บธ.001/63

บันทึกข้อความ

ส่วนงาน คณะวิทยาศาสตร์ งานบริหารและธุรการ โทร 3801
ที่ อว 69.5.1.1/1105 **วันที่** 22 สิงหาคม 2568
เรื่อง ขอรายงานสรุปเนื้อหาและการนำไปใช้ประโยชน์

เรียน คณบดีคณะวิทยาศาสตร์

ตามคำสั่ง/บันทึกข้อความ ที่ อว. 69.5.1.2.1/1598 ลงวันที่ 13 สิงหาคม 2568 อนุมัติให้ข้าพเจ้าพร้อมด้วยเจ้าหน้าที่ จำนวน 0 คน เดินทางไปปฏิบัติงานเรื่อง เข้าอบรมเชิงปฏิบัติการ เรื่อง RAG Implementation using Various Models & Platforms ณ สถาบัน IMC (IMC Institute) ถนน สุรวงศ์ แขวงสุริยวงศ์ เขตบางรัก กรุงเทพมหานคร นั้น

บัดนี้ ข้าพเจ้าได้เข้าร่วมปฏิบัติงานเรื่อง เข้าอบรมเชิงปฏิบัติการเรื่อง RAG Implementation using Various Models & Platforms เป็นที่เรียบร้อยแล้ว ดังนั้นจึงขอรายงานสรุปเนื้อหาและประโยชน์ที่ได้รับ ดังนี้

1. ทฤษฎีที่เกี่ยวข้องกับ Retrieval Augmented Generation (RAG)

1) Retrieval Augmented Generation (RAG) เป็นเทคนิคที่ผสานการค้นคืนข้อมูลด้วย vector search โดยใช้ embedding เพื่อแปลงข้อความเป็นเวกเตอร์และวัดความเหมือนด้วย similarity metrics เช่น cosine similarity ใน vector database อย่าง Pinecone จากนั้นนำข้อมูลที่ดึงมาเป็นบริบทให้ large language model (LLM) สร้างคำตอบที่แม่นยำและเกี่ยวข้อง โดย tokenization ช่วยแบ่งข้อความเป็นหน่วยย่อยเพื่อให้ LLM ประมวลผลได้ RAG ลดการ hallucinate ของ LLM และเพิ่มประสิทธิภาพโดยไม่ต้อง fine-tune โมเดล เหมาะสำหรับงานที่ต้องการข้อมูลเฉพาะเจาะจง เช่น การตอบคำถามจากเอกสาร

2) Generative AI Foundation Models คือโมเดลภาษาขนาดใหญ่ (LLM) เช่น GPT, LLaMA หรือ BERT ที่ได้รับการฝึกบนข้อมูลจำนวนมหาศาลเพื่อสร้างข้อความ, รูปภาพ, หรือเนื้อหาอื่น ๆ โดยเป็นรากฐานสำหรับงาน Generative AI การนำ Generative AI ไปใช้งาน (Implementation) เกี่ยวข้องกับการเลือกโมเดล, การปรับแต่ง (fine-tuning) หรือใช้เทคนิคอย่าง Retrieval Augmented Generation (RAG) ซึ่งผสานการค้นคืนข้อมูลจาก vector database (โดยใช้ embedding และ similarity metrics) กับการสร้างคำตอบโดย LLM เพื่อเพิ่มความแม่นยำและลดการ hallucinate RAG vs. Fine-Tuning: RAG ใช้ข้อมูลภายนอกเป็นบริบทโดยไม่ต้องปรับโมเดล ในขณะที่ fine-tuning ปรับโมเดลให้เหมาะกับงานเฉพาะแต่ใช้ทรัพยากรมากกว่า การสร้าง Gen AI Apps บน Cloud Platforms อาศัยแพลตฟอร์มอย่าง AWS, Google Cloud, หรือ Azure ซึ่งมีเครื่องมือและ API สำหรับจัดการ LLM, vector database และการประมวลผลข้อมูลขนาดใหญ่ การสร้าง RAG ด้วย AI Agent Service ผ่าน n8n ใช้ n8n ซึ่งเป็นเครื่องมือ workflow automation เพื่อเชื่อมต่อ LLM, vector database และแหล่งข้อมูลต่าง ๆ ในการสร้าง RAG pipeline ที่ทำงานอัตโนมัติ RAG Architectures ประกอบด้วยส่วนหลัก ได้แก่ retriever (ค้นหา

ข้อมูลด้วย vector search), generator (LLM สร้างคำตอบ) และ data store (เช่น vector database) ซึ่งออกแบบให้เหมาะกับงาน เช่น แชทบอทหรือระบบตอบคำถาม โดยรวมช่วยให้แอปพลิเคชัน Generative AI มีประสิทธิภาพและแม่นยำยิ่งขึ้น.

2. หลักการปฏิบัติสำหรับการจัดการ RAG ด้วย Python และ n8n

1) **การเตรียมและเก็บข้อมูลแบบไม่มีโครงสร้าง (Non-Structured Data)**
    - **หลักการ**: รวบรวมข้อมูล เช่น เอกสารข้อความ, PDF, หรือไฟล์อื่น ๆ ที่ไม่มีโครงสร้างชัดเจน และจัดเก็บในที่ที่เข้าถึงได้ เช่น Google Drive.
    - **แนวทางปฏิบัติ**:
        - อัปโหลดไฟล์ไปยัง Google Drive และใช้ Google Colab เพื่อเชื่อมต่อผ่าน API หรือ gdown เพื่อดึงข้อมูล.
        - แปลงไฟล์ (เช่น PDF) เป็นข้อความด้วยเครื่องมืออย่าง PyPDF2 หรือ pdfplumber.
2) **การแบ่งข้อความเป็นส่วน (Text Chunking)**
    - **หลักการ**: แบ่งข้อความยาวเป็นส่วนย่อย (chunks) เพื่อให้เหมาะกับการประมวลผลและการสร้าง embedding.
    - **แนวทางปฏิบัติ**:
        - ใช้ไลบรารี เช่น langchain หรือ sentence-transformers เพื่อแบ่งข้อความด้วย chunk_size=1000 (จำนวนตัวอักษร) และ chunk_overlap=200 เพื่อรักษาความต่อเนื่องของบริบท.
3) **การสร้าง Embedding เพื่อแปลงข้อความเป็นเวกเตอร์**
    - **หลักการ**: ใช้โมเดล embedding เพื่อแปลงข้อความเป็นเวกเตอร์ที่มีความหมายเชิงในมิติสูง.
    - **แนวทางปฏิบัติ**:
        - ใช้โมเดล **BAAI/bge-m3** จาก Hugging Face ซึ่งให้เวกเตอร์มิติ 1024.
4) **การจัดเก็บเวกเตอร์ในฐานข้อมูลเวกเตอร์ (Vector Database)**
    - **หลักการ**: เก็บเวกเตอร์ที่ได้จาก embedding ในฐานข้อมูลเวกเตอร์เพื่อให้ค้นหาได้รวดเร็ว.
    - **แนวทางปฏิบัติ**:
        - ใช้ **ChromaDB** เป็นฐานข้อมูลเวกเตอร์ที่ติดตั้งง่ายใน Python

5) **การทดสอบโมเดลแชทสำหรับ RAG**
    - **หลักการ**: ใช้โมเดลภาษาขนาดใหญ่ (LLM) เพื่อสร้างคำตอบโดยผสานข้อมูลที่ค้นคืนจากฐานข้อมูลเวกเตอร์.
    - **แนวทางปฏิบัติ**:

        - **ตัวเลือกที่เสียค่าใช้จ่าย**: ใช้ **OpenAI** หรือ **OpenRouter** ผ่าน API key สำหรับโมเดล เช่น GPT-4 ซึ่งทำงานบน CPU ได้ดี.
    - ตัวเลือกฟรี: ใช้โมเดล เช่น Ollama, DeepSeek R1, หรือ Gemma ซึ่งต้องติดตั้งบนเครื่องที่มี GPU และทรัพยากรเพียงพอ (Self-Host).
6) **การสร้าง RAG แบบ No-Code ด้วย n8n**
    - **หลักการ**: ใช้ n8n เพื่อสร้าง workflow อัตโนมัติสำหรับ RAG โดยไม่ต้องเขียนโค้ด.
    - **แนวทางปฏิบัติ**:
        - ตั้งค่า n8n บนเซิร์ฟเวอร์หรือใช้เวอร์ชัน cloud.
        - สร้าง workflow ดังนี้:
            1. **Input Node**: รับคำถามจากผู้ใช้ (เช่น ผ่าน webhook หรือ form).
            2. **Embedding Node**: ใช้ HTTP Request เพื่อเรียก API ของ Hugging Face (BAAI/bge-m3) เพื่อสร้าง embedding.
            3. **Vector Search Node**: เชื่อมต่อกับ ChromaDB หรือ vector database อื่นเพื่อค้นหาข้อมูลที่เกี่ยวข้อง.
            4. **LLM Node**: ส่งบริบทและคำถามไปยัง LLM (เช่น OpenAI, Ollama) เพื่อสร้างคำตอบ.
            5. **Output Node**: ส่งคำตอบกลับไปยังผู้ใช้.
        - ใช้ n8n nodes เช่น HTTP Request, Merge, และ Function เพื่อจัดการข้อมูล

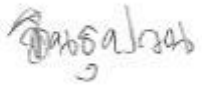
1. ประโยชน์ต่อการปฏิบัติงานในตำแหน่งหน้าที่

1) สามารถความรู้ที่ได้รับไปประยุกต์ใช้ในวิชา 10301374 การประมวลผลภาษาธรรมชาติ
2) สามารถความรู้ที่ได้รับไปประยุกต์ใช้ในวิชา 10301225 วิศวกรรมซอฟต์แวร์

2. ประโยชน์ต่อหน่วยงาน (ระดับงาน/หลักสูตร/คณะ)

1) สามารถนำองค์ความรู้จากการนำเสนอผลงานวิจัยไปใช้ในการปรับปรุงหลักสูตรสาขาวิชาวิทยาการคอมพิวเตอร์

จึงเรียนมาเพื่อโปรดทราบ



(นายสมนึก สินธุปวน)
22 สิงหาคม 2568

ความคิดเห็นของผู้บังคับบัญชาชั้นต้น (ประธานอาจารย์ผู้รับผิดชอบหลักสูตร/ผู้อำนวยการสำนักงาน/หัวหน้างาน)

บุคลากรดังกล่าวไปนำความรู้ไปใช้ประโยชน์ ดังนี้(โปรดระบุรายละเอียด)

.....

.....

.....

(อาจารย์อรรถวิท ชังคมานนท์)
ประธานอาจารย์ผู้รับผิดชอบหลักสูตร
สาขาวิชาวิทยาการคอมพิวเตอร์
...../...../.....

หมายเหตุ : 1. เอกสารแนบเช่น สำเนาบทคัดย่อ หรือโปสเตอร์(ย่อขนาด A4) หรือบทความฯ ฉบับเต็มสำเนาใบรับรองหรือหนังสือรับรองหรือใบประกาศนียบัตรหรือวุฒิบัตรฯลฯ ซึ่งเป็นหลักฐานว่าได้เข้าร่วมงานจริง

2. กรณีที่ประสงค์จะรายงานฯ กรณีไม่ได้งบพัฒนาบุคลากรหรือไม่ใช้งบประมาณ ให้ใช้แบบฟอร์มฯ นี้

3. ให้จัดรูปแบบและขยายพื้นที่ตามรายละเอียดเนื้อหาหรือข้อความ ตามความเหมาะสม