Project Functionality

Feature: PDF/DOC to Text

* แยกข้อความภาษาไทยในแต่ละย่อหน้าออกมาจากไฟล์ PDF หรือ DOC,DOCX โดย Text ภาษาไทยที่ได้จากเครื่องมือจะมีความถูกต้อง 50% ขึ้นไปสำหรับไฟล์ PDF และ เกือบสมบูรณ์สำหรับไฟล์ DOC,DOCX

Feature: Word Segmentation

* แยกประโยคภาษาไทยออกมาเป็นคำเดี่ยวในภาษาไทย เพื่อนำไปใช้ในการสร้าง vector

Feature: Machine Learning

* Model จะเป็นใช้วิธีการทำ supervised classification
* Train Model โดยใส่ Text/Tag ของ Paragraph โดยเนื้อหาจะมีภาษาไทยและภาษาอังกฤษ เพื่อทำการสร้าง Model ออกมา ซึ่ง Tag จะมาจากผู้เชี่ยวชาญที่จะทำการให้ tag ของ dataset ต่างๆ
* ตัว Model จะทำการ Generate Tag ออกมา โดยใช้ Text จาก Paragraph โดยเนื้อหาของ Paragraph จะมีภาษาไทยและภาษาอังกฤษ

Feature: Web Application

* Upload PDF/DOC ภาษาไทยไปเก็บไว้ใน Web Server
* สามารถให้ผู้เชี่ยวชาญติด Tag ของเอกสารต่างๆ ที่มีอยู่ใน Web Server โดยแต่ละ Username จะมีการเก็บข้อมูล Tag ที่แตกต่างกัน
* สามารถสืบค้นข้อมูลตัวอย่างเอกสาร โดยผ่านการพิมพ์ข้อความหรือคีย์เวิร์ดลงในช่อง Search หรือผ่านการคลิกเลือกจากเมนู Advance Filter
* สามารถแสดงผลการสืบค้นในหลายระดับ คือประเภทเอกสาร หมวดหมู่ของ tag และตัวอย่างข้อความที่สำคัญ นอกจากนี้ผู้ใช้ระบบจะสามารถดาวน์โหลดไฟล์ต้นฉบับออกจากระบบได้หากต้องการ

**Advisor :** Aj.Tiranee Achalakul

**Committee 1:** Aj.Santitham Prom-on

**Committee 2:** Aj.Phond Phunchongharn

**Committee 3:** Aj.Jumpol Polvichai

**Members:**

1. **Name**: นายศุภณัฐ ทัตตินาพานิช  
   **Email**: [zarkzaki@hotmail.com](mailto:zarkzaki@hotmail.com)  
   **ID**: 56070501053
2. **Name**: นายอินทัช แสงกระจ่าง  
   **Email**: [artkrub7@gmail.com](mailto:artkrub7@gmail.com)  
   **ID**: 56070501068