เอกสารข้อมูลของ CAIKA-RAG-Typhoon

ภาพรวม

CAIKA-RAG-Typhoon (Corporate AI Knowledge Assistant) เป็นระบบ Retrieval-Augmented Generation (RAG) ที่ออกแบบมาเพื่อให้คำตอบที่แม่นยำและตรงบริบทต่อคำถามเกี่ยวกับความรู้ขององค์กรโดยใช้คลังเอกสารท้องถิ่น

ข้อมูลผู้พัฒนา

- **ผู้พัฒนา**: ฐาปกรณ์ หาญรักษ์ (Thapagorn Hanrak)
- อีเมล: thapagorn.h@outlook.com

- **โมเดลที่ใช้**: scb10x_llama3.2-typhoon2-t1-3b-research-preview-gguf

สถาปัตยกรรมระบบ

องค์ประกอบหลัก

1. **กระบวนการประมวลผลเอกสาร**
    - ตัวโหลด: ใช้ PyPDFLoader สำหรับประมวลผลเอกสาร PDF
    - การแบ่งข้อความ: RecursiveCharacterTextSplitter ด้วยชิ้นส่วนขนาด 1,000 ตัวอักษร และการซ้อนทับ 200 ตัวอักษร
    - การจัดเก็บ: เอกสารถูกเก็บไว้ในไดเรกทอรี `data/documents`
2. **ระบบการฝังข้อมูล (Embedding)**
    - โมเดล: HuggingFaceEmbeddings (ค่าเริ่มต้น: "scb10x/llama3.2-typhoon2-t1-3b-research-preview")
    - คลังเวกเตอร์: FAISS สำหรับการค้นหาความคล้ายคลึงอย่างมีประสิทธิภาพ
    - การจัดเก็บ: การฝังข้อมูลเวกเตอร์เก็บไว้ใน `data/vector_store`
3. **โมเดลภาษา**
    - แบ็กเอนด์: LlamaCpp สำหรับการอนุมานในเครื่องที่มีประสิทธิภาพ
    - โมเดล: "scb10x_llama3.2-typhoon2-t1-3b-research-preview-gguf"
    - การกำหนดค่า: เปิดใช้งานการเร่งความเร็ว GPU ด้วยความแม่นยำครึ่งหนึ่ง (f16_kv) สำหรับการใช้หน่วยความจำอย่างมีประสิทธิภาพ
4. **กระบวนการ RAG**
    - ตัวค้นคืน: การค้นหาความคล้ายคลึงโดยใช้ FAISS

    - การตั้งค่าเริ่มต้น: k=4 เอกสารที่ดึงต่อคำถาม
    - ประเภทเชน: "stuff" (รวมเอกสารที่เกี่ยวข้องทั้งหมดเข้าไปในบริบทเดียว)

ข้อกำหนดทางเทคนิค

การปรับแต่งประสิทธิภาพ

- ปรับให้เหมาะสำหรับ Apple Silicon (M1) ด้วยการสนับสนุน Metal Performance Shaders
- ลดขนาดหน้าต่างบริบท (2048 โทเค็น) เพื่อประหยัดหน่วยความจำ
- แคชคีย์/ค่าความแม่นยำครึ่งหนึ่งสำหรับประสิทธิภาพที่ดีที่สุด
- การประมวลผลหลายเธรด (ค่าเริ่มต้น 4 เธรด)

ส่วนติดต่อผู้ใช้

- สร้างด้วย Streamlit สำหรับการเข้าถึงแบบโต้ตอบผ่านเว็บ
- คุณสมบัติ:
    - การอัปโหลดและประมวลผลเอกสาร
    - การเลือกโมเดลและการปรับพารามิเตอร์
    - การติดตามประวัติการสนทนา
    - การแสดงเอกสารต้นฉบับเพื่อความโปร่งใส

Dependencies

- langchain & langchain_community
- sentence-transformers
- faiss-cpu
- pypdf
- tiktoken
- llama-cpp-python
- streamlit

พารามิเตอร์การใช้งาน

พารามิเตอร์การสร้าง LLM

- อุณหภูมิ: 0.7 (ค่าเริ่มต้น, ปรับได้ผ่าน UI)
- Top-p: 0.95 (ค่าเริ่มต้น, ปรับได้ผ่าน UI)
- โทเค็นสูงสุด: 1024 (ค่าเริ่มต้น, ปรับได้ผ่าน UI)

พารามิเตอร์การดึงเอกสาร

- เอกสารที่ดึง: 4 (ค่าเริ่มต้น)
- พารามิเตอร์การดึงที่ปรับได้:

    - k: จำนวนเอกสารที่จะดึง
    - fetch_k: จำนวนที่จะดึงในเบื้องต้นก่อนการกรอง
    - score_threshold: คะแนนความคล้ายคลึงขั้นต่ำ

ข้อกำหนดการติดตั้ง

ความต้องการของระบบ

- MacOS กับ Apple Silicon (M1/M2) แนะนำ
- RAM ขั้นต่ำ 8GB
- พื้นที่ดิสก์ 1GB+ สำหรับโมเดลและคลังเวกเตอร์

โครงสร้างไดเรกทอรี

- `models/`: พื้นที่เก็บไฟล์โมเดล LLM
- `data/documents/`: พื้นที่เก็บเอกสาร PDF
- `data/vector_store/`: พื้นที่เก็บการฝังข้อมูลเวกเตอร์ที่สร้างขึ้น
- `logs/`: บันทึกแอปพลิเคชัน

หมายเหตุการปฏิบัติการ

- ระบบต้องเริ่มต้นก่อนการใช้งานผ่านปุ่ม "Initialize CAIKA"
- เอกสารต้องได้รับการประมวลผลก่อนที่จะสามารถตอบคำถามได้
- คุณภาพการตอบสนองขึ้นอยู่กับคุณภาพและความครอบคลุมของเอกสารที่ให้
- การประมวลผลทั้งหมดเกิดขึ้นในเครื่อง เพื่อความเป็นส่วนตัวของข้อมูล

CAIKA-RAG-Typhoon เหมาะสำหรับบุคคลและองค์กรในหลากหลายรูปแบบ ดังต่อไปนี้:

เหมาะสำหรับบุคคล:

1. **นักวิเคราะห์ข้อมูลองค์กร** - ผู้ที่ต้องการสกัดข้อมูลและความเข้าใจอย่างรวดเร็วจากเอกสารจำนวนมาก
2. **ผู้บริหารและผู้จัดการ** - ผู้ที่ต้องการค้นหาข้อมูลเชิงลึกจากเอกสารภายในโดยไม่ต้องอ่านเอกสารทั้งหมด
3. **ทีมฝ่ายวิจัยและพัฒนา** - ผู้ที่ต้องการเข้าถึงข้อมูลวิจัยภายในองค์กรอย่างรวดเร็ว
4. **เจ้าหน้าที่ฝ่ายสนับสนุนลูกค้า** - ผู้ที่ต้องการตอบคำถามลูกค้าโดยอิงจากเอกสารผลิตภัณฑ์

เหมาะสำหรับองค์กร:

1. **องค์กรที่มีเอกสารภายในจำนวนมาก** - เช่น บริษัทที่มีคู่มือ นโยบาย และรายงานจำนวนมาก
2. **องค์กรที่ให้ความสำคัญกับความเป็นส่วนตัวของข้อมูล** - เนื่องจากประมวลผลข้อมูลในเครื่องเท่านั้น ไม่ส่งข้อมูลไปยังระบบคลาวด์
3. **บริษัทที่มีข้อจำกัดในการเชื่อมต่ออินเทอร์เน็ต** - เพราะสามารถทำงานได้แบบออฟไลน์
4. **องค์กรในอุตสาหกรรมที่มีกฎระเบียบเข้มงวด** - เช่น การเงิน การแพทย์ หรือการทหาร ที่ไม่สามารถใช้ AI บนคลาวด์ได้
5. **บริษัทขนาดกลางและเล็ก** - ที่ต้องการระบบ AI ภายในองค์กรโดยไม่มีค่าใช้จ่ายรายเดือนสูง
6. **สถาบันการศึกษาและวิจัย** - ที่ต้องการเครื่องมือสำหรับวิเคราะห์และสืบค้นเอกสารวิชาการ

ข้อได้เปรียบเฉพาะสำหรับองค์กรไทย:

1. **รองรับการวิเคราะห์เอกสารภาษาไทย** - หากใช้กับโมเดลที่รองรับภาษาไทย
2. **ประหยัดค่าใช้จ่าย** - องค์กรไทยสามารถหลีกเลี่ยงค่าใช้จ่ายในการใช้บริการ AI ต่างประเทศที่มีราคาสูง
3. **ความเป็นส่วนตัวของข้อมูล** - ข้อมูลองค์กรไม่ต้องออกนอกประเทศไทย ลดความเสี่ยงด้านความปลอดภัย
4. **ปรับแต่งตามบริบทเฉพาะของธุรกิจไทย** - สามารถปรับใช้กับเอกสารที่มีลักษณะเฉพาะของภาคธุรกิจไทย

CAIKA-RAG-Typhoon จึงเป็นทางเลือกที่เหมาะสมสำหรับองค์กรที่ต้องการระบบผู้ช่วย AI ที่ปลอดภัย เป็นส่วนตัว มีประสิทธิภาพ และปรับแต่งได้ตามความต้องการเฉพาะขององค์กร