บทที่ 5
สรุปผลการศึกษา และข้อเสนอแนะ

ในบทนี้จะกล่าวถึงการสรุปผลการศึกษาการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ ซึ่งมีรายละเอียดของการปฏิบัติงานในลักษณะต่าง ๆ ดังต่อไปนี้

5.1 สรุปผลการค้นคว้า

การค้นคว้าแบบอิสระเรื่อง การวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาตินี้ ผู้ศึกษาได้ศึกษาค้นคว้าทำความเข้าใจและมีการดำเนินงานโดยเริ่มจากการศึกษาหลักการวัดความคล้ายคลึง การนับความถี่ของคำในเอกสาร การให้ค่าน้ำหนักคำ และการวัดความคล้ายคลึงของเอกสารโดยใช้การวัดความคล้ายคลึงเชิงมุมโคไซน์

การทำงานของซอฟต์แวร์จะแบ่งออกเป็น 2 ส่วน คือ ในส่วนของการประมวลผลจะแบ่งย่อยออกเป็น 2 ขั้นตอน คือ ขั้นตอนการเตรียมเอกสารก่อนการประมวลผล และขั้นตอนการประมวลผล ในขั้นตอนการเตรียมเอกสารก่อนการประมวลผลนั้น เอกสารที่นำมาประมวลผลจะถูกเตรียมให้อยู่ในรูปแบบของแบบจำลองเวกเตอร์สเปซ ซึ่งเอกสารจะต้องผ่านกระบวนการตัดคำ การนับความถี่ของคำในเอกสาร เพื่อเตรียมเอกสารเข้าสู่ขั้นตอนการประมวลผล และในขั้นตอนการประมวลผลจะเป็นขั้นตอนในการให้ค่าน้ำหนักคำแล้วจึงทำการวัดความคล้ายคลึงของเอกสาร สำหรับในส่วนของการพัฒนาซอฟต์แวร์จะทำการพัฒนาซอฟต์แวร์โดยใช้ภาษาซีชาร์ปดอทเน็ต

การวัดประสิทธิภาพของซอฟต์แวร์การวัดความคล้ายคลึงของเอกสารนั้น จะอาศัยการวัดค่าความถูกต้องของข้อมูล และการวัดค่าความแม่นยำของข้อมูล ณ ที่ค่าความครบถ้วนต่าง ๆ

5.2 ข้อจำกัดของระบบ

(1) การวัดความคล้ายคลึงของเอกสารจะอาศัยจำนวนความถี่ของคำและการให้ค่าน้ำหนักคำในเอกสารเป็นหลัก ดังนั้นถ้าเอกสารที่มีเนื้อหาคล้ายคลึงกัน แต่มีองค์ประกอบของคำในเอกสารไม่เหมือนกัน การวัดความคล้ายคลึงของเอกสารจะไม่สามารถทำได้เต็มประสิทธิภาพ

(2) การวัดความคล้ายคลึงของเอกสารโดยมีการให้ค่าน้ำหนักคำก่อนทำการวัดความคล้ายคลึงนั้น เหมาะสำหรับเอกสารที่มีองค์ประกอบของคำที่เหมือน ๆ กัน และเหมาะสำหรับใช้ในการวัดความคล้ายคลึงของเอกสารเท่านั้น

(3) การให้ค่าน้ำหนักคำก่อนทำการวัดความคล้ายคลึงของเอกสารนั้น เหมาะสำหรับเอกสารที่มีองค์ประกอบของคำที่เหมือนหรือคล้ายคลึงกัน เพราะซอฟต์แวร์การวัดความคล้ายคลึงของเอกสารโดยการให้ค่าน้ำหนักคำก่อนทำการวัดความคล้ายคลึงนี้ จะไม่สนใจในเรื่องความหมายของคำ แต่จะมองว่าคำในเอกสารคือองค์ประกอบอย่างหนึ่ง ดังนั้นถ้าเอกสารสองเอกสารจะมีความคล้างคลึงกันก็จะต้องมีองค์ประกอบของคำที่เหมือนกัน

(4) ความเร็วของโปรแกรมการวัดความคล้ายคลึงของเอกสาร จะขึ้นกับจำนวนคำทั้งหมดในเอกสารเป็นสำคัญ เพราะถ้าเอกสารประกอบด้วยจำนวนคำมาก โปรแกรมจะต้องใช้เวลาในการอ่านและนับความถี่ของคำแต่ละคำในเอกสารซึ่งจะใช้เวลาระยะหนึ่ง

5.3 ปัญหาและอุปสรรค

(1) การตัดคำไม่ถูกต้อง หรือมีการสลับตำแหน่งของคำจะมีผลอย่างมากในการวัดความคล้ายคลึงของเอกสาร รวมถึงการเอาคำหยุดออก เพราะคำ ๆ หนึ่งถ้าปกติตัดได้ไม่เป็นคำหยุด แต่ถ้ามีการสลับตำแหน่งของคำแล้วทำให้พบว่าคำนั้นเป็นคำหยุดแล้วนำคำ ๆ นั้นออก ทำให้ผลการวัดความคล้ายคลึงของเอกสารที่มีการสลับตำแหน่งคำได้ค่าความคล้ายคลึงไม่เท่ากัน

(2) เนื้อหาของเอกสารที่นำมาเป็นเอกสารทดสอบมีการใช้ศัพท์แสลง หรือสำนวนเฉพาะมาก คำที่ใช้ในหมวดหมู่หนึ่งอาจมีความหมายที่แตกต่างออกไปสำหรับอีกหมวดหมู่หนึ่ง มีผลทำให้การวัดความคล้ายคลึงของเอกสารมีความคลาดเคลื่อนได้

5.4 ข้อเสนอแนะ

จากข้อจำกัดและปัญหาดังที่กล่าวมาข้างต้น ผู้ศึกษาได้แบ่งข้อเสนอแนะ ดังนี้

(1) มีฝึกหัด (Training) ให้ซอฟต์แวร์รู้จักคำที่เป็นคำเฉพาะมากขึ้น

(2) อาจเพิ่มการคำนวณค่าน้ำหนักคำโดยใช้ค่าน้ำหนักคำมาตรฐาน และใช้การคำนวณค่าความคล้ายคลึงรูปแบบอื่นมาช่วยในการวัดความคล้ายคลึงของเอกสาร เนื่องจากความหลากหลายของเอกสารและความหลากหลายของคำจะมีผลต่อการวัดความคล้ายคลึงของเอกสารที่จะนำไปใช้ในการจัดกลุ่มเอกสาร

(3) อาจเพิ่มการวิเคราะห์โครงสร้างประโยคภาษาไทยและการพิจารณาการปรากฏร่วมกันของคำในเอกสารก่อนที่จะทำการวัดความคล้ายคลึงของเอกสาร เพื่อให้การวัดความคล้ายคลึงของเอกสารภาษาไทยมีประสิทธิภาพมากขึ้น