Experimenting with Classifying News from BBC News using the Text Analytics Process

1. บทนำ

ข้อมูลที่ได้มาจากการรายงานข่าวของ BBC โดย BBC เป็นสำนักข่าวแห่งหนึ่งที่มีความน่าเชื่อถือ และได้รับการรับรองจากทั่วโลก

รายงานนี้จัดทำขึ้นโดยมีวัตถุประสงค์ เพื่อทำการทดลองและการเลือกใช้เทคนิคต่างๆเพื่อสร้างแบบจำลองทำนายการแยกประเภทของข่าวได้อย่างอัตโนมัติที่มีประสิทธิภาพ โดยใช้ข้อมูลจากข่าวแต่ละประเภท รวมถึงการทดสอบประสิทธิภาพของวิธีการทำ Text Clustering ในการแยกกลุ่ม เมื่อเปรียบเทียบกับกลุ่ม (Categories) ที่เราได้เตรียมไว้อยู่แล้ว

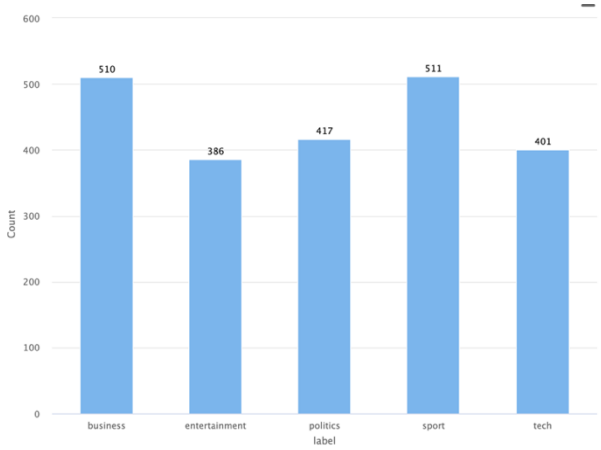
การวิเคราะห์จะใช้กระบวนการ Text Analytics ในการวิเคราะห์ผ่านการทำ Classification และ Clustering ของข้อมูล ผ่านโปรแกรม RapidMiner Studio

2. รายละเอียดชุดข้อมูล

ชุดข้อมูลเริ่มต้นมีทั้งหมด 2,225 ไฟล์ นามสกุลไฟล์ .txt แบ่ง Categories ออกเป็น 5 กลุ่ม 1. Business 2. Entertainment 3. Politics 4. Sport 5. Tech ซึ่งทำการดาวน์โหลดชุดข้อมูลมาจาก https://www.kaggle.com/datasets/shivamkushwaha/bbc-full-text-document-classification/data

มีการแบ่งชุดข้อมูลสำหรับการประมวลผลออกเป็น ชุดข้อมูลสำหรับการให้โปรแกรมเรียนรู้ (Training Data) 9 ส่วน และชุดข้อมูลสำหรับทดสอบ (Test Data) 1 ส่วน และใช้ Classification Algorithms ได้แก่ Naïve Bayes, Decision Tree และ k-NN รวมถึงใช้ Clustering Algorithm ได้แก่ k-Means ในการประมวลผล และทดสอบประสิทธิภาพของการทำนายการแยกประเภทของเอกสาร

3. การกระจายตัวของข้อมูล



รูปภาพที่ 1 : แสดง Bar Chart แจกแจงข้อมูลการนับ (Count) ของข้อมูลภายในแต่ละเอกสาร

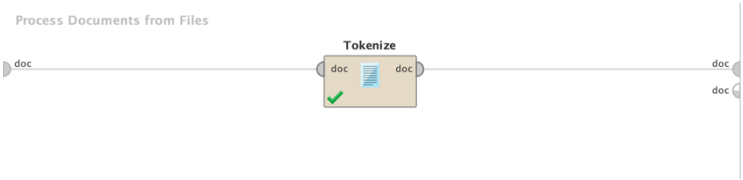
4. กระบวนการทำ Text Analytics ผ่านโปรแกรม RapidMiner Studio

4.1 Text Preprocessing

เป็นกระบวนเตรียมข้อมูลให้พร้อมก่อนนำเข้าโมเดลเพื่อประมวลผลข้อมูลอย่างมีประสิทธิภาพ โดยมีกระบวนการดังนี้

4.1.1 การแบ่งคำ (Tokinization)

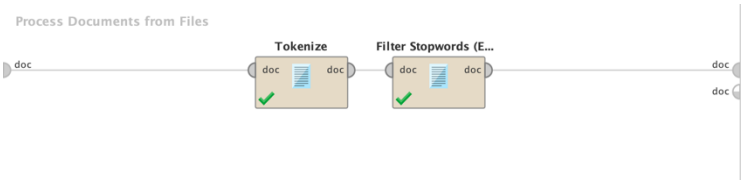
แสดงจำนวนคำหรือจำนวน Attributes ทั้งหมด 32,082 Attributes



รูปภาพที่ 2 : แสดงกระบวนการทำการแบ่งคำ (Tokinization)

4.1.2 การตัดคำโดยใช้ Stopwords

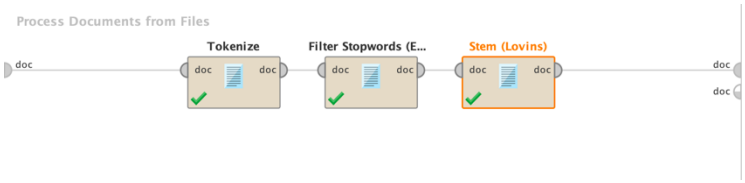
แสดงจำนวน Attributes ลดลงเหลือ 31,463 Attributes



รูปภาพที่ 3 : แสดงกระบวนการทำการตัดคำโดยใช้ Stopwords

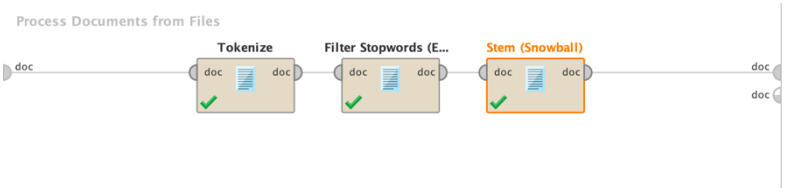
4.1.3 การรวมคำที่มีความหมายเดียวกัน (Stemming)

รูปแบบ Lovins แสดงจำนวน Attributes ลดลงเหลือ 16,955 Attribute



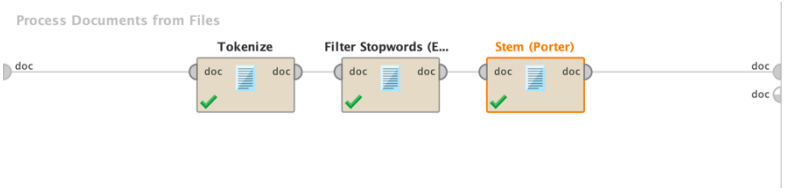
รูปภาพที่ 4 : แสดงกระบวนการรวมคำในรูปแบบ Lovins

รูปแบบ Snowball แสดงจำนวน Attributes ลดลงเหลือ 18,764 Attributes



รูปภาพที่ 5 : แสดงกระบวนการรวมคำในรูปแบบ Snowball

รูปแบบ Porter แสดงจำนวน Attributes ลดลงเหลือ 18,912 Attributes

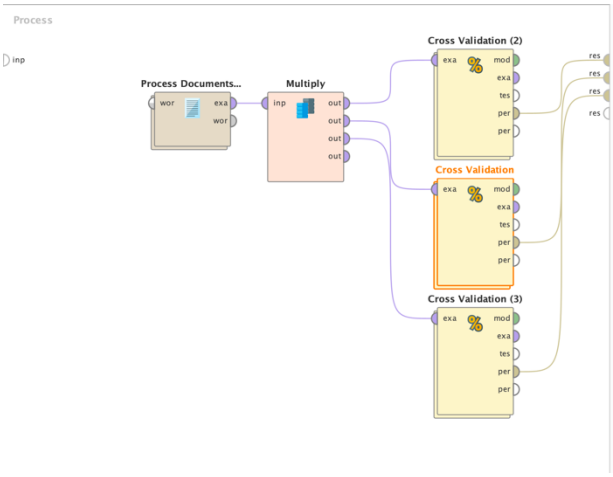


รูปภาพที่ 6 : แสดงกระบวนการรวมคำในรูปแบบ Porter

ผลสรุปจากกระบวนการรวมคำนั้น จะใช้รูปแบบ Lovins เนื่องจากมีประสิทธิภาพในการลดจำนวน Attributes ลงได้ดีที่สุด

4.2 การวิเคราะห์โดยใช้ Text Classification

กระบวนการทำ Text Classification



รูปภาพที่ 7 : แสดงกระบวนการทำ Text Classification

4.2.1 การทดสอบที่ 1 : กำหนดรูปแบบ Vector Creation คือ TF-IDF โดยยังไม่ทำการใช้กระบวนการ Pruning Method

accuracy: 93.03% +/- 1.63% (micro average: 93.03%)

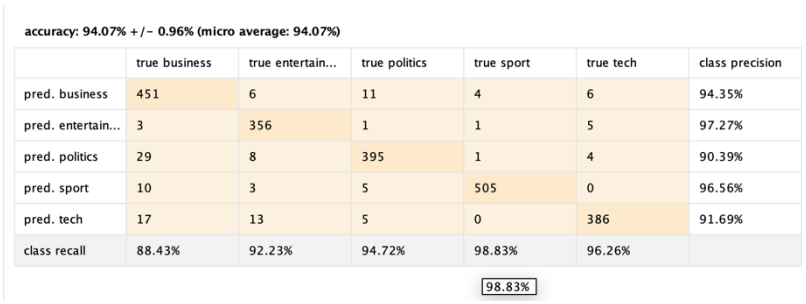
	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	442	3	14	6	5	94.04%
pred. entertain...	3	352	4	1	4	96.70%
pred. politics	36	11	391	3	6	87.47%
pred. sport	8	3	5	501	2	96.53%
pred. tech	21	17	3	0	384	90.35%
class recall	86.67%	91.19%	93.76%	98.04%	95.76%	

รูปภาพที่ 8 : แสดงกระบวนการทำ Text Classification โดยวิธี k-NN โดยกำหนดค่า k=5

accuracy: 93.26% +/- 1.67% (micro average: 93.26%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	445	5	9	4	5	95.09%
pred. entertai...	4	351	4	1	6	95.90%
pred. politics	34	13	394	2	8	87.36%
pred. sport	7	3	4	504	1	97.11%
pred. tech	20	14	6	0	381	90.50%
class recall	87.25%	90.93%	94.48%	98.63%	95.01%	

รูปภาพที่ 9 : แสดงกระบวนการทำ Text Classification โดยวิธี k-NN โดยกำหนดค่า k=10

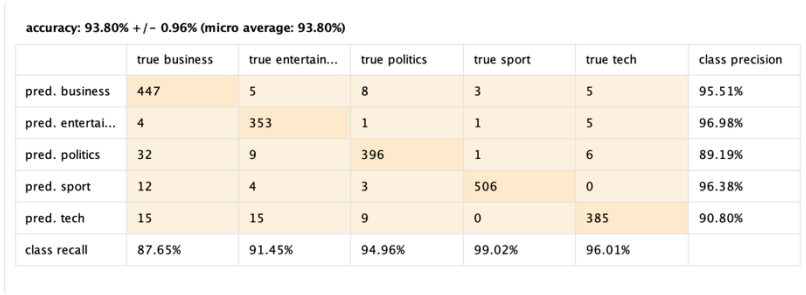


accuracy: 94.07% +/- 0.96% (micro average: 94.07%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	451	6	11	4	6	94.35%
pred. entertain...	3	356	1	1	5	97.27%
pred. politics	29	8	395	1	4	90.39%
pred. sport	10	3	5	505	0	96.56%
pred. tech	17	13	5	0	386	91.69%
class recall	88.43%	92.23%	94.72%	98.83%	96.26%	

98.83%

รูปภาพที่ 10 : แสดงกระบวนการทำ Text Classification โดยวิธี k-NN โดยกำหนดค่า k=15



accuracy: 93.80% +/- 0.96% (micro average: 93.80%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	447	5	8	3	5	95.51%
pred. entertai...	4	353	1	1	5	96.98%
pred. politics	32	9	396	1	6	89.19%
pred. sport	12	4	3	506	0	96.38%
pred. tech	15	15	9	0	385	90.80%
class recall	87.65%	91.45%	94.96%	99.02%	96.01%	

รูปภาพที่ 11 : แสดงกระบวนการทำ Text Classification โดยวิธี k-NN โดยกำหนดค่า k=20

accuracy: 90.74% +/- 2.00% (micro average: 90.74%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	429	5	23	8	6	91.08%
pred. entertai...	16	355	13	8	10	88.31%
pred. politics	27	8	370	4	10	88.31%
pred. sport	7	0	4	490	0	97.80%
pred. tech	31	18	7	1	375	86.81%
class recall	84.12%	91.97%	88.73%	95.89%	93.52%	

รูปภาพที่ 12 : แสดงกระบวนการทำ Text Classification โดยวิธี Naive Bayes

accuracy: 47.41% +/- 5.19% (micro average: 47.42%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	4	0	1	0	0	80.00%
pred. entertai...	2	174	3	0	11	91.58%
pred. politics	1	5	126	4	1	91.97%
pred. sport	494	206	285	507	145	30.97%
pred. tech	9	1	2	0	244	95.31%
class recall	0.78%	45.08%	30.22%	99.22%	60.85%	

รูปภาพที่ 13 : แสดงกระบวนการทำ Text Classification โดยวิธี Decision Tree

4.2.2 การทดสอบที่ 2 : กำหนดรูปแบบ Vector Creation คือ TF-IDF โดยทำการใช้กระบวนการ Pruning Method แบบ Percental กำหนดค่า 10% - 80%

accuracy: 90.56% +/- 2.49% (micro average: 90.56%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	463	9	24	5	17	89.38%
pred. entertain...	4	338	2	4	15	93.11%
pred. politics	19	8	370	6	8	90.02%
pred. sport	10	16	10	486	3	92.57%
pred. tech	14	15	11	10	358	87.75%
class recall	90.78%	87.56%	88.73%	95.11%	89.28%	

รูปภาพที่ 14 : แสดงกระบวนการทำ Text Classification โดยวิธี k-NN

accuracy: 90.83% +/- 1.30% (micro average: 90.83%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	464	8	25	0	17	90.27%
pred. entertain...	2	343	6	10	13	91.71%
pred. politics	22	2	377	14	7	89.34%
pred. sport	0	3	2	477	4	98.15%
pred. tech	22	30	7	10	360	83.92%
class recall	90.98%	88.86%	90.41%	93.35%	89.78%	

รูปภาพที่ 15 : แสดงกระบวนการทำ Text Classification โดยวิธี Naive Bayes

4.2.3 การทดสอบที่ 3 : กำหนดรูปแบบ Vector Creation คือ Term of Frequency โดยทำการใช้กระบวนการ Pruning Method แบบ Percental กำหนดค่า 10% - 80%

accuracy: 89.35% +/- 1.53% (micro average: 89.35%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	462	12	17	3	16	90.59%
pred. entertain...	6	325	2	6	18	91.04%
pred. politics	25	10	370	2	14	87.89%
pred. sport	4	28	16	491	13	88.95%
pred. tech	13	11	12	9	340	88.31%
class recall	90.59%	84.20%	88.73%	96.09%	84.79%	

รูปภาพที่ 16 : แสดงกระบวนการทำ Text Classification โดยวิธี k-NN

accuracy: 90.16% +/- 1.49% (micro average: 90.16%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	464	7	23	2	18	90.27%
pred. entertain...	4	339	4	15	14	90.16%
pred. politics	20	1	378	13	6	90.43%
pred. sport	0	3	3	468	6	97.50%
pred. tech	22	36	9	13	357	81.69%
class recall	90.98%	87.82%	90.65%	91.59%	89.03%	

รูปภาพที่ 17 : แสดงกระบวนการทำ Text Classification โดยวิธี Naive Bayes

4.2.4 การทดสอบที่ 4 : กำหนดรูปแบบ Vector Creation คือ Term Occurrences โดยทำการใช้กระบวนการ Pruning Method แบบ Absolute

accuracy: 55.73% +/- 5.30% (micro average: 55.73%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	310	23	27	7	52	73.99%
pred. entertain...	0	83	0	0	7	92.22%
pred. politics	26	23	253	4	44	72.29%
pred. sport	173	257	137	500	204	39.34%
pred. tech	1	0	0	0	94	98.95%
class recall	60.78%	21.50%	60.67%	97.85%	23.44%	

รูปภาพที่ 18 : แสดงกระบวนการทำ Text Classification โดยวิธี k-NN

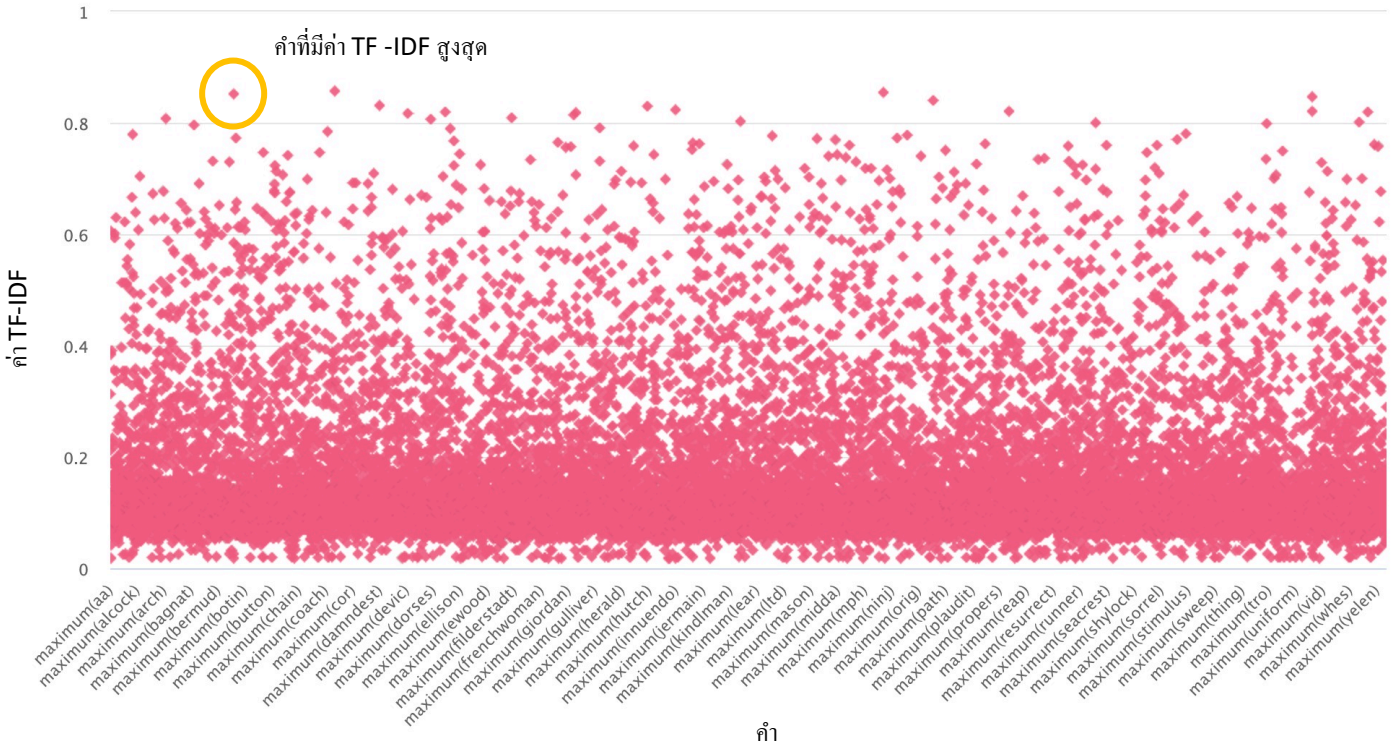
accuracy: 90.43% +/- 1.90% (micro average: 90.43%)

	true business	true entertain...	true politics	true sport	true tech	class precision
pred. business	435	2	22	8	8	91.58%
pred. entertain...	11	352	11	10	15	88.22%
pred. politics	21	9	368	2	9	89.98%
pred. sport	5	2	7	489	1	97.02%
pred. tech	38	21	9	2	368	84.02%
class recall	85.29%	91.19%	88.25%	95.69%	91.77%	

รูปภาพที่ 19 : แสดงกระบวนการทำ Text Classification โดยวิธี Naive Bayes

สรุปผลการทำ Text Classification ได้ว่า การประมวลผลข้อมูลที่ดีที่สุดจากการทดสอบค่า Accuracy คือการใช้ k-NN Algorithm ในเงื่อนไขค่า k=15 (รูปภาพที่ 10) โดยกำหนดรูปแบบ Vector Creation ในรูปแบบ TF-IDF

4.3 แสดงผลคำที่มีความสำคัญในการแยกประเภทเอกสาร



รูปภาพที่ 20 : แสดงกราฟแจกแจงค่า TF-IDF ของแต่ละคำ

ความสำคัญของคำนั้น เป็นคำที่ใช้แยกเอกสารในแต่ละ Categories ออกจากกัน โดยเราวัดความสำคัญจากค่า TF-IDF (รูปภาพที่ 20) ซึ่งคำที่มีความสำคัญที่สุด (วงกลมสีเหลือง) หรือคำที่มีค่า TF-IDF สูงที่สุด คือคำว่า blog และมีค่า TF-IDF เท่ากับ 0.851

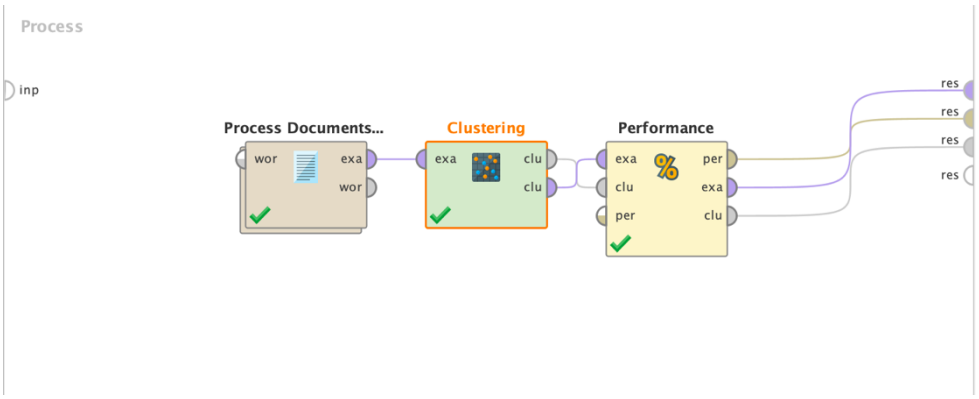
Word	Attribu...	Tota... ↓	Docum...	busine...	enterta...	politics	sport	tech
win	win	1013	526	55	194	109	609	46
film	film	1178	288	12	999	10	1	156
part	part	1492	694	183	112	901	87	209
bank	bank	545	187	459	18	12	3	53
computer	computer	444	181	11	7	5	0	421

รูปภาพที่ 21 : แสดงตาราง Total Occurrences และแจกแจงจำนวนคำในแต่ละเอกสาร

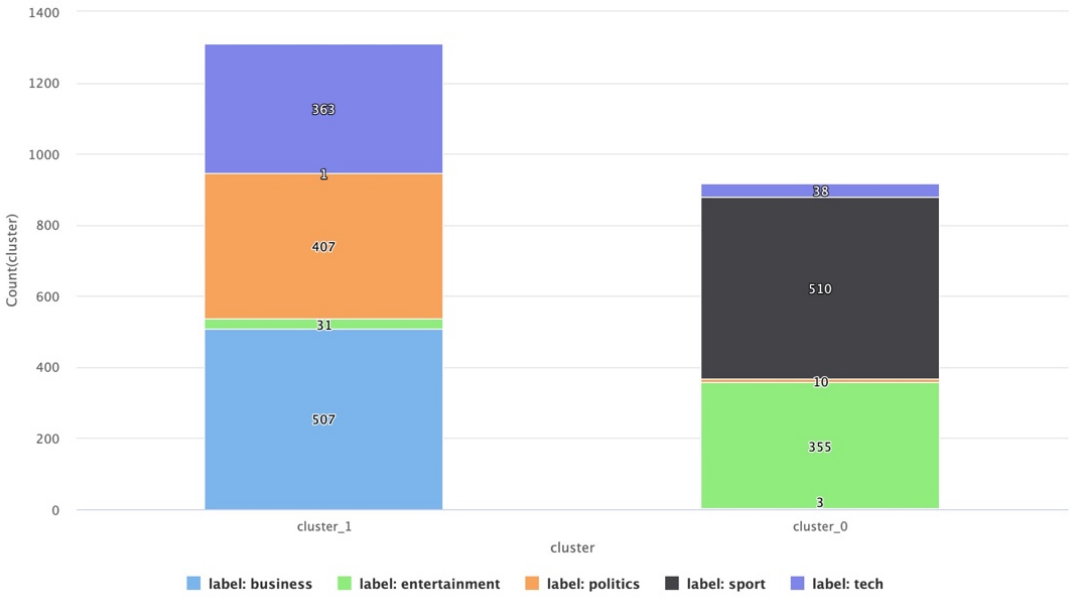
นอกจากข้อมูลในรูปภาพที่ 20 ยังมีอีกหลายคำที่สามารถนำมาใช้เป็นตัวแยกประเภทเอกสาร หรือเรียกว่าเป็นคำที่มีความสำคัญ ตามรูปภาพที่ 21 ซึ่งสามารถวิเคราะห์จากการพิจารณาดูว่ามีคำไหนบ้างที่เอกสารอื่นนั้นมีน้อย

4.4 การวิเคราะห์โดยใช้ Text Clustering

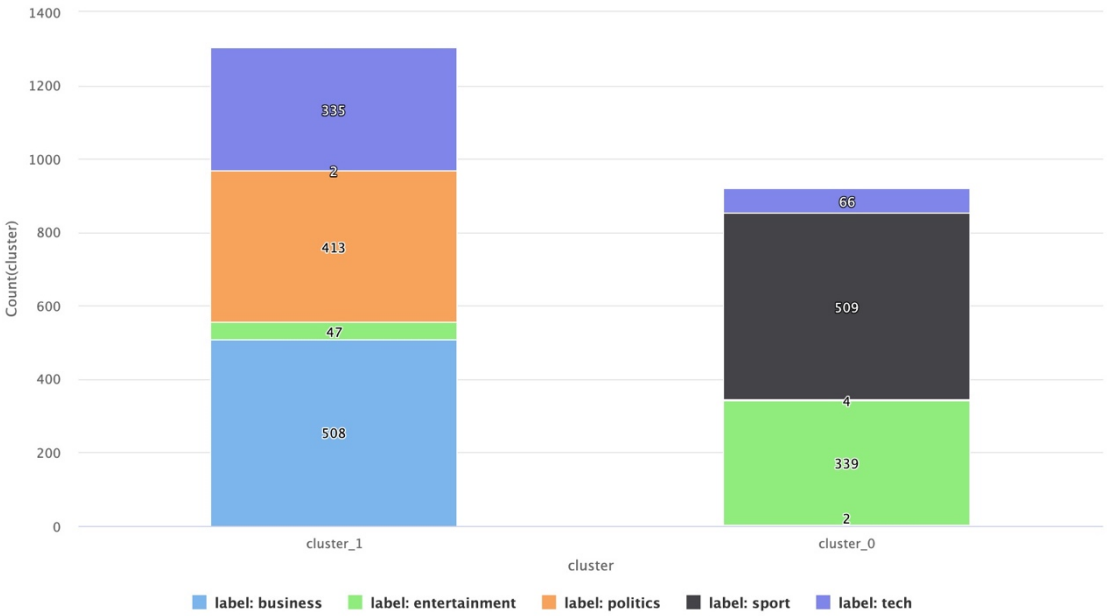
กระบวนการทำ Text Clustering



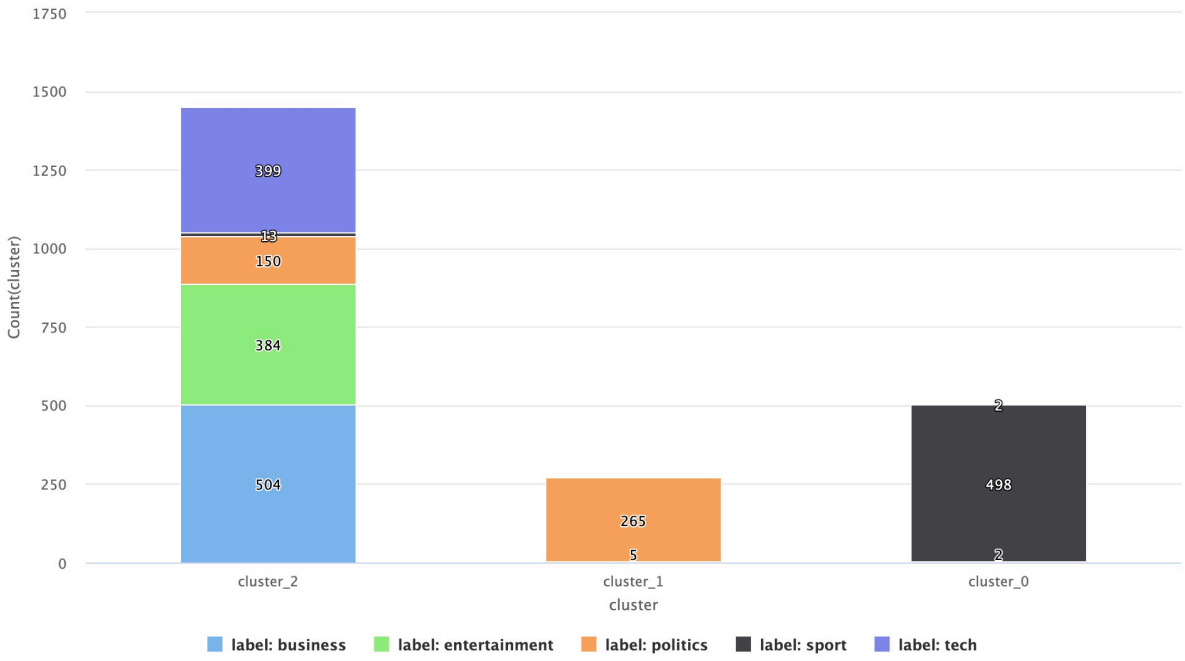
รูปภาพที่ 22 : แสดงกระบวนการทำ Text Clustering



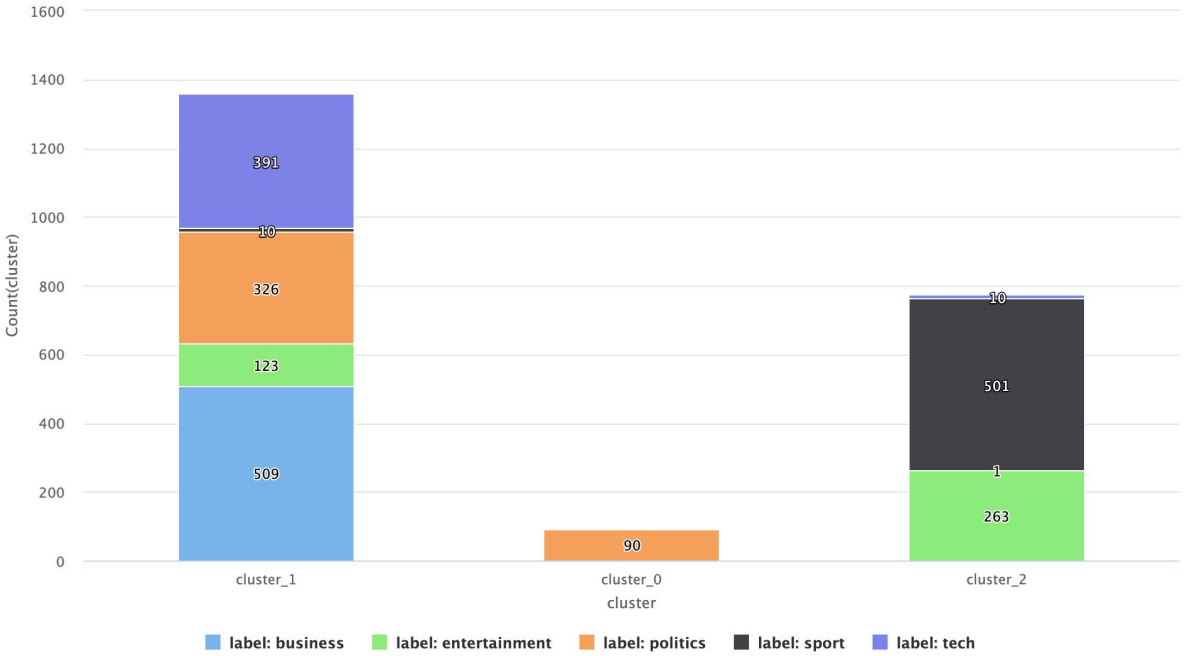
รูปภาพที่ 23 : แสดงกราฟจำนวน Categories ในแต่ละ Cluster โดยมี Parameter ค่า k=2 และวัดระยะห่างรูปแบบ Euclidian Distances



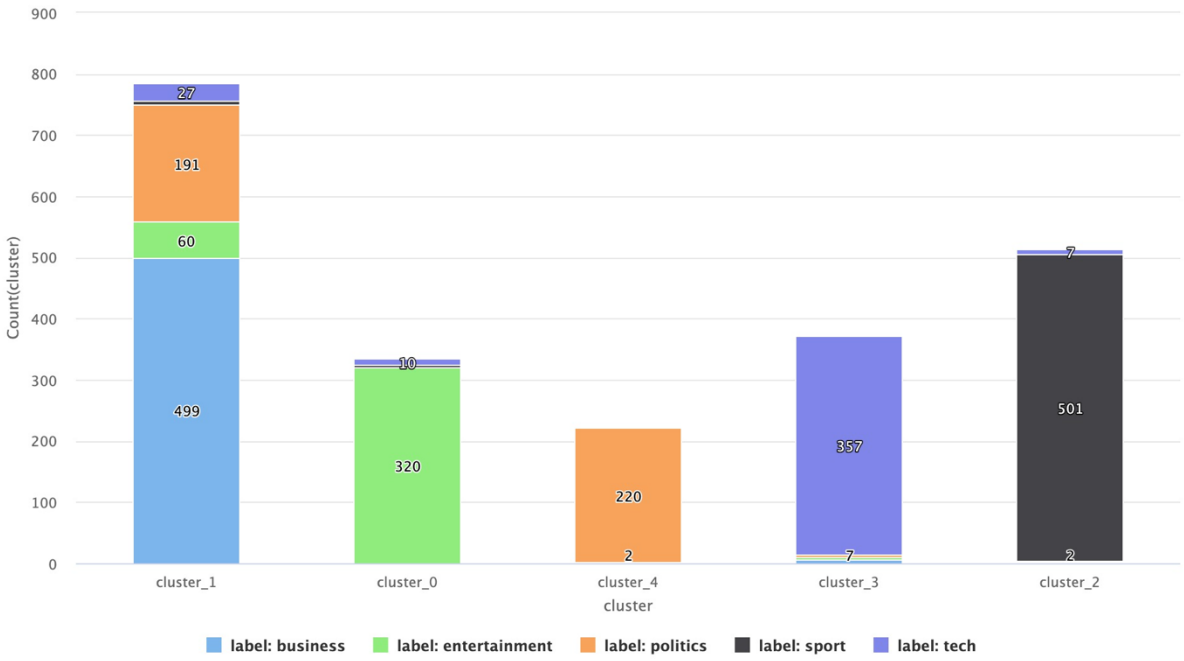
รูปภาพที่ 24 : แสดงกราฟจำนวน Categories ในแต่ละ Cluster โดยมี Parameter ค่า k=2 และวัดระยะห่างรูปแบบ Manhattan Distances



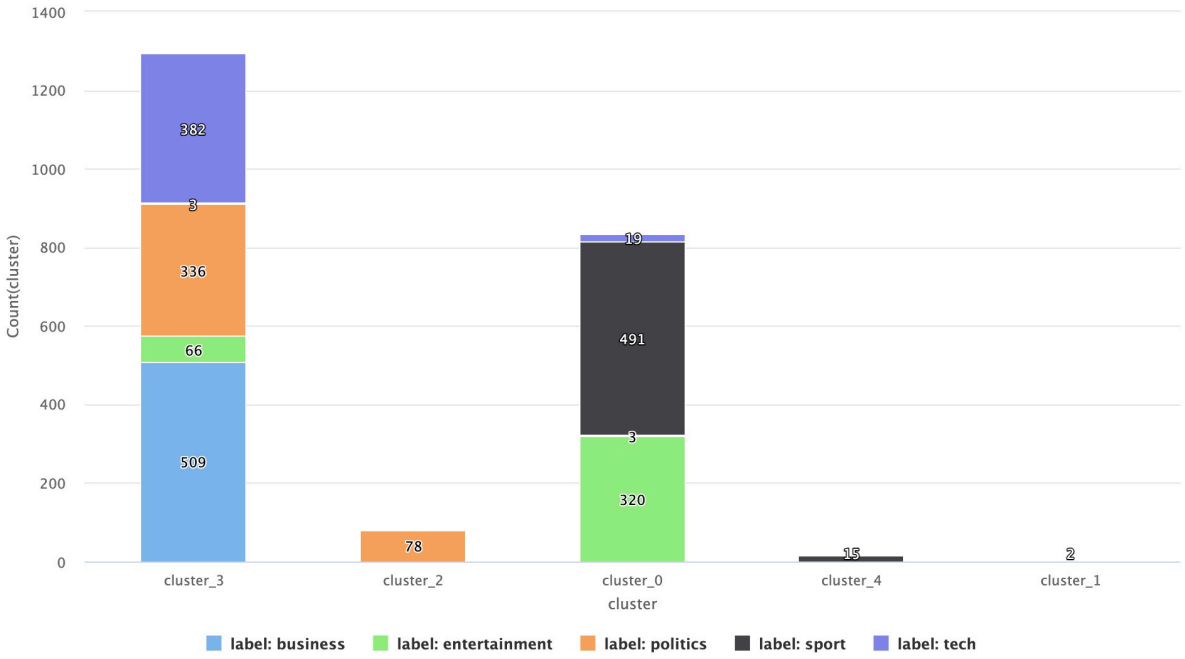
รูปภาพที่ 25 : แสดงกราฟจำนวน Categories ในแต่ละ Cluster โดยมี Parameter ค่า k=3 และวัดระยะห่างรูปแบบ Euclidian Distances



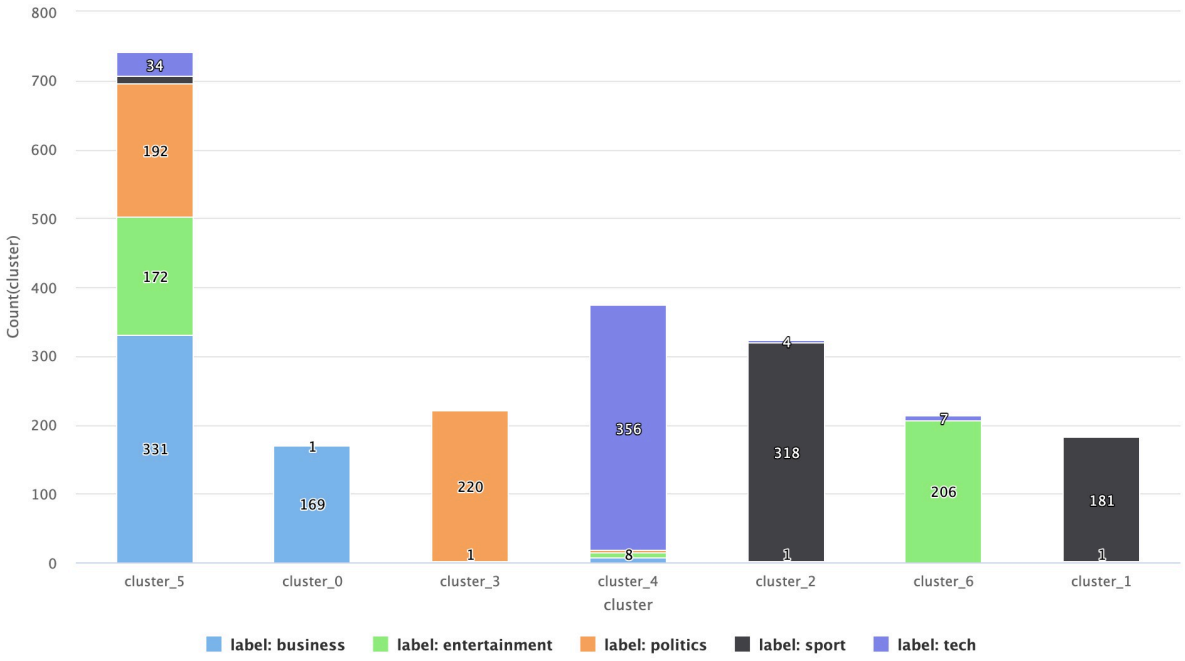
รูปภาพที่ 26 : แสดงกราฟจำนวน Categories ในแต่ละ Cluster โดยมี Parameter ค่า k=3
และวัดระยะห่างรูปแบบ Manhattan Distances



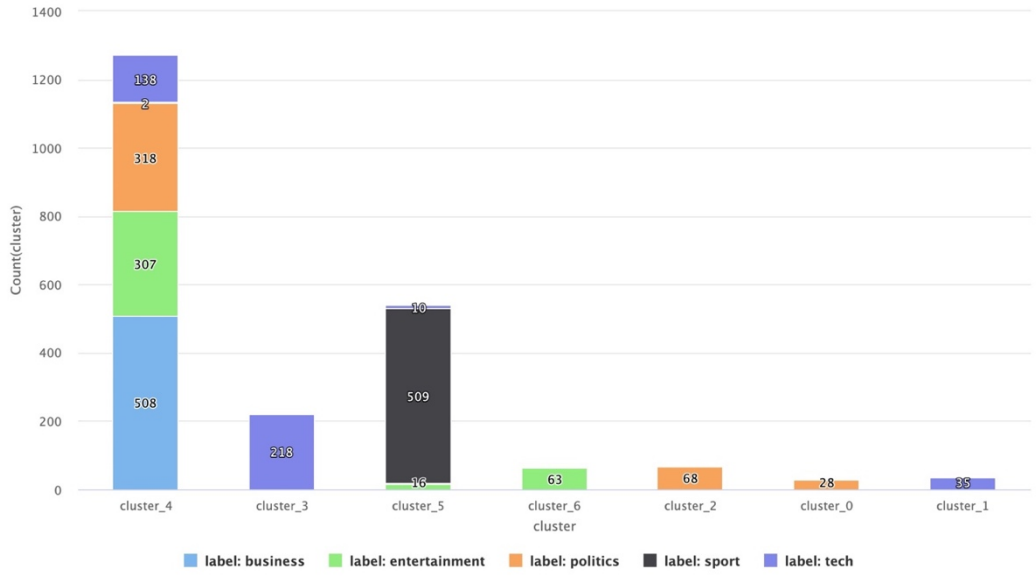
รูปภาพที่ 27 : แสดงกราฟจำนวน Categories ในแต่ละ Cluster โดยมี Parameter ค่า k=5
และวัดระยะห่างรูปแบบ Euclidian Distances



รูปภาพที่ 28 : แสดงกราฟจำนวน Categories ในแต่ละ Cluster โดยมี Parameter ค่า k=5 และวัดระยะห่างรูปแบบ Manhattan Distance



รูปภาพที่ 29 : แสดงกราฟจำนวน Categories ในแต่ละ Cluster โดยมี Parameter ค่า k=7 และวัดระยะห่างรูปแบบ Euclidian Distances



รูปภาพที่ 30 : แสดงกราฟจำนวน Categories ในแต่ละ Cluster โดยมี Parameter ค่า k=7 และวัดระยะห่างรูปแบบ Manhattan Distance

การอธิบายผลลัพธ์จากการทำ Bar Chart เพื่อดูประสิทธิภาพของการทำ Clustering เมื่อเปรียบเทียบกับ Categories ที่เราทราบอยู่แล้ว โดยมี Parameter คือค่า k และรูปแบบการวัดระยะห่าง ซึ่งผลลัพธ์ค่อนข้างชัดเจนในเรื่องของการแยก Categories ตามแต่ละ Cluster

การรายงานผลลัพธ์ที่หนึ่ง ที่ค่า k=2 จะแสดงแบ่ง Categories ในแต่ละ Cluster ค่อนข้างชัดเจน ได้มากกว่าค่า k อื่นๆ (รูปภาพที่ 23, รูปภาพที่ 24)

ผลลัพธ์ที่สอง รูปแบบการวัดระยะห่างแบบ Euclidian Distances ทำการจัดกลุ่มได้ชัดเจนกว่า แบบ Manhattan ในทุกๆการเปลี่ยนค่า k

ผลลัพธ์ที่สาม หากเปรียบเทียบค่า k ตามจำนวน Categories ตามที่เราทราบอยู่แล้ว พบว่าในค่า k=5 ที่ Euclidian Distances มีการแยก Clusters ที่เกือบชัดเจน แต่จะมี Cluster_1 ที่จะทำการแยกตาม Categories ได้ไม่ชัดเจนเหมือนกัน Clusters อื่นๆ (รูปภาพที่ 27)

	Euclidian Distances	Manhatthan Distances
k = 2	-11.118	-11.185
k = 3	-8.764	-8.962
k = 5	-8.419	-6.667
k =7	-8.089	-6.564

รูปภาพที่ 31 : แสดงค่า Davies Bouldin จากการเลือกค่า k และการเปลี่ยนการคำนวณระยะห่าง โดยกรอบสีแดงแสดงค่า Davies Boudin ที่ต่ำที่สุด

เพื่อเพิ่มความน่าเชื่อถือของข้อมูลจากการสังเกตจากกราฟ โดยการคำนวณหาค่า Davies Bouldin (รูปภาพที่ 31) โดยอ้างอิงจาก Parameter คือค่า k กับการเปลี่ยนวิธีการคำนวณหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลาง โดยค่า Davies Boudin เป็นตัวประเมินประสิทธิภาพของการทำ Clustering โดยยิ่งค่าน้อยแสดงถึงการจัดกลุ่มได้ดี นั่นทำให้สรุปได้ว่า Parameter ที่ควรกำหนดเพื่อได้โมเดลที่มีประสิทธิภาพ คือที่ค่า k=2 ตามผลลัพธ์ที่ได้จากการสังเกตกราฟ

5. สรุปผล

จากวัตถุประสงค์ที่กล่าวมาข้างต้น รายงานนี้ได้ทำการทดลองและสร้างแบบจำลอง โดยการเปลี่ยน Parameter อย่างเป็นระบบ เพื่อให้สามารถเปรียบเทียบผลลัพธ์ของการประมวลผลได้อย่างมีประสิทธิภาพ

ผลลัพธ์ที่ได้จากการสร้างโมเดล Text Classification และทดสอบประสิทธิภาพพบว่า โมเดลมีประสิทธิภาพสูงในการแบ่งกลุ่มตาม Categories ที่ได้เตรียมไว้อย่างอัตโนมัติ ผ่านตัวชี้วัด ค่า Accuracy และได้ทำการหาคำที่สำคัญที่สุดที่ใช้ในการแยกประเภทเอกสารจากค่า TF-IDF

รวมถึงการทำ Text Clustering เพื่อทดสอบประสิทธิภาพในการจัดกลุ่ม ที่อ้างอิงจาก Categories ที่มีอยู่แล้วนั้น พบว่ามีประสิทธิภาพในการจัดกลุ่มอย่างชัดเจนจากผลลัพธ์ที่ได้แสดง ผ่านตัวชี้วัดในรูปแบบของกราฟ และการคำนวณหาค่า Davies Bouldin

6. อ้างอิง

ขอบคุณชุดข้อมูลจากเว็บไซต์ https://www.kaggle.com/datasets/shivamkushwaha/bbc-full-text-document-classification/data