วริศรา พิลาสุข 6809036087

Project Proposal

ชื่อโครงงาน: การเปรียบเทียบ PageIndex และ PageIndex ร่วมกับ Light Knowledge Graph สำหรับระบบแชทบอทตอบคำถามกฎหมายอาญาไทย

1. ปัญหาและความสำคัญ (Problem Statement)

ระบบ Retrieval-Augmented Generation (RAG) เป็นเทคนิคสำคัญในการสร้าง AI ที่สามารถตอบคำถามจากเอกสาร โดยระบบ RAG แบบดั้งเดิมใช้ Vector Embeddings ในการค้นหาข้อความที่มีความคล้ายคลึงทางความหมาย อย่างไรก็ตาม วิธีนี้มีข้อจำกัดสำคัญสำหรับเอกสารกฎหมาย ได้แก่ (1) ข้อความที่มีความคล้ายคลึงกันแต่ไม่ได้สอดคล้องกันทางความหมาย เช่น คำถามเรื่อง "โทษลักทรัพย์" อาจดึงข้อความเรื่อง "โทษชิงทรัพย์" มาแทน (2) การแบ่งเนื้อหาเอกสารออกเป็นส่วนที่เท่ากันทำลายโครงสร้างลำดับชั้นของกฎหมาย และ (3) การค้นหาแบบ Vector ไม่สามารถให้เหตุผลหลายขั้นตอนสำหรับคำถามกฎหมายที่ซับซ้อนได้

PageIndex จึงเป็นแนวทาง RAG แบบที่ไม่ต้องใช้ Vector ที่เสนอขึ้นใหม่เพื่อแก้ปัญหาดังกล่าว โดยการสร้างดัชนีแบบต้นไม้จากเอกสารและใช้ LLM เข้ามาช่วยในการให้เหตุผลเพื่อค้นหา อย่างไรก็ตาม ในปัจจุบันยังไม่มีการศึกษาว่า PageIndex จะทำงานได้ดีเพียงใดกับเอกสารกฎหมายภาษาไทย และยังไม่มีงานวิจัยที่ศึกษาการนำ Knowledge Graph มาช่วยเพิ่มความแม่นยำของระบบ

2. วัตถุประสงค์ (Objectives)

2.1 พัฒนาระบบ PageIndex สำหรับประมวลกฎหมายอาญาไทย พร้อมรองรับภาษาไทย

2.2 สร้าง Light Knowledge Graph ที่ประกอบด้วย entities ทางกฎหมาย เช่น มาตรา, ความผิด, โทษ และความสัมพันธ์

2.3 เปรียบเทียบประสิทธิภาพระหว่าง PageIndex อย่างเดียว กับ PageIndex ร่วมกับ Light Knowledge Graph ในงานแชทบอทตอบคำถามกฎหมายไทย

3. ชุดข้อมูล (Dataset)

3.1 คลังข้อมูลกฎหมายไทยประเภทพระราชบัญญัติ ซึ่งรวบรวมมาจากเว็บไซต์ของสำนักงานคณะกรรมการกฤษฎีกา (https://www.krisdika.go.th/) แหล่งที่มา https://github.com/PyThaiNLP/thai-law?tab=readme-ov-file

3.2 ชุดข้อมูลทดสอบ Q&A ที่สร้างขึ้น พร้อมคำตอบมาตรฐานและการอ้างอิงมาตรา

4. เทคนิค Machine Learning ที่ใช้ (Proposed ML Techniques)

องค์ประกอบ	เทคนิค	วัตถุประสงค์
Document Indexing	PageIndex Tree Construction	สร้างโครงสร้างต้นไม้ตามลำดับชั้นของ
Text	LLM-based Summarization	สร้างสรุปสำหรับแต่ละ node ในต้นไม้
Entity Extraction	Named Entity Recognition ผ่าน	ดึง entities ทางกฎหมาย (มาตรา, ความผิด,
Relation Extraction	LLM-based Relation Extraction	ระบุความสัมพันธ์ระหว่าง entities
Query	Intent Classification	จำแนกประเภทคำถามเพื่อเลือกกลยุทธ์การ
Retrieval	Tree Search + KG Lookup	การค้นหาแบบผสมผสาน tree navigation กับ KG index
Answer Generation	RAG with LLM	สร้างคำตอบพร้อมอ้างอิงมาตรากฎหมาย
Evaluation	LLM-as-Judge, ROUGE,	ประเมินคุณภาพและความถูกต้องของคำตอบ

5. ความท้าทายที่คาดว่าจะพบ (Expected Challenges)

1) การประมวลผลภาษาไทยและศัพท์กฎหมายเฉพาะทาง
2) การแยกวิเคราะห์โครงสร้างมาตรากฎหมาย
3) ความแม่นยำในการดึง Entity จาก LLM
4) ข้อจำกัดของชุดข้อมูลทดสอบขนาดเล็ก
5) การกำหนด Ground Truth สำหรับคำถามกฎหมายที่อาจมีคำตอบถูกต้องได้หลายแบบ