บทที่ 1
บทนำ

1.1 ปัญหาและความเป็นมา

ด้วยจำนวนเอกสารที่มีปริมาณเพิ่มมากขึ้นอย่างรวดเร็ว ทำให้งานด้านการวิเคราะห์ความหมายเอกสารจึงมีความยากเพิ่มมากขึ้นตามไปด้วย โดยเอกสารแต่ละประเภทก็มีเนื้อหารายละเอียดที่แตกต่างกันไป เช่น เอกสารการประเภทพยากรณ์อากาศ สิ่งที่ต้องการวิเคราะห์คือ เกิดสภาพอากาศอะไร ที่ไหน เมื่อไหร่ มีผลกระทบต่อใครบ้างและทำให้เกิดผลอะไร และเอกสารประเภทกีฬา สิ่งที่ต้องการวิเคราะห์คือ ใครแข่งกับใคร ผลการแข่งขันเป็นอย่างไร รายละเอียดของการแข่งขันเป็นอย่างไร และมีเหตุการณ์อะไรเกิดขึ้นบ้างในระหว่างการแข่งขัน การกำหนดหัวข้อให้กับเอกสารจึงเป็นงานที่สำคัญในการวิเคราะห์ความหมายเอกสาร โดยเมื่อมีการระบุหัวข้อเอกสารแล้วก็จะทำให้ทราบว่าข้อมูลที่ต้องการวิเคราะห์คืออะไร ทำให้สามารถวิเคราะห์ความหมายได้สอดคล้องกับเนื้อความในเอกสาร

การกำหนดหัวข้อเอกสารนั้นงานวิจัยโดยส่วนใหญ่แล้วจะใช้แนวทางการวิเคราะห์โดยใช้วิธีทางสถิติ และการเรียนรู้จากกลุ่มเอกสารฝึกสอน โดยวิเคราะห์จากเทอมที่ปรากฏในเอกสาร ซึ่งในการวิเคราะห์ความสำคัญของเทอมนั้นจะมีการกำหนดค่าน้ำหนักให้กับแต่ละเทอม โดยเอกสารที่ใช้ในการฝึกสอนจะคัดคำที่เป็นคำฟุ่มเฟือยออกก่อนแล้วจึงทำการหาค่าน้ำหนักของคำที่เหลือที่เรียกว่า content words จากนั้นจึงกำหนดเทอมที่สำคัญที่สามารถใช้เป็นตัวแทนของหัวข้อเอกสารเพื่อระบุหัวข้อให้กับเอกสารใหม่ โดยเรียกเทอมเหล่านั้นว่า topic keywords เช่น เอกสารประเภทข่าวกีฬามักประกอบด้วยกลุ่มคำสำคัญ เช่น tournament, champion, won, defeat, final, player, match, coach, team และเอกสารประเภทสุขภาพมักประกอบด้วยคำสำคัญ เช่น health, disease, medical, patients, medicine, blood, drug ซึ่งเทอมที่มีความน้ำหนักมากก็จะสามารถเป็นตัวแทนที่ดีของหัวข้อนั้น โดยค่าน้ำหนักที่ใช้ในการวิเคราะห์ความสำคัญของเทอมนั้น มีด้วยกันหลายวิธี เช่น Term Frequency Inverse Document Frequency (TFIDF) (Salto and McGill, 1983), Information Gain (Wang et al., 2007, Li et al, 2009), Chi-Square (Caropreso et al., 2001, Li et al, 2009), และ Term Frequency Inverse Class Frequency (TFICF) (Kim et al., 2005)

ดังนั้นในงานวิจัยนี้จึงได้นำเสนอการออกแบบและทำการทดลองเพื่อเปรียบเทียบเทคนิคต่าง ๆ ที่ใช้ในการกำหนดหัวข้อข่าวให้กับเอกสารดังที่กล่าวข้างต้น โดยการประยุกต์ใช้การวิเคราะห์ตาม Zipf's Law () ในการกำหนดกลุ่มตัวแทนของหัวข้อ ซึ่งผลการทดลองนี้สามารถกำหนดกลุ่มของคำสำคัญแต่ละหัวข้อเพื่อใช้ในการระบุหัวข้อข่าวให้กับเอกสารใหม่ และใช้เทคนิคการจัดกลุ่มเอกสารซัพพอร์ทเวกเตอร์แมชชีนในการวัดประสิทธิภาพกลุ่มของเทอมที่สำคัญ นอกจากนี้การกำหนดกลุ่มคำสำคัญของหัวข้อนี้สามารถนำไปเป็นขั้นตอนหนึ่งในการทำงานทางด้านการสกัดความรู้ (Information Extraction) การแบ่งกลุ่มเอกสาร

(Document Clustering) การสรุปเอกสาร (Document Summarization) การวิเคราะห์เนื้อความเอกสาร (Document Content Analysis)

1.2 วัตถุประสงค์ในการวิจัย

เพื่อออกแบบ ทดลอง และเปรียบเทียบเทคนิควิธีที่เหมาะสมในการกำหนดหัวข้อข่าวให้กับเอกสาร ซึ่งเทคนิคที่ใช้ในการเปรียบเทียบได้แก่ TFIDF, TFICF, Chi-Square, และ Information Gain

1.3 ขอบเขตการศึกษา

เพื่อเปรียบเทียบประสิทธิภาพเทคนิคที่ใช้ในการกำหนดหัวข้อให้กับเอกสาร โดยกลุ่มเอกสารที่ใช้ในการทดลองประกอบด้วยเอกสารประเภท กีฬา สภาพอากาศ ธุรกิจ การเมือง สุขภาพ และบันเทิง โดยคำนึงถึงปัจจัยต่าง ๆ ที่มีผลกระทบต่อการกำหนดกลุ่มคำสำคัญที่ใช้เป็นตัวแทนของหัวข้อ ซึ่งได้แก่ จำนวนเอกสาร เทอมในกลุ่มคำสำคัญ และค่า Threshold ที่กำหนดค่าความถี่ต่ำสุดของคำสำคัญ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ผลสรุปจากการทดลองทำให้ได้เทคนิคที่เหมาะสมในการกำหนดกลุ่มคำที่สำคัญที่ใช้ในการกำหนดหัวข้อข่าว
2. สามารถนำใช้วิธีการกำหนดคำสำคัญนี้ไปใช้ในงานทางด้านการแบ่งกลุ่มเอกสาร การจัดกลุ่มเอกสาร การสกัดข้อมูล การสรุป และการวิเคราะห์ความหมายเอกสารได้