

ระบบการจัดการเอกสารโดยการ
แบ่งแยกหมวดหมู่อัตโนมัติ

DOCUMENT MANAGEMENT
SYSTEM WITH AUTOMATIC
CLASSIFICATION

ผู้จัดทำโครงการ

นายปิ่นภัทร วุฒิอาภรณ์

613020224-3

นายรัชชานนท์ ศิริสาร

613020595-8

ที่มาของปัญหาและความสำคัญ

เนื่องจากปัจจุบันเอกสารราชการมีการจัดเก็บในรูปแบบไฟล์เอกสารจำนวนมาก การค้นหาเอกสารเหล่านั้น แม้ว่าจะมีการกำหนดหมายเลขเอกสารซึ่งมีการจำแนกเป็นหมวดหมู่อยู่แล้ว แต่อย่างไรก็ตามในกรณีที่มีเอกสารจำนวนมาก ก็ยังต้องใช้แรงงานมนุษย์เป็นจำนวนมากและยังต้องใช้เวลามากด้วยเช่นกัน

ดังนั้นโครงการนี้มีแนวคิดที่จะพัฒนาระบบเพื่อสแกนหาหมายเลขเอกสาร จากนั้นสกัดออกมาจากเอกสาร เพื่อนำมาใช้ในการค้นหาเอกสาร นอกจากนี้ยังมีการพัฒนาระบบสืบค้นตามหมวดหมู่ของระบบสารบรรณหนังสือราชการไทย

วัตถุประสงค์

- สร้างระบบจัดกลุ่มคำสำคัญโดยอัตโนมัติจากไฟล์PDF เพื่อการค้นหาเอกสาร
- เพื่อพัฒนาระบบรู้จำเลขเอกสารและระบบการค้นหาเอกสารโดยอัตโนมัติ

ขอบเขตงานวิจัย

ส่วนของเว็บไซต์

พัฒนาเว็บไซต์ระบบเอกสารที่ง่ายต่อการใช้งาน พร้อมระบบโอซีอาร์ที่สามารถอ่านไฟล์ภาพหรือไฟล์พีดีเอฟ และแปลงให้อยู่ในรูปแบบของข้อความเพื่อหาคีย์เวิร์ดต่างๆที่ต้องการมาใช้เป็นข้อมูลกำกับของไฟล์นั้นๆ เช่น ชื่อเรื่อง เลขเอกสาร วันเวลา จากนั้นจัดเก็บลงฐานข้อมูล

ส่วนของผู้ใช้งานเว็บไซต์

ผู้ใช้งานสามารถใช้งานเว็บไซต์ได้อย่างมีประสิทธิภาพ และเพิ่มความสะดวกสบายให้ผู้ใช้งาน เช่น ดึงคีย์เวิร์ดต่างๆออกมาใส่ในช่องอินพุทข้อมูลของไฟล์ให้ผู้ใช้อัตโนมัติ เพื่อให้ผู้ใช้ตรวจสอบความถูกต้องและยืนยันเพื่อบันทึกลงฐานข้อมูล

ข้อจำกัดของระบบ

- การสกัดคำที่เขียนเป็นลายมือนั้น ยังมีข้อบกพร่องอีกมาก
- หากตัวอักษรมีการขาดหรือหายทำให้การสกัดคำมีปัญหาได้
- ในการสกัดคำของ Tesseract ในภาษาไทยนั้น มีความแม่นยำไม่มาก

งานวิจัยที่เกี่ยวข้อง

งานวิจัยเรื่อง ระบบสารสนเทศสำหรับการจัดการอุตสาหกรรม

งานวิจัยนี้นำเสนอวิธีการแปลงข้อมูลการผลิตรายปี โดยแปลงภาพตัวอักษร ให้เป็นข้อมูลตัวอักษร ปรากฏว่าการแปลงข้อมูลจากรูปนั้นให้เป็นตัวอักษรมีทั้งข้อดีและข้อเสียที่เป็นอุปสรรคที่เกิดขึ้นดังนี้

1. กรณีต้นฉบับเป็นลายมือเขียน ยังไม่สามารถแปลงเอกสารเป็นตัวหนังสือได้
2. ต้นฉบับมีสิ่งรบกวน ที่ทำให้ไฟล์มีความไม่สมบูรณ์ ก็อาจจะทำให้เกิดข้อผิดพลาดได้
3. เอกสารที่ผ่านการสแกนที่เอียงนั้นทำให้เกิดข้อผิดพลาดในการสกัดคำได้
4. ต้นฉบับตัวอักษรติดกันเกินไปอาจทำให้การสกัดคำผิดพลาดได้

สรุปผลการศึกษาค้นคว้างานวิจัยพบว่า การทำการรู้จำตัวยังมีความผิดพลาดได้ขึ้นอยู่กับปัจจัยต่างที่ทำให้การสกัดคำมีความผิดพลาดได้

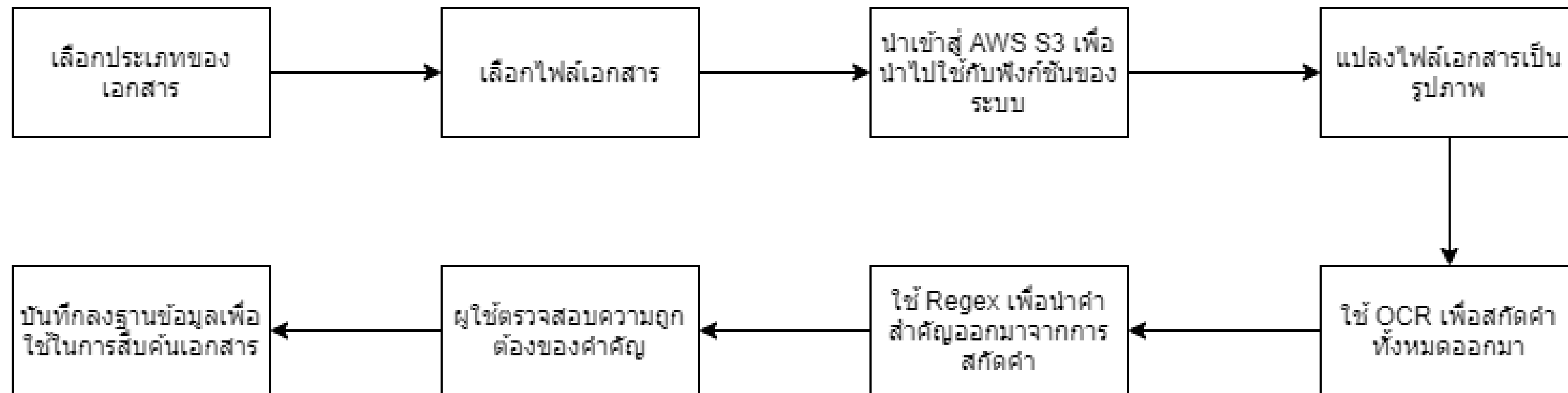
งานวิจัยที่เกี่ยวข้อง

งานวิจัยเรื่อง การเพิ่มประสิทธิภาพการรู้จำอักขระภาษาไทยด้วยแสงโดยใช้เทคนิคเปรียบเทียบสายอักขระโดยประมาณและความแตกต่างของลำดับอักขระ

งานวิจัยนี้เสนอระบบรู้จำตัวอักขระด้วยแสงที่เหมาะสมกับการทดลองที่สุดคือ การสกัดคำแบบเอพียู และพื้นที่หลังที่ให้ประสิทธิภาพในการทดลองดีที่สุดคือ สีขาวและสีของตัวอักษรที่ได้ผลที่สุดคือสีดำ ความละเอียดของภาพขั้นต่ำคือ 13 พิกเซล และมุมกล้องที่ดีที่สุดคือ 90 องศา

สรุปผลการศึกษาค้นคว้างานวิจัยพบว่าสามารถนำมาช่วยเพิ่มประสิทธิภาพในการรู้จำได้โดยการทำงานต้องมีข้อจำกัดที่มากขึ้น

การทำงานของระบบ



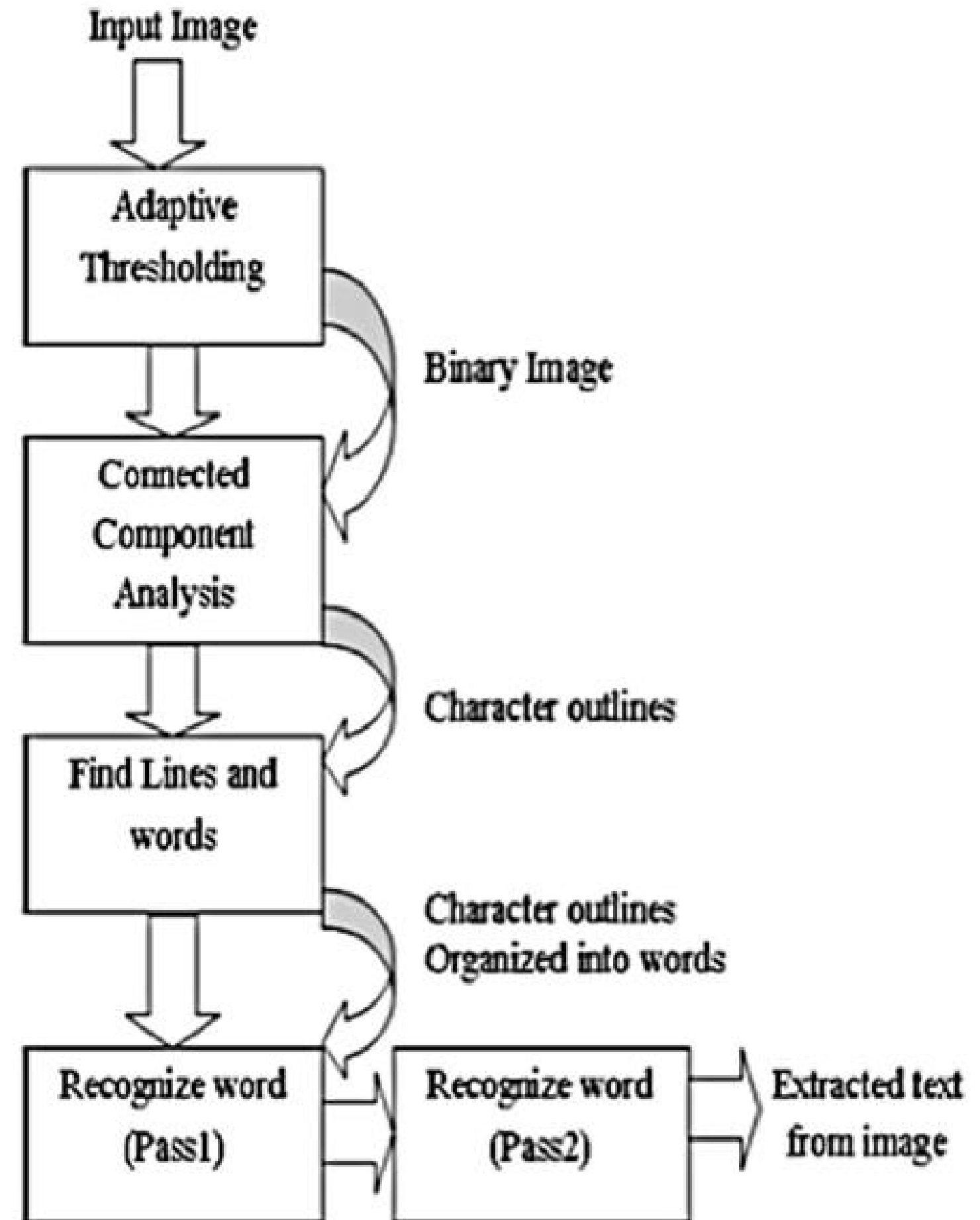
ประเภทของเอกสาร

1. หนังสือภายใน
2. หนังสือภายนอก
3. หนังสือประทับตรา
4. หนังสือคำสั่ง
5. หนังสือระเบียบ
6. หนังสือข้อบังคับ
7. หนังสือประกาศ
8. หนังสือมอบอำนาจ
9. หนังสือภาษาอังกฤษ

Amazon Web Services S3

คือที่จัดเก็บข้อมูลอ็อบเจกต์ที่สร้างขึ้นเพื่อใช้จัดเก็บและเรียกดูข้อมูลตามจำนวนที่ต้องการจากทุกที่ โดยเป็นบริการจัดเก็บข้อมูลที่เรียบง่ายซึ่งมีความทนทาน ความพร้อมใช้งาน ประสิทธิภาพ ความปลอดภัย ที่ต้องการ อีกทั้งยังสามารถ ใช้งานกับ service อื่น ๆ ที่อยู่ใน AWS

ขั้นตอนการทำงานของ Tesseract OCR



ขั้นตอนการทำงานของ Tesseract OCR

1. Adaptive Thresholding

หาค่าเฉลี่ยของทุกพิกเซลภายใต้ Moving window จากนั้นทำการหาค่าเฉลี่ยเช่นนี้ไปเรื่อย ๆ กับบริเวณที่ไม่ซ้ำกันจนกระทั่งได้มีการกำหนดค่าเฉลี่ยครบในทุก ๆ พิกเซล ถ้าค่า Gray Level ของพิกเซลนั้นมีค่ามากกว่าค่าเฉลี่ยของพิกเซลนั้นแล้วจะกำหนดให้เป็นสีขาว แต่ถ้าค่า Gray Level ของพิกเซลนั้นน้อยกว่าค่าเฉลี่ยของพิกเซลนั้นแล้วจะกำหนดให้เป็นสีดำ

ขั้นตอนการทำงานของ Tesseract OCR

2.Connected Component Analysis

ขั้นตอนการทำงานของ Tesseract OCR

3.Find Lines And Words

ขั้นตอนการทำงานของ Tesseract OCR

4. Recognize word (pass1)


การรู้จำครั้งที่1นั้นจะใช้วิธีรู้จำแบบ วิธีทางสถิติ (Statistical Approach) เป็นวิธีการที่ใช้หลักการทางสถิติ โดยนำค่าความน่าจะเป็นและ/หรือฟังก์ชันการแจกแจงความน่าจะเป็นมาใช้ในการตัดสินใจ ซึ่งได้ผลลัพธ์ออกมาเป็นค่าความน่าจะเป็นที่อินพุตเป็นตัวอักษรใด เมื่ออินพุตได้ผ่านส่วนการรู้จำครบทุกตัวแล้ว ก็นำเอาผลลัพธ์ที่ได้ทั้งหมดมาเปรียบเทียบกับกันได้ค่าความน่าจะเป็นของตัวอักษรใดมากที่สุด ผลลัพธ์จะออกเป็นตัวอักษรนั้น

ขั้นตอนการทำงานของ Tesseract OCR

5. Recognize word (pass2)

การรู้จำครั้งที่1นั้นจะใช้วิธีรู้จำแบบ Adaptive Classifier

ตัวอย่างเอกสาร



คณะกรรมการ
งานบริการการศึกษา
เลขที่ ๒๖๖
วันที่ ๒๘ ก.ย. ๒๕๕๘
เวลา ๑๖.๐๐-๑๖.๓๐

เลขที่ ๕๔๙๘
วันที่ ๒๐ ก.ย. ๒๕๕๘
เวลา ๑๖.๓๐

บันทึกข้อความ

ส่วนราชการ บัณฑิตวิทยาลัย โทร. ๔๒๔๑๙

๑. ที่ ศธ ๐๕๑๔.๑๐/๒๕๖๓๔ ๒. วันที่ ๒๕ กันยายน ๒๕๕๘

๓. เรื่อง การเสนอชื่อผู้รับทุนวิจัยสำหรับคณาจารย์บัณฑิตศึกษา ประจำปีการศึกษา ๒๕๕๘ แบบร่วมทุน (joint funding)

เรียน คณบดีคณะวิทยาศาสตร์

ตามบันทึกที่ ศธ ๐๕๑๔.๒.๑/๔๕๘๓ ลงวันที่ ๑๘ กันยายน ๒๕๕๘ เรื่อง ขอเสนอชื่อผู้รับทุนวิจัยสำหรับคณาจารย์บัณฑิตศึกษา ประจำปีการศึกษา ๒๕๕๘ แบบร่วมทุน (joint funding) ซึ่งตามประกาศบัณฑิตวิทยาลัย มหาวิทยาลัยขอนแก่น (ฉบับที่ ๖๖/๒๕๕๘) เรื่อง การรับสมัครทุนวิจัยสำหรับคณาจารย์ ประจำปีการศึกษา ๒๕๕๘ แบบร่วมทุน (joint funding) นั้น ผศ.ดร.พฤษดี ศิริแสงตระกูล เป็นผู้ได้รับอนุมัติทุนระดับปริญญาโท และได้เสนอ นางสาวปวีณา อุ่นลี รหัส ๕๘๕๐๒๐๑๑๘-๓ เพื่อรับทุนดังกล่าว จากการตรวจสอบข้อมูลนักศึกษาสำเร็จการศึกษาในระดับปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ มีผลการเรียน ๓.๑๕

บัณฑิตวิทยาลัยโดยคณะกรรมการพิจารณาทุนฯ ได้พิจารณาแล้วไม่สามารถอนุมัติทุนวิจัยสำหรับคณาจารย์ฯ ดังกล่าวได้ เนื่องจากคุณสมบัติของนักศึกษาที่ได้รับทุนในระดับปริญญาโท ไม่เป็นไปตามเกณฑ์ดังที่ระบุไว้ในประกาศ คือ เป็นนักศึกษามหาวิทยาลัยขอนแก่นที่กำลังศึกษาอยู่ในภาคการศึกษาสุดท้ายของหลักสูตรระดับปริญญาตรี หรือสำเร็จการศึกษาไปแล้ว ที่มีผลการเรียนเฉลี่ยไม่ต่ำกว่า ๓.๒๕ หรืออยู่ในกลุ่ม ๕% แรกของผู้สำเร็จการศึกษาในสาขาวิชานั้น ทั้งนี้ ผศ.ดร.พฤษดี ศิริแสงตระกูล สามารถแจ้งรายชื่อ

ตัวอย่างหน้าเว็บไซต์

Upload Document

1.

2.

3.

หน้าเว็บไซต์ (หลังนำไฟล์เอกสารเข้าสู่ระบบ)

Upload Document

PHG.JPG

↓

Choose Files | Sample.png

Or Drag It Here.

ประเภท

เลือก

เลขที่หนังสือ

วันที่

เรื่อง

Submit

หน้าเว็บไซต์ (หลังจากผ่านการทำงานของระบบ)

Upload Document

png.jpg

↓

Choose Files Sample.png

Or Drag It Here.

ประเภท

เล่มจบ

เลขที่หนังสือ

ศส 0514.10/3534

วันที่

25 กันยายน 2558

เรื่อง

การเสนอเรื่องในทุนวิจัยสำหรับผลงานวิจัยเพื่อการศึกษา

Submit

ปัญหาที่พบ

- การสกัดคำจากลายมือนั้นมีความแม่นยำที่น้อย
- การจัดเก็บสารต้องใช้เวลา เพราะเอกสารทั้งหมดมีจำนวนมาก

การพัฒนาต่อไป

จากผลการพัฒนาระบบโอซีอาร์เพื่อนำมาใช้งานกับระบบเอกสาร พบว่าการใช้ระบบโอซีอาร์กับเอกสารที่ใช้ภาษาอังกฤษนั้นได้ผลลัพธ์ที่ค่อนข้างดี แต่พอนำมาประยุกต์ใช้กับเอกสารภาษาไทยนั้นได้ผลลัพธ์ไม่เป็นที่น่าพอใจนัก จึงควรพัฒนาด้วยการนำเทคโนโลยี ดีฟ เลิร์นนิ่ง มาใช้เพื่อให้การอ่านตัวอักษร มีประสิทธิภาพและความแม่นยำมากยิ่งขึ้น

The top of the slide features a decorative header with a light pink background. On the left, there is a blue quarter-circle. On the right, there is a small pink circle.

THANK YOU

The bottom of the slide features a decorative footer with a light pink background. On the left, there is a small blue circle. On the right, there is a large pink semi-circle.

THANK YOU

[1] วีรพล มั่นสอานินาท. (2008). ระบบสารสนเทศสำหรับการจัดการข้อมูลอุตสาหกรรม. สาขาวิชาการจัดการงานวิศวกรรม ภาควิชาวิศวกรรมอุตสาหการและการจัดการ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

[2] พรศิริ ภาณุตญาณชัย. (2015). การเพิ่มประสิทธิภาพการรู้จำอักขระภาษาไทยด้วยแสงโดยใช้เทคนิคเปรียบเทียบสายอักขระโดยประมาณและความแตกต่างของลำดับอักขระ. หลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยนเรศวร

[3] รศ. นิตยา เกิดประสพม, รศ. กิตติศักดิ์ เกิดประสพ (2018). การพัฒนาวิธีการฮิวริสติกเพื่อเพิ่มประสิทธิภาพการรู้จำลายมือชื่อ. มหาวิทยาลัยเทคโนโลยีสุรนารี

[4] Regular Expression (RegEx), <https://www.bualabs.com/archives/3070/what-is-regular-expression-regex-regexp-teach-how-to-regex-python-nlp-ep-7/?fbclid=IwAR1UCuWfwClnRufYKad4XI95zPudEuNv5RbQrlWj7TJS1rClyj2ACqvhZP0>, December 03, 2019.

[5] THAI OCR, <http://thaiocr.phaisarn.com> , JULY 24, 2012.

[6] Deep learning, <https://www.thaiprogrammer.org/2018/12/deep-learning-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3/> , December 16, 2018.

“

[7] Review on Tesseract OCR Engine and Performance. www.ijiere.com , 2017

“