การวิเคราะห์ความรู้สึกด้วยวิธีการจัดกลุ่มโดเมนแบบอัตโนมัติ

วิทยานิพนธ์

ของ

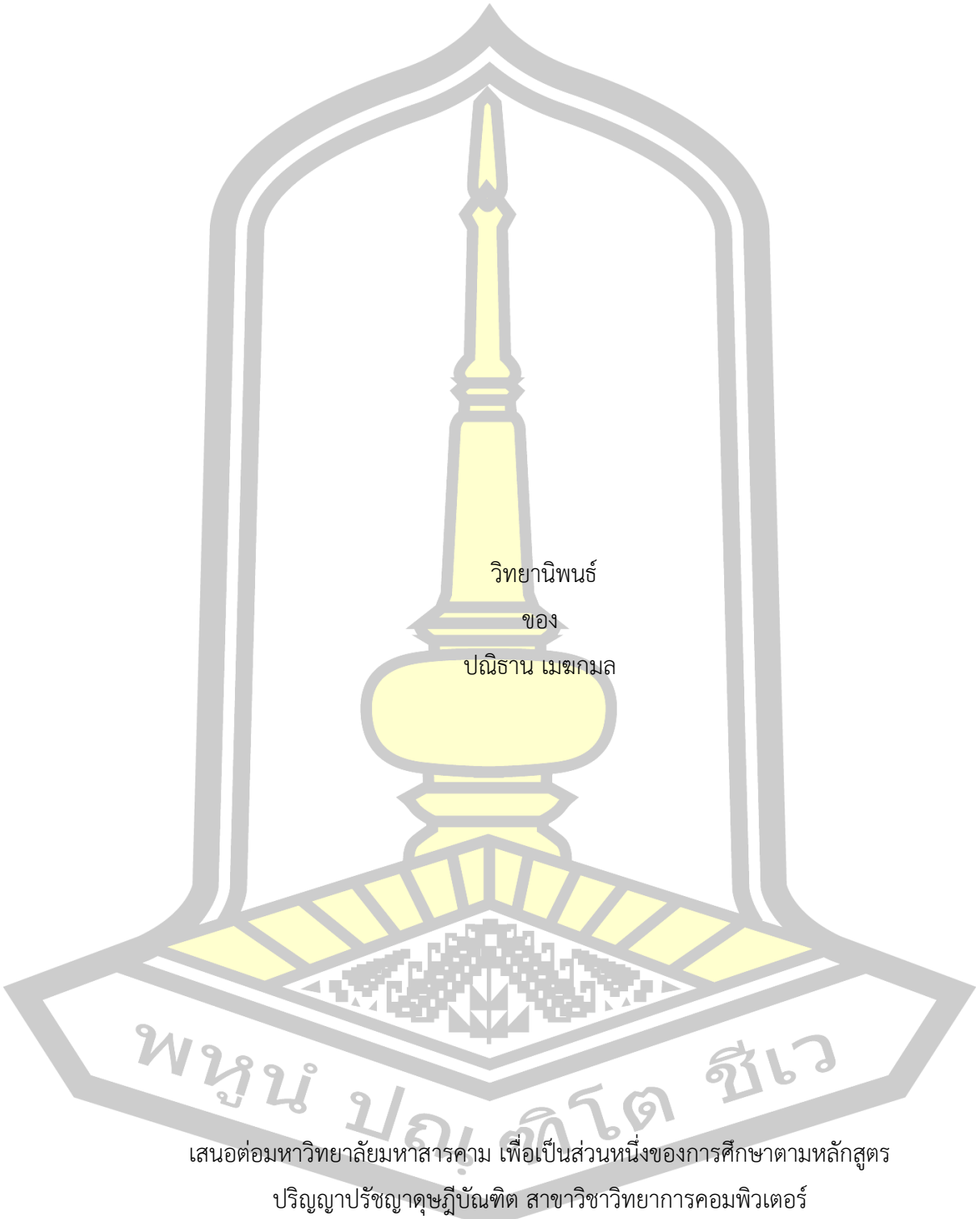
ปณิธาน เมฆกมล

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

มิถุนายน 2564

การวิเคราะห์ความรู้สึกด้วยวิธีการจัดกลุ่มโดเมนแบบอัตโนมัติ



วิทยานิพนธ์

ของ

ปณิธาน เมฆกมล

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

มิถุนายน 2564

Sentiment Analysis by Automatic Domain Clustering

Panithan Mekkamol

A Thesis Submitted in Partial Fulfillment of Requirements

for Doctor of Philosophy (Computer Science)

June 2021

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายปณิธาน เมฆกมล แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รศ. ดร. กฤษณพงศ์ สมสุข)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. ฉัตรเกล้า เจริญผล)

.....กรรมการ
(ผศ. ดร. พัฒนพงษ์ ชมภูวิเศษ)

.....กรรมการ
(ผศ. ดร. พนิดา ทรงรัมย์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา ปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

.....
(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การวิเคราะห์ความรู้สึกด้วยวิธีการจัดกลุ่มโดเมนแบบอัตโนมัติ
ผู้วิจัย ปณิธาน เมฆกมล
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. ฉัตรเกล้า เจริญผล
ปริญญา ปรัชญาดุษฎีบัณฑิต **สาขาวิชา** วิทยาการคอมพิวเตอร์
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2564

บทคัดย่อ

งานวิจัยนี้มีเป้าหมายเพื่อพัฒนาขั้นตอนวิธีจัดกลุ่มของเอกสารโดเมนที่แตกต่างกันตามความคล้ายคลึงกันของโดเมน เพื่อหาวิธีที่บ่งบอกได้เอกสารนั้นควรจะอยู่รวมในโดเมนหรือควรจะแยกเป็นกลุ่มของโดเมนใหม่ และแก้ปัญหาการจัดกลุ่มของขั้นตอนวิธี K-Means ที่จะทำการวัดความคล้ายคลึงของข้อมูลเมื่อพบว่าข้อมูลนั้นใกล้จุดศูนย์กลางของกลุ่มใด ข้อมูลจะถูกจัดให้อยู่ในกลุ่มนั้นเช่นเดียวกับข้อมูลที่มีค่าปกติ ซึ่งข้อมูลนั้นอาจจะมีความเกี่ยวข้องกับข้อมูลที่อยู่ในกลุ่มน้อยมากหรือไม่มีความเกี่ยวข้องเลยก็ได้ ในการทดลองนี้จะดูจำนวนกลุ่มที่เหมาะสมก่อนกำหนดจำนวนกลุ่มด้วยวิธี Elbow เมื่อจัดกลุ่มเอกสารแล้ว จะคำนวณหาระยะทางของเอกสารแต่ละตัวกับจุดศูนย์กลางเพื่อคำนวณหาค่า Threshold ของกลุ่ม เมื่อเอกสารใหม่เข้าไปจะคำนวณระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถ้าใกล้กลุ่มใดมากที่สุด จะนำระยะทางไปเปรียบเทียบกับค่า Threshold ของกลุ่มนั้น ผู้วิจัยนำเสนอการหาค่า Threshold ที่เหมาะสมโดยการหาระยะทางจากจุดศูนย์กลางของเอกสารในแต่ละกลุ่มแล้วหาตำแหน่งเปอเซนไทร์ของข้อมูลในกลุ่มนั้น ผู้วิจัยได้เปรียบเทียบประสิทธิภาพของการให้ค่าน้ำหนักของคุณลักษณะพบว่า TF-IDF ให้ผลลัพธ์ที่ดีกว่า BM25 การวัดความคล้ายคลึงของเอกสารที่ใช้ในขั้นตอนวิธีที่นำเสนอพบว่า Euclidean Distance ให้ผลลัพธ์ดีที่สุดเมื่อเปรียบเทียบกับวิธีอื่น ๆ ขั้นตอนวิธีที่ผู้วิจัยนำเสนอนั้นสามารถแยกเอกสารออกจากกลุ่มได้และเมื่อส่งเอกสารกลุ่มเดิมเข้าไปทดสอบสามารถจัดเอกสารเข้ากลุ่มเดิมได้ โดยกำหนดตำแหน่งเปอเซนไทล์ที่ 80 - 85 ซึ่งขั้นตอนวิธีและการกำหนดค่า Threshold ในงานวิจัยนี้มีประสิทธิภาพกว่างานวิจัยก่อนหน้า

คำสำคัญ : การจัดกลุ่มเอกสาร, การตรวจหาค่าผิดปกติ, ขั้นตอนวิธี C

TITLE Sentiment Analysis by Automatic Domain Clustering

AUTHOR Panithan Mekkamol

ADVISORS Assistant Professor Chatklaw Jareanpon , Ph.D.

DEGREE Doctor of Philosophy **MAJOR** Computer Science

UNIVERSITY Mahasarakham University **YEAR** 2021

ABSTRACT

This research aims to develop the clustering algorithm that the different domain can group using the domain similarity. This research tries to find to method that the document can classify to the previous domain or crate the new domain, and solves the problem of K-means. This problem is coming from the distance measurement of similarity from the new document to the centroid of each group. The new document will classify to the group that the relationship between groups and new document possibly are analogous or divergent. This experiment observes the proper group numbers using the Elbow before starting the process. After this process, the Threshold value will be calculated from the centroid of the document in the group and percentile. The new document will compare with the Threshold and decision to set to the group or create the new document. This research compares the performance of the weight between the TF-IDF and BM25. These results show that the best performance is came from the BM25, Euclidean distance and 80-85 percentile. The result of this research is more accuracy than the previous research.

Keyword : document clustering, outlier detection, C algorithm

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปด้วยดี ด้วยความช่วยเหลือของผู้ช่วยศาสตราจารย์ ดร.ฉัตรเกล้า เจริญผล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำ ชี้แนะแนวทางและข้อคิดเห็นต่าง ๆ อันเป็นประโยชน์อย่างยิ่งในการทำวิจัย อีกทั้งยังช่วยแก้ปัญหาต่าง ๆ ที่เกิดขึ้นระหว่างการดำเนินงานอีกด้วย

ขอกราบขอบคุณคณะกรรมการสอบวิทยานิพนธ์ประกอบด้วย รองศาสตราจารย์ ดร.กฤษณพงศ์ สมสุข ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.พัฒนพงษ์ ชมพูวิเศษ ผู้ช่วยศาสตราจารย์ ดร. พนิดา ทรงรัมย์ กรรมการสอบวิทยานิพนธ์ ที่ได้ชี้แนะแนวทางและคำแนะนำตลอดจนข้อสังเกตต่าง ๆ ทำให้ผู้เขียนได้พัฒนาแนวคิดและไตร่ตรองปัญหาต่าง ๆ ได้อย่างรอบคอบมากยิ่งขึ้นจนทำให้วิทยานิพนธ์ฉบับนี้ สำเร็จลงได้

ผู้เขียนขอกราบขอบพระคุณคณาจารย์ในหลักสูตรวิทยาการคอมพิวเตอร์ทุกท่านที่ได้อบรมสั่งสอนให้ความรู้ทางด้านวิชาการแก่ผู้เขียน ทำให้ผู้เขียนได้รู้ว่าความรู้ไม่มีที่สิ้นสุด ขอขอบคุณน้อง ๆ สาขาวิทยาการคอมพิวเตอร์ที่คอยช่วยเหลือตลอดระยะเวลาที่ศึกษา ขอบคุณพี่ ๆ น้อง ๆ ในสาขาวิชาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏอุดรธานี ซึ่งเป็นที่ทำงานของผู้เขียนที่คอยช่วยเหลือในระยะเวลาที่ผู้เขียนศึกษาต่อ

สุดท้ายนี้ผู้เขียนกราบขอบพระคุณบิดา มารดา ภรรยาและบุตรชายอันเป็นที่รักยิ่ง ที่เป็นกำลังใจสำคัญให้ตลอดเวลาในการศึกษาเล่าเรียน หากวิทยานิพนธ์ฉบับนี้มีประโยชน์และคุณค่าทางการศึกษา ผู้เขียนขอยกความดีทั้งหมดให้กับบุคคลที่ผู้เขียนได้กล่าวถึง

ปณิธาน เมฆกมล

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง ..... ฎ

สารบัญภาพ ..... ฐ

บทที่ 1 บทนำ 1

1.1 หลักการและเหตุผล 1

1.2 ปัญหาการวิจัย 2

1.3 วัตถุประสงค์การวิจัย 3

1.4 ความสำคัญของการวิจัย 3

1.5 ขอบเขตของการวิจัย 3

1.6 นิยามศัพท์เฉพาะ 4

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 5

2.1 การทำเหมืองข้อมูล 5

2.1.1 ขั้นตอนการทำเหมืองข้อมูล 5

2.2 การทำเหมืองข้อความ 6

2.3 การเตรียมข้อมูลก่อนประมวลผลสำหรับการจัดกลุ่มเอกสาร 6

2.3.1 การตัดคำ (Word Tokenize) 7

2.3.2 การหารากศัพท์ (Stemming) 8

2.3.3 การกำจัดคำหยุด (Stop-words removal) 9

2.4 แบบจำลองปริภูมิเวกเตอร์ (Vector space model : VSM) 10
2.5 การเลือกคุณลักษณะ (Feature Selection) 10
2.5.1 การคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำด้วย Term Frequency - Inverse Document Frequency 12
2.5.2 การคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำด้วย BM25 13
2.6 การจัดกลุ่มเอกสาร (Document Clustering) 14
2.6.1 การจัดกลุ่มเอกสารด้วยขั้นตอนวิธี K-Means 14
2.6.2 ขั้นตอนวิธีการจัดกลุ่มด้วย DBSCAN 16
2.7 การหาจำนวนกลุ่มที่เหมาะสม 18
2.8 การหาความคล้ายคลึงของเอกสาร 19
2.8.1 Euclidean Distance 19
2.8.2 Cosine Distance & Cosine Similarity 20
2.8.3 Manhattan 20
2.8.4 Minkowski 20
2.9 การวัดประสิทธิภาพการจัดกลุ่ม 20
2.9.1 การวัดค่าความถูกต้อง (Accuracy) 21
2.9.2 การวัดค่าความแม่นยำ (Precision) 21
2.9.3 การวัดค่าการจำได้ (Recall) 21
2.9.4 ค่าเฉลี่ยประสิทธิภาพโดยรวม (F-measure or F1 Score) 22
2.10 งานวิจัยที่เกี่ยวข้อง 22
2.10.1 การจัดกลุ่มเอกสาร 22
2.10.2 การตรวจสอบค่าผิดปกติ (Outlier Detection) 23
2.10.3 การให้ค่าน้ำหนักของคำ 24
2.10.4 การวัดความคล้ายคลึงของเอกสาร 25

บทที่ 3 วิธีการดำเนินการวิจัย 26
3.1 ชุดข้อมูล 27
3.2 ขั้นตอนวิธีในการดำเนินงานวิจัย 28
3.2.1 ขั้นตอนเตรียมข้อมูลก่อนการประมวลผล (Pre Processing) 28
3.3 การคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำ 31
3.4 การหาจำนวนกลุ่มที่เหมาะสม 32
3.5 การจัดกลุ่มเอกสาร 33
3.6 C- Algorithm 33
3.7 การหาค่า Threshold เพื่อใช้ในการแยกเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบและการจัดเอกสารเดิมเข้ากลุ่ม 35
3.8 การทดสอบประสิทธิภาพของ C-Algorithm และการวิธีวัดความคล้ายคลึงของเอกสารด้วยเทคนิคต่าง ๆ 37
3.9 การวัดประสิทธิภาพ 37
บทที่ 4 ผลการวิจัย 38
4.1 ข้อมูลที่ใช้ในการทดลอง 39
4.1.1 เครื่องมือและข้อมูลต่าง ๆ ที่ใช้ในการทดลองในงานวิจัย 39
4.1.2 การรวบรวมข้อมูลในการทดลอง 39
4.2 การเตรียมข้อมูล 40
4.3 การคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำ 43
4.3.1 การให้ค่าน้ำหนักกับคุณลักษณะด้วย TF-IDF 43
4.3.2 การให้ค่าน้ำหนักกับคุณลักษณะด้วย BM25 44
4.4 ทดสอบประสิทธิภาพขั้นตอนวิธีในการจัดกลุ่มเอกสาร 47
4.4.1 เปรียบเทียบขั้นตอนวิธีในการจัดกลุ่มเอกสารบนชุดข้อมูล Multi domain 47
4.5 ขั้นตอนวิธีในการพิจารณาเอกสารจะอยู่ในกลุ่มหรือแยกกลุ่ม 48

4.5.1 ผลการตรวจสอบจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มเอกสาร 48
4.5.2 ผลการทดลองส่งเอกสารอีกโดเมนหนึ่งเข้าไปทดสอบบนชุดข้อมูลบนชุดข้อมูล Multi-Domain Sentiment dataset 49
4.5.3 ผลการทดลองส่งเอกสารอีกโดเมนหนึ่งเข้าไปทดสอบบนชุดข้อมูล 20 Newsgroup text Dataset 53
4.5.4 ผลการทดลองด้วยเอกสารทั้งหมดของทุกกลุ่ม 57
4.6 เปรียบเทียบการจัดกลุ่มเอกสารที่ส่งเข้าไปทดสอบด้วยขั้นตอนวิธีที่นำเสนอกับวิธีการเดิม 62
4.7 แสดงคุณลักษณะข้อมูลที่อยู่ในแต่ละกลุ่ม 63
บทที่ 5 สรุปผลการวิจัย 68
5.1 สรุปและอภิปรายผล 68
5.2 ข้อเสนอแนะ 70
บรรณานุกรม 71
ประวัติผู้เขียน 78

สารบัญตาราง

หน้า

ตารางที่ 1 แสดงตัวอย่างคำหยุดในภาษาอังกฤษ 9

ตารางที่ 2 แสดงความแตกต่างระหว่าง K-means Clustering และ DBSCAN Clustering..... 17

ตารางที่ 3 Confusion Table 21

ตารางที่ 4 แสดงจำนวนชุดข้อมูลที่ใช้ในงานวิจัย..... 27

ตารางที่ 5 ตัวอย่างการหารากศัพท์ของคำ..... 31

ตารางที่ 6 แสดง Document – term matrix (DTM)..... 31

ตารางที่ 7 ตารางเปรียบเทียบเทคนิคการหาจำนวนกลุ่มที่เหมาะสม..... 32

ตารางที่ 8 ตัวอย่างการเรียงลำดับระยะทางของข้อมูลแต่ละกลุ่มจากจุดศูนย์กลาง 35

ตารางที่ 9 ตัวอย่างการหารากศัพท์ของคำบนชุดเอกสารที่ใช้ทดลอง 42

ตารางที่ 10 แสดงการให้ค่าน้ำหนักกับคุณลักษณะด้วย TF-IDF บนชุดข้อมูล Multi Domain 43

ตารางที่ 11 แสดงการให้ค่าน้ำหนักกับคุณลักษณะด้วย TF-IDF บนชุดข้อมูล 20 News Group ... 44

ตารางที่ 12 แสดงการให้ค่าน้ำหนักกับคุณลักษณะด้วย BM25 บนชุดข้อมูล Multi Domain Dataset..... 45

ตารางที่ 13 แสดงการให้ค่าน้ำหนักกับคุณลักษณะด้วย BM25 บนชุดข้อมูล 20 News Group..... 45

ตารางที่ 14 Paired-Sample T-Test บนชุดข้อมูล Multi Domain Sentiment Dataset 46

ตารางที่ 15 Paired-Sample T-Test บนชุดข้อมูล 20 News Group Dataset 47

ตารางที่ 16 แสดงผลของการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี K-Means และ DBScan..... 47

ตารางที่ 17 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Euclidean Distance บนชุดข้อมูล Multi-Domain Sentiment dataset 50

ตารางที่ 18 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Manhattan Distance บนชุดข้อมูล Multi-Domain Sentiment dataset 50

ตารางที่ 19 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Minkowski Distance บนชุดข้อมูล Multi-Domain Sentiment dataset 51

ตารางที่ 20 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Cosine บนชุดข้อมูล Multi-Domain Sentiment dataset..... 51

ตารางที่ 21 ตารางเปรียบเทียบค่า Recall ของวิธีการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi-Domain Sentiment dataset 52

ตารางที่ 22 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Euclidean Distance บนชุดข้อมูล 20 Newsgroup..... 53

ตารางที่ 23 แสดงผลการทำนายเอกสารกลุ่มที่ 3 (Test Documents) ด้วย Manhattan Distance บนชุดข้อมูล 20 Newsgroup 54

ตารางที่ 24 แสดงผลการทำนายเอกสารกลุ่มที่ 3 (Test Documents) ด้วย Minkowski Distance บนชุดข้อมูล 20 Newsgroup 54

ตารางที่ 25 แสดงผลการทำนายเอกสารกลุ่มที่ 3 (Test Documents) ด้วย Cosine บนชุดข้อมูล 20 Newsgroup..... 55

ตารางที่ 26 ตารางเปรียบเทียบค่า Recall ของวิธีการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group Text Dataset ทดลองด้วยเอกสารกลุ่มที่ 3 (Test Documents)..... 56

ตารางที่ 27 แสดงค่า Average Recall ของเอกสารทุกกลุ่มด้วยการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi Domain Sentiment..... 57

ตารางที่ 28 แสดงค่า F1 ของเอกสารทุกกลุ่มด้วยการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi Domain Sentiment 58

ตารางที่ 29 แสดงค่า Average Recall ของเอกสารทุกกลุ่มด้วยการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group Dataset 59

ตารางที่ 30 แสดงค่า F1 ของเอกสารทุกกลุ่มด้วยการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group Text Dataset..... 60

ตารางที่ 31 เปรียบเทียบการจัดกลุ่มเอกสารใหม่ด้วยขั้นตอนวิธีเดิมกับขั้นตอนวิธีที่นำเสนอ 62

สารบัญภาพ

หน้า

รูปที่ 1 แสดงการตัดคำแบบต่าง ๆ 8

รูปที่ 2 แสดง Document Term Matrix บน Vector space Model 10

รูปที่ 3 แสดงตัวอย่างการให้ค่าน้ำหนักของคำ 11

รูปที่ 4 แสดง K-Means Algorithm 15

รูปที่ 5 แสดงการการจัดกลุ่มด้วยขั้นตอนวิธี K-Means 15

รูปที่ 6 แสดงขั้นตอนวิธี DBSCAN 16

รูปที่ 7 แสดงการจัดกลุ่มด้วย DBScan 17

รูปที่ 8 elbow Method 19

รูปที่ 9 กรอบแนวคิดในการดำเนินการวิจัย 26

รูปที่ 10 แสดงขั้นตอนการเตรียมก่อนการประมวลผล 28

รูปที่ 11 ตัวอย่างเอกสารต้นฉบับ 29

รูปที่ 12 ตัวอย่างเอกสารหลังจากตัดคำ 29

รูปที่ 13 ตัวอย่างเอกสารหลังจากเปลี่ยนตัวอักษร 29

รูปที่ 14 แสดงคำที่ผสมระหว่างตัวอักษรและตัวเลข 30

รูปที่ 15 แสดง elbow Method 33

รูปที่ 16 แสดงการหาระยะทางของข้อมูลจากจุดศูนย์กลางแต่ละกลุ่ม 35

รูปที่ 17 ตัวอย่างเอกสารของชุดข้อมูล Multi Domain Sentiment Dataset 40

รูปที่ 18 ตัวอย่างเอกสารของชุดข้อมูล 20 News Group Dataset 40

รูปที่ 19 ตัวอย่างเอกสารต้นฉบับ 41

รูปที่ 20 ตัวอย่างเอกสารหลังจากตัดคำ 41

รูปที่ 21 ตัวอย่างเอกสารหลังจากเปลี่ยนตัวอักษร 41

รูปที่ 22 ตัวอย่างเอกสารหลังจากกำจัดหยุด 42

รูปที่ 23 แสดงตัวอย่างของคำที่ผสมระหว่างตัวอักษรและตัวเลขในชุดข้อมูลที่ทำการทดลอง 42

รูปที่ 24 แสดงกราฟเปรียบเทียบค่า Recall และ Accuracy ของการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี K-Means และ DBScan 48

รูปที่ 25 แสดงจำนวนกลุ่มที่เหมาะสมของชุดข้อมูล Multi-Domain Sentiment dataset 49

รูปที่ 26 แสดงจำนวนกลุ่มที่เหมาะสมของชุดข้อมูล 20NewsGroup 49

รูปที่ 27 แสดงการเปรียบเทียบค่า Recall ของวิธีวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi-Domain Sentiment dataset 53

รูปที่ 28 แสดงการเปรียบเทียบค่า Recall ของวิธีวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group 56

รูปที่ 29 แสดงแผนภูมิเปรียบเทียบค่า Recall ของขั้นตอนที่นำเสนอด้วยวิธีการวัดความคล้ายคลึงแบบต่าง ๆ บนเอกสารทั้ง 3 กลุ่ม..... 58

รูปที่ 30 แสดงการเปรียบเทียบค่า F1 ของวิธีวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi Domain Dataset ของเอกสารทั้ง 3 กลุ่ม..... 59

รูปที่ 31 แสดงแผนภูมิเปรียบเทียบค่า Recall ของขั้นตอนที่นำเสนอด้วยวิธีการวัดความคล้ายคลึงแบบต่าง ๆ บนเอกสารทั้ง 3 กลุ่ม บนชุดข้อมูล 20 News Group Dataset..... 60

รูปที่ 32 แสดงการเปรียบเทียบค่า F1 ของวิธีวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group Text Dataset 61

รูปที่ 33 แผนภูมิเปรียบเทียบการจัดกลุ่มเอกสารใหม่ด้วยขั้นตอนวิธีเดิมกับขั้นตอนวิธีที่นำเสนอ... 63

รูปที่ 34 แสดงคุณลักษณะที่สำคัญของเอกสารกลุ่มที่ 1 50 ลำดับแรก 63

รูปที่ 35 แสดงคุณลักษณะที่สำคัญของเอกสารกลุ่มที่ 2 50 ลำดับแรก 64

รูปที่ 36 แสดงคุณลักษณะที่สำคัญของเอกสารกลุ่มใหม่ 50 ลำดับแรก..... 64

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

โลกในยุคปัจจุบันอยู่ในยุคดิจิทัล การใช้งานข้อมูลในรูปแบบดิจิทัล ถือว่าเป็นผลดีเนื่องจากสามารถลดปริมาณการใช้กระดาษน้อยลง ร่นระยะเวลาในการติดต่อสื่อสาร แต่ส่งผลให้เกิดข้อมูลปริมาณมหาศาล โดยการสร้างข้อมูลดิจิทัลได้มาจากหลากหลายอุปกรณ์ ซึ่งข้อมูลมากกว่า 80% อยู่ในรูปแบบของข้อความหรือตัวอักษร [1] ข้อมูลประเภทนี้เป็นข้อมูลที่ไม่มีโครงสร้างหรือกึ่งโครงสร้าง ซึ่งไม่ได้มีการจัดประเภทไว้มีทั้งข้อมูลที่เป็นประโยชน์ ข้อมูลที่ไม่เป็นประโยชน์ ข้อมูลทางวิทยาศาสตร์ ข้อมูลทางธุรกิจ และในด้านอื่น ๆ บริษัทจำนวนร้อยละ 33 วิเคราะห์ข้อมูลเพื่อสกัดหาข้อมูลที่น่าสนใจจากรูปแบบที่ซ่อนไว้ด้วยกระบวนการทำเหมืองข้อความ [2]

การทำเหมืองข้อความ (Text Mining) จะคล้ายกับการทำเหมืองข้อมูล (Data Mining) [3] จะแตกต่างกันที่การทำเหมืองข้อมูลจะทำบนข้อมูลที่มีโครงสร้างในขณะที่เหมืองข้อความจะทำบนข้อมูลที่ไม่มีโครงสร้างหรือกึ่งโครงสร้าง [1] เช่น email เอกสาร ข้อความบนสื่อสังคมออนไลน์ เอกสาร HTML และอื่น ๆ การทำเหมืองข้อความอาจเรียกได้ในชื่อ Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT) ซึ่งโดยทั่วไปหมายถึงการสกัดข้อมูลและความรู้ที่น่าสนใจจากข้อความที่ไม่มีโครงสร้าง การทำเหมืองข้อความนำไปใช้ในงานสืบค้นสารสนเทศ การทำเหมืองข้อมูล การเรียนรู้ของเครื่อง สถิติ และภาษาศาสตร์คอมพิวเตอร์ [1]

การจัดกลุ่มเอกสารเป็นวิธีการหนึ่งของการทำเหมืองข้อความ เป็นการจัดกลุ่มโดย ใช้คุณสมบัติความคล้ายคลึงกันของคำในเอกสาร ซึ่งเอกสารที่จัดอยู่ในกลุ่มเดียวกัน จะมีกลุ่มคำดรรชนี (Index terms) เหมือนกันแต่จะต่างจากเอกสารในกลุ่มอื่น ๆ โดยจัดกลุ่มของเอกสารที่คล้ายกันไว้ในกลุ่มเดียวกันและที่ต่างกันก็ให้แยกกลุ่มกัน การที่ทำการจัดกลุ่มตามความคล้ายและต่างกันนั้น ก็เพื่อเพิ่ม Recall/Precision รวมทั้งลดคุณลักษณะที่ไม่เกี่ยวข้องในการสืบค้น กล่าวคือแทนที่จะค้นหาจากเอกสารทั้งหมด ก็เพียงจำกัดเฉพาะเอกสารที่อยู่ในกลุ่มเดียวกัน ทำให้ขอบเขตงานแคบลง การค้นหาจะยังทำได้เร็วขึ้น

เทคนิคในการจัดกลุ่มเอกสารมี 3 ประเภทหลักคือ Partitional Clustering Hierarchical Clustering และ Density based Clustering เทคนิคการจัดกลุ่มแบบ Partitional Clustering เป็นวิธีการแบ่งข้อมูลเป็นกลุ่มย่อยโดยไม่ทับซ้อนกันโดยข้อมูลจะอยู่ในกลุ่มเดียวตัวอย่างเช่น K-means K-medoids ในส่วนของ Hierarchical Clustering ชุดของกลุ่มที่ซ้อนกันจะถูกจัดให้เป็นโครงสร้างแบบลำดับขั้น [4] เรียกว่า dendogram ตัวอย่างเช่น BIRCH CURE ROCK และ Density Based Clustering จะใช้การวัดระยะทางที่แตกต่างกันและจำนวนคลัสเตอร์จะถูกกำหนดโดยอัลกอริทึมโดย

อัตโนมัติข้อมูลจะถูกแยกออกโดยขึ้นอยู่กับการเชื่อมต่อ ขอบเขตและความหนาแน่นของของมูล [5] ทั้งสามแบบมีการใช้งานที่หลากหลายในงานวิจัย [6] [7] [8]

ขั้นตอนวิธี K – Means มักจะถูกใช้ในการจัดกลุ่มเอกสารเนื่องจากให้ผลลัพธ์ที่ดี [9] เป็นที่นิยมซึ่งใช้ในการระบุความคล้ายคลึงกันระหว่างวัตถุขึ้นอยู่กับเวกเตอร์ระยะทาง (distance vectors) [10] ขั้นตอนวิธี K - Means เป็นการจัดกลุ่มแบบ centroid-based หรือ partition-based แต่ก็ยังมีข้อด้อยของการจัดกลุ่มด้วยขั้นตอนวิธี K-Means คือ ต้องระบุจำนวนกลุ่ม (คลัสเตอร์) ที่ต้องการซึ่งบางครั้งไม่รู้ว่าจำนวนกลุ่มหรือค่า K ที่เหมาะสมควรเป็นเท่าไร ซึ่งค่า K มีความสำคัญอย่างมากต่อประสิทธิภาพของอัลกอริทึม [11] เทคนิคในการหาจำนวนกลุ่มที่เหมาะสมหลายเทคนิคที่ใช้ในการจัดกลุ่มอาทิ Syakur ใช้เทคนิค Elbow ในการหาจำนวนกลุ่มที่เหมาะสมในการระบุกลุ่มลูกค้าที่มีโปรไฟล์ดีที่สุด [12] Nainggolan และคณะ [13] ใช้ Sum Of Squared Error (SSE) เพื่อหาจำนวนกลุ่มที่เหมาะสมให้กับการจัดกลุ่มด้วย K-Means นอกจากนี้แล้วในการจัดกลุ่มด้วย K-Means จะทำงานได้ไม่ดีกับชุดข้อมูลที่มีค่าผิดปกติ (outliers) และข้อมูลรบกวน (noisy) [14]

ดังนั้นการจัดกลุ่มด้วยขั้นตอนวิธี K-Means ควรมีการจัดการกับข้อมูลที่มีค่าผิดปกติก่อนเพื่อประสิทธิภาพที่ดีของการจัดกลุ่ม แต่เมื่อกระบวนการจัดกลุ่มเสร็จแล้วข้อมูลที่อยู่ในกลุ่มจะไม่ถือว่าเป็นข้อมูลที่มีค่าผิดปกติอีก เมื่อส่งข้อมูลใหม่เข้าไปตรวจสอบว่าข้อมูลนั้นควรอยู่ในกลุ่มใด ด้วยขั้นตอนวิธีของ K-Means จะทำการวัดความคล้ายคลึงของข้อมูลเมื่อพบว่าข้อมูลนั้นใกล้จุดศูนย์กลางของกลุ่มใด ข้อมูลจะถูกจัดให้อยู่ในกลุ่มนั้นเช่นเดียวกับข้อมูลที่มีค่าปกติ ซึ่งข้อมูลนั้นอาจจะมีความเกี่ยวข้องกับข้อมูลที่อยู่ในกลุ่มน้อยมากหรือไม่มีความเกี่ยวข้องเลยก็ได้ ซึ่งจะทำให้ประสิทธิภาพในการจำแนกข้อมูลว่าอยู่ในกลุ่มใดมีความผิดพลาดได้

งานวิจัยนี้ผู้วิจัยมีเป้าหมายเพื่อพัฒนาขั้นตอนวิธีการจัดกลุ่มของเอกสารในโดเมนที่แตกต่างกันด้วยการหาความคล้ายคลึงของเอกสารในแต่ละโดเมนเพื่อบ่งบอกว่าข้อมูลที่นำไปจัดกลุ่มนั้นควรจะอยู่ในกลุ่มของโดเมนนั้นหรือแยกออกมาสร้างกลุ่มของโดเมนใหม่

1.2 ปัญหาการวิจัย

ขั้นตอนวิธีใดที่ใช้ในการจัดกลุ่มของโดเมนที่แตกต่างกันตามความคล้ายคลึงกันของโดเมนเพื่อบ่งบอกว่าข้อมูลที่นำไปจัดกลุ่มนั้นควรจะอยู่ในกลุ่มของโดเมนนั้นหรือแยกออกมาสร้างกลุ่มของโดเมนใหม่และให้ประสิทธิภาพที่ดีที่สุด

1.3 วัตถุประสงค์การวิจัย

1. เพื่อพัฒนาขั้นตอนวิธีการจัดกลุ่มเอกสารบนโดเมนที่แตกต่างกันตามความคล้ายคลึงกันของโดเมน และวิธีการคัดแยกว่าเอกสารนั้นควรจะอยู่รวมในโดเมนหรือแยกออกมาสร้างกลุ่มของโดเมนใหม่
2. เพื่อเปรียบเทียบประสิทธิภาพการให้ค่าน้ำหนักของคำและการวัดความคล้ายคลึงของเอกสาร

1.4 ความสำคัญของการวิจัย

1. สามารถนำไปจัดกลุ่มของข้อความโดยการพิจารณาถึงความคล้ายคลึงของข้อมูลถึงแม้ว่าข้อมูลจะอยู่คนละโดเมน
2. เพื่อใช้ในการกำหนดโดเมนต้นทางที่จะนำไปใช้ในงานวิเคราะห์ความรู้สึกหรืองานอื่น ๆ ในการประมวลผลภาษาธรรมชาติได้
3. การพิจารณาว่าข้อมูลใดที่ไม่ควรอยู่ในกลุ่มจะสามารถเพิ่มความถูกต้องในการจัดกลุ่มข้อมูล
4. เพื่อแก้ปัญหาของการกำหนดกลุ่มด้วยขั้นตอนวิธี K-Means

1.5 ขอบเขตของการวิจัย

1. นำเสนอขั้นตอนวิธีจัดกลุ่มของโดเมนที่แตกต่างกันตามความคล้ายคลึงกันของโดเมน เพื่อหาวิธีที่บ่งบอกได้เอกสารนั้นควรจะอยู่รวมในโดเมนนั้นหรือควรจะแยกออกมาสร้างกลุ่มของโดเมนใหม่
2. งานวิจัยนี้ผู้วิจัยเลือกใช้ชุดข้อมูล 2 ชุดได้แก่
   2.1 ชุดข้อมูล Multi-Domain Sentiment dataset ซึ่งรวบรวมโดย Blitzer และคณะ จาก Amazon.com ประกอบไปด้วยสินค้าที่แตกต่างกัน 3 ประเภทคือ Book, DVDs, electronics โดยเลือกตัวอย่างรีวิวจากโดเมนทั้ง 3 อย่างละ 600 รีวิวรวมเป็น 1,800 ตัวอย่าง
   2.2 ชุดข้อมูล 20 newsgroups text dataset จำนวน 3 โดเมน ได้แก่ เอกสารโดเมน alt.atheism มีจำนวน 480 เอกสาร misc.forsale มีจำนวน 585 เอกสาร sci.electronics มีจำนวน 591 เอกสาร รวมทั้งสิ้น 1,656 เอกสาร
3. การให้น้ำหนักของคุณลักษณะที่ใช้ในงานวิจัยทำการศึกษา 2 เทคนิคคือ
   3.1 Term frequency – inverse document frequency : tf-idf
   3.2 Best Match 25 : BM 25
   3.3 เปรียบเทียบประสิทธิภาพการให้น้ำหนักทั้งสองวิธีด้วย Pair Sample T-Test
4. ขั้นตอนวิธีในการจัดกลุ่มเอกสารที่ใช้ในงานวิจัยทำการศึกษา 2 ขั้นตอนวิธีคือ
   4.1 K-Means
   4.2 DBSCAN

5. เทคนิคที่ใช้ในการวัดความคล้ายคลึงของเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบในงานวิจัยนี้ทำการศึกษา 4 เทคนิคคือ

5.1 Euclidean Distance

5.2 Cosine Similarity

5.3 Manhattan

5.4 Minkowski

6. วิธีการพิจารณาข้อมูลว่าควรอยู่ในกลุ่มหรือไม่ควรอยู่ด้วยการหาตำแหน่ง Percentile ของชุดข้อมูลในกลุ่มเพื่อหาค่า Threshold ในการพิจารณาว่าข้อมูลนั้นควรอยู่ในกลุ่มหรือนอกกลุ่ม
7. การประเมินการจัดกลุ่มเอกสารจะประเมินด้วยค่าความถูกต้อง (Accuracy), ค่าความระลึก (Recall), ค่าความแม่นยำ (Precision), ค่า F1

1.6 นิยามศัพท์เฉพาะ

โดเมน (Domain) หมายถึง กลุ่มของเอกสารที่เกี่ยวข้องกันเช่น โดเมนภาพยนตร์ ก็จะมีเอกสารที่เกี่ยวกับภาพยนต์เท่านั้น

ค่าผิดปกติ (Outlier) หมายถึง ค่าของข้อมูลที่ไม่ใกล้เคียงกับข้อมูลตัวอื่น ๆ ในกลุ่ม

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงแนวคิด ทฤษฎีที่เกี่ยวข้อง การทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องกับรายงานงานวิจัยนี้ โดยมีรายละเอียดดังต่อไปนี้

2.1 การทำเหมืองข้อมูล

การทำเหมืองข้อมูล คือการนำเทคนิคการเรียนรู้ของเครื่อง วิธีการทางสถิติ วิธีการทางปัญญาประดิษฐ์ มาวิเคราะห์และสกัดความรู้จากข้อมูลที่จัดเก็บไว้ในรูปแบบต่าง ๆ มีจุดประสงค์เพื่อวิเคราะห์ หาแนวโน้ม ความสัมพันธ์ กฎหรือรูปแบบของข้อมูล

2.1.1 ขั้นตอนการทำเหมืองข้อมูล

ขั้นตอนการทำเหมืองข้อมูล ประกอบไปด้วย 6 ขั้นตอนมาตรฐานเรียกว่า Cross-Industry Standard Process for Data Mining (CRISP-DM) ดังนี้

1. ทำความเข้าใจกับปัญหา (Business Understanding) เป็นการทำความเข้าใจกับปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปแบบของการวิเคราะห์ข้อมูลด้วยวิธีการ Data Mining และวางแผนในการดำเนินการคร่าว ๆ
2. ทำความเข้าใจกับข้อมูล (Data Understanding) เป็นการรวบรวมข้อมูลที่เกี่ยวข้อง เชื่อถือได้ ปริมาณเพียงพอต่อการนำไปใช้วิเคราะห์
3. การเตรียมข้อมูล (Data Preparation) เป็นขั้นตอนที่ใช้เวลานานที่สุด เนื่องจากโมเดลที่ได้จากการทำเหมืองข้อความจะให้ผลลัพธ์ถูกต้องหรือไม่นั้นขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ แบ่งเป็น 3 ขั้นตอนคือ
   3.1 การคัดเลือกข้อมูล (Data Selection) คือการเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่จะวิเคราะห์
   3.2 การกลั่นกรองข้อมูล (Dara Cleaning) ขั้นตอนนี้เป็นการลบข้อมูลซ้ำซ้อน แก้ไขข้อมูลที่ผิดพลาด ข้อมูลที่ผิดพลาดเช่น ข้อมูลที่ผิดรูปแบบ ข้อมูลที่หายไป ข้อมูลที่แปลกแยกจากข้อมูลอื่น
   3.3 แปลงรูปแบบข้อมูล (Data transformation) เป็นขั้นตอนการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปวิเคราะห์ตามขั้นตอนวิธีที่เลือกใช้
4. สร้างแบบจำลอง (Modeling) เป็นขั้นตอนการวิเคราะห์ด้วยเทคนิคดาต้าไมนิ่ง มี 3 เทคนิคหลักคือ การจำแนก (Classification) การจัดกลุ่ม (Clustering) และ กฎความสัมพันธ์ (Association rules) ซึ่งแต่ละเทคนิคก็จะมีเทคนิคย่อยแตกต่างกันออกไปอีก

5. การประเมิน (Evaluation) เป็นการวัดประสิทธิภาพของโมเดลวิเคราะห์ข้อมูลก่อนหน้านั้น
6. นำไปใช้งาน (Deployment) นำโมเดล หรือผลการวิเคราะห์ที่ได้ไปใช้งานจริง

2.2 การทำเหมืองข้อความ

การทำเหมืองข้อความหรือเรียกได้ว่าเป็นการค้นหาความรู้ในฐานข้อมูลข้อความ (Knowledge Discovery in Textual Databases) เป็นกระบวนการที่จะทำการค้นหาและสกัดข้อมูลที่เป็นประโยชน์จากข้อความ การทำเหมืองข้อความคล้ายกับการทำเหมืองข้อมูล แตกต่างกันตรงที่การทำเหมืองข้อมูลจะทำกับข้อมูลที่โครงสร้างแต่การทำเหมืองข้อความจะทำกับข้อความซึ่งไม่มีโครงสร้าง [15] มีการใช้การทำเหมืองข้อความในหลาย ๆ งานตัวอย่างเช่น

1) การค้นคืนสารสนเทศ (Information Retrieval) การค้นคืนสารสนเทศเป็นการค้นหาข้อความหรือสารสนเทศในเอกสารที่มีเป็นจำนวนมากให้ได้มาอย่างรวดเร็วและสอดคล้องกับความต้องการในการค้นหา
2) การสกัดสารสนเทศ (Information Extraction) เป็นวิธีระบุคำสำคัญและความสัมพันธ์ภายในข้อความ จะกระทำด้วยการมองหาลำดับที่กำหนดไว้ล่วงหน้าในข้อความเรียกว่าการจับคู่รูปแบบ (pattern matching) เพื่อที่จะระบุความสัมพันธ์ระหว่างสถานที่ ผู้คน เวลาและอื่น ๆ เพื่อที่ให้ความหมายสำหรับผู้ใช้ข้อมูล วิธีการนี้จะให้ประโยชน์เป็นอย่างมากเมื่อทำงานกับข้อความจำนวนมาก [1]
3) การประมวลผลภาษาธรรภาษาธรรมชาติ (Natural Language Processing) เป็นการวิจัยและการประยุกต์ใช้เพื่อที่จะให้คอมพิวเตอร์สามารถเข้าใจและจัดการกับข้อความที่เป็นภาษาธรรมชาติในการดำเนินงานที่ต้องการ [16]

2.3 การเตรียมข้อมูลก่อนประมวลผลสำหรับการจัดกลุ่มเอกสาร

การเตรียมข้อมูลก่อนการประมวลผลเป็นขั้นตอนสำคัญในการทำเหมืองข้อความ เนื่องจากข้อมูลที่มีอยู่ในรูปแบบที่ไม่มีโครงสร้างจึงต้องเปลี่ยนให้เป็นรูปแบบที่มีโครงสร้างก่อน ขั้นตอนในการเตรียมข้อมูลมราเป็นข้อความก่อนการนำไปประมวลผลโดยทั่วไปจะมี 6 กระบวนการได้แก่ การตัดคำ การกำจัดคำหยุด การทำความสะอาดข้อความ การหารากศัพท์ การสกัดคุณลักษณะ และการคัดเลือกคุณลักษณะ [17]

S. Vijayarani และคณะ [15] นำเสนอขั้นตอนสำคัญ 3 ขั้นตอนในการเตรียมข้อมูลก่อนการประมวลผลหลังจากตัดคำ (tokenize) แล้วประกอบด้วย การกำจัดคำหยุด (stop words removal) การหารากศัพท์ (stemming) และการให้น้ำหนักของคำด้วยขั้นตอนวิธี Term Frequency–Inverse Document Frequency (TF -IDF) Camacho-Collados และ Pilehvar [18] ทำการศึกษาเทคนิค

ในการเตรียมข้อมูลก่อนการประมวลผลโดยด้วยวิธี tokenizing, lemmatizing, lowercasing, and multiword grouping และยังพบว่าการตัดคำมีประสิทธิภาพมากกว่าการทำ lemmatization

IŞIK และ DAĞ [19] วิเคราะห์เทคนิคเตรียมข้อมูลก่อนการประมวลผลข้อความเพื่อจำแนกขั้วความคิดเห็นบนชุดข้อมูลที่รวบรวมมาจาก Yelp.com/dataset ในปี 2018 จำนวน 2 ชุดข้อมูล ชุดแรกเป็นชุดข้อมูล Review dataset ชุดที่สองเป็นชุดข้อมูล Business Dataset โดยเปรียบเทียบลำดับการใช้เทคนิคซึ่งเทคนิคที่ใช้คือ การกำจัดคำหยุด (Stopwords) การเปลี่ยนตัวอักษรเป็นตัวเล็ก (lowercasing) และการย่อคำ (lemmatization) ซึ่งผู้วิจัยสลับลำดับเทคนิคก่อนหลัง พบว่า ในการใช้เทคนิคในการเตรียมข้อมูลก่อนการประมวลผลลำดับขั้นตอนที่ให้ความถูกต้องของการจำแนกมากที่สุดคือ การกำจัดคำหยุด (Stopwords) การย่อคำ (lemmatization) การเปลี่ยนตัวอักษรเป็นตัวเล็ก (lowercasing) ตามลำดับบนตัวจำแนก K nearest neighbors (KNN), decision tree (DT), random forest (RF), stochastic gradients descent (SGD), naïve Bayes classifier (NB) และ support vector machine (SVM)

2.3.1 การตัดคำ (Word Tokenize)

การตัดคำเป็นกระบวนการแบ่งข้อความออกเป็นคำ ประโยค หรือสัญลักษณ์เรียกว่า Token เป้าหมายคือเพื่อสำรวจคำในประโยค list ของ Token จะกลายเป็นข้อมูลเข้าสำหรับการประมวลผลในการทำเหมืองข้อความ การตัดคำในภาษาอังกฤษนิยมใช้ช่องว่าง [20] กระบวนการตัดคำจะเริ่มต้นด้วยการดูข้อความทั้งหมดเพื่อหาของเขตของคำและขอบเขตของประโยค โดยจะใช้ช่องว่างที่คั่นระหว่างคำสำหรับตัดคำ และใช้จุด (.) เพื่อบ่งบอกถึงการสิ้นสุด[21]ประโยค เทคนิคในการตัดคำแบ่งออกได้เป็น 3 เทคนิคคือ การใช้กฎไวยากรณ์ทางภาษา (Rule-based) การใช้พจนานุกรมในการอ้างอิงคำ (Dictionary based) และการเรียนรู้ของเครื่องจากฐานข้อมูล

1) การตัดคำด้วยกฎไวยากรณ์ทางภาษา
   การตัดคำด้วยกฎไวยากรณ์ทางภาษาจะทำโดยการตรวจสอบกฎทางอักขระวิธีซึ่งกำหนดลักษณะของการประสมอักษร การย่อหน้า การเว้นวรรค เพื่อหาขอบเขตของคำ มีข้อดีคือมีความถูกต้องในระดับพยางค์สูง ใช้ทรัพยากรน้อย มีความเร็วในการประมวลผล แต่ความถูกต้องของผลลัพธ์ค่อนข้างต่ำ
2) การตัดคำด้วยการใช้พจนานุกรมในการอ้างอิงคำ
   การตัดคำด้วยการใช้พจนานุกรมในการอ้างอิงคำทำโดยการนำคำที่ได้มาเทียบกับคำที่มีอยู่ในพจนานุกรม ดังนั้นจึงต้องมีการสร้างพจนานุกรมขึ้นมาเพื่อเก็บคำไว้ก่อน มีข้อดีคือทำให้ได้ความถูกต้องของการตัดคำสูง แต่มีข้อจำกัดคือใช้เวลามากกว่าการตัดคำด้วยกฎไวยากรณ์และต้องใช้เวลาในสร้างพจนานุกรมด้วย
3) การตัดคำด้วยการเรียนรู้ของเครื่องจากฐานข้อมูล

การตัดคำด้วยการเรียนรู้ของเครื่องจำใช้วิธีทางสถิติมาในการประมวลผลภาษา โดยใช้คลังข้อมูลทางภาษาเป็นฐานความรู้ เก็บความถี่ที่ใช้ในการตัดคำ

วิธีการตัดคำที่ได้รับความนิยมในงานด้านเหมืองข้อความ มี 3 วิธี ได้แก่ วิธีการ Unigram วิธีการ Bigram และ Unigram + Bigram แต่ละวิธีการมีรายละเอียด ดังนี้

1) วิธีการ Unigram เป็นวิธีการหนึ่งที่ง่ายและนิยมใช้มากที่สุดในการทำเหมืองข้อความ โดยวิธีนี้ จะทำการตัดคำ 1 คำ เพื่อแทน 1 คุณลักษณะ [22]

2) วิธีการ Bigrams เป็นวิธีการตัดคำโดยการผสม 2 คำ แทนด้วย 1 คุณลักษณะ

3) วิธีการ Trigrams เป็นวิธีการตัดคำโดยการผสม 2 คำ แทนด้วย 1 คุณลักษณะ

Text	**The purpose of this study was to developed a C Algorithm**
Unigram	The/purpose/of/this/study/was/to/developed/a/C/Algorithm
Bigrams	The purpose/ purpose of/ of this/ this study/ study was/ was to/ to developed/ developed a/ a C/ C Algorithm
Trigrams	The purpose of/ purpose of this/ of this study/ this study was/ study was to/ was to developed/.../ a C Algorithm

รูปที่ 1 แสดงการตัดคำแบบต่าง ๆ

การเลือกใช้วิธีการตัดคำนั้นขึ้นอยู่กับชุดข้อมูลด้วย มีการศึกษาการตัดคำแบบต่าง ๆ เพื่อใช้ในการประมวลผลภาษาธรรมชาติร่วมกันเช่น Unigram และ Bigrams ซึ่งพบว่าให้ประสิทธิภาพในการประมวลผลดีขึ้น [23] [24]

2.3.2 การหารากศัพท์ (Stemming)

จุดประสงค์ของการหารากศัพท์เพื่อลดรูปของคำศัพท์ให้เป็นรากศัพท์ด้วยการตัดคำที่อยู่ข้างหน้า (prefixes) และคำที่อยู่ข้างหลัง (suffixes) ด้วยกฎของไวยกรณ์ เหตุผลสองประการในการหารากศัพท์คือ เพื่อประสิทธิภาพในการจัดหมวดหมู่หรือการจำแนกและลดจำนวนคุณลักษณะซึ่งจะเป็นการจำนวนมิติของข้อมูลด้วย [19] ด้วยการทำให้คำที่มีความหมายเหมือนกันเป็นคำเดียว ตัวอย่างเช่น คำว่า “Read”, “Reading” และ “Reader” จะถูกทำให้เป็นคำว่า “Read” โดยขั้นตอนวิธีในการหารากศัพท์ที่เป็นที่นิยมคือ Poster Stemmer [25] และได้พัฒนามาเป็น Snowball Stemmer [26] การทำ Stemming เป็นกระบวนตัดส่วนท้ายของคำแบบหยาบ ๆ ด้วย Heuristic ซึ่งได้ผลดีพอควร สำหรับคำในภาษาอังกฤษส่วนใหญ่ แต่ไม่ทุกคำ Stemming ทำให้ลดฟอร์มลง เหลือแต่ส่วนหน้าของคำที่เหมือน ๆ กันในคำกลุ่มเดียวกัน

2.3.3 การกำจัดคำหยุด (Stop-words removal)

คำหยุดคือคำที่ปรากฏบ่อยมาในข้อความ เช่น a, an, the, for ในภาษาอังกฤษ เนื่องจากคำหยุดมีปรากฏมากมายจนไม่มีผลในการจำแนก จึงไม่จำเป็นต้องนำมาสร้างเป็นดรรชนี ไม่ว่าจะเป็นรูปแบบใดก็ตาม การกำจัดคำหยุดจะช่วยลดมิติของพื้นที่ของคำ (term Space) คำหยุดที่พบมาที่สุดในเอกสารคือ articles prepositions และ pro-nouns ซึ่งคำเหล่านี้ไม่ได้ให้ทำให้เอกสารมีความหมายมากขึ้น คำหยุดจะถูกลบออกจากเอกสารเนื่องจากคำเหล่านี้ไม่ได้ถูกวัดเป็นคำหลักในการทำเหมืองข้อความ

อย่างไรก็ตาม คำหยุดที่ใช้ในการศึกษาโดเมนหนึ่ง อาจไม่ใช่คำหยุดในการศึกษาอีกโดเมนหนึ่ง ดังนั้นในการระบุว่าคำใดเป็นคำหยุดต้องกำหนดให้เหมาะสมกับโดเมนที่ศึกษา

ตารางที่ 1 แสดงตัวอย่างคำหยุดในภาษาอังกฤษ

List of Stop Word					
i	it	be	while	out	most
me	its	been	of	on	other
my	itself	being	at	off	some
myself	they	have	by	over	such
we	them	has	for	under	no
our	their	had	with	again	nor
ours	theirs	having	about	further	not
ourselves	themselves	do	against	then	only
you	what	does	between	once	own
your	which	did	into	here	same
yours	who	doing	through	there	so
yourself	whom	a	during	when	than
yourselves	this	an	before	where	too
he	that	the	after	why	very
him	these	and	above	how	s
his	those	but	below	all	t
himself	am	if	to	any	can
she	is	or	from	both	will

List of Stop Word					
her	are	because	up	each	just
hers	was	as	down	few	don
herself	were	until	in	more	should

2.4 แบบจำลองปริภูมิเวกเตอร์ (Vector space model : VSM)

แบบจำลองปริภูมิเวกเตอร์เป็นหนึ่งในวิธีการแทนเอกสารที่ไม่มีโครงสร้างให้มีโครงสร้างด้วยแบบจำลองเวกเตอร์สเปซ โดยกำหนดให้เอกสารแต่ละฉบับเปรียบเสมือนเวกเตอร์ของคำ ขนาดของเวกเตอร์ขึ้นอยู่กับจำนวนของคำที่ปรากฏอยู่ในเอกสารฉบับนั้นคิดค้นขึ้นโดย Salton [27] ในแต่ละเอกสารคือเวกเตอร์มิติ N โดย N คือจำนวนของคำในเอกสารที่ไม่ซ้ำกัน ดัชนีของเวกเตอร์คือคะแนนของคำในเวกเตอร์นั้น

features	terms	TF(term, document1)	TF(term, document2)
1	aid	1	0
2	all	0	1
3	back	1	0
4	brown	1	0
5	come	0	1
...	...	...	...
...	...	...	...
14	party	0	1
15	quick	1	0
16	their	0	1
17	time	0	1

รูปที่ 2 แสดง Document Term Matrix บน Vector space Model

2.5 การเลือกคุณลักษณะ (Feature Selection)

ขั้นตอนหนึ่งของกระบวนการสร้างแบบจำลองเพื่อการพยากรณ์ด้วยวิธีการเหมืองข้อมูลสำหรับการเลือกข้อมูลย่อยที่มีมิติน้อยลง (Data Reduction) กว่าข้อมูลต้นฉบับ (Original data) ซึ่งฟีเจอร์หลายฟีเจอร์อาจจะซ้ำซ้อนหรือไม่เกี่ยวข้องและไม่ได้มีนัยสำคัญสำหรับการประมวลผล กระบวนการคัดเลือกคุณลักษณะถือเป็นงานสำคัญในการปรับปรุงประสิทธิภาพในการสร้างแบบจำลอง อีกทั้งกระบวนการ Feature Selection ยังเป็นการช่วยในการเพิ่มความถูกต้องในการพยากรณ์ (Improving Prediction accuracy) [28] เนื่องจากจุดประสงค์สำคัญของการทำ Feature

Selection เพื่อลดจำนวนมิติของข้อให้เหลือเพียงชุดข้อมูล (Feature Subset) ที่มีส่งผลต่อความถูกต้องในการพยากรณ์ มากที่สุดและเป็นการเพิ่มความเร็วของอัลกอริทึมด้วย

features	terms	DF(document1 ,term)	DF(document2 ,term)	IDF(term, document1)	IDF(term, document2)
1	aid	1	0	1.41	2.10
2	all	0	1	2.10	1.41
3	back	1	0	1.41	2.10
4	brown	1	0	=LN((1+2)/(1+I6))+1	2.10
5	come	0	1	2.10	1.41
6	dog	1	0	1.41	2.10
7	fox	1	0	1.41	2.10
8	good	0	1	2.10	1.41
9	jumped	1	0	1.41	2.10
10	lazy	1	0	1.41	2.10
11	men	0	1	2.10	1.41
12	now	0	1	2.10	1.41
13	over	1	0	1.41	2.10
14	party	0	1	2.10	1.41
15	quick	1	0	1.41	2.10
16	their	0	1	2.10	1.41
17	time	0	1	2.10	1.41

รูปที่ 3 แสดงตัวอย่างการให้ค่าน้ำหนักของคำ

เมื่อได้คุณลักษณะของเอกสารหรือ Feature ที่สามารถทำมาสร้างรูปแบบได้ ฟีเจอร์ที่นิยมใช้ในการเรียนรู้ด้วยเครื่องคือการปรากฏของคำ (presence-based) หรือความถี่ (frequency) ของ n-grams ที่ได้จากขั้นตอน Pre – Processing การปรากฏของคำจะสร้างข้อมูลแต่ละตัวอย่างเป็นไบนารี่เวกเตอร์ ซึ่ง “1” หมายถึงมีคำนั้นอยู่และ 0 ไม่มีคำนั้นอยู่ สำหรับความถี่ของคำเป็นการหาจำนวนครั้งที่เกิดขึ้นของคำ ในกรณีที่มีความยาวของข้อความแตกต่างกันมาก และคุณลักษณะนั้นเกิดขึ้นเป็นจำนวนมากในทุก ๆ ข้อความแสดงว่าคุณลักษณะดังกล่าวไม่สามารถใช้เป็นตัวแทนของประโยคได้ จึงต้องใช้วิธีให้ค่าน้ำหนักกับคุณลักษณะที่ใช้เป็นตัวแทนของประโยค

เมื่อฟีเจอร์หลักของข้อความเป็น n-grams แล้ว ขนาดของข้อมูลในฟีเจอร์สเปซ (Feature Space) จะใหญ่ขึ้นตามขนาดของชุดข้อมูล (Data Set) ซึ่งการใหญ่ขึ้นของฟีเจอร์สเปซนี้ทำให้การคำนวณฟีเจอร์ทั้งหมดของกลุ่มตัวอย่างเป็นไปได้ยากและส่งผลต่อประสิทธิภาพการจำแนก ซึ่งฟีเจอร์หลายฟีเจอร์อาจจะซ้ำซ้อนหรือไม่เกี่ยวข้องและไม่ได้มีนัยสำคัญสำหรับการประมวลผล ในขั้นตอนการเลือกฟีเจอร์นี้จะเป็นกระบวนการเพื่อกำหนดกลุ่มของฟีเจอร์ที่สามารถส่งผลถึงประสิทธิภาพการทำนายให้ดีที่สุด การกำจัดฟีเจอร์ที่ไม่เกี่ยวข้องและซ้ำซ้อนออกช่วยลดขนาดของพื้นที่ของฟีเจอร์ (Feature Space) และเป็นการเพิ่มความเร็วของอัลกอริทึมด้วย

การให้ค่าน้ำหนักของคำแบบออกเป็น 2 ประเภทคือ Unsupervised term weighting และ Supervised term weighting [29] สำหรับ Unsupervised term weighting ไม่ต้องเรียนรู้การคำนวณจากเอกสารสอน (training documents) จะใช้ค่าน้ำหนักของการเกิดขึ้นของคำที่ปรากฏในเอกสาร [30] ตัวอย่างของการให้ค่าน้ำหนักเช่น TF TF-IDF BM25 ในส่วนของ Supervised term weighting จะใช้ในการเรียนรู้ในการจำแนก น้ำหนักของคำแสดงให้เห็นว่าคำศัพท์เฉพาะในเอกสารเป็นของคลาสใดโดยใช้ข้อมูลที่อยู่ในคลังมาสอน ตัวอย่างของการให้ค่าน้ำหนักเช่น Information Gain Chi

การจัดกลุ่มเอกสารเป็นการเรียนรู้ของเครื่องโดยไม่มีผู้สอนการให้ค่าน้ำหนักของคำจึงเป็นแบบ Unsupervised term weighting

2.5.1 การคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำด้วย Term Frequency - Inverse Document Frequency

TF-IDF เป็นการให้ค่าน้ำหนักของคำในเอกสารเป็นสถิติตัวเลขซึ่งแสดงให้เป็นว่าคำหนึ่ง ๆ มีความสำคัญต่อเอกสารในกลุ่มอย่างไร การให้ค่าน้ำหนักของคำด้วย TF-IDF นิยมใช้ในงานค้นคืนสารสนเทศ (information retrieval) และการทำเหมืองข้อความ (text mining) ค่าของ TF-IDF เพิ่มขึ้นตามสัดส่วนของจำนวนครั้งของคำที่ปรากฏในเอกสาร แต่ไม่ได้พิจารณาเฉพาะความถี่ของคำที่เกิดขึ้นในคลังข้อมูลเท่านั้น เพราะคำบางคำที่เกิดขึ้นบ่อยอาจเป็นคำธรรมดาไม่ได้มีผลต่อการบ่งบอกถึงความสำคัญของเอกสารนั้น [15]

TF-IDF เกิดจากผลคูณของสองค่า คือ TF (Term Frequency) กับ IDF (Inverse Document Frequency) จำนวนครั้งแต่ละคำที่เกิดขึ้นในเอกสารแต่ละฉบับจะถูกนับและรวมทั้งหมดเข้าด้วยกัน โดย Term Frequency (TF) คือ กำหนดเป็นจำนวนครั้งที่คำศัพท์เกิดขึ้นในไฟล์เอกสาร หลังจากนั้นนำค่าที่ได้แต่ละคำไปหารกับจำนวนคำทั้งหมดในเอกสาร

วิธีให้ค่าน้ำหนักกับคุณลักษณะที่ใช้เป็นตัวแทนของประโยค ด้วยวิธี Term Frequency (TF) และ Inverse Document Frequency (IDF) ด้วยสมการ

$$idf_{td} = \log(\frac{N}{D_t}) \quad (2.1)$$

โดยที่ idf_{td} คือ ส่วนกลับของความถี่เอกสารที่มีคำที่ td ปรากฏอยู่

N คือ จำนวนประโยคทั้งหมด

D_t คือ จำนวนประโยคทั้งหมดที่มีคุณลักษณะ t ปรากฏอยู่

การหาค่าความถี่และค่าความถี่ผกผัน จะคำนึงถึงความถี่ของการปรากฏคุณลักษณะในเอกสาร และค่าความถี่ผกผัน คำนวณจากผลคูณของค่าความถี่การเกิดค่าและค่าความถี่ผกผัน ดังสมการ

$$tfidf_{td} = tf_{td} \times idf_{td} \quad (2.2)$$

โดยที่ $tfidf_{td}$ คือ ค่าน้ำหนักของคำที่ d ที่ปรากฏในเอกสารที่ t

tf_{td} คือ ค่าน้ำหนักของคำที่ k ที่ปรากฏในเอกสารที่ t

2.5.2 การคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำด้วย BM25

BM25 เป็นฟังก์ชั่นการจัดอันดับของคำในถุงคำ (bag-of-word) โดยคำนึงถึงอิทธิพลของปัจจัยต่างๆ เช่นความยาวของเอกสารต่อความถี่ในการปรากฏคำซึ่งไม่ได้รับการพิจารณาในการให้น้ำหนักแบบ TF-IDF ใน BM25 เอกสารจะถูกคำนวณตามสัดส่วนล่วงหน้าแล้วคูณด้วยค่าของความถี่ของคำ ดังนั้นจึงสามารถคาดหวังการแยกคำที่มีลักษณะเฉพาะอย่างแท้จริงในเอกสารขนาดยาวได้ [31]

$$score(d,q) = IDF(q) \times \frac{f(q,D) \times (k+1)}{f(q,D) + k \times (1 - b + b \times \frac{|D|}{avgdl})} \quad (2.3)$$

โดย $f(q,D)$ คือความถี่ของคำ q ในเอกสาร D

$|D|$ คือ ความยาวของเอกสาร D

$avgdl$ คือ ความยาวเฉลี่ยของเอกสารทั้งหมด

k, b คือ พารามิเตอร์

และ

$$IDF(q) = \log \frac{N - n(q) + 0.5}{n(q) + 0.5} \quad (2.4)$$

โดย N คือ จำนวนของเอกสาร

$n(q)$ คือ จำนวนของเอกสารที่มี $q(DF)$

2.6 การจัดกลุ่มเอกสาร (Document Clustering)

การจัดกลุ่มข้อความ หมายถึงการจัดกลุ่มของเอกสารที่คล้ายกันไว้ในกลุ่มเดียวกันและที่ต่างกันก็ให้แยกกลุ่มกัน ส่วนที่จะวัดความคล้ายคลึงกันหรือแตกต่างกันนั้น การที่ทำการจัดกลุ่มตามความคล้ายและต่างกันนั้น ก็เพื่อเพิ่ม Recall/Precision รวมทั้งลดขยะในการสืบค้น กล่าวคือแทนที่จะค้นหาจากเอกสารทั้งหมด ก็เพียงจำกัดเฉพาะเอกสารที่อยู่ในกลุ่มเดียวกัน ทำให้ขอบเขตงานแคบลง การค้นหาจะยังทำได้เร็วขึ้น วิธีการ Text Clustering จะทำการหากลุ่มซึ่งเป็นเซทของ Objects ที่ตรงตามเกณฑ์การพิจารณาในที่นี้ Objects อาจจะหมายถึงข้อความ เอกสาร บทความ สิ่งพิมพ์ โดยเซทของ Objects จะหมายถึงเซทของข้อความ ชุดเอกสาร

การ clustering เป็นความต้องการที่จะค้นหาโครงสร้างใหม่ที่ไม่เคยรู้มาก่อน วิธีการ Clustering จึงเป็นแบบ Unsupervised Learning โดยไม่มีขั้นตอนของการ Training และผลลัพธ์ได้จากการค้นหากลุ่มของ Objects ให้ตรงตามเกณฑ์ที่กำหนดโดยทำการวนซ้ำหลาย ๆ รอบจนได้คำตอบ วิธีการแบ่งกลุ่มจัดได้เป็น 3 วิธีหลักคือ Partition-based clustering Hierarchical Clustering

Partition-based clustering เป็นการแบ่งกลุ่มตามแนวราบ โดยถือว่าทุก Object เท่ากันหมด หลักการจะใช้วิธี K-Means ด้วยการวัดความเหมือนของ Object กับตัวแทนกลุ่ม และทำการแบ่งกลุ่มออก แบ่งเป็น K กลุ่ม (Cluster) โดยจะแทนแต่ละ Object ด้วยเวกเตอร์ ใน Vector Space เดียวกัน การแบ่งกลุ่มด้วยวิธีนี้ จะพิจารณาจากค่าเบี่ยงเบนมาตรฐาน การแบ่งกลุ่มที่ดีควรจะต้องมีผลรวมของค่าเบี่ยงเบนมาตรฐาน Cluster ให้น้อยที่สุด

Hierarchical Clustering การจัดกลุ่มแบบลำดับชั้น เป็นเทคนิควิธีการที่จัดกลุ่มตามความคล้ายกันของข้อมูล ด้วยวิธีการวัดความคล้ายหรือความต่างเช่น Euclidean, Cityblock, Mahalanobis, Cosine เป็นต้น [10] รูปแบบการแสดงผลของ Hierarchical Clustering จะถูกแสดงในรูปของต้นไม้ โดยในแต่ละ class node จะประกอบไปด้วย child nodes เทคนิคนี้สามารถแบ่งวิธีการสร้างต้นไม้ได้เป็น 2 ประเภทคือ Agglomerative (Bottom-Up) และ Divisive (Top-Down)

2.6.1 การจัดกลุ่มเอกสารด้วยขั้นตอนวิธี K-Means

K-Means Clustering เป็นหนึ่งในขั้นตอนวิธีคือวิธีการจำแนกกลุ่มข้อมูลด้วย วิธีการแบ่งข้อมูลอัตโนมัติตามค่า k ที่กำหนด โดยกระบวนการทำงานเลือกค่า k เริ่มต้นสำหรับเป็นค่ากลาง ในการจัดกลุ่มและปรับค่าตามกระบวนการดังนี้ 1) เลือก ข้อมูลสำหรับวัดระยะห่างกับค่า K เริ่มต้นทุกค่า 2) กำหนดชุดข้อมูลให้กับ K ที่ใกล้ที่สุด และปรับค่า K ใหม่ให้เป็นค่า

กลางของกลุ่มข้อมูล และหยุดเมื่อค่า K ไม่เปลี่ยนแปลง [32] ดังนั้นการใช้ K-means clustering จึงสามารถนำ มาใช้งานเมื่อทราบจำนวนกลุ่มที่ต้องการจำแนกที่แน่นอน

Input: D = Document –by-term matrix, n : Number of Document and K- Number of Cluster

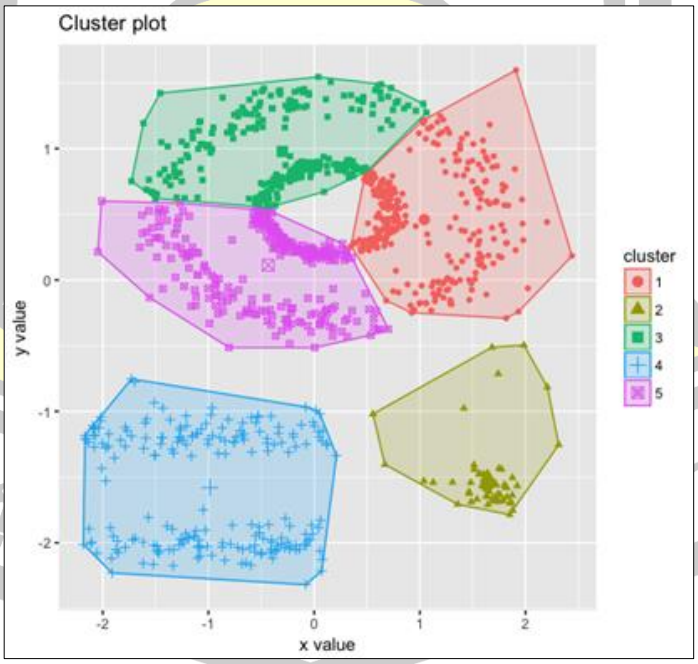
Output : K Cluster of Given Dataset

1. Randomly chose K document from document set as initial centroids
2. Repeat
   a. Assign each of the remaining document to the cluster which has closet centroid by using similarity measure.
   b. After each assignment, Calculate new cluster centroids.

Until the convergence criteria is met.

รูปที่ 4 แสดง K-Means Algorithm

ที่มา [33]



รูปที่ 5 แสดงการการจัดกลุ่มด้วยขั้นตอนวิธี K-Means

ที่มา : http://www.sthda.com/english/wiki/wiki.php?id_contents=7940

2.6.2 ขั้นตอนวิธีการจัดกลุ่มด้วย DBSCAN

DBSCAN เป็นวิธีการจัดกลุ่มแบบนำเสนอโดย Ester และคณะในปี 1996 [34] ขั้นตอนวิธี DBSCAN สามารถจัดกลุ่มด้วยรูปทรงและขนาดที่แตกต่างกันจากข้อมูลมีมี noise และ outlier ได้ DBScan จัดเป็น density-based clustering algorithm ซึ่งก็คือพื้นที่ใกล้เคียงของแต่ละจุดข้อมูลในคลัสเตอร์ซึ่งอยู่ภายในรัศมีที่กำหนด (R) ต้องมีจำนวนจุดขั้นต่ำ (M)

DBSCAN

Core Point: This is a point that has at least m points within distance n from itself.

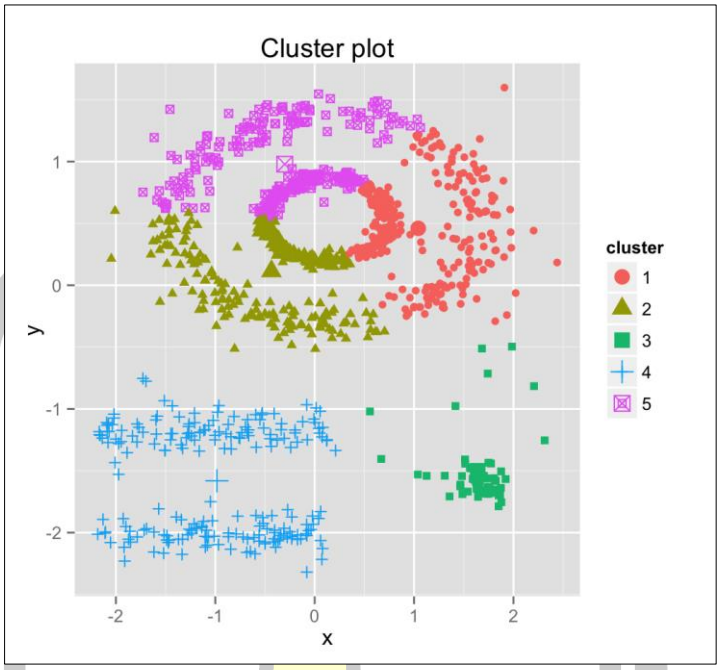
Border Point: This is a point that has at least one Core point at a distance n.

Outlier Point: This is a point that is neither a Core nor a Border. And it has less than m points within distance n from itself.

Algorithm:

1. The algorithm proceeds by arbitrarily picking up a point in the dataset (until all points have been visited).
2. If there are at least ‘minPoint’ points within a radius of ‘**ε**’ to the point then we consider all these points to be part of the same cluster.
3. The clusters are then expanded by recursively repeating the neighborhood calculation for each neighboring point

รูปที่ 6 แสดงขั้นตอนวิธี DBSCAN



รูปที่ 7 แสดงการจัดกลุ่มด้วย DBScan

ที่มา : http://www.sthda.com/english/wiki/wiki.php?id_contents=7940

ขั้นตอนวิธีในการจัดกลุ่มทั้งสองมีข้อดีและข้อเสียแตกต่างกันไปขึ้นอยู่กับลักษณะของข้อมูลที่นำมาใช้ ข้อแตกต่างระหว่าง K-means Clustering และ DBSCAN Clustering ดังตารางที่

ตารางที่ 2 แสดงความแตกต่างระหว่าง K-means Clustering และ DBSCAN Clustering

K-Means Clustering	DBSCAN Clustering
ต้องระบุจำนวนคลัสเตอร์	ไม่จำเป็นต้องระบุจำนวนคลัสเตอร์
มีประสิทธิภาพสำหรับชุดข้อมูลขนาดใหญ่	ไม่สามารถจัดการกับชุดข้อมูลที่มีมิติสูงได้อย่างมีประสิทธิภาพ
ทำงานไม่ได้ดีกับชุดข้อมูลที่มีข้อมูลที่ผิดเพี้ยนและข้อมูลรบกวน	จัดการกับค่าผิดปกติและชุดข้อมูลรบกวนได้อย่างมีประสิทธิภาพ
ในการตรวจจับความผิดของข้อมูลในโดเมนอัลกอริทึมนี้ทำให้เกิดปัญหาเนื่องจากจุดผิดปกติจะถูกกำหนดให้กับคลัสเตอร์เดียวกันกับข้อมูลปกติ	ระบุตำแหน่งพื้นที่ที่มีความหนาแน่นสูงซึ่งแยกออกจากกันโดยพื้นที่ที่มีความหนาแน่นต่ำ
ใช้พารามิเตอร์เดียว: จำนวนคลัสเตอร์ (K)	ใช้พารามิเตอร์ 2 ตัว คือ Radius(R) และ Minimum Points(M) R กำหนดรัศมีที่เลือกซึ่งหากมีจุดภายในเพียงพอแสดงว่าเป็นพื้นที่หนาแน่น M กำหนดจำนวนจุดข้อมูลขั้นต่ำที่

K-Means Clustering	DBSCAN Clustering
	จำเป็นในพื้นที่ใกล้เคียงเพื่อกำหนดเป็นคลัสเตอร์
การเปลี่ยนแปลงความหนาแน่นของจุดข้อมูลจะไม่ส่งผลต่ออัลกอริทึมการจัดกลุ่ม K-mean	การทำคลัสเตอร์ทำงานได้ไม่ดีนักสำหรับชุดข้อมูลแบบกระจัดกระจายหรือสำหรับจุดข้อมูลที่มีความหนาแน่นแตกต่างกัน

2.7 การหาจำนวนกลุ่มที่เหมาะสม

การจัดกลุ่มเอกสารด้วยวิธี K-Means ต้องมีการกำหนดจำนวนกลุ่มที่แน่นอนก่อน ซึ่งจำนวนกลุ่มที่กำหนดลงไปนั้นจะเหมาะสมหรือไม่ ต้องมีการระบุจำนวนกลุ่ม (K-value) ไว้ล่วงหน้าก่อนการจัดกลุ่ม การเลือกจำนวนกลุ่มจะส่งผลต่อผลลัพธ์ของการรวมกลุ่ม ขั้นตอนวิธีในการเลือกจำนวนกลุ่มที่เหมาะสม อาทิเช่น Elbow Method, Gap Statistic, Silhouette Coefficient และ Canopy ซึ่ง Elbow Method เป็นวิธีการหนึ่งที่มีประสิทธิภาพ [35]

แนวคิดของ elbow คือวัดระยะห่างระหว่างจะข้อมูลกับจุดศูนย์กลางของคลัสเตอร์เพื่อดูแนวโน้มของ sum of squared errors (SSE) เพื่อหาจำนวนกลุ่มที่เหมาะสมของการจัดกลุ่มเอกสารโดยวิธี elbow นั้นจะทำการจัดกลุ่มบนชุดข้อมูลตามช่วงของกลุ่มที่กำหนดไว้ สำหรับแต่ละค่าของกลุ่มจะทำการคำนวณค่า SSE แล้วแสดงออกมากเป็นกราฟเส้น ถ้าเส้นมีการหักงอ แสดงว่าจุดตรงที่งอบนเส้นนั้นคือจำนวนกลุ่มที่ดีที่สุด

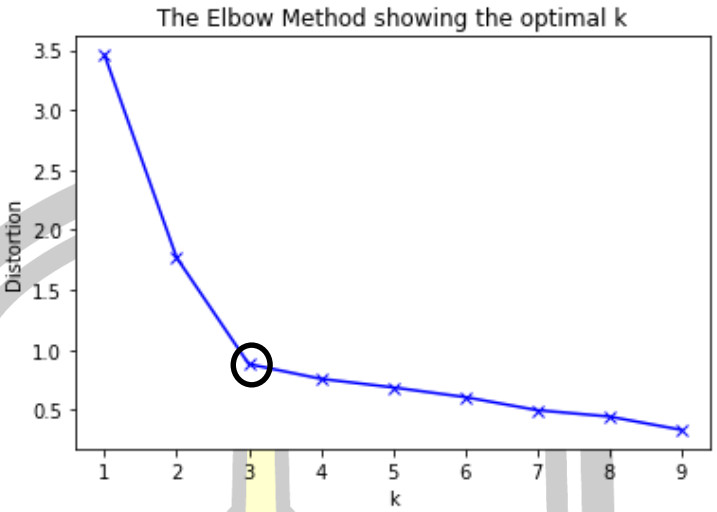
$$W_k = \sum_{r=1}^{k} \frac{1}{n_r} Dr \qquad (2.5)$$

โดย k คือ จำนวนคลัสเตอร์

n_r คือ จำนวนจุดในคลัสเตอร์ r

D_r คือ ผลรวมของระยะทางระหว่างข้อมูลทุกจุดในคลัสเตอร์ ซึ่งสามารถคำนวนได้จากสมการด้านล่าง

$$D_r = \sum_{i=1}^{n_r - 1} \sum_{j=i}^{n_r} \| d_i - d_j \|_2 \qquad (2.6)$$



รูปที่ 8 elbow Method

จากรูปที่ 8 แสดง Elbow Method ซึ่งจะแสดงจำนวนกลุ่มที่เหมาะสมที่สุดด้วยจุดที่หักงอมากที่สุดของเส้นกราฟ

2.8 การหาความคล้ายคลึงของเอกสาร

การเรียนรู้ของเครื่องแบบไม่มีผู้สอนด้วย K – Means ขึ้นอยู่กับการหาระยะทางระหว่างจุดสองจุดเพื่อทำนายผลลัพธ์ ดังนั้นการเลือกวิธีวัดความคล้ายคลึงที่เหมาะสมกับชนิดข้อมูลจึงสำคัญมากเพราะจะส่งผลถึงประสิทธิภาพของการทำนาย การวัดความคล้ายคลึงของเอกสาร (d) จะต้องเป็นไปตามเงื่อนไขสี่ประการ [36] ต่อไปนี้

โดย x และ y เป็นวัตถุที่อยู่ในชุดข้อมูลและ $d(x,y)$ คือระยะห่างระหว่าง x และ y

1. ระยะทางระหว่างสองจุดต้องไม่มีค่าเป็นลบ นั่นคือ $d(x,y) \geq 0$
2. ระยะทางระหว่างวัตถุต้องเป็น 0 ถ้าและต่อเมื่อวัตถุทั้งสองเหมือนกัน นั่นคือ $d(x, y)=0$ $d(x,y)=0$ ถ้าและต่อเมื่อ $x=y$
3. ระยะทางต้องกัน ระยะทางจาก x ไป y ต้องเท่ากันระยะทาง y ไป x นั่นคือ $d(x,y)=d(y,x)$
4. การวัดต้องเป็นไปตามอสมการสามเหลี่ยม ซึ่ง $d(x,y) \leq d(x,y)+d(z,y)$

2.8.1 Euclidean Distance

เป็นการวัดระยะทางระหว่างจุดสองจุดเป็นเส้นตรงบนระนาบเดียวกัน มีสมการคือ

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+...+(x_{in}-x_{jn})^2} \qquad (2.7)$$

โดย $i = (x_{i1}, x_{i2}, ..., x_{in})$ และ $i = (x_{j1}, x_{j2}, ..., x_{jn})$ เป็นข้อมูลหลายมิติ

2.8.2 Cosine Distance & Cosine Similarity

ความคล้ายคลึงกันของโคไซน์เป็นตัวชี้วัดความคล้ายคลึงกันระหว่างเวกเตอร์สองตัวที่วัดมุมระหว่างกัน ผลลัพธ์ที่ออกมามีค่าอยู่ระหว่าง -1 ถึง 1 โดยที่ -1 แตกต่างกันอย่างสิ้นเชิงและ 1 มีค่าใกล้เคียงกันอย่างสมบูรณ์

$$similarity(i, j) = \frac{i \cdot j}{\| i \| \times \| j \|} \quad (2.8)$$

2.8.3 Manhattan

เป็นการเป็นระยะระหว่างจุด 2 จุด โดยรวมระยะทางตามแนวแกนแนวตั้งและแนวนอน มีสมการคือ

$$d(i, j) = |(x_{i1} - x_{j1})| + |(x_{i2} - x_{j2})| + ... + |(x_{in} - x_{jn})| \quad (2.9)$$

โดย $i = (x_{i1}, x_{i2}, ..., x_{in})$ และ $i = (x_{j1}, x_{j2}, ..., x_{jn})$ เป็นข้อมูลหลายมิติ

2.8.4 Minkowski

มีสมการคือ

$$d(i, j) = (|(x_{i1} - x_{j1})|p + |(x_{i2} - x_{j2})|p + ... + |(x_{in} - x_{jn})|p)1/p \quad (2.10)$$

โดย $i = (x_{i1}, x_{i2}, ..., x_{in})$ และ $i = (x_{j1}, x_{j2}, ..., x_{jn})$ เป็นข้อมูลหลายมิติ

2.9 การวัดประสิทธิภาพการจัดกลุ่ม

วิธีการวัดประสิทธิภาพของการจัดกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยา (Precision) ค่าการจำได้ (Recall) และค่าเฉลี่ยประสิทธิภาพโดยรวม (F-Measure) อธิบายโดยใช้ตาราง Confusion เป็นตารางที่มีจำนวนแถวเท่ากับจำนวนคอลัมน์ และเท่ากับจำนวนคลาส เช่น มีคลาสคำตอบของข้อมูลชุดสอน 2 คลาส คือ ความรู้สึกเชิงบวก (Positive) และ ความรู้สึกเชิงลบ (Negative) จะได้ตารางขนาด 2x2 โดยที่ข้อมูลด้านคอลัมน์ เป็นคลาสที่อยู่ในข้อมูลชุดสอนซึ่งกำหนดโดยคนและข้อมูลด้านแถว เป็นคลาสที่ทำนายได้ ดังตารางที่ 1

ตารางที่ 3 Confusion Table

	Machine (Yes)	Machine (No)
Human (Yes)	TP	FN
Human (No)	TP	TN

ที่มา : [37]

โดยที่ TP (True Positive) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Positive

FN (False Negative) คือ จำนวนข้อมูลที่ทำนายว่าเป็นคลาส Negative แต่คำตอบคือ Positive

FP (False Positive) คือ จำนวนข้อมูลที่ทำนายว่าเป็นคลาส Positive แต่คำตอบคือ Negative

TN (True Negative) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Negative

2.9.1 การวัดค่าความถูกต้อง (Accuracy)

การวัดค่าความถูกต้องของการจำแนกดังสมการ

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.11)$$

2.9.2 การวัดค่าความแม่นยำ (Precision)

Precision เป็นการวัดความสามารถของระบบการสืบค้นในการดึงเอกสารที่เป็นคำสอบได้สอดคล้องตรงประเด็นกับคำสอบถามให้ได้มากที่สุด หากค่า Precision เป็น 1 หมายถึง คำตอบของเอกสารทุกรายการมีความเกี่ยวพันธ์กัน ถ้า Precision เป็น 0 หมายถึง ไม่มีคำตอบของเอกสารใดที่มีความเกี่ยวพันธ์กัน

การวัดค่าความแม่นยำของการจัดกลุ่มดังสมการ

$$Precision = \frac{TP}{(TP + FP)} \quad (2.12)$$

2.9.3 การวัดค่าการจำได้ (Recall)

การวัดค่าการจำได้เป็นการวัดความถูกต้องดังสมการ

$$Recall = \frac{TP}{(TP + FN)} \quad (2.13)$$

2.9.4 ค่าเฉลี่ยประสิทธิภาพโดยรวม (F-measure or F1 Score)

ค่าเฉลี่ยประสิทธิภาพโดยรวมซึ่งจะพิจารณาแยกทีละคลาส เป็นการนำค่าความระลึกและค่าความแม่นยำมาพิจารณาร่วมกัน ระบบที่มีประสิทธิภาพดีจะต้องมีค่าความระลึกและค่าความแม่นยำสูงใกล้เคียงกัน ดังสมการ

$$F-Measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (2.14)$$

2.10 งานวิจัยที่เกี่ยวข้อง

2.10.1 การจัดกลุ่มเอกสาร

Balabantaray และคณะ [38] ได้เปรียบเทียบการจัดกลุ่มเอกสารระหว่างขั้นตอนวิธี K-means และ k Medoids พบว่า K-means มีประสิทธิภาพที่ดีกว่า ADEBIYI และคณะ [39] พัฒนาต้นแบบระบบตรวจสอบงานวิจัยที่มีเนื้อหา (Area) ใกล้เคียงกันของวารสารวิจัยไนจีเรียที่สร้างขึ้นมาด้วยวิธี Semantics-based ใช้ Latent Semantic Indexing และ TF-IDF ในการสร้าง Feature Vector ใช้ขั้นตอนวิธี K-Means Clustering จัดกลุ่ม เพื่อตรวจสอบจำนวนคลัสเตอร์ที่เหมาะสมก่อนการจัดกลุ่มผู้วิจัยใช้ Elbow curve เพื่อกำหนดค่า k ประเมินผลการจัดกลุ่มด้วยเทคนิค Silhouette analysis พบว่าความคล้ายคลึงกันภายในคลัสเตอร์สูง 80% โดยเฉลี่ยในทุกจุดข้อมูล

ANAZI และคณะ [40] ได้ทดลองจัดกลุ่มเอกสารโครงการที่ใช้สำเร็จการศึกษา (graduation project documents) ด้วย k-means k-means fast และ k-medoids ร่วมกับการวัดความคล้ายคลึงของเอกสารด้วย cosine similarity Jaccard similarity และ Correlation Coefficient พบว่า การจัดกลุ่มด้วย k-means และ k-medoids ร่วมกับ cosine similarity ให้ประสิทธิภาพดีที่สุดและยังพบว่ายิ่งแบ่งจำนวนกลุ่มมากขึ้นทำให้การจัดกลุ่มของเอกสารดีขึ้นด้วย

Fry และ Manna [41] ทดลองจัดกลุ่มเอกสารด้วยขั้นตอนวิธี K-means และ Peak-searching บนชุดข้อมูลการรีวิวสินค้าตามหัวข้อ ผลการทดลองพบว่าการจัดกลุ่มด้วย K-mean ทำงานได้ดีกว่า Peak-searching

CHAKRABORTY และ คณะ [42] ได้ทดสอบประสิทธิภาพของจากจัดกลุ่มด้วยการเปลี่ยนแปลงข้อมูล (incremental) ในฐานข้อมูล โดยเปรียบเทียบ K-means กับ DBSCAN พบว่า K-means มีประสิทธิภาพดีกว่าเนื่องจากใช้เวลาน้อยกว่า DBSCAN เพราะว่า DBSCAN ต้องใช้เวลาต้องใช้เวลาในการจัดการและจัดกลุ่มข้อมูลที่มีเป็น Noise

Chunhui และ Haitao ได้เปรียบเทียบการหาจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มด้วยขั้นตอนวิธี K-Means ด้วยเทคนิค Elbow Method Gap Statistic Silhouette Coefficient และ Canopy บนชุดข้อมูล Iris Elbow Method พบว่า Elbow Method ใช้เวลาในการตรวจสอบจำนวนกลุ่มที่เหมาะสมน้อยที่สุดด้วยผลลัพธ์ของจำนวนกลุ่มที่เท่ากัน [35]

2.10.2 การตรวจสอบค่าผิดปกติ (Outlier Detection)

ผู้วิจัยได้นำแนวคิดการตรวจสอบค่าผิดปกติมาใช้ในขั้นตอนวิธีที่ผู้วิจัยนำเสนอ เพื่อใช้ในการจำแนกเอกสารว่าควรอยู่ในกลุ่มหรืออยู่นอกกลุ่ม โดยการวิเคราะห์คลัสเตอร์นั้นคล้ายคลึงกับการตรวจจับค่าผิดปกติ (Outlier detection) เนื่องจากทั้งสองเทคนิคเกี่ยวข้องกับการกำจัดวัตถุที่เกี่ยวข้องเล็กน้อยหรือไม่เกี่ยวข้อง การมีค่าผิดปกติสามารถทำให้การจัดกลุ่มผิดเพี้ยนได้ อัลกอริธึมการจัดกลุ่มโดยทั่วไปจะจัดการกับการระบุว่าข้อมูลใดเป็นค่าผิดปกติที่อาจเกิดขึ้นในระหว่างกระบวนการจัดกลุ่มและจะตัดค่านั้นออกซึ่งค่าผิดปกติเหล่านี้จะหมดไปในขั้นตอนสุดท้ายของการทำคลัสเตอร์ การกำจัดจับค่าผิดปกติมี 2 วิธีคือ Distance-Based Outlier Removal และ Cluster-Based Outlier Removal [43] โดยวิธีแรก Distance-Based Outlier Removal นั้นเป็นเทคนิคการตรวจจับค่าผิดปกติโดยจะกำหนดให้แต่ละข้อมูลมีคะแนนค่าผิดปกติ (Outlier Score) ที่กำหนดระดับข้อมูลนั้นว่าเป็นค่าผิดปกติหรือไม่ เทคนิคดังกล่าวสามารถลบเปอร์เซ็นต์ของค่าผิดปกติที่ระบุได้ เช่นข้อมูลจะถูกจัดเรียงตามสำหรับคะแนนค่าผิดปกติและข้อมูลที่อยู่ต่ำกว่าค่าคะแนนที่กำหนดจะถูกตัดออก โดยขั้นตอนวิธีนี้ใช้ Eculidean Distance หาความคล้ายคลึงระหว่างข้อมูล ถ้าข้อมูลในชุดมีระยะทางมากกว่าค่าที่กำหนดหรือ threshold จะถือว่าเป็นข้อมูลที่มีค่าผิดปรกติ (Outlier) ในส่วน Cluster-Based Outlier Removal จะตรวจหาข้อมูลผิดปกติจะเป็นผลพลอยได้ของการจัดกลุ่มซึ่งวัตถุที่อยู่ห่างไกลจากจุดศูนย์กลางของคลัสเตอร์จะถือว่าเป็นค่าผิดปกติหลังจากที่กำหนดวัตถุเข้าไปในกลุ่มแล้วจัดเรียงวัตถุด้วยระยะทางแล้ว [44]

มีหลายวิธีที่ใช้ในการจัดการกับข้อมูลที่มีค่าผิดปกติยกตัวอย่างเช่น Barai และ dey [4] นำเสนอวิธีการหาค่า Threshold เพื่อค้นหา Outlier และตัดออกก่อนที่จะมีการจัดกลุ่มเอกสาร ด้วยสมการ Threshold Value = (Maximum distance + Minimum Distance) / 2 หลังจากนั้นคำนวณระยะทางของข้อมูลด้วย Euclidean ถ้าระยะทางมากกว่าค่า Threshold จะตัดสินใจว่าข้อมูลนั้นจะอยู่นอกกลุ่ม (Outlier) แต่ถ้าไม่ใช่ข้อมูลนั้นจะถูกจัดอยู่ในกลุ่ม พบว่าประสิทธิภาพของการจัดกลุ่มดีขึ้นเล็กน้อย

Bahman และ Sattar [45] นำเสนอขั้นตอนวิธีการจัดการกลุ่มเอกสารและจัดการกับข้อมูลที่มีค่าผิดปกติของการจัดกลุ่มด้วย K-means บนชุดข้อมูล UCI ประกอบด้วย iris Bupa และ Glass เรียกว่า ขั้นตอนวิธี ODBD (ODBD algorithm) ด้วยการหาข้อมูลที่มีค่าผิดปกติและแบ่งชุดข้อมูล

ออกเป็นสองกลุ่มคือข้อมูลปกติและข้อมูลที่มีค่าผิดปกติ หลังจากนั้นคำนวณหาจุดศูนย์กลางของกลุ่มบนชุดข้อมูลปกติด้วยขั้นตอนวิธี FICBC แล้วคำนวณระยะทางระหว่างข้อมูลที่มีค่าผิดปกติแต่ละตัวกับจุดศูนย์กลางของกลุ่มด้วย Euclidean ข้อมูลที่มีค่าผิดปกติอยู่ใกล้กลุ่มใดมากที่ให้ข้อมูลไปอยู่ในกลุ่มนั้น พบว่าขั้นตอนวิธีที่นำเสนอมีความถูกต้องของการจัดกลุ่มมากกว่าวิธี K-Means ปกติ

Yu และคณะ [46] นำเสนอ OEDP k-Means Algorithm ด้วยการกำจัด Outlier จากชุดข้อมูลก่อนการจัดกลุ่ม การทำงานของขั้นตอนวิธีนี้ในขั้นแรกทำการตรวจหา Outlier และกำจัดออก ขั้นต่อแบ่งข้อมูลเข้าเป็นเซ็ทย่อยตามความหนาแน่นของข้อมูล ในการทดลองผู้วิจัยใช้ชุดข้อมูล Ecoli Iris Wine และ Climate พบว่า OEDP k-Means สามารถลดผลกระทบของค่า Outlier ได้เมื่อเลือกจุดศูนย์กลางเริ่มต้นของการจัดกลุ่มนอกจากนี้ยังทำให้ผลลัพธ์ของการจัดกลุ่มดีขึ้น

2.10.3 การให้ค่าน้ำหนักของคำ

การให้ค่าน้ำหนักของคำด้วย TF-IDF เป็นวิธีที่ใช้กันอย่างแพร่หลายในงาน information retrieval และ text mining ซึ่งจะใช้เพื่อแปลงเอกสารเป็นรูปแบบที่มีโครงสร้าง เป็นตัวเลขที่แสดงถึงความสำคัญของคำต่อเอกสาร Term Frequency (TF) คือการนับความถี่ของคำ (t) ในเอกสาร (d) จะเพิ่มขึ้นตามสัดส่วนกับจำนวนครั้งที่คำปรากฏในเอกสาร แต่ถูกชดเชยด้วยความถี่ของคำในคลังข้อมูลซึ่งช่วยในการควบคุมข้อเท็จจริงที่ว่าโดยทั่วไปแล้วคำบางที่มีมากกว่าคำอื่น ๆ แต่อาจจะไม่มีความสำคัญในการจำแนกเอกสาร [47]

Afrizal และคณะ [48] ได้ประยุกต์วิธีการให้ค่าน้ำหนักของคำด้วย TF-IDF mDFIDF และ BM25 เพื่อพัฒนาการกรองข้อมูลอัตโนมัติบนข้อมูลความคิดเห็นเกี่ยวกับผลิตภัณฑ์ท่องเที่ยว (Tourism Product reviews) ผลการทดลองพบว่าวิธีการให้ค่าน้ำหนักของคำด้วยวิธี TF-IDF ดีกว่าวิธีการอื่น ซึ่งสอดคล้องกับ Kadhim [49] ได้ทดลองเปรียบเทียบการวิธีการให้ค่าน้ำหนักของคำเพื่อการสกัดฟีเจอร์ (Feature extraction) บนข้อมูลทวิตเตอร์จำนวน 2,196 ทวีต ด้วยวิธี TF-IDF และ BM25 พบว่ามีประสิทธิภาพ (F1 - Measure) ดีกว่า BM25

Shu และคณะ [31] เปรียบเทียบวิธีการให้น้ำหนักคุณลักษณะของเอกสารระหว่าง TF-IDF และ BM25 บนชุดข้อมูลทวิตเตอร์ด้วยการรวบรวมจำนวน 18,000 ทวีตจากบริษัทจำนวน 170 บริษัท สกัดคุณลักษณะด้วย BM25 BM25+ TF และ TF-IDF ทดสอบการจัดกลุ่มเพื่อจัดกลุ่มข้อความด้วยชื่อบริษัทด้วย Fuzzy c means ตรวจสอบผลลัพธ์ของการจัดกลุ่มด้วยค่า entropy ในการกำหนดพารามิเตอร์ของ BM25 ผู้วิจัยกำหนดค่า k=1.2 และ b=0.75 พบว่าจำนวนกลุ่มถูกเปลี่ยนจาก 2 เป็น 14 กลุ่ม พบว่าผลลัพธ์ของการจัดกลุ่มด้วย BM25 ดีกว่า TF-IDF ผู้วิจัยให้ข้อคิดเห็นว่าที่เป็นเช่นนี้เพราะมีการนำความยาวของเอกสารเข้าไปพิจารณาด้วย และยังให้ข้อคิดเห็นว่าถึงแม้การให้น้ำหนักกับคุณลักษณะด้วย BM25 จะมีผลลัพธ์ที่ดีกว่าแต่ก็ไม่จริงเสมอไปกับชุดข้อมูล

อื่น ๆ เนื่องจากคุณลักษณะที่สร้างขึ้นได้รับอิทธิพลจากความแตกต่างของความยาวเอกสารซึ่งจะทำให้ความถูกต้องในการวิเคราะห์เอกสารได้รับผลกระทบในทางลบ

2.10.4 การวัดความคล้ายคลึงของเอกสาร

มีการทดลองวัดความคล้ายคลึงของเอกสารในสำหรับการจัดกลุ่มเอกสารหลายหลากวิธี ซึ่งแต่ละวิธีก็ให้ผลลัพธ์ที่แตกต่างกันโดยขึ้นอยู่กับขั้นตอนวิธีที่ใช้วัดและชุดข้อมูลที่ทดลอง เช่น

Usino และคณะ [14] นำเสนอระบบที่สามารถตรวจจับข้อมูลการลอกเลียนแบบโดยใช้ K-mean และอัลกอริธึม cosine distance ขั้นตอนก่อนการประมวลผลมีการตรวจสอบพจนานุกรมขนาดใหญ่ของอินโดนีเซียการ ออกแบบแบบจำลองพื้นที่เวกเตอร์ (vector space model) จัดกลุ่มของเอกสารด้วยขั้นตอนวิธี K - Means และวัดความคล้ายคลึงเอกสารด้วย cosine distance จากเอกสาร 17 ฉบับเป็นข้อมูลทดสอบ ผลการศึกษาความแม่นยำในการตรวจจับ 93.33%

Arzoo และ Rathod [50] ได้ทดลองจัดกลุ่มด้วยขั้นตอนวิธี K-means โดยการวัดระยะทางของข้อมูลในการจัดกลุ่มด้วยวิธีที่แตกต่างกันด้วย Euclidean Manhattan Chebychev Minkowski และได้นำเสนอขั้นตอนวิธีวัดระยะทางด้วย Euclidean ร่วมกับ Chebychev บนข้อมูลทางภูมิศาสตร์ที่มี 2 ค่า (0 และ 1) พบว่าการวัดระยะทางด้วย Euclidean ร่วมกับ Chebychev ใช้ระยะในการจัดกลุ่มเวลาน้อยกว่า Euclidean และ Minkowski ในขณะที่ Chebychev และ Manhattan ใช้ระยะในการจัดกลุ่มเวลาน้อยที่สุด โดยความถูกต้องของการจัดกลุ่มเท่ากัน

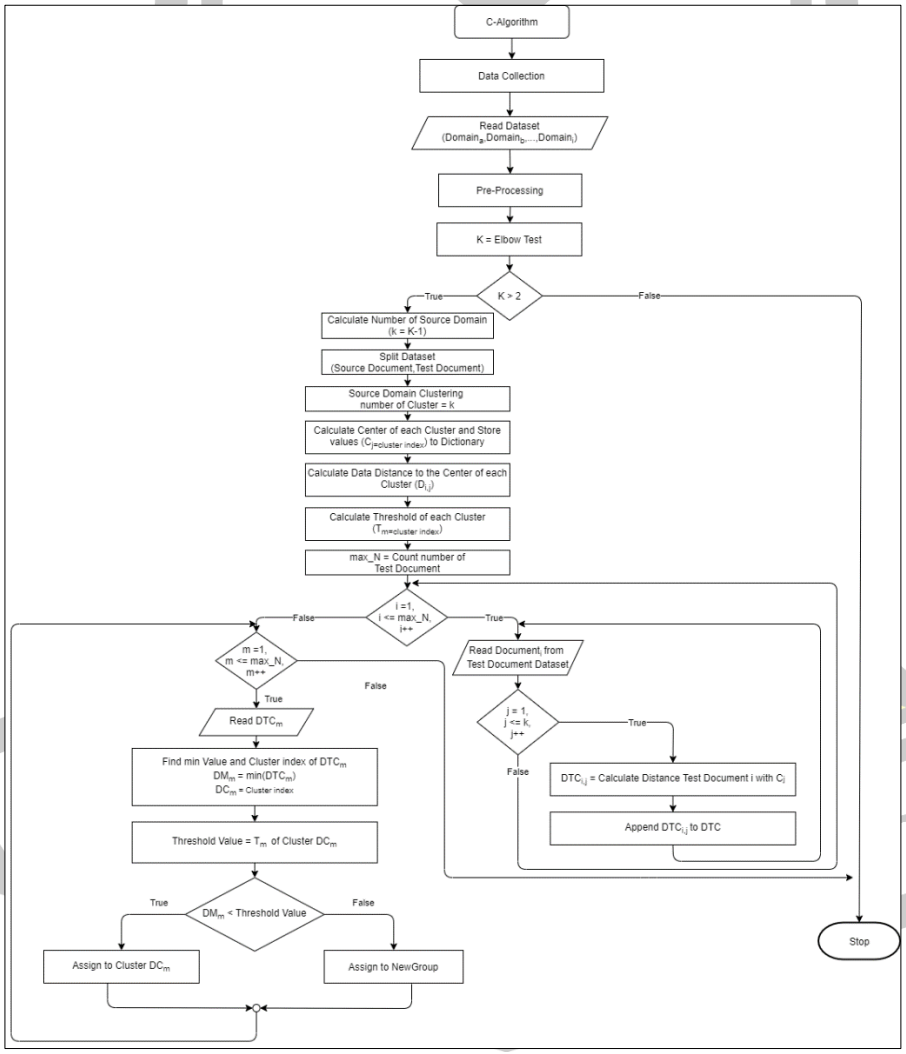
Salihu และคณะ [51] นำเสนอโมเดลสรุปบทความอัตโนมัติด้วยขั้นตอนวิธี K-Means โดยใช้วิธีวัดระยะทางที่แตกต่างกันด้วย Euclidean และ Manhattan ทั้งสองวิธีใช้ TF-IDF พบว่า Euclidean สามารถสรุปประโยคได้ 72% จากกลุ่มที่แตกต่างกันสามกลุ่มในขณะที่ Manhattan สามารถสรุปประโยคได้ 94% ของเอกสารทั้งหมดในกลุ่มเดียว

RAVINDRAN และ Malathi [52] นำเสนอวิธีการในการจัดกลุ่มด้วยขั้นตอนวิธี K-means บนชุดข้อมูลหลายมิติโดยใช้ Vector Space Model ในการนำเสนอเอกสารที่มีมิติสูง คำนวณความคล้ายคลึงด้วย Cosine ซึ่งจะได้ค่า 0 และ 1 ในกรณีค่าใกล้ 1 แสดงว่าข้อมูลนั้นมีความคล้ายคลึงกันมาก ซึ่งวิธีการนี้ให้ผลลัพธ์ดีกว่าการวัดความคล้ายคลึงด้วยระยะทางแต่ผู้วิจัยไม่มีการเปรียบเทียบการวัดความคล้ายคลึงของ Cosine กับการวัดความคล้ายคลึงด้วยวิธีการอื่น

จากแนวคิด ทฤษฎีที่เกี่ยวข้อง การทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องกับงานวิจัยนี้ ผู้วิจัยได้นำมาประยุกต์ใช้และออกแบบวิธีการดำเนินการวิจัย รวมถึงการเปรียบเทียบประสิทธิภาพของวิธีการต่าง ๆ เพื่อที่จะคัดเลือกวิธีการที่ให้ประสิทธิภาพดีที่สุด และนำไปใช้ในขั้นตอนวิธีที่ผู้วิจัยนำเสนอ ซึ่งจะนำเสนอในบทต่อไป

บทที่ 3
วิธีการดำเนินการวิจัย

การดำเนินการวิจัยให้บรรลุวัตถุประสงค์ในการจัดกลุ่มเอกสารและพิจารณาข้อมูลชุดใหม่ว่าควรอยู่ในกลุ่มหรือแยกออกจากกลุ่มนั้น ผู้วิจัยจะได้นำเสนอในแต่ละขั้นตอนดังหัวข้อต่อไปนี้ (1) ชุดข้อมูล (2) กระบวนการเตรียมการก่อนประมวลผล (3) การคัดเลือกคุณลักษณะของเอกสาร (4) การจัดกลุ่มเอกสาร (5) ขั้นตอนวิธีในการพิจารณาเอกสารว่าควรอยู่ในกลุ่มหรือแยกออกจากกลุ่ม (6) การวัดประสิทธิภาพ ดังรูปที่ 9



รูปที่ 9 กรอบแนวคิดในการดำเนินการวิจัย

จากรูปที่ 9 แสดงขั้นตอนการดำเนินการวิจัย เริ่มจากกระบวนการรวบรวมข้อมูลจากแหล่งข้อมูลมาตรฐาน แล้วเตรียมข้อมูลด้วยเพื่อแปลงข้อมูลที่อยู่ในรูปแบบไม่มีโครงสร้างให้อยู่ใน

รูปแบบมีโครงสร้าง ขั้นตอนต่อมาเป็นการหากลุ่มจำนวนที่เหมาะสมของการจัดกลุ่ม การแยกเอกสารในกลุ่มสุดท้ายออกเพื่อทดสอบ แล้วทำการจัดกลุ่มของเอกสารหลังจากนั้นเป็นขั้นตอนหาความคล้ายคลึงของโดเมนเมื่อพบว่าเอกสารที่ส่งเข้าไปทดสอบคล้ายคลึงกับกลุ่มของโดเมนใดก็จะถูกกำหนดให้อยู่ในกลุ่มนั้นซึ่งจะอธิบายขั้นตอนวิธีที่นำเสนอตามลำดับ

3.1 ชุดข้อมูล

ในงานวิจัยนี้ผู้วิจัยเลือกใช้ชุดข้อมูล 2 ชุดได้แก่

1. ชุดข้อมูล Multi-Domain Sentiment dataset ซึ่งรวบรวมโดย Blitzer และคณะ [53] จาก Amazon.com ประกอบไปด้วยสินค้าที่แตกต่างกัน 3 ประเภทคือ Book, DVDs, electronics โดยเลือกตัวอย่างรีวิวจากโดเมนทั้ง 3 อย่างละ 600 รีวิวรวมเป็น 1,800 ตัวอย่าง
2. ชุดข้อมูล 20 newsgroups text dataset ซึ่งผู้วิจัยคัดเลือกมา 3 โดเมนโดยวิธีการคัดเลือกแบบสุ่มได้แก่โดเมน alt.atheism มีจำนวน 480 เอกสาร misc.forsale มีจำนวน 585 เอกสาร sci.electronics มีจำนวน 591 เอกสาร รวมทั้งสิ้น 1,656 เอกสารจากทั้ง 3 โดเมน

ตารางที่ 4 แสดงจำนวนชุดข้อมูลที่ใช้ในงานวิจัย

Dataset	Number of Domain	Domain	Number of Documents	Min term in document	Max term in document	Avg term in document	Total
Multi-Domain Sentiment dataset	3	Book	600	5	5,176	177	1,800
		DVDs	600	10	1,374	186	
		Electronics	600	8	1,345	97	
20 newsgroups text dataset	3	alt.atheism	480	1	8,611	199	1,656
		misc.forsale	585	2	2,625	106	
		sci.electronics	591	1	11,765	134	

เหตุผลที่ผู้วิจัยเลือกใช้ชุดข้อมูล Multi-Domain Sentiment dataset เนื่องจากมีหลายงานวิจัยที่เลือกใช้ ในการจัดกลุ่มเอกสารหรือการจำแนกความรู้สึกของข้อความ นอกจากนี้จากงานวิจัยของ [54] [55] [56] พบว่าเมื่อเอกสารที่อยู่โดเมนที่แตกต่างกันแต่มีความคล้ายคลึงกัน

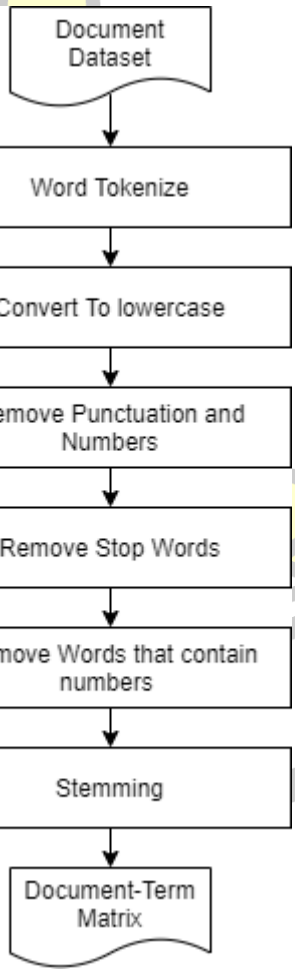
สามารถนำเอกสารในโดเมนนั้นมาเป็นโดเมนต้นทาง (Source Domain) และนำไปจำแนกเอกสารอีกโดเมนได้ (Target Domain) สามารถเข้าถึงได้จาก http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

ในส่วนชุดข้อมูล 20 newsgroups text dataset ทั้งสองเป็นชุดข้อมูลที่ถูกใช้ในงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มเอกสารอย่างแพร่หลาย [57] [58] [59]

3.2 ขั้นตอนวิธีในการดำเนินงานวิจัย

3.2.1 ขั้นตอนเตรียมข้อมูลก่อนการประมวลผล (Pre Processing)

ชุดข้อมูล Multi-Domain Sentiment dataset และชุดข้อมูล 20 newsgroup เป็นชุดข้อมูลที่นำมาใช้งานอยู่ในรูปแบบเอกสาร XML จึงได้ดำเนินการจัดรูปแบบของเอกสารของทั้ง 3 โดเมน โดยการตัด TAG ที่ไม่เกี่ยวข้องออก ให้เหลือเฉพาะ TAG <review_text> หลังจากนั้นผู้วิจัยได้เข้าสู่กระบวนการเตรียมข้อมูลก่อนการประมวลผลดังขั้นตอนดังนี้



รูปที่ 10 แสดงขั้นตอนการเตรียมก่อนการประมวลผล

1. การตัดคำ (Word Tokenize)

การตัดคำเป็นกระบวนการแบ่งข้อความออกเป็นคำ ประโยค หรือสัญลักษณ์เรียกว่า Token เป้าหมายคือเพื่อสำรวจคำในประโยค list ของ Token จะกลายเป็นข้อมูลเข้าสำหรับการประมวลผลในการทำเหมืองข้อความ การตัดคำในภาษาอังกฤษนิยมใช้ช่องว่าง [46] กระบวนการตัดคำจะเริ่มต้นด้วยการดูข้อความทั้งหมดเพื่อหาของเขตของคำและขอบเขตของประโยค โดยจะใช้ช่องว่างที่คั่นระหว่างคำสำหรับตัดคำ และใช้จุด (.) เพื่อบ่งบอกถึงการสิ้นสุดประโยค เทคนิคในการตัดคำแบ่งออกได้เป็น 3 เทคนิคคือ การใช้กฎไวยากรณ์ทางภาษา (Rule-based) การใช้พจนานุกรมในการอ้างอิงคำ (Dictionary based) และการเรียนรู้ของเครื่องจากฐานข้อมูล

```
This book has its good points ^^.If anything, it helps you p
ut into words what you want from a supervisor, but it is not
very accurate.The online test doesn't account for a differenc
e between when 2 of their options are both exactly like you,
or if they don't describe you at all 555.
```

รูปที่ 11 ตัวอย่างเอกสารต้นฉบับ

```
'This', 'book', 'has', 'its', 'good', 'points', '^^', '.', '
If', 'anything', ',', 'it', 'helps', 'you', 'put', 'into', '
words', 'what', 'you', 'want', 'from', 'a', 'supervisor', ',
', 'but', 'it', 'is', 'not', 'very', 'accurate', '.', 'The',
'online', 'test', 'doesn', "'", 't', 'account', 'for', 'a',
'difference', 'between', 'when', '2', 'of', 'their', 'option
s', 'are', 'both', 'exactly', 'like', 'you', ',', 'or', 'if'
, 'they', 'do', "n't", 'describe', 'you', 'at', 'all', '555',
'.'
```

รูปที่ 12 ตัวอย่างเอกสารหลังจากตัดคำ

2. เปลี่ยนคำเป็นอักษรตัวเล็ก

```
'this', 'book', 'has', 'its', 'good', 'points', '^^', '.', '
if', 'anything', ',', 'it', 'helps', 'you', 'put', 'into', '
words', 'what', 'you', 'want', 'from', 'a', 'supervisor', ',
', 'but', 'it', 'is', 'not', 'very', 'accurate', '.', 'the',
'online', 'test', 'doesn', "'", 't', 'account', 'for', 'a',
'difference', 'between', 'when', '2', 'of', 'their', 'option
s', 'are', 'both', 'exactly', 'like', 'you', ',', 'or', 'if'
, 'they', 'do', "n't", 'describe', 'you', 'at', 'all', '555',
'.'
```

รูปที่ 13 ตัวอย่างเอกสารหลังจากเปลี่ยนตัวอักษร

3. เลือกคำที่เป็นตัวอักษรภาษาอังกฤษเท่านั้น เนื่องจากตัวเลขและอักขระพิเศษไม่มีความจำเป็นต่อการจัดกลุ่มหรือการจำแนกเอกสาร ดังนั้นผู้วิจัยจึงเลือกคำในเอกสารที่เป็นภาษาอังกฤษเท่านั้น

4. กำจัดคำหยุด (Stop Word) คำหยุดหรือคำโหลหมายถึงคำที่เกิดขึ้นบ่อยครั้งในเอกสารซึ่งแทบจะไม่มีความสำคัญต่อการประมวลผลของเอกสาร [60] การลบคำหยุดออกจากข้อความทำให้จำนวนคุณลักษณะของเอกสารลดลงเพิ่มความเร็วในการประมวลผล การกำจัดคำหยุดจะใช้วิธีการสร้างคลังเก็บคำหยุดไว้ เมื่อมีการประมวลผลเอกสารจะนำคำในเอกสารไปเปรียบเทียบกับคำหยุดที่อยู่ในคลัง

5. การตัดเครื่องหมายวรรคตอน (punctuation) เครื่องหมายวรรคตอนไม่มีผลในการประมวลผลเอกสารผู้วิจัยจึงตัดออก เครื่องหมายวรรคตอนเช่น Full stop (.) Comma (,) Colon (:)

6. ตัดตัวเลขออกที่อยู่ในตัวอักษรออก ผู้วิจัยพบว่ามีคำที่ผสมตัวเลขและตัวอักษรอยู่เป็นจำนวนมาก ตัวอย่างเช่น

'10th', '12-in-1', '12-in-1 lexar', '12-in-1 sandisk', '128mb', '128mb jump', '150x', '150x sd', '16mm', '16mm compar', '16mm dupe', '16mm print', '17th', '18th', '18th centuri', '1920s', '1920s except', '1930s', '1960s', '1970s', '1980s', '19th', '19th centuri', '1gb', '1st', '2-disc', '2-year', '20th', '20th anniversari', '20th centuri', '21st', '21st centuri', '256mb', '28th', '2gb', '2gb elit', '2nd', '30s', '32mb', '35mm', '35mm sourc', '3d', '3d game', '3mp', '3mp digit', '3rd', '40gb', '40s', '4th', '4x', '50s', '512mb', '512mb cf', '5th', '60s', '60x', '60x 150x', '70s', '712c', '7th', '80s', '80s pop', '90-minut', '90s', 'a.k.a', 'a/c', 'a/c adaptor', 'a1000', 'a1500',

รูปที่ 14 แสดงคำที่ผสมระหว่างตัวอักษรและตัวเลข

ซึ่งคำเหล่านี้ไม่มีความหมายและไม่มีความจำเป็นที่จะต้องใช้ในการประมวลผลผู้วิจัยจึงตัดตัวเลขออกจากคำเหล่านี้

7. Stemming การทำ Stemming คือการลดรูปของคำ ๆ เดียวกันให้อยู่ในรูปพื้นฐานของคำๆนั้น โดย Stemming จะตัดส่วนท้ายของคำออกเพื่อให้เหลือแค่รากของคำนั้นๆ ในการทำ Stemming นั้นผู้วิจัยใช้ snowball Stemmer [26] ผลของการทำ Stemming ในชุดข้อมูลของผู้วิจัย แสดงดังตารางที่ 5

ตารางที่ 5 ตัวอย่างการหารากศัพท์ของคำ

คำก่อนใช้ Stemmer	หลังใช้ stemmer	คำก่อนใช้ Stemmer	หลังใช้ stemmer
'book'	'book'	'accurate'	'accur'
'good'	'good'	'online'	'onlin'
'points'	'point'	'test'	'test'
'anything'	'anyth'	'account'	'account'
'helps'	'help'	'difference'	'differ'
'put'	'put'	'options'	'option'
'word'	'word'	'exactly'	'exact'
'want'	'want'	'like'	'like'
'supervisor'	'supervisor'	'describe'	'describ'

8. ทำการแปลงข้อมูลของทุกโดเมนให้เป็น Document – term matrix โดยแถวและคอลัมน์ของ DTM จะแสดงเอกสารและคำ (Term) ตามลำดับ นอกจากนี้ความถี่ ij (องค์ประกอบของ DTM) เป็นการแสดงความถี่ของการเกิดขึ้นของ Term j ในเอกสาร i

ตารางที่ 6 แสดง Document – term matrix (DTM)

	Term 1	Term 2	...	Term p
Document1	freq11	freg12	...	freq1p
Document2	freq21	freq22	...	freg 2p
...	...	...	freq ij	...
DocumentN	freqN1	freqN2	...	freq Np

3.3 การคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำ

ผู้วิจัยคัดเลือกฟีเจอร์ด้วยวิธีให้ค่าน้ำหนักกับคุณลักษณะที่ใช้เป็นตัวแทนของประโยค ด้วยวิธี TF-IDF และ BM25 เพื่อเปรียบเทียบประสิทธิภาพการให้ค่าน้ำหนักด้วยวิธีใดให้ผลลัพธ์ที่ดีที่สุด วิธีการให้ค่าน้ำหนักของคำทั้ง 2

3.4 การหาจำนวนกลุ่มที่เหมาะสม

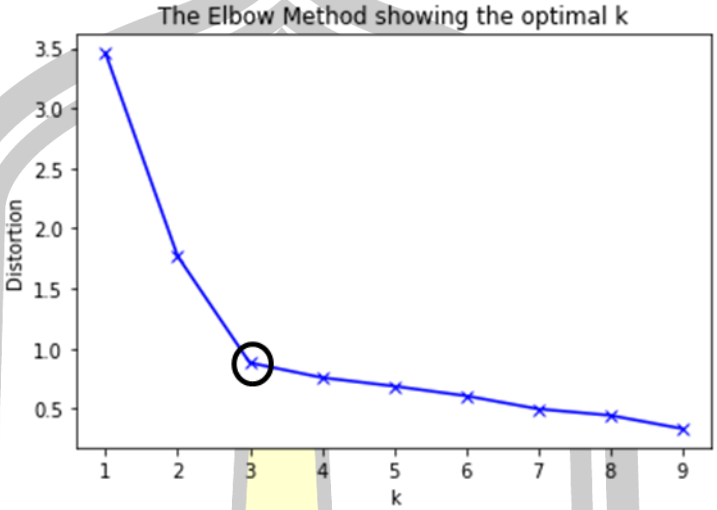
การจัดกลุ่มเอกสารด้วยวิธี K-Means ต้องมีการกำหนดจำนวนกลุ่มที่แน่นอนก่อน ซึ่งจำนวนกลุ่มที่กำหนดลงไปนั้นจะเหมาะสมหรือไม่ ต้องมีการระบุจำนวนกลุ่ม (K-value) ไว้ล่วงหน้าก่อนการจัดกลุ่ม การเลือกจำนวนกลุ่มจะส่งผลต่อผลลัพธ์ของการรวมกลุ่ม ขั้นตอนวิธีในการเลือกจำนวนกลุ่มที่เหมาะสม อาทิเช่น Elbow Method, Gap Statistic, Silhouette Coefficient ซึ่ง Elbow Method เป็นวิธีการหนึ่งที่มีประสิทธิภาพ [19] โดยจะใช้ค่า SSE เป็นตัววัดประสิทธิภาพในการหาค่า K ที่เหมาะสมในการหาจุดผันแปร (inflection point) มีความซับซ้อนน้อย จุดผันแปรขึ้นอยู่กับความสัมพันธ์ระหว่างค่า K และระยะทางถ้าจุดผันแปรไม่ชัดเจนจะไม่สามารถระบุค่า K ได้ ในส่วน Gap Statistic จะเปรียบเทียบข้อมูลด้วยค่าเฉลี่ยของชุดข้อมูลอ้างอิงกับชุดข้อมูลที่สังเกตได้เพื่อให้ค่า k ลดลงเร็วที่สุด อย่างไรก็ตามสำหรับชุดข้อมูลขนาดใหญ่วิธีนี้ไม่เหมาะสมเนื่องจากใช้เวลาและพื้นที่ในการประมวลผลมาก Silhouette Coefficient เป็นการผสานการทำงานระหว่างการจัดกลุ่มและการแยกกลุ่มเพื่อทำการวิเคราะห์คลัสเตอร์วิธีการนี้ไม่เหมาะสมกับชุดข้อมูลขนาดใหญ่เนื่องจากมีการคำนวณมาก [35]

ตารางที่ 7 ตารางเปรียบเทียบเทคนิคการหาจำนวนกลุ่มที่เหมาะสม

Elbow Method	Gap Statistic	Silhouette Coefficient
ใช้ค่า sum of squared errors (SSE)	เปรียบเทียบข้อมูลด้วยค่าเฉลี่ยของชุดข้อมูลอ้างอิงกับชุดข้อมูล	ใช้ค่า Silhouette score วัดว่าข้อมูลนั้นเหมือนกับ Cluster ที่มันอยู่มากน้อยแค่ไหน เมื่อเทียบกับ Cluster กลุ่มอื่น ๆ
มีความซับซ้อนน้อย จุดผันแปรขึ้นอยู่กับความสัมพันธ์ระหว่างค่า K และระยะทางถ้าจุดผันแปรไม่ชัดเจนจะไม่สามารถระบุค่า K ได้	ไม่เหมาะกับชุดข้อมูลขนาดใหญ่เนื่องจากใช้เวลาและพื้นที่ในการประมวลผลมาก	ไม่เหมาะสมกับชุดข้อมูลขนาดใหญ่เนื่องจากมีการคำนวณมาก

ผู้วิจัยเลือกใช้เทคนิค Elbow ในการหาจำนวนกลุ่มที่เหมาะสมเนื่องจากมีความซับซ้อนน้อยและใช้เวลาประมวลผลไม่มากเมื่อเทียบกับอีก 2 เทคนิค ซึ่งแนวคิดของ Elbow คือวัดระยะห่างระหว่างจะข้อมูลกับจุดศูนย์กลางของคลัสเตอร์เพื่อดูแนวโน้มของ sum of squared errors (SSE) เพื่อหาจำนวนกลุ่มที่เหมาะสมของการจัดกลุ่มเอกสารโดยวิธี elbow นั้นจะทำการจัดกลุ่มบนชุด

ข้อมูลตามช่วงของกลุ่มที่กำหนดไว้ สำหรับแต่ละค่าของกลุ่มจะทำการคำนวณค่า SSE แล้วแสดงออกมากเป็นกราฟเส้น ถ้าเส้นมีการหักงอ แสดงว่าจุดตรงที่งอบนเส้นนั้นคือจำนวนกลุ่มที่ดีที่สุด



รูปที่ 15 แสดง elbow Method

3.5 การจัดกลุ่มเอกสาร

ผู้วิจัยทำการเปรียบเทียบขั้นตอนวิธีในการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี K-Means และ DBSCAN เพื่อทดสอบว่าขั้นตอนวิธีใดให้ผลลัพธ์ของการจัดกลุ่มเอกสารของทั้ง 2 ชุดข้อมูลได้ดีที่สุด เพื่อคัดเลือกขั้นตอนวิธีไปใช้ในกระบวนการต่อไป ซึ่งขั้นตอนวิธีทั้งสองนั้นจะมีการกำหนดพารามิเตอร์ที่แตกต่างกันคือในขั้นตอนวิธีการจัดกลุ่มด้วย K-Means จะกำหนดพารามิเตอร์คือจำนวนกลุ่ม (k) ในส่วนของ DBSCAN พารามิเตอร์ที่กำหนดเข้าไปจะมี 2 ค่าคือค่า eps และ Min Sample ซึ่งค่า eps หมายถึงระยะทางที่ห่างที่สุดระหว่างชุดข้อมูล 2 ชุดเพื่อพิจารณาว่าอยู่ในพื้นที่ใกล้เคียงกันหรือไม่ ถ้าค่าไม่มากกว่าระยะทางที่ห่างที่สุดจะถือว่าอยู่ในกลุ่มเดียวกัน Min Sample

3.6 C- Algorithm

C-Algorithm เป็นขั้นตอนในการหาว่าเอกสารที่ส่งเข้าไปนั้นควรอยู่ในกลุ่มเอกสารใดหรือควรออกไปสร้างกลุ่มใหม่ มีขั้นตอนดังนี้

1. คำนวณหาจำนวนกลุ่มที่เหมาะสม (K) ด้วยเทคนิค Elbow
2. คำนวณหาค่า k เพื่อใช้ในการจัดกลุ่มจาก $k = K - 1$

3. ทำการแยกชุดข้อมูลโดยกลุ่มแรกเป็นชุดข้อมูลที่ใช้ในการจัดกลุ่ม (Source Domain = k) และกลุ่มชุดข้อมูลสำหรับทดสอบ (Test Document)

4. จัดกลุ่มด้วยขั้นตอนวิธี K-Means โดยกำหนดจำนวนกลุ่มเท่ากับ k

5. คำนวณหาจุดศูนย์กลางของกลุ่ม (Centroid) โดย C_j เป็นจุดศูนย์กลางของคลัสเตอร์ j

6. คำนวณหาระยะทางของข้อมูลทุกตัวที่อยู่ในกลุ่มจากจุดศูนย์กลางของกลุ่ม

7. คำนวณหาค่า Threshold กำหนด T_m = ระยะทางของข้อมูล ณ ตำแหน่งเปอเซนไทล์ที่กำหนด โดย m คือ *Cluster Index*

8. วนรอบอ่านเอกสารที่ละตัว (td_i) จากเอกสารชุดทดสอบ (TD) โดย

   8.1 วนรอบอ่านจุดศูนย์กลางของกลุ่ม C_j โดย j คือ Cluster Index

      8.1.1 คำนวณหาระยะทางของเอกสาร (td_i) กับจุดศูนย์กลางของกลุ่ม C_j ผลลัพธ์เก็บไว้ที่ $dtc_{i,j}$ โดย i คือลำดับเอกสาร j คือ Cluster Index

      8.1.2 เก็บค่าระยะทางของเอกสารไว้ใน DTC_m โดย m คือลำดับของเอกสารตัวที่ i

9. วนรอบอ่านระยะทางของเอกสารทดสอบ (dtc_m) จาก DTC_m

   9.1 หาค่าต่ำสุดและลำดับกลุ่มที่อยู่โดยกำหนด

$$Dm_m = \min(DTC_m)$$
$$DC_m = ClusterIndex$$

   9.2 ถ้า $Dm_m < T_m$

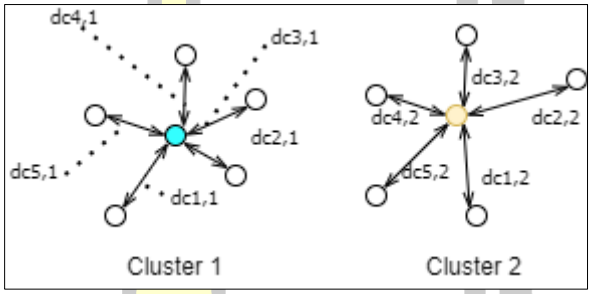
      9.2.1 กำหนดให้เอกสารอยู่ในกลุ่ม DC_m มิฉะนั้นแล้ว

      9.2.2 กำหนดให้เอกสารอยู่ในกลุ่มใหม่

3.7 การหาค่า Threshold เพื่อใช้ในการแยกเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบและการจัดเอกสารเดิมเข้ากลุ่ม

ค่า Threshold เพื่อใช้ในการพิจารณาว่าเอกสารที่ส่งเข้าไปใหม่นั้นจะสามารถแยกออกจากกลุ่มเดิมในกรณีที่มีคล้ายคล้ายคลึงกันน้อยหรือไม่มีความคล้ายคลึงกันเลย และสามารถจัดเอกสารเดิมเข้ากลุ่มเดิมได้นั้น ผู้วิจัยใช้มีวิธีการดังนี้

1. คำนวณระยะทางของข้อมูลทุกตัวที่อยู่ในกลุ่มกับจุดศูนย์กลางของกลุ่ม



รูปที่ 16 แสดงการหาระยะทางของข้อมูลจากจุดศูนย์กลางแต่ละกลุ่ม

จากรูปที่ 16 สามารถอธิบายได้ว่า ข้อมูลตัวที่ 1 ของคลัสเตอร์ที่ 1 $(dc_{1,1})$ จะทำการคำนวณระยะทางจากข้อมูลไปยังจุดศูนย์กลางของกลุ่ม และข้อมูลตัวอื่น ๆ ที่อยู่ในกลุ่มก็จะทำเช่นเดียวกัน ในส่วนคลัสเตอร์ที่ 2 ข้อมูลทุกตัวก็จะทำการคำนวณระยะทางไปยังจุดศูนย์กลางของกลุ่มเช่นเดียวกันกับคลัสเตอร์ที่ 1 ในกรณีที่มีมากกว่า 2 กลุ่มก็จะทำเช่นนี้เหมือนกัน

2. เมื่อได้ระยะทางของข้อมูลทุกตัวในแต่ละกลุ่มแล้วนำระยะทางมาจัดเรียงลำดับจากน้อยไปมาก ตัวอย่างดังตารางด้านล่าง

ตารางที่ 8 ตัวอย่างการเรียงลำดับระยะทางของข้อมูลแต่ละกลุ่มจากจุดศูนย์กลาง

เอกสาร Cluster 1	ระยะทางจากจุดศูนย์กลาง
dc2,1	0.873
dc3,1	0.921
dc5,1	0.954
dc4,1	0.982
dc1,1	1.05

เอกสาร Cluster 2	ระยะทางจากจุดศูนย์กลาง
dc4,2	0.721
dc3,2	0.752
dc5,1	0.843
dc1,2	0.876
dc2,2	0.923

3. การกำหนดค่า Threshold ผู้วิจัยใช้ตำแหน่งเปอร์เซนไทล์ในการกำหนดค่า Threshold ที่เหมาะสม ด้วยสมการ

$$k = k^{th}\ percentile \qquad (3.1)$$

$$i_j = \frac{k}{100}(n+1)$$

โดย n จำนวนข้อมูลทั้งหมดในกลุ่ม j

ถ้า i เป็นจำนวนเต็มแล้ว $k^{th}\ percentile$ คือค่าของข้อมูลที่อยู่ในตำแหน่งที่ i ของกลุ่ม j

ถ้า i ไม่เป็นจำนวนเต็ม $k^{th}\ percentile$ จะต้องนำตำแหน่งที่แตกต่างกันและระยะมาคำนวณความต่างกัน

จากตารางที่ 8 สามารถแสดงการคำนวณค่า Threshold ของกลุ่มที่ 1 ได้ดังนี้

จากสมการที่ 3.1 กำหนดตำแหน่งเปอเซนไทล์ที่ 70 จะได้ $i_j = \frac{70}{100}(5+1)$

ดังนั้น $i_j = 4.2$ จะได้ข้อมูลตำแหน่งระหว่างข้อมูลตำแหน่งที่ 4 และ 5 ซึ่ง i ไม่เป็นจำนวนเต็ม ขั้นต่อไปคือคือการหาตำแหน่งที่ต่างกันและระยะที่ต่างกันจะได้

ตำแหน่งที่ต่างกันของข้อมูล 5-4 = 1 ระยะที่ต่างกัน 1.05 – 0.982 = 0.068

ตำแหน่งที่ต่างกัน 4.2 – 4 = 0.2 ระยะที่ต่างกัน (0.068 x 0.2)/1 = 0.0136

ข้อมูลตำแหน่งที่ 4 คือ 0.982 + 0.0136 = 0.9956

ดังนั้นค่า Threshold ของกลุ่มที่ 1 ที่ตำแหน่งเปอเซนไทล์ที่ 70 คือ 0.9956 และในกลุ่มที่ 2 ก็จะใช้วิธีคำนวณเดียวกันซึ่งจะได้ 0.8852

ค่า Threshold ของแต่ละกลุ่ม (Local Threshold) เป็นค่าที่ใช้วัดระยะทางของข้อมูลที่เข้าไปใหม่กับกลุ่มนั้นซึ่งแต่ละกลุ่มจะแตกต่างกัน (Multi Threshold) การกำหนดค่า Threshold ของระยะทางแต่ละข้อมูลในกลุ่มที่มีอยู่แล้วเป็นเท่าใดจะสามารถแยกเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบออกมาได้ดีที่สุดและในการจัดเอกสารเข้ากลุ่มเดิมค่าที่ให้ผลลัพธ์ดีที่สุดควรเป็นเท่าไร ในการทดลองผู้วิจัยจะกำหนดตำแหน่งเปอเซนไทล์หลาย ๆ ค่าเพื่อหาผลลัพธ์ที่ดีที่สุด

3.8 การทดสอบประสิทธิภาพของ C-Algorithm และการวิธีวัดความคล้ายคลึงของเอกสารด้วยเทคนิคต่าง ๆ

เพื่อทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอผู้วิจัยส่งเอกสารกลุ่มใหม่เข้าไปทดสอบว่าสามารถแยกเอกสารออกจากกลุ่มที่มีอยู่แล้วหรือจัดเอกสารเข้ากลุ่มเดิมได้หรือไม่ ผู้วิจัยจะใช้ค่า Threshold ที่คำนวณได้จากขั้นตอนก่อนหน้ามาใช้ในการแยกเอกสารออกจากกลุ่ม และเปรียบเทียบการวัดความคล้ายคลึงของเอกสารด้วยวิธี Euclidean Manhattan Minkowski และ Cosine วิธีใดที่ให้ผลลัพธ์ของขั้นตอนวิธีที่นำเสนอดีที่สุดในการแยกเอกสารกลุ่มใหม่ออกจากกลุ่มที่มีอยู่แล้วและการจัดเอกสารเข้ากลุ่มเดิม

3.9 การวัดประสิทธิภาพ

วิธีการวัดประสิทธิภาพของการจัดกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าการจำได้ (Recall) และค่าเฉลี่ยประสิทธิภาพโดยรวม (F-Measure

1. การวัดค่าความถูกต้อง (Accuracy)

การวัดค่าความถูกต้องของการจำแนกดังสมการ

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (3.2)$$

2. การวัดค่าความแม่นยำ (Precision)

การวัดค่าความแม่นยำของการจำแนกดังสมการ

$$Precision = \frac{TP}{(TP + FP)} \quad (3.3)$$

3. การวัดค่าการจำได้ (Recall)

การวัดค่าการจำได้เป็นการวัดความถูกต้องของวิธีการ ดังสมการ

$$Recall = \frac{TP}{(TP + FN)} \quad (3.4)$$

4. ค่าเฉลี่ยประสิทธิภาพโดยรวม (F-measure)

ค่าเฉลี่ยประสิทธิภาพโดยรวม เป็นการนำค่าความระลึกและค่าความแม่นยำมาพิจารณาร่วมกัน ระบบที่มีประสิทธิภาพดีจะต้องมีค่าความระลึกและค่าความแม่นยำ สูงใกล้เคียงกัน ดังสมการ

$$F - Measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (3.5)$$

บทที่ 4
ผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาขั้นตอนวิธีในการจัดกลุ่มของโดเมนที่แตกต่างกันตามความคล้ายคลึงกันของโดเมนเพื่อบ่งบอกว่าข้อมูลที่นำไปจัดกลุ่มนั้นควรจะอยู่ในกลุ่มของโดเมนนั้นหรือแยกออกมาสร้างกลุ่มของโดเมนใหม่และให้ประสิทธิภาพที่ดีที่สุด ประกอบด้วยข้อมูลที่ใช้ในการทดลอง การเตรียมข้อมูล ในส่วนของการทดลองมีทั้งหมด 9 การทดลองดังนี้

1. กำหนดพารามิเตอร์ที่ใช้ในการคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำของTF-IDF และ BM25 มีวัตถุประสงค์เพื่อหาค่าพารามิเตอร์ที่ให้ผลลัพธ์ดีที่สุดในการจัดกลุ่มของทั้ง 2 ชุดข้อมูล
2. เปรียบเทียบวิธีคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำระหว่าง TF-IDF และ BM25 มีวัตถุประสงค์เพื่อหาประสิทธิภาพของวิธีคัดเลือกคุณลักษณะว่าวิธีใดให้ผลลัพธ์ดีที่สุด
3. เปรียบเทียบประสิทธิภาพการให้น้ำหนักทั้งสองวิธีด้วย Pair Sample T-Test มีวัตถุประสงค์เพื่อยืนยันประสิทธิภาพของวิธีคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำระหว่าง TF-IDF และ BM25
4. เปรียบเทียบประสิทธิภาพขั้นตอนวิธีการจัดกลุ่มระหว่าง K-Means และ DBSCAN มีวัตถุประสงค์เพื่อหาประสิทธิภาพของขั้นตอนวิธีจัดกลุ่มว่าวิธีใดให้ผลลัพธ์ดีที่สุด
5. ตรวจสอบจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มเอกสาร มีวัตถุประสงค์เพื่อตรวจสอบว่า Elbow method สามารถกำหนดจำนวนกลุ่มได้ถูกต้องตามจำนวนโดเมนเอกสารที่ส่งเข้าไปหรือไม่
6. การหาตำแหน่ง Percentile ของชุดข้อมูลในกลุ่มเพื่อหาค่า Threshold นำไปใช้ในการแยกเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบ มีวัตถุประสงค์เพื่อหาตำแหน่งเปอเซนไทล์ของระยะทางแต่ละข้อมูลในกลุ่มที่มีอยู่แล้วเป็นเท่าใดจะสามารถแยกเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบออกมาได้ดีที่สุด
7. ส่งเอกสารกลุ่มใหม่เข้าไปทดสอบ C- Algorithm และการวัดความคล้ายคลึงของเอกสารด้วยวิธีต่าง ๆ มีวัตถุประสงค์เพื่อทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอและเปรียบเทียบการวัดความคล้ายคลึงของเอกสารด้วยวิธี Euclidean Manhattan

Minkowski และ Cosine วิธีใดที่ให้ผลลัพธ์ของขั้นตอนวิธีที่นำเสนอดีที่สุดในการแยกเอกสารกลุ่มใหม่ออกจากกลุ่มเดิม

8. การหาตำแหน่ง Percentile ของชุดข้อมูลในกลุ่มเพื่อหาค่า Threshold นำไปใช้ในการการทดสอบการรวมกลุ่มการกำหนดตำแหน่งเปอเซนไทล์ของระยะทางแต่ละข้อมูลในกลุ่มที่มีอยู่แล้วเป็นเท่าใดจะสามารถแยกเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบและจัดเอกสารกลุ่มเดิมเข้าไปในกลุ่มได้ดีที่สุด
9. ส่งเอกสารทุกกลุ่มเข้าไปทดสอบ C- Algorithm และการวัดความคล้ายคลึงของเอกสารด้วยวิธีต่าง ๆ มีวัตถุประสงค์เพื่อทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอและเปรียบเทียบการวัดความคล้ายคลึงของเอกสารด้วยวิธี Euclidean Manhattan Minkowski และ Cosine วิธีใดที่ให้ผลลัพธ์ของขั้นตอนวิธีที่นำเสนอดีที่สุดในการแยกเอกสารและจัดเอกสารเข้ากลุ่ม
10. เปรียบเทียบการวัดความคล้ายคลึงของเอกสารที่ส่งเข้าไปทดสอบหลังจากจัดกลุ่มด้วยขั้นตอนวิธี K-Means มีวัตถุประสงค์เพื่อทดสอบว่าหลังจากจัดกลุ่มเอกสารแล้วเมื่อมีเอกสารใหม่เข้าไป ขั้นตอนวิธีที่นำเสนอมีผลลัพธ์แตกต่างกันอย่างไรกับการหาความคล้ายคลึงของเอกสารด้วยวิธี K-Means แบบเดิม

4.1 ข้อมูลที่ใช้ในการทดลอง

4.1.1 เครื่องมือและข้อมูลต่าง ๆ ที่ใช้ในการทดลองในงานวิจัย

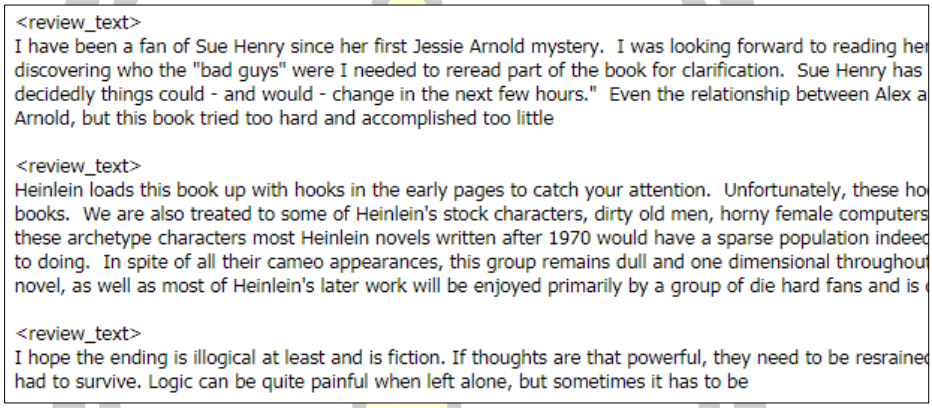
ในด้านฮาร์ดแวร์ประกอบด้วยเครื่องคอมพิวเตอร์ Intel(R) Core(TM) i7-7700 3.60 GHz RAM 8 GB ในด้านซอฟต์แวร์ที่ใช้ในการทดลองได้แก่ ระบบปฏิบัติการ Windows10 และภาษา Python ในการทดลอง

4.1.2 การรวบรวมข้อมูลในการทดลอง

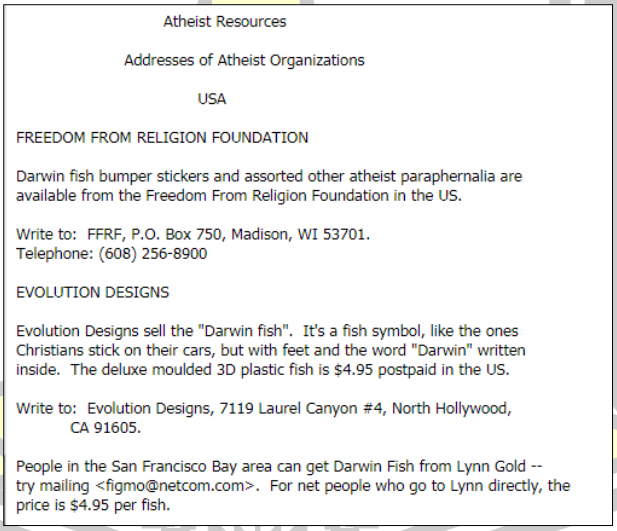
ในงานวิจัยนี้ผู้วิจัยเลือกใช้ชุดข้อมูล 2 ชุดได้แก่

1. ชุดข้อมูล Multi-Domain Sentiment dataset ซึ่งรวบรวมโดย Blitzer และคณะ [53] ซึ่งรวบรวมจาก Amazon.com ประกอบไปด้วยสินค้าจำนวน 20 ประเภท (โดเมน) สามารถเข้าถึงได้จาก https://www.cs.jhu.edu/~mdredze/datasets/sentiment/ ผู้วิจัยคัดเลือกมาจำนวน 3 โดเมนประกอบไปด้วยสินค้าที่แตกต่างกัน 3 ประเภทคือ Book, DVDs, electronics โดยเลือกตัวอย่างรีวิวจากโดเมนทั้ง 3 อย่างละ 600 รีวิว รวมเป็น 1800 ตัวอย่าง

2. ชุดข้อมูล 20 newsgroups text dataset ซึ่งเป็นชุดข้อมูลมาตรฐานที่นิยมนำมาศึกษาและเปรียบเทียบขั้นตอนวิธีในการจัดกลุ่มรวมถึงการพัฒนาขั้นตอนวิธีในการจัดกลุ่มเพื่อประสิทธิภาพที่ดีขึ้น ซึ่งผู้วิจัยคัดเลือกมา 3 โดเมนโดยวิธีการคัดเลือกแบบเจาะจง ได้แก่ โดเมน alt.atheism มีจำนวน 480 เอกสาร misc.forsale มีจำนวน 585 เอกสาร sci.electronics มีจำนวน 591 เอกสาร รวมทั้งสิ้น 1,656 เอกสารจากทั้ง 3 โดเมน



รูปที่ 17 ตัวอย่างเอกสารของชุดข้อมูล Multi Domain Sentiment Dataset



รูปที่ 18 ตัวอย่างเอกสารของชุดข้อมูล 20 News Group Dataset

4.2 การเตรียมข้อมูล

ข้อมูลที่ใช้ในการทดลองเป็นข้อความในเอกสารในรูปแบบที่ไม่มีโครงสร้าง จึงต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบที่มีโครงสร้าง เพื่อใช้ในขั้นตอนการคัดเลือกคุณลักษณะและจัดกลุ่ม งานวิจัยนี้สกัดคุณลักษณะโดยการตัดคำที่อยู่ในเอกสารด้วยวิธีการ Unigrams และ Bigrams เนื่องจากให้

ประสิทธิภาพของการจัดกลุ่มบนชุดข้อมูลที่ผู้วิจัยนำเสนอดีที่สุด หลังจากนั้นผู้วิจัยได้เข้าสู่กระบวนการเตรียมข้อมูลก่อนการประมวลผลดังขั้นตอนดังนี้

1. ตัดคำ (tokenize)

This book has its good points ^^. If anything, it helps you put into words what you want fro m a supervisor, but it is not very accurate. The online test doesn’t account for a difference between when 2 of their options are both exactly like you, or if they don't describe you at all 555.

รูปที่ 19 ตัวอย่างเอกสารต้นฉบับ

```
        'This', 'book', 'has', 'its', 'good', 'points', '^^'
, '.', 'If', 'anything', ',', 'it', 'helps', 'you', 'put', '
into', 'words', 'what', 'you', 'want', 'from', 'a', 'supervi
sor', ',', 'but', 'it', 'is', 'not', 'very', 'accurate', '.'
, 'The', 'online', 'test', 'doesn', '’', 't', 'account', 'fo
r', 'a', 'difference', 'between', 'when', '2', 'of', 'their'
, 'options', 'are', 'both', 'exactly', 'like', 'you', ',', '
or', 'if', 'they', 'do', "n't", 'describe', 'you', 'at', 'all
', '555', '.'
```

รูปที่ 20 ตัวอย่างเอกสารหลังจากตัดคำ

2. เปลี่ยนคำเป็นอักษรตัวเล็ก

'this', 'book', 'has', 'its', 'good', 'points', '^^', '.', 'if', 'anything', ',', 'it', 'helps', 'you', 'put', 'into', ' words', 'what', 'you', 'want', 'from', 'a', 'supervisor', ',', 'but', 'it', 'is', 'not', 'very', 'accurate', '.', ' the', 'online', 'test', 'doesn', '’', 't', 'account', 'for', 'a', 'difference', 'between', 'when', '2', 'of', ' their', 'options', 'are', 'both', 'exactly', 'like', 'you', ',', 'or', 'if', 'they', 'do', "n't", 'describe', 'you', 'at', 'all', '555', '.'

รูปที่ 21 ตัวอย่างเอกสารหลังจากเปลี่ยนตัวอักษร

3. การตัด punctuation และตัวเลขออก
4. กำจัดคำหยุด (Stop Word)

การลบคำหยุดออกจากข้อความทำให้จำนวนคุณลักษณะของเอกสารลดลงเพิ่มความเร็วในการประมวลผล การกำจัดคำหยุดจะใช้วิธีการสร้างคลังเก็บคำหยุดไว้ เมื่อมีการประมวลผลเอกสารจะนำคำในเอกสารไปเปรียบเทียบกับคำหยุดที่อยู่ในคลัง

'book', 'good', 'points', 'anything', 'helps', 'put', 'words', 'want', 'supervisor', 'accurate', 'online', 'test', 'account', 'difference', 'options', 'exactly', 'like', 'describe'

รูปที่ 22 ตัวอย่างเอกสารหลังจากกำจัดหยุด

5. ตัดคำที่มีตัวเลขผสมอยู่ออก

ผู้วิจัยพิจารณาดูเอกสารตัวอย่างพบว่ามีคำจำนวนหนึ่งที่มีตัวเลขผสมรวมอยู่ในตัวอักษรซึ่งกระบวนก่อนหน้าไม่สามารถจัดการซึ่งคำเหล่านี้ไม่มีความหมายและไม่มีความจำเป็นที่จะต้องใช้ในการประมวลผลผู้วิจัยจึงตัดคำทีมีตัวเลขผสมออก

'10th', '12-in-1', '12-in-1 lexar', '12-in-1 sandisk', '128mb', '128mb jump', '150x', '150x sd', '16mm', '16mm compar', '16mm dupe', '16mm print', '17th', '18th', '18th centuri', '1920s', '1920s except', '1930s', '1960s', '1970s', '1980s', '19th', '19th centuri', '1gb', '1st', '2-disc', '2-year', '20th', '20th anniversari', '20th centuri', '21st', '21st centuri', '256mb', '28th', '2gb', '2gb elit', '2nd', '30s', '32mb', '35mm', '35mm sourc', '3 d', '3 d game', '3 mp', '3 mp digit', '3 rd', '40 gb', '40 s', '4 th', '4 x', '50 s', '512mb', '512mb cf', '5th', '60s', '60x', '60x 150x', '70s', '712c', '7th', '80s', '80s pop', '90-minut', '90s', 'a.k.a', 'a/c', 'a/c adaptor', 'a1000', 'a1500'

รูปที่ 23 แสดงตัวอย่างของคำที่ผสมระหว่างตัวอักษรและตัวเลขในชุดข้อมูลที่ทำการทดลอง

6. Stemming เพื่อหารากศัพท์ของคำเพื่อลดจำนวนของคำในถุงคำ ผู้วิจัยใช้ snowball Stemmer เพื่อหารากศัพท์ของคำ

ตารางที่ 9 ตัวอย่างการหารากศัพท์ของคำบนชุดเอกสารที่ใช้ทดลอง

คำก่อนใช้ Stemmer	**หลังใช้ stemmer**	**คำก่อนใช้ Stemmer**	**หลังใช้ stemmer**
'book'	'book'	'accurate'	'accur'
'good'	'good'	'online'	'onlin'
'points'	'point'	'test'	'test'
'anything'	'anyth'	'account'	'account'
'helps'	'help'	'difference'	'differ'

คำก่อนใช้ Stemmer	หลังใช้ stemmer	คำก่อนใช้ Stemmer	หลังใช้ stemmer
'put'	'put'	'options'	'option'
'word'	'word'	'exactly'	'exact'
'want'	'want'	'like'	'like'
'supervisor'	'supervisor'	'describe'	'describ'

4.3 การคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำ

ผู้วิจัยคัดเลือกฟีเจอร์ด้วยวิธีให้ค่าน้ำหนักกับคุณลักษณะที่ใช้เป็นตัวแทนของประโยค ด้วยเทคนิค 2 เทคนิคคือ TF-IDF และ BM25 และผู้วิจัยได้ทดลองปรับพารามิเตอร์ของทั้ง 2 เทคนิคเพื่อหาว่าการกำหนดพารามิเตอร์เท่าใดจะให้ประสิทธิภาพการจัดกลุ่มดีที่สุด สำหรับขั้นตอนวิธีในการจัดกลุ่มเพื่อทดสอบประสิทธิภาพการให้ค่าน้ำหนักนั้นผู้วิจัยใช้ขั้นตอนวิธี K – Means โดยกำหนดจำนวนกลุ่ม (k) เป็น 3 กลุ่มตามจำนวนโดเมนที่ส่งเข้าไปทดสอบของทั้งสองชุดข้อมูล ได้ผลการทดลองดังนี้ ได้คุณลักษณะ (Feature) ที่เป็นตัวแทนของเอกสารดังนี้

4.3.1 การให้ค่าน้ำหนักกับคุณลักษณะด้วย TF-IDF

การให้ค่าน้ำหนักกับคุณลักษณะด้วย TF-IDF บนชุดข้อมูล Multi Domain Dataset และ 20 News group ผู้วิจัยใช้วิธีคัดเลือกคุณลักษณะด้วยการด้วยการละเว้นคำที่ปรากฏในเอกสารน้อยโดยทดลองกำหนดตั้งแต่ 2 เอกสารและเพิ่มขึ้นไปเรื่อย ๆ เพื่อหาผลลัพธ์ในการจัดกลุ่มว่าค่าที่เท่าไรให้ผลลัพธ์ดีที่สุด ผู้วิจัยได้ทดลองโดยการตัดคำที่อยู่ในเอกสารด้วยวิธีการ Unigrams และ Bigrams ผลของการคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำด้วย TF-IDF แสดงดังตารางด้านล่าง

ตารางที่ 10 แสดงการให้ค่าน้ำหนักกับคุณลักษณะด้วย TF-IDF บนชุดข้อมูล Multi Domain

Min-df	Avg. Recall	Avg. Precision	F1
2	**0.88**	**0.90**	**0.88**
3	0.86	0.90	0.86
4	0.87	0.89	0.87
5	0.86	0.89	0.86
6	0.86	0.89	0.86

Min-df	Avg. Recall	Avg. Precision	F1
7	**0.88**	**0.90**	**0.88**
8	0.83	0.87	0.83
9	0.85	0.89	0.85
10	0.84	0.88	0.85

ตารางที่ 11 แสดงการให้ค่าน้ำหนักกับคุณลักษณะด้วย TF-IDF บนชุดข้อมูล 20 News Group

Min-df	Avg. Recall	Avg. Precision	F1
2	0.79	0.86	0.80
3	0.78	0.86	0.79
4	0.79	0.86	0.82
5	0.78	0.83	0.78
6	0.80	0.87	0.81
7	0.77	0.86	0.78
8	**0.81**	**0.87**	**0.82**
9	0.78	0.86	0.79
10	0.79	0.86	0.80

จากตารางที่ 10 และ ตารางที่ 11 แสดงผลลัพธ์แสดงผลการการให้ค่าน้ำหนักกับคุณลักษณะด้วย TF-IDF บนชุดข้อมูลทั้งสองชุดข้อมูลด้วยการปรับพารามิเตอร์เพื่อดูค่าพารามิเตอร์ใดดีที่สุด ผลการทดลองพบว่าบนชุดข้อมูล Multi Domain การกำหนดพารามิเตอร์ Min DF ที่ให้ค่า Recall และ F1 ดีที่สุดคือ 2 โดยให้ค่า Recall เป็น 88% Precision เป็น 90% และค่า F1 เป็น 88%
ผลการทดลองพบว่าบนชุดข้อมูล 20 News Group การกำหนดพารามิเตอร์ Min DF ที่ให้ค่า Recall และ Accuracy ดีที่สุดคือ 8 โดยให้ค่า Recall เป็น 81% ค่า Precision 87% และค่า F1 เป็น 82%

4.3.2 การให้ค่าน้ำหนักกับคุณลักษณะด้วย BM25

การให้ค่าน้ำหนักกับคุณลักษณะด้วย BM25 บนชุดข้อมูล Multi Domain Dataset และ 20 News group จะต้องส่งพารามิเตอร์ 2 ค่าเข้าไปคือ k และ b โดยทั่วไปการกำหนดพารามิเตอร์นั้นเป็น $1.2 < k_1 < 2.0$ และ $0.5 < b < 0.8$ ผู้วิจัยได้ทดลองกำหนดพารามิเตอร์

ทั้ง 2 ด้วยค่าต่าง ๆ เพื่อหาผลลัพธ์ที่ดีที่สุด คัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำด้วย TF-IDF แสดงดังตารางด้านล่าง

ตารางที่ 12 แสดงการให้ค่าน้ำหนักกับคุณลักษณะด้วย BM25 บนชุดข้อมูล Multi Domain Dataset

k	b	Avg. Recall	Avg. Precision	F1
1.2	0.5	0.48	0.65	0.37
1.3	0.6	0.55	0.37	0.44
1.4	0.7	0.48	0.66	0.37
1.5	0.8	0.64	0.48	0.53
1.6	0.9	0.65	0.48	0.54
1.7	1.0	0.63	0.43	0.51
1.8	1.1	0.64	0.82	0.53
1.9	**1.2**	**0.88**	**0.89**	**0.88**
2.0	1.3	0.48	0.41	0.40

ตารางที่ 13 แสดงการให้ค่าน้ำหนักกับคุณลักษณะด้วย BM25 บนชุดข้อมูล 20 News Group

k	b	Avg. Recall	Avg. Precision	F1
1.2	0.5	0.42	0.80	0.33
1.3	0.6	0.34	0.36	0.19
1.4	0.7	0.50	0.47	0.42
1.5	**0.8**	**0.61**	**0.76**	**0.61**
1.6	0.9	0.53	0.81	0.45
1.7	1.0	0.34	0.45	0.20
1.8	1.1	0.57	0.67	0.49
1.9	1.2	0.58	0.64	0.49
2.0	1.3	0.34	0.79	0.20

จากตารางที่ 12 และ ตารางที่ 13 แสดงผลลัพธ์แสดงผลการการให้ค่าน้ำหนักกับคุณลักษณะด้วย BM25 บนชุดข้อมูลทั้งสองชุดข้อมูลด้วยการปรับพารามิเตอร์เพื่อดูค่าพารามิเตอร์ใดดีที่สุด ผลการทดลองพบว่าบนชุดข้อมูล Multi Domain การกำหนดพารามิเตอร์ที่ให้ค่า Recall Precision และ

F1 ดีที่สุดคือ k_1 = 1.9 และ b = 1.2 โดยให้ค่าเฉลี่ย Recall เป็น 88% Precision 89% และค่า F1 เป็น 88%

ผลการทดลองบนชุดข้อมูล 20 News Group การกำหนดพารามิเตอร์ที่ให้ผลลัพธ์ดีที่สุดคือ k_1 = 1.5 และ b = 0.8 โดยให้ค่า Recall เป็น 61% Precision 76% และค่า F1 เป็น 61 %

จากผลการทดลองให้ค่าน้ำหนักกับคุณลักษณะทั้งสองวิธีพบว่า การให้ค่าน้ำหนักด้วย TF-IDF มีประสิทธิภาพดีกว่าการให้ค่าน้ำหนักด้วย BM25 บนชุดข้อมูล 20 News Group แต่สำหรับชุดข้อมูล Multi Domain เท่ากัน ผู้วิจัยจึงได้ทำการวิเคราะห์ประสิทธิภาพด้วย Paired – sample T-test ของทั้งสองชุดข้อมูล โดยผู้วิจัยตั้งสมมติฐาน

H_0 = การให้ค่าน้ำหนักกับคุณลักษณะของคำด้วย BM25 มีประสิทธิภาพของการจัดกลุ่มดีกว่า TF-IDF

H_1 = การให้ค่าน้ำหนักกับคุณลักษณะของคำด้วย TF-IDF มีประสิทธิภาพของการจัดกลุ่มดีกว่า BM25

ในการให้ค่าน้ำหนักคุณลักษณะด้วย BM25 ผู้วิจัยกำหนดค่าพารามิเตอร์ k_1 = 1.9 และ b = 1.2 บนชุดข้อมูล Multi Domain และ k_1 = 1.5 และ b = 0.8 บนชุดข้อมูล 20 News Group เนื่องจากให้ผลลัพธ์ของการจัดกลุ่มดีที่สุด

สำหรับการให้ค่าน้ำหนักคุณลักษณะด้วย TF-IDF ผู้วิจัยกำหนดค่าพารามิเตอร์ min df = 2 บนชุดข้อมูล Multi Domain และ min df = 8 บนชุดข้อมูล 20 News Group เนื่องจากให้ผลลัพธ์ของการจัดกลุ่มดีที่สุด

ผู้วิจัยทดสอบการจัดกลุ่มด้วยขั้นตอนวิธี K-Means จำนวน 10 รอบบนแต่ละชุดข้อมูลเนื่องผลการทดลองได้ดังตาราง

ตารางที่ 14 Paired-Sample T-Test บนชุดข้อมูล Multi Domain Sentiment Dataset

Muti Domain dataset	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. error Mean	95% Confidence Interval of the difference				
				Lower	Upper			
BM25	0.442	0.156	0.049	-0.5389	-0.304	8.34	9	0.0001
TF-IDF	0.866	0.019	0.0062					

ตารางที่ 15 Paired-Sample T-Test บนชุดข้อมูล 20 News Group Dataset

20 News Group Dataset	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. error Mean	95% Confidence Interval of the difference				
				Lower	Upper			
BM25	0.396	0.112	0.035	-0.3608	-0.1425	5.21	9	0.0006
TF-IDF	0.648	0.137	0.043					

จากตารางที่ 14 และตารางที่ 15 ค่า Sig (P Value) น้อยกว่า 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0 และยอมรับ H_1 สรุปได้ว่าการให้ค่าน้ำหนักกับคุณลักษณะของคำด้วย TF-IDF มีประสิทธิภาพของการจัดกลุ่มดีกว่า BM25 ดังนั้นในการทดลองขั้นต่อไปผู้วิจัยเลือกจะใช้การให้ค่าน้ำหนักด้วย TF-IDF ในการจัดกลุ่มเอกสาร

4.4 ทดสอบประสิทธิภาพขั้นตอนวิธีในการจัดกลุ่มเอกสาร

ผู้วิจัยทำการเปรียบเทียบขั้นตอนวิธีในการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี K-Means และ DBSCAN เพื่อทดสอบว่าขั้นตอนวิธีใดให้ผลลัพธ์ของการจัดกลุ่มเอกสารของทั้ง 2 ชุดข้อมูลได้ดีที่สุด เพื่อคัดเลือกขั้นตอนวิธีไปใช้ในกระบวนการต่อไป ซึ่งขั้นตอนวิธีทั้งสองนั้นจะมีการกำหนดพารามิเตอร์ที่แตกต่างกันคือในขั้นตอนวิธีการจัดกลุ่มด้วย K-Means จะกำหนดพารามิเตอร์คือจำนวนกลุ่ม (k) ในส่วนของ DBSCAN พารามิเตอร์ที่กำหนดเข้าไปจะมี 2 ค่าคือค่า eps และ Min Sample ค่า eps หมายถึงระยะทางที่ห่างที่สุดระหว่างชุดข้อมูล 2 ชุดเพื่อพิจารณาว่าอยู่ในพื้นที่ใกล้เคียงกันหรือไม่ ถ้าค่าไม่มากกว่าระยะทางที่ห่างที่สุดจะถือว่าอยู่ในกลุ่มเดียวกัน Min Sample ผู้วิจัยได้ทำผลการทดลองเป็นดังนี้

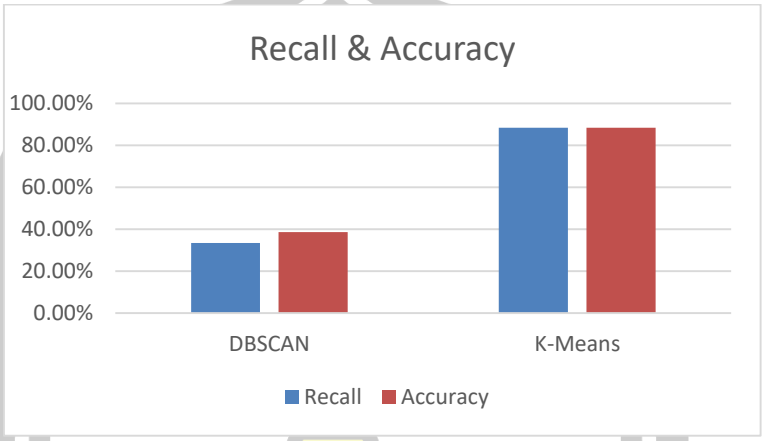
4.4.1 เปรียบเทียบขั้นตอนวิธีในการจัดกลุ่มเอกสารบนชุดข้อมูล Multi domain

ด้วยการกำหนดพารามิเตอร์ eps = 1.3 และ min_sample = 35 (0.5 กับ 5 เป็นค่าดีฟอนต์) สาเหตุที่ต้องกำหนดพารามิเตอร์ดังนี้เนื่องจากในการจัดกลุ่มของเอกสารพารามิเตอร์ข้างต้นสามารถแบ่งกลุ่มได้ 3 กลุ่มตามจำนวนโดเมน แสดงผลของการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี DBScan ด้วยการให้ค่าน้ำหนักของคำด้วย TF-IDF บนชุดข้อมูล multi domain

ตารางที่ 16 แสดงผลของการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี K-Means และ DBScan

ขั้นตอนวิธี	**Accuracy**
DBSCAN	38.5%

ขั้นตอนวิธี	Accuracy
K-Means	88.33%



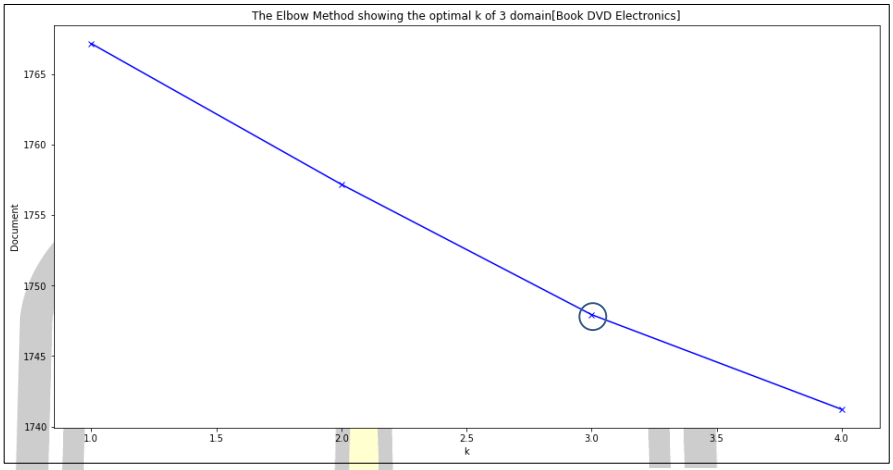
รูปที่ 24 แสดงกราฟเปรียบเทียบค่า Recall และ Accuracy ของการจัดกลุ่มเอกสารด้วยขั้นตอนวิธี K-Means และ DBScan

4.5 ขั้นตอนวิธีในการพิจารณาเอกสารจะอยู่ในกลุ่มหรือแยกกลุ่ม

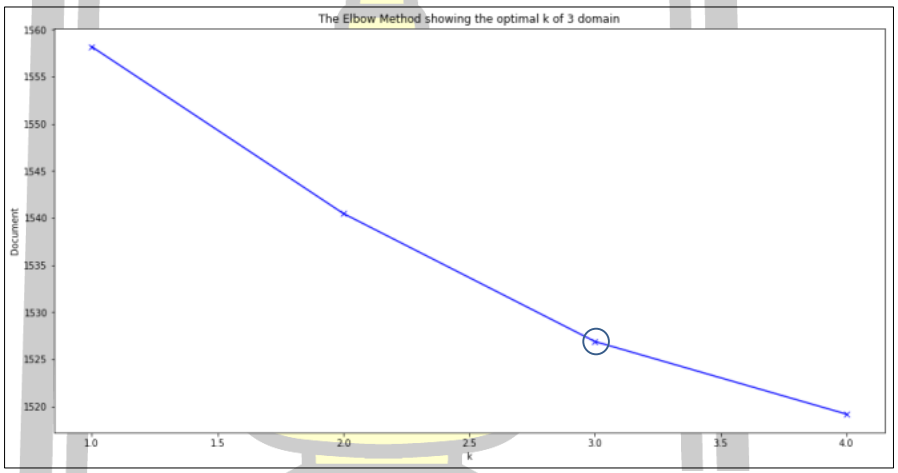
หลังจากขั้นตอนเตรียมการก่อนการประมวลผลแล้วจะเข้าสู่ขั้นตอนวิธีในการทดลองพิจารณาว่าถ้ามีเอกสารกลุ่มใหม่เข้าไปจะสามารถยุบรวมกลุ่มหรือแยกกลุ่มซึ่งได้อธิบายในบทที่ 3 จากขั้นตอนวิธีที่ผู้วิจัยนำเสนอข้างต้นได้ผลการทดลองดังนี้

4.5.1 ผลการตรวจสอบจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มเอกสาร

การจัดกลุ่มเอกสารด้วยวิธี K-Means คือการกำหนดจำนวนกลุ่มที่แน่นอนในขั้นตอนการจัดกลุ่ม ซึ่งจำนวนกลุ่มที่กำหนดลงไปนั้นจะเหมาะสมหรือไม่ แนวคิดของ Elbow คือวัดระยะห่างระหว่างจะข้อมูลกับจุดศูนย์กลางของคลัสเตอร์เพื่อดูแนวโน้มของ sum of squared errors (SSE) เพื่อหาจำนวนกลุ่มที่เหมาะสมของการจัดกลุ่มเอกสารโดยวิธี Elbow นั้นจะทำการจัดกลุ่มบนชุดข้อมูลตามช่วงของกลุ่มที่กำหนดไว้ สำหรับแต่ละค่าของกลุ่มจะทำการคำนวณค่า SSE แล้วแสดงออกมากเป็นกราฟเส้น ถ้าเส้นมีการหักงอ แสดงว่าจุดตรงที่หักงอบนเส้นนั้นคือจำนวนกลุ่มที่ดีที่สุด ผู้วิจัยใช้วิธีการ Elbow เพื่อดูจำนวนกลุ่มที่เหมาะสมของการจัดกลุ่มเอกสารโดยการส่งเอกสารทั้ง 3 โดเมนของชุดข้อมูลทั้ง 2 ชุดข้อมูล ผลของการใช้ Elbow เพื่อหาจำนวนกลุ่มที่เหมาะสมเพื่อกำหนดค่า K ได้ดังนี้



รูปที่ 25 แสดงจำนวนกลุ่มที่เหมาะสมของชุดข้อมูล Multi-Domain Sentiment dataset



รูปที่ 26 แสดงจำนวนกลุ่มที่เหมาะสมของชุดข้อมูล 20NewsGroup

จากรูปที่ 25 และรูปที่ 26 จะเห็นว่าจำนวนกลุ่มที่เหมาะสมมีจำนวน 3 กลุ่มดังนั้นผู้วิจัยจึงกำหนดค่า K = 3 แล้วทำการตัดเอกสารกลุ่มสุดท้ายออกแล้วส่งเข้าไปทดสอบขั้นตอนวิธีที่นำเสนอว่าสามารถกำหนดเอกสารให้อยู่ในกลุ่มหรือออกไปสร้างกลุ่มใหม่ได้หรือไม่

4.5.2 ผลการทดลองส่งเอกสารอีกโดเมนหนึ่งเข้าไปทดสอบบนชุดข้อมูลบนชุดข้อมูล Multi-Domain Sentiment dataset

ผู้วิจัยได้แยกเอกสารของกลุ่มที่ 3 ของชุดข้อมูล Multi Domian Sentiment Dataset ออกมาจำนวน 476 เอกสารแล้วส่ง เข้าไปทดสอบว่าสามารถแบ่งกลุ่มได้ถูกต้องหรือไม่ โดยผู้วิจัยได้คำนวณค่า Threshold เพื่อจะใช้เป็นเกณฑ์ในการแบ่งกลุ่มของเอกสารโดยคำนวณค่าเปอร์เซ็นไทร์ของระยะทางของเอกสารกลุ่มที่ 1 และกลุ่มที่ 2 ที่ 45 50 55 60 65 70 75 80 85 90 95 และ 100

แล้วส่งเอกสารกลุ่มที่ 3เข้าไปทดสอบตามขั้นตอนวิธีที่ผู้วิจัยนำเสนอ โดยใช้วิธีวัดความคล้ายคลึงของเอกสารด้วยวิธี Euclidean Distance Manhattan Minkowski และ Cosine เพื่อทดสอบว่าวิธีวัดความคล้ายคลึงของเอกสารวิธีการใดให้ผลลัพธ์ที่ดีที่สุด ผลการทดลองดังตารางด้านล่าง

ตารางที่ 17 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Euclidean Distance บนชุดข้อมูล Multi-Domain Sentiment dataset

Percentile (Threshold)	Prediction with Euclidean distance			Recall
	Cluster1	Cluster2	New Group	
45	**24**	**12**	**440**	**0.92**
50	32	18	426	0.89
55	43	23	410	0.86
60	56	30	390	0.82
65	69	34	373	0.78
70	83	39	354	0.74
75	110	51	315	0.66
80	129	62	285	0.60
85	174	74	228	0.48
90	248	86	142	0.30
95	311	95	70	0.15
100	364	110	2	0.00

ตารางที่ 18 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Manhattan Distance บนชุดข้อมูล Multi-Domain Sentiment dataset

Percentile (Threshold)	Prediction with Manhattan distance			Recall
	Cluster1	Cluster2	New Group	
45	**75**	**0**	**401**	**0.84**
50	80	0	396	0.83
55	98	0	378	0.79
60	116	0	360	0.76
65	128	0	348	0.73
70	147	0	329	0.69

Percentile (Threshold)	Prediction with Manhattan distance			Recall
	Cluster1	Cluster2	New Group	
75	172	0	304	0.64
80	199	0	277	0.58
85	232	0	244	0.51
90	289	0	187	0.39
95	341	0	135	0.28
100	438	0	38	0.08

ตารางที่ 19 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Minkowski Distance บนชุดข้อมูล Multi-Domain Sentiment dataset

Percentile (Threshold)	Prediction with Minkowski distance			Recall
	Cluster1	Cluster2	New Group	
45	**206**	**112**	**158**	**0.33**
50	216	122	138	0.29
55	226	125	125	0.26
60	236	134	106	0.22
65	246	140	90	0.19
70	251	144	81	0.17
75	261	149	66	0.14
80	267	155	54	0.11
85	273	159	44	0.09
90	278	169	29	0.06
95	284	172	20	0.04
100	296	180	0	0.00

ตารางที่ 20 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Cosine บนชุดข้อมูล Multi-Domain Sentiment dataset

Percentile (Threshold)	Prediction with Cosine			Recall
	Cluster1	Cluster2	New Group	
45	239	195	42	0.09
50	246	196	34	0.07
55	252	197	27	0.06
60	255	197	24	0.05

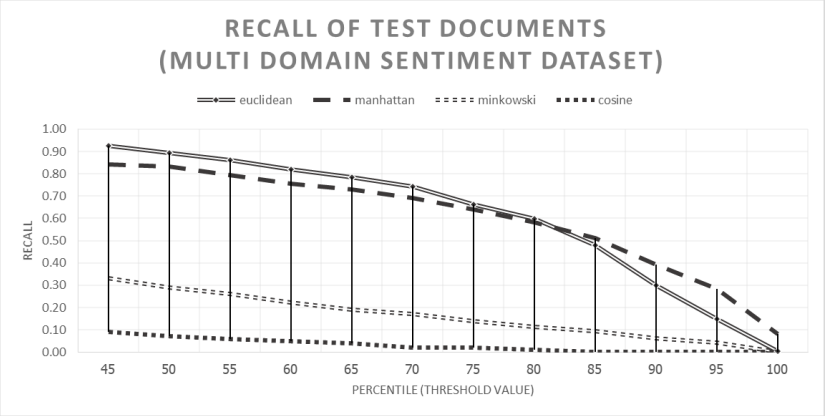
Percentile (Threshold)	Prediction with Cosine			Recall
	Cluster1	Cluster2	New Group	
65	262	197	17	0.04
70	268	197	11	0.02
75	271	197	8	0.02
80	274	197	5	0.01
85	277	197	2	0.00
90	277	197	2	0.00
95	279	197	0	0.00
100	279	197	0	0.00

จากตารางที่ 18 ตารางที่ 19 ตารางที่ 20 และตารางที่ 20 พบว่าเอกสารกลุ่มที่ 3 ที่ส่งเข้าไปทดสอบด้วยขั้นตอนวิธีที่ออกแบบมาสามารถแยกออกจากกลุ่มที่ 1 และ 2 ได้ถูกต้องที่สุดที่เปอร์เซนไทร์ที่ 45 ด้วยวิธี Euclidean, Manhattan Minkowski และ Cosine ในส่วนค่า Recall สำหรับการดึงเอกสารที่เกี่ยวข้องออกมาพบว่าวิธีวัดระยะทางของเอกสารที่ดีที่สุดคือ Euclidean (92%) รองลงมาคือ Manhattan (84%) Minkowski (33%) และ Cosine (9%)

ตารางที่ 21 ตารางเปรียบเทียบค่า Recall ของวิธีการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi-Domain Sentiment dataset

Percentile	Recall			
	euclidean	manhattan	minkowski	cosine
45	**0.9243697**	**0.842437**	0.331933	0.09
50	0.894958	0.831933	0.289916	0.07
55	0.8613445	0.794118	0.262605	0.06
60	0.8193277	0.756303	0.222689	0.05
65	0.7836134	0.731092	0.189076	0.04
70	0.7436975	0.691176	0.170168	0.02
75	0.6617647	0.638655	0.138655	0.02
80	0.5987395	0.581933	0.113445	0.01
85	0.4789916	0.512605	0.092437	0

Percentile	Recall			
	euclidean	manhattan	minkowski	cosine
90	0.2983193	0.392857	0.060924	0
95	0.1470588	0.283613	0.042017	0
100	0.0042017	0.079832	0	0



รูปที่ 27 แสดงการเปรียบเทียบค่า Recall ของวิธีวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi-Domain Sentiment dataset

4.5.3 ผลการทดลองส่งเอกสารอีกโดเมนหนึ่งเข้าไปทดสอบบนชุดข้อมูล 20 Newsgroup text Dataset

ผู้วิจัยได้แยกเอกสารของกลุ่มที่ 3 ของชุดข้อมูล 20 Newsgroup ออกมาซึ่งจำนวน 476 เอกสารแล้วส่ง เข้าไปทดสอบว่าสามารถแยกออกจากกลุ่มได้ถูกต้องหรือไม่ โดยผู้วิจัยได้คำนวณค่า Threshold เพื่อจะใช้เป็นเกณฑ์ในการแบ่งกลุ่มของเอกสารโดยคำนวณค่าเปอร์เซ็นไทร์ของระยะทางของเอกสารกลุ่มที่ 1 และกลุ่มที่ 2 ที่ 45 50 55 60 65 70 75 80 85 90 95 และ 100 เช่นเดียวกับชุดข้อมูล Multi-Domain Sentiment dataset แล้วส่งเอกสารกลุ่มที่ 3 เข้าไปทดสอบตามขั้นตอนวิธีที่ผู้วิจัยนำเสนอ โดยใช้วิธีวัดความคล้ายคลึงของเอกสารด้วย Euclidean Distance Manhattan Minkowski และ Cosine เพื่อทดสอบว่าวิธีวัดความคล้ายคลึงของเอกสารวิธีการใดให้ผลลัพธ์ที่ดีที่สุด ผลการทดลองดังตารางด้านล่าง

ตารางที่ 22 แสดงผลการทำนายของเอกสาร (Test Documents) ที่ส่งเข้าไปทดสอบ ด้วย Euclidean Distance บนชุดข้อมูล 20 Newsgroup

Percentile (Threshold)	Prediction with Euclidean distance			Recall
	Cluster1	Cluster2	New Group	
45	**45**	**1**	**859**	**0.95**
50	45	2	858	0.95
55	46	2	857	0.95
60	46	3	856	0.95
65	47	5	853	0.94
70	48	9	848	0.94
75	51	16	838	0.93
80	55	25	825	0.91
85	61	37	807	0.89
90	74	60	771	0.85
95	90	104	711	0.79
100	245	280	380	0.42

ตารางที่ 23 แสดงผลการทำนายเอกสารกลุ่มที่ 3 (Test Documents) ด้วย Manhattan Distance บนชุดข้อมูล 20 Newsgroup

Percentile (Threshold)	Prediction with Manhattan distance			Recall
	Cluster1	Cluster2	New Group	
45	**367**	**0**	**538**	**0.59**
50	407	2	496	0.55
55	441	2	462	0.51
60	472	2	431	0.48
65	512	2	391	0.43
70	588	2	315	0.35
75	588	2	315	0.35
80	655	2	248	0.27
85	699	3	203	0.22
90	762	3	140	0.15
95	804	3	98	0.11
100	887	3	15	0.02

ตารางที่ 24 แสดงผลการทำนายเอกสารกลุ่มที่ 3 (Test Documents) ด้วย Minkowski Distance บนชุดข้อมูล 20 Newsgroup

Percentile (Threshold)	Prediction with Minkowski distance			Recall
	Cluster1	Cluster2	New Group	
45	**109**	**133**	**663**	**0.73**
50	133	152	620	0.69
55	154	167	584	0.65
60	180	185	540	0.60
65	206	198	501	0.55
70	239	218	448	0.50
75	263	251	391	0.43
80	301	266	338	0.37
85	337	291	277	0.31
90	376	321	208	0.23
95	399	358	148	0.16
100	461	425	19	0.02

ตารางที่ 25 แสดงผลการทำนายเอกสารกลุ่มที่ 3 (Test Documents) ด้วย Cosine บนชุดข้อมูล 20 Newsgroup

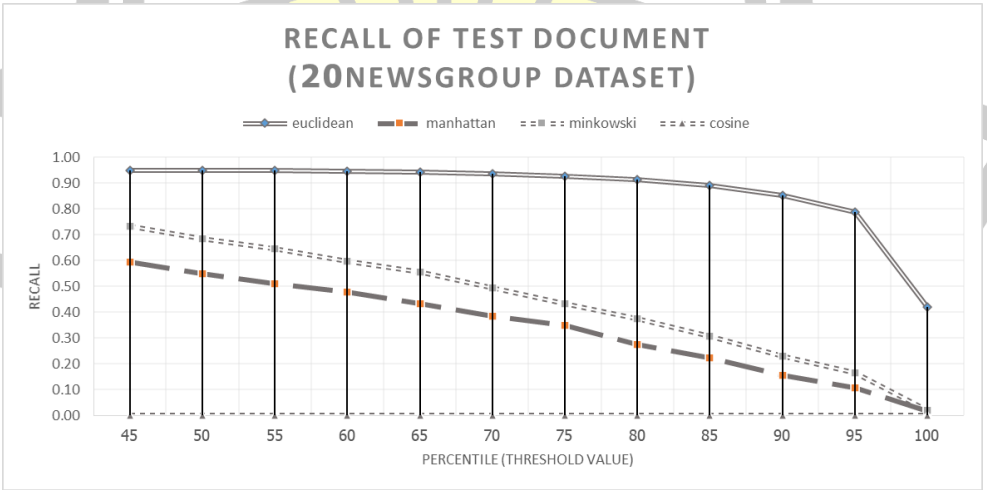
Percentile (Threshold)	Prediction with Cosine			Recall
	Cluster1	Cluster2	New Group	
45	493	410	2	0.00
50	493	411	1	0.00
55	494	411	0	0.00
60	494	411	0	0.00
65	494	411	0	0.00
70	494	411	0	0.00
75	494	411	0	0.00
80	494	411	0	0.00
85	494	411	0	0.00
90	494	411	0	0.00
95	494	411	0	0.00
100	494	411	0	0.00

จากตารางที่ 23 ตารางที่ 24 ตารางที่ 25 และตารางที่ 25 พบว่าเอกสารกลุ่มที่ 3 (Test Documents) ที่ส่งเข้าไปทดสอบด้วยขั้นตอนวิธีที่ออกแบบมาสามารถแยกออกจากกลุ่มที่มีอยู่เดิม ได้

ถูกต้องที่สุดที่เปอร์เซนไทร์ระหว่าง 45 – 60 ด้วยวิธี Euclidean, Manhattan และ Minkowski ส่วน Cosine ไม่สามารถแยกเอกสารได้ ในส่วนค่า Recall สำหรับการดึงเอกสารที่เกี่ยวข้องออกมา พบว่าวิธีวัดระยะทางของเอกสารที่ดีที่สุดคือ Euclidean (95%) รองลงมาคือ Minkowski (73%) Manhattan (59%) และ ในส่วนของ Cosine ไม่สามารถดึงเอกสารออกมาได้เลย

ตารางที่ 26 ตารางเปรียบเทียบค่า Recall ของวิธีการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group Text Dataset ทดลองด้วยเอกสารกลุ่มที่ 3 (Test Documents)

Percentile	Recall			
	euclidean	manhattan	minkowski	cosine
45	0.95	0.59448	0.7326	0
50	0.95	0.54807	0.68508	0
55	0.95	0.5105	0.6453	0
60	0.95	0.47624	0.59669	0
65	0.94	0.43204	0.55359	0
70	0.94	0.38453	0.49503	0
75	0.93	0.34807	0.43204	0
80	0.91	0.27403	0.37348	0
85	0.89	0.22431	0.30608	0
90	0.85	0.1547	0.22983	0
95	0.79	0.10829	0.16354	0
100	0.42	0.01657	0.02099	0



รูปที่ 28 แสดงการเปรียบเทียบค่า Recall ของวิธีวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group

4.5.4 ผลการทดลองด้วยเอกสารทั้งหมดของทุกกลุ่ม

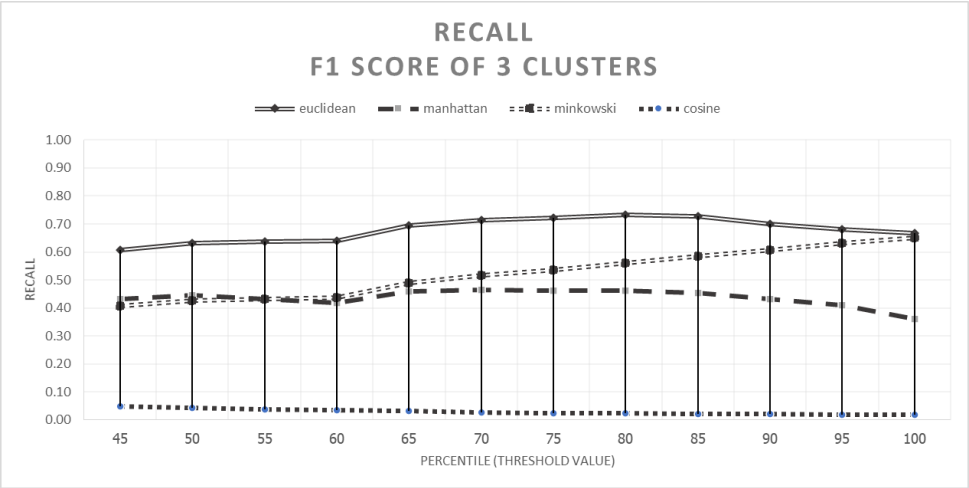
เพื่อเป็นการตรวจสอบขั้นตอนวิธีของผู้วิจัย ผู้วิจัยได้ส่งเอกสารของชุดข้อมูล Multi-Domain Sentiment dataset และ 20 News Group Text Dataset กลุ่มที่ 1 2 และ 3 ตามลำดับเข้าไปทดสอบว่าทุกกลุ่มสามารถจัดกลุ่มเข้ากลุ่มเดิมและแยกเอกสารออกจากกลุ่มได้ถูกต้องหรือไม่ โดยเปรียบเทียบด้วยวิธีวัดความคล้ายคลึงด้วยวิธี Euclidean Manhattan Minkowski และ Cosine ผลการทดลองเป็นดังนี้

1. ผลการทดลองบนชุดข้อมูล Multi Domain Sentiment Dataset

ตารางที่ 27 แสดงค่า Average Recall ของเอกสารทุกกลุ่มด้วยการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi Domain Sentiment

Percentile	Multi domain Sentiment Dataset Recall			
	euclidean	manhattan	minkowski	cosine
45	0.61	0.43	0.41	**0.05**
50	0.63	0.44	0.43	0.04
55	0.64	0.43	0.43	0.04
60	0.64	0.42	0.44	0.04
65	0.69	**0.46**	0.49	0.03
70	0.71	**0.46**	0.52	0.03
75	0.72	**0.46**	0.54	0.02
80	0.73	**0.46**	0.56	0.02
85	**0.73**	0.45	**0.58**	0.02
90	0.70	0.43	0.61	0.02
95	0.68	0.41	0.63	0.02
100	0.67	0.36	0.65	0.02

จากตารางที่ 27 ผลการทดลองพบว่าเมื่อส่งเอกสารทั้งสามกลุ่มเพื่อพิจารณาว่าสามารถจัดเข้ากลุ่มเดิมหรือแยกเอกสารออกจากกลุ่มได้ถูกต้องหรือไม่ พบว่า ค่า Recall ของทุกเอกสารของการวัดความคล้ายคลึง Euclidean Distance ทั้งสามกลุ่มได้ดีที่สุด 73 % ที่เปอร์เซนไทร์ที่ 85 แต่เมื่อพิจารณาจากผลการทดลองก่อนหน้าจะพบว่าค่า Recall ของเอกสารกลุ่มที่ 3 ซึ่งจะต้องแยกออกมาจากกลุ่มที่เปอร์เซนไทร์ที่ 85 ทำได้ถูกต้อง 85% ด้วยวิธีการวัดความคล้ายคลึงของเอกสารด้วย Euclidean Distance เช่นเดียวกัน



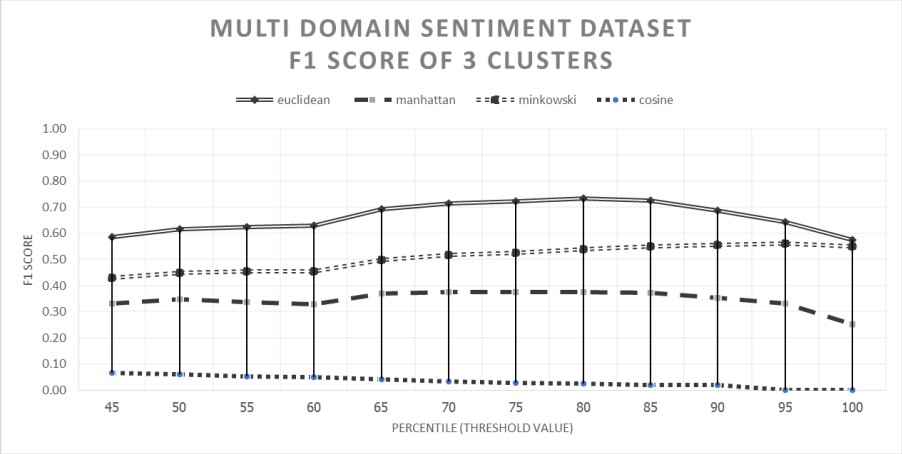
รูปที่ 29 แสดงแผนภูมิเปรียบเทียบค่า Recall ของขั้นตอนที่นำเสนอด้วยวิธีการวัดความคล้ายคลึงแบบต่าง ๆ บนเอกสารทั้ง 3 กลุ่ม

นอกจากผู้วิจัยจะพิจารณาจากค่า Recall แล้ว ผู้วิจัยยังพิจารณาค่า F1 ของการวัดความคล้ายคลึงของเอกสารทั้ง 3 กลุ่มด้วยวิธีการวัดแบบต่าง ๆ ด้วยผลการทดลองเป็นดังนี้

ตารางที่ 28 แสดงค่า F1 ของเอกสารทุกกลุ่มด้วยการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi Domain Sentiment

Percentile	F1 of Multi domain Sentiment Dataset			
	euclidean	manhattan	minkowski	cosine
45	0.58	0.33	0.43	**0.07**
50	0.62	0.35	0.45	0.06
55	0.62	0.34	0.45	0.05
60	0.63	0.33	0.46	0.05
65	0.69	0.37	0.50	0.04
70	0.71	0.37	0.52	0.03
75	0.72	0.38	0.53	0.03
80	**0.74**	**0.38**	0.54	0.02
85	0.73	0.37	0.55	0.02
90	0.69	0.35	0.55	0.02
95	0.64	0.33	**0.56**	0.00
100	0.58	0.25	0.55	0.00

จากตารางที่ 28 ผลการทดลองพบว่าเมื่อส่งเอกสารทั้งสามกลุ่มเพื่อพิจารณาว่าสามารถจัดเข้ากลุ่มเดิมหรือแยกเอกสารออกจากกลุ่มได้ถูกต้องหรือไม่ พบว่า ค่า F1 ของทุกเอกสารของการวัดความคล้ายคลึง Euclidean Distance ทั้งสามกลุ่มได้ดีที่สุด 74 % ที่เปอร์เซนไทร์ที่ 80



รูปที่ 30 แสดงการเปรียบเทียบค่า F1 ของวิธีวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล Multi Domain Dataset ของเอกสารทั้ง 3 กลุ่ม

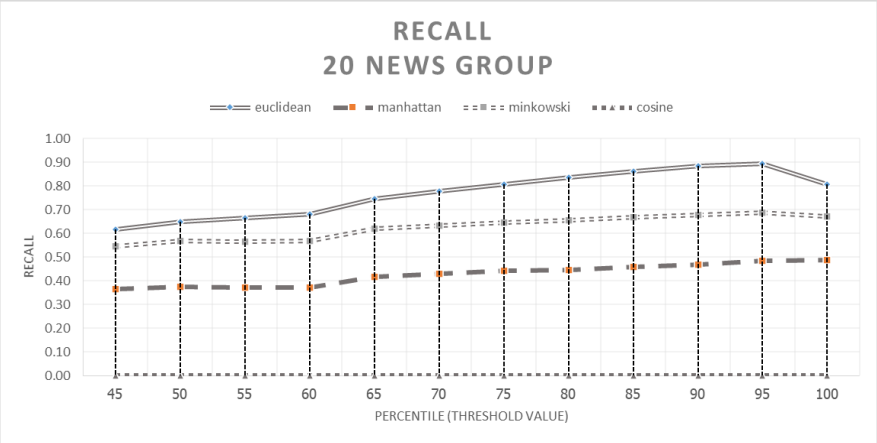
2. ผลการทดลองบนชุดข้อมูล 20 News Group Dataset

ตารางที่ 29 แสดงค่า Average Recall ของเอกสารทุกกลุ่มด้วยการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group Dataset

Percentile	20 News Group Dataset Recall			
	euclidean	manhattan	minkowski	cosine
45	0.62	0.36	0.55	0
50	0.65	0.38	0.57	0
55	0.67	0.37	0.57	0
60	0.68	0.37	0.57	0
65	0.75	0.42	0.62	0
70	0.78	0.43	0.63	0
75	0.81	0.44	0.65	0
80	0.84	0.45	0.66	0
85	0.86	0.46	0.67	0
90	0.88	0.47	0.68	0
95	**0.89**	**0.49**	**0.69**	0

Percentile	20 News Group Dataset Recall			
	euclidean	manhattan	minkowski	cosine
100	0.81	0.49	0.67	0

จากตารางที่ 29 ผลการทดลองพบว่าเมื่อส่งเอกสารทั้งสามกลุ่มเพื่อพิจารณาว่าสามารถจัดเข้ากลุ่มเดิมหรือแยกเอกสารออกจากกลุ่มได้ถูกต้องหรือไม่ พบว่า ค่า Recall ของทุกเอกสารของการวัดความคล้ายคลึง Euclidean Distance ทั้งสามกลุ่มได้ดีที่สุด 89 % ที่เปอร์เซนไทร์ที่ 95 แต่เมื่อพิจารณาจากผลการทดลองก่อนหน้าจะพบว่าค่า Recall ของเอกสารกลุ่มที่ 3 ซึ่งจะต้องแยกออกมาจากกลุ่มที่เปอร์เซนไทร์ที่ 95 สามารถดึงเอกสารที่อยู่ในกลุ่มได้ 79% ด้วยวิธีการวัดความคล้ายคลึงของเอกสารด้วย Euclidean Distance เช่นเดียวกัน



รูปที่ 31 แสดงแผนภูมิเปรียบเทียบค่า Recall ของขั้นตอนที่นำเสนอด้วยวิธีการวัดความคล้ายคลึงแบบต่าง ๆ บนเอกสารทั้ง 3 กลุ่ม บนชุดข้อมูล 20 News Group Dataset

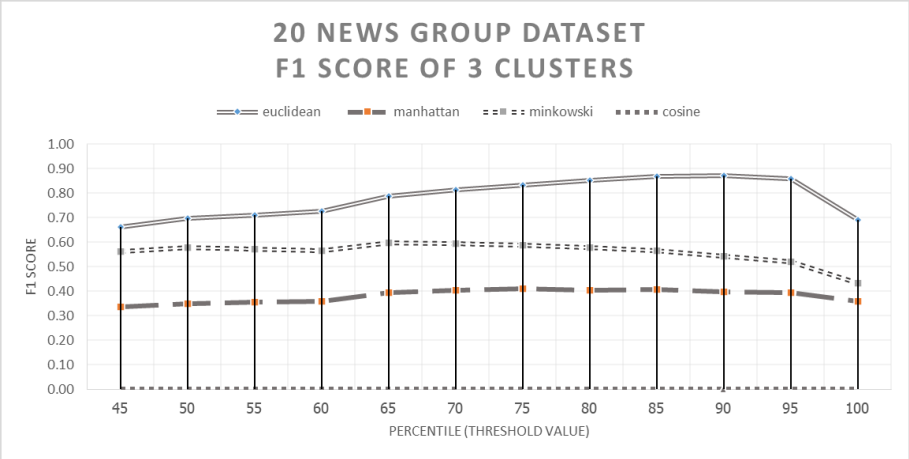
นอกจากผู้วิจัยจะพิจารณาจากค่า Recall แล้ว ผู้วิจัยยังพิจารณาค่า F1 ของการวัดความคล้ายคลึงของเอกสารทั้ง 3 กลุ่มด้วยวิธีการวัดแบบต่าง ๆ ด้วยผลการทดลองเป็นดังนี้

ตารางที่ 30 แสดงค่า F1 ของเอกสารทุกกลุ่มด้วยการวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group Text Dataset

Percentile	F1 Score : 20 News Group Dataset			
	euclidean	manhattan	minkowski	cosine
45	0.66	0.34	0.56	0
50	0.70	0.35	0.58	0

55	0.71	0.35	0.57	0
60	0.73	0.36	0.56	0
65	0.79	0.39	0.60	0
70	0.81	0.40	0.59	0
75	0.83	0.41	0.59	0
80	0.85	0.40	**0.58**	0
85	**0.87**	**0.41**	0.56	0
90	0.87	0.40	0.54	0
95	0.86	0.40	0.52	0
100	0.69	0.36	0.43	0

จากตารางที่ 30 ผลการทดลองพบว่าเมื่อส่งเอกสารทั้งสามกลุ่มเพื่อพิจารณาว่าสามารถจัดเข้ากลุ่มเดิมหรือแยกเอกสารออกจากกลุ่มได้ถูกต้องหรือไม่ พบว่า ค่า F1 ของทุกเอกสารของการวัดความคล้ายคลึงด้วย Euclidean Distance ทั้งสามกลุ่มได้ดีที่สุด 87% ที่เปอร์เซนไทร์ที่ 85 และ 90



รูปที่ 32 แสดงการเปรียบเทียบค่า F1 ของวิธีวัดความคล้ายคลึงด้วยวิธีต่าง ๆ บนชุดข้อมูล 20 News Group Text Dataset

ผลการทดลองพบว่าเมื่อส่งเอกสารทั้งสามกลุ่มของทั้งสองโดเมนเพื่อพิจารณาว่าสามารถจัดเข้ากลุ่มเดิมหรือแยกเอกสารออกจากกลุ่มได้ถูกต้องหรือไม่

พบว่าบนชุดข้อมูล Multi Domain Sentiment Dataset พบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอดีที่สุดจากการวัดด้วย F1 Score เป็น 74% ที่เปอร์เซนไทร์ที่ 80

สำหรับชุดข้อมูล 20 News Group Dataset พบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอดีที่สุดจากการวัดด้วย F1 Score เป็น 87% ที่เปอร์เซนไทร์ที่ 85

4.6 เปรียบเทียบการจัดกลุ่มเอกสารที่ส่งเข้าไปทดสอบด้วยขั้นตอนวิธีที่นำเสนอกับวิธีการเดิม

ในการทดลองนี้มีวัตถุประสงค์เพื่อทดสอบว่าหลังจากจัดกลุ่มเอกสารแล้วเมื่อมีเอกสารใหม่เข้าไป ขั้นตอนวิธีที่นำเสนอมีผลลัพธ์แตกต่างกันอย่างไรกับการหาความคล้ายคลึงของเอกสารด้วยวิธี การแบบเดิม

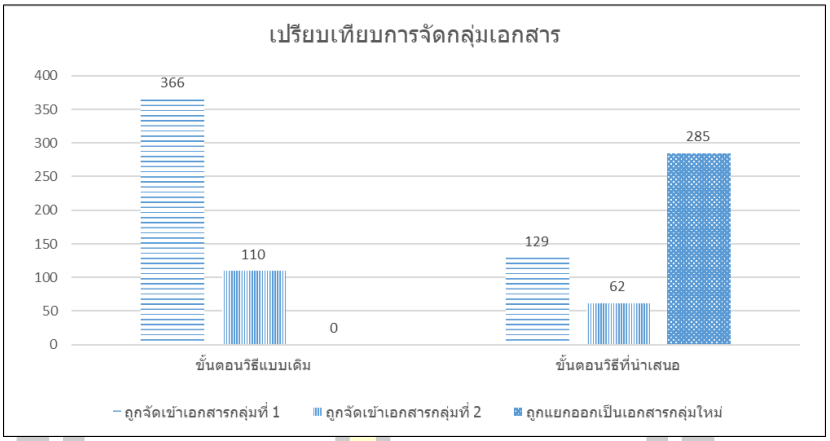
ซึ่งวิธีการแบบเดิมนั้นการหาความคล้ายคลึงเพื่อจัดกลุ่มของเอกสารที่ส่งเข้าไปใหม่นั้นจะทำการวัดระยะทางของข้อมูลใหม่กับจุดศูนย์กลางของกลุ่มถ้าใกล้จุดศูนย์กลางของกลุ่มใดมากที่สุดจะกำหนดให้ข้อมูลอยู่ในกลุ่มนั้น ซึ่งการวัดระยะทางนั้นอาจจะใช้การวัดด้วย Euclidean Distance Cosine Similarity Manhattan Distance หรือ Minkowki Distance ก็ได้ขึ้นอยู่กับลักษณะของข้อมูลนั้น ๆ

สำหรับขั้นตอนวิธีที่นำเสนอนั้นจะทำการวัดระยะทางของข้อมูลที่เข้าไปใหม่กับจุดศูนย์กลางของกลุ่มถ้าใกล้จุดศูนย์กลางของกลุ่มใดมากที่สุด และจะนำระยะทางของข้อมูลใหม่ไปเปรียบเทียบกับค่า Threshold ของกลุ่มนั้นก่อนถ้าระยะทางของข้อมูลจากจุดศูนย์กลางของกลุ่มนั้นน้อยกว่าค่า Threshold ของกลุ่มจะกำหนดให้ข้อมูลอยู่ในกลุ่มนั้น แต่ถ้ามากกว่าค่า Threshold ของกลุ่มนั้นจะถูกกำหนดให้ออกมาอยู่กลุ่มใหม่

ในการทดลองนี้ผู้วิจัยได้ใช้วิธีการกำหนดค่า Threshold ที่ตำแหน่งเปอเซนไทล์ที่ 80 และใช้วิธีการวัดระยะทางด้วย Euclidean Distance เนื่องจากผลการทดลองก่อนหน้าได้ให้ผลลัพธ์ที่ดีกว่าวิธีการอื่น ผู้วิจัยส่งเอกสารกลุ่มที่ 3 จำนวน 476 เอกสารบนชุดข้อมูล Multi Domain Sentiment Dataset เข้าไปทดสอบการจัดกลุ่ม โดยมีกลุ่มเดิมอยู่สองกลุ่ม ผลลัพธ์แสดงดังตาราง

ตารางที่ 31 เปรียบเทียบการจัดกลุ่มเอกสารใหม่ด้วยขั้นตอนวิธีเดิมกับขั้นตอนวิธีที่นำเสนอ

การจัดกลุ่มเอกสาร	**ขั้นตอนวิธีแบบเดิม**	**ขั้นตอนวิธีที่นำเสนอ**
ถูกจัดเข้าเอกสารกลุ่ม 1	366	129
ถูกจัดเข้าเอกสารกลุ่ม 2	110	62
ถูกแยกออกเป็นเอกสารกลุ่มใหม่	**0**	**285**



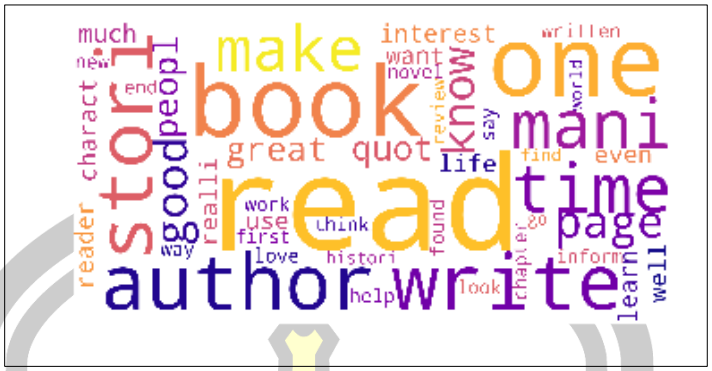
รูปที่ 33 แผนภูมิเปรียบเทียบการจัดกลุ่มเอกสารใหม่ด้วยขั้นตอนวิธีเดิมกับขั้นตอนวิธีที่นำเสนอ

จากผลการทดลองจะเห็นถึงความแตกต่างของการจัดกลุ่มเอกสารด้วยขั้นตอนที่นำเสนอ เมื่อส่งเอกสารกลุ่มใหม่เข้าไปจำนวน 476 เอกสารจะพบว่าสามารถแยกเอกสารออกจากกลุ่มเดิมที่มีอยู่แล้วทั้งสองกลุ่มได้จำนวน 285 เอกสาร และเอกสารจำนวน 221 จะถูกจัดเข้าไปอยู่ในทั้งสองกลุ่ม ซึ่งผลของการทดลองทำให้เห็นว่าการจัดกลุ่มของเอกสารด้วยขั้นตอนวิธีที่นำเสนอนั้นสามารถแยกเอกสารออกมาได้เป็นการแก้ปัญหาของการจัดกลุ่มด้วย K-Means ซึ่งจะได้จำนวนกลุ่มตามที่กำหนดและข้อมูลในกลุ่มอาจจะไม่สมควรอยู่ในกลุ่มนั้น ขั้นตอนวิธีที่นำเสนอจะพิจารณาเอาข้อมูลที่มีความคล้ายคลึงในกลุ่มน้อยที่สุดของแต่ละกลุ่มมารวมในกลุ่มในกลุ่มเดียวกันเพื่อพิจารณาข้อมูลในกลุ่มใหม่อีกที

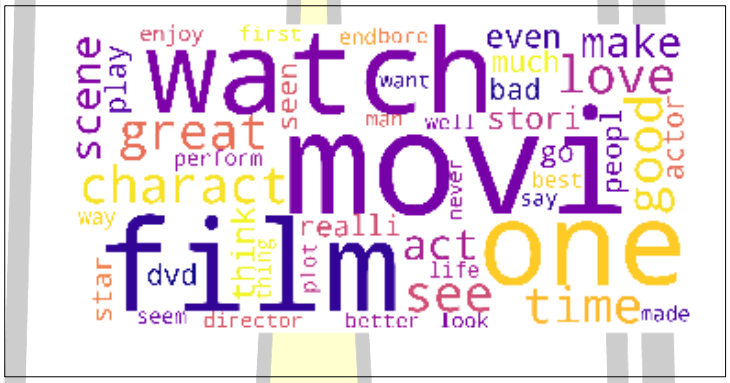
4.7 แสดงคุณลักษณะข้อมูลที่อยู่ในแต่ละกลุ่ม

เนื่องจากข้อมูลที่อยู่ในแต่กลุ่มมีคุณลักษณะเป็นคำ ผู้วิจัยได้แสดงคำที่มีความสำคัญสูงสุดในแต่ละกลุ่ม กลุ่มละ 50 คำ เพื่อแสดงให้เป็นว่าในแต่ละกลุ่มมีคำใดเป็นคุณลักษณะที่สำคัญในการพิจารณาจัดกลุ่มดังรูป

รูปที่ 34 แสดงคุณลักษณะที่สำคัญของเอกสารกลุ่มที่ 1 50 ลำดับแรก



รูปที่ 35 แสดงคุณลักษณะที่สำคัญของเอกสารกลุ่มที่ 2 50 ลำดับแรก



รูปที่ 36 แสดงคุณลักษณะที่สำคัญของเอกสารกลุ่มใหม่ 50 ลำดับแรก

จะเห็นได้ว่า คำที่อยู่ในกลุ่มใหม่ที่ถูกคัดแยกออกมาไม่พบในทั้งสองกลุ่มเดิม (Unseen Data) หรือพบได้น้อย แสดงถึงความคล้ายคลึงของเอกสารต่ำ จะพบว่ามีบางคุณลักษณะที่เหมือนกับในกลุ่มที่มีอยู่แต่ก็ไม่เพียงพอต่อการถูกจัดเข้าไปอยู่ในกลุ่มนั้น เนื่องจากมีคุณลักษณะที่สำคัญอื่นปรากฏอยู่มากกว่า

จากการทดลองที่ผู้วิจัยนำเสนอสามารถสรุปผลการทดลองได้ดังนี้

1. การทดลองกำหนดพารามิเตอร์ที่ใช้ในการคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำของ TF-IDF และ BM25 เพื่อหาค่าพารามิเตอร์ที่ให้ผลลัพธ์ดีที่สุดในการจัดกลุ่มของทั้ง 2 ชุดข้อมูล ผลการทดลองพบว่า

   1.1 การให้ค่าน้ำหนักของคำของ TF-IDF บนชุดข้อมูล Multi Domain การกำหนดพารามิเตอร์ Min DF ที่ให้ค่า Recall และ F1 ดีที่สุดคือ 2 โดยให้ค่า Recall เป็น 88% Precision เป็น 90% และค่า F1 เป็น 88% บนชุดข้อมูล 20 News Group

การกำหนดพารามิเตอร์ Min DF ที่ให้ค่า Recall และ Accuracy ดีที่สุดคือ 8 โดยให้ค่า Recall เป็น 81% ค่า Precision 87% และค่า F1 เป็น 82%

1.2 การให้ค่าน้ำหนักของคำด้วย BM25 บนชุดข้อมูล Multi Domain การกำหนดพารามิเตอร์ที่ให้ค่า Recall Precision และ F1 ดีที่สุดคือ k1 = 1.9 และ b = 1.2 โดยให้ค่าเฉลี่ย Recall เป็น 88% Precision 89% และค่า F1 เป็น 88% ผลการทดลองบนชุดข้อมูล 20 News Group การกำหนดพารามิเตอร์ที่ให้ผลลัพธ์ดีที่สุดคือ k1 = 1.5 และ b = 0.8 โดยให้ค่า Recall เป็น 61% Precision 76% และค่า F1 เป็น 61 %

2. การเปรียบเทียบวิธีคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำระหว่าง TF-IDF และ BM25 เพื่อหาประสิทธิภาพของวิธีคัดเลือกคุณลักษณะว่าวิธีใดให้ผลลัพธ์ดีที่สุด พบว่าการให้ค่าน้ำหนักด้วย TF-IDF มีประสิทธิภาพดีกว่าการให้ค่าน้ำหนักด้วย BM25 บนชุดข้อมูล 20 News Group แต่สำหรับชุดข้อมูล Multi Domain เท่ากัน
3. ทดสอบสมมติฐานของการให้ค่าน้ำหนักของคำระหว่าง TF-IDF และ BM25 ด้วย Paired – sample T-test เพื่อยืนยันประสิทธิภาพของวิธีคัดเลือกคุณลักษณะด้วยการให้ค่าน้ำหนักของคำผลการทดลองพบว่า ให้ค่าน้ำหนักกับคุณลักษณะของคำด้วย TF-IDF มีประสิทธิภาพของการจัดกลุ่มดีกว่า BM25 อย่างมีนัยสำคัญที่ 0.05
4. เปรียบเทียบประสิทธิภาพขั้นตอนวิธีการจัดกลุ่มระหว่าง K-Means และ DBSCAN เพื่อหาประสิทธิภาพของขั้นตอนวิธีจัดกลุ่มว่าวิธีใดให้ผลลัพธ์ดีที่สุด ผลการทดลองพบว่าขั้นตอนวิธี K-Means ความถูกต้อง (Accuracy) ของการจัดกลุ่มได้ 88.33% ในส่วน DBSCAN ได้ 38.5% โดยจำนวนกลุ่มที่ได้เท่ากันคือ 3 กลุ่ม
5. การตรวจสอบจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่มเอกสารด้วย Elbow method สามารถกำหนดจำนวนกลุ่มได้ถูกต้องตามจำนวนโดเมนเอกสารที่ส่งเข้าไปหรือไม่ ผลการทดสอบพบว่า Elbow method สามารถแสดงจำนวนกลุ่มที่เหมาะสมได้ถูกต้องตามจำนวนโดเมนที่ส่งเข้าไปทดสอบ
6. การกำหนดค่า Threshold ด้วยค่าต่าง ๆ เพื่อใช้ในการแยกเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบด้วยกำหนดเปอเซนไทล์ของระยะทางแต่ละข้อมูลในกลุ่มที่มีอยู่แล้วเป็นเท่าใดจะสามารถแยกเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบออกมาได้ดีที่สุด ผลการทดลองพบว่า

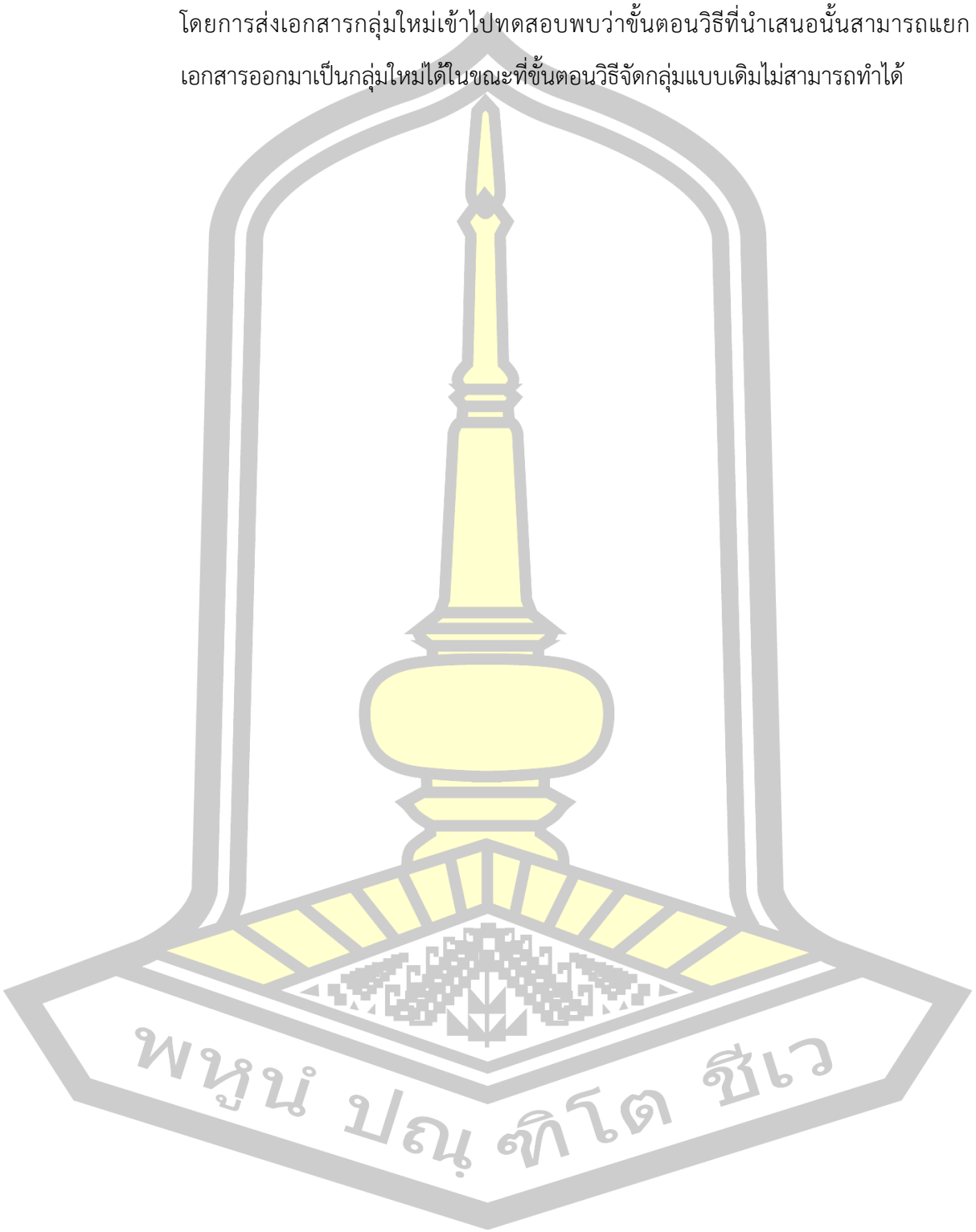
สำหรับเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบด้วยขั้นตอนวิธีที่ออกแบบมาสามารถแยกออกจากกลุ่มที่มีอยู่ได้ผลลัพธ์ดีที่สุดที่เปอเซนไทล์ที่ 45 บนทั้งสองชุดข้อมูล

7. การทดลองส่งเอกสารกลุ่มใหม่เข้าไปทดสอบ C- Algorithm เพื่อทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอและเปรียบเทียบการวัดความคล้ายคลึงของเอกสารด้วยวิธี Euclidean Manhattan Minkowski และ Cosine วิธีใดที่ให้ผลลัพธ์ของขั้นตอนวิธีที่นำเสนอดีที่สุดในการแยกเอกสารกลุ่มใหม่ออกจากกลุ่มเดิม พบว่าวิธีวัดวัดความคล้ายคลึงของเอกสารที่ดีที่สุดคือ Euclidean (95%) รองลงมาคือ Minkowski (73%) Manhattan (59%) และ ในส่วนของ Cosine ไม่สามารถดึงเอกสารออกมาได้เลย
8. การทดลองกำหนดค่า Threshold ด้วยค่าต่าง ๆ เพื่อใช้ในการทดสอบการรวมกลุ่มของเอกสารกลุ่มเดิมและแยกออกจากกลุ่มของเอกสารกลุ่มใหม่การกำหนดตำแหน่งเปอเซนไทล์ของระยะทางแต่ละข้อมูลในกลุ่มที่มีอยู่แล้วเป็นเท่าใดจะสามารถแยกเอกสารกลุ่มใหม่ที่ส่งเข้าไปทดสอบและจัดเอกสารกลุ่มเดิมเข้าไปในกลุ่มได้ดีที่สุดการพบว่าการกำหนดค่า Threshold สำหรับเอกสารทุกกลุ่มส่งเข้าไปทดสอบด้วยขั้นตอนวิธีที่ออกแบบมาสามารถแยกออกจากกลุ่มและรวมเข้ากับกลุ่มเดิมโดยการกำหนดตำแหน่งเปอเซนไทล์อยู่ระหว่าง 80-85 จะมีประสิทธิภาพโดยรวม (F1 Score) ดีที่สุดบนทั้ง 2 ชุดข้อมูล
9. ส่งเอกสารทุกกลุ่มเข้าไปทดสอบ C- Algorithm และการวัดความคล้ายคลึงของเอกสารด้วยวิธีต่าง ๆ เพื่อทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอและเปรียบเทียบการวัดความคล้ายคลึงของเอกสารด้วยวิธี Euclidean Manhattan Minkowski และ Cosine วิธีใดที่ให้ผลลัพธ์ของขั้นตอนวิธีที่นำเสนอดีที่สุดในการแยกเอกสารและจัดเอกสารเข้ากลุ่ม วิธีวัดความคล้ายคลึงของเอกสารที่ให้ประสิทธิภาพโดยรวม (F1-Score) ของขั้นตอนวิธีที่นำเสนอดีที่สุดในการแยกเอกสารและจัดเอกสารเข้ากลุ่ม

   9.1 บนชุดข้อมูล Multi Domain Dataset โดยใช้ Euclidean (74%) รองลงมาคือ Minkowski (56%) Manhattan (38%) และ ในส่วนของ Cosine ไม่สามารถดึงเอกสารออกและรวมกลุ่มเอกสารในขั้นตอนวิธีที่นำเสนอได้

   9.2 บนชุดข้อมูล 20Newsgroup โดยใช้ Euclidean (87%) รองลงมาคือ Minkowski (58%) Manhattan (41%) และ ในส่วนของ Cosine ไม่สามารถดึงเอกสารออกและรวมกลุ่มเอกสารในขั้นตอนวิธีที่นำเสนอได้

10. เปรียบเทียบการจัดกลุ่มเอกสารที่ส่งเข้าไปทดสอบด้วยขั้นตอนวิธีที่นำเสนอกับวิธีการเดิม โดยการส่งเอกสารกลุ่มใหม่เข้าไปทดสอบพบว่าขั้นตอนวิธีที่นำเสนอนั้นสามารถแยกเอกสารออกมาเป็นกลุ่มใหม่ได้ในขณะที่ขั้นตอนวิธีจัดกลุ่มแบบเดิมไม่สามารถทำได้



บทที่ 5
สรุปผลการวิจัย

งานวิจัยนี้ผู้วิจัยมีเป้าหมายเพื่อพัฒนาขั้นตอนวิธีการจัดกลุ่มของโดเมนที่แตกต่างกันตามความคล้ายคลึงกันของโดเมนเพื่อบ่งบอกว่าข้อมูลที่นำไปจัดกลุ่มนั้นควรจะอยู่ในกลุ่มของโดเมนนั้นหรือแยกออกมาสร้างกลุ่มของโดเมนใหม่ โดยมีวัตถุประสงค์การวิจัยคือ

1. เพื่อพัฒนาขั้นตอนวิธีการจัดกลุ่มเอกสารบนโดเมนที่แตกต่างกันตามความคล้ายคลึงกันของโดเมน และหาบ่งบอกได้เอกสารนั้นควรจะอยู่รวมในโดเมนนั้นหรือควรจะแยกออกมาสร้างกลุ่มของโดเมนใหม่
2. เพื่อเปรียบเทียบประสิทธิภาพการให้ค่าน้ำหนักของคำและการวัดความคล้ายคลึงของเอกสาร

5.1 สรุปและอภิปรายผล

1. ขั้นตอนวิธีที่นำเสนอสามารถแยกกลุ่มของเอกสารโดเมนใหม่ที่ส่งเข้าไปทดสอบออกจากกลุ่มเอกสารจากโดเมนอื่นได้ซึ่งสามารถแก้ปัญหาการจัดกลุ่มของขั้นตอนวิธี K-Means ที่จะทำการวัดความคล้ายคลึงของข้อมูลเมื่อพบว่าข้อมูลนั้นใกล้จุดศูนย์กลางของกลุ่มใดข้อมูลจะถูกจัดให้อยู่ในกลุ่มนั้นเช่นเดียวกับข้อมูลที่มีค่าปกติ ซึ่งข้อมูลนั้นอาจจะมีความเกี่ยวข้องกับข้อมูลที่อยู่ในกลุ่มน้อยมากหรือไม่มีความเกี่ยวข้องเลยก็ได้ ซึ่งจะทำให้ประสิทธิภาพในการจำแนกข้อมูลว่าอยู่ในกลุ่มใดมีความผิดพลาดได้ ดังนั้นการใช้ขั้นตอนวิธีที่นำเสนอสามารถแก้ปัญหาได้
2. การให้น้ำหนักของคุณลักษณะด้วย TF-IDF ให้ประสิทธิภาพดีกว่าการให้น้ำหนักของคุณลักษณะด้วย BM25 บนทั้งสองชุดข้อมูลอย่างมีนัยสำคัญที่ 0.05 ซึ่งสอดคล้องกับการทดลองของ Afrizal และคณะ [48] และ Kadhim [49] ที่พบว่าการให้น้ำหนักของคุณลักษณะด้วย TF-IDF มีประสิทธิภาพดีกว่า BM25
3. การเปรียบเทียบประสิทธิภาพของการจัดกลุ่มระหว่าง K-Means และ DBSCAN พบว่า K-Means มีประสิทธิภาพดีกว่าเมื่อพิจารณาจากจำนวนโดเมนที่ส่งเข้าไปทดสอบ เนื่องจาก K-Means มีการกำหนดจำนวนกลุ่มไว้ล่วงหน้าแล้วและสามารถตรวจสอบจำนวนกลุ่มที่เหมาะสมได้ด้วยเทคนิค Elbow ในขณะที่การจัดกลุ่มด้วยขั้นตอนวิธี DBSCAN เหมาะกับการจัดกลุ่มที่ไม่รู้จำนวนกลุ่มล่วงหน้าและเหมาะกับการตรวจสอบข้อมูลที่ผิดปกติมากกว่านอกจาก DBSCAN ยังมีความยุ่งยากในการกำหนด

ค่าพารามิเตอร์เพื่อให้ได้จำนวนกลุ่มตามที่ต้องการอีกด้วย เนื่องจากเมื่อ DBSCAN พบข้อมูลที่อยู่นอกความหนหนาแน่นและรัศมีที่กำหนดจะถูกมองเป็นอีกกลุ่มทันที

4. ผู้วิจัยนำแนวคิดการตรวจจับข้อมูลที่ผิดปกติมาใช้ในการจำแนกว่าเอกสารควรอยู่ในกลุ่มหรือสร้างกลุ่มเอกสารใหม่ ดังนั้นเอกสารกลุ่มใหม่หรือเอกสารกลุ่มเดิมที่ถูกย้ายไปอยู่กลุ่มใหม่ไม่ถือว่าเป็นค่าผิดปกติ เพียงแต่เป็นเอกสารที่มีความคล้ายคลึงของเอกสารในกลุ่มต่ำ
5. ประสิทธิภาพการวัดความคล้ายคลึงของเอกสารด้วยวิธีการวัดความคล้ายคลึงแบบต่าง ๆ ในขั้นตอนวิธีที่ผู้วิจัยนำเสนอพบว่าขั้นตอนวิธีที่ผู้วิจัยนำเสนอร่วมกับการวัดความคล้ายคลึงของเอกสารด้วย Euclidean Distance ให้ผลลัพธ์ดีที่สุด
6. การกำหนดว่าจะให้เอกสารที่ส่งเข้าไปใหม่จะอยู่ในกลุ่มหรือการแยกออกจากกลุ่มที่มีอยู่ก่อนหน้าคือการคำนวณระยะทางของเอกสารที่เข้าไปใหม่ไปยังจุดศูนย์กลางของทุกกลุ่ม ถ้าระยะทางของเอกสารใหม่ใกล้กับจุดศูนย์กลางของกลุ่มใดมากที่สุดแสดงว่าเอกสารควรจะอยู่กลุ่มนั้น แต่ต้องเปรียบเทียบกับ Threshold ของแต่ละกลุ่มด้วยถ้าระยะทางมากกว่าค่า Threshold ของกลุ่มนั้นแสดงว่าให้เอกสารที่ส่งเข้าไปใหม่นั้นออกจากกลุ่ม
7. ในการกำหนดค่า Threshold ที่สามารถแยกเอกสารใหม่ออกจากกลุ่มได้ดีที่สุดคือค่ากลางของระยะทางของเอกสารในกลุ่มนั้นกับจุดศูนย์กลางซึ่งสอดคล้องกับผลการทดลองของ Barai และ dey [4] แต่ในขณะเดียวกันถ้าส่งเอกสารกลุ่มเดิมเข้าไปทดสอบจะพบว่าผลของการรวมกลุ่มเอกสารเข้ากลุ่มเดิมทำได้ไม่ดี ดังนั้นวิธีการกำหนดค่า Threshold ที่ให้ประสิทธิภาพที่ดีที่สุดคือการกำหนดตำแหน่งเปอเซนไทล์ของชุดข้อมูลในกลุ่มโดยการกำหนดตำแหน่งเปอเซนไทล์อยู่ระหว่าง 80-85 จะให้ประสิทธิภาพโดยรวมของเอกสารทุกกลุ่มดีที่สุด
8. ประสิทธิภาพของขั้นตอนวิธีที่ทำเสนอนั้นสามารถแก้ปัญหาของขั้นตอนวิธีจัดกลุ่มด้วย K-means ได้
9. กลุ่มของเอกสารใหม่ที่ถูกแยกออกเป็นกลุ่มของเอกสารที่มีความคล้ายคลึงกับเอกสารในกลุ่มที่มีอยู่น้อยซึ่งหมายความว่าคำที่มีความสำคัญที่ปรากฏอยู่ในเอกสารกลุ่มใหม่นั้นไม่พบอยู่ในเอกสารกลุ่มเดิมหรือพบเพียงบางคำแต่น้อยมากจนไม่เพียงพอต่อการเข้าไปอยู่ในกลุ่มของเอกสารกลุ่มที่มีอยู่แล้ว เนื่องจากคำที่มีความสำคัญต่อการระบุเอกสารใหม่นั้นไม่ใช่คำที่มีความสำคัญต่อการจำแนกเอกสารกลุ่มเดิม
10. ประสิทธิภาพของการจัดกลุ่มขึ้นอยู่กับเลือกคุณลักษณะ ซึ่งงานวิจัยนี้ได้มีการทดลองเลือกคุณลักษณะของชุดข้อมูลด้วยการละเว้นคำที่ปรากฏในเอกสารน้อยกว่า 2 เอกสารในชุดข้อมูล Multi Domain Sentiment และละเว้นคำที่ปรากฏในเอกสารน้อยกว่า 8

เอกสารในชุดข้อมูล 20 News Group พบว่าให้ประสิทธิภาพสูงสุดของทั้งสองชุดข้อมูล ดังนั้นก่อนที่จะนำขั้นตอนวิธีที่นำเสนอนี้ไปใช้ควรมีการทดลองหาค่าที่เหมาะสมกับชุดข้อมูลตัวอย่างก่อนหรือใช้วิธีคัดเลือกคุณลักษณะที่เหมาะสมกับชุดข้อมูลที่นำมา

11. ก่อนการจัดกลุ่มด้วยขั้นตอนวิธีที่นำเสนอควรดูการกระจายของชุดข้อมูลก่อนเพื่อที่จะได้กำหนดค่า Threshold ได้เหมาะสม
12. กลุ่มเอกสารที่ถูกแยกออกมาเป็นกลุ่มใหม่ ไม่ได้หมายถึงเอกสารเหล่านั้นเป็นเอกสารที่มีความคล้ายคลึงกันภายในกลุ่ม แต่หมายถึงกลุ่มของเอกสารที่มีความคล้ายคลึงกับเอกสารในกลุ่มที่มีอยู่แล้วต่ำ ดังนั้นควรจะพิจารณาลักษณะของเอกสารในกลุ่มใหม่อีกทีหรืออาจจะใช้ขั้นตอนวิธีที่ผู้วิจัยนำเสนอเข้าไปใช้ในเอกสารกลุ่มใหม่นั้นก็ได้

5.2 ข้อเสนอแนะ

1. ขั้นตอนวิธีที่ได้นำเสนอ เป็นขั้นตอนวิธีการที่มีประสิทธิภาพเมื่อทดสอบกับข้อมูลเอกสารที่เป็นเอกสาร ผู้วิจัยเสนอแนะว่าควรนำแนวคิดไปทดสอบกับข้อมูลที่หลากหลาย เช่น ชุดข้อมูลที่เป็นรูปภาพ หรือชุดข้อมูลที่เป็นตัวเลข ซึ่งคาดว่าขั้นตอนวิธีที่นำเสนอนั้นจะมีประสิทธิในการรวมกลุ่มและแยกกลุ่มข้อมูลที่นำมาใช้งานได้
2. ผู้วิจัยใช้วิธีคัดเลือกคุณลักษณะด้วยการปรากฏขึ้นของคำในเอกสาร (Document Frequency) โดยการกำหนดคำที่เกิดขึ้นอย่างน้อยเป็นจำนวนกี่เอกสาร วิธีการคัดเลือกคุณลักษณะมีหลากหลายวิธีซึ่งอาจจะใช้วิธีที่แตกต่างจากผู้วิจัยก็ได้
3. ผู้วิจัยใช้วิธีหาค่า Threshold ของแต่ละกลุ่มด้วยวิธีการหาตำแหน่งเปอเซนไทล์ของชุดข้อมูลในกลุ่ม ซึ่งอาจจะใช้วิธีอื่นในการหาค่า Threshold แต่อย่างไรก็ตามค่า Threshold นั้นควรจะต้องเป็นค่าที่คำนวณจากข้อมูลในกลุ่มนั้น ๆ
4. ด้วยความหลากหลายของแหล่งข้อมูลทำให้ยากในการออกแบบการจำแนกหรือจัดกลุ่มเอกสารที่มีประสิทธิภาพ การแสดงความคิดเห็นเกี่ยวกับสินค้าหรือบริการที่หลากหลาย คำศัพท์ที่ใช้ก็มีความหมายที่แตกต่างกันด้วย ดังนั้นการใช้ข้อมูลจะแตกต่างกันเมื่ออยู่คนละโดเมนจะต้องมีการรู้ในความหมายของคำที่แตกต่างกันสำหรับแต่ละโดเมน อย่างไรก็ตามวิธีการนี้จะตามมาด้วยต้นทุนในการทั้งด้านเวลา คน เป็นจำนวนมากในการสร้างข้อมูลที่จะใช้สอนให้ได้ทุกโดเมน การทำให้ระบบสามารถที่จะค้นพบคำที่ใช้ร่วมกันระหว่างโดเมนนี้เรียกว่า การปรับตัวของโดเมน (Domain Adaptation) [61]

บรรณานุกรม

[1] Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence 2009; 1[1]: 60-76.

[2] Kaushik A, Naithani S. A comprehensive study of text mining approach. International Journal of Computer Science and Network Security (IJCSNS) 2016; 16[2]: 69.

[3] Navathe SB, Ramez E. Data warehousing and data mining. Fundamentals of Database Systems 2000; 841-872.

[4] Barai A, Dey L. Outlier detection and removal algorithm in k-means and hierarchical clustering. World Journal of Computer Application and Technology 2017; 5[2]: 24-29.

[5] Gulati H, Singh P. Clustering techniques in data mining: A comparison. 2015 2nd international conference on computing for sustainable global development (INDIACom); IEEE; 410-415.

[6] Xiong C, Hua Z, Lv K, Li X. An Improved K-means text clustering algorithm By Optimizing initial cluster centers. 2016 7th International Conference on Cloud Computing and Big Data (CCBD); IEEE; 265-268.

[7] Singh VK, Tiwari N, Garg S. Document clustering using k-means, heuristic k-means and fuzzy c-means. 2011 International Conference on Computational Intelligence and Communication Networks; IEEE; 297-301.

[8] Lomakina LS, Rodionov V, Surkova AS. Hierarchical clustering of text documents. Automation and Remote Control 2014; 75[7]: 1309-1315.

[9] Gaikwad SV, Chaugule A, Patil P. Text mining methods and techniques. International Journal of Computer Applications 2014; 85[17]:

[10] Sreedhar C, Kasiviswanath N, Reddy PC. Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop. Journal of Big Data 2017; 4[1]: 1-19.

[11] Bholowalia P, Kumar A. EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications

2014; 105[9]:

[12] Syakur M, Khotimah B, Rochman E, Satoto B. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering; IOP Publishing; 012017.

[13] Nainggolan R, Perangin-angin R, Simarmata E, Tarigan AF. Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. Journal of Physics: Conference Series; IOP Publishing; 012015.

[14] Usino W, Prabuwono A, Hamed K, Allehaibi S, Bramantoro A, Hasniaty A, et al. Document similarity detection using k-means and cosine distance. Intl J on Advanced Computer Science and Applications 2019; 10[2]: 165-170.

[15] Vijayarani S, Ilamathi MJ, Nithya M. Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks 2015; 5[1]: 7-16.

[16] Jusoh S, Alfawareh HM. Techniques, applications and challenging issue in text mining. International Journal of Computer Science Issues (IJCSI) 2012; 9[6]: 431.

[17] Kumar AA, Chandrasekhar S. Text data pre-processing and dimensionality reduction techniques for document clustering. International Journal of Engineering Research and Technology (IJERT) 2012; 1

[18] Camacho-Collados J, Pilehvar MT. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. arXiv preprint arXiv:170701780 2017;

[19] IŞIK M, DAĞ H. The impact of text preprocessing on the prediction of review ratings. Turkish Journal of Electrical Engineering & Computer Sciences 2020; 28[3]: 1405-1421.

[20] Selvam B, Abirami S. A survey on opinion mining framework. International Journal of Advanced Research in computer and communication Engineering 2013; 2[9]: 3544-3549.

[21] Mumu T. Social Network Opinion and posts mining for community preference discovery. 2013;

[22] Louwerse M, Lewis GA, Wu J. Unigrams, bigrams and LSA: Corpus linguistic

explorations of genres in Shakespeare's plays. *New directions in literary studies*: Cambridge Scholars Publishing 2008:108-129.

[23] Tripathy A, Agrawal A, Rath SK. Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications 2016; 57117-126.

[24] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. CS224N project report, Stanford 2009; 1[12]: 2009.

[25] Turney PD. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. arXiv preprint cs/0212032 2002;

[26] Porter MF. Snowball: A language for stemming algorithms. 2001.

[27] Salton G, Wong A, Yang C-S. A vector space model for automatic indexing. Communications of the ACM 1975; 18[11]: 613-620.

[28] พิมพาภรณ์ วย, สัจ พม. การ เปรียบเทียบ ประสิทธิภาพ การ จัด กลุ่ม ข้อมูล โดย วิธี การ เลือก ลักษณะ สำคัญ แบบ พลวัต เพื่อ เพิ่ม ประสิทธิภาพ ของ อั ลก อริ ทึม การ จัด กลุ่ม บน ปริภูมิ ย่อย. Information Technology Journal 2014; 10[2]: 43-51.

[29] Alsmadi I, Hoon GK. Term weighting scheme for short-text classification: Twitter corpuses. Neural Computing and Applications 2019; 31[8]: 3819-3831.

[30] Tsai FS, Kwee AT. Experiments in term weighting for novelty mining. Expert Systems with Applications 2011; 38[11]: 14094-14101.

[31] Ochikubo S, Komiya K, Saitoh F, Ishizu S. Evaluation of knowledge acquisition from document clustering based on information retrieval scales. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM); IEEE; 220-224.

[32] Wagstaff K, Cardie C, Rogers S, Schroedl S. Constrained k-means clustering with background knowledge. Icml; 577-584.

[33] Lakshmi R, Baskar S. Novel term weighting schemes for document representation based on ranking of terms and Fuzzy logic with semantic relationship of terms. Expert Systems with Applications 2019; 137493-503.

[34] Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd; 226-231.

[35] Yuan C, Yang H. Research on K-value selection method of K-means clustering algorithm. J—Multidisciplinary Scientific Journal 2019; 2[2]: 226-235.

[36] Huang A. Similarity measures for text document clustering. Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand; 9-56.

[37] Singh SK, Paul S, Kumar D, Arfi H. Sentiment analysis of twitter data set: survey. Int J Appl Eng Res 2014; 913925-13936.

[38] Balabantaray RC, Sarma C, Jha M. Document clustering using k-means and k-medoids. arXiv preprint arXiv:150207938 2015;

[39] Adebiyi MO, Adigun EB, Ogundokun RO, Adeniyi AE, Ayegba P, Oladipupo OO. Semantics-based clustering approach for similar research area detection. Telkomnika 2020; 18[4]: 1874-1883.

[40] Al-Anazi S, AlMahmoud H, Al-Turaiki I. Finding similar documents using different clustering techniques. Procedia Computer Science 2016; 8228-34.

[41] Fry C, Manna S. Can we group similar amazon reviews: A case study with different clustering algorithms. 2016 IEEE Tenth International Conference on Semantic Computing (ICSC); IEEE; 374-377.

[42] Chakraborty S, Nagwani NK, Dey L. Performance comparison of incremental k-means and incremental dbscan algorithms. arXiv preprint arXiv:14064751 2014;

[43] Christy A, Gandhi GM, Vaithyasubramanian S. Cluster based outlier detection algorithm for healthcare data. Procedia Computer Science 2015; 50209-215.

[44] Xiong H, Pandey G, Steinbach M, Kumar V. Enhancing data analysis with noise removal. IEEE Transactions on Knowledge and Data Engineering 2006; 18[3]: 304-319.

[45] Askari B, Hashemi S. Text Document Clustering by Using Semi-supervised Learning and Outlier Detection.

[46] Yu Q, Luo Y, Chen C, Ding X. Outlier-eliminated k-means clustering algorithm based on differential privacy preservation. Applied Intelligence 2016; 45[4]: 1179-1191.

[47] Sitaula C. Semantic text clustering using enhanced vector space model using nepali language. Computer Sciences and Telecommunications 2012; [4]: 41-46.

[48] Afrizal AD, Rakhmawati NA, Tjahyanto A. New Filtering Scheme Based on Term Weighting to Improve Object Based Opinion Mining on Tourism Product Reviews.

Procedia Computer Science 2019; 161805-812.

[49] Kadhim AI. Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF. 2019 International Conference on Advanced Science and Engineering (ICOASE); IEEE; 124-128.

[50] Arzoo MK, Prof A, Rathod K. K-Means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8. 2. Int Res J Eng Technol 2017; 4[4]: 2363-2368.

[51] Salihu SA, Onyekwere IP, Mabayoje MA, Mojeed HA. Performance Evaluation Of Manhattan And Euclidean Distance Measures For Clustering Based Automatic Text Summarization. 2019;

[52] Ravindran RM, Thanamani AS. K-means document clustering using vector space model. Bonfring International Journal of Data Mining 2015; 5[2]: 10-14.

[53] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Proceedings of the 45th annual meeting of the association of computational linguistics; 440-447.

[54] Ponomareva N, Thelwall M. Semi-supervised vs. cross-domain graphs for sentiment analysis. Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013; 571-578.

[55] Li S, Zong C. Multi-domain adaptation for sentiment classification: Using multiple classifier combining methods. 2008 International Conference on Natural Language Processing and Knowledge Engineering; IEEE; 1-8.

[56] Pan SJ, Ni X, Sun J-T, Yang Q, Chen Z. Cross-domain sentiment classification via spectral feature alignment. Proceedings of the 19th international conference on World wide web; 751-760.

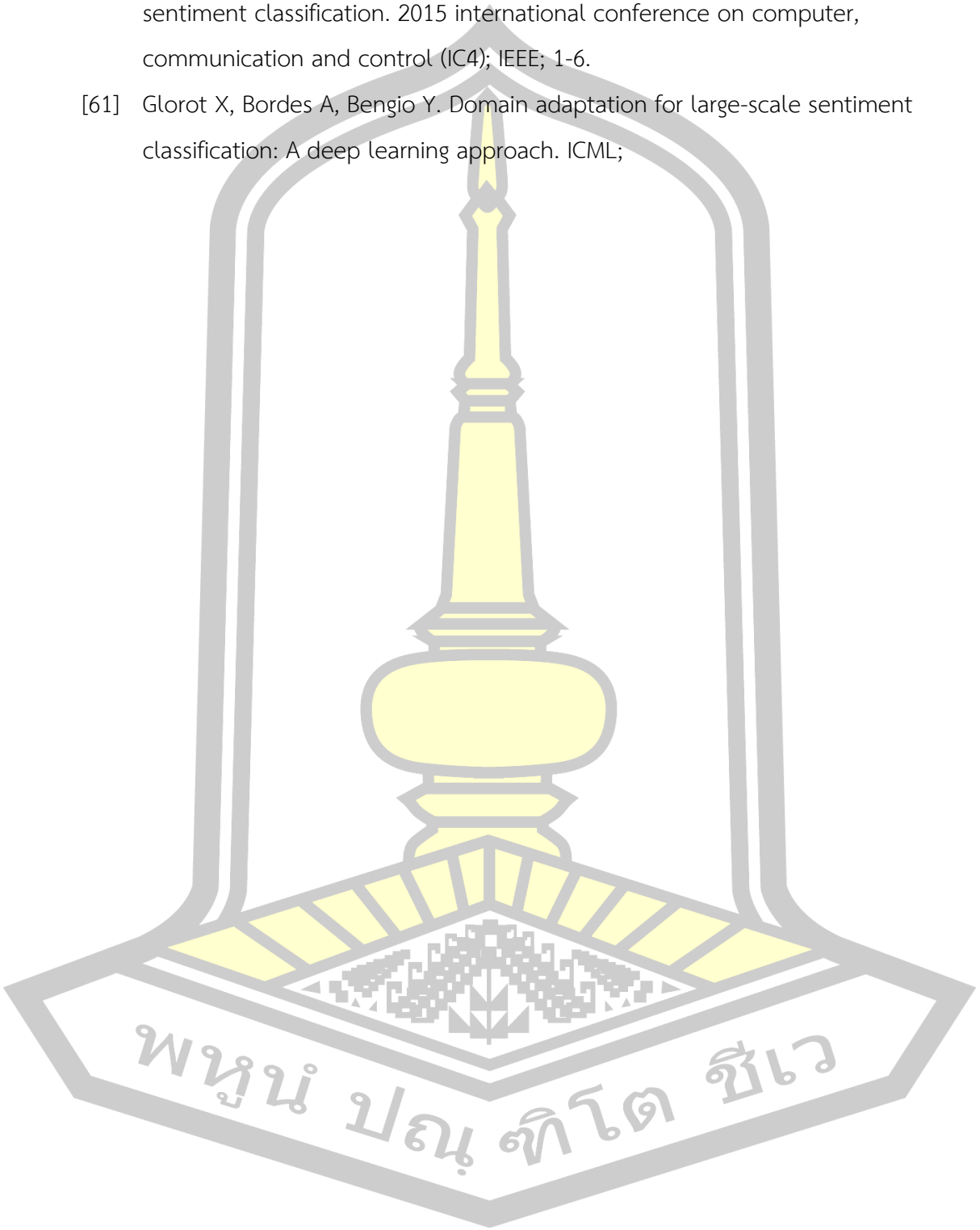
[57] Dhamija K. Hierarchical approach to document classification of 20 newsgroup dataset. Indian Statistical Institute, Kolkata; 2016.

[58] Asim MN, Khan MUG, Malik MI, Dengel A, Ahmed S. A robust hybrid approach for textual document classification. 2019 International conference on document analysis and recognition (ICDAR); IEEE; 1390-1396.

[59] Ding C, He X. K-means clustering via principal component analysis. Proceedings of the twenty-first international conference on Machine learning; 29.

[60] Ghag KV, Shah K. Comparative analysis of effect of stopwords removal on sentiment classification. 2015 international conference on computer, communication and control (IC4); IEEE; 1-6.

[61] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. ICML;



พหุน ปณฺ ฑิโต ชีเว

ประวัติผู้เขียน

ชื่อ	นายปณิธาน เมฆกมล
วันเกิด	14/09/2519
สถานที่เกิด	อุดรธานี
สถานที่อยู่ปัจจุบัน	369/4 ต.หมากแข้ง อ.เมือง จ.อุดรธานี 41000
ตำแหน่งหน้าที่การงาน	พนักงานในสถาบันอุดมศึกษา สายวิชาการ
สถานที่ทำงานปัจจุบัน	คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏอุดรธานี ต.หมากแข้ง อ.เมือง จ.อุดรธานี
ประวัติการศึกษา	พ.ศ. 2542 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิทยาการคอมพิวเตอร์ สถาบันราชภัฏพระนคร พ.ศ. 2548 ปริญญาครุศาสตรอุตสาหกรรมมหาบัณฑิต (ค.อ.ม.) สาขาคอมพิวเตอร์และเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี พ.ศ. 2564 ปริญญาดุษฎีบัณฑิต (ปร.ด) สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยสารคาม