# บทที่ 1
# บทนำ

การค้นคว้าแบบอิสระนี้ต้องการที่จะนำเสนอแนวคิดและกระบวนวิธีในการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ จึงต้องการแสดงให้เห็นถึงหลักการและทฤษฎีที่เกี่ยวข้อง เหตุผลและแรงจูงใจที่นำมาสู่การค้นคว้าแบบอิสระในครั้งนี้ พร้อมทั้งแสดงวัตถุประสงค์และประโยชน์ที่คาดว่าจะได้รับจากการศึกษาค้นคว้า นอกจากนี้ยังได้แสดงขอบเขตงานวิจัยและวิธีการวิจัยอย่างเป็นขั้นเป็นตอน เพื่อนำไปสู่การบรรลุผลสำเร็จในการวิจัย รวมถึงเครื่องมือและอุปกรณ์ที่มีความจำเป็นที่ต้องใช้ในการค้นคว้าแบบอิสระในครั้งนี้

## 1.1 หลักการ ทฤษฎี เหตุผล และ/หรือสมมุติฐาน

ยุคที่คอมพิวเตอร์และอินเทอร์เน็ตมีความเจริญก้าวหน้าไปอย่างรวดเร็ว เอกสารอิเล็กทรอนิกส์ในรูปแบบต่าง ๆ มีจำนวนเพิ่มมากขึ้นพร้อมกับความต้องการในการใช้ข้อมูลที่เพิ่มมากขึ้นตามไปด้วย ทำให้เกิดระบบที่ช่วยให้ผู้ใช้งานสามารถเข้าถึงฐานข้อมูลได้ง่าย เพื่อให้เกิดความสะดวกในการค้นหาเอกสารที่ต้องการได้อย่างรวดเร็ว ปัจจุบันได้มีผู้พัฒนาระบบเหล่านี้ให้มีความสามารถต่าง ๆ เพิ่มขึ้น นอกเหนือจากการสืบค้นเอกสารที่ปรากฏคำหรือกลุ่มคำที่ผู้ใช้ระบุเท่านั้น ความสามารถใหม่ ๆ ที่สร้างขึ้นเพื่ออำนวยความสะดวกให้แก่ผู้ใช้ เช่น การค้นหาเอกสารที่คล้ายคลึงกับเอกสารที่ผู้ใช้ให้เป็นตัวอย่าง (Querying by Example) หรือการจัดกลุ่มเอกสารที่ได้จากการสืบค้น (Retrieved Document Clustering) ตัวอย่างความสามารถของระบบที่เพิ่มขึ้นมานี้จำเป็นต้องใช้กระบวนการในการวัดความคล้ายคลึงระหว่างเอกสาร (Similarity Measure) เข้ามาช่วย ทำให้นักวิจัยและนักพัฒนาในปัจจุบันให้ความสนใจในเรื่องของการวัดความคล้ายคลึงระหว่างเอกสาร เพื่อนำมาช่วยเพิ่มประสิทธิภาพให้แก่งานด้านการจัดกลุ่มเอกสาร หรือการสืบค้นเอกสาร

แบบจำลองเวกเตอร์สเปซ (Vector Space Model) เป็นรูปแบบหนึ่งที่นิยมนำมาใช้ในการวัดความคล้ายคลึงระหว่างเอกสาร โดยแต่ละเอกสารจะแสดงในรูปแบบของเวกเตอร์ หรือชุดของคำที่มีการจัดเรียงลำดับ วิธีการประเมินค่าขึ้นอยู่กับเวกเตอร์ 0 - 1 ซึ่งแต่ละองค์ประกอบจะเป็น 0 ถ้าคำนั้น ๆ ไม่ปรากฏ หรือมีค่าเป็น 1 ถ้าคำนั้น ๆ ปรากฏในเอกสารตามที่ร้องขอ และการประเมินอีกทางหนึ่งอยู่บนพื้นฐานของเวกเตอร์และค่าน้ำหนักคำเป็นองค์ประกอบ ความสำเร็จในการใช้แบบจำลองเวกเตอร์สเปซอยู่ที่ความเข้ากันได้ในเชิงมิติ (Dimensional Compatibility) ของการ

เปรียบเทียบระหว่างเอกสารสองเอกสาร หรือระหว่างเอกสารกับคำขอ โดยจะอยู่บนพื้นฐานของการเปรียบเทียบคำที่เหมือนกันในแต่ละเอกสารเสมอ แบบจำลองเวกเตอร์สเปซมีข้อจำกัด คือ แบบจำลองเวกเตอร์สเปซจะไม่สนใจความหมายของคำ วลี โครงสร้างของคำ หรือคำที่มีความหมายเหมือนกัน ทำให้เอกสารที่มีองค์ประกอบเหมือนกัน และแต่ละองค์ประกอบมีความถี่ในการปรากฏของคำเท่ากัน จะได้เวกเตอร์สเปซที่มีลักษณะเหมือนกัน การใช้การคำนวณความคล้ายคลึงในลักษณะนี้ไม่สามารถที่จะเปรียบเทียบความคล้ายคลึงระหว่างเอกสารได้เพียงพอและจากปัญหาด้านข้อจำกัดของแบบจำลองเวกเตอร์สเปซ ทำให้มีการคิดค้นหาวิธีในการคำนวณความคล้ายคลึงระหว่างเอกสารให้มีความหลากหลายและยืดหยุ่นมากขึ้น โดยเริ่มจากการกำหนดค่าน้ำหนักให้กับคำในเวกเตอร์ พร้อมกับให้ความสำคัญกับความถี่ของคำที่ปรากฏ ผลที่ได้จากการวัดความคล้ายคลึงตามแนวคิดนี้มีความถูกต้องมากขึ้น และเป็นแรงผลักดันให้เกิดงานวิจัยหลาย ๆ งานที่ทำการพัฒนาแนวความคิดในการวัดความคล้ายคลึงของเอกสาร งานวิจัยในต่างประเทศหลายงานได้ทดลองนำวิธีการวัดความคล้ายคลึงเชิงมุมโคไซน์ (Cosine Similarity) ที่เป็นวิธีการวัดความคล้ายคลึงของเอกสารรูปแบบหนึ่งเข้ามาช่วย แต่ในประเทศไทยงานวิจัยในเรื่องของแนวคิดในการวัดความคล้ายคลึงของเอกสารภาษาไทยยังมีไม่มาก โดยเฉพาะแนวคิดในการนำหลักการของแบบจำลองเวกเตอร์สเปซมาใช้ร่วมกับการนับค่าความถี่ในการปรากฏของคำและการให้ค่าน้ำหนักคำ แล้วจึงทำการวัดความคล้ายคลึงด้วยวิธีการวัดความคล้ายคลึงเชิงมุมโคไซน์

การค้นคว้าแบบอิสระนี้มีวัตถุประสงค์เพื่อพัฒนาเครื่องมือที่ช่วยในการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยนำหลักการวัดความคล้ายคลึงเชิงมุมโคไซน์ การให้ค่าน้ำหนักคำมาใช้กับหลักการของแบบจำลองเวกเตอร์สเปซในการวัดความคล้ายคลึง และนำมาประยุกต์ใช้กับเอกสารภาษาไทยที่ทำการตัดคำภายใต้แนวทางของการประมวลผลภาษาธรรมชาติ

**1.2 วัตถุประสงค์ของการศึกษา**

การค้นคว้าแบบอิสระนี้มีวัตถุประสงค์เพื่อ

1) เพื่อพัฒนาเครื่องมือที่ช่วยในการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยนำหลักการวัดความคล้ายคลึงเชิงมุมโคไซน์ การให้ค่าน้ำหนักคำมาใช้กับหลักการของแบบจำลองเวกเตอร์สเปซในการวัดความคล้ายคลึงของเอกสาร

2) เพื่อศึกษาการจัดกลุ่มของเอกสารภาษาไทยที่มีความคล้ายคลึงกัน และทำการตัดคำภายใต้แนวความคิดการประมวลผลภาษาธรรมชาติ

**1.3 ประโยชน์ที่ได้รับจากการศึกษาเชิงทฤษฎีและ/ หรือเชิงประยุกต์**

จากการศึกษาค้นคว้าแบบอิสระนี้ คาดว่าจะได้รับประโยชน์ดังนี้

1) สามารถเปรียบเทียบความคล้ายคลึงของเอกสารได้

2) แนวคิดในการวัดความคล้ายคลึงของเอกสารสามารถนำไปปรับประยุกต์ใช้กับการจัดกลุ่มเอกสาร หรือการสืบค้นเอกสาร

3) ได้เครื่องมือที่ช่วยในการวัดความคล้ายคลึงของเอกสาร

**1.4 แผนการดำเนินงาน และขอบเขตของการศึกษา**

แผนการดำเนินงาน และขอบเขตของการศึกษาจะอธิบายถึงขั้นตอนการวางแผน หรือขั้นตอนดำเนินงาน และขอบเขตของการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ ดังนี้

**1.4.1 แผนการดำเนินงาน**

แผนการดำเนินงานการศึกษาการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ มีดังนี้

1) ศึกษาหลักของการจัดกลุ่มเอกสาร

2) ศึกษาหลักของแบบจำลองเวกเตอร์สเปซ

3) ศึกษาหลักการวัดความคล้ายคลึงของเอกสารและการให้ค่าน้ำหนักคำ

4) วิเคราะห์ ออกแบบ และดำเนินการพัฒนาระบบ

5) ทดสอบ ปรับปรุงและประเมินผล

6) สรุปผลและจัดทำเอกสารประกอบระบบ

**1.4.2 ขอบเขตของการศึกษา**

ขอบเขตของการศึกษาการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ มีดังนี้

(ก) ขอบเขตข้อมูล

ข้อมูลที่นำมาใช้ในการเปรียบเทียบจะเป็นข้อมูลภาษาไทยเท่านั้น

(ข) ขอบเขตฟังก์ชันงาน

1) ระบบรับไฟล์เอกสารมา เพื่อจะทำการตัดคำและสร้างโครงสร้างที่ทำการตัดคำ แล้วเก็บเป็นต้นฉบับ

2) ระบบรับไฟล์เอกสารที่ต้องการเปรียบเทียบในรูปแบบเดียวกัน เพื่อทำการตัดคำ และสร้างโครงสร้างแบบเดียวกันกับ 1) และนำไปเปรียบเทียบกับต้นฉบับ

3) ระบบสามารถคำนวณค่าความถี่ของคำ และค่าน้ำหนักคำได้

4) ระบบสามารถวัดความคล้ายคลึงของเอกสารได้

**1.5 วิธีการศึกษาและเครื่องมือที่ใช้ในการพัฒนา**

วิธีการศึกษาการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ จะอธิบายถึงวิธีการศึกษาและเครื่องมือที่ใช้ในการศึกษา มีดังนี้

**1.5.1 วิธีการวิจัย**

วิธีการวิจัยของการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติมีดังนี้

(1) ศึกษาทฤษฎีและหลักการต่าง ๆ ที่เกี่ยวข้อง

- หลักการตัดคำภาษาไทย
- หลักการแบบจำลองเวกเตอร์สเปซ
- การให้ค่าน้ำหนักคำ
- การนับค่าความถี่ของคำ
- หลักการวัดความคล้ายคลึงของเอกสาร

(2) จัดเตรียมข้อมูลที่จะทำการทดสอบ โดยแบ่งกลุ่มข้อมูลออกเป็นสองส่วน คือ ข้อมูลส่วนที่จะใช้เป็นข้อมูลตั้งต้นในการทดสอบ และข้อมูลส่วนที่จะนำมาทำการเปรียบเทียบ

(3) วิเคราะห์และออกแบบโปรแกรม

(4) กำหนดขอบเขตและขั้นตอนการทำงานของโปรแกรม

- แนววิธีคำนวณความคล้ายคลึงของเอกสาร
- รูปแบบการรายงานหรือแสดงผลที่ได้จากการคำนวณความคล้ายคลึง

(5) พัฒนาโปรแกรม ทดสอบผล และทำการปรับปรุงระบบ

(6) สรุปผลการวัดออกมาในลักษณะทางสถิติ เช่น ร้อยละ

**1.5.2 เครื่องมือที่ใช้ในการศึกษาและพัฒนา**

เครื่องมือที่ใช้ในการศึกษาการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ มีดังนี้

(1) ฮาร์ดแวร์

- เครื่องไมโครคอมพิวเตอร์จำนวนหนึ่งเครื่องภายใต้ระบบปฏิบัติการไมโครซอฟท์วินโดว์วิสต้า

(2) ซอฟต์แวร์

- ไมโครซอฟท์วิชวล ซีชาร์ป 2008 เอ็กซ์เพส

**1.6 สถานที่ที่ใช้ดำเนินการ**

สถานที่ที่ใช้ในการดำเนินการวิจัยและรวบรวมข้อมูล ได้แก่

(1) ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

(2) สำนักหอสมุด มหาวิทยาลัยเชียงใหม่

(3) สำนักบริการเทคโนโลยีสารสนเทศ มหาวิทยาลัยเชียงใหม่