ชื่อเรื่องการค้นคว้าแบบอิสระ การวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ

ผู้เขียน นางสาววรวีร์ เกษร

ปริญญา วิทยาศาสตรมหาบัณฑิต(วิทยาการคอมพิวเตอร์)

อาจารย์ที่ปรึกษาการค้นคว้าแบบอิสระ

ผู้ช่วยศาสตราจารย์ ดร.รัฐสิทธิ์ สุขะหุต

บทคัดย่อ

การค้นคว้าแบบอิสระเรื่อง การวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ มีวัตถุประสงค์เพื่อพัฒนาเครื่องมือที่ช่วยในการวัดค่าความคล้ายคลึงระหว่างเอกสารภาษาไทยโดยนำหลักการวัดค่าความคล้ายคลึงเชิงมุมโคไซน์ การให้ค่าน้ำหนักคำและการนับความถี่ของคำมาใช้กับหลักการของแบบจำลองเวกเตอร์สเปซในการวัดค่าความคล้ายคลึง และเพื่อศึกษาการจัดกลุ่มของเอกสารภาษาไทยที่มีความคล้ายคลึงกันและทำการตัดคำภายใต้แนวความคิดการประมวลผลภาษาธรรมชาติ

การวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติ แบ่งออกเป็น 2 ส่วน คือ ส่วนการเตรียมเอกสารก่อนการประมวลผลและส่วนการประมวลผล โดยเอกสารที่นำมาใช้ในการวัดความคล้ายคลึงนั้น จะเป็นเอกสารภาษาไทยจำนวน 150 เอกสาร จำกัดจำนวนคำอยู่ที่เอกสารละ 300 คำ การพัฒนาระบบงานจะแบ่งออกเป็น 2 ขั้นตอน คือ 1) ศึกษาวิธีการวัดความคล้ายคลึงของเอกสาร 2) พัฒนาเครื่องมือการวัดความคล้ายคลึงของเอกสารโดยใช้ภาษาซีชาร์ป ดอทเน็ต

ผลการทดสอบประสิทธิภาพการวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้การประมวลผลภาษาธรรมชาติสามารถวัดได้จากค่าความแม่นยำ และค่าความถูกต้องของข้อมูลที่ระดับค่าแตกต่างกัน การทดสอบประสิทธิภาพพบว่าระบบสามารถทำงานได้ถูกต้องเป็นไปตามวัตถุประสงค์ของการค้นคว้า

Independent Study Title Similarity Measurement of Thai Documents Using Natural Language Processing

Author Miss Warawee Kesorn

Degree Master of Science (Computer Science)

Independent Study Advisor

Asst. Prof. Dr. Rattasit Sukhahuta

Abstract

The objectives of the independent study entitled "Similarity Measurement of Thai Documents Using Natural Language Processing" are to develop a tool to help measure the similarity of Thai documents by applying cosine similarity, term weighting and term frequency Methods to a Vector Space Model in the order to measure similarity, and to study the grouping of similar Thai documents and the word segmentation revealed, using natural language processing.

The measurement of the similarity of Thai documents using natural language processing was divided into two parts, these begin: preparation of documents prior to processing, and processing. In total of 150 Thai documents were used as part of the measurement, each limited to no more than 300 words. This research system was divided into two processes, these begin; 1) the measurement and study of document similarity, and 2) the development tool of a document similarity measure using C#.net

The effectiveness of the similarity measurement for Thai documents using natural language processing, can be tested through the accuracy and correctness of the data on different levels. The results of the test showed that the measuring system works properly in accordance with the objectives of this research.