**1. การวัดความคล้ายของข้อมูลแบบ cosine similarity และ Pearson's similarity แตกต่างกันอย่างไร ผลลัพธ์ที่ได้มีผลต่อการวิเคราะห์ข้อมูลหรือไม่ อย่างไร \***

จากการทดลองพบว่า cosine similarity มีค่าอยู่ช่วง [0, 1] โดยเป็นการดูองศา cosine ของข้อมูล โดยถ้ามีค่าเท่ากับ 0 แสดงว่าข้อมูลไม่มีความสัมพันธ์กันเลย และเมื่อมีค่าเข้าใกล้ 1 แสดงว่าข้อมูลมีความสัมพันธ์กันสำหรับ Pearson's similarity [-1, 1] เป็นการหาความสัมพันธ์ระหว่าง X และ Y โดยถ้ามีค่าน้อยแสดงว่าข้อมูลไม่มีความสัมพันธ์กันมาก และเมื่อมีค่าเข้าใกล้ 1 แสดงว่าข้อมูลมีความสัมพันธ์กันมาก

ผลลัพธ์ที่ได้พบว่าค่าที่ได้จากทั้งสองวิธีมีความแตกต่างกัน คือ ผู้ใช้ที่มีความใกล้เคียงกันไม่เหมือนกันเมื่อใช้วิธีที่ต่างกัน ทำให้มีผลต่อการนำไปทดลองหรือการแนะนำได้

**2. จากการทดลอง นศ. คิดว่า การแนะนำรายการ movie ด้วยการวิเคราะห์ข้อมูลแบบใด จะตรงกับความชอบของ user เพราะเหตุใด (เปรียบเทียบการวิเคราะห์ข้อมูลอย่างน้อย 2 รูปแบบ) \***

จากการทดลองการแนะนำของคนที่มีความชอบคล้ายกันที่สุด และแนะนำที่หนังใกล้เคียงกันโดย user ให้คะแนนมากกว่าหรือเท่ากับ 3

คิดว่าแนะนำที่หนังใกล้เคียงกันโดย user ให้คะแนนมากกว่าหรือเท่ากับ 3 จะตรงกับความชอบของ user เนื่องจากมีการดูความคล้ายกันของหนังที่ดูจาก genre และยังต้องของหนังมากว่าเท่ากับ 3 ซึ่งคะแนนเมากกว่าเท่ากับ 3 ถือว่าเป็นคะแนนระดับกลางถึงดี ซึ่งสมควรที่จะเอาหนังที่คล้ายกันมาใช้แนะนำได้ด้วย ส่วนการแนะนำของคนที่มีความชอบคล้ายกันที่สุด อาจตรงกับความชอบของ user น้อยกว่า เนื่องจากหนังที่ user ที่คล้ายกันชอบโดยมีคะแนนมากที่สุด user อาจไม่ชอบได้ เนื่องจาก genre ของหนัง เป็นต้น

**3. ให้ลองยกตัวอย่างข้อมูลที่ควรนำมาวิเคราะห์เพิ่มเติม เพื่อแนะนำ movie ได้ตรงกับความต้องการของ user มากยิ่งขึ้น \***

อาจลองใช้ tag.csv ประกอบด้วย tag ของหนัง ซึ่งเป็นกลุ่มของหนัง อาจมาให้ในการแนะนำโดยดูจาก tag ได้

**4. แนะนำรูปแบบการ visualize ข้อมูล นอกจากที่ได้ทำการทดลองไปแล้ว อีกอย่างน้อย 2 รูปแบบข้อมูล ที่จะทำให้เข้าใจพื้นฐานข้อมูลได้ดีมากขึ้น \***

1. อาจใช้การ visualize ของ cosine similarity ในรูปแบบของ vector เพื่อให้เห็นความใกล้เคียงของข้อมูลได้ดียิ่งขึ้นด้วย

2. การใช้การไล่ระดับสี(heatmap) ในตารางการแสดงความเหมือนของข้อมูล โดยอาจกำหนดค่ากลางเท่ากับ 0 ให้เป็นสีดำ เมื่อมีค่าน้อยให้เป็นสีฟ้าและเมื่อมีค่าให้มีสีแดง เป็นต้น

3. การใช้ Box plots ในการดูการกระจายของข้อมูล

**5. จากการทดลองจัดกลุ่มข้อมูลด้วย K-mean clustering ให้วิเคราะห์ผลลัพธ์การจัดกลุ่มเมื่อปรับจำนวนกลุ่ม พร้อมสรุปจำนวนกลุ่มที่เหมาะสม ที่สามารถแบ่งแยกกลุ่มของ user ได้ดีที่สุด \***

จากการทดลองด้วย k = 3, 4 และ 5 พบว่าการจัดกลุ่มด้วย k = 3 สามารถแบ่งกลุ่มของ user ได้ดีที่สุด พบว่าเมื่อ k = 3 จุด centroid ที่อยู่ภายในกลุ่มนั้น ๆ มากกว่า k= 4 และ 5

**6. การนำเสนอผลการแบ่งกลุ่มด้วย K-mean clustering ด้วยกราฟของข้อมูล genre feature คู่ใด (X-Y axis) แสดงให้เห็นผลการแบ่งกลุ่มได้ชัดเจนที่สุด \***

คู่ 9 และ 13

**7. จากการทดลองจัดกลุ่มข้อมูลด้วย Gaussian Mixture Model ให้วิเคราะห์ผลลัพธ์การจัดกลุ่มเมื่อปรับจำนวนกลุ่ม พร้อมสรุปจำนวนกลุ่มที่เหมาะสม ที่สามารถแบ่งแยกกลุ่มของ user ได้ดีที่สุด \***

จากการทดลองด้วย n\_components = 3, 4 และ 5 พบว่าการจัดกลุ่มด้วย components = 4 สามารถแบ่งกลุ่มของ user ได้ดีที่สุด พบว่าเมื่อ components = 4 ข้อมูลอยุ่ภายในขอบเขตมากกว่า components = 3 และ 5

**8. การนำเสนอผลการแบ่งกลุ่มด้วย Gaussian Mixture Model ด้วยกราฟของข้อมูล genre feature คู่ใด (X-Y axis) แสดงให้เห็นผลการแบ่งกลุ่มได้ชัดเจนที่สุด \***

คู่ 9 และ 13

**9. ให้นศ. วิเคราะห์ความเหมาะสมของข้อมูลที่ใช้ในการทดลองเพื่อจัดกลุ่ม user ว่าให้ผลการจัดกลุ่มมีประสิทธิภาพหรือไม่ เพราะเหตุใด และให้แนะนำรูปแบบของข้อมูลที่คิดว่าจะให้ประสิทธิภาพในการจัดกลุ่ม user ได้ดีขึ้น เพิ่มเติมอีกอย่างน้อย 2 รูปแบบ \***

อาจยังไม่มีประสิทธิภาพมากทเที่ควร เนื่องจากขอ้มูลที่ใช้มาจากการสุ่มเพียง 30 sample เท่านั้นอาจยังไม่สามารถแบ่งกลุ่มได้ดีเท่าที่ควร

1. ใช้ข้อมูลความชอบของ user ตาม tag ของหนัง โดยมี column เป็น tag ประเภทต่างๆ และ row เป็น userId
2. ใช้ข้อมูลความชอบของ user ตามปีของหนัง โดยมี column เป็นปีของหนังต่างๆ และ row เป็น userId เพราะอาจจะมี user เป็นกลุ่มที่ชอบดูหนังเก่าหรือหนังใหม่ได้