Toogle

- [x] ~~**หัวข้อ**~~
- [ ] Idea ของการนำ LLM models มาใช้

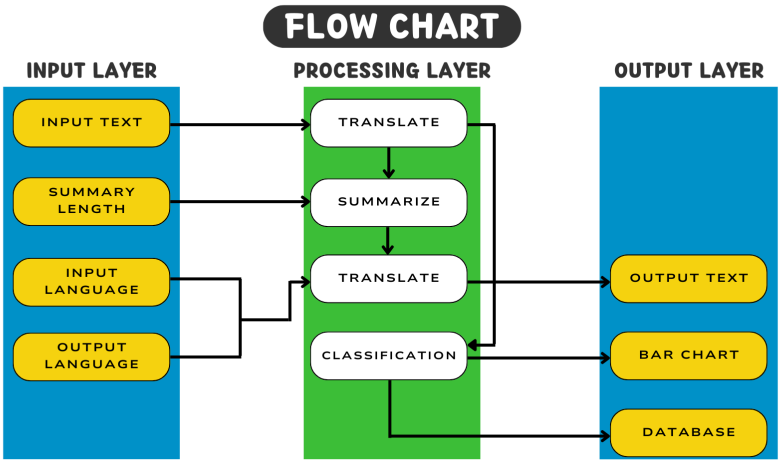
- [ ] **วิธีการรวบรวมข้อมูล** (python, web scraping, interview, questionnaire, etc)
- [x] ~~**วิธีการเก็บข้อมูล (database, csv, etc)**~~
- [ ] **ผลของการวิเคราะห์ข้อมูล และ นำเสนอข้อมูล** (Plotly and Dash)

TOOGLE: text summarizer

1. ที่มาและความสำคัญ

การนำเสนอเครื่องมือ Toogle เพื่อให้เห็นถึงความสำคัญของข้อมูลและการจัดการข้อมูลในยุคปัจจุบันมีที่มาและความสำคัญอย่างมาก โดยปัจจุบันข้อมูลมีบทบาทสำคัญมากขึ้นเนื่องจากข้อมูลเป็นแหล่งข้อมูลที่มีค่ามากสำหรับการตัดสินใจในธุรกิจและการวิเคราะห์ทางกายภาพและสังคม ดังนั้น กลุ่มของเราเล็งเห็นถึงความสำคัญของการวิเคราะห์ข้อมูลในสถานการณ์ที่ข้อมูลมีขนาดใหญ่มากและมีความซับซ้อน

เครื่องมือ Toogle คือการพัฒนา text summarizer ที่ใช้ในการคัดกรองข้อมูลอย่างสรุปและรวดเร็ว โดยสร้างเนื้อหาสรุปที่กระชับและเชื่อถือได้จากเอกสารหรือข้อความที่มีปริมาณมาก ซึ่งมีเป้าหมายในการลดความยุ่งยากและเพิ่มประสิทธิภาพในการทำงานของผู้วิเคราะห์ข้อมูล โดยการให้ผู้วิเคราะห์สามารถมุ่งเน้นข้อมูลสำคัญได้อย่างสะดวกสบายและรวดเร็ว ไม่ใช้เวลานานในการค้นหาข้อมูลที่สำคัญในการตัดสินใจและวิเคราะห์ข้อมูลให้เป็นไปอย่างมีประสิทธิภาพและมีคุณภาพในการจัดการข้อมูลในยุคปัจจุบัน



โมเดลที่ใช้ในการประมวลผลข้อมูล:

- 'facebook/bart-large-cnn' ใช้สำหรับการสร้างสรุปข้อความ
- 'facebook/mbart-large-50-many-to-many-mmt' ใช้สำหรับการแปลภาษา
- 'cardiffnlp/tweet-topic-21-multi' ใช้สำหรับการจำแนกประเภทข้อความ

2. วัตถุประสงค์

1) เพื่อให้เข้าใจและเรียนรู้การใช้งาน Dash, Plotly, Python ซึ่งเป็นเครื่องมือที่ใช้ในการสร้างและแสดงผลกราฟและข้อมูลอย่างกระชับและสวยงาม
2) เพื่อสร้างเครื่องมือที่สามารถสร้างเนื้อหาสรุปที่กระชับและเชื่อถือได้จากเอกสารหรือข้อความที่มีปริมาณมาก โดยทำให้ผู้ใช้สามารถลดความยุ่งยากและเพิ่มประสิทธิภาพในการใช้ประโยชน์จากข้อมูลที่สำคัญได้โดยมีการสร้างเนื้อหาสรุปที่มีคุณภาพและเชื่อถือได้ในขณะเดียวกัน

3. วิธีการรวบรวมข้อมูล

รวบรวมข้อมูลสำหรับการประมวลผลจากการที่ผู้ใช้กรอกข้อความที่ต้องการให้วิเคราะห์ผ่านช่องใส่ข้อความ (Textarea) บนหน้าเว็บไซต์ Dash โดยผู้ใช้สามารถเลือกภาษาของข้อความได้จากเมนูแบบ ดรอปดาวน์(Dropdown menu) ที่เตรียมไว้ หลังจากนั้น ระบบจะใช้โมดูล Langid เพื่อจำแนกและตรวจสอบภาษาของข้อความโดยอัตโนมัติ

4.วิธีการเก็บข้อมูล

ข้อมูลที่ได้รับจากการวิเคราะห์และประมวลผลข้อความจะถูกบันทึกลงในฐานข้อมูล PostgreSQL ด้วยการใช้ psycopg2 ซึ่งประกอบด้วยข้อมูลที่มีรายละเอียดดังนี้

- id: เป็น UUID ที่สร้างขึ้นใหม่ทุกครั้งที่มีการบันทึกข้อมูลเข้าฐานข้อมูล เพื่อระบุข้อมูลแต่ละแถวอย่างชัดเจน
- timestamp: บันทึกช่วงเวลาที่เกิดเหตุการณ์ขึ้น เพื่อวิเคราะห์ข้อมูลเกี่ยวกับเวลาที่ผู้ใช้เข้าถึงบ่อยที่สุด
- topic: บันทึกหัวข้อหรือประเภทของข้อมูลที่ถูกจำแนก ช่วยในการวิเคราะห์เพื่อทราบว่าผู้ใช้ส่วนใหญ่สนใจเนื้อหาประเภทใดมากที่สุด
- probability: บันทึกความน่าจะเป็นของข้อมูลที่ถูกจำแนกเข้าหมวดหมู่นั้น เพื่อให้เข้าใจถึงระดับความน่าจะเป็นของข้อมูลในแต่ละหัวข้อง่ายๆ

5.ผลที่คาดว่าจะได้รับและนำเสนอข้อมูล

ผลลัพธ์ที่คาดว่าจะได้รับจากการใช้เครื่องมือ Toogle ในการวิเคราะห์ข้อมูลคือการสร้างข้อมูลที่ถูกคัดสรรและสรุปอย่างมีนัยสำคัญ ที่ช่วยให้ผู้วิเคราะห์สามารถมุ่งเน้นไปที่ข้อมูลสำคัญได้อย่างรวดเร็วและแม่นยำ เพื่อเพิ่มความเข้าใจในข้อมูลและลดเวลาในการค้นหาข้อมูลที่สำคัญลงได้อย่างมีประสิทธิภาพเมื่อเทียบกับวิธีการคัดกรองข้อมูลแบบที่ไม่ใช้เครื่องมือช่วย ทั้งนี้รวมถึงการที่ผู้ใช้ Toogle มีระดับความพึงพอใจที่สูงในการใช้เครื่องมือสรุปโดยจะทดสอบผ่านการสำรวจผู้ใช้งานจริงเพื่อความน่าเชื่อถือและความพึงพอใจในการใช้งานสูงสุด

การนำเสนอข้อมูลนั้นจะมีการแสดงผล Topic & Probabilities ในรูปแบบกราฟหรือพล็อตเพื่อให้เข้าใจข้อมูลได้ง่ายขึ้น โดยใช้ Plotly เพื่อสร้างกราฟแท่งแนวนอน (horizontal bar chart) สำหรับแสดงความน่าจะเป็นของหัวข้อที่เกี่ยวข้องกับข้อความที่ผ่านการแยกประเภท

6. ข้อจำกัดและข้อเสนอแนะ

การเลือกใช้โมเดลจาก Hugging Face ช่วยเพิ่มประสิทธิภาพให้กับโปรเจกต์ในการสร้างสรุปข้อมูลได้อย่างมีประสิทธิภาพ แต่ก็มีข้อจำกัดบางอย่างที่ต้องพิจารณาดังนี้

1) ข้อจำกัดในการสรุปความ: โมเดล "facebook/bart-large-cnn" สามารถสรุปข้อความได้เฉพาะในภาษาอังกฤษเท่านั้น ซึ่งเป็นข้อจำกัดสำหรับผู้ใช้ที่ต้องการสรุปข้อความในภาษาอื่น ดังนั้นควรพิจารณาการใช้โมเดลที่รองรับหลายภาษาเพื่อเพิ่มความสามารถในการสรุปข้อความในภาษาที่ต้องการ
2) ข้อจำกัดในการแปลภาษา: โมเดล "facebook/mbart-large-50-many-to-many-mmt" มีข้อจำกัดในการแปลภาษาจากภาษาหนึ่งไปยังอีกภาษาหนึ่ง หากต้องการการแปลภาษาที่หลากหลาย ควรพิจารณาโมเดลที่มีความสามารถในการแปลภาษามากขึ้น
3) ข้อจำกัดทางด้านความยาวของเนื้อหาที่สรุป: ในการออกแบบระบบมีการจำกัดให้ผู้ใช้เลือกช่วงความยาวของเนื้อหาที่สรุปได้ตั้งแต่ 20-200 ตัวอักษร ซึ่งอาจทำให้สูญเสียข้อมูลสำคัญในบางกรณีที่ข้อความมีความยาวมากเกินไปหรือน้อยเกินไป แนะนำให้พิจารณาการปรับแต่งระบบโดยการเพิ่มความยืดหยุ่นในการกำหนดความยาวของเนื้อหาที่สรุป เพื่อให้ผู้ใช้จะสามารถปรับแต่งขนาดของเนื้อหาเพื่อให้ได้ข้อมูลที่ต้องการอย่างเหมาะสมและครอบคลุมเนื้อหาที่คาดหวังไว้ได้

นางสาววริสรา อินทแสง 6610412001
นางสาวชลิดา เอี่ยมลีลาพร 6610412002
นางสาวพรไพลิน ไทยสุริโย 6610412006
นางสาวมนสิชา ประดิษฐ์รอด 6610412007