สอบ Midterm

Project-Based Examination

EDA – Scopus Data

จัดทำโดย

นาย กฤษณะ เทพาฤทธิ์ 66011211002

นาย ยุติธรรม ปั่นกลาง 66011211049

นางสาว นัฐกานต์ ทองเชื้อ 6601121074

นาย อดิศร บุญเหมาะ 66011211097

นางสาว ณัฐกานต์ ตาทอง 66011211162

รายงานฉบับนี้เป็นส่วนหนึ่งของการสอบกลางภาค (Midterm)

รายวิชา 1201333 การเรียนรู้เครื่องจักร

ภาคเรียนที่ 1 ปีการศึกษา 2569

มหาวิทยาลัยมหาสารคาม

1.ภาพรวมโครงงาน

โครงงานนี้มีวัตถุประสงค์เพื่อรวบรวม สำรวจ ทำความสะอาด วิเคราะห์ และเตรียมข้อมูลบทความวิจัยจากฐานข้อมูล Scopus ให้พร้อมสำหรับการนำไปวิเคราะห์ด้วย Machine Learning โดยดำเนินงานตั้งแต่การรวมไฟล์ CSV หลายชุด การทำ EDA และ Data Cleaning การสร้าง Visualization เพื่อทำความเข้าใจแนวโน้มของ Keyword และการสร้าง Text Embedding/Feature Vector สำหรับนำไปใช้ในขั้นตอนการเรียนรู้

2.วัตถุประสงค์

- สำรวจโครงสร้างและคุณภาพของข้อมูล Scopus ด้วยกระบวนการ Exploratory Data Analysis (EDA)
- ทำความสะอาดข้อมูล เช่น การลบข้อมูลซ้ำ การจัดการข้อความที่ไม่จำเป็น และค่าว่าง
- สร้าง Visualization เพื่ออธิบายความถี่ แนวโน้มตามเวลา การเติบโต และความสัมพันธ์ของ Keyword
- รวมข้อมูลข้อความจากงานวิจัยและแปลงเป็น Feature Vector ด้วยโมเดล SPECTER2
- เตรียมชุดข้อมูลให้สามารถนำไปใช้กับ Machine Learning เช่น Clustering หรือ Similarity Search ได้

3.ขอบเขตข้อมูล

แหล่งข้อมูล Scopus

จำนวน 10 Subject

จำนวนข้อมูลดิบ 6,000 รายการ

จำนวนหลังลบชื่อเรื่องซ้ำ 5,459 รายการ

ช่วงปี 2012–2027

4. อธิบายกระบวนการทำ EDA (Exploratory Data Analysis) และ Data Cleaning

กลุ่มได้รวบรวมข้อมูลจาก Scopus ทั้งสิ้น 10 ไฟล์ และใช้ภาษา Python ใน Google Colab ร่วมกับไลบรารี pandas และ glob เพื่อรวม สำรวจ และทำความสะอาดข้อมูลก่อนนำไปใช้งานในขั้นตอน Visualization และ Text Embedding

4.1 การรวมข้อมูล (Data Integration)

เริ่มต้นด้วยการค้นหาไฟล์ CSV ที่อยู่ในชุดข้อมูลทั้งหมด จากนั้นอ่านแต่ละไฟล์ด้วย pandas และนำมาต่อกันด้วยการ Concatenate แต่ละไฟล์มี 600 รายการ รวมข้อมูลดิบทั้งหมด 6,000 รายการ โดยในแต่ละไฟล์มี 22 คอลัมน์

ชุดข้อมูล	จำนวนรายการ	ช่วงปี	ไฟล์ export (ย่อ)
ชุดที่ 1	600	2017–2026	64b6db96
ชุดที่ 2	600	2012–2026	a7dc511a
ชุดที่ 3	600	2023–2027	73147acf
ชุดที่ 4	600	2026–2027	f945e7ad
ชุดที่ 5	600	2025–2027	513a647a
ชุดที่ 6	600	2026–2027	bd749209
ชุดที่ 7	600	2026–2027	2c0be3f7
ชุดที่ 8	600	2026–2027	0b559870
ชุดที่ 9	600	2027–2027	9b5c606f
ชุดที่ 10	600	2026–2027	17a933c5

4.2 การทำความสะอาดข้อมูล (Data Cleaning)

หลังจากสำรวจข้อมูลเบื้องต้น กลุ่มทำ Data Cleaning เพื่อให้ข้อมูลพร้อมสำหรับการรวมข้อความและสร้าง Embedding โดยดำเนินการดังนี้

4.2.1) การลบข้อมูลซ้ำ (Drop Duplicates) ตรวจสอบชื่อเรื่องจากคอลัมน์ Title และกำหนด keep=False เพื่อให้ทุกแถวที่อยู่ในกลุ่มชื่อเรื่องซ้ำถูกตัดออกทั้งหมด

4.2.2) การจัดการข้อความขยะ (Text Cleaning) ลบข้อความลิขสิทธิ์หรือประโยคมาตรฐานที่ซ้ำกันใน Abstract ซึ่งไม่ช่วยในการอธิบายเนื้อหาของงานวิจัย

4.2.3) การจัดการค่าว่าง (Missing Values) หลังการลบข้อมูลซ้ำ คอลัมน์ Author Keywords ยังมีค่าว่างบางส่วน และ Index Keywords มีค่าว่างบางส่วน จึงแทนค่า Null ด้วยช่องว่าง (Empty string) เพื่อป้องกัน Error ในขั้นตอนรวมข้อความ

4.2.4) การตรวจสอบผลหลัง Cleaning ใช้ info() และการนับค่าว่างอีกครั้ง เพื่อยืนยันว่าคอลัมน์ที่ต้องนำไปสร้าง embedding_text สามารถประมวลผลต่อได้

ขั้นตอน	จำนวนก่อน	จำนวนที่กระทบ	จำนวนหลัง
รวมไฟล์ CSV	-	-	6,000
ลบ Title ที่ซ้ำแบบ keep=False	6,000	541 แถว	5,459
fillna สำหรับ Text fields	5,459	Author Keywords 349 / Index Keywords 161	5,459

ผลหลัง Data Cleaning

จากข้อมูลดิบ 6,000 รายการ เมื่อใช้ Title เป็นเกณฑ์และ keep=False พบแถวที่อยู่ในกลุ่มชื่อเรื่องซ้ำ 541 แถว จึงเหลือข้อมูล 5,459 รายการสำหรับขั้นตอน Text Preparation และ Text Embedding

4.3 การบันทึกผลลัพธ์ (Data Export)

หลังจากเสร็จสิ้นกระบวนการ Data Cleaningกลุ่มบันทึกข้อมูลที่ผ่านการทำความสะอาดเป็นไฟล์ CSV ชื่อ scopus_cleaned_data_final.csv โดยไม่บันทึก Index ของ DataFrame เพื่อใช้เป็นชุดข้อมูลกลางสำหรับ Text Preparation, Text Embedding และการนำไปประมวลผลด้วย Machine Learning ต่อไป

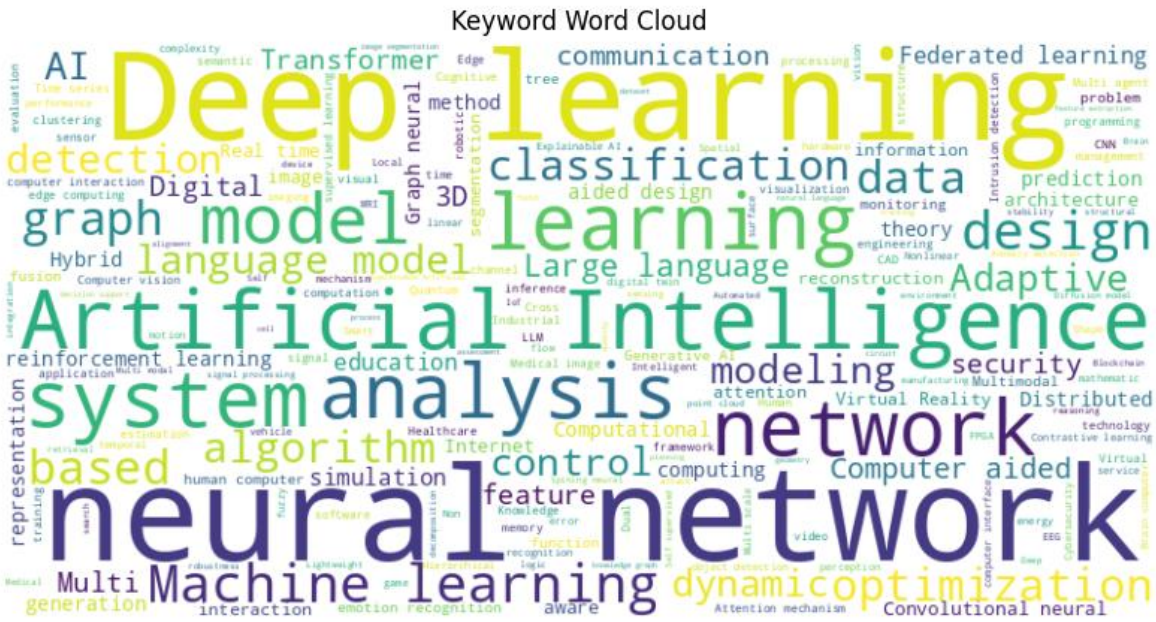
5. อธิบายการทำ Visualization

กลุ่มใช้ Visualization เพื่อทำความเข้าใจ Keyword ในหลายมิติ ได้แก่ ภาพรวมความถี่ สัดส่วนความนิยม แนวโน้มตามเวลา ความหนาแน่นรายปี อัตราการเติบโต และความสัมพันธ์ระหว่าง Keyword โดยมีทั้งกราฟที่สร้างจาก Python และกราฟอินเทอร์แอกทีฟจาก Keyword Analytics Dashboard

การระบุชุดข้อมูลที่ใช้ให้ชัดเจน

Word Cloud ใช้ข้อมูลข้อความที่ผ่านการเตรียม/ทำความสะอาด ส่วนกราฟ Dashboard 5 รูปแบบที่ปรากฏในรายงานเดิมมีค่าความถี่สอดคล้องกับข้อมูลรวมก่อนลบ Title ซ้ำจำนวน 6,000 รายการ จึงระบุแหล่งข้อมูลของกราฟ Dashboard เป็นชุดข้อมูลรวมก่อน deduplication เพื่อให้คำอธิบายตรงกับตัวเลขที่แสดงในกราฟจริง

5.1 ภาพรวมกลุ่มคำศัพท์ (Keyword Word Cloud)



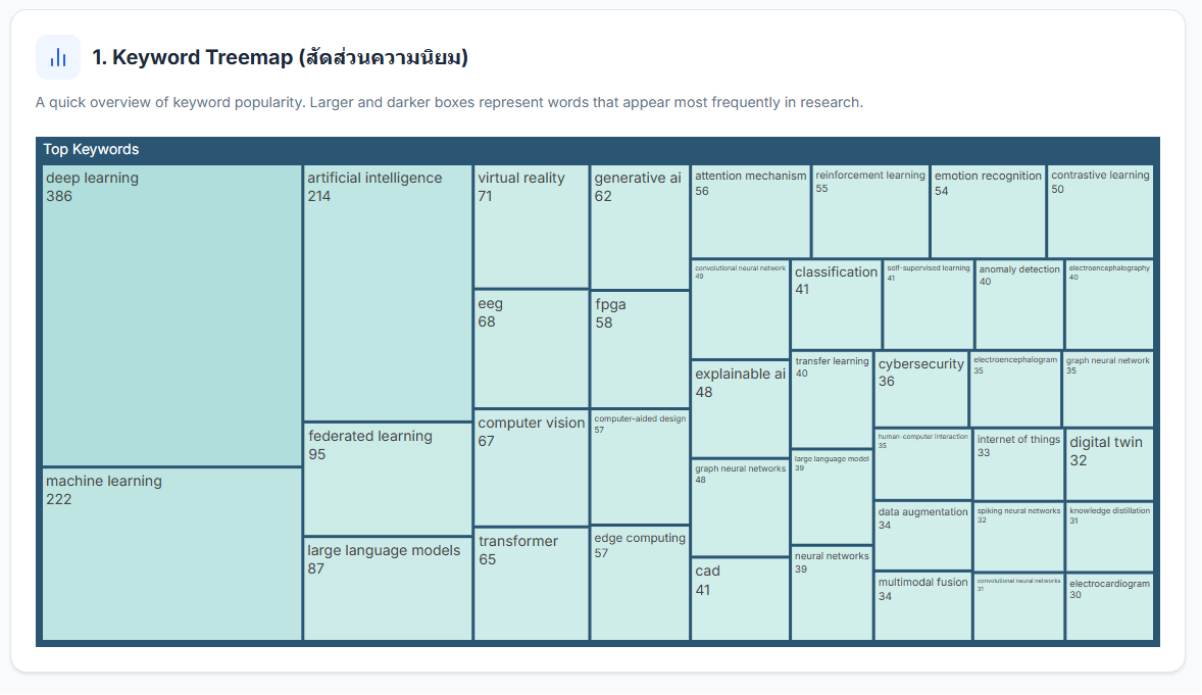
รูปที่ 1 Keyword Word Cloud จาก Author Keywords

- กราฟนี้เกี่ยวกับอะไร: Word Cloud นำข้อความจากคอลัมน์ Author Keywords มานับความถี่และแสดงเป็นคำ โดยคำที่ปรากฏบ่อยจะมีขนาดใหญ่กว่าคำที่พบไม่บ่อย
- ความสัมพันธ์กับข้อมูล: คำขนาดใหญ่ เช่น Deep Learning, Machine Learning, Artificial Intelligence และ Neural Network แสดงว่ากลุ่มหัวข้อด้าน AI และการเรียนรู้ของเครื่องเป็นแกนสำคัญของชุดข้อมูล
- ประโยชน์ในการวิเคราะห์: ช่วยให้เห็นภาพรวมของหัวข้อวิจัยได้รวดเร็ว ก่อนลงรายละเอียดด้วยกราฟเชิงปริมาณประเภทอื่น

5.2 การวิเคราะห์เชิงลึกผ่าน Keyword Analytics Dashboard

Dashboard ประกอบด้วยกราฟอินเทอร์แอกทีฟ 5 รูปแบบ โดยแต่ละกราฟตอบคำถามที่ต่างกัน ตั้งแต่คำใดได้รับความนิยมสูง เทรนด์เปลี่ยนไปตามเวลาอย่างไร คำใดกำลังเติบโต และ Keyword ใดถูกศึกษาเชื่อมโยงกัน

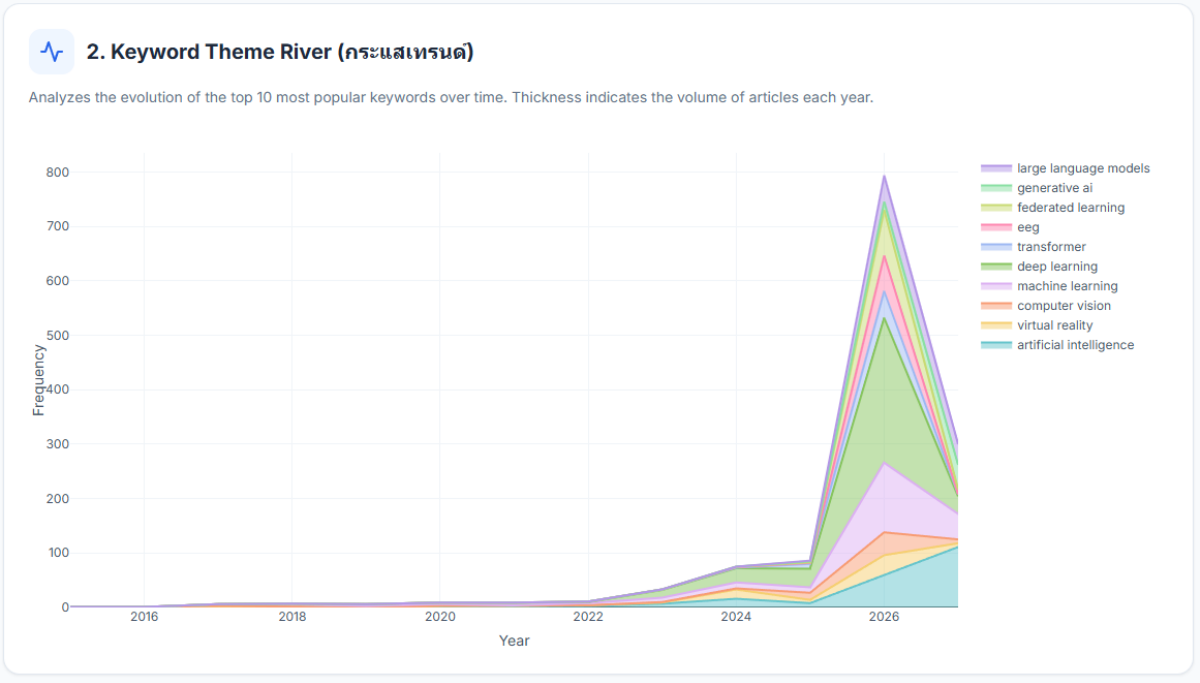
5.2.1 Keyword Treemap (สัดส่วนความนิยม)



รูปที่ 2 Keyword Treemap แสดงสัดส่วนความถี่ของ Keyword

- กราฟนี้เกี่ยวกับอะไร: Treemap แทน Keyword แต่ละคำด้วยสี่เหลี่ยม ขนาดพื้นที่ของสี่เหลี่ยมสัมพันธ์กับจำนวนครั้งที่ Keyword ปรากฏ ยิ่งพื้นที่ใหญ่ยิ่งพบคำนั้นบ่อย
- ความสัมพันธ์กับข้อมูล: จากกราฟ Deep Learning มีค่าความถี่ 386 ครั้ง, Machine Learning 222 ครั้ง และ Artificial Intelligence 214 ครั้ง จึงเป็นกลุ่มคำที่มีสัดส่วนเด่นที่สุดในข้อมูล Dashboard
- การตีความ: คำอย่าง Federated Learning, Large Language Models, Virtual Reality และ Generative AI มีขนาดรองลงมา สะท้อนหัวข้อเฉพาะทางและเทคโนโลยีที่กำลังได้รับความสนใจ

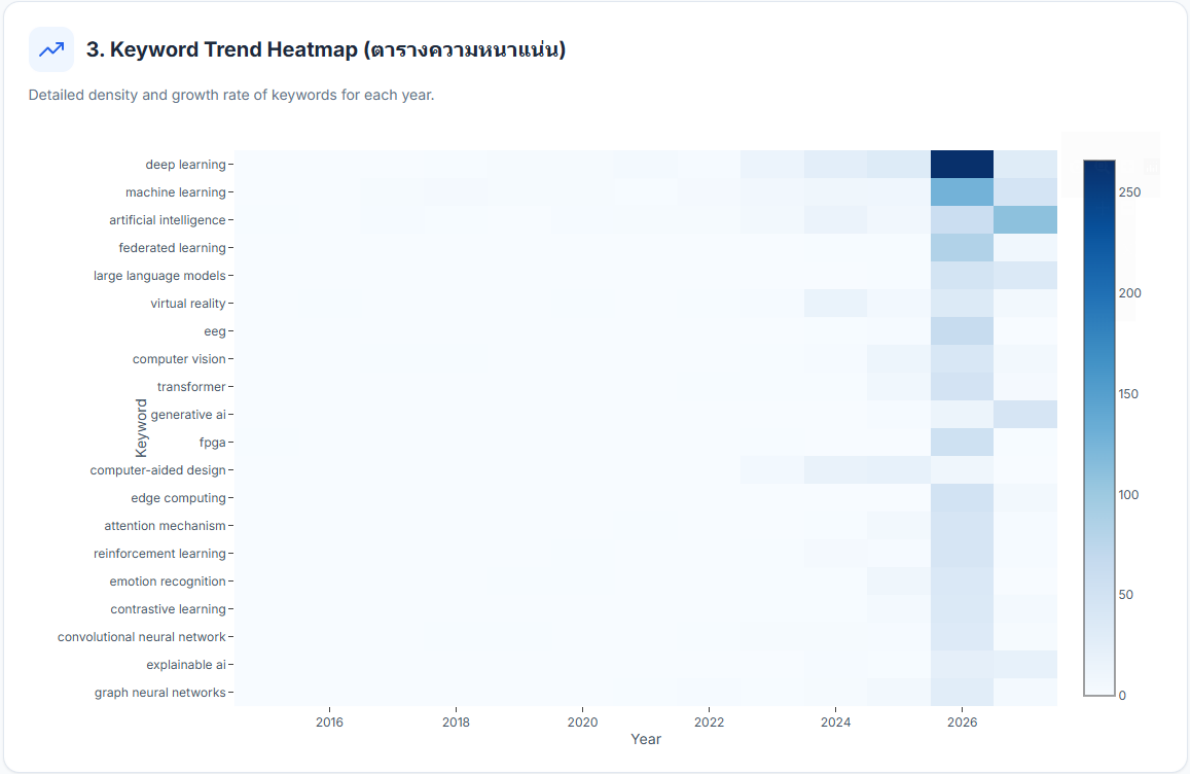
5.2.2 Keyword Theme River (กระแสเทรนด์)



รูปที่ 3 Keyword Theme River แสดงการเปลี่ยนแปลงของ Keyword ตามปี

- กราฟนี้เกี่ยวกับอะไร: Theme River แสดงวิวัฒนาการของ Keyword ที่ได้รับความนิยมตามช่วงเวลา แกน X คือปีที่ตีพิมพ์ และความหนาของแถบสีแสดงจำนวนครั้งที่ Keyword ปรากฏในแต่ละปี
- ความสัมพันธ์กับข้อมูล: ปริมาณงานวิจัยเพิ่มขึ้นชัดเจนในช่วงปีหลัง โดยเฉพาะปี 2026 ซึ่งเป็นปีที่มีจำนวนระเบียนในชุดข้อมูลสูงที่สุด
- การตีความ: Keyword เช่น Large Language Models, Generative AI และ Federated Learning มีพื้นที่เพิ่มขึ้นในช่วงปีหลัง แสดงถึงหัวข้อที่ได้รับความสนใจมากขึ้นในงานวิจัยสมัยใหม่
- ข้อควรระวัง: ความสูงของกราฟได้รับอิทธิพลจากจำนวนบทความในแต่ละปีด้วย จึงควรอ่านควบคู่กับจำนวนระเบียนรายปีจาก EDA

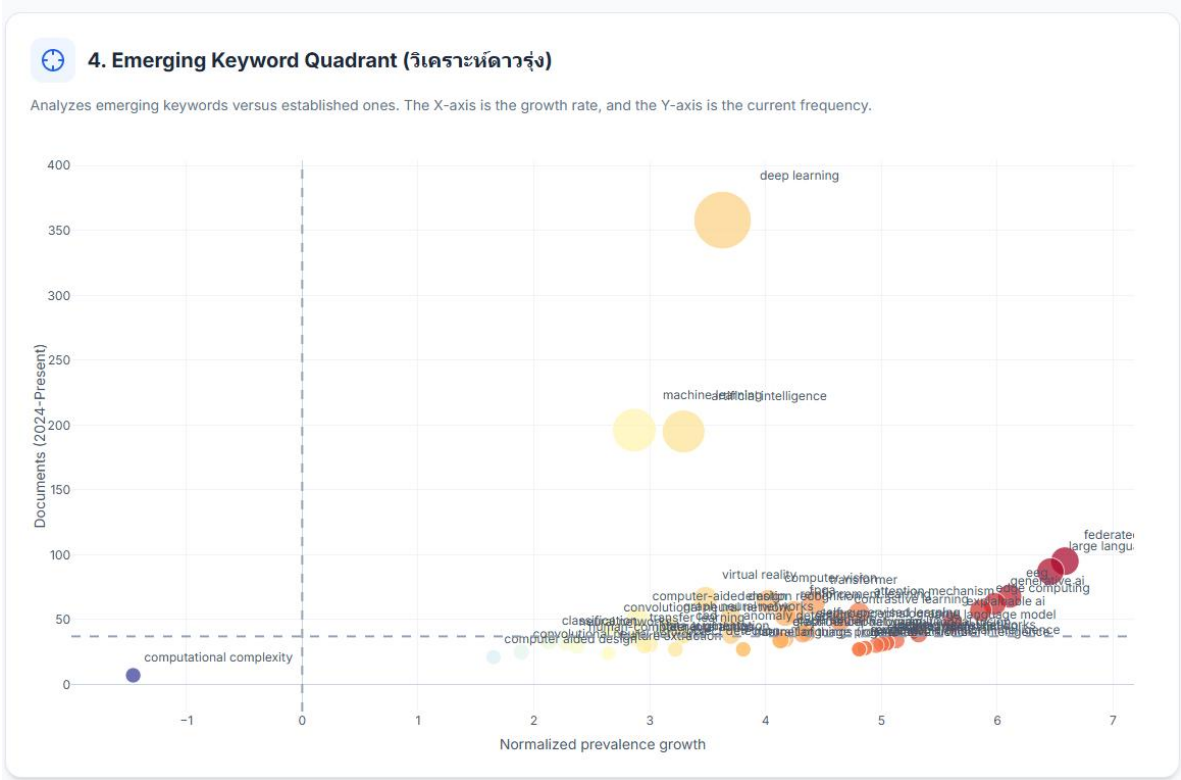
5.2.3 Keyword Trend Heatmap (ตารางความหนาแน่น)



รูปที่ 4 Keyword Trend Heatmap แสดงความหนาแน่นของ Keyword ในแต่ละปี

- กราฟนี้เกี่ยวกับอะไร: Heatmap ใช้ระดับความเข้มของสีแสดงจำนวนครั้งที่ Keyword ปรากฏในแต่ละปี โดยสีที่เข้มกว่าหมายถึงมีความถี่สูงกว่า
- ความสัมพันธ์กับข้อมูล: บริเวณปี 2026 มีความเข้มสูงที่สุดในหลาย Keyword โดยเฉพาะ Deep Learning และ Machine Learning ซึ่งสอดคล้องกับจำนวนระเบียนของปี 2026 ที่สูงในชุดข้อมูล
- การตีความ: Heatmap ช่วยเปรียบเทียบ Keyword หลายคำพร้อมกันได้สะดวก และช่วยยืนยันรูปแบบที่เห็นจาก Theme River ในลักษณะตาราง/เมทริกซ์
- ประโยชน์: เหมาะกับการหาว่า Keyword ใดเด่นในปีใด และช่วงเวลาใดมีความหนาแน่นของงานวิจัยสูง

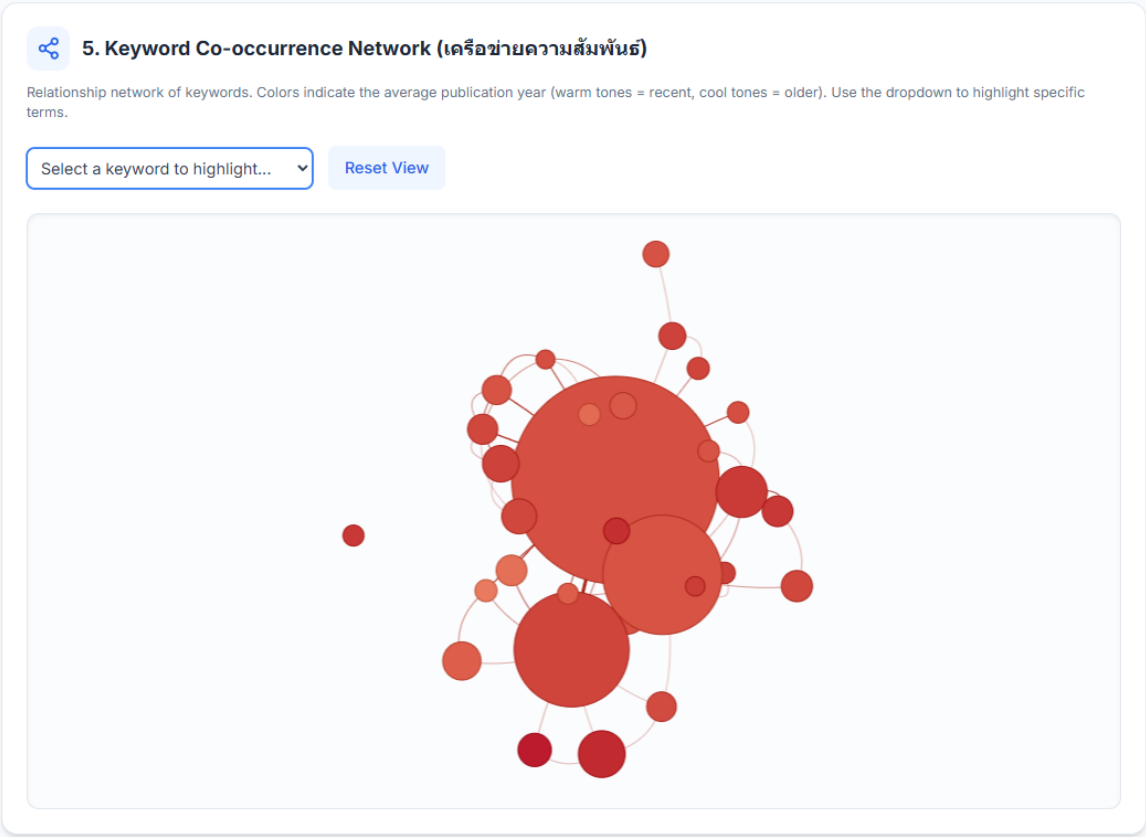
5.2.4 Emerging Keyword Quadrant (วิเคราะห์ดาวรุ่ง)



รูปที่ 5 Emerging Keyword Quadrant วิเคราะห์ความถี่ปัจจุบันและอัตราการเติบโต

- กราฟนี้เกี่ยวกับอะไร: เป็น Scatter Plot ที่วาง Keyword ตาม 2 มิติ คือแกน X แสดงอัตราการเติบโตของความนิยม และแกน Y แสดงความถี่ในช่วงปัจจุบัน
- ความสัมพันธ์กับข้อมูล: Deep Learning อยู่ในตำแหน่งที่มีความถี่สูง แสดงว่าเป็นหัวข้อขนาดใหญ่ของชุดข้อมูล ขณะที่ Federated Learning และคำที่อยู่ด้านขวามีอัตราการเติบโตสูงกว่า
- การตีความ: กราฟช่วยแยกคำที่ “นิยมอยู่แล้ว” ออกจากคำที่ “กำลังเติบโต” ทำให้มองเห็นหัวข้อที่อาจเป็นแนวโน้มสำคัญในอนาคต
- ข้อควรระวัง: Keyword ที่มีจำนวนเริ่มต้นน้อยอาจมีอัตราการเติบโตสูงได้ จึงควรพิจารณาทั้งแกน X และแกน Y ไม่ควรดู Growth rate เพียงอย่างเดียว

5.2.5 Keyword Co-occurrence Network (เครือข่ายความสัมพันธ์)



รูปที่ 6 Keyword Co-occurrence Network แสดงการปรากฏร่วมกันของคำสำคัญ

- กราฟนี้เกี่ยวกับอะไร: โหนด (Node) แทน Keyword ขนาดของโหนดสัมพันธ์กับความถี่ของคำนั้น และเส้นเชื่อม (Edge) แสดงว่า Keyword สองคำปรากฏร่วมกันในงานวิจัย
- ความสัมพันธ์กับข้อมูล: โหนดขนาดใหญ่บริเวณศูนย์กลางเป็นคำที่พบมากและเชื่อมโยงกับหลาย Keyword ส่วนโหนดที่เชื่อมกันเป็นกลุ่มสะท้อนหัวข้อที่มักถูกศึกษาในบริบทเดียวกัน
- สีของโหนด: Dashboard ใช้โทนสีเพื่อช่วยบอกความใหม่ของปีตีพิมพ์โดยเฉลี่ย ทำให้สามารถมองเห็นทั้งโครงสร้างความสัมพันธ์และมิติด้านเวลาได้พร้อมกัน
- ประโยชน์: ช่วยให้เห็นว่าการวิจัยด้าน AI ไม่ได้แยกเป็นหัวข้อเดี่ยว แต่มีการเชื่อมโยงระหว่าง Deep Learning, Machine Learning, Generative AI, Computer Vision และเทคโนโลยีอื่น

5.3 สรุปความสัมพันธ์ของ Visualization กับข้อมูล

จาก Visualization ทั้งหมดสามารถสรุปได้ว่า Deep Learning, Machine Learning และ Artificial Intelligence เป็นหัวข้อหลักที่มีความถี่สูง ขณะเดียวกันหัวข้ออย่าง Large Language Models, Generative AI และ Federated Learning แสดงสัญญาณความสนใจในช่วงปีหลัง ส่วน Co-occurrence Network

แสดงให้เห็นว่าหัวข้อเหล่านี้มีความสัมพันธ์และถูกนำมาศึกษาร่วมกันในงานวิจัยจำนวนมาก

6. สรุปข้อมูลที่จะนำไปทำ Text Embedding หรือ Feature Vector

เพื่อให้โมเดล Machine Learning สามารถเข้าใจเนื้อหาของงานวิจัยและนำไปประมวลผลต่อได้ กลุ่มของเราได้เตรียมข้อความและแปลงเป็น Feature Vector อย่างเป็นระบบ โดยมีรายละเอียดดังนี้

6.1 การคัดเลือกและเตรียมข้อมูลข้อความ (Text Preparation)

- **คอลัมน์ที่เลือกใช้งาน:** เราดึงข้อมูล 4 คอลัมน์ที่มีเนื้อหาสำคัญที่สุดของงานวิจัยแต่ละชิ้น ได้แก่ Title (ชื่อเรื่อง), Abstract (บทคัดย่อ), Author Keywords (คำสำคัญของผู้แต่ง), และ Index Keywords (คำสำคัญของฐานข้อมูล)
- **การรวมข้อความ (Concatenation):** กลุ่มของเราได้เขียนฟังก์ชันเพื่อนำข้อมูลทั้ง 4 ส่วนมารวมกันเป็นข้อความเดียว โดยเทคนิคสำคัญคือการใส่ **"คำนำหน้า (Prefix)"** ให้กับแต่ละส่วน (เช่น พิมพ์คำว่า Title: นำหน้าชื่อเรื่อง และ Abstract: นำหน้าบทคัดย่อ) เพื่อให้โมเดล AI สามารถแยกแยะบริบทของประโยคได้ว่ากำลังอ่านส่วนไหนของเปเปอร์อยู่ จากนั้นเก็บข้อความที่สมบูรณ์ไว้ในคอลัมน์ใหม่ชื่อ embedding_text

6.2 กระบวนการแปลงข้อความเป็นเวกเตอร์ (Text Embedding)

- **โมเดลที่เลือกใช้:** เราเลือกใช้โมเดล **SPECTER2** (allenai/specter2_base) ทำงานร่วมกับ Adapter ผ่านไลบรารี Transformers
- **เหตุผลที่ใช้ SPECTER2:** เป็นโมเดล AI ขั้นสูงที่ถูกฝึกสอน (Pre-trained) มาให้อ่านและทำความเข้าใจ "บทความวิจัยทางวิทยาศาสตร์และวิชาการ" โดยเฉพาะ ทำให้สามารถจับใจความสำคัญของเปเปอร์สายเทคโนโลยีได้ดีกว่าโมเดลภาษาทั่วไป
- **การสกัด Feature:** ข้อความจากคอลัมน์ embedding_text จะถูกนำไปหั่นเป็นคำๆ (Tokenization) และป้อนเข้าสู่โมเดลเพื่อคำนวณสกัดคุณลักษณะเด่น (Features) ออกมาเป็นชุดตัวเลขเวกเตอร์

6.3 การปรับบรรทัดฐานข้อมูล (Normalization)

- เวกเตอร์ที่สกัดออกมาได้ จะถูกนำไปผ่านกระบวนการ **L2 Normalization** เพื่อปรับให้เวกเตอร์ของเปเปอร์ทุกฉบับมีสเกล (ความยาว) อยู่ในมาตรฐานเดียวกันทั้งหมด ขั้นตอนนี้สำคัญมากเพราะจะช่วยให้อัลกอริทึม Machine Learning วัดระยะห่างหรือความคล้ายคลึงของเปเปอร์แต่ละฉบับได้อย่างแม่นยำ

6.4 ผลลัพธ์ที่พร้อมนำไปใช้งาน (Final Output)

- ผลลัพธ์สุดท้ายที่ได้คือ ชุดข้อมูล Feature Vector รูปแบบเมทริกซ์ขนาด **(5459, 768)** ซึ่งหมายความว่า เปเปอร์ทั้ง 5,459 ฉบับ ได้ถูกแปลงสภาพเป็นชุดตัวเลข 768 มิติ (768 Features) เรียบร้อยแล้ว
- ข้อมูลเวกเตอร์ชุดนี้พร้อมนำไปเป็น **ข้อมูลสอน (Train Data)** ให้กับอัลกอริทึม Machine Learning ในขั้นตอนต่อไปทันที เช่น นำไปใช้กับ K-Means เพื่อจัดกลุ่มงานวิจัยที่มีเรื่องราวคล้ายกัน (Clustering) หรือใช้กับ KNN เพื่อหาระยะห่างของเปเปอร์ที่มีเนื้อหาใกล้เคียงกันที่สุดได้

7. สรุปผลการดำเนินงาน

โครงงานนี้รวบรวมข้อมูลจาก Scopus จำนวน 10 ชุด ชุดละ 600 รายการ รวมทั้งหมด 6,000 รายการ จากนั้นทำ EDA เพื่อตรวจสอบโครงสร้าง ค่าว่าง ช่วงปี และประเภทเอกสาร ก่อนทำ Data Cleaning ด้วยการลบ Title ที่ซ้ำแบบ keep=False ทำให้เหลือข้อมูล 5,459 รายการสำหรับการเตรียมข้อความและ Text Embedding

ในส่วน Visualization กลุ่มใช้ Word Cloud และ Keyword Analytics Dashboard เพื่ออธิบายข้อมูลในหลายมิติ ตั้งแต่ความถี่ของ Keyword สัดส่วนความนิยม แนวโน้มตามปี ความหนาแน่น การเติบโต และความสัมพันธ์ระหว่างคำสำคัญ ผลโดยรวมชี้ให้เห็นว่า Deep Learning,

Machine Learning และ Artificial Intelligence เป็นหัวข้อหลักของชุดข้อมูล ขณะที่ Large Language Models, Generative AI และ Federated Learning เป็นหัวข้อที่มีบทบาทเด่นในช่วงปีหลัง

สุดท้าย กลุ่มรวม Title, Abstract, Author Keywords และ Index Keywords เป็น embedding_text และใช้โมเดล SPECTER2 เพื่อสร้าง Feature Vector ขนาด (5459, 768) พร้อมทำ L2 Normalization ทำให้ข้อมูลพร้อมสำหรับการนำไปใช้กับ Machine Learning ต่อไป

7.1 ข้อจำกัดและข้อควรระวัง

- CSV export ที่ได้รับไม่เก็บชื่อ Subject ที่ใช้เป็นตัวกรองค้นหาไว้ในแต่ละไฟล์ จึงต้องอ้างอิงชื่อ Subject จาก Search history หรือขั้นตอนการค้นหาของกลุ่มหากต้องการระบุชื่อครบทุก Subject
- การลบข้อมูลซ้ำแบบ keep=False จะลบทุกสำเนาของ Title ที่ซ้ำ ไม่ได้เก็บไว้หนึ่งรายการ จึงควรใช้วิธีนี้เฉพาะเมื่อกลุ่มต้องการตัดชื่อเรื่องที่ซ้ำทั้งหมดจริง
- Author Keywords และ Index Keywords มีค่าว่างบางส่วนก่อน fillna จึงทำให้บทความบางฉบับมีข้อมูล Keyword น้อยกว่าบทความอื่น
- จำนวนบทความในแต่ละปีไม่เท่ากัน โดยเฉพาะปี 2026 มีจำนวนสูงมาก ดังนั้นกราฟแนวโน้มควรตีความร่วมกับจำนวนระเบียนรายปี
- กราฟ Dashboard เดิมใช้ข้อมูลรวมก่อน deduplication ขณะที่ Text Embedding ใช้ข้อมูลหลัง Cleaning จึงต้องระบุชุดข้อมูลของแต่ละผลลัพธ์ให้ชัดเจนเพื่อไม่ให้เกิดความสับสน

7.2 ไฟล์ผลลัพธ์ที่ใช้ในการส่งงาน

ลิงก์ GitHub

https://github.com/topno2547/EDA-Scopus-Data-Midterm.git

เว็บไซต์ Visualization

https://eonstar1.github.io/Scopus-Data-Visualization/