เช็คความเหมือนของประโยคหรือข้อความ

Text similarity

จัดทำโดย

นางสาว พิมพ์ลักษณ์ สกุลเจริญกิจ 643020630-5

นาย พัชรพล พันทวี 683380431-8

รายงานฉบับนี้เป็นส่วนหนึ่งของรายวิชา SC351101 วิทยาการคำนวณ

ภาคเรียนที่ 2 ปีการศึกษา 2568

ภาควิชาวิทยาการคอมพิวเตอร์ วิทยาลัยการคอมพิวเตอร์

1. ความเป็นมาของโครงงาน

เทคโนโลยีสารสนเทศและอินเทอร์เน็ตมีบทบาทสำคัญต่อการสื่อสารและการจัดการข้อมูล โดยเฉพาะข้อมูลในรูปแบบข้อความ (Text Data) ซึ่งมีปริมาณเพิ่มขึ้นอย่างรวดเร็ว เช่น บทความ ข่าว งานวิชาการ รีวิวสินค้า และโพสต์บนสื่อสังคมออนไลน์ การจัดการและวิเคราะห์ข้อมูลข้อความจำนวนมากด้วยวิธีการแบบดั้งเดิมจึงทำได้ยากและใช้เวลามาก

หนึ่งในปัญหาที่พบได้บ่อยคือการตรวจสอบความคล้ายคลึงกันของข้อความ (Text Similarity) เช่น การตรวจสอบการคัดลอกผลงาน (Plagiarism Detection) การจัดกลุ่มเอกสารที่มีเนื้อหาใกล้เคียงกัน การค้นหาข้อมูลที่เกี่ยวข้อง หรือการเปรียบเทียบคำตอบที่มีความหมายคล้ายกัน แม้จะใช้ถ้อยคำแตกต่างกันก็ตาม หากไม่มีระบบอัตโนมัติ การวิเคราะห์ลักษณะนี้จะต้องอาศัยมนุษย์ซึ่งอาจเกิดความคลาดเคลื่อนและไม่สามารถรองรับข้อมูลจำนวนมากได้

ดังนั้น โครงงานนี้จึงมีวัตถุประสงค์เพื่อศึกษาและพัฒนาระบบวิเคราะห์ความคล้ายคลึงกันของข้อความ (Text Similarity) โดยอาศัยเทคนิคทางด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) และการเรียนรู้ของเครื่อง (Machine Learning) เพื่อให้ระบบสามารถเปรียบเทียบข้อความและคำนวณระดับความคล้ายคลึงได้อย่างมีประสิทธิภาพ โครงงานนี้จะช่วยลดเวลาและเพิ่มความแม่นยำในการวิเคราะห์ข้อความ และสามารถนำไปประยุกต์ใช้ในงานด้านการศึกษา ธุรกิจ และเทคโนโลยีสารสนเทศในอนาคต

2. วัตถุประสงค์ของโครงงาน

1. เพื่อพัฒนาโปรแกรมตรวจสอบความคล้ายคลึงของเอกสารข้อความ (Text Similarity) โดยใช้หลักการ Cosine Similarity

2. เพื่อศึกษาและประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ขั้นพื้นฐาน เช่น

- การตัดคำ
- การลบคำหยุด (Stop Words)
- การทำ Stemming อย่างง่าย

3. เพื่อสร้างระบบที่สามารถรับข้อมูลได้ทั้งจาก **ไฟล์ PDF และไฟล์ TXT** รวมถึงการป้อนข้อความด้วยตนเอง

4. เพื่อออกแบบส่วนติดต่อผู้ใช้ (GUI) ที่ใช้งานง่าย ด้วยภาษา Python และไลบรารี CustomTkinter

5. เพื่อแสดงผลลัพธ์ความคล้ายคลึงในรูปแบบเปอร์เซ็นต์ เพื่อช่วยประเมินการคัดลอกหรือความซ้ำของเอกสาร

3. ขอบเขตของโครงงาน

1. โครงงานรองรับการเปรียบเทียบเอกสาร **ครั้งละ 2 เอกสาร** เท่านั้น
2. ประเภทไฟล์ที่รองรับ ได้แก่
    - ไฟล์ข้อความ (.txt)
    - ไฟล์เอกสาร PDF (.pdf)
3. ใช้วิธีการวิเคราะห์ความคล้ายคลึงแบบ **Bag of Words** และ **Cosine Similarity**
4. มีการปรับปรุงข้อความก่อนประมวลผล โดย
    - แปลงตัวอักษรเป็นตัวพิมพ์เล็กทั้งหมด
    - ตัดเครื่องหมายวรรคตอน
    - ลบคำที่ไม่จำเป็น (Stop Words)
    - ทำ Stemming แบบง่าย
5. ระบบไม่รองรับ
    - การเปรียบเทียบเอกสารหลายไฟล์พร้อมกัน
    - การตรวจสอบฐานข้อมูลออนไลน์หรืออินเทอร์เน็ต
    - การประมวลผลภาษาไทย (ออกแบบสำหรับภาษาอังกฤษเป็นหลัก)
6. ผลลัพธ์จะแสดงเป็นเปอร์เซ็นต์ความคล้ายคลึง และใช้สีแสดงระดับความเสี่ยงของการคัดลอก

4. แผนการดำเนินงาน

ลำดับ	กิจกรรม	ระยะเวลา
1	กำหนดหัวข้อที่จะศึกษา	6 มกราคม 2569
2	วางแผนการทำงาน	6 มกราคม 2569
3	เริ่มเขียนโค้ด, เขียนโปรแกรมโดยใช้ภาษาpython	7 มกราคม 2569
4	รวบรวมข้อมูล	8 มกราคม 2569
5	นำเสนอ	
6	ส่งรายงานฉบับสมบูรณ์	

5. สรุปผลการดำเนินงาน

จากการพัฒนาโครงงาน **Plagiarism Checker System** พบว่าสามารถสร้างระบบที่ตรวจสอบความคล้ายคลึงของข้อความได้อย่างมีประสิทธิภาพในระดับพื้นฐาน โดยระบบสามารถอ่านข้อมูลจากไฟล์ PDF และ TXT ได้อย่างถูกต้อง และนำข้อความมาประมวลผลเพื่อคำนวณค่าความคล้ายคลึงด้วยวิธี Cosine Similarity

ผลลัพธ์ที่ได้สามารถแสดงเป็นเปอร์เซ็นต์ความคล้ายคลึงของเอกสารทั้งสองฉบับ พร้อมการแสดงสีเพื่อช่วยให้ผู้ใช้งานเข้าใจระดับความซ้ำของเอกสารได้ง่ายขึ้น ระบบมีส่วนติดต่อผู้ใช้ที่ใช้งานสะดวก ไม่ซับซ้อน และเหมาะสำหรับการใช้งานในเชิงการศึกษา เช่น การตรวจสอบงานรายงานหรืองานเขียนเบื้องต้น

อย่างไรก็ตาม ระบบยังมีข้อจำกัดในด้านความแม่นยำเมื่อเอกสารถูกเขียนใหม่ด้วยคำที่มีความหมายใกล้เคียงกัน (Paraphrase) และยังไม่รองรับภาษาไทยหรือการเชื่อมต่อกับแหล่งข้อมูลภายนอก ซึ่งสามารถนำไปพัฒนาต่อยอดในอนาคตได้