) ให้เข้าใจไม่ว่าจะเป็นภาพนิ่ง (Image) หรือจะเป็นภาพเคลื่อนไหว (Video) โดยภายในโปรเจคนี้ได้นำ OpenCV มาเป็นตัวทำ Image processing โดยที่นำรูปภาพที่มีมาจัดແղนหนังสือ / เอกสาร มาทำการปรับปรุงคุณภาพรูปภาพให้เหมาะสมกับการทำส่วน Optical character recognition (OCR) ให้มีความแม่นยำมากยิ่งขึ้นจากการลบรูปภาพ การลบสิ่งที่ลับกวน การลบกรอบตาราง การหมุนรูป

### 2.3.2 Tesseract OCR

เป็นหนึ่งใน library ที่เกี่ยวกับการทำ Optical character recognition (OCR) ที่ถูกพัฒนาโดย Google โดยเป็น Library Open Source ที่ใช้ในการทำเทียบกับ Text Detection โดยสามารถเรียกใช้งานได้ผ่าน Command line หรือจะเป็นการเรียก API ภายในโปรแกรมที่ทำได้นอกจากนั้น Tesseract เวอร์ชัน 5.0.0 beta มีการใช้ Convolutional Neural Network (CNN) [7] ร่วมกันกับ Long short-term memory (LSTM) เพื่อให้การทำงานやすลัดขึ้นโดยเราจะนำตัว Tesseract มาทำเป็น OCR ภายในโปรเจคนี้

### 2.3.3 DeepCut

เป็น library ในภาษา python ที่สร้างมาจาก True Corporation โดยมีลักษณะเด่นที่ใช้ CNN (Convolutional Neural network) [7] มาช่วยทำให้ผลลัพธ์ที่ได้ออกมามีความแม่นยำที่ค่อนข้างสูง ซึ่งโปรเจกของเราต้องการ DeepCut เพื่อที่จะสามารถแบ่งคำจากรูปประโยคภาษาไทยที่มีความซับซ้อน และไม่แบ่งแยกชัดเจนเหมือนภาษาอังกฤษ

### 2.3.4 ReactJS

เป็นหนึ่งใน library หรือจะเรียกว่าเป็น Framework ที่ Facebook เป็นคนสร้างขึ้นโดยทีมหน้าที่เป็นการสร้าง UI โดยมีความคิดมากจากรูปแบบ MVC [14] (Model View Controller) หรือก็คือเป็นตัวจัดการกับ Model กับ View ของตัวเว็บไซต์ โดยในโปรเจคนี้ได้เลือกใช้ ReactJS เป็น Front End สำหรับการทำ platform Web Application

### 2.3.5 Python

Python เป็นภาษาทางโปรแกรมซึ่งเป็นภาษาทางคอมพิวเตอร์ระดับสูงที่ออกแบบมาให้ใกล้เคียงกับภาษามนุษย์มากที่สุดเพื่อให้สามารถเข้าใจได้ง่ายมากทีนั้น ซึ่งในโปรเจค มีข้อมูลที่ต้องประมวลในแต่ละครั้งมีขนาดใหญ่อาจจะทำให้เกิดความล่าช้าในแต่ละการประมวล ทางผู้จัดทำจึงเลือกใช้ python เนื่องจากรองรับในส่วนของการทำ thread รวมถึงนำมาใช้ในการทำ Data preparation ทั้งการทำ image processing และการเตรียมข้อมูลต่างๆ หลังจากการทำ OCR นอกจากนี้ยังใช้ในการทำ Web server อีกด้วย

#### 2.3.5.1 Django

เป็น REST Framework ที่ใช้ภาษา python เป็นฐาน โดยในโปรเจคนี้เราจะนำมาระบุ REST API เพื่อใช้ในการใช้ library อย่างเช่น DeepCut หรือ OCR ที่สามารถใช้ร่วมการแบ่ง multi thread ได้อย่างมีประสิทธิภาพ และยังสามารถจัดการข้อมูลใน database สำหรับโปรเจคนี้

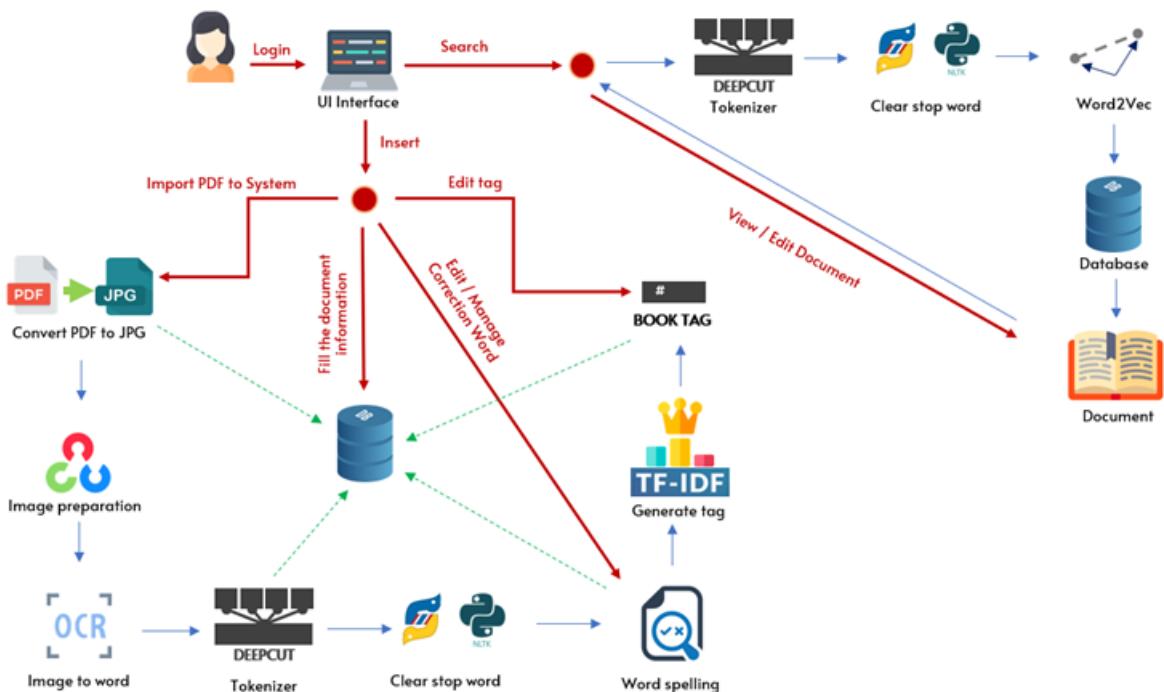
### 2.3.6 NodeJS

NodeJS เป็นสมุดโน้ตฟอร์มที่ใช้ภาษา JavaScript ที่มี library สำหรับใช้จัดการกับฝั่ง Server ซึ่ง NodeJS นั้นมีความยืดหยุ่นสูงที่สำหรับการจัดการ Web server โดย library ที่นำมาใช้คือ express เป็น web server ที่เป็น RESTful API ได้

## บทที่ 3 การออกแบบและระบบวิธีวิจัย

### 3.1 System Overview

บทนี้จะกล่าวถึงภาพรวมของระบบโดยแสดงเป็นโครงสร้างแบบตัวอย่างรูปที่ 3.1 ซึ่งประกอบไปด้วยการออกแบบระบบฐานข้อมูล ระบบการตัดคำ ระบบการประมวลรูปภาพ และการออกแบบ interface สำหรับการใช้งาน



รูปที่ 3.1 System Overview

### 3.2 Feature lists

#### 3.2.1 การแปลงเอกสารเป็นรูปภาพ

สำหรับการแปลงเอกสารผู้ใช้จะต้องทำการอัปโหลดไฟล์ PDF ของเอกสารเข้าสู่ระบบหลังจากนั้นจะระบบจะทำการแปลงแต่ละหน้าเป็นรูปภาพ JPG เพื่อนำไปใช้ต่อในขั้นตอนต่อไปและนำใช้แสดงภายใน web application

#### 3.2.2 Image preparation

ในส่วนของการจัดการรูปภาพที่จะทำการ OCR ซึ่งรูปภาพนั้นมา OCR นั้นมาจากการสแกนทำให้ภาพส่วนใหญ่อยู่ในสภาพดีแต่บางครั้งมี noise และมีความผิดพลาดจากการสแกน เช่น ภาพเอียง หรือตัวหนังสือไม่ชัดเกิดจากการขยายบ้านระหว่างการสแกน หรือมีพื้นหลังสีที่ทำให้ OCR ไม่มีประสิทธิภาพ ดังนั้นจึงต้องมีการทำ Image processing ก่อนที่จะผ่านไปทำ OCR ซึ่งในการทำ Image processing นั้นทางคณะผู้จัดทำได้ออกแบบไว้ว่าจะทำการแยกกระหว่างรูปและตัวหนังสือออกจากกัน โดยการใช้ contour เข้ามาช่วยในการคัดแยกรูปออกจากตัวอักษร โดยดูจากพื้นที่สีเหลี่ยมที่ได้จากการสแกน กลับพื้นที่ contour ว่ามีความต่างขนาดและความแตกต่างกันมากเท่าไร หรือใช้ขนาดความกว้าง

และพยายามดูว่ามีขนาดเดกินเท่าไรถึงจะตัดให้เป็นรูปภาพ นอกจากนั้นอักแบบการหมุนโดยสร้าง contour บรรทัดและวัดความเอียงของแต่ละบรรทัดว่าเอียงเท่าไรจากนั้นจึงหมุนกลับในองค์ตรงข้าม

### 3.2.3 Image to word

สำหรับการทำแปลงเอกสารเป็นข้อมูลดิจิตอลจะใช้ Tesseract OCR โดยจะใช้รูปภาพที่ผ่านกระบวนการ Image processing และประโภคที่แปลงออกมาได้จัดเก็บไว้ใช้งานต่อไป

### 3.2.4 Text preprocessing

สำหรับการทำ Text preprocessing จะประกอบไปด้วยการทำ Tokenizer หรือก็คือการตัดคำออกมาจากประโยคโดยการใช้อัลกอริทึม DeepCut และนำคำไปทำ Lemmatization หรือก็คือการลดรูปให้อยู่ในรูปแบบพื้นฐานของคำศัพท์เฉพาะภาษาอังกฤษโดยใช้ library nltk เป็นตัวจัดการก่อนจะนำไปลบ stop word คือการลบคำที่ไม่มีความหมายออกไปโดยใช้กลุ่มข้อมูลของ library pythianlp ก่อนนำไปตรวจสอบคำผิดก่อนโดยใช้อัลกอริทึมของ pythianlp และตรวจสอบคำเฉพาะที่คณผู้จัดทำได้กำหนดไว้โดยใช้ Minimum edit distance ก่อนที่จะนำไปใช้งานต่อไป

### 3.2.5 Tag generated

หลังจากที่นำข้อมูลที่ได้จากการทำ OCR และทำการเตรียมข้อมูลเสร็จเรียบร้อยแล้ว ระบบจะทำการคืนค่าแต่ละหน้าให้กับผู้ใช้เพื่อที่จะให้ผู้ใช้สามารถเข้าคำทำให้ระบบอ่าน และแก้ไขคำเหล่านี้ให้หลังจากนั้นเมื่อผู้ใช้เข้าคำเสร็จแล้ว ระบบจะนำคำทำหัวหน้าที่ได้ไปคิดคำนวนเพื่อทำการสร้างคลาสแนนให้แต่ละคำและทำการจัดลำดับคลาสแนนให้กับหนังสือเล่มนั้น ๆ โดยใช้การคิดคลาสแนนด้วยอัลกอริทึม TF-IDF

### 3.2.6 Search

ในส่วนของระบบการค้นหาหน้าที่ผู้ใช้ทำการกรอกคำค้นหาระบบจะทำการนำคำที่ผู้ใช้กรอกมาทำ Text preprocessing อีกครั้งหนึ่งแต่จะไม่ดำเนินส่วนของการตรวจสอบคำผิด และคำที่ได้จะถูกนำไปเข้าโมเดล Word2Vec เพื่อนำไปค้นหาคำใกล้เคียงของคำค้นหาก่อนที่จะนำคำที่ได้ไปค้นในฐานข้อมูลเพื่อค้นหาหนังสือที่มีความใกล้เคียงกับคำค้นมากที่สุด เมื่อได้หนังสือมาระบบจะทำการส่งข้อมูลหนังสือกลับไปให้ผู้ใช้

### 3.2.7 Manage Book

ในการจัดการข้อมูลเอกสารภายในระบบจะแบ่งทั้ง 3 ส่วนนั้นคือ 1. การเพิ่มเอกสารเข้าสู่ระบบ 2. การแก้ไขเอกสารภายในระบบ 3. การลบเอกสารออกจากระบบ ส่วนที่ 1. ใน การเพิ่มเอกสารเข้าสู่ระบบ ผู้ใช้งานจะต้องอัปโหลดไฟล์เอกสารในรูปแบบ PDF และกรอกรายละเอียดของเอกสารเพื่อเข้าสู่กระบวนการแปลงเอกสารเป็นรูปภาพต่อไปจะเป็นการทำ Image processing ก่อนที่จะนำมาทำการแปลงภาพเป็นตัวอักษรเพื่อที่จะได้ข้อมูลดิจิตอลจากเอกสารที่ผู้ใช้เพิ่มเข้าสู่ระบบหลังจากนั้นจะเป็นการทำ Text preprocessing และให้ผู้ใช้งานได้ตรวจสอบแก้ไขหรือเพิ่มเติมก่อนจะสั่งสุดการเพิ่มเอกสารเข้าสู่ระบบ ในส่วนที่ 2 การแก้ไขเอกสารภายในระบบผู้ใช้งานสามารถค้นหาเอกสารภายในระบบเพื่อนำมาแก้ไขรายละเอียดที่ผู้ใช้งานกรอกเท่านั้นแต่สามารถแก้ไขคำที่ถูกแปลงออกมาเป็นตัวอักษรได้ และส่วนสุดท้ายการลบเอกสารในระบบผู้ใช้งานสามารถลบเอกสารภายในระบบได้โดยการค้นหาเอกสารที่ต้องการและกดลบเอกสารนั้นออกจากระบบโดยเมื่อมีการลบเอกสารออกก็จะลบคำที่มีอยู่ในเอกสารออกไปจากระบบทั้งกัน

### 3.2.8 Login

ผู้ใช้งานสามารถเข้าสู่ระบบเพื่อใช้งานฟังก์ชันต่าง ๆ ภายในระบบโดยเมื่อผู้ใช้งานทำการเข้าสู่ระบบด้วยชื่อผู้ใช้งานและรหัสผ่านแล้วจะได้รับ “token” เพื่อที่จะใช้สำหรับการยืนยันตัวในการใช้ฟังก์ชันต่าง ๆ ภายในระบบและผู้ใช้งานจะสามารถออกจากระบบได้

## 3.3 System requirements

### ผู้ใช้งาน

ใช้งานได้บนระบบ web browser

- Google Chrome เวอร์ชัน 84.0 ขึ้นไป
- Microsoft Edge เวอร์ชัน 83.0 ขึ้นไป
- Firefox เวอร์ชัน 75.0 ขึ้นไป

### ผู้เชื่อมโยง

ทางด้าน Hardware

- CPU: Intel or AMD processor with 64-bit โดยที่ต้องมี 2 Core ขึ้นไป
- GPU: NVIDIA 1050ti or higher
- Disk Storage: 10 GB
- RAM: 8GB or higher

ทางด้าน Software แบ่งเป็น 2 ส่วนคือ Python และ JavaScript

#### 1. Python Backend

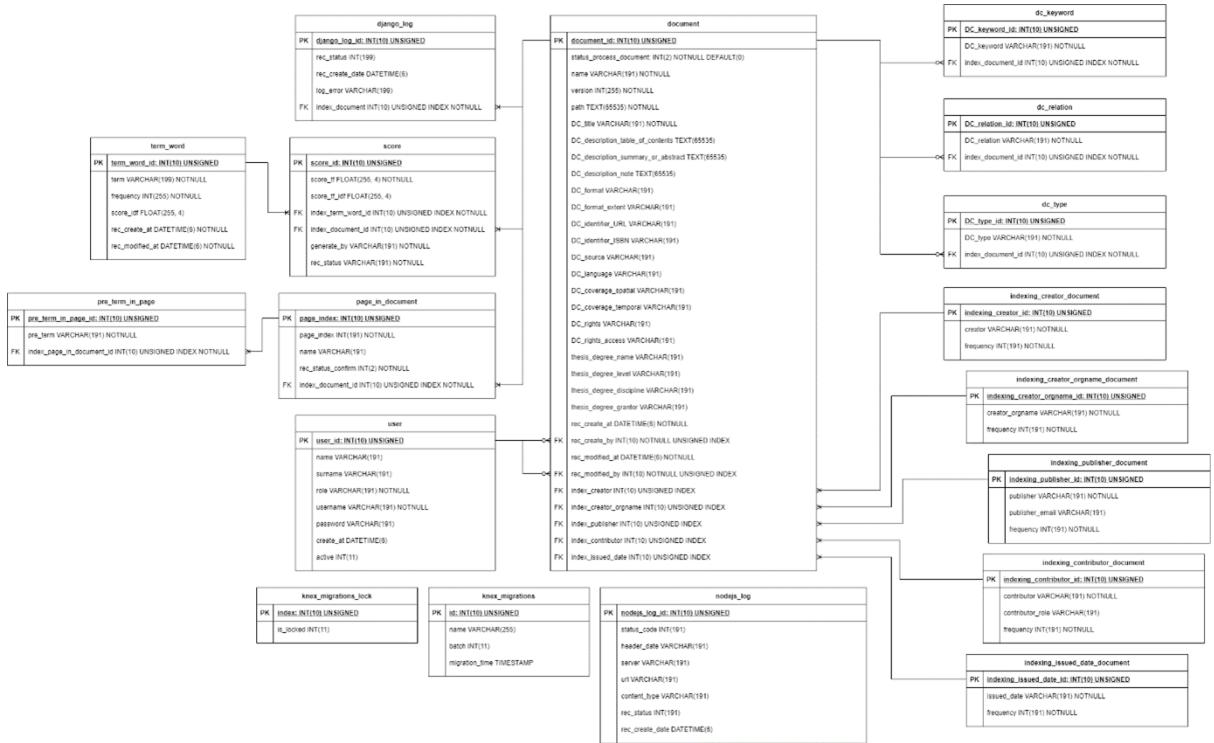
- Python เวอร์ชัน 3.7.5
- Tensorflow เวอร์ชัน 2.3.1
- DeepCut เวอร์ชัน 0.7
- Django เวอร์ชัน 3.1.3
- Djangorestframework เวอร์ชัน 3.12.2
- Django-cors-headers เวอร์ชัน 3.5.0
- Pythainlp เวอร์ชัน 2.2.5
- Pyspellchecker เวอร์ชัน 0.5.5
- nltk เวอร์ชัน 3.5.0
- mysqlclient เวอร์ชัน 2.0.1
- pillow เวอร์ชัน 8.0.1

- shapely เวอร์ชัน 1.7.1
- pytesseract เวอร์ชัน 5.0.0 beta
- opencv-python เวอร์ชัน 4.4.0.46
- pdf2image เวอร์ชัน 1.14.0
- scipy เวอร์ชัน 1.5.4

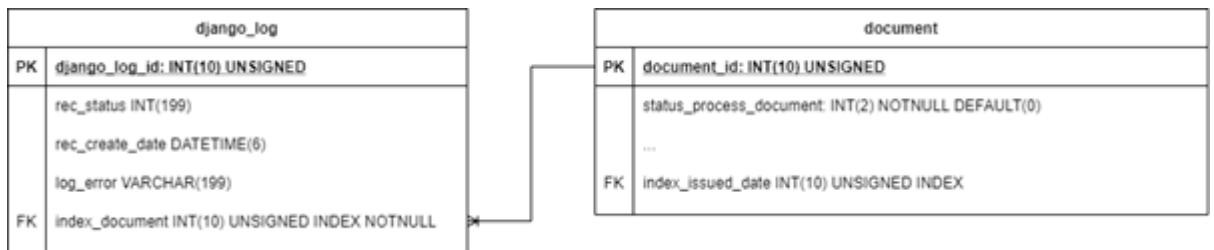
## 2. JavaScript Backend and Frontend

- nodejs เวอร์ชัน 12.16.3
- apollo-server-express เวอร์ชัน 2.19.0
- axios เวอร์ชัน 0.20.0
- cors เวอร์ชัน 2.8.5
- dotenv เวอร์ชัน 8.2.0
- express เวอร์ชัน 4.17.1
- graphql เวอร์ชัน 15.4.0
- jsonwebtoken เวอร์ชัน 8.5.1
- knex เวอร์ชัน 0.21.5
- morgan เวอร์ชัน 1.10.0
- mysql2 เวอร์ชัน 2.2.1
- password-hash เวอร์ชัน 1.2.2
- react เวอร์ชัน 16.13.1
- react-hook-form เวอร์ชัน 6.3.1
- react-router-dom เวอร์ชัน 5.2.0
- styled-components เวอร์ชัน 5.1.1
- props-types เวอร์ชัน 15.7.2

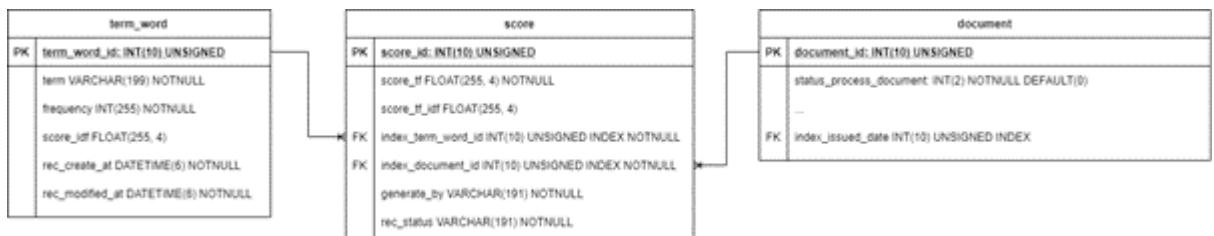
### 3.4 Database Design



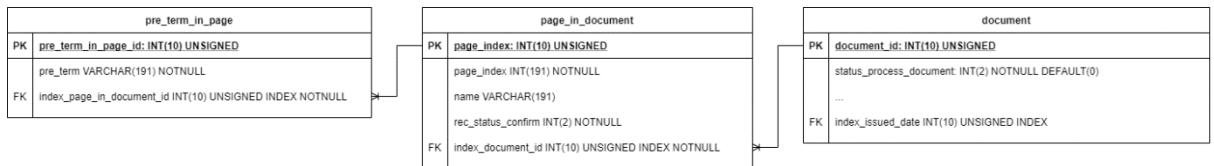
รูปที่ 3.2 แสดง ER Diagram ของฐานข้อมูล



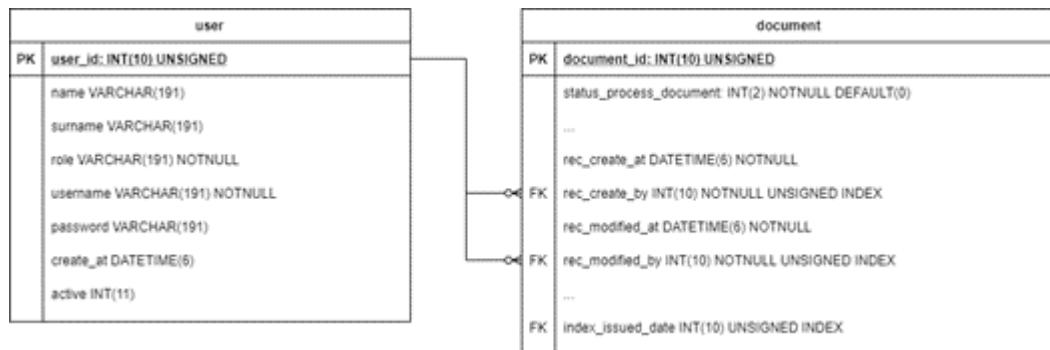
รูปที่ 3.3 แสดง ER Diagram ส่วนของการเก็บความผิดพลาดในการสร้างคีย์เวิร์ดจากเอกสาร



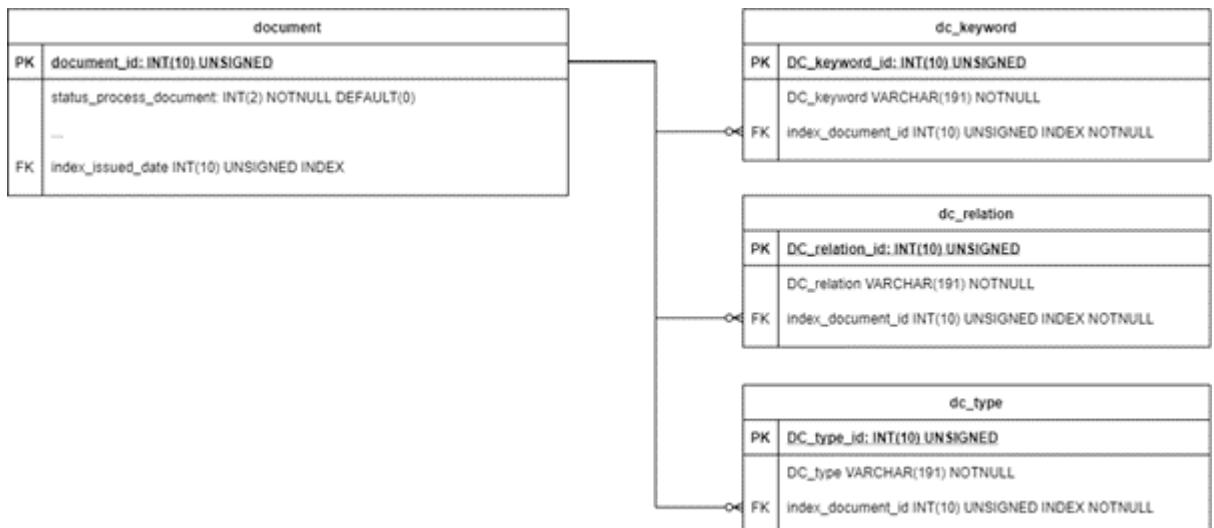
รูปที่ 3.4 แสดง ER Diagram ส่วนของคีย์เวิร์ดและคะแนนความสำคัญในระบบ



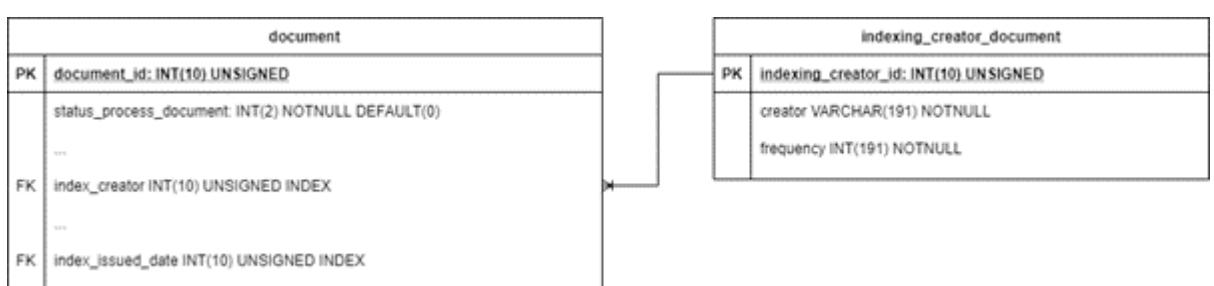
รูปที่ 3.5 แสดง ER Diagram ส่วนของการเก็บคำจากแต่ละหน้าที่แปลงมาจากการอ่านเอกสาร



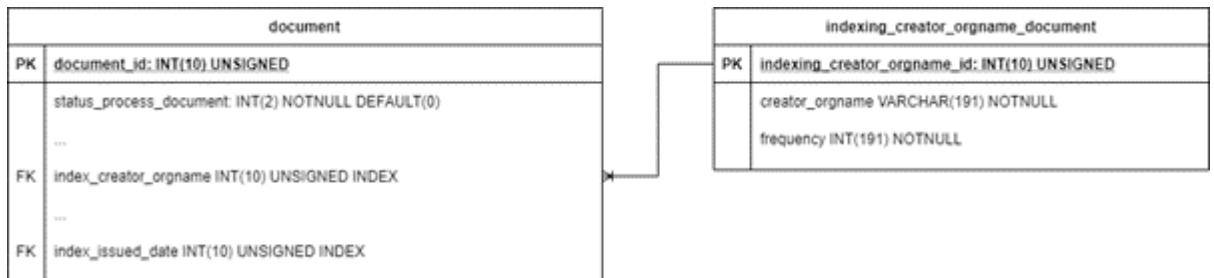
รูปที่ 3.6 แสดง ER Diagram ส่วนของประวัติของผู้ใช้งานมีการสร้างหรือแก้ไขเอกสาร



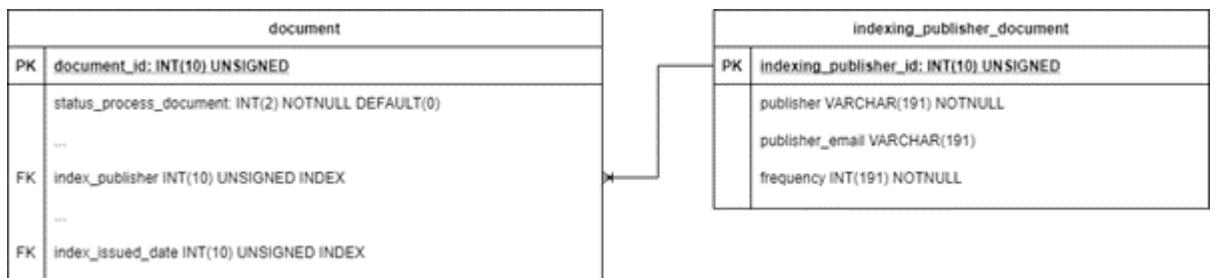
รูปที่ 3.7 แสดง ER Diagram ส่วนของการเก็บข้อมูล keyword, relation, type ของเอกสาร



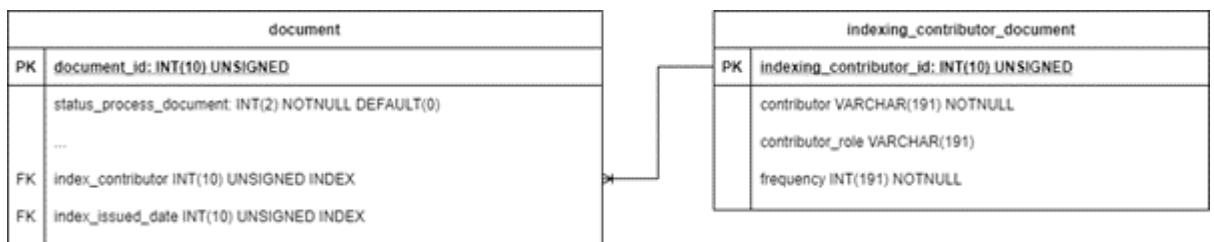
รูปที่ 3.8 แสดง ER Diagram ส่วนของ Creator มีความเกี่ยวข้องกับเอกสารในหน้าบัง



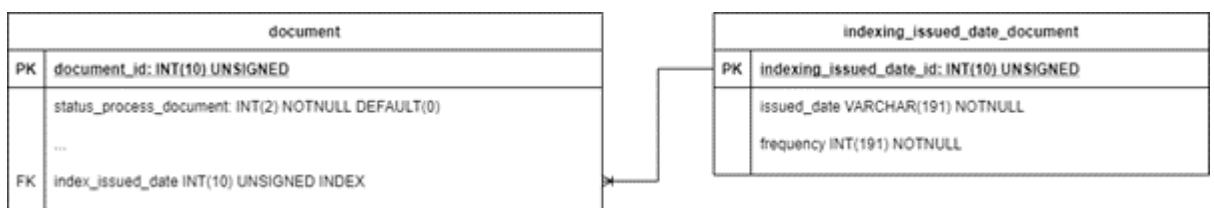
รูปที่ 3.9 แสดง ER Diagram ส่วนของ Creator Organized Name มีความเกี่ยวข้องกับเอกสารในบัง



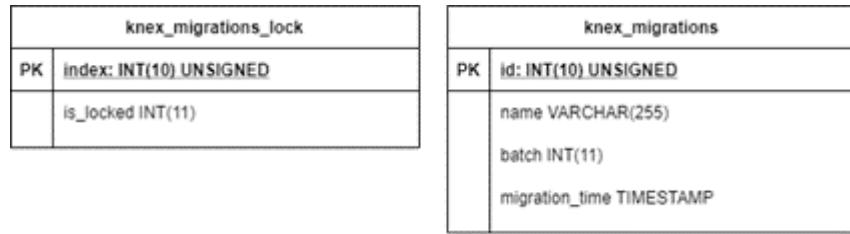
รูปที่ 3.10 แสดง ER Diagram ส่วนของ Publisher มีความเกี่ยวข้องกับเอกสารในบัง



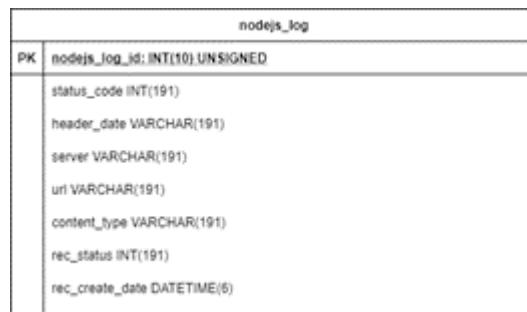
รูปที่ 3.11 แสดง ER Diagram ส่วนของ Contributor มีความเกี่ยวข้องกับเอกสารในบัง



รูปที่ 3.12 แสดง ER Diagram ส่วนของ Issued Date มีความเกี่ยวข้องกับเอกสารในบัง



รูปที่ 3.13 แสดง ER Diagram ส่วนของ Knex module ที่ใช้สำหรับ Migration ฐานข้อมูล



รูปที่ 3.14 แสดง ER Diagram ส่วนของการเก็บประวัติการ HTTP Request NodeJS ไปยัง Django

### 3.4.1 Database Structure

รูปที่ 3.2 แสดงฐานข้อมูลของทั้งระบบโดยจะมีหลัก ๆ ทั้งหมดสามส่วน ทางด้านฝั่งขวาของตาราง document จะเป็นตารางที่เก็บข้อมูลเพิ่มเติมจากตาราง document และส่วนทางด้านฝั่งซ้ายของตาราง document สำหรับการเก็บข้อมูลในด้านของการทำระบบการเก็บคำจากเอกสารที่ถูกใส่ลงในระบบ ระบบการแปลงคำเป็นคีย์เวิร์ดและคะแนน TF-IDF ที่นำมาใช้สำหรับการค้นหาเอกสาร ระบบจัดการฐานข้อมูลผู้ใช้งาน และการตรวจสอบความผิดพลาดที่มีจากการสร้างคีย์เวิร์ด และส่วนสุดท้ายที่เป็นตารางที่ไม่มีการเชื่อมโยงกับตารางใด ๆ จะมีไว้สำหรับการทำระบบฐานข้อมูล และระบบตรวจสอบ HTTP Request ของทาง NodeJS

รูปที่ 3.3 จะเป็นส่วนของการเก็บข้อผิดพลาดที่มาจากการระหว่างการสร้างคีย์เวิร์ด และการสร้างคะแนน Term-Frequency โดยในส่วนนี้มีตาราง document ที่มีความสัมพันธ์ 1 to many กับตาราง django\_log กล่าวก็คือในหนึ่งเอกสารมีได้หลายบันทึกของ Django เนื่องมีโอกาสที่เกิดความผิดพลาดแล้วต้องทำใหม่จนกว่าจะไม่มีความผิดพลาดเกิดขึ้น

รูปที่ 3.4 จะเป็นส่วนของคีย์เวิร์ด และคะแนนเพื่อนำมาใช้สำหรับการค้นหาเอกสารของระบบนี้ โดยจะมีทั้งหมดสามตาราง document, term\_word, score ตาราง document จะเป็นตารางที่เก็บข้อมูลของเอกสารไว้ ส่วนตาราง term\_word จะเป็นการเก็บคีย์เวิร์ด และคะแนน IDF สำหรับการลดความสำคัญของคีย์เวิร์ดนั้น ๆ ไปซึ่งทั้งสองตารางนี้จะเป็นความสัมพันธ์แบบ one to many กับตาราง score ที่จะมีคะแนนสำหรับระบบการค้นหาเก็บเอาไว้ ที่มีความสัมพันธ์แบบนี้เนื่องจากในแต่ละคีย์เวิร์ดมีโอกาสพบได้ในหลายเอกสาร และเอกสารเองก็สามารถมีได้หลายคีย์เวิร์ด เนื่องจากแต่ละคีย์เวิร์ดที่อยู่ต่างเอกสารกันจะมีคะแนนไม่เท่ากัน

รูปที่ 3.5 จะเป็นส่วนของการเก็บคำที่แปลงมาจากเอกสารไว้โดยเริ่มที่ตาราง document จะที่สามารถอัดได้ว่าเอกสารไหน ที่จะมีความพันธ์ one to many ไปยังตาราง page\_in\_document ที่จะเป็นตารางที่บอกถึงหน้าต่าง ๆ ในเอกสารนั้น และยังมีความสัมพันธ์ one to many ต่อไปยังตาราง per\_term\_in\_page ที่จะมีคำต่าง ๆ เก็บเอาไว้ ดังนั้นจะเป็นความสัมพันธ์ที่เอกสารนั้นจะสามารถมีได้หลายหน้า แล้วแต่ละหน้าเองก็จะมีคำต่าง ๆ ที่แปลงออกมากถูกเก็บเอาไว้

รูปที่ 3.6 จะเป็นความสัมพันธ์ของบัญชีผู้ใช้กับเอกสาร โดยจะมีตาราง user ที่จะเก็บข้อมูลของผู้ใช้งานที่มีความสัมพันธ์แบบ one to many ไปยังตาราง document ที่จะเก็บต้องเก็บข้อมูลของผู้ใช้ว่าผู้ใช้คนไหนเป็นคนสร้าง หรือแก้ไขเอกสารนี้ ซึ่งบัญชีผู้ใช้สามารถสร้างหรือแก้ไขเอกสารได้หลายเอกสาร

รูปที่ 3.7 จะเป็นส่วนของข้อมูลของตาราง Document เมื่ອันกันแต่เนื่องจากข้อมูลนี้มากกว่าหนึ่งทำให้ต้องสร้างความสัมพันธ์แบบ one to many กับตาราง dc\_keyword, dc\_relation, dc\_type ซึ่งจะเป็นข้อมูลคีย์เวิร์ด ความสัมพันธ์ และประเภทของเอกสารตามลำดับ

รูปที่ 3.8 จะเป็นส่วนของการเก็บความสัมพันธ์ระหว่าง Creator กับเอกสาร เนื่องจาก Creator สามารถมีได้หลายเอกสารทำให้ตาราง indexing\_creator\_document จะเป็นความสัมพันธ์แบบ one to many กับตาราง document

รูปที่ 3.9 จะเป็นส่วนของการเก็บความสัมพันธ์ระหว่าง Creator orgname กับเอกสารนี้ 既然 Creator orgname สามารถมีได้หลายเอกสารทำให้ตาราง indexing\_creator\_document จะเป็นความสัมพันธ์แบบ one to many กับตาราง document

รูปที่ 3.10 จะเป็นส่วนของการเก็บความสัมพันธ์ระหว่าง Publisher กับเอกสาร เนื่องจาก Publisher สามารถมีได้หลายเอกสารทำให้ตาราง indexing\_publisher\_document จะเป็นความสัมพันธ์แบบ one to many กับตาราง document

รูปที่ 3.11 จะเป็นส่วนของการเก็บความสัมพันธ์ระหว่าง Contributor กับเอกสาร เนื่องจาก Contributor สามารถมีได้หลายเอกสารทำให้ตาราง indexing\_contributor\_document จะเป็นความสัมพันธ์แบบ one to many กับตาราง document

รูปที่ 3.12 จะเป็นส่วนของการเก็บความสัมพันธ์ระหว่าง Issued Date กับเอกสาร เนื่องจาก Issued Date สามารถมีได้หลายเอกสารทำให้ตาราง indexing\_issued\_date\_document จะเป็นความสัมพันธ์แบบ one to many กับตาราง document

รูปที่ 3.13 จะเป็นสองตารางที่บันทึกการจัดการฐานข้อมูลของเครื่องมือที่ชื่อว่า Knex ที่ทำการจัดการสร้างฐานข้อมูล ด้วยคำสั่ง Migration และหลังจากทำคำสั่งเสร็จสิ้นจะเก็บบันทึกไว้

รูปที่ 3.14 จะเป็นตารางสำหรับการเก็บ HTTP Request จาก NodeJS ที่ส่งไปทางฟรอนท์ของ Django ซึ่งจะถูกเก็บข้อมูลไว้ในตารางนี้

### 3.4.2 Database Dictionary

อธิบายถึงชื่อของคอลัมน์ ความหมายและลักษณะการเก็บข้อมูลภายใต้ฐานข้อมูลโดยที่ตารางมีทั้งหมด 18 ตารางดังนี้

ตารางที่ 3.1 ตารางอธิบายความหมายตาราง term\_word

ชื่อคอลัมน์	ความหมาย	ประเภท
term_word_id	id สำหรับบ่งบอกคำศัพท์	INT (10) PK Auto_Increment
term	คำศัพท์	VARCHAR (191)
frequency	จำนวนความถี่ของเอกสารที่มีคำศัพทน้อย	INT (191)
score_idf	คะแนน idf ของคำศัพท์	FLOAT (255,4)
rec_create_at	วันเวลาของการเพิ่มคำศัพทนี้เข้าสู่ระบบ	DATETIME (6) current_timestamp
rec_modified_at	วันเวลาที่อัปเดทข้อมูลของคำศัพท์	DATETIME (6) current_timestamp

ตารางที่ 3.2 ตารางอธิบายความหมายตาราง user

ชื่อคอลัมน์	ความหมาย	ประเภท
user_id	id สำหรับบ่งบอกผู้ใช้งาน	INT (10) PK Auto_Increment
name	ชื่อของผู้ใช้งาน	VARCHAR (50)
surname	นามสกุลของผู้ใช้งาน	VARCHAR (191)
role	ตำแหน่งของผู้ใช้งาน	VARCHAR (191)
username	ชื่อผู้ใช้งานสำหรับทำการ login	VARCHAR (191)
password	รหัสผ่านผู้ใช้งานสำหรับทำการ login	VARCHAR (191)
create_at	วันเวลาของผู้ใช้งานของการเพิ่มเข้าสู่ระบบ	DATETIME (6) current_timestamp
active	สถานะการรับบัญชีผู้ใช้งาน	INT (11) Default 1

ตารางที่ 3.3 ตารางอธิบายความหมายตาราง score

ชื่อคอลัมน์	ความหมาย	ประเภท
score_id	id สำหรับบ่งบอกคะแนนของคำศัพท์	INT (10) PK Auto_Increment
score_tf	คะแนน tf ของคำศัพท์	FLOAT (255,4)
score_tf_idf	คะแนน tf-idf ของคำศัพท์	FLOAT (255,4)
index_term_word_id	id สำหรับบ่งบอกคำศัพท์	INT (10)
index_document_id	id สำหรับบ่งบอกเอกสาร	INT (10)
generate_by	คะแนนถูกคำนวณโดยใคร	VARCHAR (191) Default 'default'
rec_status	สถานะการใช้คะแนนนี้	INT (191) Default 1

ตารางที่ 3.4 ตารางอธิบายความหมายตาราง pre\_term\_in\_page

ชื่อคอลัมน์	ความหมาย	ประเภท
pre_term_in_page_id	id สำหรับบ่งบอกคำศัพท์ช่วงคราวที่รอให้ผู้ใช้งานตรวจสอบ	INT (10) PK Auto_Increment
pre_term	คำศัพท์ช่วงคราวที่รอให้ผู้ใช้ตรวจสอบ	VARCHAR (191)
index_page_in_document_id	id สำหรับบ่งบอกที่อยู่ของคำศัพท์ช่วงคราวที่รอให้ผู้ใช้งานตรวจสอบ	INT (10) FK

ตารางที่ 3.5 ตารางอธิบายความหมายตาราง page\_in\_document

ชื่อคอลัมน์	ความหมาย	ประเภท
page_in_document_id	id สำหรับบ่งบอกที่อยู่ของคำศัพท์ช่วงคราวที่รอให้ผู้ใช้งานตรวจสอบ	INT (10) PK Auto_Increment
page_index	หน้าของเอกสาร	INT (191)
name	ชื่อ File ของข้อมูล	VARCHAR (191)
rec_status_confirm	สถานะการยืนยันโดยผู้ใช้งาน	INT (2) Default 2
index_document_id	id สำหรับบ่งบอกเอกสาร	INT (10) FK

ตารางที่ 3.6 ตารางอธิบายความหมายตาราง nodejs\_log

ชื่อคอลัมน์	ความหมาย	ประเภท
nodejs_log_id	id สำหรับการจัดเก็บประวัติการทำงานฝั่ง nodejs	INT (10) PK Auto_Increment
status_code	เก็บสถานะ HTTP หลังจากที่ส่งไปแล้วว่าได้สถานะใด	INT (191)
header_date	เก็บข้อมูล header ของ HTTP ที่ส่งไป	VARCHAR (191)
server	ชื่อรูปแบบของเซิฟเวอร์ที่ส่งไป	VARCHAR (191)
url	ตำแหน่งโดเมนหรือ IP ที่ส่งไป	INT (10) FK
content_type	รูปแบบเนื้อหาที่ส่งไป	VARCHAR (191)
rec_status	สถานะที่บอกร่วมกับการส่งเกิดข้อผิดพลาดระหว่างทาง	INT (191)
rec_create_date	วันเวลาที่ทำการส่ง ณ ตอนนั้น	DATETIME (6) current_timestamp

ตารางที่ 3.7 ตารางอธิบายความหมายตาราง knex\_migrations\_lock

ชื่อคอลัมน์	ความหมาย	ประเภท
index	บ่งบอกลำดับของไฟล์ migration ของ knex	INT (10) PK Auto_Increment
is_locked	สถานะของไฟล์ migration	INT (11)

ตารางที่ 3.8 ตารางอธิบายความหมายตาราง knex\_migrations

ชื่อคอลัมน์	ความหมาย	ประเภท
id	บ่งบอกลำดับการทำงานของไฟล์ migration ของ knex	INT (10) PK Auto_Increment
name	ชื่อไฟล์ migration ที่ถูกทำงานเรียบร้อย	VARCHAR (255)
batch	ลำดับที่	INT (11)
migration_time	เวลาที่ถูกสั่งให้ทำงาน	TIMESTAMP current_timestamp

ตารางที่ 3.9 ตารางอธิบายความหมายตาราง indexing\_publisher\_document

ชื่อคอลัมน์	ความหมาย	ประเภท
indexing_publisher_id	id สำหรับบ่งบอกสำนักพิมพ์	INT (10) PK Auto_Increment
publisher	ชื่อสำนักพิมพ์	VARCHAR (191)
publisher_email	e-mail ของสำนักพิมพ์	VARCHAR (191)
frequency	จำนวนของสำนักพิมพ์ที่ถูกอ้างอิง	INT (191)

ตารางที่ 3.10 ตารางอธิบายความหมายตาราง indexing\_issued\_date\_document

ชื่อคอลัมน์	ความหมาย	ประเภท
indexing_issued_date_id	id สำหรับบ่งบอกปีที่เขียน	INT (10) PK Auto_Increment
issued_date	วันเวลาของปีที่เขียนเอกสาร	DATE
frequency	จำนวนของวันเวลาของปีที่เขียนที่ถูกอ้างอิง	INT (191)

ตารางที่ 3.11 ตารางอธิบายความหมายตาราง indexing\_creator\_orgname\_document

ชื่อคอลัมน์	ความหมาย	ประเภท
indexing_creator_orgname_id	id สำหรับบ่งบอกชื่อหน่วยงานรับผิดชอบสังกัด	INT (10) PK Auto_Increment
creator_orgname	ชื่อหน่วยงานรับผิดชอบสังกัด	VARCHAR (191)
frequency	จำนวนของหน่วยงานรับผิดชอบสังกัดที่ถูกอ้างอิง	INT (191)

ตารางที่ 3.12 ตารางอธิบายความหมายตาราง indexing\_creator\_document

ชื่อคอลัมน์	ความหมาย	ประเภท
indexing_creator_id	id สำหรับบ่งบอกผู้เขียนเอกสาร	INT (10) PK Auto_Increment
creator	ชื่อของผู้เขียนเอกสาร	VARCHAR (191)
frequency	จำนวนของผู้เขียนเอกสารที่ถูกอ้างอิง	INT (191)

ตารางที่ 3.13 ตารางอธิบายความหมายตาราง indexing\_contributor\_document

ชื่อคอลัมน์	ความหมาย	ประเภท
indexing_contributor_id	id สำหรับบ่งบอกชื่อหน่วยข้อมูลผู้ร่วมงาน	INT (10) PK Auto_Increment
contributor	ชื่อหน่วยข้อมูลผู้ร่วมงาน	VARCHAR (191)
contributor_role	ตำแหน่งของหน่วยข้อมูลผู้ร่วมงาน	VARCHAR (191)
frequency	จำนวนของหน่วยข้อมูลผู้ร่วมงานที่ถูกอ้างอิง	INT (191)

ตารางที่ 3.14 ตารางอธิบายความหมายตาราง document

ชื่อคอลัมน์	ความหมาย	ประเภท
document_id	id สำหรับบ่งบอกเอกสาร	INT (10) PK Auto_Increment
status_process_document	สถานะการทำงานของเอกสาร	INT (2)
name	ชื่อไฟล์ PDF เอกสาร	VARCHAR (191)
version	ครั้งที่ติดพิมพ์	INT (255)
path	ตำแหน่งไฟล์ PDF ที่ผู้ใช้งานอัปโหลดเข้าสู่ระบบ	TEXT
DC_title	ชื่อเอกสาร	VARCHAR (191)
DC_title_alternative	ชื่อรองของเอกสาร	VARCHAR (191)
DC_description_table_of_contents	สารสารคัญที่มาจากการรับัญ	TEXT
DC_description_summary_or_abstract	บทสรุปสารสารคัญของหนังสือแต่ละเล่ม	TEXT
DC_description_note	รายละเอียดทั่วไปของเอกสาร	TEXT
DC_format	รูปแบบข้อมูลที่ถูกจัดเก็บในระบบ	VARCHAR (191)
DC_format_extent	ขนาดของไฟล์เอกสาร	VARCHAR (191)
DC_identifier_URL	แหล่งที่มาของเอกสาร	VARCHAR (191)
DC_identifier_ISBN	เลขมาตรฐานสากลของเอกสาร	VARCHAR (191)
DC_source	หน่วยข้อมูลต้นฉบับ	VARCHAR (191)
DC_language	ภาษาของเอกสาร	VARCHAR (191)
DC_coverage_spatial	สถานที่ของเอกสารที่เป็นเจ้าของ	VARCHAR (191)
DC_coverage_temporal	ช่วงเวลาในหน่วยปีของเอกสาร	VARCHAR (191)
DC_rights	ระดับการเข้าถึงของข้อมูล	VARCHAR (191)
DC_rights_access	ตำแหน่งที่มีสิทธิ์ในการเข้าถึงข้อมูล	VARCHAR (191)
thesis_degree_name	ชื่อเต็มของปริญญา	VARCHAR (191)
thesis_degree_level	ระดับของปริญญา	VARCHAR (191)
thesis_degree_discipline	สาขาวิชา	VARCHAR (191)
thesis_degree_grantor	มหาวิทยาลัย	VARCHAR (191)
rec_create_at	วันเวลาของเอกสารที่ถูกนำเข้าสู่ระบบ	DATETIME (6) current_timestamp
rec_create_by	id สำหรับบ่งบอกผู้ใช้งานที่นำเอกสารเข้าสู่ระบบ	INT (10) FK
rec_modified_at	วันเวลาของเอกสารที่ถูกแก้ไขข้อมูล	DATETIME (6) current_timestamp
rec_modified_by	id สำหรับบ่งบอกผู้ใช้งานที่แก้ไขเอกสารในระบบ	INT (10) FK
index_creator	id สำหรับบ่งบอกผู้เขียนเอกสาร	INT (10) FK
index_creator_orgname	id สำหรับบ่งบอกชื่อหน่วยงานรับผิดชอบสังกัด	INT (10) FK

ชื่อคอลัมน์	ความหมาย	ประเภท
index_publisher	id สำหรับบ่งบอกสำนักพิมพ์	INT (10) FK
index_contributor	id สำหรับบ่งบอกชื่อหน่วยข้อมูลผู้ร่วมงาน	INT (10) FK
index_issued_date	id สำหรับบ่งบอกปีที่เขียน	INT (10) FK

ตารางที่ 3.16 ตารางอธิบายความหมายตาราง django\_log

ชื่อคอลัมน์	ความหมาย	ประเภท
django_log_id	id สำหรับการจัดเก็บประวัติการทำงานฝั่ง django	INT (10) PK Auto_Increment
rec_status	สถานะการทำงานที่เกิดขึ้น	INT (191)
rec_create_date	วันเวลาของการทำงานที่เกิดขึ้น	DATETIME (6) current_timestamp
log_error	ข้อมูลข้อผิดพลาดที่เกิดขึ้น	VARCHAR (191)
index_document	id สำหรับบ่งบอกเอกสารที่ทำงาน	INT (10) FK

ตารางที่ 3.17 ตารางอธิบายความหมายตาราง dc\_type

ชื่อคอลัมน์	ความหมาย	ประเภท
DC_type_id	id สำหรับบ่งบอกประเภทของเอกสาร	INT (10) PK Auto_Increment
DC_type	ประเภทของเอกสาร	VARCHAR (191)
index_document_id	id สำหรับบ่งบอกเอกสาร	INT (10)

ตารางที่ 3.18 ตารางอธิบายความหมายตาราง dc\_relation

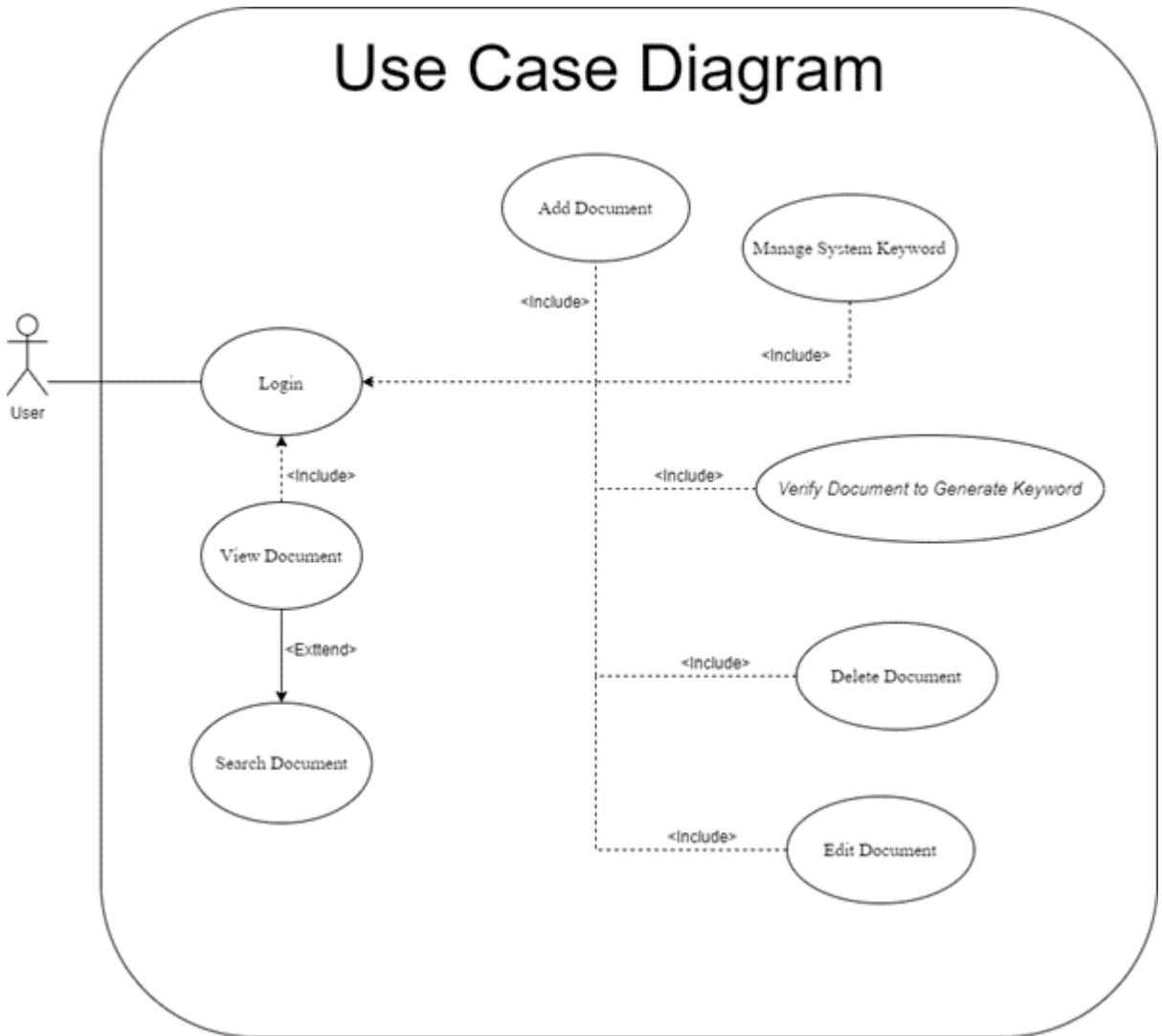
ชื่อคอลัมน์	ความหมาย	ประเภท
DC_relation_id	id สำหรับบ่งบอกเอกสารที่เกี่ยวข้อง	INT (10) PK Auto_Increment
DC_relation	ชื่อเอกสารที่เกี่ยวข้อง	VARCHAR (191)
index_document_id	id สำหรับบ่งบอกเอกสาร	INT (10)

ตารางที่ 3.19 ตารางอธิบายความหมายตาราง dc\_keyword

ชื่อคอลัมน์	ความหมาย	ประเภท
DC_keyword_id	id สำหรับบ่งบอก tag	INT (10) PK Auto_Increment
DC_keyword	คำศัพท์	VARCHAR (191)
index_document_id	id สำหรับบ่งบอกเอกสาร	INT (10)

### 3.5 UML Design

#### 3.5.1 Use case diagram



รูปที่ 3.15 Use case diagram

#### 3.5.2 Sequence diagram

##### 3.5.2.1 Use case Add Document

Scenario 1: เพิ่มหนังสือ/เอกสารเข้าสู่ระบบ

Goal: เพิ่มข้อมูลของเอกสารเข้าไปอยู่ในระบบ

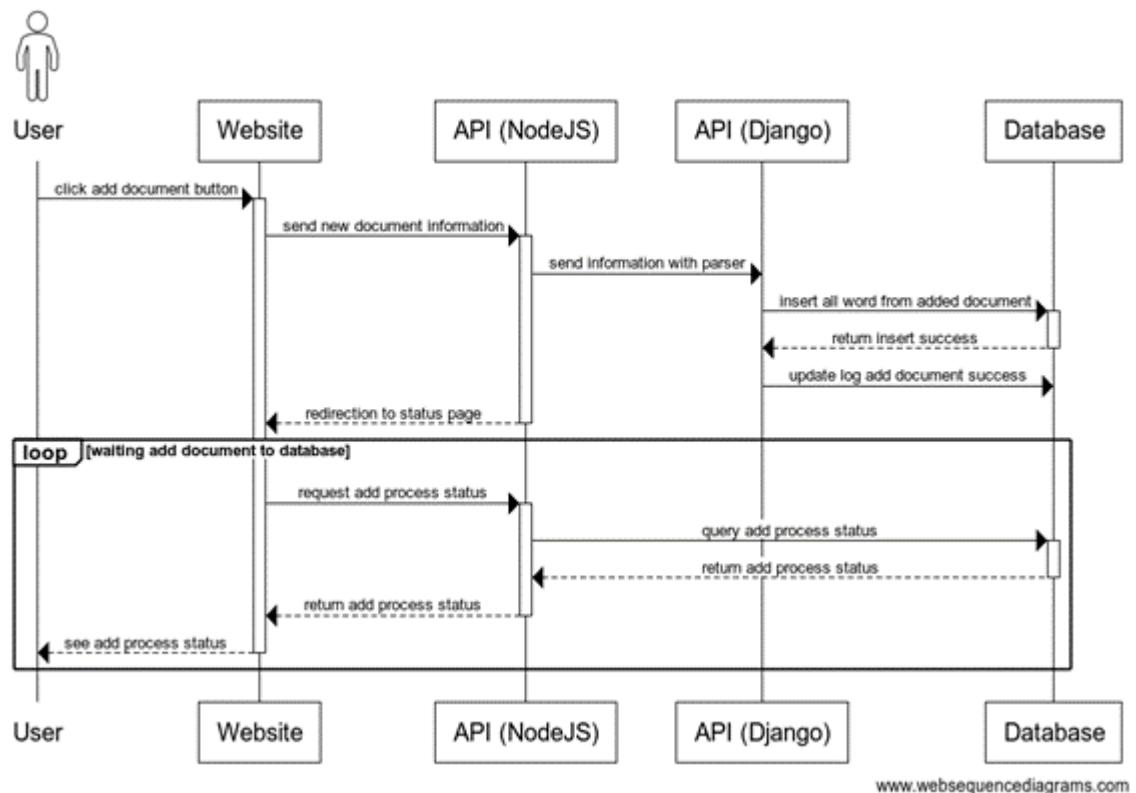
Precondition: กดไปที่หัวข้อ INSERT BOOK ใน Web Application

Main success scenario:

1. อัปโหลดเอกสาร/หนังสือเลือกหน้าที่จะให้เริ่มต้นการแปลง

2. กรอกข้อมูลรายละเอียดที่ต้องการลงในระบบ
3. แสดงสถานะของการเพิ่มข้อมูล
4. เพิ่มเอกสาร/หนังสือเข้าสู่ระบบ

### Use case Add Document



รูปที่ 3.16 แสดง Scenario 1 เพิ่มเอกสารเข้าระบบ

#### 3.5.2.2 Use case Manage word in document

Scenario 2: การตรวจสอบและแก้ไขคำก่อนนำเข้าสู่ระบบ

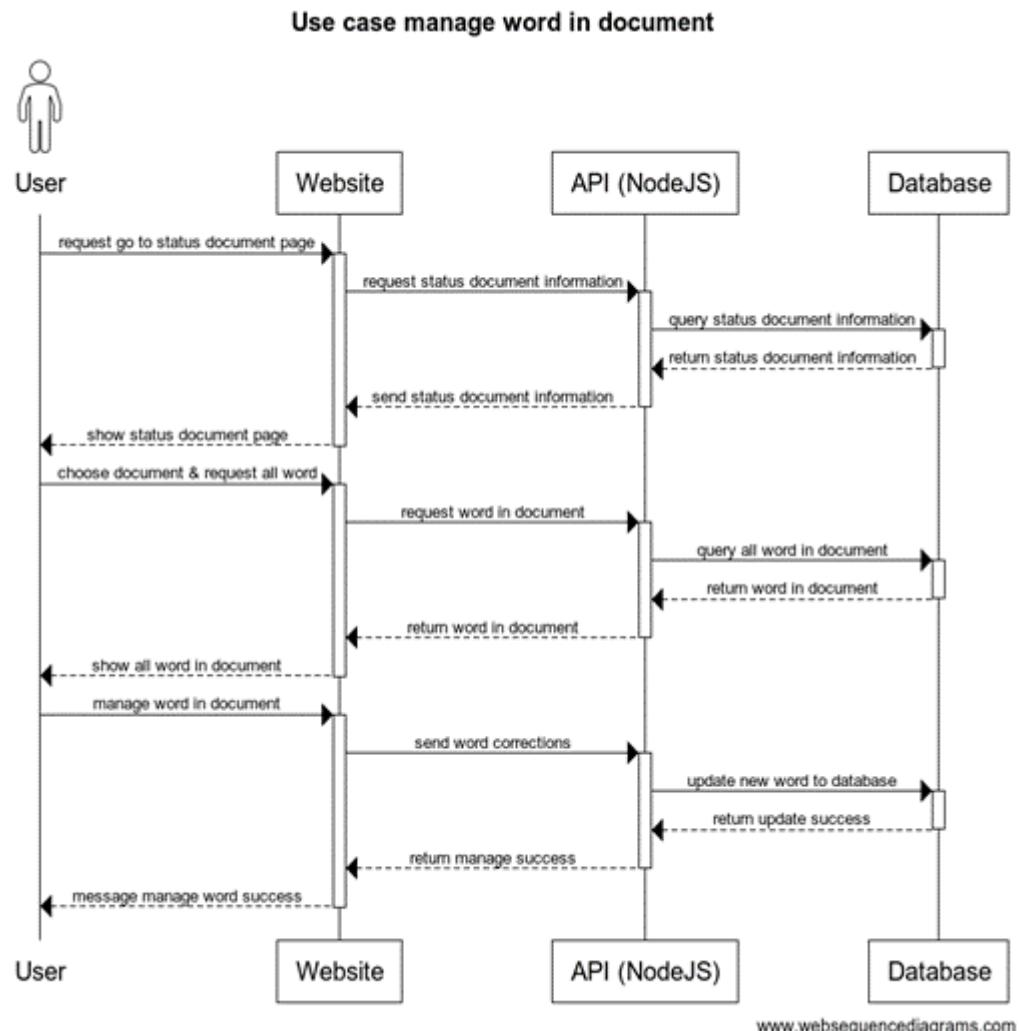
Goal: ผู้ใช้งานเห็นคำที่จะถูกแปลงเป็นดิจิตอลแล้วสามารถจัดการคำเหล่านี้ได้

Precondition: อยู่ภายในขั้นตอนการเพิ่มหนังสือ/เอกสารลงในระบบ

Main success scenario:

1. ผู้ใช้เข้าไปยังหน้าดูสถานะการเพิ่มเอกสาร
2. ผู้ใช้เลือกเอกสารที่อยู่ในสถานะตรวจสอบคำ
3. ระบบแสดงคำทั้งหมดที่ถูกแปลงมาได้จากเอกสารแต่ละหน้า
4. ผู้ใช้ตรวจสอบ แก้ไขคำที่แสดงขึ้นมา

5. ยืนยันขั้นตอนการตรวจสอบและแก้ไขคำ



รูปที่ 3.17 แสดง Scenario 2 การจัดการคำที่ถูกเก็บได้จากเอกสารในระบบ

### 3.5.2.3 Use case Verify Document to Generate Keyword

Scenario 3: ยืนเอกสารว่าพร้อมสำหรับการถูกนำไปสร้างคีย์เวิร์ด

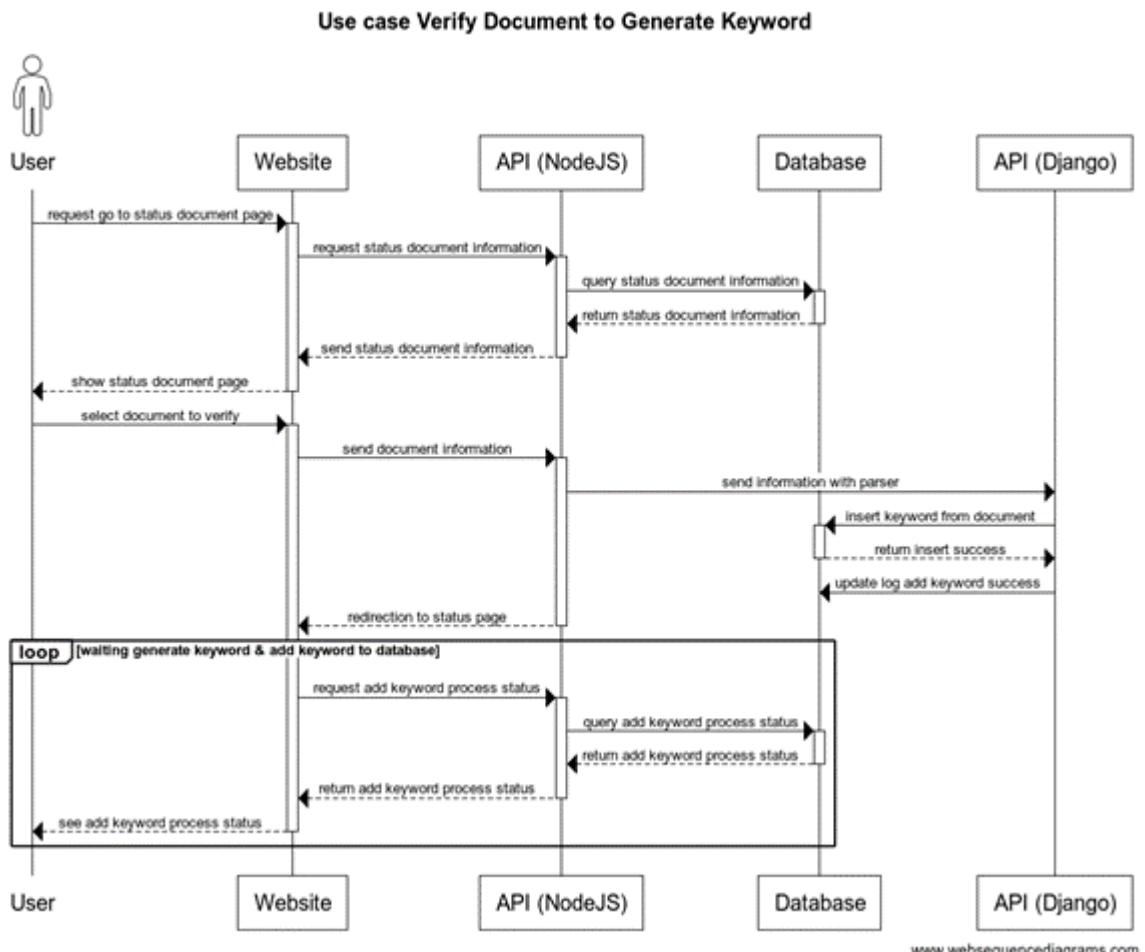
Goal: เอกสารถูกยืนยันพร้อมกับสร้างคีย์เวิร์ดเพื่อเพิ่มเข้าไปในระบบ

Precondition: ไปยังหน้าสถานะของเอกสารแล้วด้ไปยังปุ่มยืนเอกสารถูกต้อง

Main success scenario:

1. ผู้ใช้เข้าไปยังหน้าสถานะการเพิ่มเอกสาร
2. ระบบแสดงสถานะเอกสารว่าเอกสารไหนอยู่สถานะได้แล้วบ้าง
3. ผู้ใช้กดปุ่มยืนยันว่าเอกสารถูกต้อง

4. ระบบย้ายไปหน้าสถานะเอกสารอีกครั้งเพื่อรอผลการทำงาน
5. ระบบแสดงการยืนยันเอกสาร และถูกเพื่อคีย์เวิร์ดเสร็จสิ้น



รูปที่ 3.18 แสดง Scenario 3 ยืนเอกสารว่าพร้อมสำหรับการถูกนำไปสร้างคีย์เวิร์ด

#### 3.5.2.4 Use case Edit Document

Scenario 4: การแก้ไขรายละเอียดของเอกสาร/หนังสือที่อยู่ภายในระบบ

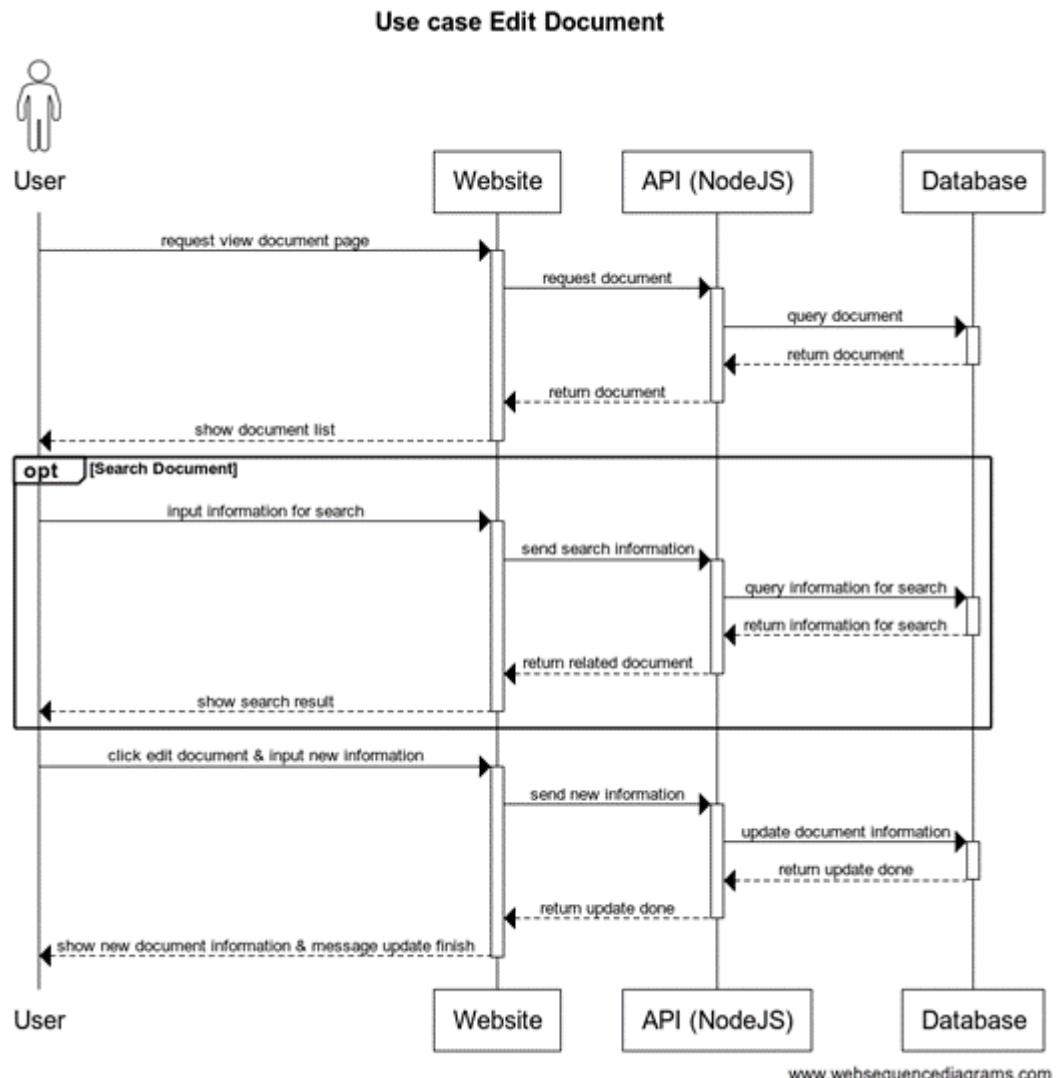
Goal: รายละเอียดเอกสารถูกแก้ไขตามผู้ใช้งานต้องการ

Precondition: กดไปที่หัวข้อ MANAGE BOOK ใน Web Application

Main success scenario:

1. ผู้ใช้ค้นหาเอกสารที่ต้องการแก้ไขรายละเอียด
2. แสดงผลลัพธ์ในการค้นหาเอกสาร/หนังสือ
3. เลือกเอกสาร/หนังสือที่ต้องการแก้ไขรายละเอียด

4. แก้ไขรายละเอียดที่ต้องการ
5. กดบันทึกข้อมูลลงในระบบ



รูปที่ 3.19 แสดง Scenario 4 แก้ไขข้อมูลเอกสาร

### 3.5.2.5 Use case Delete Document

Scenario 5: ลบเอกสาร/หนังสือภายในระบบ

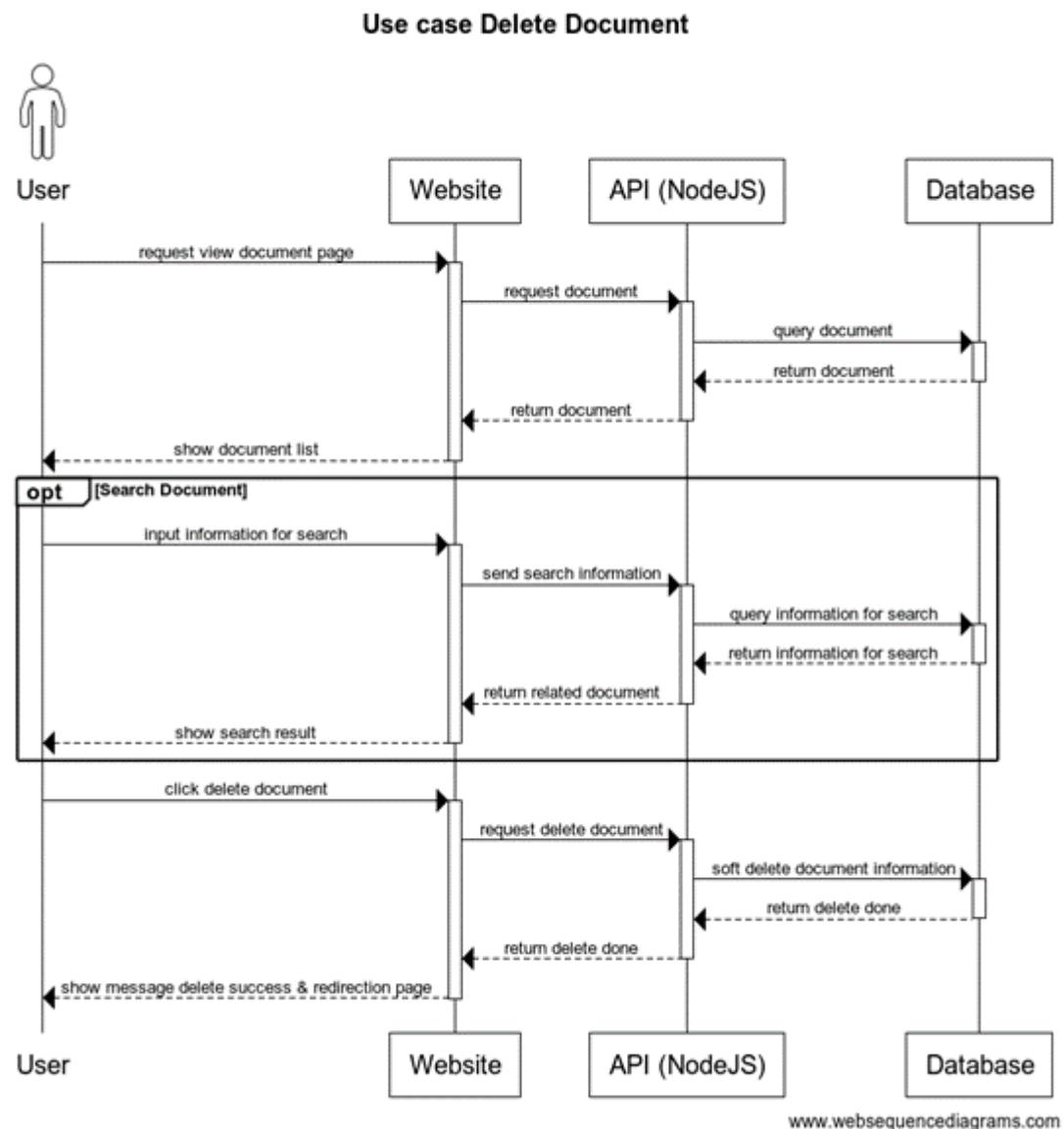
Goal: เอกสาร/หนังสือถูกนำออกจากระบบ

Precondition: กดเลือกทััวว้า MANAGE BOOK ใน Web Application

Main success scenario:

1. ผู้ใช้ทำการค้นหาเอกสารหนังสือที่ต้องการจะลบออกจากระบบ

2. แสดงผลลัพธ์ในการค้นหาเอกสาร/หนังสือ
3. กดลบเอกสาร/หนังสือที่ต้องการ
4. กดยืนยันคำสั่งลบเพื่อบันทึกลงระบบ



รูปที่ 3.20 แสดง Scenario 5 ลบเอกสาร

### 3.5.2.6 Use case View Document & Search Document

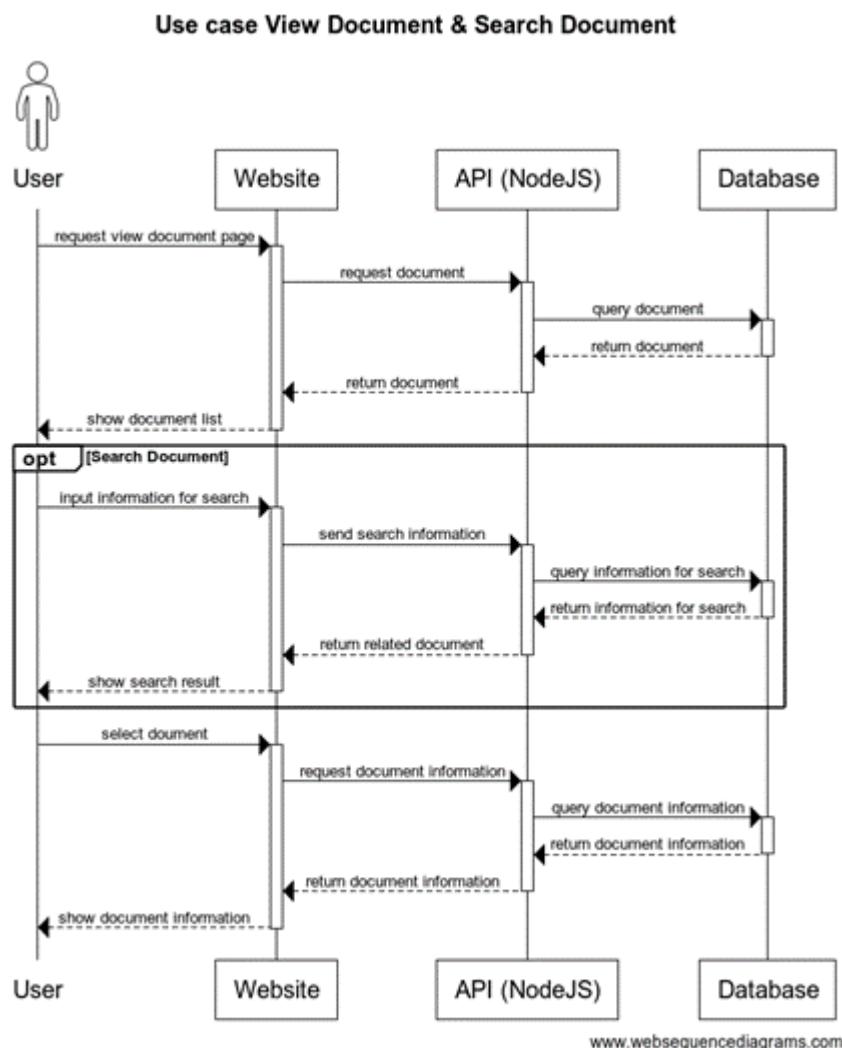
Scenario 6: ดูข้อมูลเอกสาร และการค้นหาเอกสาร

Goal: ผู้ใช้เจอเอกสารที่ต้องการ

Precondition: กดไปที่หัวข้อ SEARCH ใน Web Application

Main success scenario:

1. กรอกรายละเอียดข้อมูลที่ต้องการจะค้นหา
2. แสดงผลลัพธ์ในการค้นหา
3. ผู้ใช้เลือกเอกสารที่ต้องการที่จะดูข้อมูล
4. ระบบย้ายไปยังหน้าแสดงข้อมูลเอกสารที่ถูกเลือก



รูปที่ 3.21 แสดง Scenario 6 ดูข้อมูลเอกสาร และการค้นหาเอกสาร

### 3.5.2.7 Use case Login

Scenario 7: ระบบล็อกอิน

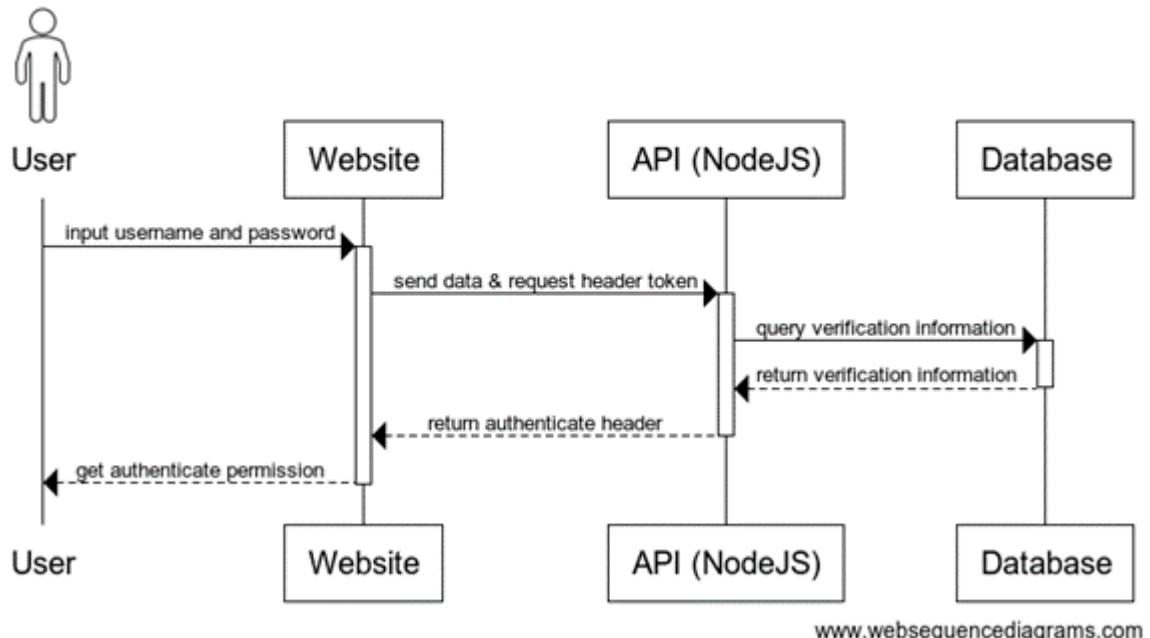
Goal: เพื่อเข้าสู่ระบบให้สามารถใช้ฟังก์ชันภายใน Web Application เพื่อเติมได้

Precondition: กดหัวข้อ LOGIN ใน Web Application

Main success scenario:

1. ผู้ใช้กรอกชื่อผู้ใช้งานและรหัสผ่าน
2. กดเข้าสู่ระบบ
3. เข้าสู่ระบบสำเร็จ ส่งผู้ใช้กลับไปสู่ Homepage
4. สามารถเข้าใช้งานฟังก์ชันของ Web Application ได้

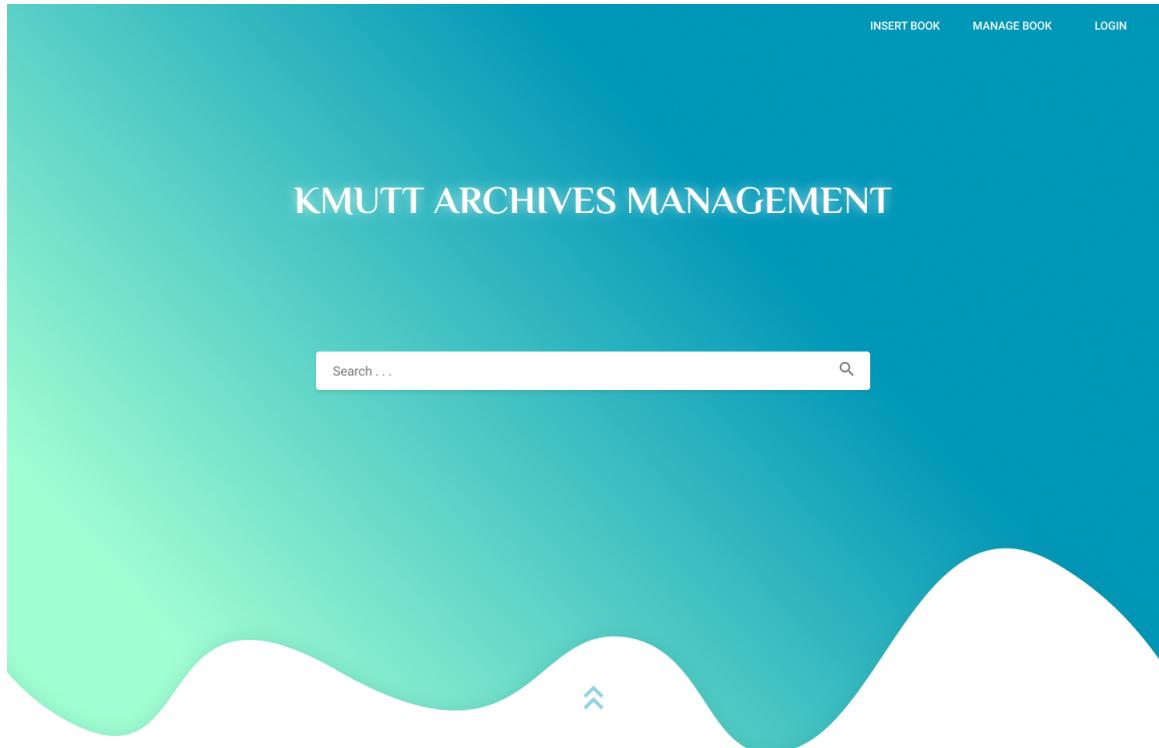
### Use case Login



รูปที่ 3.22 แสดง Scenario 7 ระบบล็อกอิน

### 3.6 GUI Design

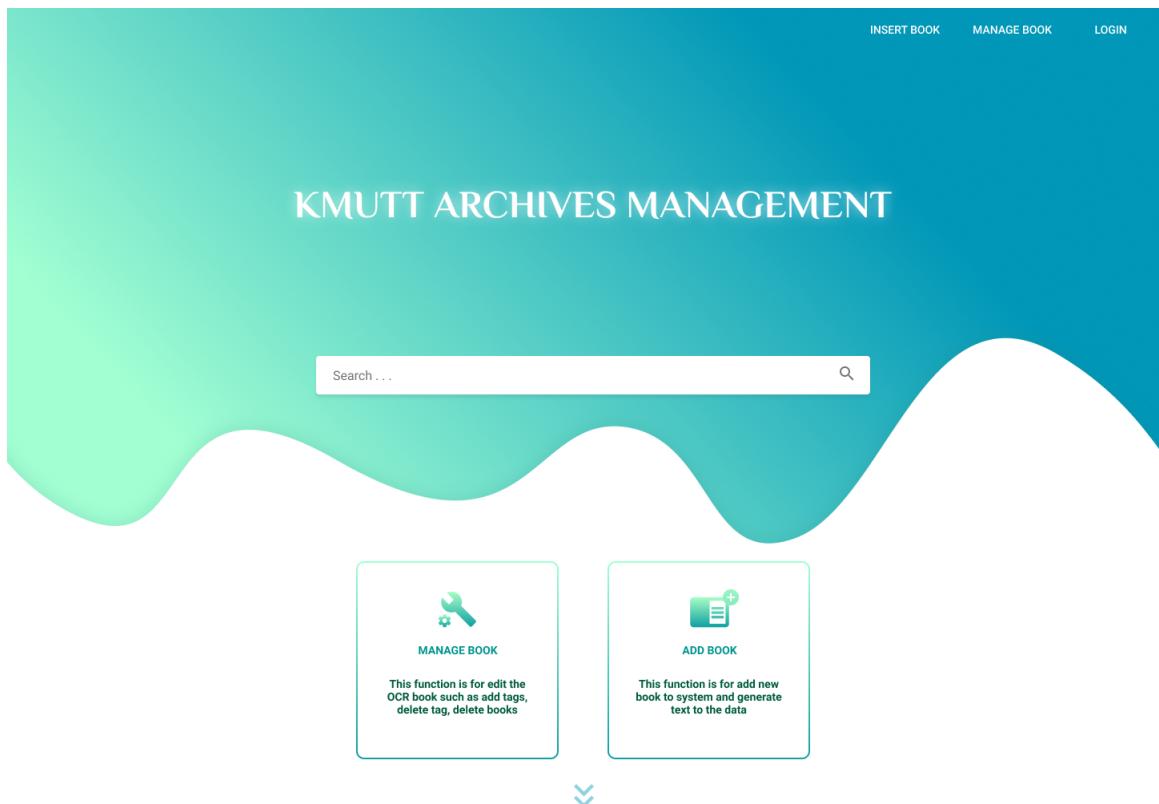
#### 3.6.1 Homepage



รูปที่ 3.23 ภาพแสดงหน้าหลักของเว็บไซต์

หน้าหลักของเว็บไซต์จะเป็นหน้าที่เน้นการค้นหาเป็นหลัก ที่ผู้ใช้สามารถเข้าถึงเมนูการเพิ่มหนังสือ การจัดการ และการเข้าสู่ระบบได้ที่แถบ Navigation ด้านบนของเว็บไซต์ดังรูปที่ 3.23

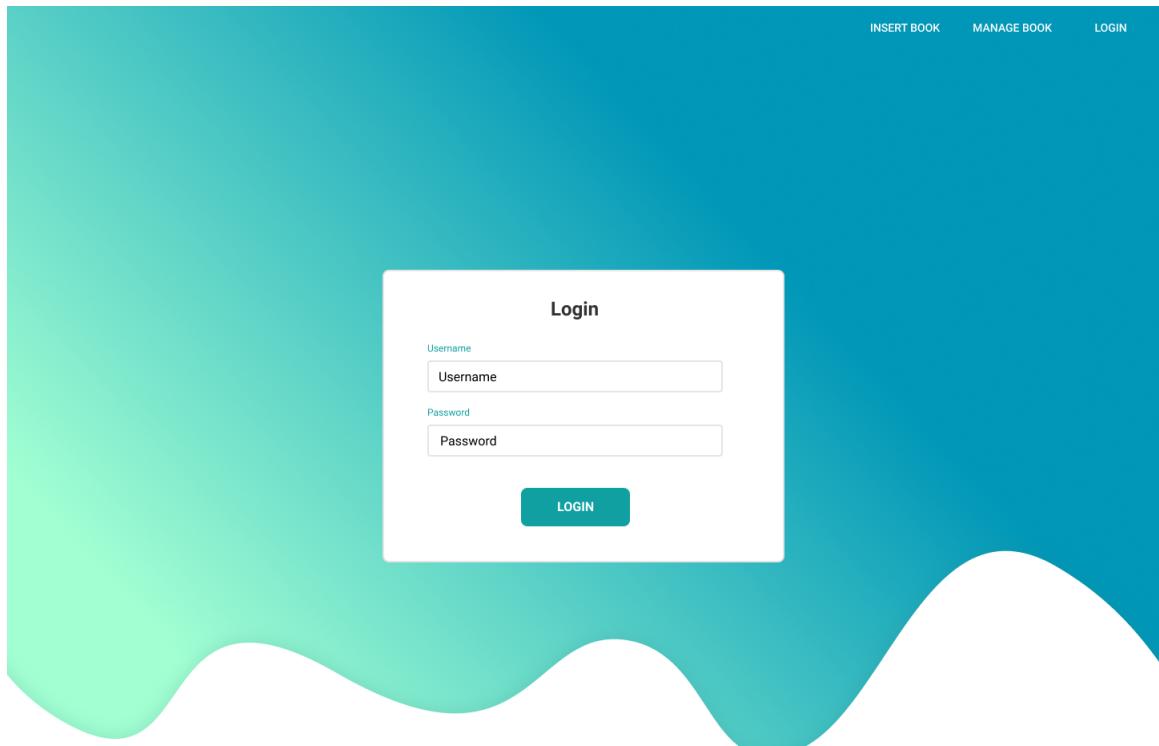
### 3.6.2 Homepage2



รูปที่ 3.24 ภาพแสดงหน้าหลักของเว็บไซต์หลังจากการกดเปิดเมนู

เมื่อกดปุ่มลูกศรที่ด้านล่างของรูป 3.23 จะมีเมนูเพิ่มเติมขึ้นมากลายเป็นรูปที่ 3.24 ซึ่งจะแสดงรายละเอียดในแต่ละฟังก์ชันเพิ่มเติม

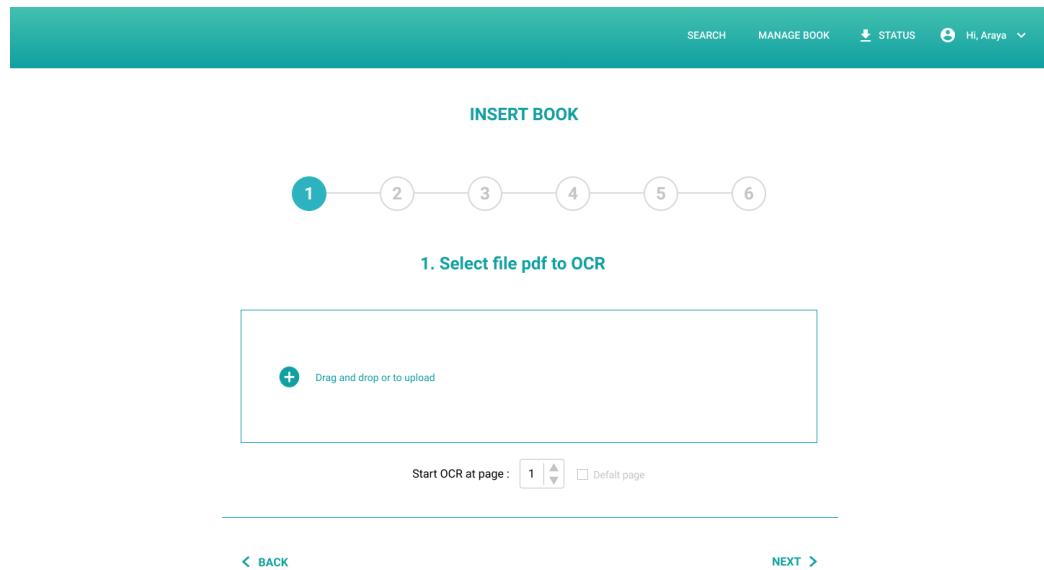
### 3.6.3 Login



รูปที่ 3.25 ภาพแสดงหน้าเข้าสู่ระบบ

ก่อนที่จะทำการเพิ่มหนังสือหรือจัดการกับหนังสือผู้ใช้นั้นจะต้องเข้าสู่ระบบก่อนเสมอ ถ้าเกิดกดเข้าฟังก์ชันการเพิ่มหนังสือหรือค้นหาโดยที่ยังไม่ได้เข้าสู่ระบบ ระบบจะบังคับให้ผู้ใช้เข้ามาในหน้าเข้าสู่ระบบดังรูป 3.25 เพื่อทำการเข้าสู่ระบบหรือจะเข้ามาโดยการกด log in ที่ปุ่มขวาบนได้

### 3.6.4 Insert Book(1)



รูปที่ 3.26 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าสู่ระบบขั้นเลือกไฟล์

หน้าเพิ่มหนังสือขึ้นแรกจะเป็นการเลือกไฟล์เอกสารที่ต้องการโดยที่จะมีส่วนของการเพิ่มไฟล์ที่อยู่รูปของ pdf เพื่อทำ OCR จากนั้นจะสามารถเลือกได้ว่าจะทำการ OCR ตั้งแต่หน้าไหนดังรูปที่ 3.26

### 3.6.5 Insert Book (2)

**INSERT BOOK**

1 2 3 4 5 6

**2. Fill the data**

**Title**

Title \*

Title Alternative

**Creator**

Creator name

Creator Organization name

**Description**

Table of contents

Summary

Abstract

Note

**Publisher**

Publisher

Publisher E-mail

**Contributor**

Contributor

Contributor Role

**Date**

Issued date

**Coverage**

Coverage Spatial

Coverage Temporal

**Rights**

Rights

Rights Access



---

◀ BACK

NEXT ▶

หน้าเพิ่มหนังสือขั้นตอนที่ 2 เป็นหน้าที่ต้องใส่ข้อมูลที่จำเป็นของหนังสือ โดยที่จำเป็นต้องใส่จะมีสัญลักษณ์กำกับไว้หรือคือชื่อหนังสือดังรูป 3.27 โดยในหน้านี้จะมีกล่องใส่ข้อมูลที่ถูกกรอกบ่อย ๆ สำหรับผู้ใช้(เจ้าหน้าที่)

### 3.6.6 Insert Book (3)

**INSERT BOOK**

SEARCH MANAGE BOOK STATUS Hi, Araya ▾

1 2 3 4 5 6

**3. Optional data**

**Identifier**

Identifier URL  
Input

Identifier ISBN  
Input

**Source**

Source  
Input

**Relation**

Relation + ADD  
Input

**Thesis**

Degree name  
Input

Degree level  
Input

Degree discipline  
Input

Degree grantor  
Input

**Type**

Type  
Text

**Language**

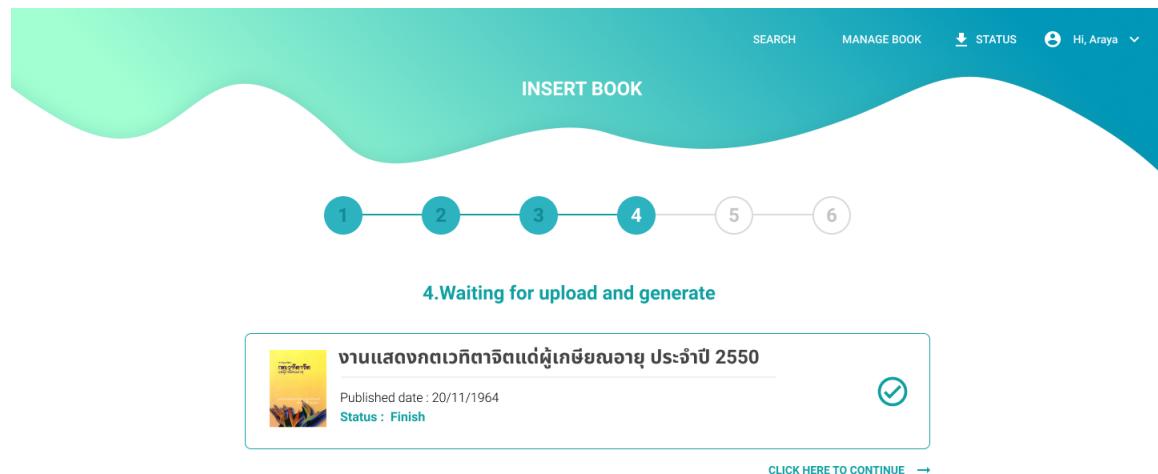
Language  
Thai

< BACK NEXT >

รูปที่ 3.28 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าสู่ระบบขั้นกรอกข้อมูลขั้นที่ 2

ในขั้นตอนที่ 3 จากรูปที่ 3.28 จะเป็นหน้าที่ใส่ข้อมูลที่ส่วนใหญ่ผู้ใช้จะไม่ค่อยกรอกมากนัก ซึ่งไม่มีกล่องข้อมูลให้กรอกเพื่อใช้สามารถข้ามไปขั้นตอนถัดไปได้เลย

### 3.6.7 Insert Book (4)



รูปที่ 3.29 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าขั้นโหลดข้อมูลเข้าสู่ระบบ

หลังจากที่ทำการใส่ข้อมูลอุปกรณ์ทั้งหมดแล้วมาถึงหน้าที่เป็นหน้าโหลดข้อมูลดังรูป 3.29 ที่ระบบจะทำการ OCR และทำการเตรียมชุดข้อมูลที่ได้จากการ OCR โดยการนำคำมาตัดและเช็คคำพิเศษ เมื่อโหลดข้อมูลเสร็จแล้วระบบจะทำการเปลี่ยนสถานะการโหลดและขึ้นล็อปเพื่อเข้าสู่ขั้นตอนถัดไปได้

### 3.6.8 Insert Book (5)



รูปที่ 3.30 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าสู่ระบบขั้นแก้ไขคำผิด

หลังจากโหลดและเตรียมข้อมูลเรียบร้อยแล้ว ระบบจะทำการแสดงข้อมูลที่ถูกแปลงมาโดยที่ผู้ใช้สามารถแก้ไขคำได้ดังรูป 3.30 หรือสามารถข้ามໄเด้เลย เช่น กัน โดยเมื่อคลิกไปที่กล่องข้อความจะขึ้นให้แก้ແร์แต่ละคำและเมื่อเปลี่ยนหน้าจะทำการเก็บข้อมูลที่เปลี่ยนไว้ และจะบันทึกการแก้ไขข้อมูลทั้งหมดที่แก้ไขเมื่อข้ามไปขั้นตอนถัดไป

### 3.6.9 Insert Book (6)

The screenshot shows a software interface titled 'INSERT BOOK' with a teal header bar. In the top right corner, there are buttons for 'SEARCH', 'MANAGE BOOK', 'STATUS', and a user profile 'Hi, Araya'. Below the header, a horizontal progress bar consists of six numbered circles (1 to 6) connected by lines. The circle at position 6 is highlighted in teal and labeled '6. Edit tag'. Below this, there is a book cover thumbnail for 'งานแสดงออกเต็มจิตแด่ผู้เกียรติยศ ประจำปี 2550'. Underneath the thumbnail, the text 'Tag / Keyword' is displayed, followed by three buttons: 'ก朵บกิตา X', 'เกี่ยน X', and 'ประลักษ์ X'. Below these buttons is a search input field with the placeholder 'Input' and a '+ ADD' button. At the bottom left is a 'BACK' button with a left arrow icon, and at the bottom right is a teal 'Finish' button with a checkmark icon.

รูปที่ 3.31 ภาพแสดงขั้นตอนการเพิ่มหนังสือเข้าสู่ระบบขึ้นแก้ไขและเพิ่มคำสำคัญ

หน้าสุดท้ายของการเพิ่มหนังสือจะเป็นหน้าที่ให้ผู้ใช้สามารถจัดการกับ Keyword ได้ดังรูปที่ 3.31 โดยเมื่อผู้ใช้ต้องการใส่คำสำคัญเพิ่มสามารถกด ADD เพื่อเพิ่มคำที่ต้องการใส่ได้ และสามารถลบเมื่อคลิกที่ปุ่มกากรากที่คำสำคัญที่ระบบทำกรสร้างมาให้ เมื่อแก้ไขเสร็จแล้วสามารถกดปุ่ม Finish เพื่อทำการบันทึกข้อมูล

### 3.6.10 Search

The screenshot shows a library search interface with a teal header bar. On the right side of the header are buttons for 'EDIT BOOK', 'MANAGE BOOK', 'STATUS' (with a download icon), and a user profile 'Hi, Araya'. Below the header is a search bar with placeholder text 'Search ...' and a magnifying glass icon. To the right of the search bar is a 'Filter' section with several checkboxes and a 'Creator' input field containing 'Creator'. Other filter options include 'Creator Organization Name', 'Contributor', 'Contributor Role', 'Issued Date', 'Publisher', and 'Publisher Email'. A blue 'APPLY' button is located at the bottom right of the filter section. The main content area displays four search results, each consisting of a thumbnail image, a title, and a brief description. The titles are all identical: 'งานแสดงกตเวกิตาจิตแด่ผู้เกียรติยศ ประจำปี 2550'. Each result includes the creator 'Joe, Bryan', coverage temporal '1998', and tags '(กตเวกิตา) (เกียรติยศ) (1964)'. The thumbnails show a yellow book cover with a blue floral design.

Search results : KMUTT

**งานแสดงกตเวกิตาจิตแด่ผู้เกียรติยศ ประจำปี 2550**

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : (กตเวกิตา) (เกียรติยศ) (1964)

**งานแสดงกตเวกิตาจิตแด่ผู้เกียรติยศ ประจำปี 2550**

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : (กตเวกิตา) (เกียรติยศ) (1964)

**งานแสดงกตเวกิตาจิตแด่ผู้เกียรติยศ ประจำปี 2550**

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : (กตเวกิตา) (เกียรติยศ) (1964)

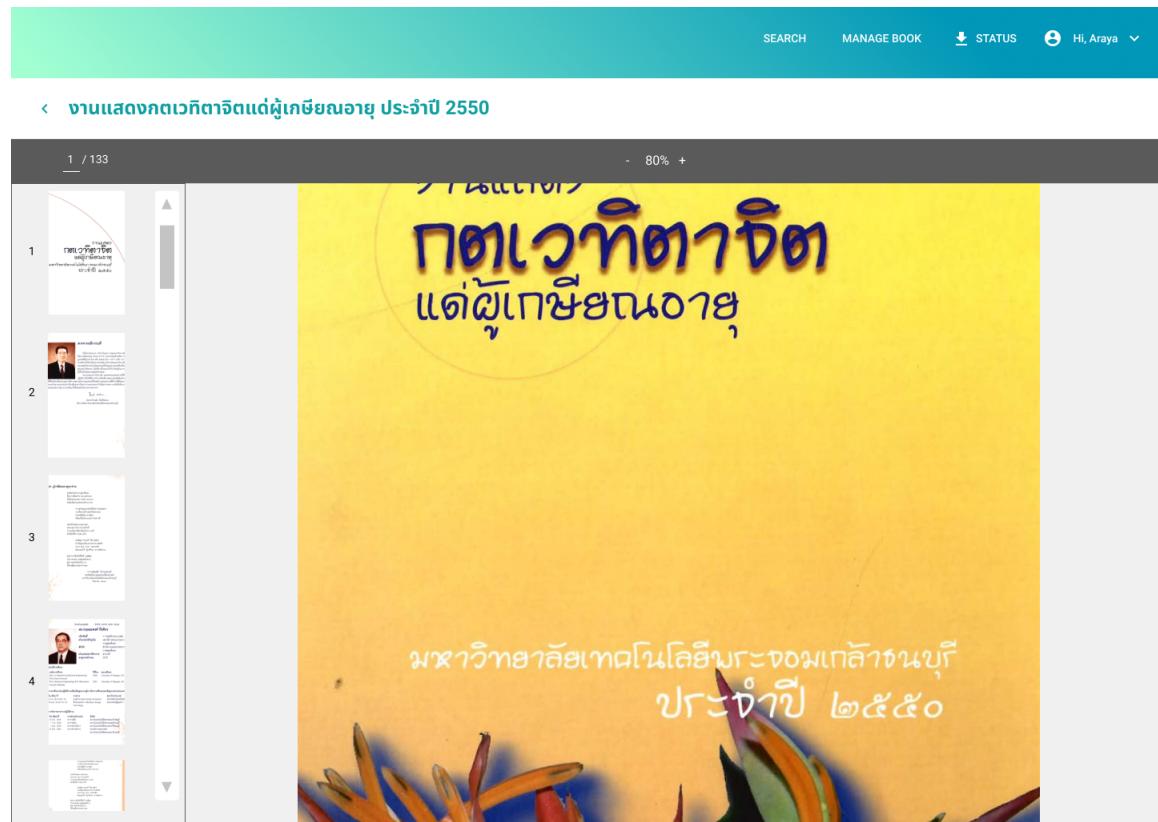
**งานแสดงกตเวกิตาจิตแด่ผู้เกียรติยศ ประจำปี 2550**

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : (กตเวกิตา) (เกียรติยศ) (1964)

รูปที่ 3.32 ภาพแสดงหน้าค้นหาข้อมูล

หน้าแสดงข้อมูลการค้นหาเมื่อทำการค้นหาข้อมูลจากหน้าแรก (รูปที่ 3.23 หรือ 3.24) จะทำการแสดงข้อมูลหนังสือที่ตรงกับ keyword โดยเรียงคะแนนของหนังสือที่เกี่ยวข้องกับคำค้นหามากที่สุดดังรูปที่ 3.32 เมื่อกดเข้าไปที่รายชื่อหนังสือจะทำการนำทางผู้ใช้ไปยังหน้าดูหนังสือ ดังรูปที่ 3.33

### 3.6.11 Document View



รูปที่ 3.33 ภาพแสดงหน้าจอหนังสือ

เมื่อเราค้นหาและเลือกหนังสือ ก็จะมีหน้าหนังสือ (รูปที่ 3.33) ขึ้นมาให้ดูนี้อหงายโนโดยที่ผู้ใช้สามารถปรับขนาดภาพและสามารถเลือกหน้าที่ต้องการจะเปิดได้และสามารถย้อนหลับไปยังหน้าเดิมได้ที่ปุ่มลูกครรภ์ทางด้านซ้ายบน

### 3.6.12 Manage book

Search results : KMUTT

**งานแสดงกตเวกิตาอิตเด่อผู้เกียญนอยุ ประจำปี 2550**

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : กตเวกิตา, เกียญ, 1994

**งานแสดงกตเวกิตาอิตเด่อผู้เกียญนอยุ ประจำปี 2550**

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : กตเวกิตา, เกียญ, 1994

**งานแสดงกตเวกิตาอิตเด่อผู้เกียญนอยุ ประจำปี 2550**

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : กตเวกิตา, เกียญ, 1994

**งานแสดงกตเวกิตาอิตเด่อผู้เกียญนอยุ ประจำปี 2550**

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : กตเวกิตา, เกียญ, 1994

รูปที่ 3.34 ภาพแสดงหน้าการจัดการหนังสือที่เพิ่มเข้าสู่ระบบ

ในหน้าของการจัดการหนังสือดังรูปที่ 3.34 จะมีลักษณะคล้ายกับหน้าการค้นหาเพียงแต่ว่าจะมีฟังก์ชันสำหรับการแก้ไขเนื้อหนังสือภายในที่ผู้ใช้เคยกรอกไว้ตอน OCR หนังสือมา เมื่อกดปุ่มลบจะมีหน้าต่างแจ้งเตือนเพื่อถามความแนใจในการลบเอกสาร หรือกดปุ่ม Edit เพื่อทำการเข้าสู่การแก้ไขข้อมูลของเอกสารนั้นๆดังรูปที่ 3.35 - 3.37

### 3.6.13 Edit Book

**SEARCH** **MANAGE BOOK** **STATUS** **Hi, Araya ▾**

**INSERT BOOK**



หน้าแสดงผลเว็บไซต์เดียวเกี่ยวน้ำยา ประจำปี 2550

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : กอบกิจ เมือง 1964

**Title**

Title \*

Title Alternative

**Creator**

Creator name

Creator Organization name

**Description**

Table of contents

Summary

Abstract

Note

**Publisher**

Publisher

Publisher E-mail

**Contributor**

Contributor

Contributor Role

**Date**

Issued date

**Coverage**

Coverage Spatial

Coverage Temporal

**Rights**

Rights

Rights Access

**INSERT BOOK**

---



**งานแสดงกตเวกิตาอิตแล่ญูกเซยณวาตุ ประจำปี 2550**

Creator : Joe, Bryan  
Coverage temporal : 1998  
Tag : กงบะกิ หนังสือ 1964

---

**Identifier**

Identifier URL

Identifier ISBN

**Source**

Source

**Relation**

Relation + ADD

**Thesis**

Degree name

Degree level

Degree discipline

Degree grantor

**Type**

Type Text

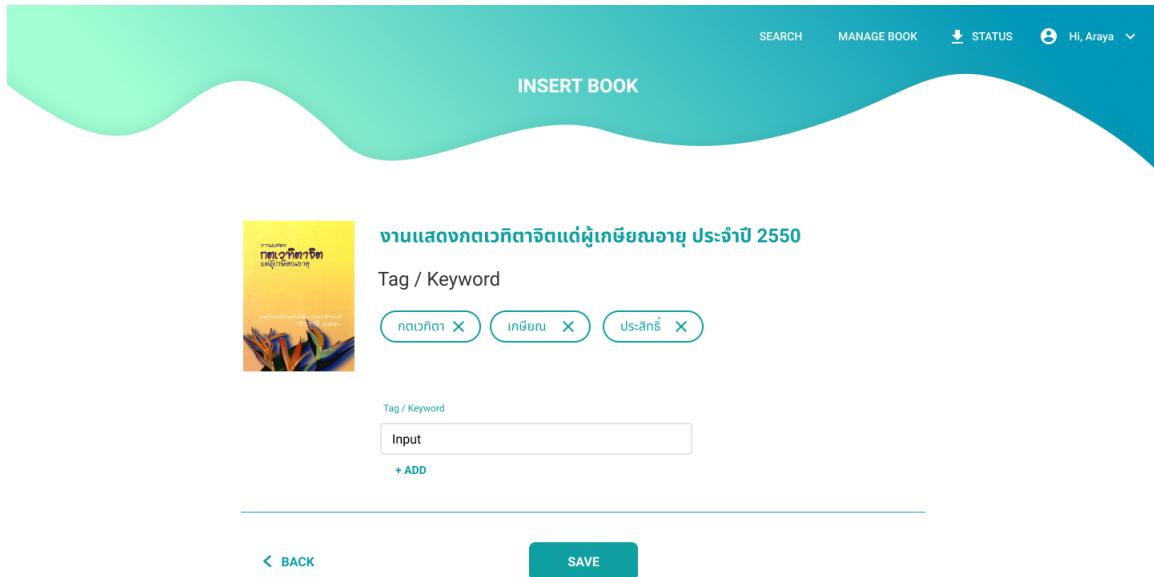
**Language**

Language Thai

---

[\*\*< BACK\*\*](#)
**SAVE**
[\*\*NEXT >\*\*](#)

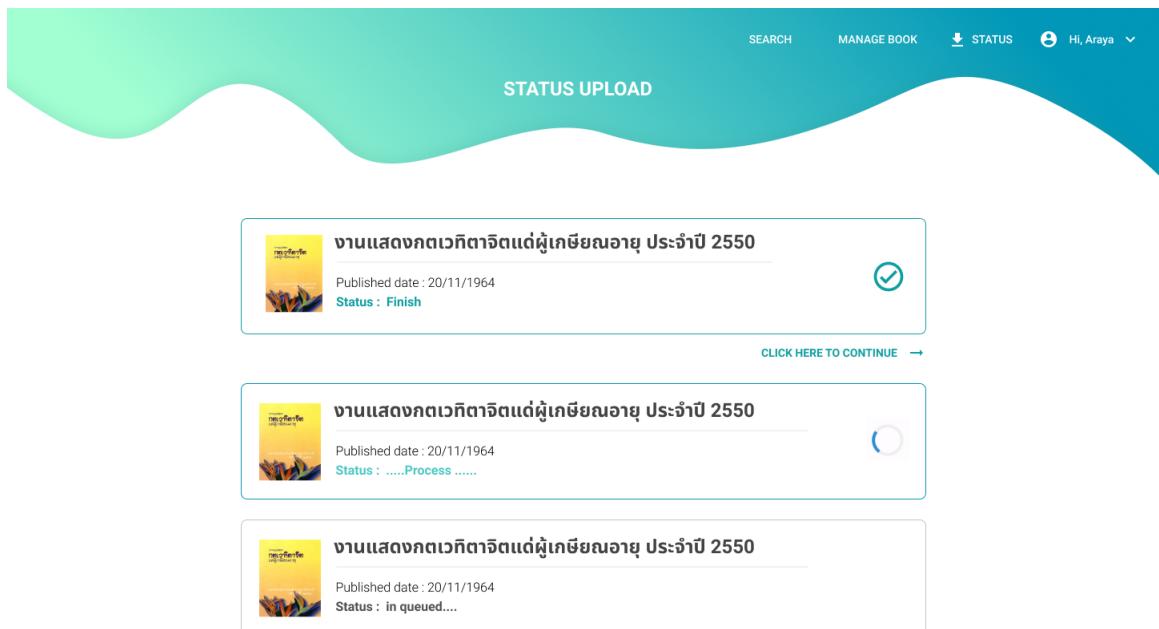
รูปที่ 3.36 ภาพแสดงขั้นตอนการแก้ไขหนังสือขั้นที่ 2



รูปที่ 3.37 ภาพแสดงขั้นตอนการแก้ไขหนังสือขั้นที่ 3

หน้าแก้ไขหนังสือแบบอ打เป็น 3 ขั้นตอนดังรูป 3.35 - 3.37 ซึ่งจะมีให้แก้ไข ข้อมูลที่เคยกรอกไว้ตอนเพิ่มหนังสือเข้ามา โดยจะมีรูปปักหนังสือและข้อหนังสืออยู่กว่ากำลังแก้ไขหนังสือเล่มใหม่อยู่ และในทุกหน้าจะมีปุ่มสำหรับบันทึกในทุกหน้าเพื่อที่จะสามารถบันทึกโดยที่ไม่ต้องรอไปหน้าสุดท้ายเพื่อบันทึกข้อมูล

### 3.6.14 Upload Status Page



รูปที่ 3.38 ภาพแสดงหน้าการโหลดข้อมูล

จากรูป 3.38 สำหรับผู้ใช้ที่ทำการเพิ่มเอกสารเข้าสู่ระบบจะมีหน้าสำหรับโหลดกรณีที่กดออกมากลังจากผ่านขั้นตอนการเพิ่มหนังสือขั้นตอนที่ 4 จะสามารถเข้ามาดูสถานะและทำการดำเนินการต่อได้โดยไม่ต้องผ่านการเพิ่มหนังสือเข้าสู่ระบบใหม่

### 3.6.15 Evaluate Process Design

ในส่วนของการประเมินผลการทำงานนั้นจะแบ่งออกเป็น 2 ส่วนคือ ส่วนของการทำ image processing จะช่วยให้การทำ OCR มีประสิทธิภาพมากเท่าไร และส่วนของระบบการค้นหา โดยในส่วนของ OCR จะทำการประเมินจากการเลือกเช็คคำจาก 2 หน้าของแต่ละเอกสารมาเช็คว่าแต่ละหน้ามีคำพิเศษเท่าไร โดยจะเลือกวัดเอกสารทั้งหมด 5 เล่มแบบสุ่มและเทียบการทำ Image processing ว่าทำแบบไหนได้ผลลัพธ์ดีที่สุด

ตารางที่ 3.20 ตารางประเมินการทำ OCR

ตารางประเมินการทำ OCR				
หนังสือ	หน้า	จำนวนคำทั้งหมด	คำที่ผิด(%)	คำเกิน(คำ)

ระบบการค้นหา จะเช็คโดยให้ผู้ใช้เป็นผู้ประเมินว่าได้รับเอกสารตรงตามที่ต้องการหรือไม่โดยจะให้เจ้าหน้าที่บรรณาธิการคัดเลือกหนังสือจำนวน 3 เล่มที่คาดหวังว่าจะขึ้นมาเมื่อค้นหาทั้งหมด 10 ครั้ง

ตารางที่ 3.21 ตารางประเมินระบบการค้นหา

ตารางประเมินระบบการค้นหา		
คำค้นหา	หนังสือที่คาดหวัง	การค้นหา
		<p>คะแนน 5 ระดับ</p> <p>5 = ค้นหาหนังสือได้ตรงตามที่ต้องการ และมีหนังสือที่เกี่ยวข้องกับคำค้นหาขึ้นมาอย่างถูกต้อง</p> <p>4 = ค้นหาหนังสือได้ถูกต้องตามที่ต้องการ บางเล่มและมีหนังสือที่เกี่ยวข้องกับคำค้นหาขึ้นมา</p> <p>3 = 'ไม่สามารถค้นหาหนังสือที่ต้องการแต่' มีหนังสือที่เกี่ยวข้องกับคำค้นหาขึ้นมา</p> <p>2 = สามารถค้นหาหนังสือที่มีความเกี่ยวข้องกับคำค้นหา และมีหนังสือที่ไม่เกี่ยวข้องกับการค้นหาแสดงในผลลัพธ์</p> <p>1 = 'ไม่มีหนังสือที่เกี่ยวข้องขึ้นมาในผลลัพธ์'</p>

ตารางที่ 3.22 ตารางประเมิน Design

ตารางประเมิน Design				
เกณฑ์การประเมิน	ผลลัพธ์		หมายเหตุ	คำเกิน(คำ)
	ผ่าน	ไม่ผ่าน		
1.หน้าเข้าสู่ระบบ				
2.Insert Book				
3.Search				
4.Manage Book				
5.View Book				
6.Home page				
7.Status Page				

ตารางที่ 3.23 ตารางประเมิน test

เกณฑ์การประเมิน	ผลลัพธ์		หมายเหตุ
	ผ่าน	ไม่ผ่าน	
1. สามารถเข้าสู่ระบบและออกจาก ระบบได้			
2. สามารถเพิ่มเอกสารเข้าสู่ระบบได้			
3. สามารถแก้ไขรายละเอียดเอกสารที่อยู่ในระบบได้			
4. สามารถตรวจสอบและแก้ไขคำที่เพิ่มเข้ามาในระบบในขั้นตอนเพิ่มเอกสารได้			
5. สามารถลบเอกสารที่อยู่ในระบบได้			
6. สามารถค้นหาข้อมูลเอกสารภายในระบบได้			
7. สามารถเรียกดูเอกสารที่ต้องการได้			

## บรรณานุกรม

1. Doxygen, 2020, “OpenCV,” [https://docs.opencv.org/3.4/d4/d73/tutorial\\_py\\_contours\\_begin.html](https://docs.opencv.org/3.4/d4/d73/tutorial_py_contours_begin.html), [Online; accessed 12-October-2020].
2. Hongxiang Fan, Mingliang Jiang, Ligang Xu, Hua Zhu, Junxiang Cheng, and Jiahu Jiang, 2020, “Comparison of Long Short Term Memory Networks and the Hydrological Model in Runoff Simulation,” **Water**, vol. 12, no. 1, 2020.
3. Fasttext, 2018, “English word vectors,” <https://fasttext.cc/docs/en/english-vectors.html>.
4. Y. Goldberg and O. Levy, 2014, “word2vec Explained: Deriving Mikolov et al.’s,” <https://arxiv.org/pdf/1402.3722v1.pdf>.
5. Google, 2020, “Tesseract OCR,” <https://opensource.google/projects/tesseract>, [Online; accessed 22-November-2020].
6. NECTEC, “AI For Thai,” <https://aiforthai.in.th/index.php#home>, [Online; accessed 22-November-2020].
7. Keiron O’Shea and Ryan Nash, 2015, “An Introduction to Convolutional Neural Networks,” **CoRR**, vol. abs/1511.08458, 2015.
8. P. T. Perez, 2018, “Deep Learning: Recurrent Neural Networks,” <https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010>, [Online; accessed 10-October-2020].
9. Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 2018, “Deep contextualized word representations,” **CoRR**, vol. abs/1802.05365, 2018.
10. Ritambhara, “Minimum edit distance of two strings,” <https://www.ritambhara.in/minimum-edit-distance-of-two-strings/>, [Online; accessed 12-October-2020].
11. Xin Rong, 2014, “word2vec Parameter Learning Explained,” **CoRR**, vol. abs/1411.2738, 2014.
12. Saixiii, 2017, “RESTful คืออะไร REST คือ การสื่อสารแลกเปลี่ยนข้อมูลผ่าน webservice,” <https://saixiii.com/what-is-restful/#:~:text=Representational%20state%20transfer%20%E0%B8%AB%E0%B8%A3%E0%B8%B7%E0%B8%AD%20REST,XML%2C%20HTML%2C%20JSON%20%E0%B9%82%E0%B8%94%E0%B8%A2%20response%2F>.
13. Nattanon Sornchumni, 2018, “ImageProcessing4: Edges and Contours ตามท่าเส้นของภาพ(ไม่ใช่เส้นขอบฟ้า!),” <https://medium.com/nattanon.s/imageprocessing-4-edges-and-contours-%E0%B8%95%E0%B8%B2%E0%B8%A1%E0%B8%AB%E0%B8%B2%E0%B9%80%E0%B8%AA%E0%B9%89%E0%B8%99%E0%B8%82%E0%B8%AD%E0%B8%9A%E0%B8%A0%E0%B8%B2%E0%B8%9E-%E0%B9%84%E0%B8%A1%E0%B9%88%E0%B9%83%E0%B8%8A%E0%B9%/>.
14. techterms, 2018, “MVC,” <https://techterms.com/definition/mvc>, [Online; accessed 10-October-2020].
15. NARESUAN UNIVERSITY, 2018, “Image Morphology,” <https://humancominteracg1.wixsite.com/group1/w2-morphological>, [Online; accessed 22-October-2020].