

## สารบัญ

|                                                                | หน้า |
|----------------------------------------------------------------|------|
| กิตติกรรมประกาศ                                                | ก    |
| บทคัดย่อภาษาไทย                                                | ง    |
| บทคัดย่อภาษาอังกฤษ                                             | จ    |
| สารบัญ                                                         | ฉ    |
| สารบัญตาราง                                                    | ช    |
| สารบัญภาพ                                                      | ฌ    |
| บทที่ 1 บทนำ                                                   |      |
| 1.1 หลักการ ทฤษฎี เหตุผล และ/หรือสมมุติฐาน                     | 1    |
| 1.2 วัตถุประสงค์ของการศึกษา                                    | 2    |
| 1.3 ประโยชน์ที่ได้รับจากการศึกษาเชิงทฤษฎีและ/ หรือเชิงประยุกต์ | 3    |
| 1.4 แผนการดำเนินงาน และขอบเขตของการศึกษา                       | 3    |
| 1.5 วิธีการศึกษาและเครื่องมือที่ใช้ในการพัฒนา                  | 4    |
| 1.6 สถานที่ที่ใช้ดำเนินการ                                     | 5    |
| บทที่ 2 แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง                  |      |
| 2.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)      | 6    |
| 2.2 แบบจำลองเวกเตอร์สเปซ (Vector Space Model)                  | 20   |
| 2.3 การวัดประสิทธิภาพของระบบ (Retrieval Effectiveness)         | 23   |
| 2.4 งานวิจัยที่เกี่ยวข้อง                                      | 25   |
| บทที่ 3 หลักการของแบบจำลองเวกเตอร์สเปซ                         |      |
| 3.1 กระบวนการแบบจำลองเวกเตอร์สเปซ                              | 30   |
| 3.2 การให้ค่าน้ำหนักคำ (Term Weighting)                        | 33   |
| 3.3 การวัดความคล้ายคลึง (Similarity Measure)                   | 39   |

## สารบัญ (ต่อ)

|                                                                  | หน้า |
|------------------------------------------------------------------|------|
| บทที่ 4 การพัฒนาซอฟต์แวร์เพื่อวัดความคล้ายคลึงของเอกสารภาษาไทย   |      |
| 4.1 กระบวนการเตรียมเอกสาร (Pre-Processing)                       | 42   |
| 4.2 การประมวลผลเอกสาร (Document Processing)                      | 45   |
| 4.3 การวัดความคล้ายคลึงเชิงมุมโคไซน์ (Cosine Similarity Measure) | 48   |
| 4.4 การประเมินผลซอฟต์แวร์การวัดความคล้ายคลึงของเอกสาร            | 50   |
| บทที่ 5 สรุปผลการศึกษา และข้อเสนอแนะ                             |      |
| 5.1 สรุปผลการค้นคว้า                                             | 54   |
| 5.2 ข้อจำกัดของระบบ                                              | 54   |
| 5.3 ปัญหาและอุปสรรค                                              | 54   |
| 5.4 ข้อเสนอแนะ                                                   | 55   |
| บรรณานุกรม                                                       | 56   |
| ภาคผนวก                                                          |      |
| ภาคผนวก ก คำหยุดภาษาไทย (Thai Stop Words)                        | 59   |
| ภาคผนวก ข ตัวอย่างข้อมูลที่นำมาทดสอบ                             | 68   |
| ภาคผนวก ค คู่มือการใช้งานซอฟต์แวร์การวัดความคล้ายคลึงของเอกสาร   | 75   |
| ประวัติผู้เขียน                                                  | 82   |

## สารบัญตาราง

| ตาราง                                                                         | หน้า |
|-------------------------------------------------------------------------------|------|
| 2.1 ผลการตัดคำทั้งหมดด้วยวิธีเทียบคำที่ยาวที่สุด                              | 16   |
| 2.2 ผลการตัดคำด้วยวิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่มีในพจนานุกรมน้อยที่สุด | 18   |
| 3.1 การจัดรูปแบบของเอกสารแบบถุงของคำ                                          | 29   |
| 3.2 แสดงความถี่ของคำในเอกสาร A                                                | 30   |
| 3.3 แสดงความถี่ของคำในเอกสาร B                                                | 31   |
| 3.4 แสดงค่าของเอกสาร A เมื่อนำมาทำเป็นเวกเตอร์                                | 31   |
| 3.5 แสดงค่าของเอกสาร B เมื่อนำมาทำเป็นเวกเตอร์                                | 31   |
| 3.6 แสดงค่าของข้อสอบถามเมื่อนำมาทำเป็นเวกเตอร์                                | 32   |
| 3.7 แสดงคำในเอกสารทั้งหมด เมื่อนำมาทำเป็นดัชนีคำศัพท์                         | 35   |
| 3.8 แสดงความถี่ของคำแต่ละคำที่พบในเอกสารแต่ละเอกสาร                           | 36   |
| 3.9 คำที่ได้หลังจากการคำนวณค่าน้ำหนักคำในเอกสารแล้ว                           | 38   |
| 3.10 คำที่ได้หลังจากการคำนวณค่าน้ำหนักคำของข้อสอบถามที่รับเข้ามา              | 38   |
| 3.11 ตัวอย่างสัมประสิทธิ์ความสัมพันธ์แบบต่าง ๆ                                | 39   |
| 4.1 ผลลัพธ์จากการหาความถี่ของคำและค่าความถี่ส่วนกลับของเอกสาร                 | 47   |
| 4.2 ผลลัพธ์ที่ได้จากการหาค่าน้ำหนักคำในเอกสาร                                 | 47   |
| 4.3 รายละเอียดขั้นตอนการทดสอบประสิทธิภาพ                                      | 50   |

## สารบัญภาพ

| รูป                                                                        | หน้า |
|----------------------------------------------------------------------------|------|
| 2.1 แสดงกลุ่มตัวอักษรที่เป็นไปได้สำหรับการตัดคำ                            | 13   |
| 2.2 แสดงเวกเตอร์ของเอกสาร 1 เอกสาร 2 และข้อสอบถาม                          | 21   |
| 2.3 แสดงเมตริกของเอกสาร 1 เอกสาร 2 และข้อสอบถาม                            | 22   |
| 2.4 แสดงความสัมพันธ์ระหว่างกลุ่มข้อมูลในฐานข้อมูลที่เกี่ยวข้องกับข้อสอบถาม | 23   |
| 2.5 แสดงความสัมพันธ์ระหว่างกลุ่มของข้อมูลที่ใช้สืบค้นกับข้อสอบถาม          | 23   |
| 2.6 แสดงการคำนวณการหาค่าความถูกต้องของข้อมูล                               | 24   |
| 2.7 แสดงการคำนวณหาค่าความแม่นยำของข้อมูล                                   | 24   |
| 2.8 ความสัมพันธ์ระหว่างค่าความแม่นยำและค่าความถูกต้อง                      | 25   |
| 3.1 การนับความถี่ของคำเมื่อนำมาเขียนในรูปเมตริกซ์                          | 36   |
| 3.2 ความถี่ของเอกสารที่พบของดัชนีคำศัพท์แต่ละคำในเอกสาร                    | 36   |
| 3.3 คำที่ได้หลังจากการคำนวณหาค่าส่วนกลับความถี่ของเอกสารแล้ว               | 37   |
| 3.4 การหาค่าน้ำหนักคำแต่ละคำในเอกสาร                                       | 38   |
| 4.1 แสดงขั้นตอนการทำงานของซอฟต์แวร์                                        | 44   |
| 4.2 ขั้นตอนการเตรียมเอกสารในการประมวลผลขั้นแรก                             | 45   |
| 4.3 ผลการวัดประสิทธิภาพของการวัดความคล้ายคลึงระหว่างเอกสารชุด 1 และ 2      | 51   |
| 4.4 ผลการวัดประสิทธิภาพของการวัดความคล้ายคลึงระหว่างเอกสารชุด 1 และ 3      | 51   |
| 4.5 ผลการวัดประสิทธิภาพของการวัดความคล้ายคลึงระหว่างเอกสารชุด 1 และ 4      | 52   |
| ค.1 แสดงไอคอนการเข้าโปรแกรม                                                | 75   |
| ค.2 ซอฟต์แวร์การวัดความคล้ายคลึงของเอกสาร                                  | 75   |
| ค.3 แสดงส่วนเมนูบาร์                                                       | 76   |
| ค.4 แสดงส่วนเมนูการตั้งค่า                                                 | 76   |
| ค. 5 แสดงส่วนการแสดงผล                                                     | 76   |
| ค.6 แสดงเมนูย่อยในเมนูบาร์                                                 | 77   |
| ค. 7 แสดงเมนูย่อยของเมนูไฟล์                                               | 77   |
| ค.8 แสดงเมนูย่อยของเมนูคำนวณ                                               | 77   |
| ค.9 แสดงเมนูย่อยของเมนูการตั้งค่า                                          | 78   |
| ค.10 แสดงส่วนย่อยของการแสดงผล                                              | 78   |

ญ

## สารบัญภาพ (ต่อ)

| รูป                                                               | หน้า |
|-------------------------------------------------------------------|------|
| ค.11 แสดงหน้าจอย่อยของการแสดงผลก่อนทำการวัดความคล้ายคลึงของเอกสาร | 79   |
| ค.12 การแสดงผลหลังทำการวัดความคล้ายคลึงของเอกสาร                  | 80   |



ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่  
Copyright© by Chiang Mai University  
All rights reserved