บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงแนวคิด วิธีการ และอัลกอริทึมที่เกี่ยวข้อง ต่องานวิจัยและการพัฒนาระบบการสรุปความแบบอัตโนมัติสำหรับเอกสารคดีความ โดยแนวคิดและอัลกอริทึมที่เกี่ยวข้องมีดังนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ความหมายของคดีความ

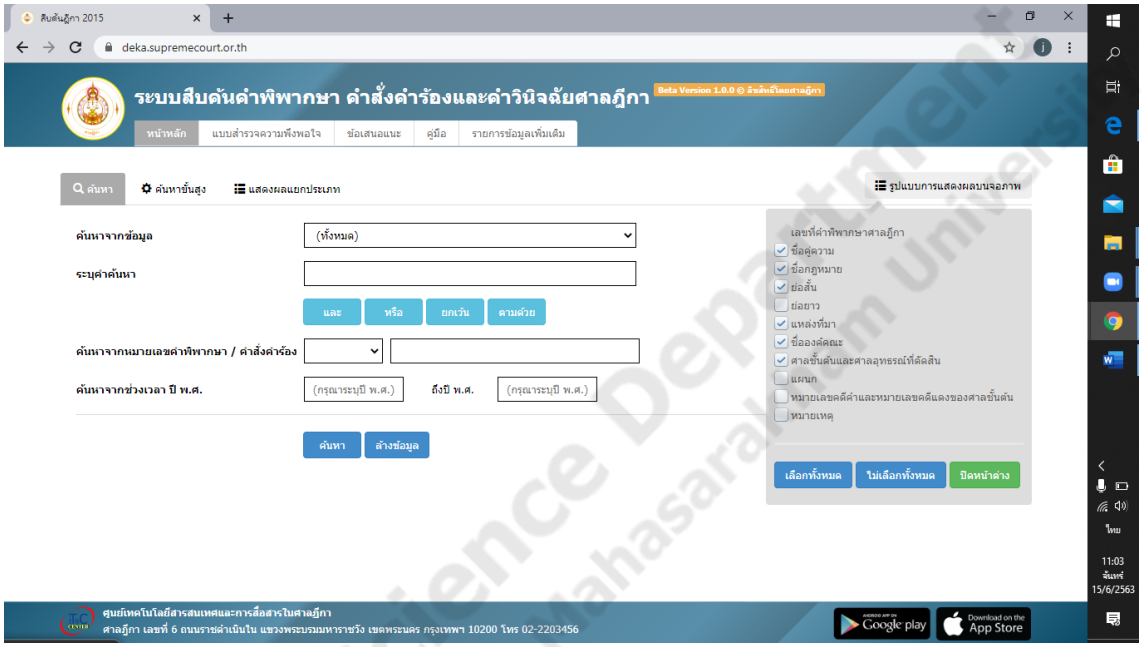
คดีความ (Lawsuit) [6] หมายถึง การฟ้องร้องดำเนินคดีโดยคู่กรณีหรือคู่ความอีกฝ่ายหนึ่งในศาลแพ่งของกฎหมาย คำว่า "คดีความ" ถูกนำมาใช้ในการอ้างถึงคดีแพ่งที่เกิดขึ้นในศาลยุติธรรมซึ่งโจทก์เป็นฝ่ายที่อ้างว่ามีความสูญเสียอันเนื่องมาจากการกระทำของจำเลยเรียกร้องให้มีการเยียวยาตามกฎหมายหรือเป็นธรรม จำเลยจะต้องตอบสนองต่อการร้องเรียนของโจทก์ หากโจทก์เป็นฝ่ายถูกอาจมีคำสั่งศาลที่หลากหลายเพื่อบังคับใช้สิทธิ ชดใช้ค่าเสียหายหรือกำหนดคำสั่งห้ามชั่วคราวหรือถาวรเพื่อป้องกันการกระทำหรือบังคับใช้การกระทำ อาจมีการออกคำตัดสินเพื่อป้องกันข้อพิพาททางกฎหมายในอนาคต

การดำเนินการของคดีความที่เรียกว่าการดำเนินคดี โจทก์และจำเลยถูกเรียกว่าเป็นคู่ความและทนายความที่เป็นตัวแทนของโจทก์จะถูกเรียกว่าเป็นผู้ดำเนินคดี [7] การฟ้องร้องดำเนินคดีในระยะนี้อาจอ้างถึงกระบวนการทางอาญา

คดีเริ่มต้นเมื่อมีการร้องเรียน (Claim) และยื่นต่อศาล การร้องเรียนควรระบุอย่างชัดเจนว่าโจทก์ได้รับความเสียหายหรือต้องการให้จำเลยชดใช้ค่าเสียหายอย่างไรและควรระบุข้อกล่าวหาหรือข้อเท็จจริงที่เกี่ยวข้องที่จะสนับสนุนการร้องเรียนทางกฎหมายที่โจทก์นำมาด้วย การร้องเรียนเริ่มต้นเป็นขั้นตอนที่สำคัญที่สุดในคดีแพ่ง เนื่องจากการร้องเรียนเป็นการกำหนดพื้นฐานทางกฎหมายและข้อเท็จจริงสำหรับคดีทั้งหมด

2.1.2 แหล่งข้อมูลคดีความแพ่งและพาณิชย์ที่มีการเผยแพร่ในปัจจุบัน

เอกสารคดีความที่ใช้ในการศึกษาได้จากเว็บไซต์ของศาลฎีกา ลักษณะไฟล์ที่ใช้เป็นไฟล์ข้อความที่ได้จากการดาวน์โหลดเมื่อวันที่ 27 พฤษภาคม พ.ศ.2563 ซึ่งเป็นข้อมูลที่มีการเปิดเผยให้กับบุคคลทั่วไปที่สนใจที่จะศึกษา ดังภาพประกอบที่ 2.1



ภาพประกอบที่ 2.1 เว็บไซต์ของศาลฎีกา
ที่มา: https://deka.supremecourt.or.th/

ในส่วนของเอกสารคดีความที่ใช้ในการศึกษานั้นเป็นการแสดงเอกสารคดีความที่มีการสรุปแล้ว โดยมีจำนวนคำในแต่ละเอกสารอย่างน้อย 1,200 คำและสูงสุดอยู่ที่ 2,500 คำ ซึ่งแสดงรายละเอียดดังภาพประกอบที่ 2.1

ภาพประกอบที่ 2.2 ลักษณะชุดข้อมูลคำพิพากษาของศาลฎีกาที่แบ่งออกเป็น 4 ส่วน

ที่มา: https://deka.supremecourt.or.th/

โดยในแต่ละส่วนนั้นผู้วิจัยได้มีการอธิบายถึงความสำคัญที่จะนำมาใช้ในการวิเคราะห์เป็นแนวทางในการดำเนินงานการสรุปความแบบอัตโนมัติ ดังนี้

ส่วนที่ 1 : ข้อพิพาทของคดีความ (Dispute)

ข้อพิพาทของคดีความ คือ ส่วนของเนื้อหาที่แสดงถึงเรื่องราวที่ฟ้องร้องกันระหว่างโจทย์และจำเลยหรือคู่กรณี

ส่วนที่ 2 : ข้อเท็จจริงของคดีความ (Fact)

ข้อเท็จจริงของคดีความ คือ ส่วนที่บอกถึงความเป็นจริงที่เกี่ยวข้องกับเหตุการณ์ที่เกิดขึ้นที่สามารถนำมาใช้เป็นหลักฐานหรือข้ออ้างอิงสำหรับการพิสูจน์ที่เกิดขึ้นในข้อพิพาทของคดีความ

ส่วนที่ 3 : ข้อวินิจฉัยของคดีความ (Decision)

ข้อวินิจฉัยของคดีความ คือ เนื้อหาที่แสดงถึงการวิเคราะห์และสังเคราะห์จากข้อเท็จจริงในคดีความ รวมถึงการตีความเพื่อให้ได้คำตัดสินคดี

ส่วนที่ 4 : คำตัดสินของคดีความ (Judgment)

คำตัดสินของคดีความ คือ ข้อสรุปของคดีความภายหลังจากที่ศาลได้วินิจฉัยแล้ว

2.2 การสรุปความแบบอัตโนมัติ (Automatic Text Summarization)

2.2.1 ความหมายของการสรุปความ

การสรุปความ (Text Summarization) [15,16] คือ กระบวนการสรุปสาระสำคัญของข้อมูลที่สำคัญที่สุด จากหนึ่งข้อความหรือจากหลายๆข้อความที่มีข้อมูลคล้ายคลึงกันกับต้นฉบับของข้อความนั้นและจะมีความยาวของเนื้อหาไม่เกินครึ่งของต้นฉบับ มีวัตถุประสงค์เพื่อลดความยาว และความซับซ้อนของข้อมูลโดยยังคงใจความสำคัญไว้และเพื่อให้ได้ภาพรวมโดยย่อของเอกสารข้อความขนาดใหญ่หรือชุดของเอกสาร

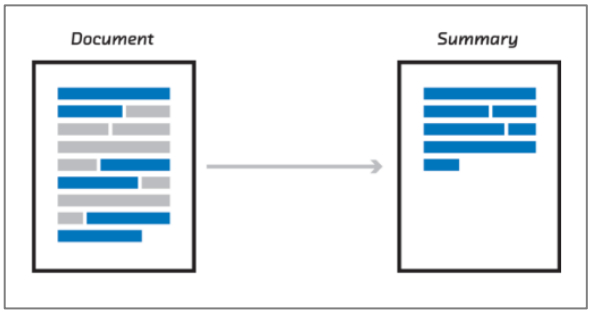
2.2.2 ประเภทของการสรุปความ

ประเภทของการสรุปความนั้นมีหลายลักษณะ โดยทั่วไปจะเป็นการพิจารณาใน 2 รูปแบบ ได้แก่ การพิจารณาตามจำนวนเอกสารที่ทำการสรุปความ และการพิจารณาตามรูปแบบของการสรุปความ ดังนี้

2.2.2.1 ประเภทของการสรุปความที่พิจารณาตามจำนวนเอกสารที่ทำการสรุปความ [8]

โดยในส่วนของการสรุปความจากจำนวนเอกสารที่ใช้ในการสรุปความจะแบ่งออกเป็น 2 ประเภท ได้แก่

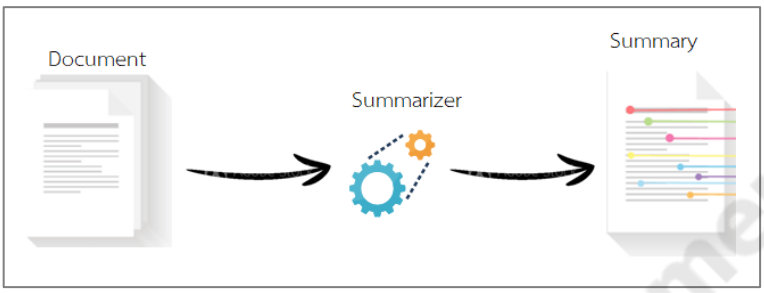
- การสรุปความจากเอกสารเดียว (Single-Document Summarization) [8] คือลักษณะของการสรุปสาระสำคัญจากเอกสารทีละเอกสาร



ภาพประกอบที่ 2.3 ลักษณะการสรุปความจากเอกสารเดียว

ที่มา: https://www.mc.ai/text-summarization

- การสรุปความจากหลายเอกสาร (Multi-Document Summarization) [8] คือลักษณะของการสรุปสาระสำคัญจากเอกสารหลาย ๆ เอกสารที่เขียนเกี่ยวกับหัวข้อเดียวกัน แล้วเอามารวมกันเพื่อสรุปเป็นใจความเดียว



ภาพประกอบที่ 2.4 ลักษณะการสรุปความจากหลายเอกสาร

2.2.2.2 ประเภทของการสรุปความที่พิจารณาตามรูปแบบของการสรุปความ

การสรุปความตามลักษณะรูปแบบและวิธีการในการสรุปความ จะแบ่งออกเป็น 2 ประเภท คือการสรุปความด้วยการสกัดสาระสำคัญ (Extraction-Based Summarization) และการสรุปความที่มีการนำมาเรียบเรียงใหม่ (Abstraction-Based Summarization) [24] ซึ่งมีรายละเอียดดังนี้

1) การสรุปความด้วยการสกัดสาระสำคัญ (Extraction-Based Summarization)

การสรุปความด้วยการสกัดสาระสำคัญ [25] จะเป็นการสรุปความจากเอกสารโดยที่ดึงข้อมูลจากเอกสารหรือประโยคที่คล้ายคลึงกัน โดยที่ประโยคไม่เปลี่ยนแปลงและนำมารวมกันเพื่อทำการสรุปความ

2) การสรุปความที่มีการนำมาเรียบเรียงใหม่ (Abstraction-Based Summarization)

เป็นการสรุปความที่มีการนำมาเรียบเรียงใหม่ [26] จะเป็นการสรุปความจากเอกสารโดยเทคนิคการเรียนรู้เชิงลึกขั้นสูงถูกนำไปใช้กับการถอดความ และทำให้เอกสารสั้นลงโดยการสร้างประโยคใหม่ที่อาจไม่ได้เป็นส่วนหนึ่งของเอกสาร และเนื่องจากอัลกอริทึมการเรียนรู้ด้วยเครื่องแบบนำมาเรียบเรียงใหม่ สามารถสร้างวลีและประโยคใหม่ที่แสดงข้อมูลที่สำคัญที่สุดจากเอกสารต้นฉบับ จึงทำให้สามารถช่วยให้ข้อมูลถูกต้องทางไวยากรณ์ ต่างจากการสรุปความด้วยการสกัดสาระสำคัญที่ข้อมูลทางไวยากรณ์อาจจะไม่ถูกต้องทั้งหมด

2.3 การประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP)

การประมวลผลภาษาธรรมชาติ หรือ NLP [9] เป็นสาขาย่อยของภาษาศาสตร์วิทยาการคอมพิวเตอร์และปัญญาประดิษฐ์ที่เกี่ยวข้องกับการโต้ตอบระหว่างคอมพิวเตอร์และภาษามนุษย์

(ธรรมชาติ) โดยการใช้โปรแกรมคอมพิวเตอร์ในการประมวลผลและวิเคราะห์ข้อมูลภาษาธรรมชาติจำนวนมาก เช่น การรู้จำเสียงพูด เทคนิคที่ใช้ในการประมวลผลภาษาธรรมชาติที่ผู้วิจัยได้นำมาใช้ในงานวิจัยมีดังนี้

2.3.1 การตัดคำ (Words segmentation)

การตัดคำ [10] เป็นการแบ่งส่วนคำหรือการโทเค็น (Tokenization) คำถือเป็นงานพื้นฐานในการประมวลผลภาษาธรรมชาติ (NLP) สำหรับการตัดคำภาษาไทยนั้นจะมีความยุ่งยากมากกว่า เช่นเดียวกับภาษาจีน ญี่ปุ่น และเกาหลี เป็นภาษาที่ไม่มีการแบ่งกลุ่ม มักเขียนเป็นคำเรียงและต่อเนื่อง

หากถ้าเป็นการตัดคำในภาษาอังกฤษ ฝรั่งเศส หรือสเปน ซึ่งเป็นภาษาที่มีรากฐานมาจากภาษาละตินมักจะไม่มีปัญหาและทำการตัดคำได้ง่ายกว่าภาษาที่ไม่ได้มีรากฐานมาจากภาษาละติน เพราะภาษาเหล่านี้มักมีเครื่องหมายคั่นคำ (Delimiting) เช่น ช่องว่าง (Spaces) เซมิโคลอน (Semi-Colon) จุลภาค (Comma) เครื่องหมายคำพูด (Quote) และจุด (Period)

สำหรับภาษาที่ไม่แบ่งส่วนมีการศึกษาเทคนิคเพื่อแก้ไขปัญหาการแบ่งคำ สามารถแบ่งออกเป็น 2 วิธี คือ ตามพจนานุกรม (Dictionary-Based: DCB) และการเรียนรู้ด้วยเครื่อง (Machine Learning Based: MLB)

จากการศึกษาพบว่าในการตัดคำภาษาไทยที่มีประสิทธิภาพและเป็นที่น่าพอใจ มักจะเป็นการตัดคำที่ได้จากวิธีการตัดคำตามพจนานุกรม โดยวิธีการตัดคำตามพจนานุกรม จะเป็นการใช้ชุดคำศัพท์จากพจนานุกรมในการแยกและการแบ่งข้อความขั้นตอนการแยกจะค้นหาชุดของอักขระตามพจนานุกรมเพื่อค้นหาคำที่ตรงกัน ประสิทธิภาพการตัดคำตามพจนานุกรมจะขึ้นอยู่กับคุณภาพและขนาดของคำที่ต้องการตัด พจนานุกรมที่ใช้มีคำที่ค่อนข้างง่ายและตรงไปตรงมาซึ่งมักจะมีปัญหา เช่น ปัญหาคำที่ไม่รู้จัก เป็นคำที่ไม่พบในพจนานุกรม หรือปัญหาความกำกวมของคำที่ทำการตัด ปัญหาเหล่านี้สามารถแก้ไขด้วยเทคนิคต่างๆ เช่น

2.3.1.1 เทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบเปรียบเทียบคำที่ยาวที่สุด (Longest Matching) [11]

ในภาษาไทยมีการเขียนตัวอักษรโดยไม่มีขอบเขตของคำที่ชัดเจน โดยคำที่เขียนมักขึ้นอยู่กับบริบทซึ่งมีหลากหลายวิธีในการแบ่งเป็นคำ ยกตัวอย่างเช่น

คำว่า “*อาจอง*” สามารถแบ่งออกเป็น

“*อา – จอง*” หรือ “*อาจ – อง*”

หรือคำว่า “*นั่งตากลม*” สามารถแบ่งออกเป็น

“*นั่ง - ตา – กลม*” หรือ “*นั่ง - ตาก – ลม*”

จากตัวอย่างดังกล่าวจะเห็นได้ว่ามีความซับซ้อนในการระบุคำจึงได้มีการนำเทคนิคการเปรียบเทียบคำที่ยาวที่สุดมาใช้ในการแบ่งคำ การทำงานจะอ่านข้อความจากซ้ายไปขวาแล้วนำคำที่ได้ไปเปรียบเทียบกับคำในพจนานุกรมและเลือกคำที่ยาวที่สุด อย่างไรก็ตามคำที่ยาวที่สุดที่ได้อาจไม่สอดคล้องกับความเป็นจริง

2.3.1.2 เทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบที่สอดคล้องมากที่สุด (Maximal Matching) [11]

เทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบที่สอดคล้องมากที่สุด [12] เป็นอีกเทคนิคที่ใช้ในการแก้ไขข้อบกพร่องของเทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบเปรียบเทียบคำที่ยาวที่สุด โดยเป็นวิธีการตัดคำที่ทำให้คำที่ได้มีจำนวนคำที่น้อยที่สุด และเลือกตัดคำที่เป็นไปได้ทั้งหมดของประโยคนั้นๆ ก่อน จากนั้นจึงจะทำการเลือกรูปแบบที่เหมาะสมที่สุดโดยการพิจารณาจาก จำนวนคำที่ตัดได้ และรูปแบบของประโยคที่มีจำนวนคำน้อยที่สุดจะถูกคัดเลือกให้เป็นรูปแบบที่สอดคล้องที่สุด

สมมติให้เอกสารเป็นคดีแดงที่ถูกตัดสินในศาลฎีกา เป็นคดีแดงในกลุ่มคดีแพ่งและพาณิชย์ ในเรื่องของการซื้อขายที่ดินจำนวน 3 ฉบับ โดยกำหนดให้เอกสารชุดที่ 1 แทนด้วย D1 เอกสารชุดที่ 2 แทนด้วย D2 และเอกสารชุดที่ 3 แทนด้วย D3 โดยกำหนดให้เอกสาร D1 เป็นเอกสารที่นำมาใช้เป็นต้นแบบในการเปรียบเทียบ ดังนี้

D1: พิพากษายืน ค่าฤชาธรรมเนียมชั้นฎีกาให้เป็นพับ

D2: พิพากษากลับ ให้ยกคำร้องขอของผู้ร้องทั้งสาม ค่าฤชาธรรมเนียมทั้งสามศาลให้เป็นพับ

D3: พิพากษายืน ค่าฤชาธรรมเนียมในศาลชั้นต้นและชั้นฎีกาให้เป็นพับ

ตารางที่ 2.1 แสดงการตัดคำ

เอกสารต้นฉบับ	เอกสารที่ผ่านการตัดคำ
พิพากษายืน ค่าฤชาธรรมเนียมชั้นฎีกาให้เป็นพับ	พิพากษา/ยืน/ค่าฤชาธรรมเนียม/ชั้น/ฎีกา/ให้/เป็น/พับ
พิพากษากลับ ให้ยกคำร้องขอของผู้ร้องทั้งสาม ค่าฤชาธรรมเนียมทั้งสามศาลให้เป็นพับ	พิพากษา/กลับ/ให้/ยก/คำร้องขอ/ของ/ผู้/ร้อง/ทั้ง/สาม/ค่าฤชาธรรมเนียม/ทั้ง/สาม/ศาล/ให้/เป็น/พับ
พิพากษายืน ค่าฤชาธรรมเนียมในศาลชั้นต้นและชั้นฎีกาให้เป็นพับ	พิพากษา/ยืน/ค่าฤชาธรรมเนียม/ใน/ศาลชั้นต้น/และ/ชั้น/ฎีกา/ให้/เป็น/พับ

2.3.2 การตัดคำหยุด (Stop-Word Removal)

การตัดคำหยุด [13] เป็นการตัดคำหรือสัญลักษณ์ที่พบบ่อยในเอกสาร แต่คำหรือสัญลักษณ์เหล่านั้นไม่ได้ส่งผลต่อใจความสำคัญและไม่มีนัยสำคัญต่อการวิเคราะห์ข้อมูลในเอกสาร ดังนั้นเมื่อทำการตัดคำเหล่านั้นออกไปแล้วก็ไม่ทำให้ใจความสำคัญในเอกสารนั้นๆ เปลี่ยนแปลงไป ตัวอย่างคำหยุดที่มักปรากฏในเอกสาร เช่น

- คำในกลุ่มคำบุพบท (Prepositions) เป็นคำที่นำหน้าคำนามเพื่อแสดงความสัมพันธ์ของคำนามอีกคำในประโยค เช่น in, on, with, so, ใต้, บน, ริม เป็นต้น
- คำในกลุ่มคำสันธาน (Conjunction) เป็นคำที่เชื่อมต่อคำอื่นหรือกลุ่มคำเช่น and, or, but, ทั้ง...และ, ทั้ง...หรือ, ...หรือ...และอื่นๆ
- คำในกลุ่มคำคุณศัพท์ (Adjective) เป็นคำที่ใช้บอกลักษณะและคุณสมบัติต่างๆ ของคำนามว่ามีลักษณะอย่างไร เช่น one, two, many, little, เล็ก, ใหญ่, น้อย เป็นต้น
- คำในกลุ่มคำสรรพนาม (Pronoun) เป็นคำที่ใช้เรียกแทนค านาม อันได้แก่ คน สัตว์ สิ่งของ สถานที่ เพื่อหลีกเลี่ยงการเรียกชื่อนั้นซ้ำๆ ตัวอย่างคำในกลุ่มนี้ เช่น ผม, ฉัน, ข้าพเจ้า, I, me, it, mine เป็นต้น

2.3.3 การสร้างตัวแทนข้อความ (Text Representation) [14]

เนื่องจากปัจจุบันคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของข้อความที่เป็นภาษาธรรมชาติได้โดยตรง จึงต้องมีการจำลองข้อความให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถอ่านเข้าใจและสามารถเรียนรู้ได้ โดยการสร้างตัวแทนข้อความที่นิยมใช้ก็คือการจำลองเอกสารให้อยู่ในแบบจำลองปริภูมิเวกเตอร์ (Vector Space Model: VSM)

การสร้างตัวแทนข้อความในรูปแบบของเวกเตอร์ เป็นหนึ่งในวิธีการแทนเอกสารที่ไม่มีโครงสร้าง (Unstructured Text Document) โดยการแบ่งข้อความให้อยู่ในรูปของถุงคำ (Bag-of-Words: BOW) ซึ่งจะเก็บอยู่ในรูปแบบของเวกตอร์ โดยกำหนดให้เอกสารแต่ละฉบับเปรียบเสมือนเวกเตอร์ของคำ ขนาดของเวกเตอร์ขึ้นอยู่กับจำนวนของคำที่ปรากฏอยู่ในเอกสาร โดยใช้วิธีการหาค่าความถี่ของคำ หรือเรียกว่าการหาค่าน้ำหนักของคำ (Term Weighting) มักแทนค่าด้วยเลขฐานสอง คือ จะมีค่าตั้งแต่ 0 ถึง 1 หากค่าเป็น 0 หมายความว่าไม่มีคำนั้นอยู่ในเอกสาร และถ้าหากค่าเป็น 1 ก็หมายความว่าพบคำนั้นในเอกสาร ซึ่งจะได้รูปแบบที่มีลักษณะของการแทนความสัมพันธ์ระหว่างคำ (Words: W) และเอกสาร (Documents: D) ด้วยเวกเตอร์ 2 มิติ

	w_1	w_2	w_3	w_4	...	w_i
D_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	$w_{1,4}$	...	$w_{1,i}$
D_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	$w_{2,4}$	...	$w_{2,i}$
...	...	...	...	...	...	...
D_j	$w_{j,1}$	$w_{j,2}$	$w_{j,3}$	$w_{j,4}$	...	$w_{j,i}$

ภาพประกอบที่ 2.5 ตัวอย่างของเวกเตอร์สเปซในรูปแบบของเมตริกซ์

2.3.4 การให้น้ำหนักคำ (Term Weighting)

ในการจัดหมวดหมู่เอกสารหรือข้อความมักจะถูกจำลองให้อยู่ในรูปแบบของเวกเตอร์ และเอกสารแต่ละฉบับจะแสดงเป็นเวกเตอร์ ซึ่งประกอบด้วยน้ำหนักของคำศัพท์หลายคำ การจัดหมวดหมู่มักจะเริ่มต้นด้วยการดูเอกสารและค้นหาคำสำคัญในเอกสารเหล่านั้น โดยใช้วิธีการให้น้ำหนักคำ (Term Weighting) ซึ่งจะแสดงให้เห็นถึงความสัมพันธ์ของเอกสารหรือข้อความที่เกี่ยวข้องกันได้ชัดเจนยิ่งขึ้น

การให้น้ำหนักคำ [12-15] คือ การกำหนดค่าน้ำหนักให้กับคำหรือเอกสาร เพื่อแสดงถึงความสำคัญของคำ ซึ่งจะจัดให้อยู่ในรูปแบบของ Vector Space Model (VSM) หรือ Bag-of-Words (BOW) ซึ่งหากคำใดที่พบเป็นจำนวนมากในเอกสารหรือคำที่พบบ่อย แสดงว่าคำเหล่านั้นไม่มีความสำคัญจึงไม่สามารถนำมาใช้เป็นตัวแทนของเอกสารได้

จากตารางที่ 2.1 จะสามารถแสดงความสัมพันธ์ระหว่าง *"คำสำคัญ"* และ *"เอกสาร"* ในรูปแบบของ BOW ได้ดังนี้

ตารางที่ 2.2 แสดงความสัมพันธ์ระหว่างคำสำคัญและเอกสาร

คำ (word)	D1	D2	D3
พิพากษา	1	1	1
ยืน	1	0	1
ค่าฤชาธรรมเนียม	1	1	1
ชั้น	1	0	1
ฎีกา	1	0	1
ให้	1	2	1
เป็น	1	1	1
พับ	1	1	1
กลับ	0	1	0
ยก	0	1	0
คำร้องขอ	0	1	0
ของ	0	1	0
ผู้	0	1	0
ร้อง	0	1	0
ทั้ง	0	2	0
สาม	0	2	0
ศาล	0	1	0
ใน	0	0	1
ศาลชั้นต้น	0	0	1
และ	0	0	1

ในการให้น้ำหนักคำๆ หนึ่งในเอกสารฉบับหนึ่งจะพิจารณาจากความถี่ของคำ (Term Frequency) ที่ปรากฏในเอกสารนั้นและจำนวนของเอกสารทั้งหมดที่มีคำๆ นั้นปรากฏอยู่ โดยวิธีการให้น้ำหนักคำมีอยู่หลายวิธี ดังนี้

2.3.4.1 การให้น้ำหนักคำตามความถี่ (Term Frequency: tf)

การให้น้ำหนักคำตามความถี่ [12,13] เป็นการกำหนดค่าความถี่ให้กับคำในแต่ละเอกสาร ค่าความถี่ที่ได้จะขึ้นอยู่กับจำนวนครั้งของการปรากฏคำนั้น ๆ ซึ่งสามารถคำนวณได้จากสมการที่ (2.1)

$$tf(t,d) = 1 + \log(f_{t,d}) \tag{2.1}$$

โดยที่ t คือ คำศัพท์ (keywords or term)

d คือ เอกสาร (document)

$f(t,d)$ คือ จำนวนของคำที่พบในเอกสาร d

ตัวอย่างการคำนวณหาค่า tf ของ D1 D2 และ D3 ด้วยสมการที่ (2.1)

$tf_{(\text{“พิพากษา”,D1})} = 1 + \log(1) = 1$
$tf_{(\text{“ยืน”,D1})} = 1 + \log(1) = 1$
$tf_{(\text{“ค่าฤชาธรรมเนียม”,D1})} = 1 + \log(1) = 1$
$tf_{(\text{“ชั้น”,D1})} = 1 + \log(1) = 1$
$tf_{(\text{“ฎีกา”,D1})} = 1 + \log(1) = 1$
$tf_{(\text{“ให้”,D1})} = 1 + \log(1) = 1$
$tf_{(\text{“เป็น”,D1})} = 1 + \log(1) = 1$
$tf_{(\text{“พับ”,D1})} = 1 + \log(1) = 1$
$tf_{(\text{“พิพากษา”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“กลับ”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“ให้”,D2})} = 1 + \log(2) = 1.301$
$tf_{(\text{“ยก”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“คำร้องขอ”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“ของ”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“ผู้”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“ร้อง”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“ทั้ง”,D2})} = 1 + \log(2) = 1.301$
$tf_{(\text{“สาม”,D2})} = 1 + \log(2) = 1.301$
$tf_{(\text{“ค่าฤชาธรรมเนียม”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“ศาล”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“เป็น”,D2})} = 1 + \log(1) = 1$
$tf_{(\text{“พับ”,D2})} = 1 + \log(1) = 1$

$tf_{(\text{“พิพากษา”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“ยืน”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“ค่าฤชาธรรมเนียม”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“ใน”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“ศาลชั้นต้น”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“และ”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“ชั้น”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“ฎีกา”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“ให้”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“เป็น”},D3)} = 1 + \log(1) = 1$
$tf_{(\text{“พับ”},D3)} = 1 + \log(1) = 1$

จากการคำนวณดังกล่าวแสดงให้เห็นถึงค่าน้ำหนักของคำโดยใช้ความถี่ของคำในแต่ละเอกสาร

2.3.4.2 การให้น้ำหนักคำแบบความถี่ผกผัน

(Term Frequency – Inverted Document Frequency : tf – idf)

การให้น้ำหนักคำแบบความถี่ผกผัน [12,13] เป็นอีกหนึ่งวิธีที่มีการนำมาใช้ในการให้น้ำหนักคำ โดย *tf* เป็นการกำหนดค่าความถี่ของคำหนึ่งๆ ที่พบในแต่ละเอกสาร ส่วน *idf* เป็นการนำความถี่ของเอกสารเข้ามาใช้ในการพิจารณา วิธีนี้จะมาในการแก้ไขปัญหาสำหรับคำที่มีความถี่สูงในเอกสารใดเอกสารหนึ่งเท่านั้น แต่ไม่ค่อยปรากฏในเอกสารอื่นๆ ซึ่งสามารถคำนวณได้จากจากสมการที่ (2.2)

$$Idf = \log\left(\frac{1+N}{df}\right) \quad (2.2)$$

โดยที่ N คือ จำนวนเอกสารทั้งหมดในคลัง
df คือ จำนวนเอกสารที่มีคำๆ นั้นปรากฏอยู่

สมมติว่ามีเอกสารในคลังเอกสารจำนวน 3 เอกสาร ดังนั้นค่า N จึงเท่ากับ 3 ขณะที่ df คือจำนวนเอกสารที่มีคำๆ นั้นปรากฏอยู่ ดังนั้นค่า Idf ของแต่ละคำสารมารถคำนวณได้ ดังนี้

$idf_{(\text{พิพากษา})} = \log((1+3)/3) = 0.125$
$idf_{(\text{ยืน})} = \log((1+3)/2) = 0.301$
$idf_{(\text{ค่าฤชาธรรมเนียม})} = \log((1+3)/3) = 0.125$
$idf_{(\text{ชั้น})} = \log((1+3)/2) = 0.301$
$idf_{(\text{ฎีกา})} = \log((1+3)/2) = 0.301$
$idf_{(\text{ให้})} = \log((1+3)/3) = 0.125$
$idf_{(\text{เป็น})} = \log((1+3)/3) = 0.125$
$idf_{(\text{พับ})} = \log((1+3)/3) = 0.125$

$$
\begin{aligned}
idf_{(\text{กลับ})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{ยก})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{คำร้องขอ})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{ของ})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{ผู้})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{ร้อง})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{ทั้ง})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{สาม})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{ศาล})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{ใน})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{ศาลชั้นต้น})} &= \log((1+3)/1) &&= 0.602\\
idf_{(\text{และ})} &= \log((1+3)/1) &&= 0.602
\end{aligned}
$$

เมื่อเราทำการคำนวณค่า Idf เสร็จเรียบร้อยแล้ว ต่อมานำค่าที่ได้มาทำการคำนวณหาค่า $tf - idf$ ดังต่อไปนี้

$$tf - idf = tf \times idf \qquad (2.3)$$

ตารางที่ 2.3 ตารางแสดงค่าน้ำหนักของคำ โดยการให้น้ำหนักแบบความถี่ผกผันของเอกสาร D1

คำที่ปรากฏในเอกสาร D1	ค่า tf	ค่า idf	$tf - idf$
พิพากษา	1	0.125	= 0.125
ยืน	1	0.301	= 0. 301
ค่าฤชาธรรมเนียม	1	0.125	= 0.125
ชั้น	1	0.301	= 0. 301
ฎีกา	1	0.301	= 0. 301
ให้	1	0.125	= 0.125
เป็น	1	0.125	= 0.125
พับ	1	0.125	= 0.125

ตารางที่ 2.4 ตารางแสดงค่าน้ำหนักของคำ โดยการให้น้ำหนักแบบความถี่ผกผันของเอกสาร D2

คำที่ปรากฏในเอกสาร D2	ค่า tf	ค่า idf	$tf - idf$
พิพากษา	1	0.125	= 0.125
กลับ	1	0.602	= 0.602
ให้	1.301	0.125	= 0.162
ยก	1	0.602	= 0.602
คำร้องขอ	1	0.602	= 0.602

ตารางที่ 2.4 ตารางแสดงค่าน้ำหนักของคำ โดยการให้น้ำหนักแบบความถี่ผกผันของเอกสาร D2 (ต่อ)

คำที่ปรากฏในเอกสาร D2	ค่า tf	ค่า idf	$tf - idf$
ของ	1	0.602	= 0.602
ผู้	1	0.602	= 0.602
ร้อง	1	0.602	= 0.602
ทั้ง	1.301	0.602	= 0.783
สาม	1.301	0. 602	= 0.783
ค่าฤชาธรรมเนียม	1	0.125	= 0.125
ศาล	1	0. 602	= 0.602
เป็น	1	0.125	= 0.125
พับ	1	0.125	= 0.125

ตารางที่ 2.5 ตารางแสดงค่าน้ำหนักของคำ โดยการให้น้ำหนักแบบความถี่ผกผันของเอกสาร D3

คำที่ปรากฏในเอกสาร D3	ค่า tf	ค่า idf	$tf - idf$
พิพากษา	1	0.125	= 0.125
ยืน	1	0.301	= 0.301
ค่าฤชาธรรมเนียม	1	0.125	= 0.125
ใน	1	0.602	= 0.602
ศาลชั้นต้น	1	0.602	= 0.602
และ	1	0.602	= 0.602
ชั้น	1	0.301	= 0.301
ฎีกา	1	0.301	= 0.301
ให้	1	0.125	= 0.125
เป็น	1	0.125	= 0.125
พับ	1	0.125	= 0.125

จากสมการที่ (2.3) การให้น้ำหนักคำแบบความถี่ผกผันเป็นวิธีที่นำมาช่วยในการแก้ปัญหาของการให้น้ำหนักคำแบบ tf และการให้น้ำหนักคำแบบ idf ให้มีประสิทธิภาพในการทำงานมากยิ่งขึ้น

2.3.5 เอ็นเซมเบิล (Ensemble) [17]

เป็นวิธีการที่อาศัยตัวจำแนกข้อมูลมากกว่าหนึ่งตัว ซึ่งแต่ละตัวที่จะใช้ในการจำแนกข้อมูลจะมีกระบวนการทำงานที่แตกต่างกัน และทุกตัวจำแนกข้อมูลจะกระทำกับข้อมูลเดียวกัน เมื่อได้ผลการจำแนกของแต่ละตัวที่ทำการจำแนกข้อมูลแล้ว ก็จะนำผลเหล่านั้นมาผ่านกระบวนการรวบรวม และตัดสินสุดท้าย เพื่อให้ได้เพียงผลจำแนกเดียว ซึ่งสามารถทำได้หลายวิธี เช่น

- การโหวตคะแนน (Vote Ensemble)
  - การลงคะแนนเสียงข้างมาก (Majority Voting)

การลงคะแนนเสียงข้างมากมีสามประเภทขึ้นอยู่กับการตัดสินใจว่าจะใช้รูปแบบใด

(1) ตัวจำแนกข้อมูลทั้งหมดเห็นด้วย (การลงคะแนนเป็นเอกฉันท์)
(2) มากกว่าครึ่งหนึ่งของตัวจำแนกเห็นด้วย (การลงคะแนนเสียงส่วนมาก)
(3) ได้รับคะแนนเสียงสูงสุด ไม่ว่าผลรวมของคะแนนเสียงเหล่านั้นจะเกิน 50% หรือไม่ (การลงคะแนนเสียงหลายเสียง)

โดยกำหนดให้ $d_{t,c}$ แทนเอกสารที่ทำการจำแนก และ max_c แทนค่าของผลโหวตของกลุ่มที่มากที่สุด ซึ่งสามารถแสดงได้ดังสมการที่ (2.4)

$$\sum_{t=1}^{T} d_{t,c^*} = max_c \sum_{t=1} d_{t,c} \tag{2.4}$$

หากตัวจำแนกข้อมูลที่ออกมาเป็นกลางจะแสดงให้เห็นว่าการลงคะแนนเสียงส่วนใหญ่เป็นกฎการรวมกันที่เหมาะสมที่สุด

- การถ่วงน้ำหนัก (Weight Vote Ensemble)

ถ้าตัวจำแนกข้อมูลบางตัวมีแนวโน้มที่จะถูกต้องมากกว่าคำอื่นๆ การให้น้ำหนักการตัดสินใจของตัวจำแนกข้อมูลเหล่านั้นมากขึ้นจะสามารถปรับปรุงประสิทธิภาพโดยรวมได้มากขึ้นเมื่อเทียบกับการลงคะแนนเสียงข้างมาก และสามารถกำหนดน้ำหนัก W_1 ให้กับตัวจำแนกข้อมูล h_1 ตามสัดส่วนของประสิทธิภาพการสรุปทั่วไป ซึ่งสามารถแสดงได้ดังสมการที่ (2.5)

$$\sum_{t=1}^{T} W_1 d_{1,c^*} = \max_{c} \sum_{t=1}^{T} W_1 d_{t,c} \tag{2.5}$$

2.3.6 การวิเคราะห์ด้วยค่าเทรสโฮลด์ (Threshold-Based Analysis) [18]

Threshold-based [19] คือ การกำหนดค่าเกณฑ์เพื่อใช้ในการแบ่งกลุ่มของข้อมูล โดยในการสร้างแบบจำลองทางคณิตศาสตร์หรือสถิติมักมีการใช้ค่าเทรสโฮลด์หรือชุดของคค่าเทรสโฮลด์ เพื่อใช้ในการแยกแยะช่วงของค่าที่คาดการณ์โดยโมเดลที่แตกต่างกันไป

การวิเคราะห์ด้วยค่าเทรสโฮลด์ คือ การวิเคราะห์อย่างใดอย่างหนึ่งโดยอาศัยค่าเทรสโฮลด์เพื่อช่วยในการตัดสินใจว่า ประโยคแต่ละประโยคนั้นมีความคล้ายคลึงกันมากน้อยเพียงใด

ตัวอย่างเช่น ในงานโครงงานฯ นี้จะใช้ค่าเทรสโฮลด์และค่าความคลายคลึงกัน โดยสมมติว่าสนใจประโยคที่มีค่าความคล้ายคลึงมากกว่า 0.4 หากประโยคใดก็ตามมีค่าความคล้ายคลึงน้อยกว่า 0.4 ก็จะไม่ถูกคัดเลือกให้อยู่ในเนื้อหาของการสรุปความ แต่ถ้าประโยคมีค่าความคล้ายคลึงกันมากกว่า 0.4 ก็จะถูกคัดเลือกให้อยู่ในเนื้อหาของการสรุปความ

2.3.7 การวิเคราะห์ความคล้ายคลึง (Similarity Analysis) [20]

การวิเคราะห์ความคล้ายคลึงเป็นวิธีการในการวิเคราะห์ว่าวัตถุ (Object) 2 ชิ้นมีความความคล้ายคลึงหรือความสอดคล้องกันหรือไม่ เทคนิคที่ใช้ในการวิเคราะห์ความคล้ายคลึงมีหลายเทคนิคที่เป็นที่นิยม เช่น ความคล้ายคลึงกันของโคไซน์ (Cosine Similarity) และ การวัดค่าความคล้ายคลึงด้วย BM25 เป็นต้น

1. ความคล้ายคลึงกันของโคไซน์ (Cosine Similarity)

ความคล้ายคลึงกันของโคไซน์ [21] เป็นการวัดความคล้ายคลึงกันที่ใช้กันทั่วไปสำหรับเวกเตอร์ที่มีค่าความจริงซึ่งใช้ในการดึงข้อมูล เพื่อให้คะแนนความคล้ายคลึงกันของเอกสารในแบบจำลองปริภูมิเวกเตอร์ ซึ่งสามารถคำนวณดังสมการที่ (2.6)

$$similarity = cos(\theta) = \frac{A \times B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2.6)$$

สมมติพิจารณาจากข้อมูลในตารางที่ 2.2 แสดงความสัมพันธ์ระหว่างคำสำคัญและเอกสารแสดงให้ความสัมพันธ์ระหว่างคำสำคัญในแต่ละเอกสาร จะสามารถแสดงในรูปแบบของ Array ได้ ซึ่งสามารถแสดงได้ดังนี้

D1 = [1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0]

D2 = [1,0,1,0,0,2,1,1,1,1,1,1,1,1,2,2,1,0,0,0]

D3 = [1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,1,1,1]

จากสมการที่ (2.6) สามารถแสดงการคำนวณความคล้ายคลึงของเอกสาร D1, D2 และ D3 ได้ดังนี้

$$D1 \cdot D2 = 6$$

$$|D1| = \sqrt{\begin{pmatrix} 1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+0^2+0^2 \\ +0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2 \end{pmatrix}} = 2.828$$

$$|D2| = \sqrt{\begin{pmatrix} 1^2+0^2+1^2+0^2+0^2+2^2+1^2+1^2+1^2+1^2 \\ +1^2+1^2+1^2+1^2+2^2+2^2+1^2+0^2+0^2+0^2 \end{pmatrix}} = 4.796$$

$$|D1| \times |D2| = 13.563$$

$$similarity = cos(\theta) = \frac{D1 \times D2}{\|D1\|\|D2\|} = \frac{6}{13.563} = 0.443$$

ดังนั้นการวัดความคล้ายคลึงกันด้วยโคไซน์ระหว่าง A และ B มีค่าเท่ากับ 0.443

$$D1 \cdot D3 = 8$$

$$|D1| = \sqrt{\begin{pmatrix} 1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+0^2+0^2 \\ +0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2 \end{pmatrix}} = 2.828$$

$$|D3| = \sqrt{\begin{pmatrix} 1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+0^2+0^2 \\ +0^2+0^2+0^2+0^2+0^2+0^2+0^2+1^2+1^2+1^2 \end{pmatrix}} = 3.317$$

$$|D1| \times |D3| = 9.380$$

$$similarity = cos(\theta) = \frac{A \times C}{\|A\|\|C\|} = \frac{8}{9.380} = 0.852$$

ดังนั้นการวัดความคล้ายคลึงกันด้วยโคไซน์ระหว่าง A และ C มีค่าเท่ากับ 0.852

จากการคำนวณข้างต้นแสดงการวัดความคล้ายคลึงกันของโคไซน์ [21] หากค่าความคล้ายคลึงที่ได้มีค่าเข้าใกล้ 1 จะแสดงว่าเอกสารทั้งสองที่ทำการเปรียบเทียบนั้นมีความคล้ายคลึงกันมาก โดยถ้าหากค่าที่ได้เป็น 0 แสดงว่าเอกสารทั้งสองนั้นมีความคล้ายคลึงกันน้อยมาก

2.3.7.1 การวัดค่าความคล้ายคลึงด้วย BM25

BM25 [22] หรือที่เรียกว่า Okapi BM25 เป็นฟังก์ชันที่ใช้ในการจัดอันดับเอกสาร เป็นเทคนิคหนึ่งที่มีประสิทธิภาพและนิยมใช้อย่างแพร่หลายเพื่อใช้พิจารณาความคล้าคลึงกันของเอกสาร

โดยจะใช้ได้ดีกับเอกสารที่มีลักษณะยาว โดยเอกสารที่ใชเป็นตัวแทน จะแทนด้วย q และเอกสารในคลังจะแทนด้วย d ซึ่งสามารถแสดงการคำนวณได้จากสมการที่ (2.7)

$$S(q,d) = \sum_{i=1}^{n} \left(IDF_i \frac{n_i(k_1+1)}{n_i + k_1\left(1 - b + \frac{b|D|}{\overline{D}}\right)} \right) \quad (2.7)$$

โดยที่ n_i คือ ค่า tf ที่ได้จากสมการที่ (2.1)

b คือ ค่าที่ใช้ควบคุมการปรับมาตรฐานความยาวของเอกสาร โดยปกติค่าจะอยู่ระหว่าง $0 \le b \le 1$

k_1 คือ โดยที่ค่า k [23] คือ ค่าที่ใช้ควบคุมอัตราความอิ่มตัวความถี่ของคำจะอยู่ระหว่าง $k1 \ge 0$

$|D|$ คือ ความยาวของเอกสารนั้น ๆ

$\overline{D}$ คือ ค่าความยาวเฉลี่ยของเอกสารทั้งหมดใcนคลัง

IDF_i คือ ค่า global weight หาได้จากสมการที่ (2.8)

$$IDF = Log\left(\frac{N-n+k}{n+k}\right) \quad (2.8)$$

โดยที่ N คือ จำนวนเอกสารทั้งหมดในคลัง

n คือ จำนวนเอกสารในคลังที่มีคำนั้นปรากฏอยู่

k คือ ค่าปรับ Smoothing โดยปกติแล้วจะมีค่าเท่ากับ 0.5

อย่างไรก็ตาม BM25 ยังมีข้อบกพร่องอยู่ จึงมีการนำ BM25Plus หรือ BM25+ [22] ที่ได้รับการพัฒนาเพื่อแก้ไขข้อบกพร่องประการหนึ่งของมาตรฐาน BM25 เนื่องจากการค้นหาคำในเอกสารที่ยาวมากและได้ผลลัพธ์ที่ไม่น่าพอใจ จึงทำให้การที่ใช้ BM25 กับเอกสารที่มีความยาวมากไม่มีประสิทธิภาพมากพอ โดยมีการบวกค่า δ เพิ่มเข้ามา ซึ่งค่า δ ที่แนะนำควรจะมีค่าเท่ากับ 1 จากการศึกษาพบว่าค่าพารามิเตอร์ k1 และ b [24] ที่นิยมใช้ คือ 2.0 และ 0.75 ตามลำดับ สูตรการให้คะแนนของ BM25+ สามารถแสดงได้ดังสมการที่ (2.9)

$$BM25F_{Sim_q} = \sum_{t \in q} \log\left(\frac{N+1}{df_t}\right) \times \left(\frac{(k_1+1) \times tf_{td}}{k_1 \times \left((1-b) + b \times \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} + \delta \right) \quad (2.9)$$

จากเอกสาร D1 และเอกสาร D2 สามารถคำนวณหาค่าความคล้ายคลึงกันด้วย BM25+ ได้ดังนี้

D1 : พิพากษา/ยืน/ค่าฤชาธรรมเนียม/ชั้น/ฎีกา/ให้/เป็น/พับ

D2 : พิพากษา/กลับ/ให้/ยก/คำร้องขอ/ของ/ผู้/ร้อง/ทั้ง/สาม/ค่าฤชาธรรมเนียม/ทั้ง/สาม/ศาล/ให้/เป็น/พับ

$$
\begin{aligned}
BM25F\_Similarity(D1, D2) \;=\; & 0.125 \times \left(\frac{(2.0+1)\times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{14}{12.5}\right)\right) + 1} + 1 \right) \\
& + 0.125 \times \left(\frac{(2.0+1)\times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{14}{12.5}\right)\right) + 1} + 1 \right) \\
& + 0.125 \times \left(\frac{(2.0+1)\times 1.301}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{14}{12.5}\right)\right) + 1.301} + 1 \right) \\
& + 0.125 \times \left(\frac{(2.0+1)\times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{14}{12.5}\right)\right) + 1} + 1 \right) \\
& + 0.125 \times \left(\frac{(2.0+1)\times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{14}{12.5}\right)\right) + 1} + 1 \right) \\
& = 0.243 + 0.243 + 0.265 + 0.243 + 0.243 \\
& = 1.237
\end{aligned}
$$

ดังนั้นการวัดความคล้ายคลึงกันด้วย MB25+ ระหว่าง D1 และ D2 มีค่าประมาณ **1.237**

จากเอกสาร D1 และเอกสาร D3 สามารถคำนวณหาค่าความคล้ายคลึงกันด้วย BM25+ ได้ดังนี้

D1 : พิพากษา/ยืน/ค่าฤชาธรรมเนียม/ชั้น/ฎีกา/ให้/เป็น/พับ

D3 : พิพากษา/ยืน/ค่าฤชาธรรมเนียม/ใน/ศาลชั้นต้น/และ/ชั้น/ฎีกา/ให้/เป็น/พับ

$$BM25F\_Similarity(D1, D3) = 0.125 \times \left(\frac{(2.0+1) \times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{11}{12.5}\right)\right) + 1} + 1 \right)$$

$$+0.301 \times \left(\frac{(2.0+1) \times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{11}{12.5}\right)\right) + 1} + 1 \right)$$

$$+0.125 \times \left(\frac{(2.0+1) \times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{11}{12.5}\right)\right) + 1} + 1 \right)$$

$$+0.301 \times \left(\frac{(2.0+1) \times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{11}{12.5}\right)\right) + 1} + 1 \right)$$

$$+0.301 \times \left(\frac{(2.0+1) \times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{11}{12.5}\right)\right) + 1} + 1 \right)$$

$$+0.125 \times \left(\frac{(2.0+1) \times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{11}{12.5}\right)\right) + 1} + 1 \right)$$

$$+0.125 \times \left(\frac{(2.0+1) \times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{11}{12.5}\right)\right) + 1} + 1 \right)$$

$$+0.125 \times \left(\frac{(2.0+1) \times 1}{2.0 \times \left((1-0.75) + 0.75 \times \left(\frac{11}{12.5}\right)\right) + 1} + 1 \right)$$

$$= 0.258 + 0.621 + 0.258 + 0.621 + 0.621 + 0.258 + 0.258 + 0.258$$

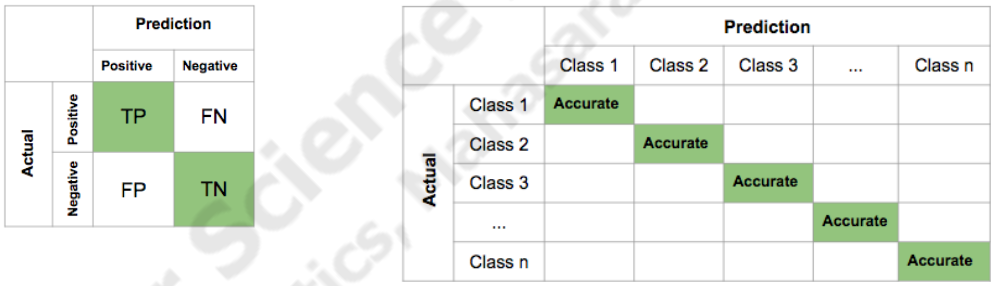
$$= 3.153$$

ดังนั้นการวัดความคล้ายคลึงกันด้วย BM25+ ระหว่าง D1 และ D3 มีค่าประมาณ **3.153**

2.4 การวัดประสิทธิภาพของโมเดล (Evaluation)

การประเมินผลสรุป [25] เป็นงานที่ยากเนื่องจากไม่มีการสรุปที่สมบูรณ์แบบสำหรับเอกสารที่กำหนดหรือชุดของเอกสาร เนื่องจากตัวชี้วัดที่แตกต่างกัน การไม่มีตัวชี้วัดการประเมินผลแบบมนุษย์หรือแบบอัตโนมัติทำให้ยากที่จะเปรียบเทียบ มนุษย์จึงมักจะหลีกเลี่ยงกระบวนการประเมินผลด้วยตนเอง โดยการนำเทคนิคต่าง ๆ มาใช้ในการประเมินผล ในงานวิจัยนี้ผู้วิจัยได้นำเทคนิคมาตราฐาน ที่เรียกว่า การวัดค่าความลึก (Recall) การวัดค่าความแม่นยำ (Precision) และการวัดค่า F-Measure มาใช้ในการประเมิณโมเดล

ซึ่งจากภาพประกอบที่ 2.6 (ก) แสดงตาราง Confusion Metrix สำหรับการจำแนกข้อมูลแบบ 2 กลุ่ม ในขณะที่ภาพประกอบที่ 2.6 (ข) เป็นการแสดงตาราง Confusion Metrix สำหรับการจำแนกข้อมูลแบบหลายกลุ่ม



(ก) Confusion Metrix (ข) Confusion Metrix for multi class

ภาพประกอบที่ 2.6 รูปแบบตาราง Confusion Metrix

ที่มา: https://docs.wso2.com/display/ML110/Model+Evaluation+Measures

โดยในการใช้ตาราง Confusion Metrix สำหรับประเมินการจำแนกข้อมูลแบบ 2 กลุ่ม สามารถอธิบายได้ดังต่อไปนี้

- True Positive (TP) คือ สิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้นจริง ในกรณี ทำนายว่า “*จริง*” และสิ่งที่เกิดขึ้น คือ “*จริง*”
- True Negative (TN) คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า “*ไม่จริง*” และสิ่งที่เกิดขึ้น คือ “*ไม่จริง*”
- False Positive (FP) คือ สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่า “*จริง*” แต่สิ่งที่เกิดขึ้น คือ “*ไม่จริง*”

- False Negative (FN) คือ สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือทำนายว่า “*ไม่จริง*” แต่สิ่งที่เกิดขึ้น คือ “*จริง*”

อย่างไรก็ตาม โดยในการใช้ตาราง Confusion Metrix สำหรับประเมินการจำแนกข้อมูลแบบหลายกลุ่ม สามารถอธิบายได้ดังต่อไปนี้

Recall (ความถูกต้องของการทำนายว่าจะเป็น “*จริง*” เทียบกับ จำนวนครั้งของเหตุการณ์ทั้งที่ทำนาย และ เกิดขึ้น ว่า “*เป็นจริง*”)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.10}$$

2.5 งานวิจัยที่เกี่ยวข้อง (Related Word)

จากการศึกษางานวิจัยที่ผ่านมาพบว่า การสรุปความกับเอกสารคดีความนั้นมีการศึกษา ทดลองและอธิบายถึงเทคนิควิธีการในการสรุปความด้านกฎหมายเป็นจำนวนมาก เช่น Grover และคณะ ได้เริ่มทำการศึกษา และประยุกต์เทคนิคการสรุปความแบบอัตโนมัติสำหรับโดเมนทางด้านกฎหมาย โดยเริ่มศึกษาเมื่อปี ค.ศ.2003 – ค.ศ.2005 ซึ่งเป็นงานวิจัยที่ค้นหาเทคนิควิธีการ และการประยุกต์ใช้เทคนิคทางด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) โดยศึกษาบนพื้นฐานงานวิจัยของ Teufal และ Moens [26] ซึ่งมีการนำมาประยุกต์ใช้กับการสรุปความด้านกฎหมายดังนี้

งานวิจัยของ Grover และคณะ [27] ได้ทำการศึกษาเกี่ยวกับโครงสร้างประโยคตาม Tense เพื่อให้ทราบว่าประโยคในเอกสารคดีความนั้นมีลักษณะเป็นประโยคในรูป Tense ใดบ้าง และใน 3 ส่วนหลักของเอกสารคดีความที่ต้องพิจารณา คือ (1) ข้อพิพาทของคดีความ (2) ข้อเท็จจริงของคดีความ และ (3) ข้อวินิจฉัยคดีความ

ซึ่งข้อมูลที่นำมาใช้ทดลองในงานวิจัยของ Grover และคณะ [27] คือ คำตัดสินคดีความจากศาลสภาขุนนาง ปะเทศอังกฤษ (House of Lords Judgments) ซึ่งผลการศึกษานั้นมีข้อจำกัดในเรื่องของความกำกวมของคำในประโยคของเอกสารคดีความ เนื่องจากเป็นคำเฉพาะทางด้านกฎหมายทำให้การจัดกลุ่มคำไม่ชัดเจน และจำนวนของหมวดหมู่ที่แยกนั้นไม่ครอบคลุมสำหรับการดำเนินคดีที่กว้างขึ้น ซึ่งจากการศึกษานี้ทำให้สามารถสรุปส่วนของการสรุปความเพื่อให้เกิดความเหมาะสมได้ 7 ส่วน คือ

(1) ข้อเท็จจริงของคดีความหรือการบอกถึงความเป็นจริงที่เกี่ยวกับเหตุการณ์ที่เกิดขึ้น (Fact)

(2) การดำเนินการของกฎหมายที่ผ่านการพิจารณาจากสารล่าง (Proceedings)

(3) ข้อพิพาทหรือมูลเหตุของคดีความ (Background)

(4) คำพิพากษาที่เกี่ยวข้องกับคดีความ (Proximation)

(5) คำพิพากษาที่ใช้ตัดสินคดีความ (Distancing)

(6) หากไม่มีคำพิพากษาก่อนหน้าและไม่มีจารีตประเพณีที่เคยใช้มาศาลจะใช้หลักเหตุผลที่คำนึงถึงความถูกต้องและเป็นธรรมหรือหลักความยุติธรรมในการตัดสินคดี (Framing)

(7) การเพิ่มความคิดเห็นเพิ่มเติมของศาล (Disposal)

ต่อมาในปี ค.ศ.2004 Hachey และ Grover [28] ได้นำผลการศึกษาของ Grover [29] ในส่วนของการแบ่งส่วนของคำตัดสิน 7 ส่วนที่ได้มาวิเคราะห์ถึงตัวจำแนกคำตัดสินคดีจากประโยคข้อความจากคำตัดสินคดีจากศาลสภาขุนนาง ประเทศอังกฤษ (House of Lords Judgments) จากปี ค.ศ. 2001 – ค.ศ.2003 จำนวน 40 เอกสาร ซึ่งในการศึกษานี้ผู้วิจัยได้ทดสอบโดยให้ผู้เชี่ยวชาญด้านกฎหมายจำนวน 2 คน ทำการจำแนกคำตัดสินคดีความตามการแบ่งส่วน 7 ส่วนโดยใช้ข้อมูลเพิ่มเติมที่นอกเหนือจากเอกสาร 40 เอกสารที่มี และวัดความน่าเชื่อถือที่ทำการจำแนกด้วยสัมประสิทธิ์คัปปา (Kappa co-efficient) ซึ่งมีค่าความน่าเชื่อถืออยู่ในระดับที่ดี จากการทดลองสรุปด้วยผู้เชี่ยวชาญด้านกฎหมายโดยใช้ 7 ส่วนที่ได้ทำให้เห็นถึงความเป็นไปได้ในการที่จะทำการสรุปความแบบอัตโนมัติจากการแบ่งส่วนเอกสาร 7 ส่วน แล้วทำการทดลองจำแนกคำตัดสินด้วยเทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP)

ต่อมาในปี ค.ศ.2005 Hachey และ Grover [29] ได้นำอัลกอริทึม C4.5 decision tree, Naïve Bayes, windows algorithm, support vector machines (SVM) และ maximum entropy วัดประสิทธิภาพด้วย micro-averaged F-score เพื่อศึกษากรอบการจำแนกประเภท มาตรฐานและการติดชื่อลำดับตัวจำแนกจากการบรรยายเหตุการณ์ โดยใช้ข้อมูลเดียวกับในงานของ Grover และนำเทคนิคด้านการประมวลผลภาษาธรรมชาติ (NLP) ในการดำเนินงานเพื่อหา Name entity และใช้งานวิจัยของ Teufal และ Moens เป็นแนวทางในการศึกษาการสรุปความในครั้งนี้ [19] และได้นำวิธี Feature Sets, Location, Thematic Words, Sentence Length, Quotation, Entities และ Cue Phrases ผลปรากฏว่าผลการสร้างแบบจำลองลำดับ Quotation ให้ผลมากที่สุดวัดจากค่า Maximum entropy F-score และในการสร้างแบบจำลองลำดับใช้ hidden Markov models โดยใช้ Maximum entropy Markov models (MEMMs) โดยจำแนกหมวดหมู่ดังนี้ Fact, Disposal, Textual และ Other ซึ่งผลการจำแนกที่ได้คือ Textual ให้ค่าความถูกต้องมากที่สุดในการทดลอง