

สรุปเนื้อหาเกี่ยวกับการวิเคราะห์ความคิดเห็นจากไฟล์ "Introduction to Sentiment Analysis" มีดังนี้:

1. การวิเคราะห์ความคิดเห็นคืออะไร? What is sentiment analysis? หน้า 2

- การวิเคราะห์ความคิดเห็น (Sentiment Analysis) คือกระบวนการตรวจสอบและทำความเข้าใจความคิดเห็นหรืออารมณ์ของผู้เขียนต่อหัวข้อหนึ่งๆ ไม่ว่าจะเป็นเชิงบวก กลาง หรือเชิงลบ โดยอารมณ์ยังสามารถแบ่งออกเป็นเชิงคุณภาพหรือเชิงปริมาณได้
- **ความหมาย:** การวิเคราะห์ความคิดเห็น (Sentiment Analysis) หรือที่เรียกว่า “การทำเหมืองความคิดเห็น” คือกระบวนการตรวจสอบและทำความเข้าใจว่าผู้เขียนมีความรู้สึกหรือความคิดเห็นอย่างไรต่อหัวข้อที่กำลังพูดถึง เช่น เชิงบวก เชิงลบ หรือกลาง
- **การใช้งาน:** ใช้ในการสแกนข้อความในสื่อสังคมออนไลน์ บทวิจารณ์ผลิตภัณฑ์ บล็อก หรือข่าวสาร เพื่อวิเคราะห์ว่าคนส่วนใหญ่มีความเห็นอย่างไร
- **ความสำคัญ:** ช่วยให้ธุรกิจสามารถติดตามความรู้สึกของลูกค้าและตอบสนองได้อย่างรวดเร็ว
- **ข้อดี:** สามารถวิเคราะห์ข้อมูลจำนวนมากได้รวดเร็วและมีประสิทธิภาพ
- **ข้อเสีย:** การวิเคราะห์อาจผิดพลาดเมื่อพบข้อความที่มีการประชดหรือข้อความที่ใช้คำหลายความหมาย
- **ฟังก์ชันใน Python:** ฟังก์ชันอย่าง TextBlob สามารถช่วยวิเคราะห์ความคิดเห็นได้ โดยคืนค่าความรู้สึก (polarity) และความเป็นอัตวิสัย (subjectivity)

2. องค์ประกอบของระบบวิเคราะห์ความคิดเห็น: หน้า 3-4

- **ความคิดเห็น/อารมณ์:** ความคิดเห็น (polarity) ว่าสิ่งนั้นเป็นบวก กลาง หรือ ลบ อารมณ์ก็อาจเป็นเชิงคุณภาพ (เช่น ความสุข ความประหลาดใจ ความโกรธ) หรือเชิงปริมาณ (เช่น คะแนนที่ให้หนังจาก 1 ถึง 10)
 - **หัวข้อที่พูดถึง:** สิ่งที่คุณเขียนแสดงความคิดเห็นเกี่ยวกับ เช่น หนังสือ ภาพยนตร์ หรือสินค้า
 - **ผู้ถือความคิดเห็น:** ผู้ที่แสดงความคิดเห็นอาจเป็นบุคคลหรือองค์กร
- **องค์ประกอบ:**
 1. ความคิดเห็นหรืออารมณ์ (Opinion/Emotion): สามารถเป็นเชิงบวก กลาง หรือลบ
 2. หัวข้อ (Subject): สิ่งที่ถูกพูดถึง เช่น สินค้า ภาพยนตร์
 3. ผู้แสดงความคิดเห็น (Opinion Holder): บุคคลหรือองค์กรที่เป็นเจ้าของความคิดเห็น
- **ความสำคัญ:** การระบุทั้งความคิดเห็นและอารมณ์ช่วยให้เข้าใจเนื้อหาเชิงลึกมากขึ้น
- **ข้อดี:** สามารถให้ข้อมูลเชิงลึกเกี่ยวกับสิ่งที่ผู้คนชื่นชอบหรือไม่ชอบได้อย่างละเอียด
- **ข้อเสีย:** อาจเจอปัญหาเมื่อข้อมูลไม่ชัดเจนหรือคลุมเครือ
- **ฟังก์ชันใน Python:** สามารถใช้ TextBlob และ VADER เพื่อตรวจสอบอารมณ์และความคิดเห็นได้

3. เหตุผลในการวิเคราะห์ความคิดเห็น: หน้า 5

- การวิเคราะห์ความคิดเห็นมีการใช้งานในหลายด้าน เช่น การติดตามสื่อสังคมออนไลน์ การวิเคราะห์แบรนด์ ช่วยให้เข้าใจว่าลูกค้ามีปฏิสัมพันธ์กับแบรนด์อย่างไร หรือสิ่งที่ลูกค้าพอใจหรือไม่พอใจ
- **การใช้งาน:** เหมาะสำหรับการตรวจสอบสื่อสังคมออนไลน์และการวิเคราะห์แบรนด์ ช่วยให้รู้ว่าผู้คนพูดถึงแบรนด์หรือผลิตภัณฑ์อย่างไร
- **ข้อดี:** ทำให้สามารถตอบสนองต่อความคิดเห็นของลูกค้าได้อย่างรวดเร็ว เช่น การปรับปรุงบริการ
- **ข้อเสีย:** การวิเคราะห์อัตโนมัติอาจให้ผลลัพธ์ที่ไม่ตรงกับความเป็นจริงหากข้อมูลมีความซับซ้อนหรือการประชดประชัน
- **ตัวอย่าง:** การติดตามความคิดเห็นเกี่ยวกับสินค้าบน Twitter

4. ตัวอย่างการวิเคราะห์ความคิดเห็นจากรีวิวภาพยนตร์: หน้า 6

- ในการวิเคราะห์ความคิดเห็นสามารถใช้ชุดข้อมูลเช่น รีวิวภาพยนตร์จาก IMDB ที่แบ่งเป็นคอลัมน์ข้อความรีวิว และคอลัมน์ "label" ซึ่ง 1 หมายถึงบวก และ 0 หมายถึงลบ
- **การใช้งาน:** การวิเคราะห์ชุดข้อมูลรีวิวภาพยนตร์โดยแบ่งความคิดเห็นเป็นบวกและลบ โดยใช้การคำนวณสัดส่วนของรีวิวทั้งสองแบบ
- **ความสำคัญ:** ช่วยให้รู้ว่าภาพยนตร์ได้รับความนิยมในเชิงบวกหรือลบมากแค่ไหน
- **ข้อดี:** สามารถประมวลผลข้อมูลรีวิวจำนวนมากได้อย่างรวดเร็ว
- **ข้อเสีย:** อาจเกิดความผิดพลาดถ้ารีวิวมีการใช้ภาษาประชดประชันหรือมีความคิดเห็นหลายด้าน
- **ฟังก์ชันใน Python:** ใช้ pandas เพื่ออ่านข้อมูลและคำนวณสัดส่วน

5. ประเภทของการวิเคราะห์ความคิดเห็น: หน้า 15

- **การใช้งาน:** Document Level: การพิจารณาความคิดเห็นทั้งเอกสาร

☐ ความหมาย:

- การวิเคราะห์ความคิดเห็นในระดับเอกสาร (Document Level) เป็นการพิจารณาความคิดเห็นของผู้เขียนต่อเอกสารทั้งหมด เช่น การรีวิวภาพยนตร์ทั้งเรื่อง การรีวิวสินค้าทั้งชิ้น หรือความคิดเห็นต่อบริการของบริษัท

☐ ความสำคัญ:

- มีความสำคัญในการสรุปความรู้สึกหรือความคิดเห็นรวมของเอกสารนั้น ๆ โดยไม่ต้องเจาะลึกในระดับประโยคหรือแง่มุม ช่วยให้เข้าใจว่าความคิดเห็นโดยรวมของผู้เขียนต่อเรื่องนั้นเป็นบวก ลบ หรือกลาง

☐ ประโยชน์/ข้อดี:

- **ประหยัดเวลา:** สามารถวิเคราะห์ข้อความยาว ๆ ได้รวดเร็ว โดยดูเพียงความคิดเห็นโดยรวมจากเอกสาร
- **ง่ายต่อการตีความ:** ช่วยให้ผู้อ่านเข้าใจความคิดเห็นหลักจากเอกสารโดยไม่ต้องลงรายละเอียดในแต่ละประโยค
- **เหมาะสำหรับงานที่ไม่ต้องการความละเอียดสูง:** เช่น การวิเคราะห์รีวิวภาพยนตร์หรือรีวิวผลิตภัณฑ์โดยรวม

☐ ข้อเสีย:

- ขาดความแม่นยำในรายละเอียด: ไม่สามารถวิเคราะห์ความรู้สึกในแต่ละประโยคหรือแง่มุมได้ ดังนั้นอาจไม่แสดงความคิดเห็นที่มีหลายมุมมอง เช่น ในเอกสารที่มีทั้งส่วนที่ดีและไม่ดี
- ไม่สามารถระบุจุดเด่นและจุดด้อยของสินค้า/บริการได้ชัดเจน: เนื่องจากการเป็นารดูภาพรวม
 - Sentence Level: การพิจารณาว่าความคิดเห็นในแต่ละประโยคเป็นบวก ลบ หรือกลาง

☐ ความหมาย:

- การวิเคราะห์ความคิดเห็นระดับประโยค (Sentence Level) คือการพิจารณาความรู้สึกในแต่ละประโยค เช่น ประโยคนี้เป็นเชิงบวก ลบ หรือกลาง ซึ่งจะช่วยให้เข้าใจว่าแต่ละส่วนของข้อความมีความหมายและอารมณ์อย่างไร

☐ ความสำคัญ:

- ช่วยให้สามารถเข้าใจรายละเอียดของข้อความได้มากขึ้น เช่น ถ้าผู้ใช้รีวิวสินค้าที่มีทั้งข้อดีและข้อเสีย การวิเคราะห์ระดับประโยคจะช่วยให้เห็นว่าแต่ละประโยคสื่อถึงอะไร
- มีความสำคัญเมื่อข้อมูลมีความหลากหลายในแต่ละประโยค เช่น การรีวิวสินค้าที่มีหลายความคิดเห็นในข้อความเดียว

☐ ประโยชน์/ข้อดี:

- ให้อารยะเอียดที่ลึกกว่า: สามารถเจาะลึกและวิเคราะห์แต่ละประโยคได้ว่ามีความหมายเชิงบวก ลบ หรือกลาง
- เพิ่มความแม่นยำ: ช่วยให้การวิเคราะห์ความคิดเห็นของผู้ใช้แม่นยำขึ้น โดยเฉพาะในข้อความที่ซับซ้อนหรือมีหลายแง่มุม

☐ ข้อเสีย:

- ต้องใช้ทรัพยากรมากขึ้น: การวิเคราะห์แต่ละประโยคต้องการประมวลผลที่มากกว่า เนื่องจากต้องพิจารณาแยกจากกัน
- อาจยุ่งยากเมื่อประโยคไม่ชัดเจน: ข้อความบางประโยคอาจมีความคลุมเครือหรือใช้ภาษาประชด ทำให้การวิเคราะห์ไม่แม่นยำ
 - Aspect Level: การแยกวิเคราะห์ความเห็นในแต่ละแง่มุมของสินค้า เช่น กล้องดี แต่แบตเตอรี่แย่

☐ ความหมาย:

- การวิเคราะห์ความคิดเห็นในระดับแง่มุม (Aspect Level) คือการเจาะลึกความคิดเห็นในแง่มุมหรือคุณสมบัติเฉพาะของสินค้า เช่น กล้องของโทรศัพท์ดี แต่แบตเตอรี่ไม่ดี

☐ **ความสำคัญ:**

- สำคัญมากสำหรับการวิเคราะห์ความคิดเห็นที่มีหลายแง่มุม เช่น การรีวิวผลิตภัณฑ์ที่มีคุณสมบัติหลากหลาย การวิเคราะห์ระดับนี้จะช่วยให้เข้าใจผู้ใช้ชอบหรือตำหนิในจุดใดของผลิตภัณฑ์
- ช่วยให้ผู้พัฒนาหรือผู้ผลิตสินค้าสามารถปรับปรุงผลิตภัณฑ์ในจุดที่ลูกค้าตำหนิได้อย่างตรงจุด

☐ **ประโยชน์/ข้อดี:**

- สามารถระบุจุดเด่นและจุดด้อยได้ชัดเจน: ช่วยให้รู้ว่าลูกค้าชอบหรือไม่ชอบคุณสมบัติใดของสินค้า เช่น ลูกค้าชอบกลิ่น แต่ไม่ชอบแบตเตอรี่
- เหมาะสำหรับการวิเคราะห์รีวิวสินค้าและบริการที่มีหลายคุณสมบัติ: เช่น โทรศัพท์, รถยนต์, โรงแรม เป็นต้น

☐ **ข้อเสีย:**

- ซับซ้อนและใช้เวลา: การวิเคราะห์ในแต่ละแง่มุมต้องใช้เวลามากกว่า เนื่องจากต้องแยกประโยคและพิจารณาแต่ละแง่มุมในรายละเอียด
- ต้องมีการตั้งค่าสำหรับการวิเคราะห์เฉพาะด้าน: เช่น ต้องรู้ว่าจะแยกแง่มุมไหนออกมาเป็นหัวข้อการวิเคราะห์

6. อัลกอริทึมการวิเคราะห์ความคิดเห็น:

- **แบบใช้กฎ/พจนานุกรม:** ใช้คำที่กำหนดคะแนนไว้ล่วงหน้า เช่น "ดี" ได้ +1 "แย่" ได้ -3 เพื่อคำนวณคะแนนรวมของความคิดเห็น
- **ระบบอัตโนมัติ/การเรียนรู้ของเครื่อง:** ใช้ข้อมูลที่มีป้ายกำกับในอดีตเพื่อคาดการณ์ความคิดเห็นของข้อความใหม่

☐ **การใช้งาน:** มีทั้งแบบใช้กฎ (Rule-based) และการเรียนรู้ของเครื่อง (Machine Learning)

1. **Rule-based:** ใช้พจนานุกรมของคำพร้อมคะแนน เช่น "ดี" ได้ +1 "แย่" ได้ -1
2. **Machine Learning:** ใช้ข้อมูลในอดีตเพื่อฝึกโมเดล

☐ **ข้อดีของ Rule-based:** ง่ายและรวดเร็ว

☐ **ข้อเสียของ Rule-based:** ต้องใช้พจนานุกรมที่ครอบคลุมทุกสถานการณ์

☐ **ข้อดีของ Machine Learning:** แม่นยำเมื่อมีข้อมูลมากพอ

☐ **ข้อเสียของ Machine Learning:** ต้องใช้เวลาฝึกโมเดล

☐ **ฟังก์ชันใน Python:** TextBlob และ VADER เป็นเครื่องมือสำคัญในการทำงานทั้งสองแบบ

7. การคำนวณคะแนน Valence ของประโยค:

- การคำนวณ valence คือการวัดว่าประโยคมีความรู้สึกเชิงบวกหรือลบอย่างไร โดยใช้เครื่องมืออย่าง TextBlob ค่าจะอยู่ระหว่าง -1 (ลบมาก) ถึง +1 (บวกมาก)

- ☐ **ความหมาย:** คะแนน Valence คือการคำนวณความรู้สึกในข้อความ โดยมีค่าตั้งแต่ -1 ถึง +1 (เชิงลบถึงเชิงบวก)
- ☐ **การใช้งาน:** ใช้ในการประเมินความรู้สึกเชิงบวก ลบ หรือกลางของประโยค
- ☐ **ข้อดี:** ให้คะแนนเชิงปริมาณทำให้วัดได้ชัดเจน
- ☐ **ข้อเสีย:** การวัดความรู้สึกที่มีความหมายหลายด้านอาจยาก
- ☐ **ฟังก์ชันใน Python:** ใช้ TextBlob เพื่อตรวจสอบคะแนน valence

8. การคำนวณคะแนนความรู้สึก (Sentiment Polarity Scores):

- ใช้ VADER lexicon เพื่อกำหนดคะแนนความรู้สึกให้กับข้อความ โดยคะแนนจะอยู่ในช่วง -1 ถึง +1 เพื่อบ่งบอกความรุนแรงของความรู้สึกเชิงลบหรือเชิงบวก

- ☐ **การใช้งาน:** VADER เป็นเครื่องมือสำหรับวิเคราะห์ความคิดเห็น โดยจะให้คะแนนแต่ละประโยคเป็นค่าระหว่าง -1 ถึง +1
- ☐ **ข้อดี:** เหมาะสำหรับการวิเคราะห์ข้อความสั้นๆ เช่น โพสต์ในสื่อสังคมออนไลน์
- ☐ **ข้อเสีย:** อาจไม่เหมาะกับข้อความที่ซับซ้อนมาก
- ☐ **ฟังก์ชันใน Python:** nltk.sentiment.vader.SentimentIntensityAnalyzer สามารถวิเคราะห์และให้คะแนนได้

9. ไลบรารีใน Python ที่ใช้ในการวิเคราะห์ความคิดเห็น:

- ไลบรารีที่ใช้ ได้แก่ TextBlob, VADER, pandas, re, matplotlib, nltk ที่มีเครื่องมือสำหรับการประมวลผลและแสดงผลข้อมูลความคิดเห็น

ในการวิเคราะห์ความคิดเห็น (Sentiment Analysis) มีไลบรารีหลายตัวที่ถูกใช้ใน Python เพื่อช่วยในการประมวลผลข้อความและวิเคราะห์ความคิดเห็นเชิงบวก เชิงลบ หรือกลาง โดยไลบรารีหลักที่ใช้มีดังนี้:

TextBlob:

- **ความหมาย:** TextBlob เป็นไลบรารีที่ใช้งานง่ายสำหรับการประมวลผลภาษาธรรมชาติ (NLP) ช่วยในการวิเคราะห์ความคิดเห็นได้โดยตรง และสามารถคำนวณค่า **polarity** (เชิงบวก-ลบ) และ **subjectivity** (อัตวิสัย) ของประโยคได้
- **ความหมาย:** TextBlob เป็นไลบรารีสำหรับการประมวลผลภาษาธรรมชาติ (NLP) ที่ใช้งานง่าย มันสามารถทำงานกับข้อความที่เป็นประโยคหรือเอกสารเพื่อประเมินความรู้สึก (Sentiment) โดยใช้พจนานุกรมที่มีอยู่แล้ว
- **ฟังก์ชันหลัก:** sentiment ซึ่งสามารถคืนค่าความรู้สึก (polarity) และความเป็นอัตวิสัย (subjectivity)
- **ข้อดี:** ใช้งานง่ายและสามารถทำงานได้ทันทีโดยไม่ต้องเตรียมข้อมูลมาก
 - ใช้งานง่ายและไม่ต้องเตรียมข้อมูลมาก
 - เหมาะสำหรับการวิเคราะห์ข้อความที่เป็นภาษาอังกฤษ
- **ข้อเสีย:** อาจไม่สามารถจัดการกับข้อความที่ซับซ้อนได้ดีเท่ากับโมเดลขั้นสูง
 - ประสิทธิภาพอาจต่ำกว่าวิธีการที่ใช้ Machine Learning ในการวิเคราะห์ความคิดเห็นที่ซับซ้อน

การใช้งาน:

- **การวิเคราะห์ความรู้สึก:** TextBlob สามารถวิเคราะห์ความคิดเห็นเชิงบวก, เชิงลบ หรือกลาง โดยใช้ฟังก์ชัน sentiment.polarity ที่ให้ค่าเป็นตัวเลขระหว่าง -1 ถึง +1
- **การแยกคำ (Tokenization):** TextBlob มีฟังก์ชันแยกคำจากประโยคให้อัตโนมัติ เช่น words สำหรับแยกคำและ sentences สำหรับแยกประโยค
- **การแก้ไขไวยากรณ์:** TextBlob สามารถตรวจสอบและแก้ไขการสะกดคำผิดในข้อความได้ด้วยฟังก์ชัน correct()

ความสำคัญ:

- TextBlob ช่วยให้ผู้ใช้สามารถประมวลผลความคิดเห็นจากข้อความได้อย่างง่ายดายโดยไม่ต้องใช้ข้อมูลขนาดใหญ่หรือโมเดลที่ซับซ้อน ด้วยการใช้งานที่ตรงไปตรงมา TextBlob เหมาะสำหรับการวิเคราะห์เบื้องต้นหรือการประเมินความรู้สึกของข้อความในระดับเอกสารหรือประโยค
- มีความสำคัญมากสำหรับผู้เริ่มต้นใช้งาน NLP เพราะช่วยลดความซับซ้อนในการเริ่มต้นประมวลผลภาษาธรรมชาติ

VADER (Valence Aware Dictionary and Sentiment Reasoner):

- **ความหมาย:** VADER เป็นเครื่องมือวิเคราะห์ความคิดเห็นแบบพจนานุกรมและกฎ (lexicon and rule-based) โดยเน้นการวิเคราะห์ความรู้สึกในข้อความที่พบได้ทั่วไป เช่น โพสต์สื่อสังคมออนไลน์หรือบทวิจารณ์
- **ความหมาย:** VADER เป็นเครื่องมือวิเคราะห์ความคิดเห็นที่ออกแบบมาสำหรับข้อความในสื่อสังคมออนไลน์ โพสต์บล็อก รีวิวสินค้า ฯลฯ โดย VADER ใช้พจนานุกรมที่มีการให้คะแนนไว้ล่วงหน้าเพื่อคำนวณคะแนนเชิงบวก-ลบ

- **การใช้งาน:**

-VADER สามารถใช้ฟังก์ชัน `polarity_scores()` เพื่อให้คะแนนแต่ละประโยคว่าเป็นเชิงบวก, เชิงลบ, กลาง หรือ คำนวณคะแนนรวม (compound score)

-คะแนนรวม (Compound Score) จะเป็นคะแนนที่บ่งบอกว่าอารมณ์ในข้อความนั้น ๆ เอียงไปทางบวกหรือลบ มากน้อยแค่ไหน

- **ฟังก์ชันหลัก:** `SentimentIntensityAnalyzer` สามารถให้คะแนนความคิดเห็นในรูปแบบ **positive, neutral, negative** และ **compound** (คะแนนรวม)

- **ข้อดี:** แม่นยำกับข้อความสั้น ๆ เช่น โพสต์ในโซเชียล

-เหมาะกับข้อความสั้น เช่น ทวิต โพสต์ในสื่อสังคมออนไลน์

-รวดเร็วและใช้งานง่าย

- **ข้อเสีย:** ไม่เหมาะกับข้อความยาวหรือที่ซับซ้อนมาก

ความสำคัญ:

- VADER ถูกออกแบบมาเพื่อจัดการกับข้อความในสื่อสังคมออนไลน์และรีวิว ซึ่งข้อความเหล่านี้มักสั้นและมีการใช้ภาษาไม่เป็นทางการ จึงเหมาะสมอย่างยิ่งสำหรับการวิเคราะห์ความคิดเห็นที่มาจากแหล่งข้อมูลเช่น Twitter, Facebook หรือรีวิวผลิตภัณฑ์
- ด้วยความสามารถในการให้คะแนนข้อความในรูปแบบที่ละเอียด (Positive, Neutral, Negative และ Compound) VADER ช่วยให้การวิเคราะห์มีความชัดเจนและแม่นยำมากขึ้น โดยเฉพาะเมื่อข้อมูลมีขนาดใหญ่

pandas:

- **ความหมาย:** pandas เป็นไลบรารีที่ถูกใช้ในการจัดการข้อมูลที่เป็นตาราง (dataframe) ซึ่งช่วยให้นำเข้า แยกแยะ และจัดการข้อมูลขนาดใหญ่ได้

- **ฟังก์ชันหลัก:** `read_csv`, `value_counts`, `merge`

- **การใช้งาน:**

-pandas ช่วยให้นำเข้าข้อมูลจากแหล่งข้อมูลหลายรูปแบบ เช่น CSV, Excel, SQL

-สามารถทำการวิเคราะห์ข้อมูลพื้นฐาน เช่น การนับจำนวนข้อมูล การกรองข้อมูล การรวมข้อมูล

- **ข้อดี:** เหมาะสำหรับการจัดการข้อมูลขนาดใหญ่และซับซ้อน

- **ข้อเสีย:** อาจมีประสิทธิภาพต่ำกับข้อมูลที่มีขนาดใหญ่มาก

ความสำคัญ:

- การจัดการข้อมูลที่มีโครงสร้างและเป็นตารางมีความสำคัญมากในกระบวนการวิเคราะห์ข้อมูล Pandas ช่วยในการจัดการข้อมูลที่มีขนาดใหญ่ได้อย่างมีประสิทธิภาพ ช่วยให้การคำนวณ สรุปผล และการแสดงผลข้อมูลสามารถทำได้อย่างรวดเร็ว
- Pandas มีบทบาทสำคัญในขั้นตอนการเตรียมข้อมูล เช่น การทำความสะอาด การกรองข้อมูล หรือการรวมข้อมูล ซึ่งเป็นกระบวนการที่สำคัญมากก่อนการวิเคราะห์ความคิดเห็น

re (Regular Expressions):

- **ความหมาย:** ไวยากรณ์นี้ใช้สำหรับการจัดการและค้นหา pattern ในข้อความ เช่น การลบสัญลักษณ์พิเศษหรือการแยกข้อความที่ต้องการ
- **ความหมาย:** ไวยากรณ์ re ใช้ในการค้นหาและจัดการกับข้อความที่มีรูปแบบเฉพาะ โดยใช้กฎการจับคู่รูปแบบ (Pattern Matching) เช่น การค้นหาคำเฉพาะหรือการลบสัญลักษณ์พิเศษในข้อความ
- **การใช้งาน:** ใช้สำหรับทำความสะอาดข้อมูลก่อนการวิเคราะห์ เช่น ลบสัญลักษณ์พิเศษ (@, #, ! ฯลฯ) หรือคำนวณลักษณะของข้อความตามที่ต้องการ
- **ฟังก์ชันหลัก:** re.sub, re.findall
- **ข้อดี:** ใช้งานง่ายสำหรับการทำความสะอาดข้อความ
-ยืดหยุ่นในการค้นหาและจัดการกับข้อความที่ซับซ้อน
- **ข้อเสีย:** ต้องใช้ความรู้เกี่ยวกับ regular expressions

ความสำคัญ:

- re มีความสำคัญในการทำสะอาดข้อมูลและการค้นหา pattern ในข้อความ เช่น การลบสัญลักษณ์พิเศษหรือการแยกคำ สิ่งนี้จำเป็นมากในกระบวนการเตรียมข้อมูลก่อนการวิเคราะห์ความคิดเห็น เพราะข้อมูลที่มีสัญลักษณ์พิเศษหรือข้อมูลที่ไม่ได้อยู่ในรูปแบบที่ต้องการ อาจทำให้ผลลัพธ์คลาดเคลื่อนได้
- ความสามารถในการจัดการข้อมูลที่ไม่มีโครงสร้างทำให้ re สำคัญในกรณีที่ต้องการประมวลผลข้อความที่ซับซ้อน

matplotlib:

- **ความหมาย:** ไวยากรณ์ที่ใช้สร้างกราฟและแผนภูมิ เพื่อช่วยในการแสดงผลข้อมูลที่ได้จากการวิเคราะห์ความคิดเห็น
- **การใช้งาน:** สามารถสร้างแผนภูมิหลากหลายรูปแบบ เช่น แผนภูมิแท่ง (Bar Chart), แผนภูมิวงกลม (Pie Chart), กราฟเส้น (Line Graph) เพื่อแสดงข้อมูลที่ได้จากการวิเคราะห์
- **ฟังก์ชันหลัก:** plot, bar, pie

- **ข้อดี:** สามารถสร้างกราฟและแผนภูมิที่มีความซับซ้อนได้
-ช่วยในการแสดงข้อมูลที่น่าสนใจและเข้าใจง่ายขึ้น
- **ข้อเสีย:** มีความซับซ้อนในการใช้งานหากต้องการกราฟขั้นสูง
-การปรับแต่งกราฟบางอย่างอาจใช้เวลาและมีความซับซ้อน

ความสำคัญ:

- การแสดงผลข้อมูลที่ได้รับการวิเคราะห์เป็นส่วนสำคัญในการทำความเข้าใจข้อมูล Matplotlib ช่วยให้การนำเสนอผลลัพธ์มีความชัดเจน เข้าใจง่าย และสามารถอธิบายข้อมูลที่ซับซ้อนได้อย่างมีประสิทธิภาพ
- การสร้างกราฟหรือแผนภูมิเพื่อแสดงสัดส่วนความคิดเห็นเชิงบวกหรือเชิงลบสามารถช่วยให้ผู้บริหารหรือผู้ใช้งานเห็นภาพรวมและแนวโน้มของความคิดเห็นได้อย่างชัดเจน

nlk (Natural Language Toolkit):

- **ความหมาย:** NLTK เป็นไลบรารีที่ใช้สำหรับการประมวลผลภาษาธรรมชาติ เช่น การแยกคำ (tokenization), การวิเคราะห์คำ (stemming), และการจัดประเภทข้อความ
- **ฟังก์ชันหลัก:** word_tokenize, pos_tag, SentimentIntensityAnalyzer
- **ข้อดี:** มีเครื่องมือครบครันสำหรับการประมวลผลข้อความ
- **ข้อเสีย:** การใช้งานบางส่วนต้องการการเตรียมข้อมูลและโมเดลที่ซับซ้อน

ความสำคัญ:

- NLTK เป็นไลบรารีที่ทรงพลังสำหรับการประมวลผลภาษาธรรมชาติในขั้นสูง ช่วยให้นักพัฒนาและนักวิจัยสามารถทำงานกับข้อความได้ตั้งแต่การแยกคำ (tokenization), การทำ stemming, POS tagging ไปจนถึงการจัดประเภทข้อความ
- ความสามารถในการทำงานกับข้อความหลายรูปแบบทำให้ NLTK มีความสำคัญในโครงการที่ซับซ้อนหรือข้อมูลที่ต้องการการวิเคราะห์ละเอียด เช่น การวิเคราะห์ความคิดเห็นในเอกสารขนาดใหญ่ หรือข้อความที่มีหลายมิติ (เช่น ความเห็นที่มีหลายแง่มุม)

1. แบบอัตโนมัติหรือการเรียนรู้ของเครื่อง (Automated/Machine Learning-based)

- ใช้ข้อมูลที่ถูกติดป้ายกำกับ (labeled data) เพื่อฝึกโมเดล โดยโมเดลจะเรียนรู้จากข้อมูลเก่าเพื่อทำนายความรู้สึกของข้อความใหม่
- ใช้เวลาฝึกโมเดล แต่มีความแม่นยำและประสิทธิภาพที่สูงกว่าหากมีข้อมูลมากพอ
- **การทำงาน:** ใช้โมเดลที่ถูกฝึกจากข้อมูลตัวอย่างที่มีการติดป้ายกำกับ (labeled data) เช่น ข้อความที่ระบุว่าเป็นเชิงบวกหรือเชิงลบ ระบบจะเรียนรู้จากข้อมูลเหล่านี้เพื่อคาดการณ์ความรู้สึกของข้อความใหม่ ๆ โดยวิธีนี้โมเดลมักจะถูกใช้ในปัญหาการจำแนกประเภท (classification problem) และต้องใช้ข้อมูลจำนวนมากในการฝึก
- **ข้อดี:** มีความแม่นยำสูงหากมีการฝึกโมเดลด้วยข้อมูลที่มากพอ, สามารถเรียนรู้บริบทที่ซับซ้อนได้ดี เช่น การประชดประชัน, สามารถพัฒนาและปรับปรุงคุณภาพได้เรื่อย ๆ ตามการเพิ่มข้อมูลใหม่
- **ข้อเสีย:** ต้องใช้เวลาฝึกโมเดล และต้องการข้อมูลจำนวนมาก, ถ้าไม่มีข้อมูลฝึกที่เพียงพอ อาจเกิดการคาดการณ์ผิดพลาดได้

2.แบบใช้กฎหรือพจนานุกรม (Rule or Lexicon-based)

- อิงกับการสร้างกฎหรือพจนานุกรมที่กำหนดความรู้สึกของคำต่าง ๆ ล่วงหน้า
- ทำงานได้รวดเร็ว แต่ความแม่นยำอาจต่ำเมื่อข้อความมีบริบทหรือคำที่ไม่อยู่ในพจนานุกรม
- **การทำงาน:** ใช้ชุดของกฎหรือพจนานุกรมที่มีการกำหนดคะแนนความรู้สึกล่วงหน้า คำแต่ละคำจะมีคะแนนความรู้สึก เช่น คำว่า "ดี" อาจมีคะแนน +1 และคำว่า "แย่" อาจมีคะแนน -2 เมื่อตรงกับคำที่อยู่ในพจนานุกรม ระบบจะนำคะแนนเหล่านี้มารวมกันเพื่อหาผลลัพธ์ว่าข้อความนั้นเป็นบวกหรือลบ
- **ข้อดี:** ทำงานได้รวดเร็ว เนื่องจากไม่ต้องฝึกโมเดล, ง่ายต่อการพัฒนาและบำรุงรักษา โดยการเพิ่มหรือแก้ไขคำในพจนานุกรม, ไม่ต้องใช้ข้อมูลตัวอย่างในการฝึก
- **ข้อเสีย:** ไม่แม่นยำในบริบทที่ซับซ้อน เช่น การใช้คำประชดประชัน หรือคำที่มีหลายความหมายในแต่ละบริบท, ขาดความยืดหยุ่นเมื่อต้องรับมือกับคำศัพท์หรือสถานการณ์ใหม่ ๆ ที่ไม่อยู่ในพจนานุกรม

ความแตกต่าง

- **การเรียนรู้ของเครื่อง:** เน้นการฝึกโมเดลจากข้อมูลจำนวนมาก สามารถเรียนรู้และเข้าใจบริบทที่ซับซ้อนได้ แต่ต้องใช้เวลาและข้อมูลมาก
- **การใช้กฎหรือพจนานุกรม:** ทำงานได้รวดเร็ว แต่ความแม่นยำขึ้นอยู่กับความครบถ้วนของพจนานุกรม ไม่สามารถปรับตัวกับคำศัพท์ใหม่หรือบริบทที่ซับซ้อนได้ดีเท่าวิธีการเรียนรู้ของเครื่อง