บทที่ 4
การพัฒนาซอฟต์แวร์เพื่อวัดความคล้ายคลึงของเอกสารภาษาไทย

การพัฒนาซอฟต์แวร์เพื่อวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้หลักการประมวลผลภาษาธรรมชาติ ขั้นตอนของการพัฒนางานการค้นคว้าแบบอิสระ เริ่มจากกระบวนการเตรียมข้อมูลเอกสาร ซึ่งจะเป็นการตัดคำ การกำจัดคำหยุด และการนับความถี่ของคำในเอกสาร เพื่อจัดรูปแบบของเอกสารให้อยู่ในรูปแบบของแบบจำลองเวกเตอร์สเปซ จากนั้นจึงนำเอกสารที่ผ่านกระบวนการเตรียมข้อมูลแล้วไปทำการหาค่าความถี่ของคำและหาค่าส่วนกลับความถี่ของเอกสาร แล้วจึงทำการวัดความคล้ายคลึงของเอกสารด้วยวิธีการวัดความคล้ายคลึงเชิงมุมโคไซน์ ทั้งนี้เพื่อให้เข้าใจถึงขั้นตอนการทำงานของระบบงานจึงขออธิบายดังรายละเอียดต่อไปนี้

4.1 กระบวนการเตรียมเอกสาร (Pre-Processing)

กระบวนการเตรียมเอกสารเป็นขั้นตอนการเตรียมเอกสารก่อนการประมวลผลให้อยู่ในรูปแบบที่คอมพิวเตอร์เข้าใจในภาษาธรรมชาติของมนุษย์ การค้นคว้าแบบอิสระนี้จะเน้นการประมวลกับเอกสารภาษาไทย ซึ่งปัญหาสำคัญของการประมวลผลเกี่ยวกับเอกสารภาษาไทยมีอยู่ 3 ประการ คือ

(1) ความกำกวมในเรื่องขอบเขตของคำ เนื่องจากลักษณะของภาษาไทยที่มีการเขียนคำที่เรียงต่อกันไปเป็นประโยคโดยที่ไม่มีเครื่องหมายแสดงขอบเขตของคำ วลี หรือประโยค เมื่อเปรียบเทียบกับภาษาอื่น ๆ แล้ว หลายภาษาในโลกนี้จะมีจุดฟลูสตอป (Full Stop) บอกจุดจบของประโยค เช่น ภาษาอังกฤษ ภาษาฝรั่งเศส ภาษาญี่ปุ่นและภาษาจีน หรือมีช่องว่าง (Space) บอกขอบเขตของคำ เช่น ภาษาอังกฤษและภาษาฝรั่งเศส หรือมีจุลภาค (Comma) บอกขอบเขตของวลี เช่น ภาษาอังกฤษ ภาษาฝรั่งเศส และภาษาญี่ปุ่น เป็นต้น

(2) หน้าที่คำที่กำกวม เพราะระบบการสร้างคำในภาษาไทย ซึ่งมีวิธีการในการสร้างคำอยู่หลายวิธีดังที่ได้กล่าวไว้ในบทที่ 2 ไม่ว่าจะเป็นการประสมคำ การซ้อนคำ หรือการซ้ำคำ เพื่อให้ได้คำใหม่ มาใช้เพิ่มเติมในภาษาและตอบสนองต่อความจำเป็นในการสื่อสาร โดยเฉพาะคำประสมนั้นถือว่าเป็นปัญหามาก เนื่องจากเกิดการนำคำมูลอย่างน้อย 2 คำมาประกอบกัน ทำให้ยากต่อการระบุว่าคำมูลที่พบนั้นเป็นคำ ๆ เดียวหรือเป็นส่วนหนึ่งของคำประสม เช่น แม่น้ำ พ่อตา แขกยาม และนาฬิกาทราย เป็นต้น รวมถึงโครงสร้างประโยคในภาษาไทยซับซ้อน บางส่วนของประโยคในภาษาไทยสามารถจะละได้ เช่น คำว่า “ถูก” (Passive Voice) หรือ บางครั้งกรรมของประโยคย้ายมา

อยู่ด้านหน้าได้ หรือวลีต่าง ๆ ของประโยคสามารถสลับตำแหน่งได้ ประกอบกับภาษาไทยไม่มีสัญลักษณ์ระบุขอบเขตของประโยคจึงทำให้เกิดความกำกวมว่าวลีนั้นควรขยายส่วนนำของวลีนั้นหรือขยายตามส่วนของวลีนั้น

(3) ความหมายของคำ สาเหตุนี้ถือเป็นสาเหตุของปัญหาที่สำคัญที่สุดในงานด้านการตัดคำและทำให้งานวิจัยด้านนี้ยังคงได้รับการพัฒนาอย่างต่อเนื่อง กล่าวคือ มนุษย์สามารถรับรู้ได้ว่าสายอักขระใดถือเป็นคำก็ต่อเมื่อ สายอักขระนั้นสื่อความหมายอย่างใดให้เป็นที่เข้าใจได้ แต่ความหมายเป็นสิ่งที่ซับซ้อน และไม่ปรากฏออกมาให้สังเกตได้ชัดเจน เป็นเพียงความเข้าใจในเชิงลึกจากผู้ที่ใช้ภาษานั้น ๆ มีอยู่ร่วมกัน กระนั้นก็ดี แม้มนุษย์ด้วยกันเองก็ตามบางครั้งก็ให้คำตอบในการตัดคำที่ต่างกันออกไปอันเนื่องมาจากปัจจัยหลาย ๆ ประการ เช่น ประสบการณ์ บริบท คำแวดล้อม และการตีความ เป็นต้น

ปัจจัยทั้ง 3 ข้อ ข้างต้นเป็นเครื่องบ่งชี้ได้อย่างชัดเจนว่าการประมวลผลภาษาธรรมชาติเป็นสิ่งที่ละเอียดอ่อนและลึกซึ้งโดยเฉพาะในภาษาไทย การจะหาเกณฑ์ที่บอกขอบเขตของคำให้แน่นอนตายตัวคงเป็นเรื่องที่ลำบาก แม้แต่คนไทยซึ่งเป็นเจ้าของภาษาก็ยังไม่สามารถตกลงเห็นพ้องต้องกันได้เสมอไปว่าสายอักขระที่ปรากฏนั้นเป็นหนึ่งคำหรือไม่ ดังนั้นที่ทำให้คอมพิวเตอร์ซึ่งเป็นเครื่องกลสามารถตัดคำได้อย่างถูกต้องทุกคำ โดยเลียนแบบกระบวนการทางคิดของนักภาษาศาสตร์จึงเป็นสิ่งที่ยาก เพราะคอมพิวเตอร์ไม่มีทั้งประสบการณ์สั่งสมและทักษะในการวิเคราะห์บริบทหรือตีความ

สำหรับการค้นคว้าแบบอิสระนี้ เอกสารที่นำมาใช้ในการทดสอบจะเก็บข้อมูลมาจากบทความบนอินเทอร์เน็ต ซึ่งจะแบ่งชุดเอกสารออกเป็น 5 ชุด คือ

(1) เอกสารชุดที่ 1 จะเป็นเอกสารตั้งต้น 30 เอกสาร โดยที่แต่ละเอกสารจะมีจำนวนคำทั้งหมด 300 คำ เนื้อหาในเอกสารจะประกอบด้วย

- บทความด้านการเมือง 10 เอกสาร
- บทความด้านกีฬา 10 เอกสาร
- บทความด้านสุขภาพ 10 เอกสาร

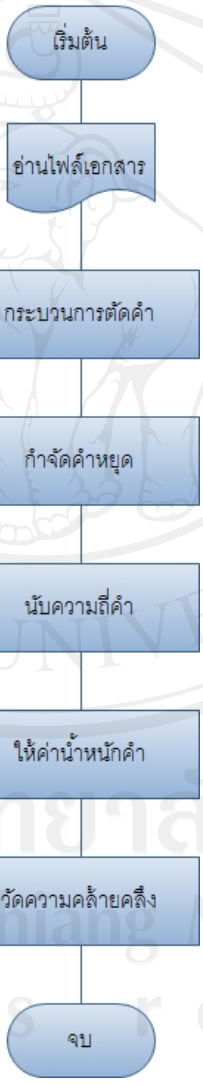
(2) เอกสารชุดที่ 2 จะเอาเอกสารจากเอกสารชุดที่ 1 มาทำการตัดคำทิ้งไป 10% จำนวน 30 เอกสาร

(3) เอกสารชุดที่ 3 จะเอาเอกสารจากเอกสารชุดที่ 1 มาทำการตัดคำทิ้งไป 50% จำนวน 30 เอกสาร

(4) เอกสารชุดที่ 4 จะเอาเอกสารจากเอกสารชุดที่ 1 ทำการผสมเนื้อหาระหว่าง 2 บทความ จำนวน 30 เอกสาร

(5) เอกสารชุด 5 จะเป็นเอกสารที่มีเนื้อหาแตกต่างไปจากเอกสารชุดที่ 1 จำนวน 30 เอกสาร ซึ่งเนื้อหาจะมาจากหลาย ๆ กลุ่มบทความ

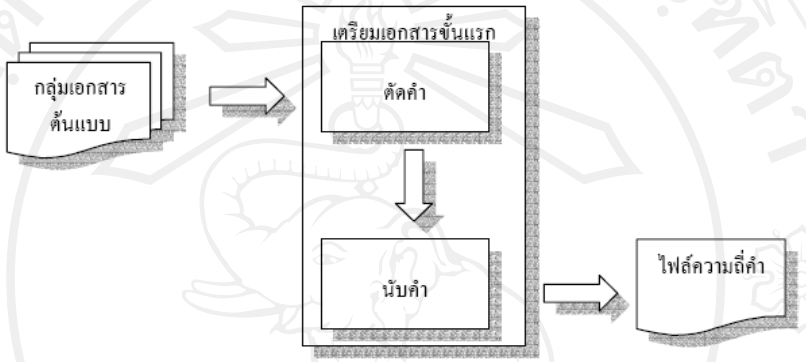
การทำงานของซอฟต์แวร์จะเริ่มจากการอ่านไฟล์เอกสารที่มีในระบบ จากนั้นจะนำเอกสารที่อ่านเข้าสู่กระบวนการตัดคำ แล้วจะทำการกำจัดคำหยุดออก จากนั้นจึงทำการนับความถี่ของคำในเอกสาร ทำการให้ค่าน้ำหนักคำ แล้วจึงจะเข้าสู่กระบวนการวัดค่าความคล้ายคลึง ซึ่งจะแสดงได้ดังรูป 4.1



รูป 4.1 แสดงขั้นตอนการทำงานของซอฟต์แวร์

4.2 การประมวลผลเอกสาร (Document Processing)

ในกระบวนการประมวลผลเอกสารจะเป็นกระบวนการในการกำหนดค่าน้ำหนักคำที่จะใช้เป็นดัชนีคำศัพท์ของเอกสาร คำเหล่านี้จะต้องถูกกำหนดและบันทึกเก็บไว้ โดยจะต้องสกัดคำเหล่านี้รวมถึงคำหยุดออกมาจากเอกสาร การสกัดคำเหล่านี้จากเอกสารจะต้องผ่านกระบวนการต่าง ๆ เพื่อให้ได้คำซึ่งจะนำมาเป็นตัวชี้หรือดัชนีคำศัพท์ของเอกสาร ดังรูป 4.2



รูป 4.2 ขั้นตอนการเตรียมเอกสารในการประมวลผลขั้นแรก

จากรูป 4.2 จะเห็นว่าขั้นตอนการเตรียมเอกสารในการประมวลผล ซึ่งการเตรียมเอกสารขั้นแรกจะประกอบด้วยขั้นตอน 2 ขั้นตอน คือ

(ก) การตัดคำ

ขั้นตอนนี้จะเป็นขั้นตอนการอ่านเอกสารจากเอกสารไฟล์ข้อความที่ได้เตรียมไว้เข้ามาทำการตัดคำจากรูปประโยคออกเป็นหน่วยของคำ โดยใช้โปรแกรมตัดคำ CMU WordSeg ที่จะตัดคำโดยอาศัยสถิติการเกิดตัวอักษร 3 ตัวที่ยาวติดต่อกัน (3 Grams) จากนั้นจะทำการกำจัดคำหยุด (Stop Word Removal) หรือคำที่ปรากฏในเอกสารบ่อยครั้งแต่ไม่ได้เป็นประเด็นสำคัญในเอกสาร ซึ่งประเภทของคำที่จัดว่าเป็นคำหยุดในภาษาไทยมีดังต่อไปนี้

1) คำบุพบท (Preposition) ตัวอย่างของคำบุพบทที่เป็นคำหยุดได้แก่ ของ ใน แก่ แต่ ต่อ ตั้งแต่ โดย เมื่อ กว่า กับ เป็น ดูก่อน ช้าแต่ ทาง สู่ แค่ เป็นต้น

2) คำสันธาน (Conjunction) ตัวอย่างของคำสันธานที่เป็นคำหยุดได้แก่ เพราะ เพราะว่า และ หรือ จึง ดังนั้น มิฉะนั้น ทั้ง แต่ แต่ว่า ครั้น หรือไม่ก็ เป็นต้น

3) คำสรรพนาม (Pronoun) ตัวอย่างของคำสรรพนามที่เป็นคำหยุดได้แก่ ฉัน เรา เขา ดิฉัน กระผม กู คุณ ท่าน เธอ ใต้เท้า มัน ใคร ตัวนั้น อันนั้น เป็นต้น

4) คำวิเศษณ์ (Adverb) ตัวอย่างของคำวิเศษณ์ที่เป็นคำหยุดได้แก่ มาก น้อย ใหญ่ เล็ก อ้วน โต สูง หลาย เยอะ หอม นุ่ม เผ็ด นั้น เอง ทั้งนี้ ค่ะ ครับ ขอรับ จ้า ทำไม เป็นต้น

5) คำอุทาน (Interject) ตัวอย่างของคำอุทานที่เป็นคำหยุดได้แก่ เอ๊ะ อ๊ะ อ๋อ โธ่ โถ อนิจจา ตายละ โอ้โห เป็นต้น

ในการกำจัดคำหยุดนี้จะรวมถึงสัญลักษณ์พิเศษต่าง ๆ เช่น เครื่องหมายคำพูด (“ ”) หรือ เครื่องหมายคำถาม (?) และเครื่องหมายที่ใช้ในการคำนวณต่าง ๆ เป็นต้น

(ข) การนับความถี่ของคำ

ในขั้นตอนนี้จะเป็นขั้นตอนการนับความถี่ของคำที่ปรากฏในเอกสาร แล้วสร้างไฟล์ใหม่ซึ่งประกอบไปด้วยคำต่าง ๆ และจำนวนคำที่ปรากฏในเอกสารแต่ละเอกสาร เพื่อที่จะนำไปคำนวณต่อไปในขั้นตอนต่าง ๆ ในขั้นตอนนี้นอกจากจะทำการเก็บคำและจำนวนคำแล้ว ยังต้องเก็บจำนวนเอกสารที่พบคำคำนั้นอีกด้วย เพื่อใช้สำหรับขั้นตอนการคำนวณค่าน้ำหนักคำของเอกสารในขั้นตอนต่อไป

4.2.1 การหาค่าส่วนกลับความถี่ของเอกสาร

ในขั้นตอนนี้จะเป็นกระบวนการที่จะกระทำภายหลังจากที่ผ่านกระบวนการในหัวข้อ (ข) จากนั้นจะนำมาหาจำนวนเอกสารที่มีคำ ๆ นั้นปรากฏอยู่เพื่อนำไปสู่การหาค่าน้ำหนักคำรวมถึงการคำนวณค่าส่วนกลับความถี่ของเอกสาร ตามสมการ (3.1) เพื่อจะนำค่าส่วนกลับความถี่ของเอกสารไปใช้ในการคำนวณหาค่าน้ำหนักคำแต่ละคำในแต่ละเอกสารตามสมการ (3.2) คือ ค่าน้ำหนักคำนั้น ๆ ที่ปรากฏอยู่ในแต่ละเอกสาร ซึ่งตัวอย่างการคำนวณค่าน้ำหนักคำสามารถแสดงได้ดังตัวอย่าง 4.1

ตัวอย่าง 4.1 ตัวอย่างการเตรียมเอกสารก่อนนำเข้าสู่การประมวลผล สมมุติว่าในคลังเอกสาร (Document Collection) มีเอกสาร 4 ฉบับ ได้แก่

เอกสาร 1 ความคิดเป็นจุดเริ่มต้นของความรู้

เอกสาร 2 ความรู้อาจทำให้คนเราเปลี่ยนแปลงความคิด

เอกสาร 3 คนเราต้องปรับปรุงในสิ่งที่เคยทำผิดเพื่อใครอีกคน

เอกสาร 4 วันนี้เธอบอกเขาว่าไปกินข้าวร้านเดิมไม่อร่อย

นำคำทั้งหมดที่อยู่ในเอกสารทั้ง 4 เอกสาร มาตัดคำตามหัวข้อ (ก) ซึ่งจะได้คำทั้งหมดที่อยู่ในเอกสาร จากนั้นทำการนับความถี่ของคำทั้งหมดที่อยู่ในเอกสารทั้งหมด จะได้ตามตาราง 4.1

ตาราง 4.1 ผลลัพธ์จากการหาความถี่ของคำและค่าความถี่ส่วนกลับของเอกสาร

ลำดับ	Term Word	D1	D2	D3	D4	df	D/df	IDF
1	ความคิด	1	1	0	0	2	4/2	0.301
2	จุดเริ่มต้น	1	0	0	0	1	4/1	0.602
3	ความรู้	1	0	0	0	1	4/1	0.602
4	รู้	0	0	0	0	1	4/1	0.602
5	คนเรา	0	1	1	0	2	4/2	0.301
6	ปรับปรุง	0	0	1	0	1	4/1	0.602
7	คน	0	0	1	0	1	4/1	0.602
8	กิน	0	0	0	1	1	4/1	0.602
9	ข้าว	0	0	0	1	1	4/1	0.602
10	ร้าน	0	0	0	1	1	4/1	0.602
11	อร่อย	0	0	0	1	1	4/1	0.602

จากตาราง 4.1 ค่าส่วนกลับความถี่ของเอกสารคำนวณได้จากสมการ $idf = \log\left(\frac{N}{df}\right)$ โดยอ้างอิงจากทฤษฎีที่กล่าวมาแล้วในบทที่ 3 จากนั้นจะหาค่าน้ำหนักคำในเอกสาร ตามสมการ (3.2) ในบทที่ 3 จะได้ผลลัพธ์ตามตาราง 4.2

ตาราง 4.2 ผลลัพธ์ที่ได้จากการหาค่าน้ำหนักคำในเอกสาร

ลำดับคำ	D1	D2	D3	D4
1	0.301	0.301	0	0
2	0.602	0	0	0
3	0.602	0	0	0
4	0	0.602	0	0
5	0	0.301	0.301	0
6	0	0	0.602	0
7	0	0	0.602	0
8	0	0	0	0.602
9	0	0	0	0.602
10	0	0	0	0.602
11	0	0	0	0.602

4.3 การวัดความคล้ายคลึงเชิงมุมโคไซน์ (Cosine Similarity Measures)

ขั้นตอนการวัดความคล้ายคลึงเชิงมุมโคไซน์นั้น เราจะคำนวณหลังจากได้ค่าจากขั้นตอนที่ 4.2.1 โดยเทอมที่มีค่าเป็น 0 จะไม่นำมาใช้ในการคำนวณ และจะคำนวณตามสมการ (3.3) ที่กล่าวในบทที่ 3

$$|D1| = \sqrt{0.301^2 + 0.602^2 + 0.602^2}$$
$$= \sqrt{0.815}$$
$$|D2| = \sqrt{0.301^2 + 0.602^2 + 0.301^2}$$
$$= \sqrt{0.544}$$
$$|D3| = \sqrt{0.301^2 + 0.602^2 + 0.602^2}$$
$$= \sqrt{0.815}$$
$$|D4| = \sqrt{0.602^2 + 0.602^2 + 0.602^2 + 0.602^2}$$
$$= \sqrt{1.450}$$

เพราะฉะนั้นค่า $|D_i| = \sqrt{\sum_i W_{i,j}^2}$

จากนั้นจะคำนวณค่าผลคูณสเกลาร์ (Dot Product) ระหว่างเวกเตอร์ เทอมที่มีค่าเป็น 0 จะไม่นำมาใช้ในการคำนวณ

$$D1 \bullet D1 = (0.301 * 0.301) + (0.602 * 0.602) + (0.602 * 0.602)$$
$$= 0.815$$

$$D1 \bullet D2 = (0.301 * 0.301)$$
$$= 0.091$$

$$D1 \bullet D3 = 0$$

$$D1 \bullet D4 = 0$$

$$D2 \bullet D2 = (0.301 * 0.301) + (0.602 * 0.602) + (0.301 * 0.301)$$
$$= 0.544$$

$$D2 \bullet D3 = (0.301 * 0.301)$$
$$= 0.091$$

$$D2 \bullet D4 = 0$$

$$D3 \bullet D3 = (0.301 * 0.301) + (0.602 * 0.602) + (0.602 * 0.602)$$
$$= 0.815$$

$$D3 \bullet D4 = 0$$

$$D4 \bullet D4 = (0.602 * 0.602) + (0.602 * 0.602) + (0.602 * 0.602) + (0.602 * 0.602)$$
$$= 1.450$$

เพราะฉะนั้นจะได้ค่า $D \bullet D_i = \sum_i W_{d,j} W_{i,j}$

จากนั้นจะทำการวัดความคล้ายคลึงเชิงมุมโคไซน์จาก

$$\text{Cosine } D_i = Sim(D, D_i)$$

$$Sim(D, D_i) = \frac{\sum_i W_{d,j} W_{i,j}}{\sqrt{\sum_j W_{d,j}^2} \sqrt{\sum_i W_{i,j}^2}}$$

จะได้ค่าการวัดความคล้ายคลึง ดังนี้

$$Sim(D_1, D_1) = \frac{0.815}{\sqrt{0.815} * \sqrt{0.815}}$$
$$= 1$$

$$Sim(D_1, D_2) = \frac{0.091}{\sqrt{0.815} * \sqrt{0.544}}$$
$$= 0.136$$

$$Sim(D_1, D_3) = \frac{0}{\sqrt{0.815} * \sqrt{0.815}}$$
$$= 0$$

$$Sim(D_1, D_4) = \frac{0}{\sqrt{0.815} * \sqrt{1.450}}$$
$$= 0$$

$$Sim(D_2, D_2) = \frac{0.544}{\sqrt{0.544} * \sqrt{0.544}}$$
$$= 1$$

$$Sim(D_2, D_3) = \frac{0.091}{\sqrt{0.544} * \sqrt{0.816}}$$

$$= 0.136$$

$$Sim(D_2, D_4) = \frac{0}{\sqrt{0.544} * \sqrt{1.450}}$$

$$= 0$$

$$Sim(D_3, D_3) = \frac{0.815}{\sqrt{0.815} * \sqrt{0.815}}$$

$$= 1$$

$$Sim(D_3, D_4) = \frac{0}{\sqrt{0.815} * \sqrt{1.450}}$$

$$= 0$$

$$Sim(D_4, D_4) = \frac{1.450}{\sqrt{1.450} * \sqrt{1.450}}$$

$$= 1$$

4.4 การประเมินผลซอฟต์แวร์การวัดความคล้ายคลึงของเอกสาร

การประเมินผลการวัดความคล้ายคลึงของเอกสารจะใช้ค่าความแม่นยำของข้อมูล และค่าความถูกต้องของข้อมูลที่เป็นค่ามาตรฐานที่ใช้ในการประเมินประสิทธิภาพมีรายละเอียดดังนี้

4.4.1 วิธีทดสอบประสิทธิภาพของซอฟต์แวร์

สำหรับวิธีทดสอบประสิทธิภาพของซอฟต์แวร์ ผู้ค้นคว้ากำหนดว่าเอกสารที่ “คล้ายคลึง” คือ เอกสารที่ผู้ใช้ควรจะได้รับเมื่อสั่งให้ซอฟต์แวร์ทำการเปรียบเทียบเอกสารที่มีอยู่ในระบบทั้งหมด

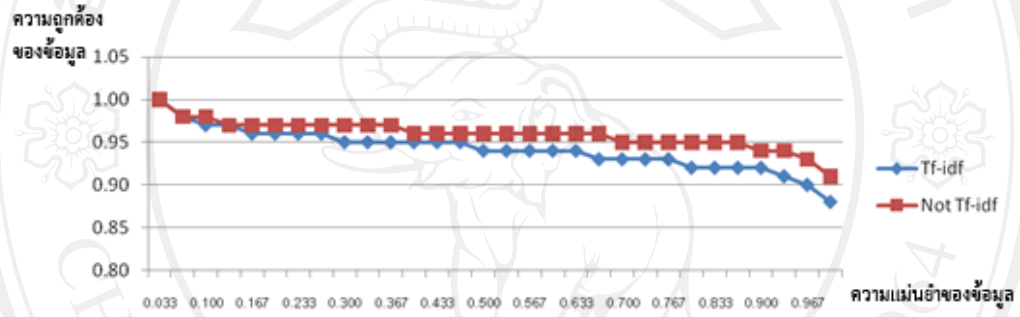
การทดสอบประสิทธิภาพจะใช้วิธีการวัดค่าความถูกต้องที่ระดับค่าความแม่นยำต่าง ๆ โดยจะกำหนดการทดสอบเป็น 3 ส่วนดังนี้

ตาราง 4.3 รายละเอียดขั้นตอนการทดสอบประสิทธิภาพ

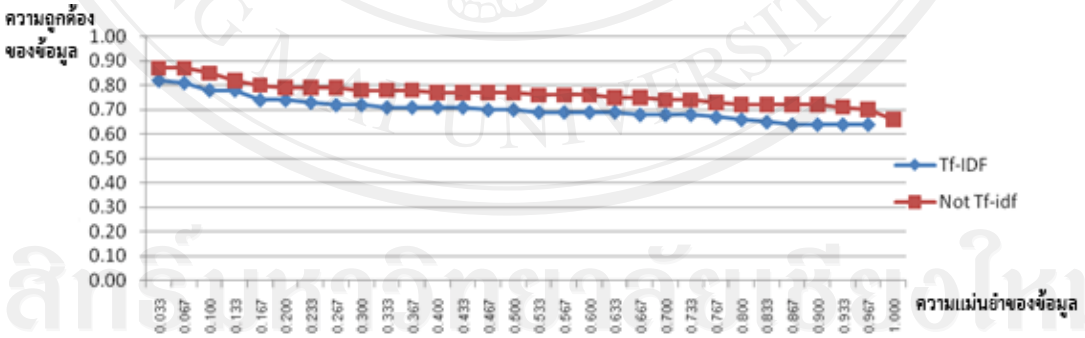
กระบวนการ	วิธีการ
การเตรียมข้อมูลเอกสาร	1. ตัดคำด้วยเครื่องหมาย “|” 2. กำจัดคำที่ไม่มีนัยสำคัญออก 3. แทนเอกสารด้วยระบบเวกเตอร์ 4. คิดค่าถ่วงน้ำหนักด้วยวิธีการให้ค่าน้ำหนักคำ
การวัดความคล้ายคลึง	วัดความคล้ายคลึงเชิงมุม
การวัดประสิทธิภาพ	วัดค่าความถูกต้องที่ระดับค่าความแม่นยำต่าง ๆ

4.4.2 ผลการทดสอบประสิทธิภาพของซอฟต์แวร์

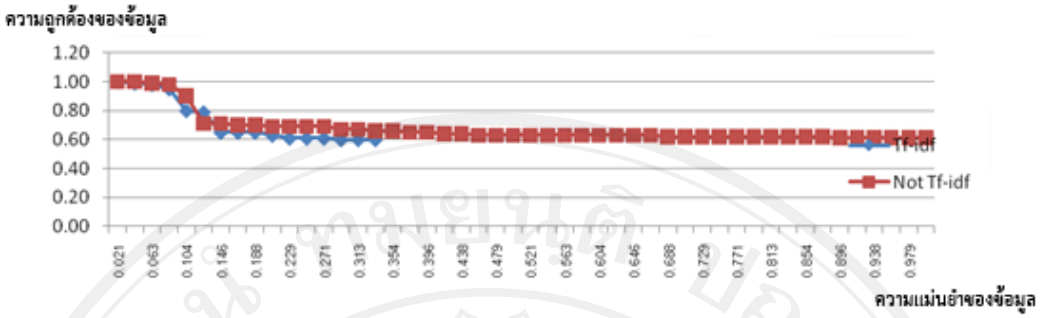
ผลการทดสอบประสิทธิภาพของซอฟต์แวร์การวัดความคล้ายคลึงของเอกสาร จะแสดงให้เห็นตามรูป ซึ่งการทดสอบประสิทธิภาพของซอฟต์แวร์ จะทำการเปรียบเทียบผลของการวัดความคล้ายคลึงของเอกสาร 2 แบบ คือ การวัดความคล้ายคลึงแบบมีการให้ค่าน้ำหนักคำก่อนทำการวัดความคล้ายคลึง และการวัดความคล้ายคลึงแบบไม่ให้ค่าน้ำหนักคำก่อนทำการวัดความคล้ายคลึง โดยเส้นกราฟสีแดงจะเป็นการวัดประสิทธิภาพของการวัดความคล้ายคลึงของเอกสารที่ไม่มีการให้ค่าน้ำหนักคำ ส่วนเส้นกราฟสีฟ้าเป็นการวัดประสิทธิภาพของการวัดความคล้ายคลึงของเอกสารที่ให้ค่าน้ำหนักคำก่อนทำการวัดความคล้ายคลึง ได้ผลจากการทดสอบตามรูป 4.3 รูป 4.4 และรูป 4.5



รูป 4.3 ผลการวัดประสิทธิภาพของการวัดความคล้ายคลึงระหว่างเอกสารชุด 1 และ 2



รูป 4.4 ผลการวัดประสิทธิภาพของการวัดความคล้ายคลึงระหว่างเอกสารชุด 1 และ 3



รูป 4.5 ผลการวัดประสิทธิภาพของการวัดความคล้ายคลึงระหว่างเอกสารชุด 1 และ 4

สำหรับเอกสารชุดที่ 5 เมื่อทำการวัดค่าความคล้ายคลึงแล้ว พบว่าไม่มีเอกสารใดมีความคล้ายคลึงกันเลย

จากรูปแสดงค่าความถูกต้องของเอกสารที่ได้รับ ณ ค่าความแม่นยำต่าง ๆ โดยที่ค่าความแม่นยำของข้อมูลและค่าความถูกต้องของข้อมูลในการทดสอบนี้จะคำนวณจาก

ค่าความถูกต้องของข้อมูล = จำนวนเอกสารที่โปรแกรมอ่านได้และมีความคล้ายคลึงกัน / จำนวนเอกสารทั้งหมดในกลุ่มที่คล้ายคลึงกัน

ค่าความแม่นยำของข้อมูล = จำนวนเอกสารที่โปรแกรมอ่านได้และมีความคล้ายคลึงกัน / จำนวนเอกสารที่โปรแกรมอ่านได้ทั้งหมด

ผลของค่าความแม่นยำที่ได้จากการประเมินนั้น พบว่าหลักการวัดความคล้ายคลึงที่มีการให้ค่าน้ำหนักคำ ให้ผลการวัดความคล้ายคลึงดีกว่าการวัดความคล้ายคลึงโดยไม่ให้ค่าน้ำหนักคำทุกกรณี

สำหรับการวัดความคล้ายคลึงที่ใช้แบบจำลองเวกเตอร์สเปซทำการวัดความคล้ายคลึงโดยไม่มีการให้ค่าน้ำหนักคำก่อนทำการวัดความคล้ายคลึงนั้น จะได้ผลการวัดความคล้ายคลึงที่มากกว่าการวัดความคล้ายคลึงโดยให้ค่าน้ำหนักคำก่อนทำการวัด เนื่องจากในขั้นตอนการให้ค่าน้ำหนักคำนั้น โปรแกรมจะนำเฉพาะคำที่มีค่าน้ำหนักคำในเทอมเดียวกันของระหว่าง 2 เอกสารมาใช้ในการคำนวณ ทำให้เอกสารที่มีองค์ประกอบของคำที่แตกต่างกัน เมื่อมีการให้ค่าน้ำหนักคำแล้วทำการวัดความคล้ายคลึงนั้น ได้ผลจากการวัดความคล้ายคลึงจึงได้ค่าความคล้ายคลึงน้อยกว่าแบบไม่ให้ค่าน้ำหนักคำ แสดงให้เห็นว่าวิธีการให้ค่าน้ำหนักคำก่อนทำการวัดความคล้ายคลึงนั้นสามารถจัดลำดับของเอกสารที่มีความคล้ายคลึงกันได้มากกว่า

ส่วนการสลับตำแหน่งของคำนั้นจะมีผลทำให้โปรแกรมมองคำ ๆ เดียวกัน เป็นคนละคำ เนื่องจากคำ ๆ หนึ่ง อย่างเช่น “ความก้าวหน้า” เมื่อมีการสลับตำแหน่งของคำเป็น “ก้าวความหน้า” โปรแกรมจะมองคำ 2 คำ นี้เป็นคำที่มีองค์ประกอบที่แตกต่างกัน เนื่องจากผลจากการตัดคำที่ได้คือ “ความ|ก้าวหน้า” และ “ก้าว|ความ|หน้า” ทำให้โปรแกรมมองคำทั้ง 2 คำ เป็นคำที่มีองค์ประกอบของที่แตกต่างกัน ส่งผลให้การนับความถี่ของคำในเอกสารแตกต่างกันด้วย