ระบบจัดหมวดหมู่สถานที่ท่องเที่ยวด้วยคำอธิบายภาษาไทยแบบอัตโนมัติผ่านเว็บเซอร์วิส
Automatic Tourist Attraction Categorization System using Thai Description via Web Service

ชูพันธุ์ รัตนโภคา (Choopan Rattanapoka)[1] และเมธาวี สุทธิกุล (Mathawee Sutikun)[2]

[1]ภาควิชาเทคโนโลยีวิศวกรรมอิเล็กทรอนิกส์ วิทยาลัยเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

[2]ภาควิชาเทคโนโลยีวิศวกรรมอิเล็กทรอนิกส์ วิทยาลัยเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

choopanr@kmutnb.ac.th, mathawee.1990@hotmail.com

บทคัดย่อ

บทความนี้เสนอการออกแบบและพัฒนาระบบจัดหมวดหมู่สถานที่ท่องเที่ยวแบบอัตโนมัติ จากคำอธิบายสถานที่ท่องเที่ยวที่เป็นภาษาไทยด้วยภาษาจาวา โดยข้อมูลคำอธิบายสถานที่ท่องเที่ยว จะถูกส่งเข้ามายังระบบผ่านช่องทางการให้บริการของ เว็บเซอร์วิส ซึ่งหมวดหมู่สถานที่ท่องเที่ยวที่ระบบรองรับในการจัดหวดหมู่ มีทั้งหมด 4 หมวดหมู่ คือ วัด ทะเล ภูเขา และที่พัก ด้วยการประยุกต์ใช้วิธีการเรียนรู้ของเครื่องผ่านโปรแกรมเวก้า จากผลการทดลองโดยการทดสอบกับสถานที่ท่องเที่ยวในจังหวัดภูเก็ตพบว่า ระบบมีความแม่นยำในการจัดหมวดหมู่สถานที่ท่องเที่ยวอยู่ที่ 95%

คำสำคัญ: เวก้า การจำแนกหมวดหมู่ การเรียนรู้ของเครื่อง เว็บเซอร์วิส

Abstract

This paper presents a design and implementation of tourism spots categorized system using Thai description of the tourist attractions. The information of tourist attraction input to our system via web service. Then, our system automatically classifies them into the properly category by using machine learning algorithm via Weka. Our system supports 4 tourism spots categories: temple, beach, mountain and accommodation. In the experiments, we use tourist attractions in Phuket and found out that our system has about 95% accuracy for the tourism attraction categorization.

Keyword: Weka, Classification, Machine Learning, Web Service

1. บทนำ

ปัจจุบันอุตสาหกรรมการท่องเที่ยวในประเทศไทย ถือว่าเป็นรายได้หลักอย่างหนึ่งของประเทศ การรวบรวมข้อมูลและการประชาสัมพันธ์สถานที่ท่องเที่ยวเพื่อให้นักท่องเที่ยวสามารถค้นหาข้อมูลได้อย่างสะดวกนั้นควรจะมีการจัดหมวดหมู่ให้กับสถานที่ท่องเที่ยวอย่างเป็นระบบ และการที่จะให้ได้มาซึ่งสถานที่ท่องเที่ยวและข้อมูลเกี่ยวกับสถานที่ต่างๆ เข้ามาเก็บในฐานข้อมูลนั้น จะสะดวกมากขึ้นถ้าเปิดโอกาสให้บุคคลทั่วไปสามารถช่วยกันเพิ่มข้อมูลต่างๆ เกี่ยวกับสถานที่ท่องเที่ยวได้ แต่อย่างไรก็ตามเมื่อมีการนำเข้าข้อมูลเกี่ยวกับสถานที่ท่องเที่ยวเป็นจำนวนมาก ซึ่งผู้นำเข้าข้อมูลสถานที่ท่องเที่ยวอาจจะไม่ได้จำแนกหมวดหมู่สถานที่ท่องเที่ยวมาด้วยหรือใส่หมวดหมู่ผิด จะทำให้เกิดภาระกับผู้ดูแลระบบที่ทำหน้าที่ตรวจสอบและจัดหมวดหมู่สถานที่ท่องเที่ยว ยิ่งไปกว่านั้นถ้าผู้ดูแลระบบไม่เคยไปสถานที่นั้นๆ มาก่อน จะยิ่งทำให้การตัดสินใจในการตรวจสอบและจำแนกหมวดหมู่ของสถานที่ท่องเที่ยวลำบากมากยิ่งขึ้น

ดังนั้นเพื่อให้การจัดสถานที่ท่องเที่ยวไปตามหมวดหมู่ที่ถูกต้อง อีกทั้งยังลดภาระและข้อผิดพลาดให้กับผู้ดูแลระบบในการจัดหมวดหมู่สถานที่ท่องเที่ยว งานวิจัยนี้จึงต้องการนำเสนอการออกแบบและพัฒนาระบบจัดหมวดหมู่สถานที่ท่องเที่ยวแบบอัตโนมัติด้วยวิธีการเรียนรู้ของเครื่องผ่านโปรแกรม Weka จากคำอธิบายสถานที่ท่องเที่ยวภาษาไทย โดยให้บริการในรูปแบบของเว็บเซอร์วิส โดยขอบเขตของระบบต้นแบบใช้สำหรับสถานที่ท่องเที่ยวภายในจังหวัดภูเก็ต และรองรับหมวดหมู่สถานที่ท่องเที่ยวได้ทั้งหมด 4 หมวดหมู่ คือ วัด ทะเล ภูเขา และที่พัก

2. วรรณกรรมที่เกี่ยวข้อง

การจัดแบ่งหมวดหมู่ข้อความ [1], [2] คือ การระบุประเภทของเอกสารหรือข้อความแบบอัตโนมัติโดยอิงจากตัวเนื้อหาในเอกสารหรือข้อความนั้น ๆ การนำการจัดหมวดหมู่ข้อความมาช่วยใช้การจัดหมวดหมู่ของสถานที่ท่องเที่ยว จะเป็นการค้นหารูปแบบและความสัมพันธ์ในชุดข้อมูล โดยมีการนำเทคนิคในการเรียนรู้ของเครื่อง (Machine Learning) มาประยุกต์ใช้และสร้างแบบจำลองเพื่อทำนายข้อมูลของสถานที่ท่องเที่ยวที่เข้ามาใหม่ ว่าควรจัดอยู่ในหมวดหมู่ใด

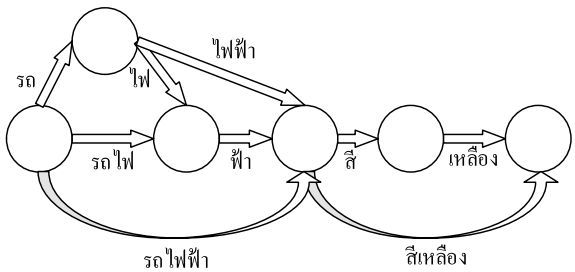
งานวิจัยที่เกี่ยวข้องกับการจัดหมวดหมู่สถานที่ท่องเที่ยวที่พบมีมากมายเช่น [3] ที่นำข้อมูลเมตาดาต้า (Meta data) ของรูปภาพสถานที่ท่องเที่ยวซึ่งอยู่ในรูปแบบข้อความ ที่ได้จากเว็บไซต์ผู้ให้บริการอัพโหลดรูปภาพฟลิคเกอร์ (Flickr) มาทำการจำแนกและระบุสถานที่ เพื่อนำมาวิเคราะห์ และแนะนำตารางการเดินทางภายในหนึ่งวันว่าไปเที่ยวสถานที่ใดได้บ้าง หรือใน [4] เป็นงานวิจัยที่มุ่งเน้นการจัดทำพจนานุกรมทางด้านภูมิศาสตร์ ซึ่งมีการจัดหมวดหมู่สถานที่ทางด้านภูมิศาสตร์ เช่น ภูเขา แม่น้ำ โบราณสถาน ฯลฯ โดยนำข้อมูลจากฟลิคเกอร์ และวิกิพีเดีย มาใช้เป็นข้อมูลประกอบการพิจารณา นอกจากนี้ยังมีงานวิจัยเกี่ยวกับการสร้างแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้เทคนิคการเรียนรู้ของเครื่อง [5] ที่ใช้จัดหมวดหมู่สถานที่ท่องเที่ยวจากคำอธิบายสถานที่ท่องเที่ยว ซึ่งนำข้อมูลมาจากเว็บไซต์โลนลี่ แพลนเนตและได้นำเสนอแบบจำลองหลายประเภท แต่อย่างไรก็ตามงานวิจัยดังกล่าวรองรับกับการจัดหมวดหมู่สถานที่ท่องเที่ยวด้วยคำอธิบายสถานที่ท่องเที่ยวด้วยภาษาอังกฤษเท่านั้น

ดังนั้นพื้นฐานในการออกแบบและพัฒนาระบบจัดหมวดหมู่สถานที่ท่องเที่ยวด้วยคำอธิบายภาษาไทยแบบอัตโนมัตินั้น มีการใช้ทฤษฎีพื้นฐานและเครื่องมือช่วยหลักๆ 3 ส่วน ได้แก่ 1) LexTo ซึ่งเป็นเครื่องมือที่ใช้สำหรับตัดแบ่งคำภาษาไทยจากคำอธิบายสถานที่ท่องเที่ยว เพื่อนำไปใช้ในการนับจำนวนคำสำคัญที่มีในคำอธิบาย 2) โปรแกรม Weka ซึ่งใช้ในการสร้างและเรียกใช้งานแบบจำลองที่ใช้ในการเรียนรู้ของเครื่องให้สามารถจัดหมวดหมู่ให้กับสถานที่ท่องเที่ยวตามคำอธิบายได้ และ 3) เทคนิคในการเรียนรู้ของเครื่องด้วยเทคนิคซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine: SVM)

2.1 LexTo (Thai Lexeme Tokenizer)

LexTo (http://www.sansarn.com/lexto/) เป็นเครื่องมือที่ใช้ในการแบ่งคำภาษาไทยที่ถูกพัฒนาขึ้นด้วยภาษาจาวา และมีการแจกจ่ายในรูปแบบของโอเพนซอร์ส ซึ่งมีต้นแบบมาจาก Lexitron เพื่อนำมาใช้ในการตัดแบ่งคำภาษาไทยโดยเฉพาะ โดย LexTo ใช้เทคนิคการแบ่งคำแบบคำที่รู้จักที่ยาวที่สุด (Longest Matching) หรือเรียกอีกอย่างหนึ่งว่าวิธีแบ่งคำแบบฐานพจนานุกรม (Dictionary-based) ซึ่งเป็นเทคนิคการแบ่งคำโดยเลือกคำที่ยาวที่สุดออกมาจากพจนานุกรม โดยพจนานุกรมที่กล่าวถึงนี้สามารถแก้ไข เพิ่มเติม และเปลี่ยนแปลงคำที่ต้องการตัดแบ่งได้อย่างอิสระ

หลักการทำงานของ Lexto คือ Lexto จะอ่านคำทั้งหมดจากพจนานุกรมเข้ามาในหน่วยความจำก่อน จากนั้นเมื่อมีการส่งข้อความที่ต้องการตัดแบ่งคำภาษาไทยเข้ามา Lexto จะนำคำในพจนานุกรมมาเทียบกับข้อความที่ถูกส่งเข้ามาโดยพิจารณาตัดแบ่งจากคำในพจนุกรมที่ยาวที่สุด ดังตัวอย่างในภาพที่ 1 เมื่อต้องการแบ่งคำว่า “รถไฟฟ้าสีเหลือง” คำที่เจอในพจนานุกรมคือ “รถ” “รถไฟ” และ “รถไฟฟ้า” ซึ่งเป็นคำที่ยาวที่สุดจึงแบ่งคำแรกที่ “รถไฟฟ้า” จากนั้นจึงแบ่งคำถัดไป พบคำว่า “สี” และ “สีเหลือง” จึงแบ่งคำจากคำที่ยาวที่สุดซึ่งก็คือ “สีเหลือง” ดังนั้นเมื่อส่งความ “รถไฟฟ้าสีเหลือง” ให้กับ Lexto ก็จะได้คำ 2 คำออกมาคือ “รถไฟฟ้า” และ “สีเหลือง”



ภาพที่ 1: การตัดแบ่งคำภาษาไทยใน Lexto

2.2 โปรแกรม Weka

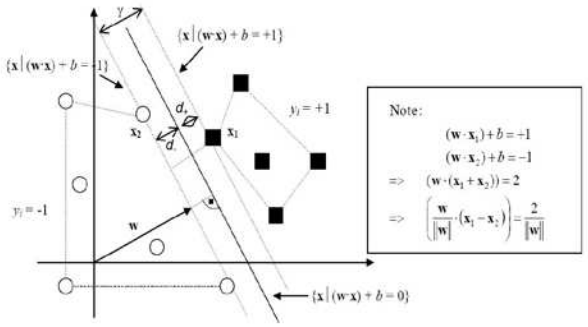
โปรแกรม Weka (Waikato Environment for Knowledge Analysis) เริ่มพัฒนามาตั้งแต่ปี 1997 โดยมหาวิทยาลัย Waikato ประเทศนิวซีแลนด์ ซึ่งเป็นฟรีแวร์ที่พัฒนาด้วยภาษาจาวา สามารถทำงานได้บนระบบปฏิบัติการหลักๆ ทุกตัว โดยเน้นกับ

งานทางด้านการเรียนรู้ของเครื่อง (Machine Learning) และการทำเหมืองข้อมูล (Data Mining) อีกทั้งยังมีส่วนติดต่อสำหรับการเขียนโปรแกรม (API) ที่ผู้พัฒนาโปรแกรมต่างๆ สามารถติดต่อเรียกใช้งาน Weka ได้โดยตรง

Weka ให้บริการหน้าที่การทำงานพื้นฐานต่างๆ เกี่ยวกับการทำเหมืองข้อมูลอย่างมากมาย เช่น data processing, clustering, classification, regression, visualization และ feature selection รวมถึงยังเปิดช่องทางให้ผู้พัฒนาโปรแกรมสามารถนำอัลกอริธึมที่พัฒนาขึ้นมาเอง มาใช้ใน Weka ได้อีกด้วย

2.3 ซัพพอร์ตเวคเตอร์แมชชีน

หลักการของซัพพอร์ตเวคเตอร์แมชชีน คือ การสร้างไฮเปอร์เพลนที่เหมาะสมบนระนาบของข้อมูลตัวอย่าง (Training data) เพื่อแบ่งแยกกลุ่มข้อมูลที่แตกต่างกัน ในการสร้างไฮเปอร์เพลนที่เหมาะสม มีการนิยามระยะห่างระหว่างจุดของข้อมูลที่อยู่ใกล้กับไฮเปอร์เพลนมากที่สุดทั้งสองด้าน คือ d_+ และ d_- โดยระยะมาร์จิน (Margin) γ เกิดจากระยะ $d_+ + d_-$ ไฮเปอร์เพลนที่เหมาะสมคือไฮเปอร์เพลนที่มีค่า มาร์จิน γ กว้างที่สุดดังแสดงในภาพที่ 2 โดยข้อมูลตัวอย่างที่อยู่บนขอบของมาร์จิน γ จะถูกเรียกว่า *ซัพพอร์ตเวคเตอร์*



ภาพที่ 2: การแบ่งกลุ่มข้อมูล ■ และ ○ ด้วยไฮเปอร์เพลนใน SVM

จากภาพที่ 2 เป็นการแบ่งกลุ่มข้อมูลแบบ 2 กลุ่ม (Binary Classification Problem) กำหนดให้กลุ่มข้อมูลที่ใช้ในการฝึกสอน (Training Dataset) ประกอบด้วย l ตัวอย่าง (Samples) ซึ่งสามารถแสดงอยู่ในรูป $\{x_k, y_k\}, k = 1,\ldots,l$ และ $x_k \in \Re^n$, $y_k \in \{-1,+1\}$ โดย x_k จะเป็น อินพุทเวคเตอร์ ในขณะที่ y_k จะเป็นคลาสของข้อมูล (Class Label) หลักการของซัพพอร์ตเวคเตอร์แมชชีน คือ การสร้างไฮเปอร์เพลนที่เหมาะสมบนระนาบของข้อมูลตัวอย่าง เพื่อแบ่งกลุ่มของข้อมูลทั้งสอง โดยไฮเปอร์เพลนดังกล่าวจะถูกกำหนดโดยพารามิเตอร์ $(\mathbf{w}, b)$ โดย $\mathbf{w}$ เป็นเวคเตอร์ที่ตั้งฉากกับไฮเปอร์เพลน และ b จะเป็นค่าคงที่ ซึ่งกำหนดตำแหน่งของเวคเตอร์ที่สัมพันธ์กับตำแหน่งดั้งเดิมในปริภูมิอินพุท (Input Space) สมการของไฮเปอร์เพลนแบบเชิงเส้น (Linear Hyperplane) จะถูกกำหนดโดยสมการ $(\mathbf{w}\cdot\mathbf{x}) + b = 0$ และเพื่อลดปัญหาในเรื่องของสเกล $\mathbf{w}$ และ b จะถูกกำหนดด้วยสมการ $|(\mathbf{w}\cdot\mathbf{x}) + b| = 1$ สำหรับจุดที่อยู่ใกล้ไฮเปอร์เพลนมากที่สุด ดังนั้นสมการของไฮเปอร์เพลนแสดงได้ดังสมการที่ (1)

$$y_i[(\mathrm{w}\cdot\mathrm{x}_i)+b]\geq 1 \qquad \forall i \tag{1}$$

ดังที่กล่าวข้างต้นการฝึกสอนด้วยเทคนิค SVM จะเป็นการคำนวณหาไฮเปอร์เพลนที่มีค่ามาร์จินกว้างที่สุดซึ่งสามารถหาได้จากการทำให้ค่า $\mathbf{w}$ มีค่าที่น้อยที่สุด โดยปัญหาดังกล่าวสามารถหาคำตอบได้โดยใช้วิธี Lagrange Multipliers ดังสมการที่ (2) โดยตัวแปร $\alpha_i \geq 0$ จะถูกเรียกว่า Positive Lagrange Multipliers

$$L(\mathrm{w},b,\alpha) = \frac{1}{2}\|\mathrm{w}\|^2 - \sum_{i=1}^{l}\alpha_i y_i(x_i\cdot\mathrm{w}+b) + \sum_{i=1}^{l}\alpha_i \tag{2}$$

จากสมการที่ (2) สิ่งที่ต้องการคือให้ค่าของ $L(\mathbf{w},b,\alpha)$ มีค่าน้อยที่สุด โดยเทียบกับค่าของ $\mathbf{w}$ และ b ในขณะเดียวกันค่าของ $L(\mathbf{w},b,\alpha)$ จะต้องมีค่ามากที่สุดเมื่อเทียบกับ $\alpha_i \geq 0$ (Dual Variables) โดยปัญหาดังกล่าวสามารถหาคำตอบได้โดยอาศัยวิธี Wolfe Dual ดังนั้นสมการที่ใช้เพื่อหาคำตอบของสมการที่ (2) สามารถแสดงได้ดังนี้

$$\text{Maximize } W(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(\mathrm{x}_i\cdot\mathrm{x}_j) \tag{3}$$

$$\text{Subject to (1) } \sum_{i=1}^{l}\alpha_i y_i = 0, \text{ and (2) } \alpha_i \geq 0 \text{ for } i = 1,\ldots,l.$$

จากสมการที่ (3) ตัวอย่างข้อมูลที่ใช้ในการฝึกสอนที่ทำให้ค่าของ $\alpha_i > 0$ จะถูกเรียกว่า *ซัพพอร์ตเวคเตอร์* ซึ่งจะวางตัวอยู่บนขอบของมาร์จิน สำหรับข้อมูลตัวอย่างที่เหลือจะมีค่าของ $\alpha_i = 0$ สามารถจะถูกตัดทิ้งได้โดยไม่ก่อให้เกิดผลกระทบต่อไฮเปอร์เพลนที่สร้างขึ้น จากอัลกอริธึมที่กล่าวมาข้างต้น จะเหมาะสำหรับในกรณีที่กลุ่มข้อมูลตัวอย่างสามารถถูกแบ่งแยกได้ด้วยไฮเปอร์เพลนแบบเชิงเส้นเท่านั้น ดังนั้นเพื่อให้

อัลกอริธึมดังกล่าวสามารถแบ่งแยกกลุ่มข้อมูลที่มีลักษณะไม่เป็นเชิงเส้น (Nonlinear Dataset) จึงจำเป็ะต้องแปลงกลุ่มข้อมูลตัวอย่างไปสู่ปริภูมิมิติที่สูงขึ้น (Higher Dimensional Space) เรียกว่าปริภูมิ Feature (Feature Space) โดยการแปลงดังกล่าวจะกระทำผ่านฟังก์ชันที่ไม่เป็นเชิงเส้นดังนี้ $\Phi: \Re^n \to F.$ กำหนดให้ Φ แสดงการแมพ (Mapping) จากปริภูมิของกลุ่มข้อมูลตัวอย่างไปสู่ปริภูมิ Feature โดยในที่นี้จะต้องรวมขั้นตอนดังกล่าวเข้ากับสมการที่ (3) ซึ่งสามารถทำได้โดยการแทนที่ทุกค่าของ $\mathbf{x}_i$ ด้วย $\Phi(\mathbf{x}_i)$ โดยในที่นี้เราจะหลีกเลี่ยงการคำนวณค่า $\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)$ โดยตรงเนื่องจากเป็นขั้นตอนที่ต้องใช้เวลาในคำนวณสูง แต่เราจะใช้เทคนิคของฟังก์ชันเคอร์แนล (Kernel Function) ในการคำนวณแทน โดยฟังก์ชันเคอร์แนลสามารถนิยามได้ดังนี้ $K(\mathbf{x}_i,\mathbf{x}_j)=\Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j).$ แทนที่เคอร์แนล K ลงในสมการที่ (3) ดังนั้นสามารถแสดงสมการซึ่งใช้ในการหาค่าของไฮเปอร์เพลนได้ดังสมการที่ (4) โดย C จะเป็นค่าคงที่เพื่อใช้ในการปรับหรือชดเชยระหว่างค่าผิดพลาดของการฝึกสอน

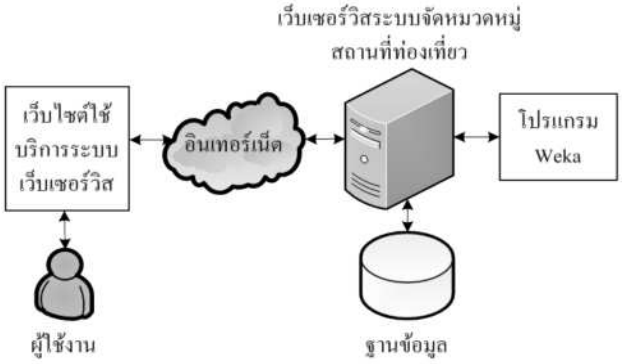
$$\text{Maximize } W(\boldsymbol{\alpha})=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \qquad (4)$$

Subject to (1) $\sum_{i=1}^{l}\alpha_i y_i=0$, and (2) $0\le\alpha_i\le C \quad \forall i$.

3. วิธีการดำเนินการวิจัย

ระบบจัดหมวดหมู่สถานที่ท่องเที่ยวในงานวิจัยนี้ ได้ถูกออกแบบให้บริการผ่านเว็บเซอร์วิส โดยมีหน้าที่หลัก 2 อย่าง คือ 1) ให้บริการข้อมูลเกี่ยวกับ ชื่อ คำอธิบาย และหมวดหมู่สถานที่ท่องเที่ยวทั้งหมดที่มีอยู่ในฐานข้อมูลกับผู้ร้องขอ และ 2) ให้บริการในการรับข้อมูลของชื่อ และคำอธิบายสถานที่ท่องเที่ยวจากผู้ใช้งานทั่วไป เพื่อนำมาจัดหมวดหมู่ผ่านโปรแกรม Weka และเก็บข้อมูลทั้งหมดลงในฐานข้อมูลของระบบ ดังแสดงในภาพที่ 3

สำหรับการออกแบบและพัฒนาแบบจำลองการเรียนรู้ของเครื่อง สามารถแบ่งขั้นตอนสำคัญๆ ออกเป็น 2 ขั้นตอน คือ การสกัดคุณลักษณะเด่นของหมวดหมู่สถานที่ท่องเที่ยว และการเลือกแบบจำลองการเรียนรู้ของเครื่องในการจัดหมวดหมู่สถานที่ท่องเที่ยว



ภาพที่ 3: โครงสร้างโดยรวมของระบบจัดหมวดหมู่สถานที่ท่องเที่ยว

3.1 การสกัดคุณลักษณะเด่นของหมวดหมู่สถานที่ท่องเที่ยว

ระบบจัดหมวดหมู่สถานที่ท่องเที่ยวต้นแบบนี้ ได้วางขอบเขตของสถานที่ท่องเที่ยวเฉพาะในจังหวัดภูเก็ต และมีความสามารถในการจำแนกหมวดหมู่สถานที่ท่องเที่ยวได้ทั้งหมด 4 หมวดหมู่ คือ วัด ทะเล ภูเขา และที่พัก

ดังนั้นขั้นตอนนี้ จึงเริ่มจากรวบรวมชื่อและคำอธิบายสถานที่ท่องเที่ยวในจังหวัดภูเก็ตจากอินเทอร์เน็ต หมวดหมู่ละ 10 สถานที่ แล้วนำมาผ่านโปรแกรมตัดคำภาษาไทย Lexto เพื่อนับจำนวนคำที่ถูกใช้ในคำอธิบายสถานที่ท่องเที่ยวมากที่สุด 15 อันดับแรกของแต่ละหมวดหมู่ ดังแสดงในตารางที่ 1

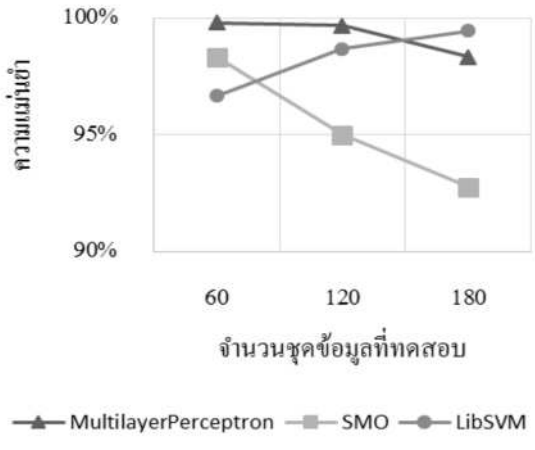
ตารางที่ 1: คำที่ถูกใช้บ่อยที่สุด 15 คำแรกในคำอธิบายสถานที่ท่องเที่ยว

อันดับ	วัด	ทะเล	ภูเขา	ที่พัก
1	วัด	หาด	ยอดเขา	รีสอร์ท
2	เทพ	ชายหาด	ป่า	สปา
3	พระ	กะตะ	ชม	ห้อง
4	เจ้าอาวาส	ทะเล	วิว	โรงแรม
5	หลวงพ่อ	หาดทราย	จุดชมวิว	บีช
6	บ้าน	เกาะ	ภูเขา	สิ่งอำนวยความสะดวก
7	ผุด	ภูเก็ต	รัง	ห้องพัก
8	ศรีลังกา	เต่า	ทัศนียภาพ	บริการ
9	ฉลอง	กิจกรรม	หอ	สระว่ายน้ำ
10	ชาวบ้าน	ป่าตอง	กรมป่าไม้	ให้บริการ
11	ท้าว	แหลม	การอนุรักษ์	ไทย
12	นิมิต	คลื่น	ช้าง	เดอะ
13	บรม	น้ำ	ตีนเขา	วิลล่า
14	พระทอง	ทราย	ทิวทัศน์	แกรนด์
15	พ่อ	ทางน้ำ	ป่าสงวน	ความสะดวก

3.2 การเลือกแบบจำลองการเรียนรู้ของเครื่อง

การเลือกแบบจำลองในการจัดหมวดหมู่ ถือเป็นหัวใจสำคัญของการเรียนรู้ของเครื่อง ผู้วิจัยจึงใช้คำสำคัญที่ถูกใช้บ่อยที่สุด 10 คำแรกของแต่ละหมวดหมู่ มาสร้างเป็นข้อมูลคุณลักษณะเด่นของสถานที่ท่องเที่ยวแต่ละสถานที่ ในการทดลองนี้ กลุ่มข้อมูลตัวอย่างของสถานที่ท่องเที่ยวที่ใช้ฝึกสอนและทดสอบความถูกต้องเป็นชุดข้อมูลเดียวกัน โดยได้ใช้ขนาดของชุดข้อมูล 3 ขนาด คือ 60, 120 และ 180 ข้อมูล และทดสอบกับแบบจำลองในการจัดหมวดหมู่ 3 แบบจำลอง คือ MultilayerPerceptron ที่เป็นแบบจำลองเครือข่ายประสาทเทียม และ ซัพพอร์ตเวคเตอร์แมชชีนที่ใช้วิธี SMO (Sequential minimal optimization) โดยทั้ง 2 แบบจำลองข้างต้นมีมาพร้อมกับ Weka และแบบจำลองที่ 3 คือ ซัพพอร์ตเวคเตอร์แมชชีนที่ติดตั้งเพิ่มเติมชื่อ LibSVM [6] ซึ่งใช้เคอร์เนลโดยปริยายคือฟังก์ชัน radial basisได้ผลการทดลองดังภาพที่ 4

จากผลการทดลองจะเห็นได้ว่า แบบจำลองที่ใช้เทคนิคของ MultilayerPerceptron ให้ความแม่นยำถึง 100% สำหรับการใช้ข้อมูลฝึกสอนและทดสอบ ชุดเดียวกันทั้ง 60 ข้อมูล แต่ความแม่นยำเริ่มลดลงเมื่อมีจำนวนข้อมูลที่ฝึกสอนมากขึ้นเรื่อยๆ โดยมีความแม่นยำเหลือ 98.33% เพื่อมีชุดข้อมูลที่ใช้ฝึกสอนจำนวน 180 ข้อมูล สำหรับแบบจำลองที่ใช้เทคนิคซัพพอร์ตเวคเตอร์แมชชีน SMO ที่มีมากับ Weka นั้นให้ผลลัพธ์ที่ไม่น่าพึงพอใจ ซึ่งตรงกันข้ามกับแบบจำลองที่ใช้เทคนิคซัพพอร์ตเวคเตอร์แมชชีนของ LibSVM ที่ให้ความแม่นยำมากขึ้นเรื่อยๆ ตามจำนวนข้อมูลที่ใช้ในการฝึกสอน โดยให้ความแม่นยำถึง 99.44% สำหรับชุดข้อมูลที่ฝึกสอนจำนวน 180 ข้อมูล



ภาพที่ 4: ความแม่นยำของแบบจำลองแบบต่างๆ

4. ผลการดำเนินงาน

จากผลการทดสอบแบบจำลองประเภทต่างๆ จะเห็นได้ว่าแบบจำลองที่ใช้เทคนิคซัพพอร์ตเวคเตอร์แมชชีนของ LibSVM ให้ความแม่นยำมากที่สุด และเพื่อให้ระบบสามารถทำงานได้อย่างรวดเร็ว จึงแบ่งการทดลองเพิ่มเติมอีก 2 การทดลอง คือ การทดลองเพื่อวิเคราะห์จำนวนคำสำคัญที่ใช้เป็นคุณลักษณะเด่นในการจัดหมวดหมู่สถานที่ท่องเที่ยว ว่าถ้าลดจำนวนคำสำคัญออกไปจะมีผลกระทบต่อความแม่นยำและเวลาที่ใช้ในการจัดหมวดหมู่อย่างไร และการทดลองเพื่อหาความแม่นยำแบบละเอียดสำหรับหมวดหมู่แต่ละประเภท

4.1 ผลกระทบของแบบจำลองกับจำนวนคำสำคัญ

ในการทดลองนี้ได้นำคำที่ถูกพบบ่อยที่สุดของสถานที่ในแต่ละหมวดหมู่มาใช้เป็นลักษณะเด่นเพื่อสร้างแบบจำลอง โดยทดลองสร้างแบบจำลองทั้งหมด 3 แบบจำลอง จากคำที่พบบ่อยที่สุดของแต่ละหมวดหมู่จำนวน 5, 10 และ 15 คำแรก และในแต่ละแบบจำลองจะใช้ชุดข้อมูลฝึกจำนวน 5 ขนาด คือ 40, 60, 80, 100 และ 120 ข้อมูล เพื่อวัดเวลาที่ใช้ในการสร้างแบบจำลอง และความแม่นยำที่ใช้ในการจัดหมวดหมู่กับชุดข้อมูลทดสอบขนาด 40 ชุดข้อมูลที่ไม่ได้ถูกใช้ในขั้นตอนการฝึก ซึ่งได้ผลลัพธ์ดังแสดงในตารางที่ 2

จากการทดลองพบว่าการใช้คำสำคัญของแต่ละหมวดหมู่สถานที่ท่องเที่ยว 10 และ 15 คำ สำหรับจำนวนข้อมูลที่ใช้ฝึก 120 ข้อมูลนั้น ให้ความแม่นยำในการจัดหมวดหมู่ถึง 95% ซึ่งดีกว่าการใช้คำสำคัญเพียง 5 คำ ที่ให้ความแม่นยำเพียง 92.5% อย่างไรก็ตามจะเห็นว่าผลลัพธ์ของการใช้คำสำคัญ 10 คำนั้นให้ผลลัพธ์เทียบเท่ากับการใช้คำสำคัญ 15 คำ และยังใช้เวลาในการสร้างแบบจำลองที่น้อยกว่า

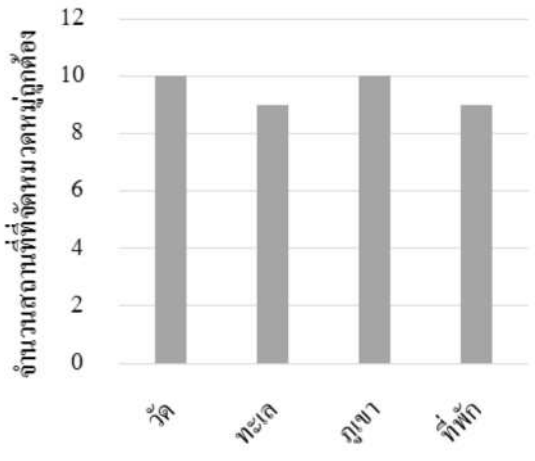
ตารางที่ 2: ความแม่นยำในการจัดหมวดหมู่ และเวลาในการสร้างแบบจำลองตามจำนวนคำสำคัญ

จำนวนข้อมูลที่ใช้ฝึก	5 คำสำคัญ		10 คำสำคัญ		15 คำสำคัญ	
	ความแม่นยำ	เวลา (ms)	ความแม่นยำ	เวลา (ms)	ความแม่นยำ	เวลา (ms)
40	77.5%	258	75.00%	304	72.5%	328
60	85.0%	314	87.5%	346	87.5%	386
80	87.5%	330	90.0%	388	90.0%	455
100	90.0%	357	92.5%	431	92.5%	500
120	92.5%	375	95.0%	488	95.0%	559

4.2 ความแม่นยำของแบบจำลองในแต่ละหมวดหมู่

จากผลการทดลองในหัวข้อที่ 4.1 พบว่าเมื่อใช้คำสำคัญจำนวน 10 คำสำคัญเพื่อเป็นคุณลักษณะเด่นในการสร้างแบบจำลอง และใช้จำนวนข้อมูลในการฝึก 120 ข้อมูลนั้น นอกจากจะทำให้ความแม่นยำดีที่สุดแล้ว ยังใช้เวลาในการสร้างแบบจำลองรวดเร็วกว่าการใช้คำสำคัญจำนวน 15 คำ ดังนั้นการทดลองนี้จะเป็นการทดสอบเพื่อดูว่าเมื่อใช้แบบจำลองที่เกิดจากการใช้คำสำคัญ 10 คำเป็นตัวกำหนดคุณลักษณะเด่นของหมวดหมู่สถานที่ท่องเที่ยวและใช้ข้อมูลจำนวน 120 ข้อมูลในการฝึกแบบจำลองแล้วนั้น เมื่อทดสอบกับข้อมูลที่ไม่ได้ใช้ในการฝึกหมวดหมู่ละ 10 ข้อมูลแล้ว ความแม่นยำในแต่ละหมวดหมู่แสดงดังภาพที่ 5

โดยการผลการทดลองพบว่าจากข้อมูลสถานที่ท่องเที่ยว 10 สถานที่ในแต่ละหมวดหมู่นั้น หมวดหมู่สถานที่ท่องเที่ยวประเภทวัดและภูเขามีความแม่นยำถึง 100% ซึ่งคาดว่าคำสำคัญที่ใช้กำหนดคุณลักษณะเด่นของสถานที่ท่องเที่ยวทั้ง 2 ประเภทนี้ค่อนข้างเด่นชัด เช่น วัด สำหรับหมวดหมู่ประเภทวัด และ *ยอดเขา* สำหรับหมวดหมู่ประเภทภูเขา แต่สำหรับหมวดหมู่ประเภทที่พักและทะเลนั้นมีการจัดหมวดหมู่ที่ผิดพลาดอยู่ประเภทละ 1 สถานที่ สำหรับหมวดหมู่ประเภททะเลคาดว่ามาจากคำสำคัญที่ใช้ในการกำหนดคุณลักษณะเด่น มีคำว่า *กะตะ* เป็นส่วนประกอบซึ่งเป็นคำเฉพาะเกินไป และสำหรับหมวดหมู่ประเภทที่พัก โดยทั่วไปจะมีคำอธิบายที่กล่าวถึงความใกล้ของที่พักกับภูเขาหรือชายหาดมากจนเกินไป ทำให้ถูกจัดหมวดหมู่ผิดประเภท



ภาพที่ 5: ความแม่นยำในการจัดหมวดหมู่สถานที่ท่องเที่ยว

5. สรุป

งานวิจัยนี้ได้ออกแบบและพัฒนาระบบจัดหมวดหมู่สถานที่ท่องเที่ยวด้วยคำอธิบายแบบอัตโนมัติผ่านเว็บเซอร์วิส โดยมีการประยุกต์นำความสามารถของเครื่องมือแบ่งคำภาษาไทย Lexto และโปรแกรม Weka มาใช้งานในการเรียนรู้ของเครื่องเพื่อช่วยเหลือการจัดหมวดหมู่สถานที่ท่องเที่ยวในจังหวัดภูเก็ต โดยมีหมวดหมู่สถานที่ท่องเที่ยวทั้งหมด 4 ประเภทได้แก่ วัด ทะเล ภูเขา และที่พัก จากการทดลองพบว่าการใช้แบบจำลองการเรียนรู้ของเครื่องจักรด้วยเทคนิคซัพพอร์ทเวคเตอร์แมชชีนด้วย LibSVM ให้ความแม่นยำในการจัดหมวดหมู่มากที่สุด อีกทั้งการใช้คำสำคัญของแต่ละหมวดหมู่สถานที่เพียงแค่ 10 คำ และจำนวนข้อมูลคำอธิบายของแต่ละสถานที่ที่ใช้ฝึกเพียง 120 ข้อมูล เพียงพอที่ทำให้แบบจำลองให้ความแม่นยำในการจัดหมวดหมู่ถึง 95%

เอกสารอ้างอิง

[1] T.W. Miller and H.T Nguyen., “Data and Text Mining: A Business Applications Approach”, *New York: Pearson Prentice Hall Upper Saddle River*, 2005.

[2] F. Sebastiani, “Text categorization**”**, *Alessandro Zanasi (ed.), Text Mining and its Applications*, WIT Press, Southampton, UK, pp. 109-129, 2005.

[3] A. Popescu, G. Grefenstette, and P-A. Moëllic, “Mining Tourist Information from User-Supplied Collections”, *Proceedings of the 18th ACM conference on Information and knowledge management*. NY, USA, pp. 1713-1716, 2009.

[4] A. Popescu, G. Grefenstette and H. Bouamor, “Mining a Multilingual Geographical Gazetteer from the Web”, *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Washington DC., USA, vol 01, pp. 58-65, 2009

[5] คมคิด ชัชราภรณ์, ธรา อั่งสกุล และจิติมนต์ อั่งสกุล, “แบบจำลองการจัดหวมดหมู่สถานที่ท่องเที่ยวโดยเทคนิคการเรียนรู้ของเครื่อง” *วารสารเทคโนโลยีสุรนารี ฉบับสังคมศาสตร์* ปีที่ 6 ฉบับที่ 2 (ธันวาคม 2555) หน้า 35-58.

[6] C.-C. Chang and C.-J. Lin., “LIBSVM : a library for support vector machines.”, *ACM Transactions on Intelligent Systems and Technology*, vol 2, pp 27:1--27:27, 2011.