

# ระบบสกัดข้อมูลจากเอกสารแบบอัตโนมัติ

## (Automated Documents Extraction System)

รหัสนักศึกษา 57070019

ผู้พัฒนา นายจิรวัฒน์ บุญกำหนด

ที่ปรึกษา ผศ.ดร. กิตติสุชาติ พสุภา

คณะ เทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

## บทคัดย่อ

เป็นระบบจัดเก็บเอกสารที่มีฟังก์ชันหลากหลายในการใช้งานเพื่อเพิ่มประสิทธิภาพและความสะดวกสบายในการจัดเก็บหรือค้นหาเอกสารในระบบ นอกจากนี้ยังสามารถสกัดข้อมูลจากเอกสารได้ด้วย ระบบนี้เป็นระบบภายในองค์กรเหมาะสำหรับองค์กรที่ต้องการจัดเก็บเอกสารรูปแบบดิจิทัลต่างๆโดยแยกหมวดหมู่ชัดเจน เพื่อลดต้นทุนการจัดเก็บในรูปแบบสิ่งพิมพ์ และเพิ่มความสะดวกสบายในการจัดเก็บ

## สิ่งที่พัฒนา

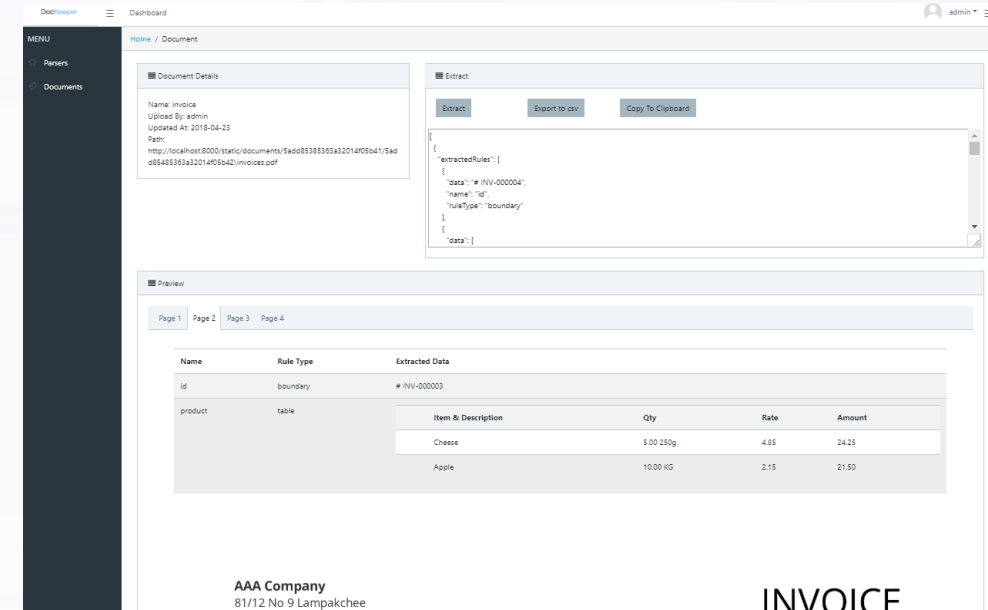
ระบบจัดเก็บเอกสารบน Cloud มีฟีเจอร์การทำงานที่อัตโนมัติช่วยเพิ่มความสะดวก และลดต้นทุนในการจัดเก็บเอกสาร ระบบนี้เป็นระบบที่พัฒนาเพื่อนำไปใช้ในองค์กรต่างๆ อย่างมีประสิทธิภาพ โดยใช้งานผ่านเว็บแอปพลิเคชัน และจัดเก็บข้อมูลลงบนฐานข้อมูล

พัฒนาวิธีการ Classification ของเอกสารที่เป็นรูปภาพ เพื่อเพิ่มประสิทธิภาพในการจัดกลุ่มของเอกสาร

## สรุปผลการพัฒนา

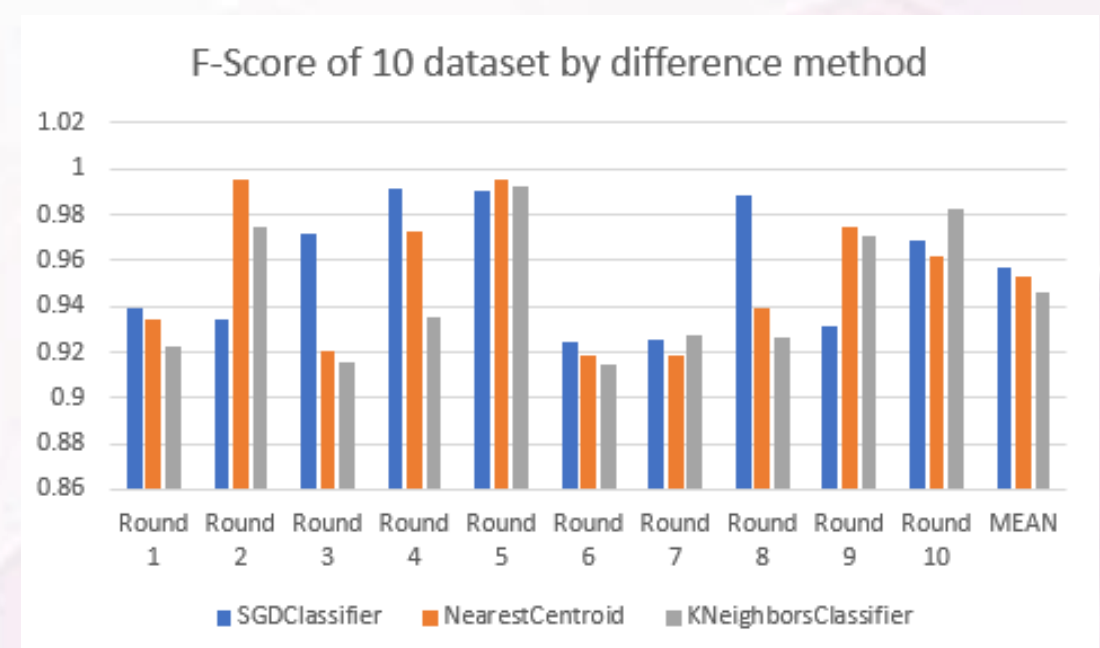
### Web Application

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

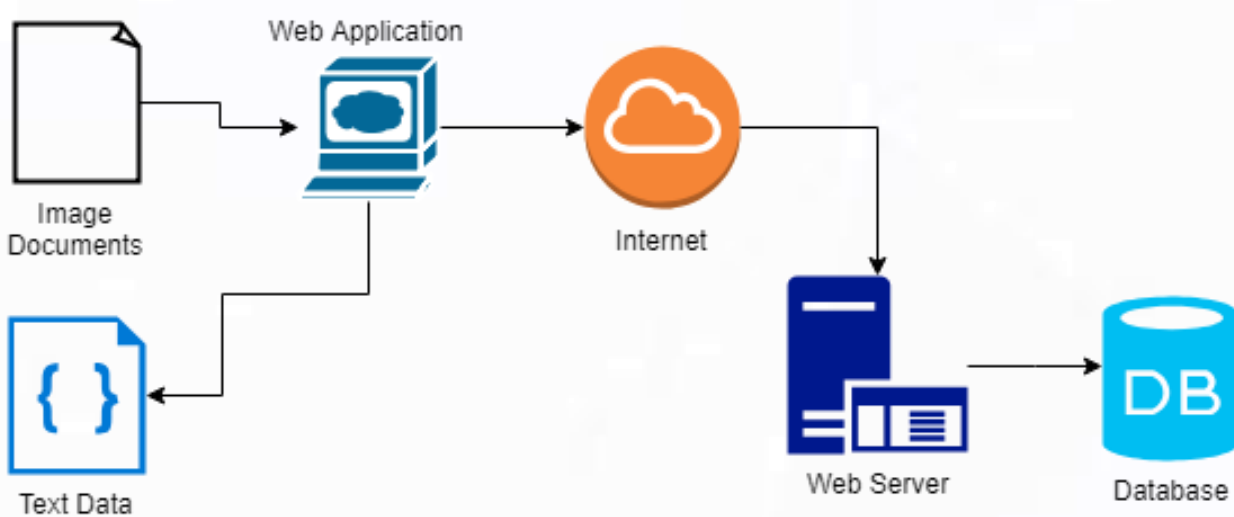
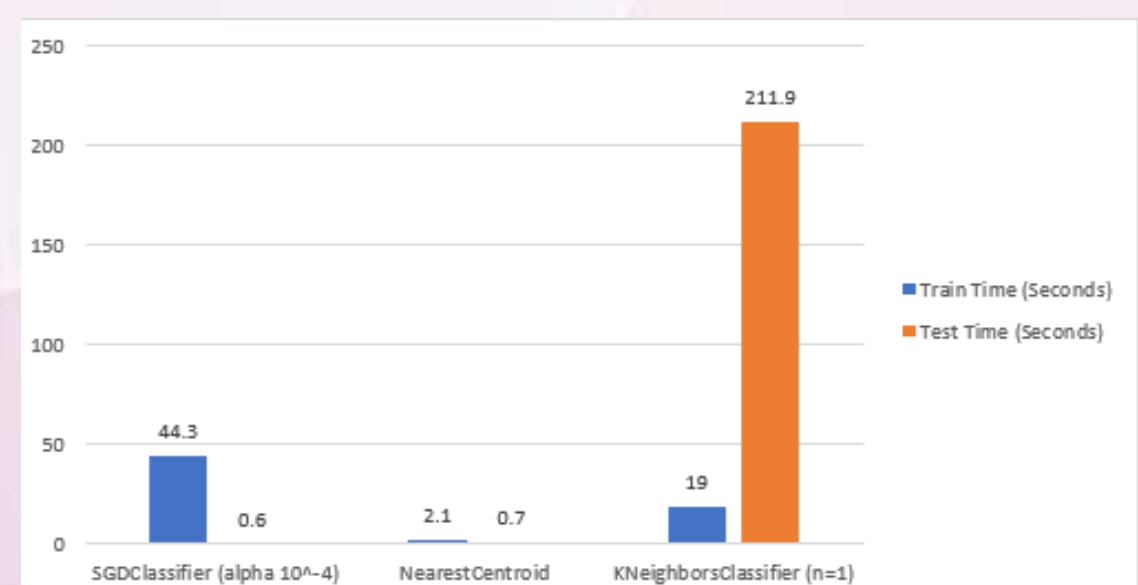


### Document Classification

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 จำนวนที่ดีที่สุดคือ 1 และ Learning rate ของ SGD ค่าที่ดีที่สุดคือ 0.0001 จากนั้นรันทดสอบ 10 รอบด้วย random test data มีผลสรุปดังนี้

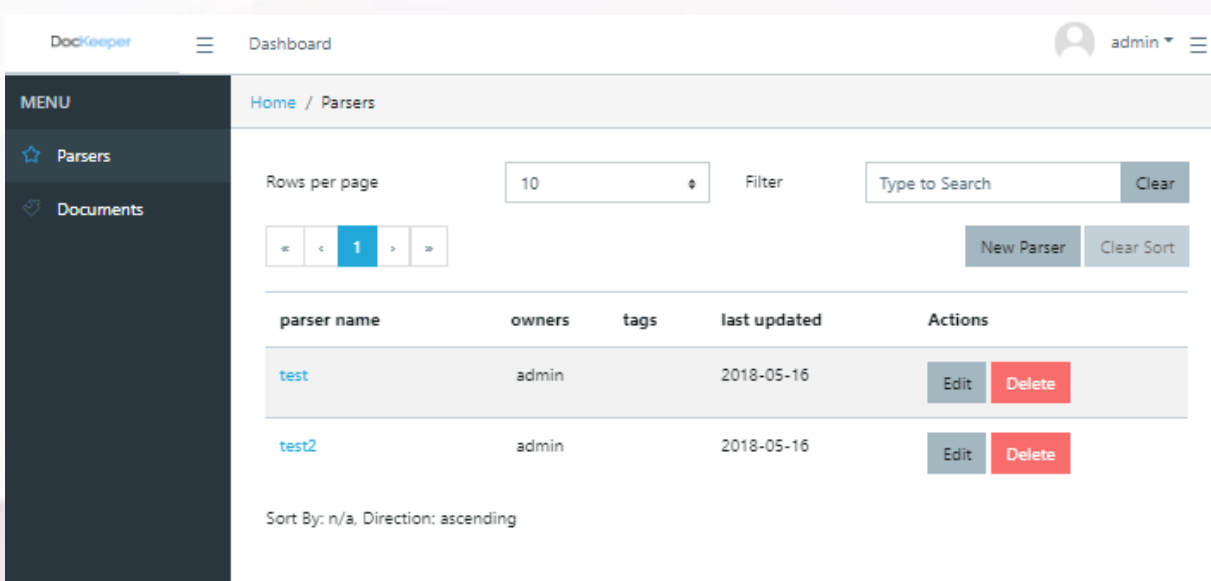


วิธีที่เลือกใช้ในระบบคือ Nearest centroid เพราะว่ามีความเร็วในการ fit มากกว่า SGD และความแม่นยำไม่ได้น้อยกว่า SGD มาก ตามระยะเวลาเฉลี่ยในการรัน 10 รอบ



## วัตถุประสงค์

- เพื่อเป็นระบบสำหรับจัดเก็บเอกสาร
- เพื่อลดปัญหาเอกสารสูญหายหรือชำรุด
- เพื่อเพิ่มความสะดวกในการจัดเก็บข้อมูลเข้าสู่ฐานข้อมูล
- เพื่อเพิ่มความเร็วในการจัดเก็บข้อมูลเข้าสู่ฐานข้อมูล
- เพื่อลดความผิดพลาดในการป้อนข้อมูลเข้าสู่ฐานข้อมูล
- เพื่อเพิ่มความสะดวกในการค้นหาเอกสาร
- เพื่อจำแนกเอกสารแบบอัตโนมัติ



# DocKeeper