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

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

1.5 วัตถุประสงค์

1.6 ขอบเขตของงานวิจัย

1.7 เนื้อหาทางวิศวกรรมที่เป็นต้นฉบับ

1.8 การแยกย่อยงาน และร่างแผนการดำเนินงาน

1.9 ตารางการดำเนินงาน

1.9.1 ผลการดำเนินงานในภาคการศึกษาที่ 1

1.9.2 ผลการดำเนินงานในภาคการศึกษาที่ 2