เว็บไซต์ช่วยวิเคราะห์เอกสารและสร้างคุณลักษณะ

Documents Analyzer and Feature Extraction Via Website

จิรวรรณ เจริญสุข[1*] อานนท์ ผ่องรัศมีเพ็ญ[1] ทรงพร ฉายศรี[1] และ อรนิตย์ บัณฑิตย์เสถียร[1]

บทคัดย่อ

งานวิจัยนี้นำเสนอการวิเคราะห์ข้อความหรือเอกสารภาษาอังกฤษผ่านเว็บไซต์ เพื่ออำนวยความสะดวกให้กับผู้วิจัยในการวิเคราะห์ข้อมูลการปรากฏของคำและความสัมพันธ์ของคำภายในเอกสาร อีกทั้งยังเป็นเครื่องมือที่ช่วยให้นิสิต นักศึกษาที่เรียนวิชาการประมวลผลภาษาธรรมชาติใช้ในการเรียนรู้วิธีการประมวลผลข้อความเบื้องต้น นอกจากนี้เว็บไซต์นี้ยังถือว่าเป็นเครื่องมือที่ช่วยจัดเตรียมข้อมูลเพื่อนำไปประยุกต์ใช้งานกับซอฟต์แวร์ด้านเหมืองข้อมูลอีกด้วยกระบวนการทำงานของเว็บไซต์จะเริ่มต้นจากผู้ใช้งานกรอกข้อความหรือนำเข้าเอกสารสู่เว็บไซต์ หลังจากนั้นข้อความดังกล่าวจะถูกตัดคำแบบอัตโนมัติโดยใช้ช่องว่างและเครื่องหมายแล้วให้ผู้ใช้งานเลือกตัวเลือกที่ใช้ในการวิเคราะห์ได้แก่การกำจัดคำหยุดและการถอดรากศัพท์ จากนั้นผู้ใช้สามารถเลือกรูปแบบการแสดงผลให้ตรงตามความต้องการ ได้แก่ การแสดงค่าความถี่ การแสดงผลค่าน้ำหนักของคำ การแสดงผลค่าการสร้างคุณลักษณะของคำด้วยการแทนค่าความจริงหรือความถี่ และการแสดงผลการวิเคราะห์คำสำคัญของเอกสาร ผลการทดสอบการใช้งานเว็บไซต์นี้จากกลุ่มผู้เรียนวิชาการประมวลผลภาษาธรรมชาติเชิงสถิตและผู้ที่สนใจเกี่ยวกับการวิเคราะห์ข้อความได้ผลสรุประดับความพึงพอใจตามค่าสถิติอยู่ในระดับดี โดยมีค่าเฉลี่ยเท่ากับ 3.95 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.909

คำสำคัญ: การวิเคราะห์เอกสาร การสกัดคุณลักษณะ การสร้างคำสำคัญ

Abstract

This research purposes English text and documents analysis via website. It is facilitated for researchersto analyze the presence of the word and correlation of words into documents. In addition, this website is a tool of students who study Natural Language Processing subject for learning about the basic methods of document analysis. Furthermore, it is usedin data preparation process and is applied in data mining software. The first process of website, user can types text in textbox or selects files in input process. These inputs are automatically segmented to be each word by using space and symbols. Then, user can selects options of analysis which are Stop-Word List Removal and Stemming. Later, users can selects output displays including Frequency of word, TF-IDF weight, Features extraction by using Boolean or frequency of word, and Keyword extraction of all documents. The results of website evaluation from students who study Statistic of Natural Language Processing subject and who interest in text analysis are a good level according to the statistics. And, mean is 3.95 and standard deviation is 0.909.

Keywords: Documents Analyzer, Feature Extraction, KeywordExtraction

[1] คณะวิทยาศาสตร์ ศรีราชา มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา ชลบุรี 20230

[1] Faculty of Science at Si Racha, Kasetsart University Si Racha Campus, ChonBuri 20230,Thailand

* Corresponding author. E-mail : sfscijrc@src.ku.ac.th

บทนำ

การประมวลผลภาษาธรรมชาติเป็นกระบวนการที่ทำให้คอมพิวเตอร์มีความสามารถในการตีความภาษาธรรมชาติซึ่งมนุษย์ใช้สื่อสารในชีวิตประจำวันได้ ถ้าเราสามารถประมวลผลภาษาธรรมชาติได้จะทำให้เข้าถึงและใช้ประโยชน์จากข้อมูลจำนวนมากได้อย่างรวดเร็วและมีประสิทธิภาพ เนื่องจากข้อมูลต่าง ๆ ในโลกนี้นับวันยิ่งเพิ่มปริมาณมากขึ้นเรื่อยๆ การจะนำข้อมูลเหล่านี้ไปใช้งานจะต้องใช้เวลาเป็นอันมาก ในการรวบรวมและวิเคราะห์ข้อมูลเหล่านั้น ถ้าสามารถนำเครื่องคอมพิวเตอร์มาช่วยในการประมวลผลได้ก็จะใช้เวลาลดลง นอกจากนี้การประมวลผลภาษาธรรมชาติยังเป็นเครื่องมือที่มนุษย์และคอมพิวเตอร์ใช้ในการสื่อสารระหว่างกันในปัจจุบันการใช้เทคโนโลยีและการแลกเปลี่ยนข้อมูลระหว่างกัน ได้เข้ามามีบทบาทในชีวิตประจำวันมากขึ้น ทำให้เกิดเอกสารและข้อมูลจำนวนมาก ซึ่งข้อมูลที่เกิดขึ้นจะมีรูปแบบที่หลากหลายแตกต่างกันออกไปไม่ว่าจะอยู่ในรูปแบบตัวหนังสือ ฐานข้อมูลหรือวีดีโอ ซึ่งข้อมูลเหล่านี้มีขนาดใหญ่และมีการเปลี่ยนแปลงอยู่ตลอดเวลา งานวิจัยนี้จึงนำเสนอการพัฒนาเว็บไซด์เพื่อใช้เป็นเครื่องมือในการวิเคราะห์ข้อความหรือเอกสาร โดยการสร้างคุณลักษณะเพื่อสนับสนุนการทำงานของซอฟต์แวร์การเรียนรู้ด้วยเครื่อง (machine learning) และยังช่วยลดค่าใช้จ่ายในการซื้อซอฟต์แวร์ที่ช่วยในการวิเคราะห์ข้อมูลจากต่างประเทศ อีกทั้งเครื่องมือดังกล่าวที่พัฒนาขึ้นนี้ยังสามารถช่วยลดเวลาในการวิเคราะห์และการหาความสัมพันธ์ของคำที่ปรากฏ การหาค่าน้ำหนักคำและการสร้างคุณลักษณะเพื่อใช้เป็นตัวแทนของเอกสาร นอกจากเว็บไซด์นี้ยังช่วยให้ ผู้วิจัย นิสิต นักศึกษา ที่สนใจและเรียนวิชาเกี่ยวกับ Natural Language Processing (NLP) สามารถเรียนรู้เกี่ยวกับการประมวลผลข้อความและสามารถจัดเตรียมข้อมูลเบื้องต้นสำหรับนำไปใช้ในงานวิจัยในขั้นตอนต่อไปได้

วิธีการศึกษา

ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง

Laddaet *al.* (2009) นำเสนอการย่อความเอกสาร โดยขั้นตอนแรกจะต้องผ่านกระบวนการกำจัดคำหยุดและการหารากศัพท์ก่อนจะเข้าสู่ขั้นตอนการย่อความเอกสารโดยใช้อัลกอริทึมฟัซซีลอจิกในการหาประโยคที่เป็นใจความสำคัญของเอกสาร

Leena.H.Patil , Mohammed Atique (2012) นำเสนอการจัดกลุ่มเอกสารโดยนำอัลกอริทึม Porter (Porter, 1980) มาใช้ในการหารากศัพท์ และมีกระบวนการกำจัดคำหยุดเพื่อลดจำนวนคำศัพท์ที่ไม่มีผลต่อการจัดกลุ่มเอกสาร แล้วทำการเลือกคุณลักษณะของเอกสารโดยใช้การให้ค่าน้ำหนักคำด้วยTF-IDF บนพื้นฐานค่าความถี่ของแต่ละคำภายในเอกสาร

KrantiGhag, Ketan Shah (2014) งานวิจัยนี้ได้นำเสนอการจำแนกทัศนคติจากเอกสารโดยเริ่มจากกระบวนการกำจัดคำหยุดเพื่อลดคุณลักษณะของเอกสารลงและเพื่อเพิ่มประสิทธิภาพในการจำแนกทัศนคติ แล้วใช้ค่าความถี่เป็นสัดส่วนเปรียบเทียบการกระจายของเอกสารและใช้การให้ค่าน้ำหนักคำด้วยTF-IDF เป็นการถ่วงค่าน้ำหนักก่อนส่งเข้าสู่เทคนิคที่ใช้สำหรับการดึงข้อมูลและการจำแนกทัศนคติเป็นความรู้สึกในเชิงบวกและความรู้สึกในเชิงลบ

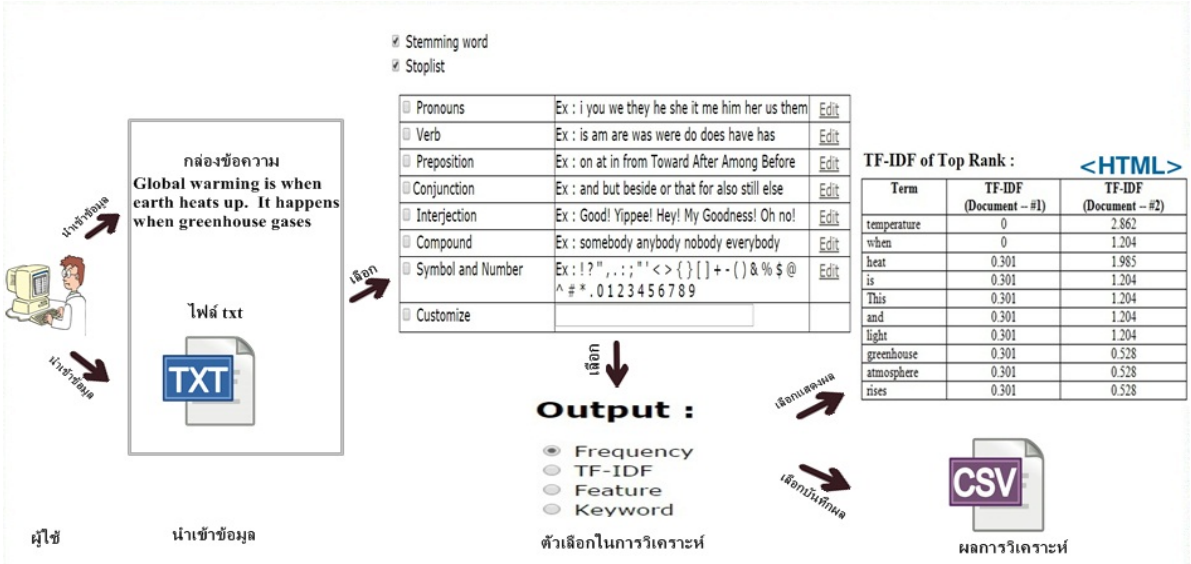
จากบทความข้างต้นแสดงให้เห็นว่าในการสร้างคุณลักษณะจากเอกสารจะมีขั้นตอนพื้นฐานซึ่งเป็นที่สำคัญและส่งผลต่อประสิทธิภาพของกระบวนการในลำดับถัดไป คือ ขั้นตอนการกำจัดคำหยุด การหารากศัพท์ ผลลัพธ์ของเว็บไซด์นี้คือการแสดงผลค่าความถี่การแสดงผลการให้ค่าน้ำหนักคำด้วยTF-IDF และการแสดงผลค่าคุณลักษณะ

ของคำซึ่งผลจากการวิเคราะห์ดังกล่าวสามารถนำไปประยุกต์ใช้ในงานวิจัยด้านการประมวลผลภาษาธรรมชาติได้อย่างมีประสิทธิภาพมากขึ้น

ศึกษาปัญหาและวิเคราะห์

จากการศึกษางานวิจัยที่เกี่ยวข้องและได้ทำการวิเคราะห์เว็บไซต์ Text analyserนั้นต้องเสียค่าใช้จ่ายในการใช้งานบางส่วน การนำเข้าข้อมูลและ การกำจัดคำหยุดนั้นสามารถทำได้น้อย อีกทั้งยังขาดการสร้างคุณลักษณะเพื่อช่วยเพิ่มประสิทธิภาพในการวิเคราะห์เอกสารให้คอมพิวเตอร์เข้าใจได้ ผู้วิจัยจึงมีแนวคิดที่จะพัฒนาเว็บไซต์ช่วยวิเคราะห์เอกสารและสร้างคุณลักษณะรายการคำศัพท์โดยเพิ่มจำนวนการรับข้อมูลในรูปแบบไฟล์ให้มากขึ้น ช่วยลดคุณลักษณะของเอกสาร โดยการกำจัดคำหยุด (Stop-Word List Removal) ทั้งนี้ผู้วิจัยทำการจำแนกคำหยุดออกเป็นหมวดหมู่ เพื่อให้ง่ายต่อการใช้งาน และยังสามารถเพิ่ม ลบ แก้ไข คำหยุด ตามความต้องการของผู้ใช้ได้ จากนั้นเพิ่มประสิทธิภาพในการวิเคราะห์โดยการหารากศัพท์ (Stemming)ในเอกสารที่รับเข้ามา และในส่วนผลการวิเคราะห์นั้นเพื่อให้ง่ายต่อการดูข้อมูลและนำไปวิเคราะห์ในด้านต่าง ๆ แบ่งออกเป็น 4 ตัวเลือกได้แก่ ค่าความถี่(Frequency) การให้ค่าน้ำหนักคำด้วยTF-IDF การสกัดคุณลักษณะ(Feature Extraction)และการหาคำสำคัญของเอกสาร (KeywordExtraction) ทั้งนี้ยังสามารถบันทึกผลการวิเคราะห์เป็นไฟล์CSV เพื่อนำไปใช้ในงานอื่นๆ ได้อีกด้วย

แผนภาพการทำงานของเว็บไซต์



ภาพที่ 1 กระบวนการทำงานของระบบ

จากภาพที่ 1 กระบวนการทำงานของเว็บไซต์จะเริ่มต้นจากผู้ใช้งานทำการกรอกข้อความลงในกล่องรับข้อความหรือนำเข้าเอกสารที่เป็นรูปแบบของไฟล์ตัวอักษร (txt) ได้จำนวนสูงสุด 5 ไฟล์ หลังจากนั้นระบบจะทำการตัดคำแบบอัตโนมัติ โดยใช้ช่องว่าง (space) และเครื่องหมายจุลภาค (.) หลังจากนั้นผู้ใช้งานสามารถเลือกตัวเลือกที่ใช้ในการวิเคราะห์โดยแบ่งเป็น 1) การกำจัดคำหยุด (Stop list) ซึ่งการกำจัดคำหยุดแบ่งออกเป็น 8 หมวดหมู่ ซึ่งผู้ใช้

สามารถเพิ่ม ลบ แก้ไข คำหยุดได้เองตามความต้องการ 2) การถอดรากศัพท์ (Stemming)เพื่อช่วยลดจำนวนคำศัพท์ของเอกสาร ส่วนของการแสดงผลการวิเคราะห์นั้น แบ่งออกเป็น 4 ส่วนได้แก่ การแสดงผลค่าความถี่ (Frequency)การแสดงผลการให้ค่าน้ำหนักคำด้วยTF-IDF และการแสดงผลค่าการสร้างคุณลักษณะของคำ(Feature Extraction) ซึ่งผลการวิเคราะห์นี้ผู้ใช้สามารถเลือกแสดงผลบนหน้าเว็บไซต์หรือเลือกบันทึกในรูปแบบไฟล์ CSVเพื่อนำไปใช้กับซอฟต์แวร์ด้านเหมืองข้อมูล เช่น Weka, Rapid Miner เป็นต้น นอกจากนี้ยังสามารถแสดงคำสำคัญ (KeywordExtraction) ของเอกสารทั้งหมดโดยรายละเอียดของแต่ละขั้นตอนแสดงไว้ในลำดับถัดไป

การตัดคำ (Word Segmentation)

การตัดคำเป็นพื้นฐานและขั้นตอนแรกของการประมวลผลข้อความ ทั้งนี้การตัดคำของแต่ละภาษาขึ้นอยู่กับลักษณะโครงสร้างของภาษานั้นๆ ในงานวิจัยนี้เป็นเอกสารภาษาอังกฤษจะใช้ช่องว่าง (Space) หรือเครื่องหมาย เช่น ; (Semi-colon) , (Comma) . (Period) ? (Question Mark) ! (Exclamation Mark) ในการแบ่งขอบเขตของคำหรือจุดสิ้นสุดของประโยค (นิเวศและคณะ, 2554)

การกำจัดคำหยุด (Stop-Word List Removal)

การกำจัดคำหยุดเป็นขั้นตอนที่นำรายการคำศัพท์ที่ไม่มีนัยสำคัญออกจากเอกสาร ซึ่งปรากฏของคำเหล่านี้ไม่เกี่ยวข้องและไม่มีประโยชน์ในการวิเคราะห์ นอกจากนี้การนำรายการคำศัพท์เหล่านี้ออกจากเอกสารแล้วไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลงไป สำหรับงานวิจัยนี้ได้แบ่งประเภทของรายการคำหยุด(คมคิดและคณะ,2555) ออกเป็น 8 ประเภท ได้แก่ คำสรรพนาม (Pronoun) คำกริยา (Verb) คำบุพบท (Preposition) คำสันธาน (Conjunction)คำอุทาน (Interjection) คำผสม (Compound) สัญลักษณ์และตัวเลข (Symbol and Numeric) คำคุณศัพท์(Adjective) คำหยุดเพิ่มเติม(Customize) การกำจัดคำหยุดถือว่าเป็นขั้นตอนที่สำคัญขั้นตอนหนึ่งในการลดขนาดคำภายในเอกสารและลดขนาดของการสร้างคุณลักษณะซึ่งจะส่งผลให้การประมวลผลเอกสารในขั้นตอนต่อไป มีประสิทธิภาพเพิ่มมากขึ้น และตัวอย่างรายการคำหยุดของงานวิจัยนี้แสดงดังตาราง (วิรัตน์ ชูนุ้ย, 2555)

ตารางที่ 1 ตารางประเภทคำหยุด

ประเภทของคำหยุด	ตัวอย่าง	ประเภทของคำหยุด	ตัวอย่าง
1. สรรพนาม (Pronouns)	you , he , her , them	5. คำอุทาน (Interjection)	ouch , oops , here
2. คำกริยา (Verb)	is , am, were , have	6. คำผสม (Compound)	somebody , anybody
3. คำบุพบท (Preposition)	on , at , in , from	7.เครื่องหมายและตัวเลข (Symbol and number)	0-9 , ! , ? , .<> { } [ ] +
4. คำสันธาน (Conjunction)	and , but , therefore	8. คำหยุดเพิ่มเติม (customize)	ผู้ใช้สามารถกำหนดคำเพิ่มเติมได้ด้วยตนเอง

การหารากศัพท์ (Stemming)

การหารากศัพท์ถือว่าเป็นตัวเลือกหนึ่งที่ผู้ใช้สามารถเลือกในการวิเคราะห์เอกสาร ซึ่งเป็นขั้นตอนการหารูปแบบเดิมของคำหรือหาคำที่มีความหมายคล้ายกัน เนื่องจากคำเหล่านั้นมีรากศัพท์เป็นคำเดียวกัน โดยในภาษาอังกฤษมีคำที่มีความหมายเหมือนกัน แต่มีรูปแบบของคำต่างกันโดยเติมคำปัจจัย (Suffixes) ต่อท้าย เช่น “connected”, “connecting”, “connection” และ “connections” จากคำตัวอย่างดังกล่าวจะมีความหมายเหมือนกันและมาจากรากศัพท์คำว่า “connect” ซึ่งงานวิจัยนี้ได้ใช้อัลกอริทึมที่นิยมใช้ในการหารากศัพท์ของคำภาษาอังกฤษ (วิรัตน์ ชูนุ้ย, 2555) คือ อัลกอริทึม Porter (Porter, 1980)และตัวอย่างการหารากศัพท์จากคำที่มีรูปแบบต่างกัน แสดงดังตารางที่ 2 ตารางคำที่ผ่านการหารากศัพท์ด้วยอัลกอริทึม Porter

ตารางที่ 2 ตารางคำที่ผ่านการหารากศัพท์ด้วยอัลกอริทึม Porter

ก่อนหารากศัพท์	หลังหารากศัพท์
Connected , Connecting , Connection	Connect
Sleeping	Sleep
Hopefulness	hopeful

การให้ค่าน้ำหนักคำด้วย TF-IDF

การให้ค่าน้ำหนักเป็นขั้นตอนการแทนค่าของคำในเอกสารด้วยค่าน้ำหนัก เพื่อใช้ในประเมินค่าความสำคัญของคำเหล่านั้น ในงานวิจัยนี้ได้ให้ค่าน้ำหนักคำด้วย TF-IDF (Term Frequency - Inverse Document Frequency) ซึ่งเป็นการพิจารณาค่าความสำคัญของคำจากความถี่ของคำที่ปรากฏในแต่ละเอกสาร โดยให้ค่าน้ำหนักของคำที่ปรากฏในเอกสารหลายๆ เอกสาร ถ้าค่า TF-IDF ของคำใดมีค่าน้ำหนักต่ำจะแสดงให้เห็นว่าคำนั้นไม่สามารถแสดงถึงลักษณะเฉพาะเป็นตัวแทนของเอกสารนั้นๆ ได้ โดยวิธีการคำนวณหาค่าน้ำหนัก W_{jk}ด้วยค่า TF-IDF แสดงดังสมการด้านล่าง (วิรัตน์ ชูนุ้ย, 2555)

$$W_{jk} = TF_{jk} \times IDF_{k} \quad (1)$$

โดยที่ W_{jk} คือ ค่าน้ำหนักของคำที่ k ปรากฏอยู่ในเอกสารที่ j ,TF_{jk}คือ ความถี่ของคำที่ k ปรากฏอยู่ในเอกสารที่ j โดยที่ F_{jk}เป็นจำนวนความถี่ของคำที่ j ในเอกสาร k

$$TF_{jk} = F_{jk} IDF_{k} \quad (2)$$

โดยที่ IDF คือส่วนกลับของความถี่เอกสารที่มีคำที่ k ปรากฏอยู่

$$IDF_{k} = \log \frac{N}{DFk} IDF_{k} \quad (3)$$

DF คือ ความถี่เอกสารที่มีคำที่ k ปรากฏอยู่และ N คือ จำนวนเอกสารทั้งหมด

การสร้างคุณลักษณะ (Feature Extraction)

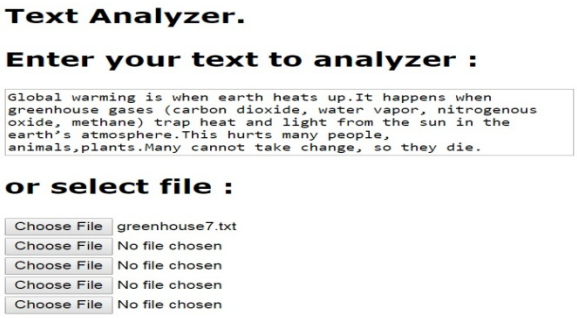
การสร้างคุณลักษณะเป็นขั้นตอนการดึงคุณลักษณะของเอกสาร เพื่อใช้เป็นตัวแทนในการส่งไปขั้นตอนกระบวนการสร้างคุณลักษณะ โดยจะต้องมีการกำหนดว่าจะใช้ค่าใดแทนคุณลักษณะของเอกสารนั้น ๆ ซึ่งอาจจะใช้ค่าความจริง (Boolean) ค่าความถี่ (Frequency) วลีหรือประโยค คุณลักษณะที่สกัดได้นั้นจะถูกจัดรูปแบบให้อยู่ในลักษณะข้อมูลเวกเตอร์ สำหรับงานวิจัยนี้ได้ทำการดึงคุณลักษณะของเอกสาร การแทนคุณลักษณะของเอกสารโดยใช้ค่าความจริง (Boolean) มีค่าเป็น 0 และ 1 โดยที่ 0 หมายถึง ไม่มีคุณลักษณะ 1 หมายถึง มีคุณลักษณะและค่าความถี่ (Frequency) ในการแทน จากนั้นนำคุณลักษณะที่สร้างขึ้นมาใช้กับการเรียนรู้ด้วยเครื่องจักรกล (นิเวศและคณะ, 2554)

การสร้างคำสำคัญ (KeywordExtraction)

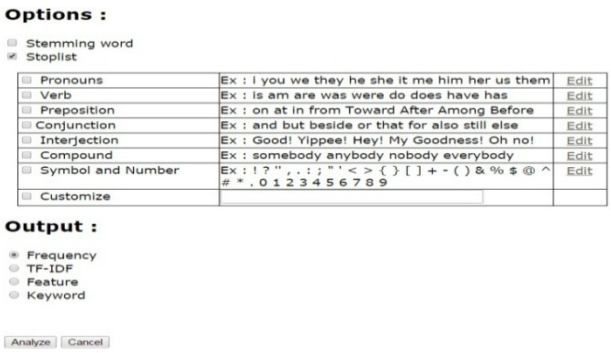
คำสำคัญ หมายถึง คำที่กำหนดขึ้นมาเพื่อใช้แทนเนื้อหาของบทความหรือเอกสาร ซึ่งต้องเป็นคำที่สั้นแต่ได้ใจความ มีความหมายเป็นคำนามหรือเป็นศัพท์เฉพาะในงานนี้ได้สร้างคำสำคัญโดยการคำนวณหาค่าน้ำหนักคำด้วยค่า TF-IDF ของทุกเอกสาร และเลือกค่าสูงสุด 1 ค่า เป็นตัวแทนของทุกเอกสาร เพื่อบอกคำสำคัญของเอกสารที่รับเข้ามา

ผลการศึกษา

จากการทดลองโดยการใส่ข้อความลงในกล่องรับข้อความและเลือกไฟล์ที่ต้องการวิเคราะห์ซึ่งสามารถเลือกใส่ข้อมูลเป็นไฟล์ได้จำนวน 5 ไฟล์ โดยตัวอย่างดังภาพที่2จากกล่องข้อความ (Text Box) ได้นำเข้าข้อมูลเกี่ยวกับภาวะโลกร้อนและในส่วนของการเลือกไฟล์ได้นำเข้าข้อมูลเกี่ยวกับการเพิ่มขึ้นของอุณหภูมิอย่างรวดเร็ว ส่วนตัวเลือกที่ใช้ในการวิเคราะห์ ผู้ใช้สามารถเลือกใช้การหารากศัพท์และการการกำจัดคำหยุดได้ รายการคำหยุดแบ่งออกเป็น 8 ประเภท โดยผู้ใช้สามารถเลือกดูหรือและลบรายการคำศัพท์ในแต่ละประเภทได้ รูปแบบในการแสดงผลของเว็บไซต์แบ่งเป็น 4 แบบ ได้แก่ ค่าความถี่ของคำ การให้น้ำหนัก การสกัดคุณลักษณะและการสร้างคำสำคัญให้กับเอกสาร ซึ่งหน้าจอที่ใช้ในเลือกตัวเลือกและการแสดงผลของการวิเคราะห์แสดงดังภาพที่ 3



ภาพที่ 2 หน้าจอหลักของเว็บไซต์



ภาพที่ 3 หน้าจอตัวเลือกในการวิเคราะห์และตัวเลือกการแสดงผล

การแสดงผลความถี่ของคำผู้ใช้สามารถเลือกเรียงจากค่าความถี่สูงสุด (Top Rank) ในแต่ละทุกเอกสารหรือรวมทั้งหมด แสดงดังภาพที่ 4 โดยมีเงื่อนไขของ Top Rank ที่ใช้ในการแสดงผล ได้แก่ 10, 20, 30, 40, 50, 60, 70, 200, All

Frequency and Top in all Documents : 10

No.	Word	Rank	Occurrences	Frequency (%)
1.	temperature	1	6	3.41 %
2.	when	1	6	3.41 %
3.	heat	1	6	3.41 %
4.	is	2	5	2.84 %
5.	This	2	5	2.84 %
6.	and	2	5	2.84 %
7.	light	2	5	2.84 %
8.	greenhouse	3	4	2.27 %
9	atmosphere	3	4	2.27 %
10.	rises	3	4	2.27 %

ภาพที่ 4 หน้าจอแสดงผลการวิเคราะห์ในรูปแบบความถี่ของคำ

การแสดงค่าน้ำหนักของคำด้วยTF-IDF ของแต่ละคำในเอกสารได้โดยการหาค่าน้ำหนักของคำด้วยTF-IDFของแต่ละเอกสารนั้น ทั้งนี้เอกสารดังภาพที่ 5 จะแสดงค่า TF-IDF ของข้อมูลเข้าทั้ง 2 เอกสาร เพื่อประเมินค่าความสำคัญของคำนั้นๆ ในกลุ่มเอกสาร

TF-IDF of Top Rank :

Term	TF-IDF (Document -- #1)	TF-IDF (Document -- #2)
temperature	0	2.862
when	0	1.204
heat	0.301	1.985
is	0.301	1.204
This	0.301	1.204
and	0.301	1.204
light	0.301	1.204
greenhouse	0.301	0.528
atmosphere	0.301	0.528
rises	0.301	0.528

ภาพที่ 5 หน้าจอแสดงผลการวิเคราะห์ในรูปแบบค่าน้ำหนักของคำด้วย TF-IDF

Feature Extraction value on : Boolean ▾ Save file as CSV

Feature Extraction in all documents.

	temperature	when	heat	is	This	and	light	greenhouse	atmosphere
S[1]	1	1	0	1	0	0	0	0	0
S[2]	1	1	1	0	0	1	1	1	1
S[3]	0	0	0	0	1	0	0	0	0
S[4]	0	0	0	0	0	0	0	0	0
S[5]	1	1	0	1	0	0	0	0	0
S[6]	1	0	1	0	0	1	1	1	1
S[7]	0	0	0	0	1	0	0	0	0
S[8]	0	0	0	0	0	0	0	0	0
S[9]	1	1	1	1	0	1	1	1	1

ภาพที่ 6 หน้าจอแสดงผลการวิเคราะห์ในรูปแบบค่าการสร้างคุณลักษณะของคำ

S[n] เมื่อ S คือลำดับของประโยค nคือ จำนวนประโยคที่ปรากฏในเอกสาร แสดงดังภาพที่ 6

การแสดงผลคำสำคัญของเอกสารทั้งหมด โดยพิจารณาจากการให้ค่าน้ำหนักคำด้วยTF-IDF ที่สูงที่สุดจากทุกเอกสารมาเป็นคำสำคัญ ดังนั้นจากข้อมูลในภาพที่ 5 คำว่า “Temperature” จึงเป็นคำสำคัญของเอกสาร

จากผลการประเมินความพึงพอใจผู้ใช้งานด้วยแบบประเมิน โดยกลุ่มเป้าหมายของการประเมินคือ นักวิจัย นิสิต นักศึกษาที่เรียนวิชาการประมวลผลภาษาธรรมชาติในระดับปริญญาตรี และผู้สนใจเกี่ยวกับการวิเคราะห์เอกสารจำนวน 42 ราย ผลระดับความพึงพอใจตามค่าสถิติอยู่ในระดับดี ผลการประเมินโดยรวม พบว่ามีความพึงพอใจในระดับดี (ค่าเฉลี่ย $\bar{x}$ = 3.95, SD = 0.909) โดยความพึงพอใจรายข้อเรียงลำดับมากที่สุดดังนี้ การจัดลำดับการทำงาน ของเว็บไซต์ มีค่าค่าเฉลี่ย$\bar{x}$เท่ากับ 4.12 และค่าเบี่ยงเบนมาตรฐานSD เท่ากับ 0.803 ลำดับรองลงมาดังนี้ เว็บไซต์มีประโยชน์ต่อผู้ใช้งานและสามารถนำไปประยุกต์ได้ มีค่าค่าเฉลี่ย$\bar{x}$เท่ากับ 4.07 และค่าเบี่ยงเบนมาตรฐาน SD เท่ากับ 0.894 ลำดับถัดมาดังนี้ เว็บไซต์ตรงตามความต้องการของผู้ใช้และเว็บไซต์มีความน่าสนใจ มีค่าเฉลี่ย$\bar{x}$ เท่ากับ 3.88 และค่าเบี่ยงเบนมาตรฐานSD เท่ากับ 0.968 และเท่ากับ 1.017 ตามลำดับ สุดท้ายเว็บไซต์ใช้งานง่ายไม่ซับซ้อน มีค่าเฉลี่ย$\bar{x}$เท่ากับ 3.81 และมีค่าค่าเฉลี่ยSD เท่ากับ0.862

สรุป

งานวิจัยนี้เป็นการนำเสนอเว็บไซต์ช่วยวิเคราะห์หรือเอกสาร เพื่อใช้เป็นเครื่องมือในการวิเคราะห์เอกสารในเชิงสถิติ โดยมุ่งเน้นให้นิสิต นักศึกษาที่เรียนวิชาการประมวลผลภาษาธรรมชาติในระดับปริญญาตรี และผู้สนใจเกี่ยวกับการวิเคราะห์เอกสารเข้าใจขั้นตอนพื้นฐานในกระบวนการก่อนประมวลผลสำหรับข้อความโดยแสดงผลเป็นค่าสถิติ ได้แก่ การหาความถี่ของคำ การหาความสัมพันธ์ของคำ การหารากศัพท์ และการกำจัดคำหยุด การให้ค่าน้ำหนักคำด้วยTF-IDF นอกจากนี้งานวิจัยนี้ยังสามารถสร้างคุณลักษณะของคำจากเอกสารที่เป็นข้อมูลเข้าของเว็บไซต์ โดยมีการแทนค่าคำภายในเอกสารได้ 2 แบบ ได้แก่ การแทนด้วยค่าคุณลักษณะของคำและด้วยค่าความถี่ ทั้งนี้เพื่ออำนวยความสะดวกต่อนักวิจัยในการนำไปใช้ในกระบวนการวิจัยขั้นต่อไป ผู้ใช้สามารถเลือก การแสดงผลการสร้างคุณลักษณะดังกล่าวให้แสดงบนหน้าเว็บไซต์ หรือเลือกบันทึกผลการวิเคราะห์เป็นไฟล์ในรูปแบบ CSV ที่นำไปใช้กับซอฟต์แวร์ด้านเหมืองข้อมูลต่อไป จากผลการประเมินความพึงพอใจผู้ใช้งานแสดงให้เห็นว่าเว็บไซต์ช่วยวิเคราะห์เอกสารและสร้างคุณลักษณะในระดับดี ซึ่งตรงกับวัตถุประสงค์กับของงานวิจัยนี้

ส่วนการพัฒนางานวิจัยนี้ในอนาคตสามารถเพิ่มประสิทธิภาพในการวิเคราะห์ข้อความในเอกสารมากขึ้นได้ เนื่องจากงานวิจัยนี้พิจารณาเพียงรูปคำ(surface form)เท่านั้นจึงควรมีการเพิ่มความรู้ให้กับระบบมากขึ้นตัวอย่างเช่น ข้อมูลทางไวยากรณ์ของคำหรือวลี เช่น หน้าที่ของคำ(Part of speech)การระบุขอบเขตของวลี การระบุนิพจน์ระบุนาม(Named Entities)การเพิ่มข้อมูลเชิงความหมายให้กับคำ เช่น การหาคำที่มีความหมายเหมือนกันแต่มีรูปแบบของคำไม่เหมือนกัน(synonym)การหาความสัมพันธ์ระหว่างคำศัพท์ จากฐานข้อมูล Word-Net เป็นต้น

เอกสารอ้างอิง

นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์และพยุง มีสัจ.2554.การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสาร ภาษาไทยแบบอัตโนมัติ. วารสารพัฒนบริหารศาสตร์ ปี 51 ฉบับที่ 3/2554.

วิรัตน์ ชูนุ้ย.2555.การลดขนาดลักษณะเฉพาะโดยใช้ FCA สำหรับการจำแนกประเภทเว็บเพจ.

วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยสงขลานครินทร์.

KranitGhag,Ketan Shah.2014.Sentiment Classification using Relative Term Frequency Inverse Document Frequency.Research.Internation Journal of Advanced Computer Science and Applications.Vol.5.No.2.

LaddaSuanmali , Mohammed Salem Binwahlan and NaomieSalim.2009. Sentence Features Fusion for Text Summarization Using Fuzzy Logic .Research. Ninth International Conference on Hybrid Intelligent Systems.

Leena. H. Patil , Mohammed Atique .2012. A Novel Approach for Feature Selection Method TFIDF in Document Clustering.Research. 2013 3rd IEEE International Advance Computing Conference(IACC).

MF.Porter.1980.AnAn algorithm for suffix stripping.Program.no. 3, pp 130-137, July 1980.

Textalyser.[สืบค้นเมื่อ 5 พฤศจิกายน 2558] จาก http://textalyser.net/