Project No. 67

ระบบจัดเก็บและจัดการเอกสารภายในหอบรรณสารสนเทศ

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 Prechaprapranwong, Ph.D.)

. . . . . Committee Member

(Asst.Prof.Sanan Srakaew)

. . . . . Committee Member

(Asst.Prof.Surapont Toomnark)

Project Title	Project No. 67 ระบบจัดเก็บและจัดการเอกสารภายในหอบรรณสารสนเทศ
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

หัวข้อปริญญานิพนธ์	ระบบจัดเก็บและจัดการเอกสารภายในหอบรรณสารสนเทศ KMUTT Archives Management Platform
หน่วยกิต	3
ผู้เขียน	นายอัครพล บุญเสริมศักดิ์กุล นางสาวธนพร ปิติอนุสรณ์ นายอรรณพ กองสมบัติเจริญ
อาจารย์ที่ปรึกษา	ผศ.ดร.สุธาทิพย์ มณีวงศ์วัฒนา
หลักสูตร	วิศวกรรมศาสตรบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
คณะ	วิศวกรรมศาสตร์
ปีการศึกษา	2563

บทคัดย่อ

การจะสืบค้นข้อมูลจากเอกสารหรือชั้นหนังสือที่มีการรวบรวมข้อมูลไว้ตั้งแต่อดีตนั้นเป็น ปัญหาอย่างหนึ่งของเจ้าหน้าที่บรรณารักษ์ที่ต้องทำการดูแลเอกสารเหล่านี้ เนื่องจาก การที่ยังไม่มีการเก็บหนังสือและเอกสารให้อยู่ในรูปแบบของข้อมูลดิจิตอลทำให้ต้อง สืบค้นโดยการค้นหาเอกสารและหนังสือแต่ละเล่มโดยการดูจากเนื้อหาสารบัญเพื่อให้ ได้หนังสือที่ตรงกับข้อมูลที่ต้องการมากที่สุด ซึ่งการที่ค้นหาจากหน้าสารบัญของ หนังสือแต่ละเล่มก็จะทำให้การค้นหาเป็นไปอย่างล่าช้า และบางครั้งการดูเพียง แค่สารบัญก็อาจจะทำให้ได้หนังสือที่ไม่ตรงกับความต้องการของผู้ที่เข้ามายืมหนังสือ ในโครงการนี้เราได้ทำการพัฒนาการระบบจัดเก็บและค้นหาเอกสารอิเล็กทรินิกส์ โดยแบ่งออกเป็น 2 ขั้นตอนคือ การนำเข้าข้อมูล และการสร้างระบบค้นหา โดยขั้นตอนการนำเข้าข้อมูล เราจะเริ่มจากการทำ image processing เพื่อเตรียมข้อมูลรูปภาพที่ได้มา ก่อนจะนำไปผ่านกระบวนการ OCR เพื่อแปลงรูปภาพเหล่านี้ให้อยู่ในรูปของข้อมูลดิจิตอล โดยการเก็บข้อมูลในรูปแบบของ Information Retrieval เพื่อช่วยให้ความเร็วการค้นหามีประสิทธิภาพมากยิ่งขึ้น และนำข้อมูลมาทำการตัดคำ และเช็คคำผิด จากนั้นจะนำมาหาคำสำคัญของหนังสือหรือเอกสารนั้น ๆ โดยการใช้การหาคะแนนแบบ TF-IDF ส่วนการสร้างระบบการค้นหาจะเริ่มจากรับคำค้นหามาจากผู้ ใช้และทำการนำคำที่ได้ไปเข้าโมเดล word2Vec เพื่อหาคำที่ใกล้เคียง จากนั้นนำคำใกล้เคียงและคำค้นหาไปดึงคะแนน 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

ขอขอบคุณนางสาวอารยา ศรีบัวบาน เจ้าหน้าที่หอบรรณสารสนเทศและ ผศ.ดร.สุธาทิพย์ มณีวงศ์วัฒนา อาจารย์ที่ปรึกษารวมทั้งเจ้าหน้าที่ภายในหอสมุดมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรีที่เสียสละเวลาให้ความรู้ความเข้าใจ ทั้งในเรื่องการเก็บข้อมูลและคอยแนะนำวิธีการจัดการกับปัญหาต่างๆที่เกิดขึ้น นำมาสู่การทำหัวข้อปริญญานิพนธ์ฉบับนี้ให้สำเร็จตามที่ต้องการ

สารบัญ

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูปภาพ

รูปที่ **หน้า**

สารบัญสัญลักษณ์

SYMBOL		UNIT
α	Test variable	m^2
λ	Interarival rate	jobs/ second
μ	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

นับตั้งแต่การก่อตั้งหอสมุดมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี ได้มีการเก็บรวบรวมองค์ความรู้จากประสบการณ์การทำงานของคณะอาจารย์ ผู้เชี่ยวชาญในทางด้านศาสตร์ต่าง ๆ ในรูปแบบลายมือและสื่อสิ่งพิมพ์ไม่ว่าจะเป็น หนังสือ เอกสาร รวมถึงบันทึกเหตุการณ์ในอดีตในรูปของจดหมายเหตุเพื่อส่งต่อประวัติศาสตร์ความรู้ไปยังคนรุ่นหลังโดยมีการจัดเก็บอยู่ภายในหอจดหมายเหตุที่มีเจ้าหน้าที่บรรณารักษ์เป็นผู้ดูแล และเนื่องจากการที่ เอกสาร หนังสือยังไม่ได้มีการจัดเก็บในรูปแบบดิจิตอลทำให้เมื่อมีบุคคลภายนอกที่ต้องการข้อมูลเพื่อนำไปทำกิจกรรมต่าง ๆ ไม่ว่าจะเป็นการทำวิจัย รายงาน หรือหาข้อมูลเพื่อประกอบการประชุมก็ตามแต่ ก็จำเป็นที่จะต้องมาติดต่อเจ้าหน้าที่บรรณารักษ์ผู้ดูแลเพื่อที่จะให้เจ้าหน้าที่บรรณารักษ์ทำการค้นหาหนังสือที่มีเนื้อหาตามที่เราต้องการ ซึ่งการค้นหาข้อมูลที่ต้องการนั้นเจ้าหน้าที่จะต้องทำการค้นหาด้วยระบบมือทำให้การค้นหาข้อมูลดำเนินการไปอย่างล่าช้า นอกจากนั้นวิธีการหาข้อมูลของเจ้าหน้าที่บรรณารักษ์จะเลือกตรวจสอบข้อมูลของหนังสือจากการดูสารบัญทำให้ข้อมูลที่ได้รับมาอาจจะตกหล่นจากข้อมูลเล่มอื่นได้

เพื่ออำนวยความสะดวกให้กับบรรณารักษ์ในการสืบค้นข้อมูลและทำให้การบริการในการสืบค้นเอกสารต่าง ๆ และให้บุคคลภายนอกสามารถทำการค้นหาข้อมูลได้ด้วยตนเองครบถ้วนทางคณะผู้จัดทำโครงการจึงได้พัฒนาระบบการจัดเก็บเอกสารและระบบการค้นหาโดยการใช้เครื่องมือในการทำ OCR เพื่อแปลงเอกสารให้อยู่ในรูปแบบของเอกสาร digital และหาคำสำคัญในการสร้าง tag ด้วยวิธี Term Frequency - Inverse Document Frequency เพื่อเพิ่มประสิทธิ์ภาพให้กับการค้นหา

1.3 ประเภทของโครงงาน

นำเสนอความต้องการของผู้มีส่วนได้ส่วนเสียเฉพาะกลุ่ม

1.4 วิธีการที่นำเสนอ

ระบบการค้นหาเอกสาร มีขั้นตอนการทำงานดังนี้

1. นำเอกสารมาแปลงเป็นรูปภาพในรูปแบบสแกน
2. นำรูปภาพเข้าสู่ระบบโดยใช้การรับส่งข้อมูลแบบ RESTful API ในระบุประเภทของการใช้งาน
3. นำรูปภาพผ่านกระบวนการ Image Processing โดยใช้ OpenCV ในการลบส่วนอื่น ๆที่ไม่ใช่ข้อความออกและตัดเฉพาะข้อความเพื่อนำไปใช้ในขั้นต่อไป
4. นำรูปที่ผ่านการทำ Image Processing มาเข้าสู่ระบบ OCR เพื่อแปลงข้อมูลจากรูปภาพมาเป็นข้อความในระบบดิจิตอล
5. นำข้อมูลที่เก็บไว้มาทำการตัดแบ่งคำภาษาไทยและแก้คำผิด
6. ค้นหาคำสำคัญโดยใช้วิธี TF-IDF เพื่อนำมาใช้ในการสร้าง Tag
7. นำข้อมูลที่ถูกแปลงเก็บและข้อมูลเกี่ยวกับ Tag ลงในดาต้าเบส
8. ทำระบบค้นหาในรูปแบบ Cosine Similarity
9. ทำระบบหาคำใกล้เคียงโดยใช้วิธี Word2Vec
10. ทำแพลตฟอร์มเว็ปไซต์เพื่อเป็น User Interface ให้กับผู้ใช้งานได้ใช้งานสำหรับการใช้งานในการค้นหาข้อมูลและเพิ่มข้อมูลหนังสือลงไปในฐานข้อมูลเพิ่ม

1.5 วัตถุประสงค์

1. สร้างระบบแปลงข้อมูลเอกสารให้อยู่ในรูปแบบดิจิตอล
2. สร้าง web platform เพื่อทำการค้นหาเอกสารจากคำค้น และพัฒนาเครื่องมือสนับสนุนการทำงานของบรรณารักษ์ประจำหอบรรณสารสนเทศ
3. สร้างระบบการค้นหาโดยการใช้วิธีการ อินฟอเมชันรีทีฟวอล ซึ่งวัดความใกล้เคียงกันระหว่างคำค้นและข้อมูลในฐานข้อมูลโดยวิธี โคซาย ซิมิลาริตี้
4. เพิ่มประสิทธิ์ภาพในการเข้าถึงข้อมูลในรูปแบบดิจิตอล
5. เรียนรู้เรื่องการทำ Image processing

1.6 ขอบเขตของงานวิจัย

1. ระบบแปลงข้อมูลจากเอกสารและหนังสือเก่า รองรับเฉพาะเอกสารที่เป็นตัวอักษรแบบพิมพ์ และรองรับไฟล์เอกสารเฉพาะ PDF เท่านั้น
2. ทำระบบตัดคำ Stop word ภาษาไทยโดยอ้างอิงมาจาก pythainip และภาษาอังกฤษจาก nltk
3. ทำระบบค้นหาแบบ Cosine Similarity ในระบบ Information retrieval
4. ข้อมูลหนังสือที่นำมาใช้คือหนังสือจำพวก งานแสดงกตเวทิตาจิต เอกสารรายงานประจำปี ตั้งแต่ปีพุทธศักราช 2527 ถึง 2560 รวมประมาณ 44 เล่ม จากหอจดหมายเหตุมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
5. ทำ platform เว็บไซต์ในรูปแบบ responsive แต่ไม่รองรับขนาดมือถือ รองรับเฉพาะคอมพิวเตอร์หรือโน๊ตบุ๊ค
6. การแปลงสิ่งพิมพ์เป็นดิจิตอลใช้ Tesseract ในการแปลงเอกสารและหนังสือเป็นรูปแบบดิจิตอล
7. การตัดคำภาษาไทยทางคณะผู้จัดทำ จะใช้ freeware เช่น DeepCut มาใช้ในส่วนของการตัดคำภาษาไทย

1.7 เนื้อหาทางวิศวกรรมที่เป็นต้นฉบับ

- การทำ Image processing สำหรับการเตรียมภาพก่อนนำไปทำ OCR

  โปรเจคของเราทำเกี่ยวกับการทำ OCR เพื่ออ่านภาพให้กลายเป็น text แต่ถึงแม้ว่าภาพที่ได้มาจะจากการสแกนหรือการถ่ายรูป แต่ถึงอย่างนั้น OCR ที่ใช้ก็ยังคงมีข้อจำกัดในเรื่องของคุณภาพของภาพที่ใช้ ถ้าเกิดว่าภาพที่ใช้เอียง หรือมี noise จะทำให้การอ่านมีประสิทธิภาพน้อยลง นอกจากนี้การตัดภาพแยกย่อหน้าแต่ละย่อหน้าทำให้การอ่านมีความถูกต้องมากยิ่งขึ้น

- การพัฒนาเว็บไซต์สำหรับการค้นหาหนังสือในหอจดหมายเหตุ

  เว็บไซต์ของเราจะใช้ ReactJS, NodeJs, python ในการพัฒนาเว็บไซต์เป็น Interface ให้กับ user สำหรับการใช้งานระบบการค้นหาหนังสือ รวมถึงการอัปโหลดเอกสารเพื่อแปลงเอกสารเข้าสู่ระบบดิจิตอลและ API ต่าง ๆ

- คัดเลือกคำสำคัญออกมาเพื่อสร้าง tag

  สำหรับแบ่งแยกหมวดหมู่ของหนังสือโดยใช้ หลักการของ TF-IDF ในการค้นหาคำสำคัญของหนังสือเพื่อนำมาสร้าง tag และใช้สำหรับการค้นหาข้อมูล

- ทำระบบค้นหาโดยใช้คำที่มีความหมายใกล้เคียง

  สำหรับการค้นหาเราจะนำคะแนน TF-IDF มาใช้เป็นคะแนนเพื่อใช้ในการค้นหาแบบ Cosine similarity และค้นหาคำใกล้เคียง (Query Expansion) เพื่อทำให้การค้นหาเจอผลลัพธ์ที่ต้องการเพิ่มมากขึ้น

1.8 การแยกย่อยงาน และร่างแผนการดำเนินงาน

1. ศึกษาและค้นคว้าปัญหาของโครงการ
2. เสนอหัวข้อโปรเจค
3. ค้นหาข้อมูลเกี่ยวกับเทคโนโลยีที่ใช้ในโปรเจค
4. ประเมินความเป็นไปได้และกำหนดขอบเขตของโปรเจค
5. จัดเก็บ requirement จากกลุ่มผู้ใช้งาน
    5.1. ติดต่อเจ้าหน้าที่ของหอสมุด
    5.2. เก็บข้อมูลที่ต้องการแปลงเข้าสู่ระบบดิจิตอล
6. นำเสนอโครงการครั้งที่ 1
7. ออกแบบ UX/UI
8. แปลงรูปภาพเป็น Full-text
    8.1. นำเอกสารมาแปลงเป็นรูปภาพในรูปแบบสแกน
    8.2. ศึกษาการใช้งาน OpenCV
    8.3. สร้างระบบ Image processing เพื่อทำการปรับแต่งรูปภาพและทำการปรับแต่งจนได้ระบบที่รองรับกับ Data ที่มี
    8.4. นำรูปที่ผ่านการทำ Image Processing มาเข้าสู่ระบบ OCR เพื่อแปลงข้อมูลจากรูปภาพมาเป็นข้อความในระบบดิจิตอล
9. นำข้อมูลที่เก็บไว้มาทำการตัดแบ่งคำภาษาไทยและหาคำสำคัญโดยใช้ TF-IDF
    9.1. ทำการตัดแบ่งคำ (Tokenization)
    9.2. ลบ stop word ออกจากข้อมูล
10. ทำระบบค้นหา
    10.1. ทำระบบค้นหาโดยใช้หลักการ Cosine Similarity
    10.2. ทำการค้นหาด้วยคำใกล้เคียงโดยใช้ Word2Vec
11. จัดทำเว็บไซต์แพลตฟอร์ม
12. ทดสอบระบบ
13. ปรับปรุงแก้ไข
14. นำเสนอโปรเจค

1.9 ตารางการดำเนินงาน

ตารางการดำเนินงาน ภาคการศึกษาที่ 1/2563																					
ที่	หัวข้อ	สิงหาคม				กันยายน				ตุลาคม				พฤศจิกายน				ธันวาคม			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	ศึกษาค้นคว้าหาปัญหาของโครงการ																				
2	เสนอหัวข้อโปรเจค																				
3	ศึกษาและหาข้อมูลเกี่ยวกับเทคโนโลยีที่ใช้ในโปรเจค																				
4	ประเมินความเป็นไปได้และกำหนดขอบเขตของโปรเจค																				
5	จัดเก็บ requirement จากกลุ่มผู้ใช้งาน																				
6	นำเสนอโครงงานครั้งที่ 1																				
7	ออกแบบ UX/UI																				
8	แปลงรูปภาพเป็น Full-text																				
9	นำข้อมูลที่เก็บไว้มาทำการตัดแบ่งคำภาษาไทยและทำการสร้าง tag โดยใช้หลักการของ TF-IDF																				
10	นำเสนอโปรเจค																				
11	จัดทำระบบการค้นหา																				

ตารางที่ 1.1 ตารางการดำเนินงาน ภาคการศึกษาที่ 1/2563

ตารางการดำเนินงาน ภาคการศึกษาที่ 2/2563																	
ที่	หัวข้อ	มกราคม				กุมภาพันธ์				มีนาคม				เมษายน			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	จัดทำระบบการค้นหา																
2	จัดทำเว็บไซต์แพลตฟอร์ม																
3	ทดสอบระบบ																
4	ปรับปรุงแก้ไข																
5	นำเสนอโปรเจค																

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