บทความวิจัย

การสรุปข้อความแบบการสกัดสาระสำคัญสำหรับเอกสารคดีแดงที่ถูกตัดสินในศาลฎีกา
Extraction-Based Text Summarization for Cases Decided in the Supreme Court

กนกวรรณ โพธิ์ชัย[1] จุฑารัตน์ ไทยสำโรง[2] และจันทิมา พลพินิจ[3]
Kanokwan Pochai[1] Jutarat Thaisamrong[2] and Jantima Polpinij[3]

บทคัดย่อ

คดีความ (Lawsuit) [1] คือเรื่องที่พิพาทหรือกล่าวหากันในทางกฎหมาย ซึ่งต้องดำเนินการตามกระบวนการพิจารณาความตามที่กฎหมายกำหนด ซึ่งคดีความที่เกี่ยวข้องกับชีวิตประจำวันของผู้คนจะมีอยู่ 2 คดีคือ คดีแพ่งและพาณิชย์ (Civil And Commercial Case) และคดีอาญา (Criminal Case)

สำหรับผู้คนที่สนใจด้านคดีความนั้น คดีแดงกลายเป็นแหล่งศึกษาความรู้และข้อมูลด้านกฎหมายที่สำคัญของประชาชน นักศึกษา หรือแม้แต่นักกฎหมายเอง เพราะในการตัดสินคดีความในแต่ละคดีนั้น ผู้พิพากษาอาจจะมีมุมมองในตัดสินคดีความที่แตกต่างกันไป อย่างไรก็ตาม คดีที่ผ่านการตัดสินแล้วและมีการเผยแพร่ มักจะอยู่ในรูปแบบเอกสารข้อความที่ค่อนข้างยาว ทำให้ยากต่อการทำความเข้าใจ แม้ว่าเอกสารคดีแดงจะใช้คำศัพท์และรูปแบบที่เป็นมาตรฐานก็ตาม ซึ่งก็รวมถึงคดีแดงของประเทศไทยด้วย จากเหตุผลดังกล่าว โครงงานฉบับนี้ จึงนำเสนอการประยุกต์เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) โดยการประยุกต์กระบวนการแบบ Extraction-Based Text Summarization ในการสรุปประเด็นสำคัญที่อยู่ในคดีแดงของคดีแพ่งและพาณิชย์ โดยเน้นคดีความเรื่องการซื้อขายที่ดิน

คำสำคัญ: การสรุปความ, คดีแดง, การประมวลผลภาษาธรรมชาติ, การสรุปความด้วยการสกัดสาระสำคัญ

บทนำ

คดีความ (Lawsuit) [1] คือเรื่องที่พิพาทหรือกล่าวหากันในทางกฎหมาย ซึ่งต้องดำเนินการตามกระบวนการพิจารณาความตามที่กฎหมายกำหนด ซึ่งคดีความที่เกี่ยวข้องกับชีวิตประจำวันของผู้คนจะมีอยู่ 2 คดีคือ คดีแพ่งและพาณิชย์ (Civil And Commercial Case) และคดีอาญา (Criminal Case) [2]

โดยคดีแพ่งและพาณิชย์ [3] เป็นคดีที่เกี่ยวกับเรื่องส่วนตัวของบุคคล 2 ฝ่าย ที่มีการทำผิดสัญญาหรือโต้แย้งสิทธิกัน เป็นเรื่องที่ผู้ได้รับการโต้แย้งสิทธิ จะฟ้องร้องอีกฝ่ายที่ทำการโต้แย้งสิทธิ หรือทำผิดสัญญา ตัวอย่างคดีแพ่งและพาณิชย์ เช่น คดีกู้ยืมเงิน คดีผิดสัญญา คดีเช่าทรัพย์ คดีตั๋วเงิน คดีจำนอง คดีซื้อขาย คดีมรดก เป็นต้น

สำหรับคดีอาญา [2] นั้น เป็นคดีที่เมื่อเกิดขึ้นจะกระทบกระเทือนถึงสาธารณชนในบ้านเมือง โดยคดีที่ฟ้องร้องกันเนื่องจากมีการกระทำความผิดทางอาญาหรือรับโทษอื่น ๆ ในทางอาญา เช่น ให้ปรับ ให้จำคุกหรือ ให้ประหารชีวิต ตัวอย่างคดีอาญา เช่น คดีทำร้ายร่างกาย คดีลักทรัพย์ คดีชิงทรัพย์ คดีปล้นทรัพย์ คดีฆ่าคนตาย คดีประมาททำให้ผู้อื่นบาดเจ็บหรือเสียชีวิต คดีรับของโจร เป็นต้น

อย่างไรก็ตาม เมื่อเกิดคดีความก็จะเกิดการร้องทุกข์หรือการฟ้องร้อง ในกรณีที่เป็นเจ้าทุกข์หรือผู้ร้องทุกข์จะเรียกว่า “โจทก์ (Plaintiff)” [4] แต่หากเป็นผู้ที่คาดว่าเป็นผู้กระทำผิดก็จะเรียกว่า “จำเลย (Defendant)” [4] โดยทั่วไปแล้ว การตัดสินวินิจฉัยคดีความจะมี 3 ศาลคือ ศาลชั้นต้น ศาลอุทธรณ์ และศาลฎีกา คดีใดก็ตามหากมีการติดสินในชั้นศาลฎีกาไม่ว่าผลจะออกมาเป็นเช่นไร ถือว่าคดีนั่นได้สิ้นสุดลงแล้ว [5] สำหรับเป็นคดีที่รู้ผลแล้วว่าใครเป็นฝ่ายผิดฝ่ายถูกเนื่องจากศาลมีคำสั่งวินิจฉัยชี้ขาด หรือพิพากษา (Judge) แล้ว จะเรียกว่า “คดีแดง (Decided case)” [4]

สำหรับผู้คนที่สนใจด้านคดีความนั้น คดีแดงกลายเป็นแหล่งศึกษาความรู้และข้อมูลด้านกฎหมายที่สำคัญของประชาชน นักศึกษา หรือแม้แต่นักกฎหมายเอง เพราะในการตัดสินคดีความในแต่ละคดีนั้น ผู้พิพากษาอาจจะมีมุมมองในตัดสินคดีความที่แตกต่างกันไป อย่างไรก็ตาม คดีที่ผ่านการตัดสินแล้วและมีการเผยแพร่ มักจะอยู่ในรูปแบบเอกสารข้อความที่ค่อนข้างยาว ทำให้ยากต่อการทำความเข้าใจ แม้ว่าเอกสารคดีแดงจะใช้คำศัพท์และรูปแบบที่เป็นมาตรฐานก็ตาม ซึ่งก็รวมถึงคดีแดงของประเทศไทยด้วย

ดังนั้นโครงงานปริญญานิพนธ์นี้ จึงนำเสนอการประยุกต์เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ในการสรุปประเด็นสำคัญที่อยู่ในคดีแดง

การทบทวนวรรณกรรม

1. การสรุปความแบบอัตโนมัติ (Automatic Text Summarization)

    1) ความหมายของการสรุปความ

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

โดยย่อของเอกสารข้อความขนาดใหญ่หรือชุดของเอกสาร

2) ประเภทของการสรุปความ

ประเภทของการสรุปความที่พิจารณาตามจำนวนเอกสารที่ทำการสรุปความ [8] โดยในส่วนของการสรุปความจากจำนวนเอกสารที่ใช้ในการสรุปความจะแบ่งออกเป็น 2 ประเภท ได้แก่ การสรุปความจากเอกสารเดียว (Single-Document Summarization) [8] คือลักษณะของการสรุปสาระสำคัญจากเอกสารทีละเอกสาร และ การสรุปความจากหลายเอกสาร (Multi-Document Summarization) [8] คือลักษณะของการสรุปสาระสำคัญจากเอกสารหลาย ๆ เอกสารที่เขียนเกี่ยวกับหัวข้อเดียวกัน แล้วเอามารวมกันเพื่อสรุปเป็นใจความเดียว

ประเภทของการสรุปความที่พิจารณาตามรูปแบบของการสรุปความ

(1) การสรุปความด้วยการสกัดสาระสำคัญ Extraction-Based Summarization) การสรุปความด้วยการสกัดสาระสำคัญ [25] จะเป็นการสรุปความจากเอกสารโดยที่ดึงข้อมูลจากเอกสารหรือประโยคที่คล้ายคลึงกัน โดยที่ประโยคไม่เปลี่ยนแปลงและนำมารวมกันเพื่อทำการสรุปความ

(2) การสรุปความที่มีการนำมาเรียบเรียงใหม่ (Abstraction-Based Summarization) เป็นการสรุปความที่มีการนำมาเรียบเรียงใหม่ [26] จะเป็นการสรุปความจากเอกสารโดยเทคนิคการเรียนรู้เชิงลึกขั้นสูงถูกนำไปใช้กับการถอดความ และทำให้เอกสารสั้นลงโดยการสร้างประโยคใหม่ที่อาจไม่ได้เป็นส่วนหนึ่งของเอกสาร และเนื่องจากอัลกอริทึมการเรียนรู้ด้วยเครื่องแบบนำมาเรียบเรียงใหม่ สามารถสร้างวลีและประโยคใหม่ที่แสดงข้อมูลที่สำคัญที่สุดจากเอกสารต้นฉบับ จึงทำให้สามารถช่วยให้ข้อมูลถูกต้องทางไวยากรณ์ ต่างจากการสรุปความด้วยการสกัดสาระสำคัญที่ข้อมูลทางไวยากรณ์อาจจะไม่ถูกต้องทั้งหมด

2. การประมวลผลภาษาธรรมชาติ

การประมวลผลภาษาธรรมชาติ หรือ NLP [9] เป็นสาขาย่อยของภาษาศาสตร์วิทยาการคอมพิวเตอร์และปัญญาประดิษฐ์ที่เกี่ยวข้องกับการโต้ตอบระหว่างคอมพิวเตอร์และภาษามนุษย์ (ธรรมชาติ) โดยการใช้โปรแกรมคอมพิวเตอร์ในการประมวลผลและวิเคราะห์ข้อมูลภาษาธรรมชาติจำนวนมาก เช่น การรู้จำเสียงพูด เทคนิคที่ใช้ในการประมวลผลภาษาธรรมชาติที่ผู้วิจัยได้นำมาใช้ในงานวิจัยมีดังนี้

1) การตัดคำ

(Words segmentation)

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

- เทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบเปรียบเทียบคำที่ยาวที่สุด (Longest Matching) [11]
- เทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบที่สอดคล้องมากที่สุด (Maximal Matching) [11]

2) การสร้างตัวแทนข้อความ

(Text Representation) [14]

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 1 ตัวอย่างของเวกเตอร์สเปซในรูปแบบของเมตริกซ์

3) การให้น้ำหนักคำ (Term Weighting)

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

3. การวิเคราะห์ด้วยค่าเทรสโฮลด์ (Threshold-Based Analysis) [18]

Threshold-based [19] คือ การกำหนดค่าเกณฑ์เพื่อใช้ในการแบ่งกลุ่มของข้อมูล โดยในการสร้างแบบจำลองทางคณิตศาสตร์หรือสถิติมักมีการใช้ค่าเทรสโฮลด์หรือชุดของคค่าเทรสโฮลด์ เพื่อใช้ในการแยกแยะช่วงของค่าที่คาดการณ์โดยโมเดลที่แตกต่างกันไป

การวิเคราะห์ด้วยค่าเทรสโฮลด์ คือ การวิเคราะห์อย่างใดอย่างหนึ่งโดยอาศัยค่าเทรสโฮลด์เพื่อช่วยในการตัดสินใจว่า ประโยคแต่ละประโยคนั้นมีความคล้ายคลึงกันมากน้อยเพียงใด

4. การวิเคราะห์ความคล้ายคลึง (Similarity Analysis) [20]

เทคนิคที่ใช้ในการวิเคราะห์ความคล้ายคลึงมีหลายเทคนิคที่เป็นที่นิยม เช่น ความคล้ายคลึงกันของโคไซน์ (Cosine Similarity) และ การวัดค่าความคล้ายคลึงด้วย BM25 เป็นต้น

1) การวัดค่าความคล้ายคลึงด้วย BM25

BM25 [22] หรือที่เรียกว่า Okapi BM25 เป็นฟังก์ชันที่ใช้ในการจัดอันดับเอกสาร เป็นเทคนิคหนึ่งที่มีประสิทธิภาพและนิยมใช้อย่างแพร่หลายเพื่อใช้พิจารณาความคล้าคลึงกันของเอกสารโดยจะใช้ได้ดีกับเอกสารที่มีลักษณะยาว โดยเอกสารที่ใชเป็นตัวแทน จะแทนด้วย q และเอกสารในคลังจะแทนด้วย d

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

5. งานวิจัยที่เกี่ยวข้อง

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

ต่อมาในปี ค.ศ.2004 Hachey และ Grover [28] ได้นำผลการศึกษาของ Grover [29] ในส่วนของการแบ่งส่วนของคำตัดสิน 7 ส่วนที่ได้มาวิเคราะห์ถึงตัวจำแนกคำตัดสินคดีจากประโยคข้อความ จากคำตัดสินคดีจากศาลสภาขุนนาง 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 ผลปรากฏว่าผลการสร้างแบบจำลองลำดับ Quotation ให้ผลมากที่สุดวัดจากค่า Maximum entropy F-score และในการสร้างแบบจำลองลำดับใช้ hidden Markov models โดย

ใช้ Maximum entropy Markov models (MEMMs) โดยจำแนกหมวดหมู่ดังนี้ Fact, Disposal, Textual และ Other ซึ่งผลการจำแนกที่ได้คือ Textual ให้ค่าความถูกต้องมากที่สุดในการทดลอง

กระบวนการวิจัย

ในบทนี้จะอธิบายถึงชุดข้อมูลเอกสารคำพิพากษาศาลฎีกาที่ใช้ในโครงงานนี้ และวิธีการดำเนินงานในการสรุปความสำหรับเอกสารคดีความ ดังนี้

1. การรวบรวมข้อมูล (Data Collection)

ในโครงงานนี้ได้ใช้ชุดข้อมูลเอกสารคำพิพากษาศาลฎีกาประเภทคดีแพ่งและพาณิชย์ สำหรับโครงงานนี้จะทำการศึกษาและทดลองโดยใช้คำพิพากษาในส่วนของคดีความที่เกี่ยวกับการซื้อขายที่ดิน ชุดข้อมูลที่นำมาใช้ในโครงงานนี้ได้จากการดาวโหลดไฟล์เอกสารคำพิพากษาศาลฎีกา (Text file) ส่วนของการซื้อขายที่ดิน จำนวน 60 เอกสาร

ภาพประกอบที่ 2 ลักษณะของชุดข้อมูล

ภาพประกอบที่ 3 ตัวอย่างชุดข้อมูลในรูปแบบ XML

2. การแยกองค์ประกอบของเอกสาร

ขั้นตอนนี้จะเป็นการแยกองค์ประกอบของเอกสาร โดยจะใช้คำสำคัญในการแยกองค์ประกอบของเอกสาร ซึ่งสามารถแบ่งออกได้เป็น 4 ส่วน คือ Dispute Fact Decision และ Judgment

ภาพประกอบที่ 4 เอกสารที่ได้หลังผ่านกระบวนการแยกองค์ประกอบ

3. การสกัดประโยคสำคัญเพื่อการสรุปความ (Summarizing Text)

โดยจะทำการสุ่มเอกสารจำนวน 5 เอกสาร ขึ้นมาเป็นเอกสารหลักสำหรับทำการเปรียบเทียบและอีก 55 เอกสารจะใช้ในการศึกษาเรื่องการสรุปความด้วยเทคนิคแบบการวิเคราะห์ความคล้ายคลึง โดยจะมีขั้นตอนการทำงาน ดังนี้

ขั้นตอนที่ 1 : การตัดคำ ขั้นตอนนี้จะเป็นการแบ่งข้อความออกเป็นคำ โดยใช้เทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบเปรียบเทียบคำที่ยาวที่สุด (Longest Matching)

ขั้นตอนที่ 2 : การแยกประโยค ขั้นตอนนี้เป็นการแยกข้อความในแต่ละส่วนออกเป็นประโยค โดยจะใช้ช่องว่างในการแยกประโยค

ขั้นตอนที่ 3 : การสร้างตัวแทนข้อความและการให้น้ำหนักคำ ขั้นตอนนี้เป็นการแสดงประโยคแต่ละประโยคในเอกสารแต่ละส่วนด้วยแบบจำลองปริภูมิเวกเตอร์ (Vector Space Model: VSM) และการให้น้ำหนักคำในประโยค เพื่อแสดงถึงความสำคัญของคำนั้นๆ ว่ามีมากน้อยเพียงใด โดยการให้น้ำหนักคำจะใช้ค่า idf ในการให้น้ำหนักคำซึ่งมีสมการดังนี้

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

โดยที่ N คือ จำนวนเอกสารทั้งหมดในคลัง

df คือ จำนวนเอกสารที่มีคำๆ นั้นปรากฏอยู่

ขั้นตอนที่ 4 : การวัดค่าความคล้ายคลึงด้วย BM25+ ขั้นตอนนี้เป็นการสกัดประโยคด้วยการวิเคราะห์ความคล้ายคลึงของข้อความโดยมีสมการดังนี้

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

ซึ่งการเปรียบเทียบประโยคจะทำการเปรียบเทียบระหว่างประโยคที่เป็นการสรุปความโดยผู้เชี่ยวชาญและประโยคในเอกสารที่กำลังจะทำความสรุปความแบบอัตโนมัติ หากประโยคใดในเอกสารที่กำลังพิจารณามีความเหมือนหรือสอดคล้องกันกับประโยคที่เป็นการสรุปความโดยผู้เชี่ยวชาญ ก็จะถูกสกัดออกมาเป็นประโยคของการสรุปความแบบอัตโนมัติ อย่างไรก็ตาม แต่ละประโยคที่จะมีเปรียบเทียบทั้งหมด 5 รอบ

หลังจากประโยคที่กำลังพิจารณา ได้ทำการวิเคราะห์ความคล้ายคลึงกับทุกประโยคที่เป็นสรุปโดยผู้เชี่ยวชาญแล้ว จากนั้นจะนำค่าความคล้ายคลึงของประโยคทั้งหมดในเอกสารที่กำลังพิจารณามาบวกกันแล้วหารด้วยจำนวนของประโยคทั้งหมดในเอกสารที่กำลังพิจารณา จึงจะได้ค่าความคล้ายคลึงเฉลี่ยของประโยคนั้นๆ จากนั้นจะนำค่าความคล้ายคลึงไปหาค่าเฉลี่ยเพื่อที่จะได้นำไปกำหนดค่าเทรสโฮลด์ เพื่อใช้ในการคัดเลือกประโยคเพื่อนำมาเป็นสรุปความ

4. การประเมิน (Evaluation)

เป็นขั้นตอนการประเมินโมเดลเพื่อใช้ในการสรุปความ ก่อนการนำไปใช้จริงซึ่งโดยทั่วไปจะใช้เทคนิคมาตราฐานที่เรียกว่า การวัดค่าความระลึก (Recall) ในการประเมินกานชรแยกองค์ประกอบของเอกสาร และจะใช้อัลกอริทึมของ ROUGE เข้ามาใช้ในการประเมินประสิทธิภาพของการสรุปความ โดยจะใช้ในส่วนของ ROUGE-N และ ROUGE-L โดยสามารแสดงสมการได้ดังนี้

$$Recall = \frac{tp}{tp + fn}$$

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

ผลการวิจัย

1. ผลการประเมินประสิทธิภาพการแยกองค์ประกอบของเอกสาร

ขั้นตอนนี้จะเป็นการแยกองค์ประกอบของเอกสาร ซึ่งสามารถแบ่งออกได้เป็น 4 ส่วน คือ Dispute Fact Decision และ Judgment โดยจะทำ

การแยกองค์ประกอบโดยการใช้ คำ หรือ วลี ที่สำคัญในการแยกองค์ประกอบเพียงเท่านั้น

แต่พบปัญหาคือ การแยกองค์ประกอบยังไม่มีประสิทธิภาพมากพอ จึงทำให้บางครั้งทำการแยกองค์ประกอบได้ไม่ถูกต้อง จึงทำการปรับวิธีการแยกองค์ประกอบคือ การนำเอา คำ หรือ วลี ที่สำคัญและย่อหน้าเข้ามาช่วยในการทำงานเพื่อให้มีประสิทธิภาพในการทำงานมา

ตารางที่ 1 การทดสอบการแยกองค์ประกอบของเอกสาร

ก่อนปรับกระบวนการ	หลังปรับกระบวนการ
0.631	0.847

ซึ่งจะเห็นได้ว่าหลังจากทำการปรับกระบวนการแล้ว การทำแบบนำเอา คำ หรือ วลี ที่สำคัญและย่อหน้าเข้ามาช่วยในการแยกองค์ประกอบทำสามารถทำให้สามารถแยกองค์ประกอบได้อย่างมีประสิทธิภาพมากขึ้น

2. ผลการประเมินประสิทธิภาพการสกัดประโยคสำคัญเพื่อสรุปความ

ตารางที่ 2 การทดสอบแบบเปรียบเทียบ 1 ประโยค

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
1	0.486	0.161	0.498	0.219
2	0.421	0.232	0.535	0.361
3	0.413	0.176	0.474	0.253
4	0.501	0.324	0.640	0.344
5	0.436	0.313	0.503	0.314
6	0.457	0.326	0.495	0.357
7	0.496	0.237	0.548	0.266
8	0.561	0.386	0.592	0.393
9	0.433	0.341	0.476	0.375
10	0.492	0.319	0.527	0.354
11	0.419	0.351	0.463	0.376
12	0.487	0.352	0.518	0.364
13	0.531	0.363	0.543	0.377

ตารางที่ 2 การทดสอบแบบเปรียบเทียบ 1 ประโยค (ต่อ)

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
14	0.455	0.341	0.475	0.359
15	0.485	0.328	0.531	0.362
16	0.563	0.347	0.575	0.394
17	0.423	0.372	0.468	0.402
18	0.352	0.251	0.422	0.337
19	0.501	0.327	0.545	0.364
20	0.523	0.413	0.537	0.422
21	0.434	0.224	0.461	0.342
22	0.413	0.332	0.436	0.387
23	0.492	0.355	0.511	0.365
24	0.521	0.367	0.533	0.384
25	0.413	0.327	0.436	0.355
26	0.448	0.241	0.471	0.387
27	0.533	0.337	0.573	0.362
28	0.437	0.375	0.458	0.396
29	0.466	0.312	0.498	0.351
30	0.447	0.314	0.483	0.347
ค่าเฉลี่ย	**0.467**	**0.314**	**0.507**	**0.355**

ตารางที่ 3 การทดสอบแบบเปรียบเทียบ 2 ประโยค

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
1	0.522	0.181	0.544	0.265
2	0.4525	0.261	0.557	0.389
3	0.4435	0.198	0.516	0.319

ตารางที่ 3 การทดสอบแบบเปรียบเทียบ 2 ประโยค (ต่อ)

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
4	0.5085	0.3645	0.598	0.373
5	0.47	0.352	0.558	0.401
6	0.491	0.3665	0.529	0.389
7	0.533	0.2665	0.561	0.313
8	0.573	0.399	0.600	0.409
9	0.465	0.369	0.508	0.394
10	0.5285	0.3185	0.547	0.368
11	0.450	0.3685	0.511	0.4055
12	0.5235	0.396	0.544	0.396
13	0.5705	0.408	0.568	0.397
14	0.489	0.3835	0.531	0.391
15	0.521	0.369	0.542	0.397
16	0.605	0.39	0.574	0.405
17	0.4545	0.4185	0.540	0.416
18	0.378	0.282	0.484	0.383
19	0.4885	0.3675	0.541	0.373
20	0.523	0.4145	0.561	0.418
21	0.4665	0.252	0.527	0.393
22	0.446	0.3735	0.489	0.381
23	0.4785	0.379	0.504	0.391
24	0.52	0.3675	0.545	0.388
25	0.4435	0.3675	0.502	0.386
26	0.4815	0.271	0.503	0.391
27	0.5725	0.379	0.561	0.394
28	0.4695	0.3865	0.534	0.408

ตารางที่ 3 การทดสอบแบบเปรียบเทียบ 2 ประโยค (ต่อ)

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
29	0.5005	0.343	0.517	0.391
30	0.4805	0.348	0.521	0.366
ค่าเฉลี่ย	**0.494**	**0.344**	**0.537**	**0.383**

ตารางที่ 4 การทดสอบแบบเปรียบเทียบ 3 ประโยค

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
1	0.558	0.201	0.591	0.311
2	0.484	0.29	0.579	0.417
3	0.474	0.22	0.558	0.385
4	0.516	0.405	0.556	0.402
5	0.504	0.391	0.614	0.489
6	0.525	0.407	0.563	0.422
7	0.570	0.296	0.575	0.361
8	0.585	0.412	0.609	0.426
9	0.497	0.397	0.541	0.413
10	0.565	0.318	0.568	0.383
11	0.481	0.386	0.559	0.435
12	0.560	0.44	0.571	0.429
13	0.610	0.453	0.593	0.417
14	0.523	0.426	0.587	0.424
15	0.557	0.41	0.553	0.433
16	0.647	0.433	0.574	0.416
17	0.486	0.465	0.613	0.431
18	0.404	0.313	0.546	0.429
19	0.476	0.408	0.537	0.382

ตารางที่ 4 การทดสอบแบบเปรียบเทียบ 3 ประโยค (ต่อ)

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
20	0.523	0.416	0.586	0.415
21	0.499	0.280	0.593	0.445
22	0.479	0.415	0.542	0.376
23	0.465	0.403	0.497	0.418
24	0.519	0.368	0.557	0.393
25	0.474	0.408	0.568	0.417
26	0.515	0.301	0.536	0.395
27	0.612	0.421	0.549	0.427
28	0.502	0.398	0.611	0.420
29	0.535	0.374	0.537	0.431
30	0.514	0.382	0.559	0.386
ค่าเฉลี่ย	**0.521**	**0.374**	**0.567**	**0.410**

ตารางที่ 5 การทดสอบ แบบเปรียบเทียบ 4 ประโยค

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
1	0.5715	0.226	0.605	0.349
2	0.496	0.326	0.593	0.469
3	0.4855	0.2475	0.571	0.433
4	0.5285	0.4555	0.569	0.452
5	0.5165	0.4395	0.629	0.550
6	0.538	0.4575	0.577	0.474
7	0.584	0.333	0.589	0.406
8	0.5995	0.4635	0.624	0.479
9	0.509	0.4465	0.554	0.464
10	0.579	0.3575	0.582	0.430

ตารางที่ 5 การทดสอบ แบบเปรียบเทียบ 4 ประโยค (ต่อ)

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
11	0.493	0.434	0.572	0.489
12	0.574	0.495	0.585	0.482
13	0.625	0.5095	0.607	0.469
14	0.536	0.479	0.601	0.477
15	0.5705	0.461	0.566	0.487
16	0.663	0.487	0.588	0.468
17	0.498	0.523	0.628	0.484
18	0.414	0.352	0.559	0.482
19	0.4875	0.459	0.550	0.429
20	0.536	0.468	0.601	0.466
21	0.511	0.315	0.607	0.503
22	0.4905	0.4665	0.555	0.423
23	0.4765	0.453	0.509	0.47
24	0.5315	0.414	0.570	0.442
25	0.4855	0.459	0.582	0.469
26	0.5275	0.3385	0.549	0.444
27	0.627	0.4735	0.562	0.480
28	0.5145	0.4475	0.626	0.472
29	0.548	0.4205	0.550	0.484
30	0.5265	0.4295	0.572	0.434
ค่าเฉลี่ย	**0.534**	**0.421**	**0.581**	**0.462**

ตารางที่ 6 การทดสอบแบบเปรียบเทียบ 5 ประโยค

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
1	0.585	0.251	0.620	0.388

ตารางที่ 6 การทดสอบแบบเปรียบเทียบ 5 ประโยค (ต่อ)

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
2	0.508	0.362	0.607	0.521
3	0.497	0.275	0.585	0.481
4	0.541	0.506	0.583	0.502
5	0.529	0.488	0.644	0.611
6	0.551	0.508	0.591	0.527
7	0.598	0.370	0.603	0.451
8	0.614	0.515	0.639	0.532
9	0.521	0.496	0.568	0.516
10	0.593	0.397	0.596	0.478
11	0.505	0.482	0.586	0.543
12	0.588	0.550	0.599	0.536
13	0.640	0.566	0.622	0.521
14	0.549	0.532	0.616	0.530
15	0.584	0.512	0.580	0.541
16	0.679	0.541	0.602	0.520
17	0.510	0.581	0.643	0.538
18	0.424	0.391	0.573	0.536
19	0.499	0.510	0.563	0.477
20	0.549	0.520	0.615	0.518
21	0.523	0.350	0.622	0.556
22	0.502	0.518	0.569	0.470
23	0.488	0.503	0.521	0.522
24	0.544	0.460	0.584	0.491
25	0.497	0.510	0.596	0.521
26	0.540	0.376	0.562	0.493
27	0.642	0.526	0.576	0.533

ตารางที่ 6 การทดสอบแบบเปรียบเทียบ 5 ประโยค (ต่อ)

เอกสารที่	Non - Threshold		Threshold	
	ROUGE-N	ROUGE-L	ROUGE-N	ROUGE-L
28	0.527	0.497	0.641	0.525
29	0.561	0.467	0.563	0.538
30	0.539	0.477	0.586	0.482
ค่าเฉลี่ย	**0.547**	**0.467**	**0.595**	**0.513**

จากผลการทดลองในตารางที่ 2 การเปรียบเทียบแบบทีละ 1 ประโยคจะเห็นว่าเมื่อทำการทดสอบแล้วค่าของ ROUGE-N และ ROUGE-L ที่ไม่มีการใช้ค่าเทรสโฮลด์มาใช้ในการคัดเลือกประโยคเพื่อเป็นการสรุปความจะมีค่าอยู่ที่ 0.467 และ 0.314 ตามลำดับ และค่า ROUGE-N และ ROUGE-L ที่มีการนำค่าเทรสโฮล์ดมาใช้ในการคัดเลือกประโยคเพื่อเป็นสรุปความมีค่าอยู่ที่ 0.507 และ 0.355 ตามลำดับ ซึ่งผลที่ได้ยังเป็นที่ไม่น่าพอใจ เพราะเนื่องจากรูปแบบของประโยคที่ตัดมาได้นั้น ไม่มีความสมเหตุสมผลจึงทำให้ค่าของ ROUGE-N และ ROUGE-L มีค่าค่อนข้างน้อย

ดังนั้นเราจึงได้ปรับปรุงกระบวนการโดยจากเดิมจะพิจารณาเพียง 1 ประโยค จึงทำการเพิ่มจำนวนประโยคที่ต้องการนำมาเป็นสรุปเป็น 2 ประโยค , 3 ประโยค , 4 ประโยค และ 5 ประโยค ตามลำดับ ซึ่งปรากฏว่าผลของการเพิ่มจำนวนของประโยคทำให้ค่าของ ROUGE-N และ ROUGE-L มีค่าเพิ่มมากขึ้นจากเดิมเป็น 0.595 และ 0.513 ตามลำดับ

เอกสารอ้างอิง

[1] วิณัฏฐา แสงสุข และ ฐิติพร ลิ้มแหลมทอง, ความรู้เบื้องต้นเกี่ยวกับกฎหมายทั่วไป. กรุงเทพมหานคร: สำนักพิมพ์มหาวิทยาลัยรามคำแหง, 2555.

[2] ธีระ สิงหพันธุ์, กฎหมายอาญาภาค 1. กรุงเทพมหานคร : สำนักพิมพ์มหาวิทยาลัยรามคำแหง, 2556

[4] สำนักงานราชบัณฑิตยสภา, “พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ.๒๕๕๔.” .

[5] สำนักงานศาลฎีกา, “ศาลฎีกา - The Supreme Court of Thailand.” .

[8] C. M. Ambrus *et al.*, “Treatment of lead poisoning with an immobilized chelator comparison with conventional therapy,” *Res. Commun. Mol. Pathol. Pharmacol.*, vol. 110, no. 3–4, pp. 253–263, 2001.

[10] C. Haruechaiyasak, S. Kongyoung, and M. Dailey, “A comparative study on thai word segmentation approaches,” *5th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. ECTI-CON 2008*, vol. 1, pp. 125–128, 2008, doi: 10.1109/ECTICON.2008.4600388.

[12] N. Durrani and S. Hussain, “Urdu word segmentation,” *NAACL HLT 2010 - Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist. Proc. Main Conf.*, no. June, pp. 528–536, 2010.

[13] W. Kunnu, N. Kaewrattanapat, E. Major, and I. M. Program, “The Automatic

Classification of Thai news by Similarity Method ."

[14] R. Zhao and K. Mao, "Fuzzy bag-of-words model for document representation," *IEEE Trans. fuzzy Syst.*, vol. 26, no. 2, pp. 794–804, 2017.

[15] R. A. García-Hernández, R. Montiel, Y. Ledeneva, E. Rendón, A. Gelbukh, and R. Cruz, "Text summarization by sentence extraction using unsupervised learning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5317 LNAI, pp. 133–143, 2008, doi: 10.1007/978-3-540-88636-512.

[20] K. Maher and M. S. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 4, pp. 1715–1720, 2016.

[22] A. Trotman, A. Puurula, and B. Burgess, "Improvements to BM25 and language models examined," *ACM Int. Conf. Proceeding Ser.*, vol. 27-28-Nove, pp. 58–65, 2014, doi: 10.1145/2682862.2682863.

[25] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[26] C. Grover, B. Hachey, I. Hughson, and C. Korycinski, "Automatic summarisation of legal documents," *Proc. Int. Conf. Artif. Intell. Law*, pp. 243–251, 2003, doi: 10.1145/1047788.1047839.

[27] B. Hachey and C. Grover, "Sentence Classification Experiments for Legal Text Summarisation," *Leg. Knowl. Inf. Syst. Jurix 2004, Seventeenth Annu. Conf.*, no. May, pp. 29–38, 2004.

[29] B. Hachey and C. Grover, "Automatic legal text summarisation: Experiments with summary structuring," *Proc. Int. Conf. Artif. Intell. Law*, no. May, pp. 75–84, 2005, doi: 10.1145/1165485.1165498.