ระบบสกัดข้อมูลจากเอกสารแบบอัตโนมัติ

นายจิรวัฒน์ บุญกำเนิด

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ

Email: 57070019@kmitl.ac.th

**บทคัดย่อ**

เป็นระบบ Cloud Application ที่ใช้ในการจัดเก็บเอกสารเข้าสู่ระบบ มีฟังก์ชัน ในการใช้งานหลากหลายเพื่อเพิ่มประสิทธิภาพในการจัดเก็บและเพิ่มความสะดวกสบาย ในการจัดเก็บหรือค้นหาเอกสารในระบบ ระบบนี้สามารถสกัดข้อมูลจากเอกสารด้วยเทคโนโลยี Optical character recognition ได้สามารถจำแนกกลุ่มของเอกสารแบบอัตโนมัติ ด้วยการทำ Classification ด้วยการทดสอบหาวิธีที่เหมาะสมที่สุดในการทำ Classification สำหรับระบบ และระบบนี้ยังสามารถจำกัดสิทธิ์การเข้าถึงของแต่ละผู้ใช้ ระบบนี้เหมาะสำหรับเป็นระบบที่ใช้ภายในองค์กรที่ต้องการจัดเก็บเอกสารรูปแบบดิจิทัลต่างๆ เช่น รูปภาพ หรือไฟล์ PDF โดยต้องการแยกหมวดหมู่ของเอกสารชัดเจน ต้องการความสะดวกสบายและความง่ายดายในการจัดเก็บ และเพื่อลดต้นทุนการจัดเก็บแทนการจัดเก็บในรูปเเบบสิ่งพิมพ์

คำสำคัญ – Document, Classification, Image Processing

**1.บทนำ**

**1.1 ความสำคัญและความเป็นมาของโครงงาน**

หน่วยงานต่างๆในปัจจุบัน ยังคงใช้วิธีในการจัดเก็บเอกสารเเบบสิ่งพิมพ์ ทำให้ มีปัญหาในการค้นหา ซึ่งทำได้อย่างยากลำบาก เเละเอกสารมีโอกาสที่จะสูญหายหรือชำรุด

ทำให้บางหน่วยงานในปัจจุบันเริ่มเปลี่ยนมาจัดเก็บเอกสารในรูปแบบดิจิทัลกันแล้ว เพื่อลดปัญหาข้อมูล สูญหายหรือชำรุด เพื่อเพิ่มความสะดวกในการค้นหา จึงทำให้บางบริษัทเริ่มมีการพัฒนา Cloud Application ที่จะช่วยในการจัดการกับเอกสารในรูปแบบดิจิทัลออกมาแล้ว เพื่อนำมาจำหน่ายบริการให้กับองค์กรต่างๆ

ด้วยเหตุผลประการนี้ดังนั้นเราจึงมีเเนวคิดที่จะพัฒนาระบบที่สามารถจัดเก็บเอกสารเป็นข้อมูลเเบบดิจิทัล ด้วยวิธีการเเสกนเอกสารเป็นรูปภาพเเล้วจึงจัดเก็บไว้บนฐานข้อมูล เเละยังมีฟีเจอร์การใช้งานต่างๆ เพื่อเพิ่มความสะดวกให้เเก่ผู้ใช้งาน เช่น สามารถเเนะนำผู้ใช้ว่าเอกสารนี้เป็นเอกสารชนิดไหนด้วยเทคนิคการจำเเนก (Classification) รูปภาพของเอกสาร สามรถสกัดข้อมูลด้วยวิธีการ Optical character recognition (OCR) ออกมาเป็นรูปแบบ Text เพื่อนำไปใช้งานอย่างอื่นได้ สามารถ Export ข้อมูลเป็นรูปแบบ Json format หรือ CSV format ได้

**1.2 วัตถุประสงค์**

- เพื่อเป็นระบบสำหรับจัดเก็บเอกสารในรูปแบบดิจิทัล

- เพื่อเพิ่มความสะดวกและรวดเร็วในการจัดเก็บข้อมูลเข้าสู่ฐานข้อมูล

- เพื่อลดความผิดพลาดในการป้อนข้อมูลเข้าสู่ฐานข้อมูล

- เพื่อทดสอบวิธีในการจำแนกเอกสาร

- เพื่อสกัดข้อมูลเอกสารและนำไปใช้งานรูปแบบอื่น

- เพื่อลดปัญหาเอกสารสูญหายหรือชำรุด

- เพื่อเพิ่มความสะดวกในการค้นหาเอกสาร

**1.3 เป้าหมายและขอบเขตของโครงการ**

1. ออกแบบฐานข้อมูล

2. จัดทำ Web Server ในการจัดเก็บเอกสาร

3. จัดทำ Web Application

4. ทำ Feature ในการ Extract ข้อมูลจากเอกสาร

5. ทดสอบวิธีที่ใช้ในการจำแนกเอกสาร จากนั้นเลือกวิธีที่เหมาะสมที่สุด

6. สามารถ Export ข้อมูลเป็นรูปแบบ Image, PDF, CSV, JSON

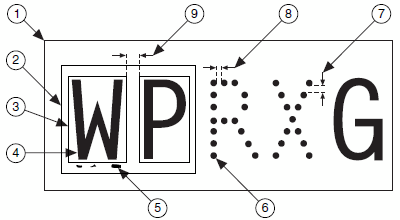
7. มีการยืนยันตัวตนก่อนเข้าใช้งานระบบ

8. สามารถกำหนดสิทธิ์การเข้าถึงของกลุ่มของเอกสาร

**2.ทฤษฏีและเทคโนโลยีที่เกี่ยวข้อง**

**Optical Character Recognition (OCR)**

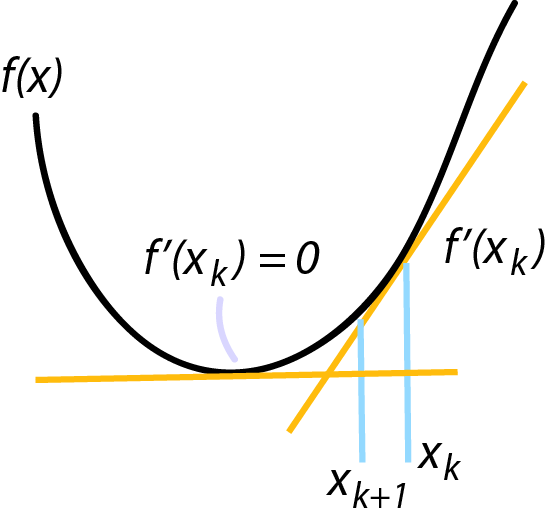
คือ กระบวนการจดจำตัวอักษรของคอมพิวเตอร์ ที่มากจากสิ่งพิมพ์หรือมาจาก การเขียนด้วยรายมือ สามารถสกัดตัวอักษรจากรูปภาแปลงเป็นข้อความในรูปแบบที่ จะนำไปใช้งานอื่นๆในคอมพิวเตอร์ได้ง่ายขึ้น ในการทำ OCR สามารถใช้อัลกอริทึม หรือวิธีที่ต่างกันไป



รุปที่ 2.1 กระบวนการทำ OCR

Gradient Descent

เป็นอัลกอริทึมในการหา cost function หรือค่าที่เหมาะสมให้กับฟังก์ชัน โดยการวนหาค่าไปเรื่อยๆเพื่อหาที่ที่ต่าที่สุด อาจใช้เพื่อหาสมการเส้นตรง



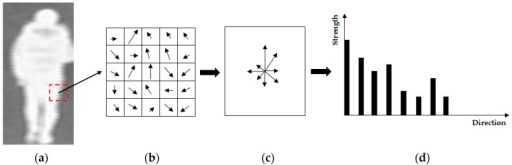
รูปที่ 2.3 วิธีการทำ Gradient Descent

Stochastic Gradient Descent (SGD)

เป็นการหา cost function คล้ายกลับ Gradient Descent แต่ทุกรอบในการคำนวณ จะสุ่มข้อมูลเพียงบางส่วนเพื่ออัพเดตค่า ไม่ได้ใช้ข้อมูลทั้งหมดเหมือน Gradient Descent

Histogram of oriented gradients

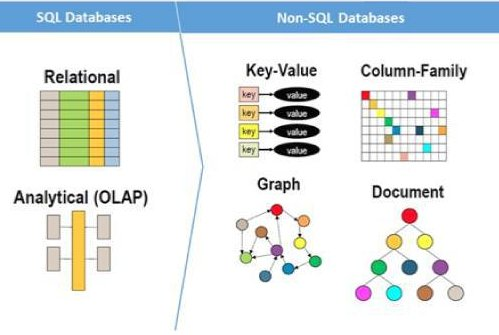
เป็นการหาทิศทางของรูปภาพโดยดูจากสีของรูปภาพ โดยแบ่งเป็น Block และแต่ละ Block แบ่งเป็นหลายๆ Cell



รูปที่ 2.3 ขั้นตอนการทำ HOG

**NoSQL**

คือเเนวทางสำหรับออกเเบบเเละจัดเก็บฐานข้อมูล ที่เหมาะกับข้อมูลที่มีขนาดใหญ่ เเละไม่มีโครงสร้างที่ชัดเจน ทำให้รองรับข้อมูลที่มีรูปเเบบหลากหลายเเละยืดหยุ่น เเละยังเหมาะสำหรับเเอพลิเคชันที่มีการใช้งานเเบบ Realtime



รูปที่ 2.4 ความแตกต่าง SQL และ Non-SQL

**Flask Web framework**

เว็บ Framework ที่เขียนด้วยภาษา Python มีต้นแบบมาจาก Werkzeug และ Jinja 2 จุดเดนตรงที่มีความใช้งานง่ายและหนักเบา

**MongoDb**

ระบบฐานข้อมูลแบบ NoSQL หรือ Non-Relational Database เป็นฐานข้อมูลที่เหมาะสำหรับจัดเก็บเอกสารที่มีความยืดหยุ่นของข้อมูลสูง และมีการเรียกข้อมูลได้เร็วกว่า SQL ทั่วไป

**Vue.js**

เป็น Web front-end framework ที่ทำงานบน Node.js สามารถ Render ไฟล์เว็บ สามารถทำ Routing URL ได้ และยังมีเครื่องมืออื่นๆให้ใช้อีกครบครัน

**OpenCV**

เป็น Library ที่ใช้สำหรับการทำ Image Processing และ Computer Vision มีเครื่องมือให้ใช้งานได้หลากหลาย Library ถูกเขียนด้วยภาษา C++

**Tesseract OCR**

เป็น Library ในการทำ OCR ที่ทาง Google ได้พัฒนาขึ้น มีความแม่นยำสูง และในรุ่นใหม่มีการเทรนโมเดลด้วย LSTM ทำให้มีความแม่นยำมากขึ้น

**ImageMagick**

เป็น Open-source สำหรับ การจัดการรูปภาพ แปลงไฟล์รูปภาพ แก้ไขรุปภาพ หรือแม้แต่บีบอัดรูปภาพ ก็สามารถทำได้

**SciPy Toolkits**

เป็นกลุ่มของ open source ซอฟแวร์ที่ใช้สำหรับในการทำการคำนวนทางวิทยาศาสตร์ต่างๆ มีซอฟแวร์ย่อยๆอีกมากมาย เช่น NumPy, Pandas, Scikit-learn , Scikit-image, Ipython

**3.การวิเคราะห์และออกแบบระบบ**

**3.1 ขั้นตอนการปฏิบัติ**

1.ทำการศึกษาวิธีการทำระบบ Web Application โดยเทคโนโลยีที่จะใช้มีดังนี้ Flask, Vue.js

2.ทำการศึกษาวิธีการจัดการฐานข้อมูล โดยชนิดของฐานข้อมูลที่เลือกใช้คือ MongoDb

3.ศึกษาวิธีการใช้งาน Python Library เช่น Numpy, Scipy, Keras, OpenCV เป็นต้น

4.ศึกษาวิธีเทคนิคการทำ Image Processing ด้วยวิธีต่างๆ เช่น Maximally Stable Extremal Regions (MSER), Text Detection Methods, HoughLines

5.ทดสอบ Histogram of Oriented Gradients Image Feature Extraction

6.ศึกษาวิธีการทำ Document Classification ด้วยวิธี

-Stochastic gradient descent

-Nearest centroid

-k-nearest neighbors

เเละวัดประสิทธิภาพเพื่อเปรียบเทียบผลของการจัดหมวดหมู่ของรูปเอกสาร แล้วจึงเลือกใช้วิธีที่ได้ผลดีที่สุด

7.ทำการออกแบบClass Diagram ของระบบ

8.ทำการออกแบบการทำงานของระบบและ Flow การทำงานต่างๆ

9. ทำการออกแบบตัวอย่างจำลอง (Mockup) ของWeb Application

10.ลงมือพัฒนาระบบ Web Application

11.ลงมือพัฒนาฟีเจอร์ในการสกัดข้อมูลจากเอกสารด้วยวิธีการ OCR

12.ลงมือพัฒนาฟีเจอร์ในการ Recommend Parser ด้วยวิธีการ Classification

13.ทำการทดสอบระบบ

14.ทำการ Deployment ระบบเพื่อสำหรับการใช้งานจริง

15.จัดทำGuide และวีดีโอสอนวิธีการใช้งาน

16.จัดทำเอกสาร

**3.2 ขอบเขตและข้อจำกัด**

-ไม่สามารถใช้บน Device ที่ไม่มีการเชื่อมต่อกับอินเทอร์เน็ตได้

-สามารถใช้งานได้เฉพาะ Device ที่สามารถใช้งาน Web Browser ได้เท่านั้น