FACULTY OF INFORMATION TECHNOLOGY
King Mongkut's Institute of Technology Ladkrabang

SENIOR PROJECT
SHOWCASE

ระบบสกัดข้อมูลจากเอกสารแบบอัตโนมัติ

(Automated Documents Extraction System)

รหัสนักศึกษา 57070019
ผู้พัฒนา นายจิรวัฒน์ บุญกำเหนิด
ที่ปรึกษา ผศ.ดร. กิติ์สุชาต พสุภา
คณะ เทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

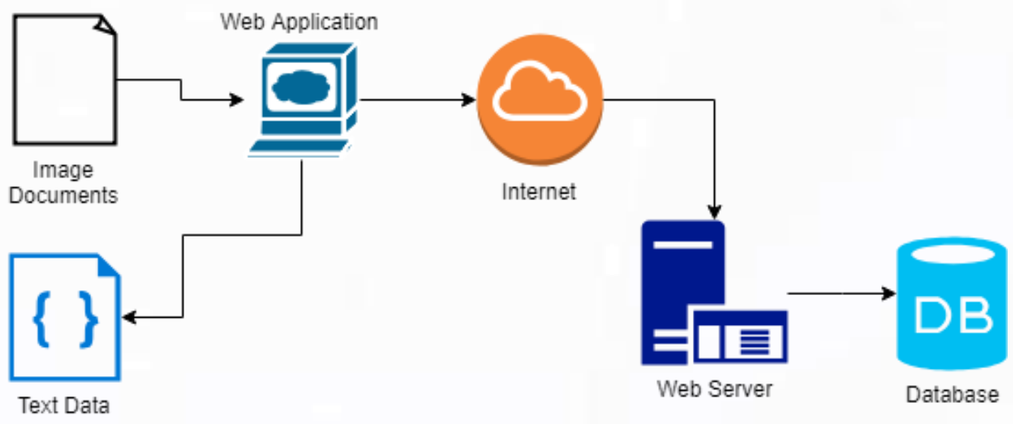
บทคัดย่อ

เป็นระบบจัดเก็บเอกสารที่มีฟังก์ชันหลากหลายในการใช้งานเพื่อเพิ่มประสิทธิภาพและความสะดวกสบายในการจัดเก็บรือค้นหาเอกสารในระบบ นอกจากนี้ยังสามารถสกัดข้อมูลจากเอกสารได้อีกด้วย ระบบนี้เป็นระบบภายในองค์กร เหมาะสำหรับองค์กรที่ต้องการจัดเก็บเอกสารรูปแบบดิจิทัลต่างๆโดยแยกหมวดหมู่ชัดเจน เพื่อลดต้นทุนการจัดเก็บในรูปแบบสิ่งพิมพ์ และเพิ่มความสะดวกสบายในการจัดเก็บ

สิ่งที่พัฒนา

ระบบจัดเก็บเอกสารบน Cloud มีฟีเจอร์การทำงานที่อัตโนมัติช่วยเพิ่มความสะดวก และลดต้นทุกในการจัดเก็บเอกสาร ระบบนี้เป็นระบบที่พัฒนาเพื่อนำไปใช้ในองค์กรต่างๆ อย่างมีประสิทธิภาพ โดยใช้งานผ่านเว็บแอพลิเคชัน และจัดเก็บข้อมูลลงบนดาต้าเบส

พัฒนาวิธีในการ Classification ของเอกสารที่เป็นรูปภาพ เพื่อเพิ่มประสิทธิ์ภาพในการจัดกลุ่มของเอกสาร



วัตถุประสงค์

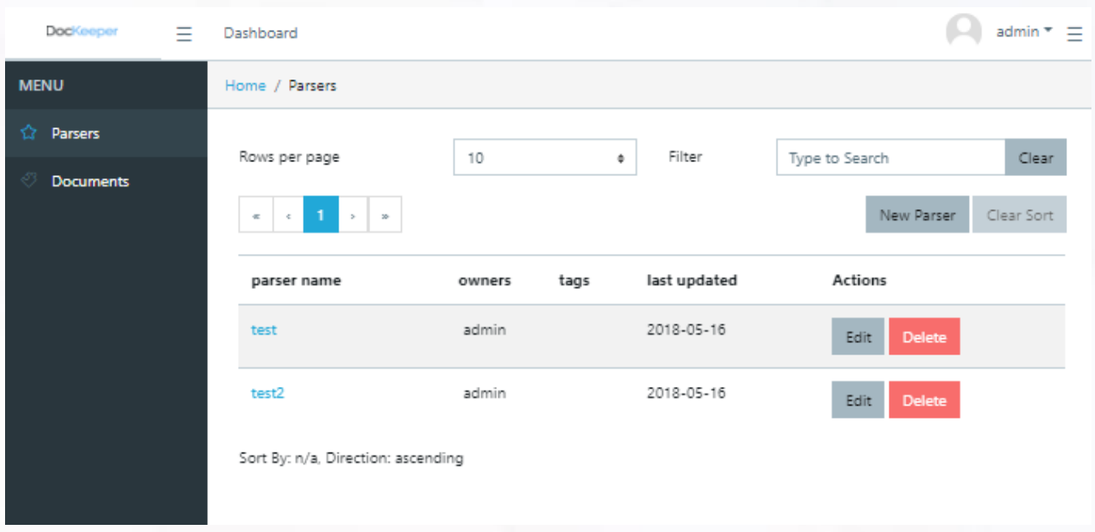
1.) เพื่อเป็นระบบสำหรับจัดเก็บเอกสาร
2.) เพื่อลดปัญหาเอกสารสูญหายหรือชำรุด
3.) เพื่อเพิ่มความสะดวกในการจัดเก็บข้อมูลเข้าสู่ฐานข้อมูล
4.) เพื่อเพิ่มความรวดเร็วในการจัดเก็บข้อมูลเข้าสู่ฐานข้อมูล
5.) เพื่อลดความผิดพลาดในการป้อนข้อมูลเข้าสู่ฐานข้อมูล
6.) เพื่อเพิ่มความสะดวกในการค้นหาเอกสาร
7.) เพื่อจำแนกเอกสารแบบอัตโนมัติ

DocKeeper

สรุปผลการพัฒนา

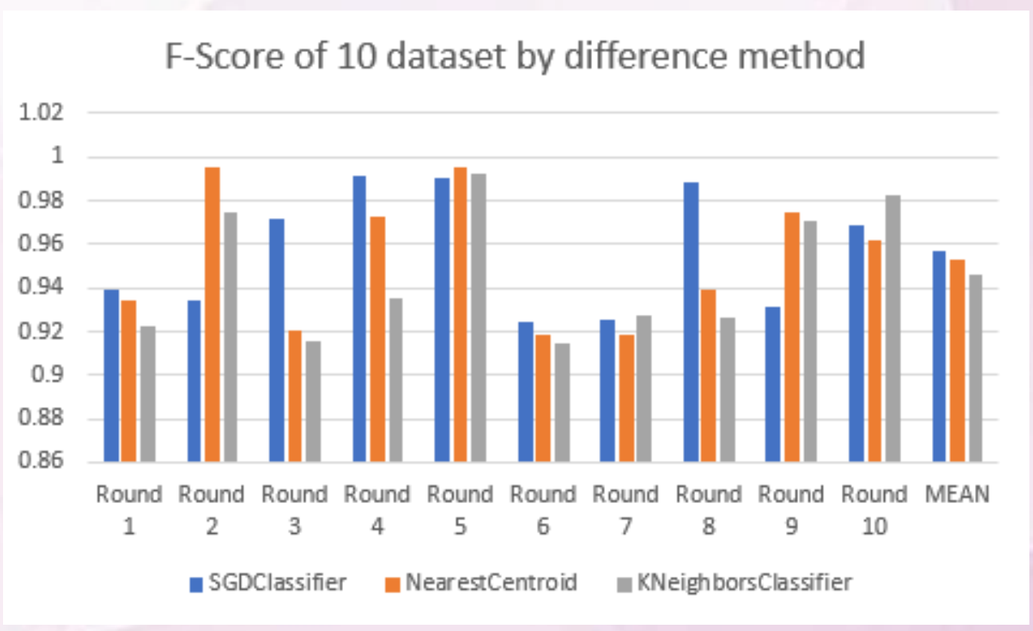
Web Application

- สามารถอัพโหลด/แก้ไข/ลบ เอกสาร
- สามารถจัดหมวดหมู่ของเอกสาร
- สามารถ Extract ด้วย OCR และกำหนดรูปแบบในการ Extract
- มีระบบยืนยันตน สามารถกำหนด Permission ในการจัดการแต่ละชนิดของเอกสาร
- สามารถ Export ข้อมูลเป็น Json, CSV, Text



Document Classification

วิธีที่ใช้ในการทดสอบมี 3 วิธี Stochastic gradient descent, Nearest centroid, k-nearest neighbors โดยมีข้อมูลทดสอบทั้งหมด 2112 รูป และจำนวน class ทั้งหมด 24 แบบ เริ่มจากนำรูปทั้งหมดไปทำการ feature extraction ด้วยวิธีการ histogram of oriented gradients โดย resize รูปภาพให้เท่ากัน ที่ตั้งไว้คือ 700*500 และกำหนดทิศทางของ HOG 8 ทิศทาง จะได้ฟีเจอร์ทั้งหมด 85,608 feature จากนั้นแบ่งข้อมูลเป็น train 70% test 30% ก่อนที่จะทดสอบได้มีการทำ k-fold cross-validation จาก train data แบ่งข้อมูลเป็น 5 set เพื่อทำการปรับ parameter neighbor ของ K-NN และ Learning rate ของ SGD จากนั้นรันทดสอบ 10 รอบด้วย random test data มีผลสรุปดังนี้



วิธีที่เลือกใช้ในระบบคือ Nearest centroid เพราะว่ามีความเร็วในการ fit มากกว่า SGD ถึง 10 เท่า และความแม่นยำไม่ได้น้อยกว่า SGD มาก