การพัฒนาและประยุกต์ใช้เทคนิคเว็บสแครปปิ้งและการประมวลผลภาษาธรรมชาติสำหรับการรวบรวมและวิเคราะห์ข้อมูลข่าวสารเกี่ยวกับน้ำท่วมในประเทศไทย

Development and Application of Web Scraping Techniques and Natural Language Processing for Collecting and Analyzing Flood-related News Data in Thailand

ชวลิต โควีระวงศ์[1*] รัตถชล อ่างมณี[2] ไพรินทร์ มีศรี[3]และดาวรถา วีระพันธ์[4]

Chavalit Koweerawong[1*], Rattachon Angmanee[2] Phairin Meesri[3] and Daoratha Weerapan[4]

[1]คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ ในพระบรมราชูปถัมภ์

1 หมู่ 20 ต.คลองหนึ่ง อ.คลองหลวง จ.ปทุมธานี 13180

[1]Faculty of Science and Technology, Valaya Alongkorn Rajabhat University under the Royal Patronage University.

1 M.20 ,Khlong Nueng, Khlong Luang, Pathum Thani 13180

*Corresponding author E-mail: chavalit@vru.ac.th

บทคัดย่อ

งานวิจัยนี้มุ่งพัฒนาระบบอัตโนมัติหรือกึ่งอัตโนมัติสำหรับรวบรวมและวิเคราะห์ข่าวสารเกี่ยวกับน้ำท่วมในประเทศไทยจากแหล่งข้อมูลออนไลน์ โดยมีวัตถุประสงค์เพื่อระบุพื้นที่เสี่ยงและแนวโน้มของสถานการณ์น้ำท่วม ผ่านการนำเสนอข้อมูลในรูปแบบแผนที่ความร้อน รวมถึงการประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติ เช่น Named Entity Recognition เพื่อสกัดข้อมูลสำคัญ เช่น ชื่อสถานที่และเวลา ผลการวิจัยแสดงให้เห็นถึงประสิทธิภาพของเทคนิคเว็บสแครปปิ้งด้วยเครื่องมือ BeautifulSoup และ Selenium ในการรวบรวมข่าวสารน้ำท่วมจำนวนมาก พร้อมด้วยการวิเคราะห์ข้อมูลด้วยโมเดล WangchanBERTa เพื่อสกัดข้อมูลที่สำคัญ ผลลัพธ์ที่ได้สามารถนำไปสรุปพื้นที่เสี่ยงและเหตุการณ์น้ำท่วมตลอดปี 2567 ได้อย่างรวดเร็วและแม่นยำ โดยใช้เวลารวมไม่เกิน 1 ชั่วโมง 33 นาที ทั้งยังแสดงผลสรุปด้วยแผนที่ความร้อนเพื่ออธิบายข้อมูลเชิงพื้นที่ อย่างไรก็ตาม การวิจัยพบข้อจำกัดด้านความน่าเชื่อถือของข้อมูลและความซับซ้อนของการประมวลผลในกรณีที่ข้อมูลมีขนาดใหญ่และหลากหลาย ข้อเสนอแนะในการพัฒนาระบบในอนาคต ได้แก่ การขยายแหล่งข้อมูลให้ครอบคลุมโซเชียลมีเดียและแพลตฟอร์มสาธารณะของหน่วยงานราชการ การนำเทคนิค AI เช่น การเรียนรู้เชิงลึกมาปรับปรุงความแม่นยำในการสกัดข้อมูล และการพัฒนาระบบแจ้งเตือนแบบเรียลไทม์ที่ผนวกกับระบบภูมิสารสนเทศ เพื่อให้การแสดงผลข้อมูลมีความชัดเจนและเข้าถึงง่าย

คำสำคัญ : ข่าวน้ำท่วม, การประมวลผลภาษาธรรมชาติ, การสกัดข้อมูล, เว็บสแครปปิ้ง, การวิเคราะห์ข้อมูล

Abstract

This research aims to develop an automated or semi-automated system for collecting and analyzing online flood-related news in Thailand to identify risk areas and flood trends. The information is presented in the form of heatmaps, utilizing Natural Language Processing (NLP) techniques such as Named Entity Recognition (NER) to extract key data, including locations and timestamps. The study demonstrates the efficiency of web scraping tools, including BeautifulSoup and Selenium, in gathering a large volume of flood-related news. Data was processed and analyzed using the WangchanBERTa model, enabling the extraction of critical information. The results provide an effective and timely summary of risk areas and flood events throughout 2024, with the entire process completed within 1 hour and 33 minutes. Heatmaps were used to visualize spatial data and convey information clearly. However, the research identified limitations regarding data reliability and processing complexity, particularly when handling large and diverse datasets. Future system enhancements should include expanding data sources to cover social media and public platforms from government agencies, incorporating advanced AI techniques such as Deep Learning to improve data extraction accuracy, and developing a real-time alert system integrated with Geographic Information Systems (GIS) to enhance clarity and accessibility of the displayed information

Keywords : Flood news,Natural Language Processing, Data extraction, Web scraping, Data analysis

บทนำ

ในปี 2566 จากการวิเคราะห์พื้นที่น้ำท่วมโดยใช้ข้อมูลจากภาพถ่ายดาวเทียม พบว่า ประเทศไทยมีพื้นที่ถูกน้ำท่วมประมาณ 4.74 ล้านไร่ กระจายตัวอยู่ในทุกภาคของประเทศ โดยมี 52 จังหวัด 419 อำเภอ 2,621 ตำบล ที่เกิดน้ำท่วมในพื้นที่ (สำนักงานทรัพยากรน้ำแห่งชาติ, 2566) น้ำท่วมเป็นหนึ่งในภัยพิบัติทางธรรมชาติที่เกิดขึ้นบ่อยครั้งในประเทศไทย โดยเฉพาะในช่วงฤดูฝน พฤษภาคม -ตุลาคม ของทุกปี ซึ่งมีปริมาณน้ำฝนสะสมอย่างต่อเนื่อง น้ำท่วมก่อให้เกิดผลกระทบอย่างรุนแรงทั้งในด้านเศรษฐกิจ สังคม และสิ่งแวดล้อม นอกจากนี้ ยังส่งผลต่อความเชื่อมั่นในการพัฒนาโครงสร้างพื้นฐานและการวางแผนเมืองในระยะยาว ด้วยเหตุนี้การจัดการข้อมูลข่าวสารที่เกี่ยวข้องกับน้ำท่วม การรายงานสถานการณ์ การระบุพื้นที่เสี่ยง และข้อมูลเชิงลึกจากข่าวที่เผยแพร่ผ่านสื่อออนไลน์ จึงเป็นปัจจัยสำคัญในการลดผลกระทบและเสริมสร้างระบบเตือนภัยหรือเฝ้าระวังที่มีประสิทธิภาพ (Aribowo & Wisudarma, 2018) ปัจจุบันแหล่งข้อมูลข่าวสารออนไลน์ที่ใช้ในการรายงานข้อมูลเกี่ยวกับน้ำท่วมแบ่งดังนี้ 1.เว็บไซต์ข่าวไทย เช่น ไทยรัฐ เดลินิวส์ และมติชน ซึ่งเป็นแหล่งข้อมูลที่มีความน่าเชื่อถือและเผยแพร่ข่าวสารเกี่ยวกับน้ำท่วมอย่างสม่ำเสมอ 2. โซเชียลมีเดีย เช่น Facebook และ X

(Twitter) (Kumar & Sebastian, 2019) ถูกนำมาใช้เป็นแหล่งข้อมูลสำหรับรายงานสถานการณ์แบบเรียลไทม์จากผู้ใช้ในพื้นที่ สำหรับข้อมูลที่กล่าวมาทั้งสองแหล่งนี้เป็นข้อมูลที่ไม่ได้มาจากทางราชการและอาจมีความสำคัญในด้านรายงานนอกพื้นที่เครื่องตรวจวัดอัตโนมัติ และ 3. แพลตฟอร์มข้อมูลสาธารณะ เช่น เว็บไซต์ของกรมอุตุนิยมวิทยาและหน่วยงานบริหารจัดการน้ำที่ให้ข้อมูลอย่างเป็นทางการเกี่ยวกับสถานการณ์น้ำท่วม

แม้ว่าข่าวสารเกี่ยวกับน้ำท่วมจะเผยแพร่ผ่านช่องทางออนไลน์อย่างกว้างขวาง ทั้งจากเว็บไซต์ข่าว โซเชียลมีเดีย และแพลตฟอร์มข้อมูลสาธารณะ แต่การรวบรวมข้อมูลเหล่านี้เพื่อให้สามารถใช้งานได้จริงกลับเป็นเรื่องท้าทาย เนื่องจากปริมาณข้อมูลที่มหาศาลซึ่งถูกเผยแพร่จากหลายแหล่งในรูปแบบต่าง ๆ ทำให้การคัดกรองและจัดระเบียบข้อมูลต้องใช้เวลาและทรัพยากรมาก นอกจากนี้คุณภาพและความน่าเชื่อถือของข้อมูล จากบางแหล่งข่าวยังเป็นปัญหา เนื่องจากบางข่าวสารอาจขาดความถูกต้องหรือมีข้อมูลคลาดเคลื่อน ส่งผลต่อความแม่นยำของการวิเคราะห์ อีกทั้งข่าวส่วนใหญ่ถูกนำเสนอในรูปแบบข้อความที่ไม่เป็นโครงสร้าง (Manning & Schütze, 2021) ซึ่งจำเป็นต้องใช้เทคนิคการประมวลผลภาษาธรรมชาติ ในการสกัดข้อมูลสำคัญเพื่อการใช้งานที่มีประสิทธิภาพ โดยนำข้อมูลที่มากมายนำมาสรุปในเวลาอันรวดเร็ว

เครื่องมือสกัดข้อมูล Scrapy, Beautiful Soup, และ Selenium เป็นเครื่องมือสำคัญที่ช่วยในการสกัดข้อมูลด้วย Python (Mitchell, 2018) โดยแต่ละเครื่องมือมีจุดเด่นและเหมาะสมกับงานที่แตกต่างกันไป Scrapy เหมาะสำหรับการรวบรวมข้อมูลในปริมาณมากจากหลายเว็บไซต์แบบอัตโนมัติ เนื่องจากมีโครงสร้างที่รองรับการจัดการและเก็บข้อมูลเป็นระบบ พร้อมด้วยความเร็วในการประมวลผล เหมาะสำหรับการดึงข้อมูลจากเว็บไซต์ที่มีโครงสร้าง HTML (Scrapy, 2021) ในทางกลับกัน Selenium ถูกออกแบบมาสำหรับเว็บไซต์ที่ใช้ JavaScript หรือมีเนื้อหาแบบไดนามิกที่ต้องการการจำลองพฤติกรรมผู้ใช้งาน ดังนั้น การเลือกใช้เครื่องมือเหล่านี้ควรพิจารณาจากลักษณะของข้อมูลและเป้าหมายของโครงการเทคนิค Named Entity Recognition (NER) เป็นหนึ่งในกระบวนการสำคัญของการประมวลผลภาษาธรรมชาติที่ช่วยในการสกัดข้อมูลสำคัญจากข้อความโดยระบุและจัดประเภทชื่อเฉพาะ เช่น ชื่อบุคคล สถานที่ องค์กร วันที่ หรือข้อมูลสำคัญอื่น ๆ NER เป็นเทคนิคที่มีความแม่นยำสูงเมื่อใช้โมเดลภาษาที่ทันสมัย เช่น BERT หรือ WangchanBERTa สำหรับภาษาไทย ซึ่งช่วยเพิ่มประสิทธิภาพในการระบุเอนทิตีจากข้อความขนาดใหญ่

งานวิจัยนี้มีวัตถุประสงค์หลักในการพัฒนาและประยุกต์ใช้เทคนิค เว็บสแครปปิ้ง (Web Scraping) สำหรับการรวบรวม วิเคราะห์และสรุปข่าวน้ำท่วมในประเทศไทย โดยมีเป้าหมายสำคัญดังนี้

1) เพื่อพัฒนาระบบบอัตโนมัติหรือกึ่งอัตโนมัติสำหรับการรวบรวมข่าวสารเกี่ยวกับน้ำท่วมจากแหล่งข้อมูลออนไลน์ต่างๆ

2) เพื่อวิเคราะห์ข้อมูลข่าวที่รวบรวมได้และสรุป เพื่อระบุพื้นที่เสี่ยงหรือแนวโน้มของสถานการณ์น้ำท่วม โดยนำเสนอในรูปแบบแผนที่ความร้อน

3) เพื่อประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติ (NLP) เช่น Named Entity Recognition (NER) เพื่อสกัดข้อมูลสำคัญตามชื่อสถานที่และเวลา

วิธีการดำเนินการวิจัย

ในการออกแบบและพัฒนาวิธีสกัดข้อมูลจากเว็บไซต์สำหรับข่าวน้ำท่วมในประเทศไทย มุ่งเน้นการออกแบบและพัฒนาขั้นตอนวิธีการสกัดข่าวแบบอัตโนมัติ คาดการณ์พื้นที่เสี่ยงภัยและสถานการณ์น้ำท่วม เพื่อความปลอดภัยของประชาชนในพื้นที่ งานวิจัยนี้มุ่งพัฒนาระบบบอัตโนมัติหรือกึ่งอัตโนมัติสำหรับสำหรับการรวบรวมข่าวสารเกี่ยวกับน้ำท่วมจากแหล่งข้อมูลออนไลน์ต่าง ๆ และวิเคราะห์ข้อมูลข่าวที่รวบรวมได้และสรุปผล โดยมีขั้นตอนการดำเนินงาน ดังนี้

1. การกำหนดเป้าหมายและรูปแบบข้อมูล โดยผู้วิจัยกำหนดประเภทข้อมูลที่ต้องการวิเคราะห์ ได้แก่ ข่าวเกี่ยวกับน้ำท่วม การแจ้งเตือนภัย หรือการโพสต์จากโซเชียลมีเดีย กำหนดรูปแบบของข้อมูล เช่น ข้อความ ประเภทของรูปภาพ วันที่ และแหล่งที่มาของข้อมูล และกำหนดความถี่ในการเก็บรวบรวมข้อมูลที่ต้องมาใช้ในการวิเคราะห์ เช่น รายวัน รายชั่วโมง

2. การเลือกแหล่งข้อมูล เกี่ยวกับน้ำท่วมช่องทางออนไลน์ที่น่าเชื่อถืออย่างเว็บไซต์ไทยรัฐออนไลน์มาทำการศึกษา ซึ่งเป็นแหล่งข้อมูลที่เผยแพร่ข่าวสารเกี่ยวกับน้ำท่วมอย่างสม่ำเสมอ โดยมีเป้าหมายหลักอยู่ที่แท็กของเว็บไซต์ “น้ำท่วม 2567”

3. การใช้โปรแกรมในการสกัดข้อมูลจากแหล่งข้อมูล โปรแกรมเว็บสแครปปิ้งได้รับการออกแบบมาเพื่อรวบรวมข้อมูลข่าวน้ำท่วมจากแหล่งข้อมูลออนไลน์ต่าง ๆ โดยใช้เครื่องมือที่เหมาะสมกับแต่ละแหล่ง BeautifulSoup ใช้สำหรับการดึงข้อมูล HTML จากเว็บไซต์ที่มีโครงสร้างคงที่ และ Selenium ใช้สำหรับเว็บไซต์ที่แสดงเนื้อหาผ่าน JavaScript และเพื่อป้องกันการละเมิดสิทธิ์การทำเว็บสแครปปิ้งได้ปฏิบัติตามการตั้งค่าในไฟล์ robots.txt ของเว็บไซต์แหล่งข่าวอย่างถูกต้องตามข้อกำหนดการเข้าถึงข้อมูลของเว็บไซต์

4. การประมวลผลข้อมูลข่าว นำลิงก์หรือ URL ของข่าวทั้งหมดจะถูกสกัดออกมาจากเว็บไซต์ด้วยโปรแกรมสกัดข้อมูล ด้วยไต่ไปตามเว็บ (Web Crawling) โดยสกัดจากแท็กของเว็บไซต์ “น้ำท่วม 2567” ของเว็บไซต์แหล่งข้อมูล และต้องระมัดระวังไม่ให้ลิงค์ซ้ำซ้อนกันเองด้วยการลบข้อมูลที่ซ้ำออก จากนั้นใช้โปรแกรมสกัดข้อมูล HTML จากลิงค์ที่ได้มาและสกัดเนื้อหาข่าวที่อยู่ภายใต้ HTML อีกที ถึงตอนนี้จะได้ข้อมูลข่าวที่รวบรวมมาผ่านกระบวนการประมวลผลเพื่อให้ได้ข้อมูลดิบที่เป็นข่าวสารพร้อมสำหรับการวิเคราะห์ สุดท้ายเทคนิค NER ถูกนำมาใช้ในการสกัดข้อมูลสำคัญจากข่าวน้ำท่วม เช่น เวลารายงานข่าว และ การระบุชื่อสถานที่ที่เกิดน้ำท่วม ได้แก่ จังหวัด อำเภอ หรือพื้นที่เฉพาะ เครื่องมือ NLP ที่ใช้ เช่น WangchanBERTa ถูกเลือกมาเพื่อเพิ่มความแม่นยำในการประมวลผลข้อมูลและการระบุข้อมูลที่เกี่ยวข้อง

5. การสรุปข้อมูลและการตีความ ข้อมูลสถานที่และเวลาที่ผ่านการประมวลผลจะถูกวิเคราะห์ทั้งในเชิงปริมาณและเชิงคุณภาพ การวิเคราะห์เชิงปริมาณเน้นการนับจำนวนข่าวในแต่ละพื้นที่และ แต่ละช่วงเวลา ส่วนการวิเคราะห์เชิงคุณภาพมุ่งเน้นการตีความบริบทของเหตุการณ์และผลกระทบที่เกิดขึ้น ผลการวิเคราะห์นี้จะแสดงในรูปแบบกราฟิก เช่น แผนภูมิหรือแผนที่ความร้อน (Heatmap) เพื่อช่วยให้การนำข้อมูลไปใช้สำหรับการวางแผนและการบริหารจัดการภัยพิบัติเป็นไปอย่างมีประสิทธิภาพ

ผลการวิจัย

การสกัดลิงก์จากข่าว

หน้าข่าวบนเว็บไซต์ไทยรัฐออนไลน์ที่เป็นแท็ก “น้ำท่วม 2567” มีลักษณะเป็นหน้าดัชนีที่รวบรวมเหตุการณ์ข่าวทั้งหมดที่เกี่ยวข้องกับแท็กดังกล่าว การสกัดลิงก์ (URL) เพื่อรวบรวมข้อมูลจึงเริ่มต้นจากหน้านี้ อย่างไรก็ตาม หน้าดังกล่าวแสดงข่าวเริ่มต้นเพียง 10 ข่าว และต้องกดปุ่ม “โหลดเพิ่มเติม” เพื่อดูข่าวเพิ่มเติม ซึ่งการโหลดข้อมูลเพิ่มเติมนี้ถูกดำเนินการผ่าน JavaScript ด้วยเหตุนี้ การใช้ Selenium จึงเหมาะสมกว่าการใช้ BeautifulSoup ในการจัดการเว็บไซต์ที่มีโครงสร้างแบบไดนามิกเช่นนี้ ดังนั้นการสกัดลิงก์ข่าวจึงเลือกใช้ Selenium แทน

การสกัดข้อมูลข่าวจากลิงก์

ข้อมูลข่าวจากแหล่งข้อมูล ข่าวน้ำท่วมออนไลน์ 555 รายการ ที่สกัดมาได้โดยใช้ BeautifulSoup และ Request ดึงข้อมูลจากลิงก์ที่ได้มาทำให้ได้ข้อความที่มาจากเนื้อข่าวโดยปราศจากโค้ด HTML และเนื้อหาเหล่านั้นจะถูกนำมาผ่านกระบวนการประมวลผลโดยใช้เทคนิค NER ด้วยโมเดล WangchanBERTa เพื่อสกัดสถานที่และเวลาที่ปรากฏในรายงานข่าว อย่างไรก็ตาม บางข่าวที่มีความยาวมากจำเป็นต้องแบ่งเป็น 2-5 ส่ว เนื่องจากข้อจำกัดด้านความยาวข้อความของ WangchanBERTa ซึ่งสามารถประมวลผลได้สูงสุดที่ 512 tokens ต่อข้อความตามการตั้งค่ามาตรฐานของ BERT โดยจำนวน characters ที่แท้จริงจะขึ้นอยู่กับกระบวนการ tokenization ซึ่งในกรณีของภาษาไทย การตัดคำที่ซับซ้อนอาจทำให้ข้อความที่มีจำนวนตัวอักษรมากแปลงเป็นจำนวน token ที่สูงกว่าภาษามาตรฐานเช่นภาษาอังกฤษ แต่ละข่าวที่ผ่านการประมวลผลจะใช้เวลาประมาณ 1-3 วินาที ข้อมูลที่สกัดมาได้ก็จะบันทึกลงฐานข้อมูลชั่วคราว

สรุปข้อมูลด้วย DataFrame และ PivotTable

ข้อมูลเกี่ยวกับสถานที่และเวลาที่ได้จะถูกนำมาวิเคราะห์ต่อโดย*ใช้* Pandas DataFrame และ PivotTable หากข่าวบางรายการระบุสถานที่มากกว่าหนึ่งแห่ง จะต้องมีการเพิ่ม record ใน DataFrame เพื่อนำมาวิเคราะห์ ซึ่งจากข้อมูลที่ได้จะทำให้สามารถสรุปผลและจัดแสดงในตารางที่ 1 และจากตารางที่ 1 และ 2 ผลการวิเคราะห์พบว่าจังหวัดที่ได้รับผลกระทบจากน้ำท่วมมากที่สุดคือเชียงรายและเชียงใหม่ ซึ่งอยู่ทางภาคเหนือของประเทศไทย โดยทั้งสองจังหวัดมีการกล่าวถึงรวมกันถึง *29*% ของข่าวทั้งหมด รายงานข่าวส่วนใหญ่ (*79*%) เกิดขึ้นในไตรมาสที่ *3* ของปี *2567* (กรกฎาคมถึงกันยายน) และอีก *20*% เกิดขึ้นในไตรมาสที่ *4* (พฤศจิกายนถึงธันวาคม) ในขณะที่ไตรมาสที่ *2* (เมษายนถึงมิถุนายน) ไม่มีรายงานข่าวน้ำท่วมเลย ส่วนข่าวในไตรมาสที่ *1* เป็นข่าวสรุปเหตุการณ์น้ำท่วมในปีก่อนหน้า ภาคที่ได้รับผลกระทบจากน้ำท่วมมากที่สุดคือภาคเหนือและภาคกลาง คิดเป็น *41*% และ *33*% ของข่าวทั้งหมดตามลำดับ ในไตรมาสที่ *3* มีรายงานน้ำท่วมจากภาคเหนือ กลาง ตะวันออก และตะวันตก แต่ลดลงมากกว่า *50*% ในไตรมาสที่ *4* ขณะที่ภาคใต้มีรายงานข่าวน้ำท่วมในไตรมาสที่ *3* และ *4* อย่างสมดุลกัน ส่วนภาคตะวันออกเฉียงเหนือมีรายงานน้ำท่วมเฉพาะในไตรมาสที่ *3* เท่านั้น

ตารางที่ 1 สรุปจังหวัดที่มีจำนวนเหตุการณ์น้ำท่วมที่รายงานในแหล่งออนไลน์ปี 2567 มากที่สุด 10 อันดับแรก แบ่งตามภูมิภาคและรายไตรมาส

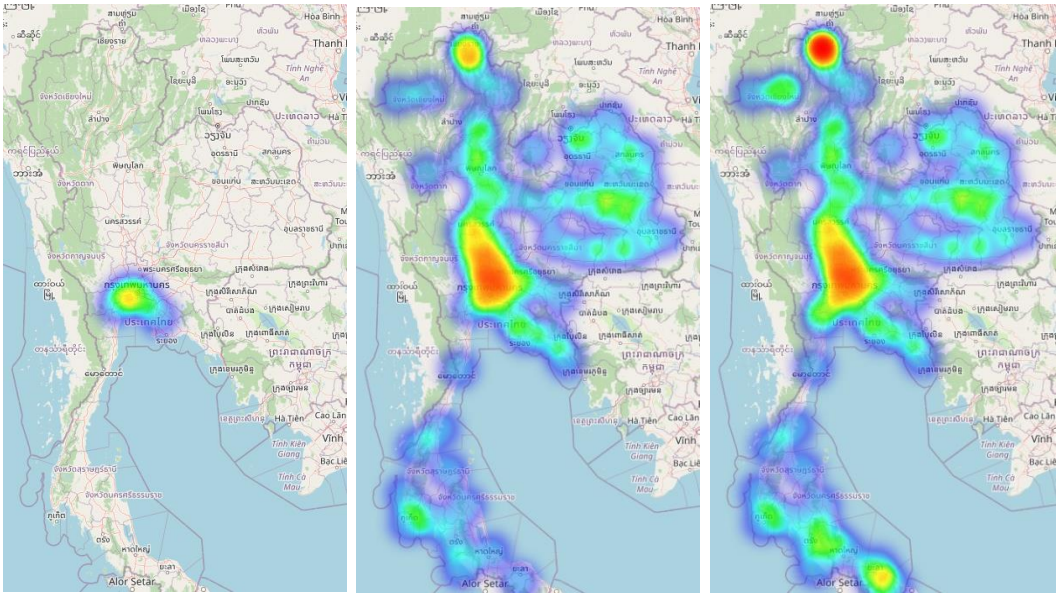
จังหวัด	ภูมิภาค	ไตรมาส 1	ไตรมาส 3	ไตรมาส 4	รวม
เชียงราย	ภาคเหนือ	0	120	25	145
เชียงใหม่	ภาคเหนือ	0	58	44	102
สุโขทัย	ภาคกลาง	0	34	5	39
ชัยนาท	ภาคกลาง	0	30	7	37
พระนครศรีอยุธยา	ภาคกลาง	0	24	9	33
แพร่	ภาคเหนือ	0	30	0	30
หนองคาย	ภาคตะวันออกเฉียงเหนือ	0	29	0	29
น่าน	ภาคเหนือ	0	29	0	29
พะเยา	ภาคเหนือ	0	25	1	26
อ่างทอง	ภาคกลาง	0	20	4	24
นนทบุรี	ภาคกลาง	1	12	10	23
ตาก	ภาคตะวันตก	0	16	2	18
นครสวรรค์	ภาคกลาง	0	11	5	16

ตารางที่ 2 สรุปจำนวนเหตุการณ์น้ำท่วมที่รายงานในแหล่งออนไลน์ปี 2567 แบ่งตามภูมิภาคและรายไตรมาส

ภูมิภาค	ไตรมาส 1	ไตรมาส 3	ไตรมาส 4	รวม
ภาคเหนือ	0	277	71	348
ภาคกลาง	7	207	61	275
ภาคตะวันออกเฉียงเหนือ	0	112	0	112
ภาคใต้	0	30	31	61
ภาคตะวันตก	0	17	5	22
ภาคตะวันออก	1	19	1	21
รวมทั้งหมด	**8**	**662**	**169**	**839**

สรุปและอธิบายข้อมูลด้วยแผนที่ความร้อน

การสร้างแผนที่ความร้อน (Heatmap) ด้วย Python โดยใช้ Folium เริ่มจากการเตรียมข้อมูลในรูปแบบ List ที่ประกอบด้วยพิกัดละติจูด ลองจิจูด และค่าความร้อนที่สะท้อนจำนวนครั้งของรายงานน้ำท่วม เพื่อแสดงข้อมูลเชิงพื้นที่ เช่น ความหนาแน่นของเหตุการณ์หรือจุดที่น่าสนใจในพื้นที่ต่างๆ แผนที่ความร้อนที่สร้างขึ้นแสดงจำนวนรายงานน้ำท่วมในแต่ละไตรมาสของปี โดยพบว่าในไตรมาสที่ 1 มีรายงานเฉพาะในภาคกลาง ส่วนไตรมาสที่ 2 ไม่มีรายงานใดๆ ในขณะที่ไตรมาสที่ 3 และ 4 เหตุการณ์น้ำท่วมกลับมาแสดงความหนาแน่นอีกครั้ง โดยจุดที่น่าสนใจของเหตุการณ์น้ำท่วมอยู่ในภาคกลางตอนล่างและภาคเหนือตอนบน ขณะที่ภาคใต้มีการรายงานเหตุการณ์ตลอดในไตรมาสที่ 3 และ 4 ดังรูปที่ 1



รูปที่ 1 แผนภาพความภูมิร้อนจำนวนรายงานน้ำท่วม (ซ้าย) รายงานไตรมาส 1 และ 2 (กลาง) รายงานไตรมาส 3 (ขวา) รายงานไตรมาส 4

ความเร็วในการสกัดข้อมูล

กระบวนการสกัดข้อมูลด้วยเว็บสแครปปิ้งเริ่มจากการสกัดลิงก์จากข่าวซึ่งใช้เวลาเพียง 2 วินาที ต่อมาข้อมูลข่าวจากลิงก์แต่ละรายการถูกสกัดออกมา โดยเฉลี่ยใช้เวลาประมาณไม่เกิน 10 วินาทีต่อข่าว รวมเวลาทั้งหมด 5550 วินาที หรือประมาณ 92 นาที หลังจากนั้น ข้อมูลที่รวบรวมได้ถูกสรุปด้วย DataFrame และ PivotTable ซึ่งใช้เวลาเพียง 8 วินาที และสุดท้าย การสร้างแผนที่ความร้อนเพื่อแสดงข้อมูลเชิงพื้นที่ใช้เวลาเพียง 1 วินาที รวมเป็นเวลาไม่เกิน 1 ชั่วโมง 33 นาที โดยเวลาทั้งหมด 99.99% หมดไปกับการสกัดข้อมูลจากหน้าแหล่งข่าวและสกัดข้อมูลด้วยเทคนิค NER

จากแหล่งข้อมูลข่าวน้ำท่วมออนไลน์ 555 รายการ และด้วยเทคนิคเว็บสแครปปิ้งพบว่าใช้เวลาในการรวบรวมข้อมูลไม่นานเมื่อเทียบกับการรวบรวมข้อมูลด้วยตนเองและสามารถดึงข้อมูลจากหลายแหล่งพร้อมกันได้

อย่างมีประสิทธิภาพ นอกจากนี้ การใช้เทคนิค BeutifulSoup และ Selenium ยังช่วยเพิ่มความยืดหยุ่นในการเก็บข้อมูลจากเว็บไซต์ที่มีการแสดงผลด้วย JavaScript การวิเคราะห์ข้อมูลที่ได้จากเว็บสแครปปิ้งยังมีความละเอียดมากขึ้น (McKinney, 2022) เนื่องจากการประยุกต์ใช้เทคนิค NLP เช่น NER ช่วยในการสกัดข้อมูลสำคัญโดยอัตโนมัติ ผลการทดลองนี้แสดงให้เห็นว่าเทคนิคเว็บสแครปปิ้งเป็นเครื่องมือที่มีประสิทธิภาพสูงสำหรับการรวบรวมและวิเคราะห์ข่าวสารเกี่ยวกับน้ำท่วมในประเทศไทย และสามารถนำไปประยุกต์ใช้ในการจัดการภัยพิบัติได้อย่างเหมาะสม

วิจารณ์ผล

การสกัดลิงก์ข่าวและการสกัดข้อมูลจากข่าว

การสกัดลิงก์จากข่าวออนไลน์ในงานวิจัยนี้ยังมีข้อจำกัดในด้านความหลากหลายของแหล่งข้อมูล เช่น สำนักข่าวต่าง ๆ โซเชียลมีเดีย และแพลตฟอร์มข้อมูลสาธารณะของทางราชการ ซึ่งอาจทำให้ข้อมูลสำคัญบางส่วนถูกมองข้าม หากไม่มีความครอบคลุมของแหล่งข้อมูล จะทำให้การประมวลผลขาดมิติของความเร็วในการตอบสนองต่อเหตุการณ์จากโซเชียลมีเดีย หรือความถูกต้องจากข้อมูลที่มาจากแหล่งทางการ อย่างไรก็ตาม งานวิจัยนี้ได้นำเสนอเฟรมเวิร์กสำหรับการสกัดข้อมูลจากเว็บไซต์ออนไลน์ ซึ่งเน้นการเลือกวิธีการสกัดที่เหมาะสมกับโครงสร้างของเว็บไซต์เฉพาะ แม้ว่าวิธีการเดียวกันอาจไม่สามารถนำไปใช้กับทุกเว็บไซต์ได้โดยตรง การใช้ BeautifulSoup เป็นเครื่องมือหลักในการสกัดข้อมูลก็แสดงให้เห็นถึงข้อจำกัดบางประการ ทำให้ในบางกรณีต้องพึ่งพา Selenium เพื่อช่วยจัดการกับปัญหา เช่น การโหลดข้อมูลด้วย JavaScript ในขณะที่ Scrapy ซึ่งเป็นเฟรมเวิร์กเฉพาะด้านการสกัดข้อมูล อาจเป็นทางเลือกที่ดีกว่าสำหรับการจัดการข้อมูลจำนวนมากและการเขียนเงื่อนไขซับซ้อน เช่น การจัดการลิงก์ซ้ำหรือข้อผิดพลาดที่ไม่คาดคิด ทั้งนี้ การใช้ Scrapy อาจต้องแลกมากับการเรียนรู้และความเข้าใจในเฟรมเวิร์ก นอกจากนี้ยังมีข้อกังวลเรื่องข้อกำหนดทางกฎหมายและจริยธรรมในการดึงข้อมูลจากบางแหล่ง อย่างไรก็ดี เทคนิคเว็บสแครปปิ้งได้ช่วยเพิ่มประสิทธิภาพในการรวบรวมข้อมูลข่าวน้ำท่วมจากหลายแหล่งได้อย่างรวดเร็ว ช่วยลดภาระการทำงานของคน และรวบรวมข้อมูลปริมาณมากในเวลาสั้น ซึ่งนำไปสู่การวิเคราะห์เชิงลึก เช่น การระบุพื้นที่เสี่ยงน้ำท่วมและความถี่ของเหตุการณ์ ทั้งนี้ช่วยสนับสนุนการตัดสินใจในการจัดการภัยพิบัติได้ดียิ่งขึ้น

การใช้ NLP ในการสกัดข้อมูล

ข้อจำกัดและข้อผิดพลาดที่สำคัญในกระบวนการวิจัยนี้เกี่ยวข้องกับการใช้เทคนิค NER และเครื่องมือวิเคราะห์ข้อมูล เช่น ข้อจำกัดในการประมวลผลข้อความยาว ซึ่งต้องแบ่งข้อความออกเป็นส่วนย่อยเพื่อให้ใช้งานกับโมเดล WangchanBERTa ที่รองรับความยาวข้อความสูงสุด 512 tokens ต่อครั้ง การแบ่งข้อความนี้อาจทำให้บริบทของข้อมูลสูญหายหรือผลลัพธ์ขาดความต่อเนื่อง นอกจากนี้ ปัญหาด้านประสิทธิภาพ เช่น เวลาที่ใช้ในการประมวลผลข้อมูลจำนวนมาก และข้อจำกัดทางเทคนิคของอุปกรณ์ อาจส่งผลต่อความแม่นยำและความครบถ้วนของผลลัพธ์ แม้ WangchanBERTa จะเป็นโมเดลที่มีศักยภาพในการสกัดข้อมูลสำคัญ เช่น "LOCATION," "DATE," "PERSON," และ "ORGANIZATION" แต่ประสิทธิภาพและความแม่นยำของข้อมูลที่สกัด

ได้ยังคงแตกต่างกันไป เช่น ข้อมูลประเภท LOCATION และ DATE มีความแม่นยำสูงและสามารถจัดการได้ด้วยการทำความสะอาดข้อมูลอัตโนมัติ ในขณะที่ข้อมูลประเภท PERSON และ ORGANIZATION ยังคงต้องใช้แรงงานคนในการคัดแยก เนื่องจากข้อมูลมีลักษณะกระจัดกระจาย นอกจากนี้ ปัญหาข้อมูลที่ไม่สะอาด เช่น การที่ชื่อจังหวัดมีคำว่า "จ." หรือ "จังหวัด" ติดมาด้วย จำเป็นต้องระมัดระวังก่อนนำไปใช้งาน อย่างไรก็ตาม การประยุกต์ใช้เทคนิค NLP ยังคงแสดงให้เห็นถึงศักยภาพในการสรุปเนื้อหาได้อย่างรวดเร็วและมีประสิทธิภาพ ข้อมูลที่ได้จากการวิจัยสามารถนำไปประยุกต์ใช้ในหลายมิติ เช่น การวางแผนและการเตือนภัยน้ำท่วมผ่านการสร้างแผนที่ความร้อน (Heatmap) ซึ่งช่วยสนับสนุนการจัดการภัยพิบัติได้อย่างมีประสิทธิภาพ

การสรุปเนื้อหาจากข้อมูลที่ได้มา

จากการวิเคราะห์ภาพรวมทั้งประเทศในปี 2567 พบว่าภาคเหนือมีรายงานการเกิดน้ำท่วมมากที่สุด รองลงมาคือภาคกลาง โดยเหตุการณ์น้ำท่วมใหญ่ที่อำเภอแม่สาย จังหวัดเชียงราย ซึ่งเริ่มตั้งแต่วันที่ 9 กันยายน 2567 และยุติลงในเดือนพฤศจิกายน สาเหตุมาจากอิทธิพลของพายุไต้ฝุ่นยางิที่เคลื่อนตัวจากทะเลจีนใต้ด้วยความรุนแรงระดับซุปเปอร์ไต้ฝุ่น ก่อนลดระดับเป็นดีเปรสชันเมื่อขึ้นฝั่งเวียดนาม ส่งผลให้ภาคเหนือตอนบนมีฝนตกหนัก เกิดดินโคลนถล่ม น้ำท่วมฉับพลัน และสร้างความเสียหายเป็นวงกว้าง ส่วนในภาคกลาง สถานการณ์การระบายน้ำยังเป็นไปตามปกติ แม้ว่าจะมีน้ำระบายลงมาจากภาคเหนือตอนบน แต่ไม่มีรายงานความผิดปกติ สำหรับภาคตะวันออกเฉียงเหนือและภาคตะวันออกมีฝนตกหนักในไตรมาส 3 ขณะที่ภาคใต้และภาคตะวันตกมีฝนกระจายในไตรมาส 3 และ 4 โดยในภาคใต้ ไตรมาส 3 ได้รับอิทธิพลจากลมมรสุมตะวันตกเฉียงใต้ และในไตรมาส 4 มีผลจากดีเปรสชันที่พัดมาจากประเทศจีน ซึ่งก่อให้เกิดฝนตกในพื้นที่ การวิเคราะห์ข้อมูลจากข่าวออนไลน์ครั้งนี้แสดงให้เห็นความสอดคล้องกับสถานการณ์จริงและเหตุผลทางวิทยาศาสตร์ที่รองรับผลการวิจัยได้อย่างชัดเจน

สรุปผล

งานวิจัยนี้เน้นย้ำถึงประสิทธิภาพของเทคนิคเว็บสแครปปิ้งใน การรวบรวมและวิเคราะห์ข่าวสารเกี่ยวกับน้ำท่วมในประเทศไทย โดยใช้เครื่องมือ BeautifulSoup และ Selenium ในการสกัดข้อมูลจากเว็บไซต์ข่าว ก่อนนำข้อมูลเหล่านี้ไปประมวลผลด้วยเทคนิค Named Entity Recognition (NER) ในระบบประมวลผลภาษาธรรมชาติ (NLP) โดยใช้โมเดล WangchanBERTa เพื่อดึงข้อมูลสำคัญ เช่น ชื่อสถานที่และเวลาของเหตุการณ์ ผลการวิเคราะห์ชี้ให้เห็นว่า ข้อมูลที่ได้สามารถนำไปสรุปเหตุการณ์พื้นที่เสี่ยงภัยได้อย่างรวดเร็วและมีประสิทธิภาพ ทั้งยังสามารถรวบรวมข้อมูลตลอดปี 2567 ภายในเวลาไม่เกิน 1 ชั่วโมง 33 นาที พร้อมแสดงผลด้วยแผนที่ความร้อนเพื่อช่วยอธิบายสถานการณ์อย่างชัดเจน อย่างไรก็ตาม งานวิจัยนี้พบข้อจำกัดในด้านความน่าเชื่อถือของข้อมูลและความซับซ้อนในการประมวลผล โดยการขยายแหล่งข้อมูลอาจสร้างความท้าทายในแง่ของโครงสร้างและคุณภาพของข้อมูล ตลอดจนพฤติกรรมของข้อมูลที่หลากหลายมากขึ้น

เพื่อพัฒนาระบบให้ดียิ่งขึ้นในอนาคต ควรเพิ่มความหลากหลายของแหล่งข้อมูล เช่น การรวมข้อมูลจากโซเชียลมีเดียและแพลตฟอร์มข้อมูลสาธารณะของทางราชการ รวมถึงการใช้เทคนิคปัญญาประดิษฐ์ขั้นสูง เช่น

Deep Learning เพื่อเพิ่มความแม่นยำในการสกัดข้อมูล นอกจากนี้ การพัฒนาระบบแจ้งเตือนแบบเรียลไทม์ที่ผสานกับระบบภูมิสารสนเทศ (GIS) จะช่วยให้การแสดงผลข้อมูลมีความชัดเจนและเข้าถึงได้ง่ายยิ่งขึ้น

กิตติกรรมประกาศ

ขอขอบพระคุณสำนักงานคณะกรรมการส่งเสริมวิทยาศาสตร์ วิจัยและนวัตกรรม (สกสว.) ที่ได้ให้ทุนอุดหนุนการวิจัยจากงบประมาณแผ่นดิน ประจำปีงบประมาณ พ.ศ. 2567 และขอขอบคุณสำนักงาน ป้องกันและบรรเทาสาธารณภัยจังหวัดสระแก้ว และโครงการชลประทานสระแก้ว ที่ให้ข้อมูลและข้อเสนอแนะสำหรับงานวิจัยนี้เพื่อให้เกิดประโยชน์แก่ผู้เกี่ยวข้อง

เอกสารอ้างอิง

สำนักงานทรัพยากรน้ำแห่งชาติ. (2566). รายงานการติดตามการบริหารจัดการทรัพยากรน้ำฤดูฝน ปี 2566 . สืบค้นเมื่อ 2 มกราคม 2568 จากแหล่งข้อมูล :http://rbmd.onwr.go.th.

Aribowo, D., & Wisudarma, G. (2018). Natural disaster risk assessment in Indonesia using web scraping and analytic hierarchy process. Procedia Computer Science, 135, 542-549.

Kumar, A., & Sebastian, T. M. (2019). Sentiment analysis on Twitter. International Journal of Computer Science Issues, 9(4), 372-378.

Mitchell, R. (2018). Web scraping with Python: Collecting more data from the modern web. O'Reilly Media.

Manning, C. D., & Schütze, H. (2021). Foundations of statistical natural language processing. MIT press.

McKinney, W. (2022). Python for data analysis: Data wrangling with pandas, NumPy, and I Python. O'Reilly Media.

Scrapy, C. (2021). Scrapy documentation. Scrapy.