# การออกแบบและพัฒนาระบบสกัดข้อมูลกึ่งมีโครงสร้างตามกระบวนการอีแอลที และค้นคืนสารสนเทศด้วยการนำเสนอภาพข้อมูล กรณีศึกษา : รายงานอุบัติเหตุทางถนนรายใหญ่บนเว็บไซต์

# The Design and Development of a Knowledge Extraction and Retrieval System via Data Visualization. Case Study: Road Major Accidents on Website

**จักรินทร์ สันติรัตนภักดี** [1]
สาขาวิชาระบบสารสนเทศคอมพิวเตอร์ คณะบริหารธุรกิจ มหาวิทยาลัยวงษ์ชวลิตกุล[1]
**Chakkarin Santirattanaphakdi**
Department of Computer Information System, Faculty of Business Administrator, Vongchavalitkul University [1]
E-mail: chakkarin_san@vu.ac.th

*Received: October 9, 2020; Revised: November 13, 2020; Accepted: December 2, 2020*

**บทคัดย่อ**

งานวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนาการสกัดข้อมูลกึ่งมีโครงสร้างรายงานอุบัติเหตุทางถนนที่รุนแรงบนเว็บไซต์ตามกระบวนการอีแอลที ด้วยการสร้างคลังข้อมูลคำศัพท์ร่วมกับการรู้จำเอนทิตี้มาสกัดข้อมูลใน 6 ประเด็น ได้แก่ จังหวัด วันที่ จำนวนผู้เสียชีวิต จำนวนผู้บาดเจ็บ ประเภทและจำนวนของยานพาหนะที่เกิดเหตุ เพื่อสร้างระบบค้นคืนสารสนเทศด้วยการนำเสนอภาพข้อมูลปฏิสัมพันธ์กับผู้ใช้ผ่านทางเว็บไซต์ เมื่อประเมินความสามารถในการค้นคืนสารสนเทศ ในประเด็นจังหวัด และวันที่ วัดผลด้วยผู้เชี่ยวชาญ 5 คน จำนวน 77 ครั้งตามจำนวนจังหวัดในประเทศไทย ทำซ้ำจนครบ 3 ครั้งแล้วนำมาหาค่าเฉลี่ย พบว่า ค่าความถูกต้อง ค่าความครบถ้วน และค่าประสิทธิภาพโดยรวม เท่ากับ 0.87, 0.85 และ 0.86 ตามลำดับ จากนั้นนำมาประเมินความถูกต้อง เลือกใช้วิธีการสุ่มตัวอย่างโดยไม่ใช้ความน่าจะเป็น จำนวน 30 คน จากการเลือกกลุ่มตัวอย่างแบบเฉพาะเจาะจง ที่ความเชื่อมั่น 100% พบว่า ค่าเฉลี่ยผลการประเมินใน 4 ประเด็นจากการทำซ้ำ 3 ครั้งคือ จำนวนผู้เสียชีวิต จำนวนผู้บาดเจ็บ ชนิดและจำนวนของยานพาหนะในที่เกิดเหตุ มีค่าความถูกต้องจากการประเมินของผู้ใช้คิดเป็นร้อยละ 93.48, 78.21, 81.74 และ 80.26 ตามลำดับ พบว่าเงื่อนไขในการสกัดข้อมูลตามลำดับก่อนหลัง ส่งผลต่อความถูกต้องของจำนวนผู้บาดเจ็บและเสียชีวิต และการสร้างคลังคำศัพท์ชนิดของยานพาหนะที่เกี่ยวข้องร่วมกับการใช้หลักการรู้จำชื่อส่งผลต่อความถูกต้องของประเภทและจำนวนของยานพาหนะในที่เกิดเหตุแบบแปรผันตรง

**คำสำคัญ:** การสกัดข้อมูล ข้อมูลกึ่งมีโครงสร้าง อุบัติเหตุทางถนน การนำเสนอภาพข้อมูล

**ABSTRACT**

The research aimed to design and develop a semi-structured data knowledge extraction system for major accidents via website based on the ELT process. By corpus-based approach with named-entity recognition to extract data in 6 issues: province, date, number of death(s), number of injured, type and number of motor vehicle(s) involved in the accident to create an information retrieval system by interactive data visualization on website. The assessment of the retrieval of information system by province and date, it was conducted by 5 expert assessors for 77 times, covering 77 provinces in Thailand, then repeated the process 3 times and calculated the mean. It was found that the precision, recall and F-measure was 0.87, 0.85 and 0.86 respectively. To determine the validity of the system, a sample size of 30 from purposive sampling with reliability 100% confidence. The assessment by 30 users, repeated 3 times, in 4 issues: the number of deaths, number of injured, types and number of motor vehicle(s) involved in the accident. The accuracy was 93.48, 78.21, 81.74 and 80.26 % respectively. The research found that the sequence of information extraction conditions, before and after, affected the accuracy of the number of injuries and deaths. And the creation of vocabulary bank of the types of vehicles involved, coupled with the corpus-based approach named entity recognition method affected the accuracy of the type and number of motor vehicle(s) involved in the accident in direct variation.

**KEYWORDS:** Data Extraction, Semi-Structure Data, Road Accidents, Data Visualization

**บทนำ**

อุบัติเหตุทางถนนนับเป็นหนึ่งในปัญหาสำคัญระดับโลก จากการรายงานขององค์การอนามัยโลก (World Health Organization: WHO, 2018) พบว่า มีผู้เสียชีวิตจากการอุบัติเหตุทางถนนเฉลี่ยวันละ 3,700 คน ซึ่งประเทศไทยมีผู้เสียชีวิตจากอุบัติเหตุทางถนนสูงที่สุดเป็นอันดับหนึ่งในเอเชียและอันดับที่ 9 ของโลก โดยมีประมาณการผู้เสียชีวิต 32.7 คนต่อประชากรหนึ่งแสนคน หรือปีละ 22,491 คน เฉลี่ย 60 คนต่อวัน ประเทศไทยเล็งเห็นถึงความสำคัญของการนำข้อมูลมาช่วยในการวางแผนป้องกันอุบัติเหตุทางถนน ด้วยข้อมูลสถานการณ์ที่สามารถมองเห็นแนวโน้มความรุนแรงของปัญหา ตลอดจนสาเหตุและปัจจัยที่ทำให้เกิดอุบัติเหตุ เพื่อนำไปสู่การแก้ไขปัญหาอย่างเป็นรูปธรรม อันจะเห็นจากมีการเก็บข้อมูลพื้นฐานสำหรับการวิเคราะห์เชิงสถิติจากสถานการณ์ และสาเหตุการเกิดอุบัติเหตุทางถนน เพื่อวิเคราะห์ถึงแนวโน้มสถานการณ์อุบัติเหตุทางถนนรวมทั้งนำมากำหนดทิศทางนโยบายต่างๆทั้งระยะเร่งด่วนและระยะยาว แต่ปัจจุบันข้อมูลที่นำมาใช้มาจากหลายหน่วยงาน และมีวัตถุประสงค์ในการเก็บข้อมูลแตกต่างกัน (สำนักโรคไม่ติดต่อ กรมควบคุมโรค, 2559) ทำให้ข้อมูลจำนวนผู้บาดเจ็บและเสียชีวิตคลาดเคลื่อน และไม่มีข้อมูลจากหน่วยงานใดที่มีความครอบคลุมครบถ้วน จากปัญหาที่เกิดขึ้นทำให้ข้อมูลอุบัติเหตุของประเทศไทยไม่เป็นที่น่าเชื่อถือ

ดังนั้น ข้อมูลที่น่าเชื่อถือและยังรอการนำไปใช้ประโยชน์อีกรูปแบบหนึ่ง นั่นคือการรายงานอุบัติเหตุในสถานการณ์จริง จากผู้รับผิดชอบในแต่ละพื้นที่ที่รายงานมายังศูนย์รับแจ้งอุบัติเหตุ และประสานการช่วยเหลือผู้ประสบภัยจากรถ บริษัท

กลางคุ้มครองผู้ประสบภัยจากรถ จำกัด ผ่านเว็บไซต์ www.thairsc.com ในลักษณะรายงานอุบัติเหตุรายใหญ่ ที่มีข้อมูลที่ปรากฏบนเว็บไซต์ตั้งแต่ปี 2557 – ปัจจุบัน ในรูปแบบของข้อมูลประเภทกึ่งมีโครงสร้าง จากการกำหนดหัวข้อในการบันทึกข้อมูล โดยแต่ละช่องเปิดโอกาสให้ผู้รายงานสามารถใส่ข้อมูลตามสถานการณ์ที่เกิดขึ้นจริง ทำให้ในแต่ละครั้งข้อมูลในการรายงานขาดความถูกต้อง ไม่เป็นมาตรฐาน และแตกต่างกันไปตามบริบทของผู้รายงาน จึงยากที่จะนำข้อมูลการรายงานอุบัติเหตุมาใช้ให้เกิดประโยชน์ในการวางแผนป้องกันและแก้ปัญหา

ผู้วิจัยทราบถึงปัญหา และเล็งเห็นถึงความสำคัญของข้อมูล จึงออกแบบและพัฒนาระบบสกัดข้อมูลรายงานอุบัติเหตุทางถนนรายใหญ่บนเว็บไซต์ ตามกระบวนการอีแอลที และค้นคืนสารสนเทศด้วยเทคนิคการนำเสนอภาพข้อมูล เพื่อนำข้อมูลดังกล่าวมาเป็นส่วนหนึ่งในการวางแผนป้องกันอุบัติเหตุ ตลอดจนเสนอแนวทางการออกแบบและพัฒนาการสกัดข้อมูลประเภทกึ่งมีโครงสร้าง และค้นคืนสารสนเทศในอนาคต

**วัตถุประสงค์**

1. เพื่อออกแบบและพัฒนาการสกัดข้อมูลกึ่งมีโครงสร้างตามกระบวนการอีแอลที
2. เพื่อค้นคืนสารสนเทศด้วยการนำเสนอภาพข้อมูลแบบปฏิสัมพันธ์กับผู้ใช้

**ทบทวนวรรณกรรม**

1. อีทีแอล (Extract, Transform and Load : ETL) เป็นกระบวนการหนึ่งในระบบคลังข้อมูล (Data Warehouse) เริ่มต้นจาก 1) การสกัดข้อมูล (Kimball and Caserta, 2004) จากแหล่งข้อมูลที่อยู่ในรูปแบบต่างๆ ซึ่งมีความซับซ้อนในการเข้าถึงที่แตกต่างกัน 2) การเปลี่ยนรูปข้อมูล (Golffarelli and Rizzi, 2009) เป็นกระบวนการทำความสะอาดข้อมูล (Data Cleaning) เพื่อให้ได้คุณภาพ และเป็นมาตรฐานสำหรับนำไปใช้งาน และ 3) การถ่ายโอนข้อมูลเข้าสู่คลังข้อมูลที่กำหนดไว้ (Ponniah, 2010)

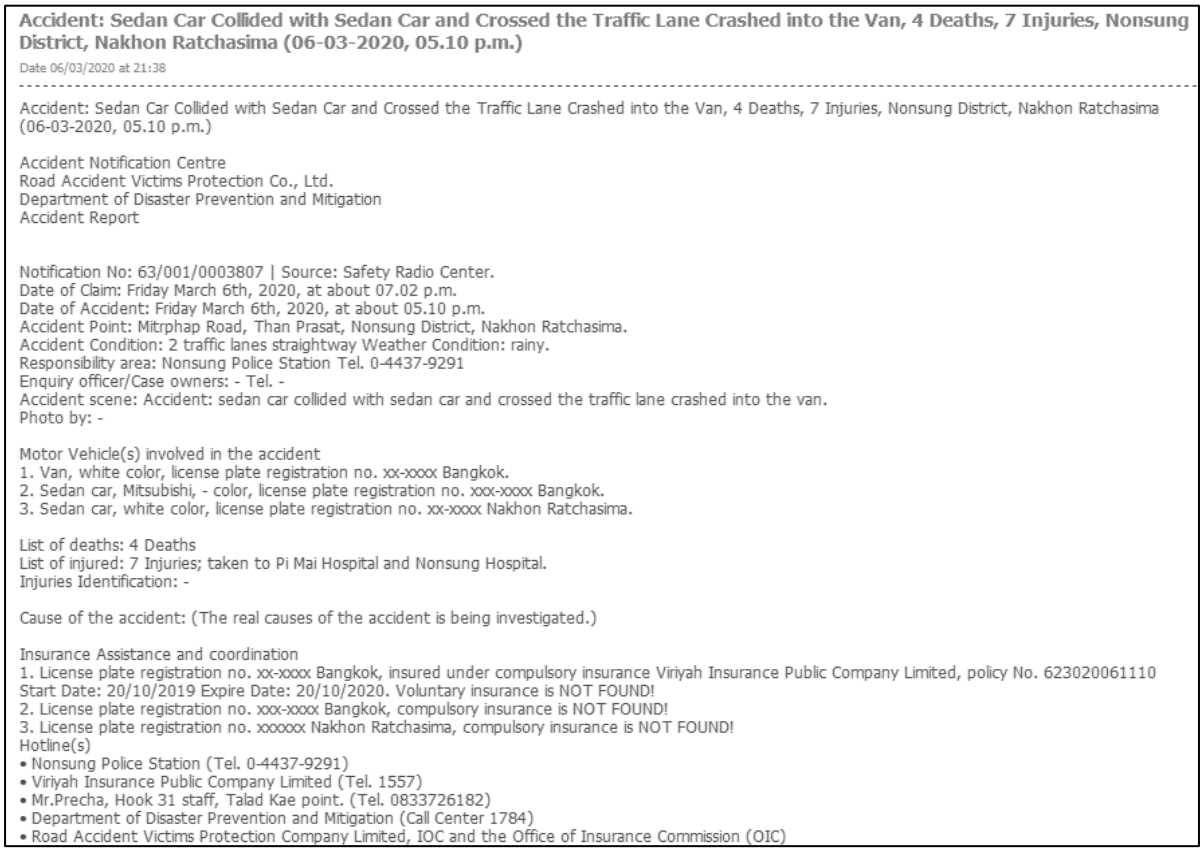
จะเห็นได้ว่ากระบวนการอีทีแอลมีการนำไปประยุกต์ใช้อย่างหลากหลาย ซึ่งอาจเปลี่ยนแปลงลำดับการทำงานตามความเหมาะสมในการดำเนินงาน ที่หลังจากการสกัดข้อมูลแล้วจึงถ่ายโอนข้อมูลไปไว้ในฐานข้อมูลก่อนที่จะทำการเปลี่ยนรูปข้อมูลเป็นขั้นตอนสุดท้ายก่อนนำไปใช้งาน เรียกว่า อีแอลที (ELT) (Rainardi, 2008) ที่ได้รับความนิยมอย่างกว้างขวาง โดยเฉพาะอย่างยิ่งกับข้อมูลแบบมีโครงสร้าง และแบบกึ่งมีโครงสร้าง เนื่องจากช่วยลดการใช้ทรัพยากร และต้นทุนของระบบ

3. ข้อมูลกึ่งมีโครงสร้าง (Semi-Structure Data) เป็นรูปแบบของข้อมูลที่มีโครงสร้างไม่เป็นไปตามโครงสร้างของแบบจำลองข้อมูลเชิงสัมพันธ์ หรือรูปแบบ ของตารางข้อมูล (Garcia-Molina, Ullman & Widom, 2009) แต่จะมีแท็ก (tag) หรือเครื่องหมายอื่นๆ เพื่อแยกองค์ประกอบของข้อมูลแต่ละส่วน ที่สื่อความหมายและแสดงถึงลำดับชั้นของระเบียนและข้อมูลภายใน ดังภาพที่ 1

| XML | JSON | YAML |
|---|---|---|
| <Servers><br><Server><br><name>Server1</name><br><owner>Steve</owner><br><created>123456</created><br></Server><br></Servers> | {<br>Servers [<br>{<br>name: Server1,<br>owner: Steve,<br>created: 123456<br>}<br>]<br>} | Servers:<br>-<br>name: Server1<br>owner: Steve<br>created: 123456 |

**ภาพที่ 1** คำสั่งภาษา XML, JSON และ YAML ที่อยู่ในรูปแบบข้อมูลกึ่งมีโครงสร้าง

อย่างไรก็ดี ข้อมูลกึ่งมีโครงสร้างมิได้จำกัดอยู่เพียงที่กล่าวมาเท่านั้น แต่ยังรวมถึงอีเมล ซึ่งเนื้อหาอยู่ในรูปแบบข้อมูลไม่มีโครงสร้าง หากแต่ชื่อผู้ส่ง ชื่อผู้รับ วันที่ และหัวเรื่องแสดงให้เห็นถึงโครงสร้างของข้อมูล ร่วมถึงเนื้อหาที่ปรากฏในสื่อสังคมออนไลน์ต่างๆ ที่มีจำนวนการไลค์ แชร์ ความคิดเห็น รวมถึงชื่อของผู้เขียน และผู้อ่านแสดงภายใต้โครงสร้างที่กำหนดไว้ ทำให้เข้าใจถึงองค์ประกอบของข้อมูล เช่นเดียวกับรายงานอุบัติเหตุรายใหญ่ที่ปรากฏบนเว็บไซต์ www.thairsc.com ประกอบด้วย 2 ส่วน คือส่วนพาดหัว และส่วนรายละเอียดที่อยู่ในรูปแบบของข้อมูลประเภทกึ่งมีโครงสร้าง ที่กำหนดหัวข้อแต่ละประเด็นให้ผู้อ่านเข้าใจ ดังภาพที่ 2



**Accident: Sedan Car Collided with Sedan Car and Crossed the Traffic Lane Crashed into the Van, 4 Deaths, 7 Injuries, Nonsung District, Nakhon Ratchasima (06-03-2020, 05.10 p.m.)**

Date 06/03/2020 at 21:38

---

Accident: Sedan Car Collided with Sedan Car and Crossed the Traffic Lane Crashed into the Van, 4 Deaths, 7 Injuries, Nonsung District, Nakhon Ratchasima (06-03-2020, 05.10 p.m.)

Accident Notification Centre
Road Accident Victims Protection Co., Ltd.
Department of Disaster Prevention and Mitigation
Accident Report

Notification No: 63/001/0003807 | Source: Safety Radio Center.
Date of Claim: Friday March 6th, 2020, at about 07.02 p.m.
Date of Accident: Friday March 6th, 2020, at about 05.10 p.m.
Accident Point: Mitrphap Road, Than Prasat, Nonsung District, Nakhon Ratchasima.
Accident Condition: 2 traffic lanes straightway Weather Condition: rainy.
Responsibility area: Nonsung Police Station Tel. 0-4437-9291
Enquiry officer/Case owners: - Tel. -
Accident scene: Accident: sedan car collided with sedan car and crossed the traffic lane crashed into the van.
Photo by: -

Motor Vehicle(s) involved in the accident
1. Van, white color, license plate registration no. xx-xxxx Bangkok.
2. Sedan car, Mitsubishi, - color, license plate registration no. xxx-xxxx Bangkok.
3. Sedan car, white color, license plate registration no. xx-xxxx Nakhon Ratchasima.

List of deaths: 4 Deaths
List of injured: 7 Injuries; taken to Pi Mai Hospital and Nonsung Hospital.
Injuries Identification: -

Cause of the accident: (The real causes of the accident is being investigated.)

Insurance Assistance and coordination
1. License plate registration no. xx-xxxx Bangkok, insured under compulsory insurance Viriyah Insurance Public Company Limited, policy No. 623020061110 Start Date: 20/10/2019 Expire Date: 20/10/2020. Voluntary insurance is NOT FOUND!
2. License plate registration no. xxx-xxxx Bangkok, compulsory insurance is NOT FOUND!
3. License plate registration no. xxxxxx Nakhon Ratchasima, compulsory insurance is NOT FOUND!
Hotline(s)
• Nonsung Police Station (Tel. 0-4437-9291)
• Viriyah Insurance Public Company Limited (Tel. 1557)
• Mr.Precha, Hook 31 staff, Talad Kae point. (Tel. 0833726182)
• Department of Disaster Prevention and Mitigation (Call Center 1784)
• Road Accident Victims Protection Company Limited, IOC and the Office of Insurance Commission (OIC)

**ภาพที่ 2** รายงานอุบัติเหตุรายใหญ่

จะเห็นได้ว่าการกำหนดหัวข้อเป็นโครงสร้างในแต่ละองค์ประกอบ แต่ข้อมูลที่ถูกบันทึกจะแตกต่างกันไปตามบริบทผู้รายงาน ส่งผลให้ขาดความถูกต้อง ไม่เป็นมาตรฐาน และยากต่อการนำไปใช้ประโยชน์

4. อุบัติเหตุ (Accident) หมายถึง เหตุการณ์ที่เกิดโดยมิได้เกิดจากความตั้งใจกระทำของบุคคล และมิได้คาดคิดซึ่งก่อให้เกิดความเสียหายทั้งทางร่างกายและจิตใจ (The World Health Organization: WHO, (2018)) ดังนั้นอุบัติเหตุทางถนน (Road Accident) ย่อมหมายถึง เหตุการณ์ที่เกิดขึ้นบนถนนอย่างไม่คาดคิดโดยไม่มีสิ่งบอกเหตุล่วงหน้าแต่มีสาเหตุและส่งผลกระทบที่สามารถชี้วัดได้ ในแต่ละครั้งอาจทำให้มีผู้เสียชีวิต ณ ที่เกิดเหตุ หรือภายใน 24 ชั่วโมง รวมถึงมีผู้บาดเจ็บในที่เกิดเหตุ ซึ่งความรุนแรงของอุบัติเหตุแบ่งได้ตามความเสียหายที่เกิดขึ้นในแต่ละครั้ง โดยอุบัติเหตุใหญ่ (Major Accident) หมายถึง อุบัติเหตุทางถนนที่มีจำนวนผู้เสียชีวิต 2 รายขึ้นไป หรือผู้บาดเจ็บตั้งแต่ 4 คนขึ้นไป หรือผู้บาดเจ็บ รวมผู้เสียชีวิต ตั้งแต่ 4 คนขึ้นไป ตลอดจนอุบัติเหตุที่เป็นที่สนใจของประชาชน เช่น บุคคลสำคัญหรือบุคคลสาธารณะเกิดบาดเจ็บแอ็ดมิท หรือเสียชีวิตจากอุบัติเหตุ หรือรถน้ำมัน รถสารเคมี เกิดอุบัติเหตุ จนเกิดเพลิงไหม้ หรือสารเคมีฟุ้งกระจายไปสู่ชุมชน เป็นต้น (ศูนย์อำนวยการความปลอดภัยทางถนน, 2560) ที่รายงานมายังศูนย์รับแจ้งอุบัติเหตุ และประสานการช่วยเหลือผู้ประสบภัยจากรถ บริษัทกลางคุ้มครองผู้ประสบภัยจากรถจำกัดเผยแพร่ต่อสาธารณะผ่านเว็บไซต์ www.thairsc.com เป็นข้อมูลเพื่อนำไปเยียวยาผู้ประสบภัยจากรถให้ได้รับการคุ้มครองตามกฎหมาย และเพื่อเป็นแนวทางในการป้องกันและลดการเกิดอุบัติเหตุทางถนน เพื่อสนับสนุนแผนปฏิบัติการทศวรรษแห่งความปลอดภัยทางถนน 2554-2563

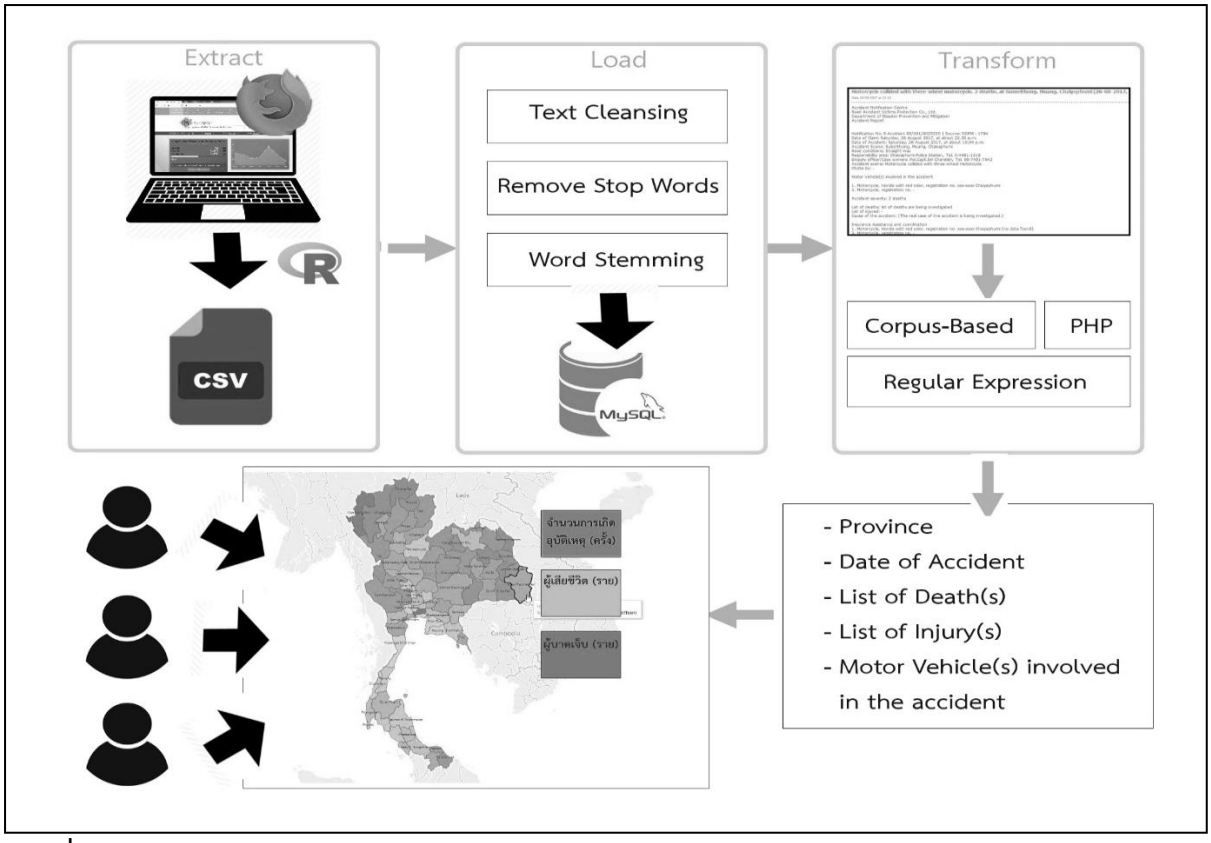
5. การนำเสนอภาพข้อมูล (Data Visualization) เป็นการแปลงข้อมูลเชิงปริมาณให้กลายเป็นภาพ เพื่อแสดงผลแทนการบรรยายเป็นตัวอักษร หรือคำพูดในลักษณะของกราฟ แผนภูมิ ภาพวาด ภาพถ่าย รวมถึงการใช้สีเพื่อจำแนกข้อมูลให้สวยงาม และง่ายต่อการทำความเข้าใจ (Wike, 2019) เนื่องจากมนุษย์มีความสามารถในการรับรู้ด้วยตามากกว่าประสาทสัมผัสอื่นๆ ดังนั้นเมื่อมองเห็นภาพก็จะสามารถเก็บข้อมูลทั้งหมดได้พร้อมกันในทันที จากนั้นสมองจะนำข้อมูลที่เห็นมาประมวลผล ต่างจากการรับข้อมูลด้วยการฟัง ซึ่งผู้ฟังจะต้องรับข้อมูลต่อเนื่องจนจบข้อความก่อนที่จะสามารถนำไปประมวลผลได้ (Chen, Härdle & Unwin (2008)) นอกจากนี้ยังทำให้มองเห็นสิ่งที่ไม่สามารถทราบได้จากการอ่านข้อมูลตัวเลขเพียงอย่างเดียว เช่น แนวโน้มหรือรูปแบบการกระจายตัวของข้อมูลผู้เสียชีวิตจากโรคระบาดที่ไม่สามารถทราบได้จากตัวเลขเชิงสถิติ เมื่อแสดงในรูปแบบแผนที่จะช่วยให้ทราบว่าโรคนั้นเกิดขึ้นบ่อยใน สภาพแวดล้อมแบบใด Fry (2007) เสนอ 7 ขั้นตอนของการนำเสนอภาพข้อมูล ได้แก่ 1) การรวบรวมข้อมูล (Acquire) 2) การแยกข้อมูล (Parse) 3) การกรองข้อมูล (Filter) 4) การใช้หลักคณิตศาสตร์ สถิติ และเหมืองข้อมูล (Mine) 5) การกำหนดรูปแบบการนำเสนอข้อมูล (Represent) 6) การใช้กราฟิกช่วยในการนำเสนอ (Refine) และ 7) การสร้างปฏิสัมพันธ์ (Interact) ในงานวิจัยชิ้นนี้สรุปเป็น 3 ขั้นตอนสอดคล้องกับกระบวนการอีแอลที ได้แก่ 1) การรวบรวมข้อมูลจากแหล่งต่างๆ (Data Acquisition) ด้วยการเข้าถึงข้อมูลรายงานอุบัติเหตุรายใหญ่ที่เผยแพร่เป็นสาธารณะบนอินเทอร์เน็ตด้วยกระบวนการขูดเว็บ 2) การเตรียมข้อมูล (Data Preparation) ปรับโครงสร้างข้อมูลให้เหมาะสมกับการนำเสนอ โดยทำความสะอาดและปรับโครงสร้างของข้อมูล และ 3) การจัดทำภาพข้อมูลให้อยู่ในรูปแบบที่เหมาะสม (Data Presentation) เพื่อให้สามารถสื่อความได้ถูกต้องตรงตามความต้องการ ในงานวิจัยชิ้นนี้ใช้ Google Charts เป็นเครื่องมือการนำเสนอภาพข้อมูล และแสดงผลปฏิสัมพันธ์กับผู้ใช้บนเว็บไซต์

6. งานวิจัยในปัจจุบันที่ออกแบบและพัฒนากระบวนการสกัดข้อมูลได้เสนอหลากหลายวิธีการ แบ่งตามกระบวนการทำงานเป็น 3 กลุ่ม ได้แก่ 1) การตัดคำโดยใช้กฎ (Rule-Based Approach) เป็นการตัดคำโดยใช้วิธีเกณฑ์ทางอักขรวิธีที่กำหนดลักษณะของการประสมอักษร โดย กานดา รุณนะพงศา และปโยธร อุราธรรมกุล (2548) ได้ทำการวิจัยเรื่องการตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่ พบว่ามีความถูกต้องมากขึ้น 2) การตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) ที่เก็บคำศัพท์ไว้ในพจนานุกรม แล้วนำข้อความป้อนเข้าไปค้นหาและเปรียบเทียบสายอักขระกับคำศัพท์ในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวควรตัดคำในบริเวณใดและประกอบด้วยคำใดบ้าง โดยสถิตย์โชค โพธิ์สอาด และ ปิติภูมิ โพสาวัง (2018) ทำการวิจัยเรื่อง การจำแนกพฤติกรรมการขับขี่รถโดยสารสาธารณะโดยใช้วิธีการสกัดข้อความ และเทคนิคการเรียนรู้ของเครื่อง โดยใช้พจนานุกรมคำศัพท์อิเล็กทรอนิกส์ (LEXiTRON) ในการตัดคำ อย่างไรก็ดี การตัดคำโดยใช้พจนานุกรมก็มีข้อจำกัดบางประการ เนื่องจากมีความเป็นไปได้ที่คำที่ปรากฏในเอกสาร อาจจะไม่ปรากฏในพจนานุกรม จึงเป็นที่มาของเอ็นแกรม (N-gram) ที่นำบางส่วนของข้อความออกมาเป็นตามค่า N และ 3) การตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) โดยเตรียมคลังข้อมูลที่มีการตัดคำและการกำกับหน้าที่ของคำไว้ล่วงหน้า ซึ่งสุรจิต ภูมิคง, ธรา อั่งสกุล และจิติมนต์ อั่งสกุล (2013) ได้นำเสนอเทคนิคและวิธีการสกัดข่าวด้านความปลอดภัยจากเว็บไซต์รวบรวมข่าว นำมาจัดประเภทและวิเคราะห์คำศัพท์ที่เกี่ยวข้อง เพื่อนำมาใช้เป็นฐานข้อมูลในการสกัดคำศัพท์ที่เกี่ยวข้องกับความไม่ปลอดภัยออกจากเนื้อข่าวแบบอัตโนมัติ

จากการทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้อง พบว่า มีหลากหลายแนวคิดในการพัฒนาวิธีการที่เพิ่มประสิทธิภาพในการสกัดข้อมูล ในงานวิจัยนี้เป็นการออกแบบและพัฒนาระบบสกัดข้อมูลรายงานอุบัติเหตุทางถนนรายใหญ่บนเว็บไซต์ด้วยการสร้างคลังข้อมูลคำศัพท์ ร่วมกับการกำหนดรูปแบบอักขระ (Regular Expression) และการโปรแกรมมิ่งภาษา PHP เพื่อสกัดข้อมูลด้วยวิธีการรู้จำชื่อเอนทิตี้ (Named Entity Recognition : NER) เพื่อนำผลลัพธ์ที่ได้มาสร้างระบบค้นคืนสารสนเทศด้วยเทคนิคการนำเสนอภาพข้อมูลผ่านบนเว็บไซต์

**วิธีดำเนินการวิจัย**

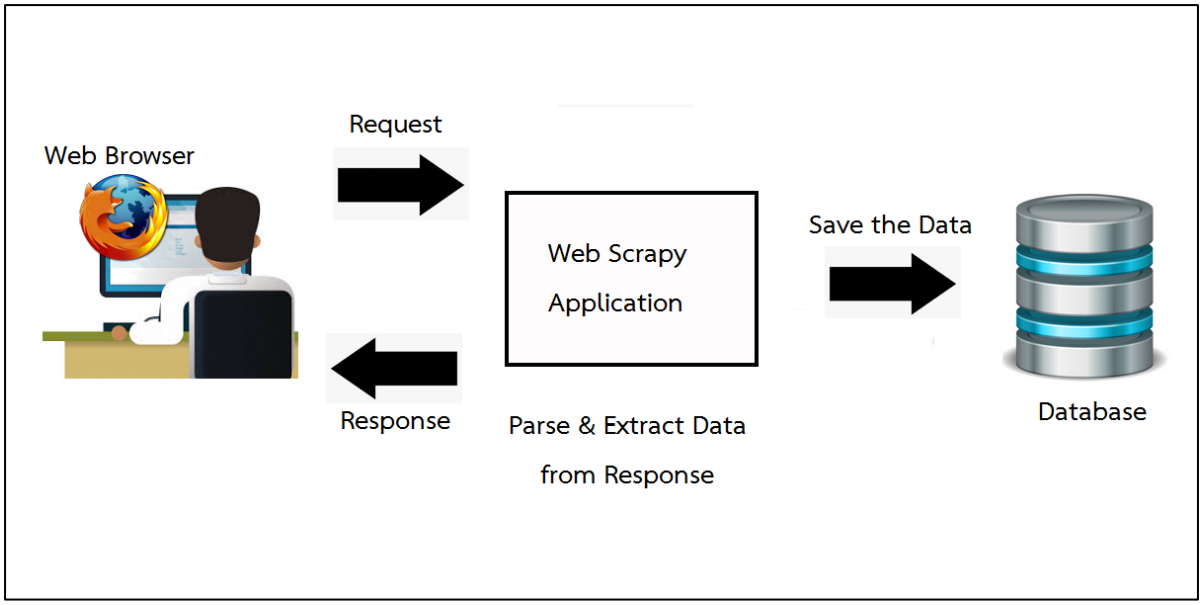
งานวิจัยชิ้นนี้ใช้กระบวนการอีแอลทีเป็นกรอบในการดำเนินงานกับข้อมูลแบบกึ่งมีโครงสร้าง ประกอบด้วย 3 ส่วน ได้แก่ 1) การสกัดข้อมูล 2) การถ่ายโอนข้อมูลเข้าสู่คลังข้อมูล และ 3) การเปลี่ยนรูปข้อมูล เพื่อค้นคืนสารสนเทศด้วยการนำเสนอภาพข้อมูล ดังภาพ 3




**ภาพที่ 3** กระบวนการสกัดข้อมูลและค้นคืนสารสนเทศตามกรอบอีแอลที

**1) การสกัดข้อมูล (Extract)** เป็นการเข้าถึงข้อมูลหรือรวบรวมข้อมูลจากแหล่งต่างๆ ก่อนที่จะนำไปใช้งาน เนื่องจากข้อมูลในเว็บไซต์มีจำนวนมาก การจะเข้าถึงข้อมูลทีละรายการด้วยตนเองไม่ใช่เรื่องง่าย อีกทั้งรูปแบบข้อมูลที่อยู่บนเว็บไซต์นั้นถูกออกแบบมาเพื่อแสดงผลให้กับผู้ใช้เท่านั้น ไม่สามารถนำมาวิเคราะห์ได้โดยตรง จึงเป็นที่มาของการขูดเว็บ (Web Scraping) ที่ช่วยดำเนินการกับข้อมูลที่ต้องการให้มาอยู่ในรูปแบบที่พร้อมนำไปวิเคราะห์และประมวลผลต่อไป

หลักการทำงานของการขูดเว็บจะเป็นแนวทางเดียวกับการเข้าใช้งานเว็บไซต์โดยทั่วไปผ่านโปรโตคอลเอชทีทีพี (Mitchell, 2015) ดังภาพที่ 4 กล่าวคือเมื่อผู้ใช้มีการร้องขอบริการจากเว็บไซต์ผ่านเว็บเบราว์เซอร์ จะทำการส่งคำขอ (Request) ไปยังเครื่องแม่ข่าย ซึ่งต้องผ่านการประมวลผลจากเครื่องบริการแปลงชื่อเว็บไซต์เป็นหมายเลขไอพี (DNS Server) ในแต่ละชั้นตามข้อกำหนด จากนั้นเมื่อเครื่องแม่ข่ายทำการประมวลผลคำขอ หากต้องมีการประมวลผลที่ฝั่งแม่ข่ายหรือการติดต่อฐานข้อมูลก็จะดำเนินการโดยใช้เวลาและทรัพยากรเครือข่ายที่แตกต่างกัน เมื่อประมวลผลเสร็จสิ้นจะตอบสนอง (Response) กลับมายังผู้ใช้ในรูปแบบของเอชทีเอ็มแอลผ่านเว็บเบราว์เซอร์อีกครั้ง ดังนั้นการขูดเว็บจะทำงานเสมือนกับว่าผู้ใช้เข้าใช้เว็บไซต์ด้วยตนเอง แต่ในทางกลับกันจะเขียนคำสั่งหรือใช้โปรแกรมสำเร็จรูปในการเข้าถึงเว็บไซต์และคัดลอกข้อมูลในตำแหน่งที่ต้องการมาอยู่ในรูปของไฟล์ที่รอการนำไปประมวลผลต่อไป




**ภาพที่ 4** กระบวนการขูดเว็บ

อย่างไรก็ดี การสกัดข้อมูลบนเว็บไซต์มีข้อควรระวังอยู่ 2 ประการ คือ ระวังมิให้สร้างภาระในการทำงานมากเกินไปจนอาจก่อให้เกิดความเสียหายแก่เครื่องแม่ข่าย และการอนุญาตให้เข้าถึงและคัดลอกข้อมูลบนเว็บไซต์ โดยรายงานอุบัติเหตุจราจรนี้เป็นข้อมูลเพื่อนำไปเยียวยาผู้ประสบภัยจากรถให้ได้รับการคุ้มครองตามกฎหมาย และเพื่อเป็นแนวทางในการป้องกันและลดการเกิดอุบัติเหตุทางถนน และเผยแพร่ผ่านเว็บไซต์แบบสาธารณะ ดังนั้นผู้วิจัยทำการทดสอบว่าเว็บไซต์ www.thairsc.com นั้นอนุญาตการใช้งานของเว็บครอว์เลอร์ (Web Crawler) ซึ่งเป็นโปรแกรมอัตโนมัติที่ใช้ในการเก็บข้อมูลจากเว็บไซต์ด้วยไลบรารี robots.txt โดยผลลัพธ์คือการคืนค่า TRUE แสดงถึงเว็บไซต์ดังกล่าวสามารถใช้เว็บครอว์เลอร์ในการเก็บข้อมูลใน 6 ประเด็น ได้แก่ จังหวัด วันที่ จำนวนผู้เสียชีวิต จำนวนผู้บาดเจ็บ ประเภท และจำนวนของยานพาหนะที่เกิดเหตุ โดยจะไม่นำข้อมูลส่วนบุคคลของผู้ประสบเหตุมาเป็นส่วนหนึ่งของผลลัพธ์จากการสกัดข้อมูล

งานวิจัยชิ้นนี้ใช้กระบวนการสกัดข้อมูลจากเว็บไซต์ www.thairsc.com จะใช้การโปรแกรมมิ่งภาษา R (Khalil & Fakir, 2017) ด้วยไลบรารี Rvest และ xml2 ผ่านโปรแกรม Rstudio ที่มีประสิทธิภาพสูงในการวิเคราะห์และจัดการกับข้อมูล ซึ่งทั้งหมดเป็นซอฟต์แวร์โอเพนซอร์ส (Open Source) ในการเข้าถึงเว็บเพจแต่ละหน้าที่กำหนดไว้แบบวนลูป จากนั้นจะคัดลอกข้อมูลรายงานอุบัติเหตุทางถนนรายใหญ่จากเว็บไซต์ ในรูปแบบภาษาอังกฤษที่ปรากฏบนเว็บไซต์ตั้งแต่ 1 มกราคม 2559 – 31 พฤษภาคม 2562 รวมทั้งสิ้น 2,897 รายการ จากนั้นนำมาบันทึกเป็นไฟล์ CSV (Comma Separated Value) ดังภาพที่ 5 เพื่อรอนำไปประมวลผลในขั้นตอนต่อไป

| A | B | C | D |
|---|---|---|---|
| 3001 | Accident Sedan Car and Sedan Car Crossed the Traffic Lane Collided with Van, 4 Deaths, 7 Injuries, Nonsung District, Nakhon Ratchasima (06-03-2020, 05.10 p.m.) | Accident: Sedan Car and Sedan Car Crossed the Traffic Lane Collided with Van 4 Deaths 7 Injuries Nonsung District Nakhon Ratchasima (06032020 05.10 p.m.)Accident Notification CentreRoad Accident Victims Protection Co. Ltd. Department of Disaster Prevention and MitigationAccident ReportNotification No: 63/001/0003807 \| Source: Safety Radio Center. Date of Claim: Friday March 6th 2020 at about 07.02 p.m. Date of Accident: Friday March 6th 2020 at about 05.10 p.m.Accident Point: Mitrphap Road Than Prasat Nonsung District Nakhon Ratchasima. Accident Condition: 2lane traffic straightway Weather Condition: rain. Responsibility area: Nonsung Police Station Tel. 044379291Enquiry officer/Case owners: Tel. Accident scene: Accident: sedan car and sedan car crossed the traffic lane collided with van.Photo by: Motor Vehicle(s) involved in the accident1.Van white color license plate registration no. xxxxxx Bangkok.2.Sedan car Mitsubishi color license plate registration no. xxxxxxx Bangkok.3.Sedan car white color license plate registration no. xxxxxx Nakhon Ratchasima.List of deaths: 4 Deaths List of injured: 7 Injuries; taken to Pi Mai Hospital and Nonsung Hospital. Injuries Identification: Cause of the accident: (The real causes of the accident is being investigated.)Insurance Assistance and coordination1.License plate registration no. xxxxxx Bangkok insured under compulsory insurance Viriyah Insurance Public Company Limited policy No. 623020061110 Start Date: 20/10/2019 Expire Date: 20/10/2020. Voluntary insurance is NOT FOUND 2.License plate registration no. xxxxxxx Bangkok compulsory insurance is NOT FOUND 3.License plate registration no. xxxxxx Nakhon Ratchasima compulsory insurance is NOT FOUNDHotline(s)Nonsung Police Station (Tel. 044379291)Viriyah Insurance Public Company Limited (Tel. 1557)Mr.Precha Hook 31 staff Talad Kae point. (Tel. 0833726182)Department of Disaster Prevention and Mitigation (Call Center 1784)Road Accident Victims Protection Company Limited IOC and the Office of Insurance Commission (OIC)Road Accident Victims Protection Co. Ltd.44/1 Rungrojthanakul Building 11th Floor Ratchadapisek Rd. Huaykwang Bangkok 10310Call Center 1791 Tel: 021009191 Fax: 0264302934www.rvp.co.th www.thairsc.com | http://www.thairsc.com/th/BigAccDetail.aspx?l=en&qid=48011 |

**ภาพที่ 5** รายงานอุบัติเหตุทางถนนรายใหญ่จาก www.thairsc.com ที่อยู่ในรูปของ CSV

**2) การถ่ายโอนข้อมูลเข้าสู่คลังข้อมูล (Load)** ในเบื้องต้นผู้วิจัยทำการคัดกรองข้อมูล (Data Filtering) เพื่อให้ได้ข้อมูลที่มีความสมบูรณ์ โดยตรวจสอบโครงสร้างของข้อมูลจากขั้นตอนที่ผ่านมา พบว่า รายงานอุบัติเหตุรายใหญ่ที่อยู่ในรูปแบบภาษาอังกฤษจำนวนหนึ่งยังขาดความสมบูรณ์ในด้านเนื้อหาเมื่อเทียบกับรายงานในรูปแบบภาษาไทย ดังนั้นจึงพิจารณาตัดข้อมูลที่ไม่ตรงตามเงื่อนไขออก คงเหลือข้อมูลทั้งสิ้น 1,362 รายการ จากนั้นนำข้อมูลเข้าสู่โปรแกรม Rstudio อีกครั้ง เพื่อเข้าสู่กระบวนการเตรียมข้อมูลรูปแบบข้อความก่อนนำไปประมวลผล (Text Preprocessing) ด้วยการเขียนโปรแกรมภาษา R เพื่อให้ได้ข้อมูลที่มีคุณภาพ และอยู่ในอยู่ในรูปแบบที่เป็นมาตรฐานพร้อมสำหรับการนำไปใช้งานในขั้นตอนถัดไป แบ่งออกเป็น 3 ขั้นตอน ดังนี้

2.1) การทำความสะอาดข้อความ (Text Cleaning) ก่อนจะนำไปใช้งานจำเป็นต้องการจัดการกับเครื่องหมาย หรืออักขระพิเศษที่ใช้ในการแสดงผล ตลอดจนเปลี่ยนตัวอักษรภาษาไทยให้เป็นภาษาอังกฤษทั้งหมด เพื่อให้เหลือเฉพาะข้อความที่พร้อมนำไปประมวลผล ตลอดจนตัดข้อความส่วนท้ายของทุกๆรายการออก เนื่องจากเป็นเพียงรายละเอียดที่ตั้งและช่องทางการติดต่อ มิได้มีผลต่อการนำไปประมวลผล

2.2) การกำจัดคำหยุด (Remove Stop Words) คือการกำจัดคำที่มักพบบ่อยๆ ในประโยค แต่ไม่ค่อยช่วยในการสื่อความหมาย เช่น a, an, the, also, just, etc. เป็นต้น มักอยู่ในรูปของคำบุพบท หรือคำสันธาน ที่ทำหน้าที่เชื่อมคำหรือเชื่อมประโยค ดังนั้นสามารถลบคำเหล่านี้ออกจากข้อมูลได้ทันที โดยไม่กระทบต่อเนื้อหาในการประมวลผล โดยงานวิจัยชิ้นนี้ใช้การกำจัดคำหยุดด้วยไลบรารี stopwords ของภาษา R

2.3) การหารากศัพท์ (Word Stemming) คือการแปลงคำให้อยู่ในรูปแบบของรากศัพท์ การตัดโดยส่วนหน้า (Prefix) หรือส่วนท้าย (Suffix) ของคำออกให้เหลือแค่รากศัพท์

ของคำนั้นๆ ซึ่งคำศัพท์ที่มีรากศัพท์เดียวกันจะมีลักษณะที่คล้ายคลึงกัน เช่น crashed เมื่อตัด ed จะได้รากศัพท์คือ crash ในงานวิจัยชิ้นนี้ใช้ไลบรารี tm_map ของภาษา R ในการเรียกใช้ฟังค์ชัน stemDocument ในการแปลงคำศัพท์ให้อยู่ในรูปแบบรากศัพท์ หลังจากนั้นทำการแปลงตัวอักษรภาษาอังกฤษที่เป็นตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็กทั้งหมด ด้วยฟังค์ชัน tolower เพื่อให้สะดวกแก่การวิเคราะห์ข้อมูลที่อยู่ในรูปแบบข้อความ

เมื่อดำเนินการตามกระบวนการดังกล่าวเสร็จสิ้น จะทำการบันทึกข้อมูลอีกครั้งในรูปแบบไฟล์ CSV จากนั้นทำการโอนย้ายข้อมูลที่ได้เข้าสู่ฐานข้อมูล MySQL เพื่อนำไปประมวลผลในขั้นตอนต่อไป เมื่อตรวจสอบความถูกต้องของข้อมูลที่นำเข้า (Import) อีกครั้ง พบว่า ข้อมูลทั้งหมดที่ผ่านกระบวนเตรียมข้อมูลรูปแบบข้อความก่อนนำไปประมวลผลมีจำนวน 1,362 รายการเข้าสู่ฐานข้อมูลครบถ้วน

**3) การเปลี่ยนรูปข้อมูล (Transform)** ในงานวิจัยชิ้นนี้จะเปลี่ยนรูปแบบข้อมูลจากเดิมในลักษณะข้อมูลกึ่งมีโครงสร้าง ด้วยการสร้างคลังข้อมูลคำศัพท์ ร่วมกับการกำหนดรูปแบบอักขระ และการโปรแกรมมิ่งด้วยภาษา PHP ให้กลายเป็นข้อมูลที่มีโครงสร้าง เพื่อนำมาสร้างระบบค้นคืนสารสนเทศด้วยการนำเสนอภาพข้อมูลผ่านเว็บไซต์ โดยมีรายละเอียดดังนี้

3.1) การเปลี่ยนรูปแบบวันที่ให้เป็นมาตรฐานเดียวกันก่อนนำไปประมวลผล เนื่องจากรูปแบบการระบุวันที่ในรายงานอุบัติเหตุนั้นมีความแตกต่าง เช่น การใส่ชื่อเดือนแทนการใส่เลขเดือน การระบุปี พ.ศ.แทนที่จะเป็นปี ค.ศ. ดังนั้นจึงต้องเปลี่ยนรูปข้อมูลให้เป็นมาตรฐานเดียวกัน เพื่อให้ผลลัพธ์จากการสกัดข้อมูลมีความถูกต้องมากขึ้น

3.2) การสร้างคลังข้อมูลคำศัพท์ เมื่อพิจารณาในรายละเอียดข้อมูลรายงานอุบัติเหตุรายใหญ่ พบว่า มีจำนวนมากที่ยังมีความผิดพลาด เช่น รายชื่อจังหวัดที่สะกดผิด อันจะเห็นได้จาก “Khon Kean” และ “Khon Kaen” หมายถึงจังหวัดขอนแก่น รวมถึงชนิดของยานพาหนะในที่เกิดเหตุที่เรียกแตกต่างกันไปตามบริบทของผู้รายงาน ทั้ง “Bike” และ “Bicycle” หมายถึงรถจักรยานเช่นเดียวกัน จากที่กล่าวมาจะเห็นได้ว่าเป็นอุปสรรคต่อการนำไปใช้ประโยชน์ ดังนั้นผู้วิจัยทำการรวบรวมข้อมูลในประเด็นที่จะนำมาประมวลผลมาสร้างคลังข้อมูลคำศัพท์ จากการสังเคราะห์เอกสารที่เกี่ยวข้อง ได้แก่ รายชื่อจังหวัดในประเทศไทย (Province) ในรูปแบบภาษาอังกฤษ (สุพีร์พัฒน์ จองพานิช, รุ่ง สพสมัย และคณะ, 2550) ซึ่งสถานที่เกิดอุบัติเหตุ อันจะนำมาเป็นเงื่อนไขให้กับผู้ใช้ในการค้นคืนสารสนเทศ ร่วมกับประเภทของยานพาหนะในที่เกิดเหตุ (Motor Vehicle(s) involved in the accident) ตามการจำแนกประเภทของยานพาหนะที่เกิดอุบัติเหตุจราจรทางบกที่ได้รับแจ้งผ่านระบบ CRIMES จากรายงานการวิเคราะห์สถานการณ์อุบัติเหตุทางถนนของกระทรวงคมนาคม พ.ศ. 2561 จำนวน 10 ลำดับแรก (สำนักงานนโยบายและแผนการขนส่งและจราจร กลุ่มพัฒนาความปลอดภัย, 2562)

จากนั้นจึงนำรายชื่อจังหวัด มาสร้างเป็นตารางในฐานข้อมูล ประกอบด้วยชื่อจังหวัดที่ถูกต้องและไม่ถูกต้อง และตารางชนิดของยานพาหนะแต่ละประเภท จากนั้นกำหนดน้ำหนักของคำศัพท์ตามจำนวนตัวอักษร เพื่อใช้เป็นเงื่อนไขการเรียงลำดับในการประมวลผลจากจำนวนตัวอักษร

3.3) สร้างเงื่อนไขในการค้นหารายชื่อจังหวัดจากส่วนหัวของรายงานอุบัติเหตุแต่ละรายการ หากไม่พบข้อมูลจะไปค้นหาในส่วนของรายละเอียด โดยยึดตามความถี่ที่พบมากที่สุดในเขตข้อมูล และสร้างเงื่อนไขในการค้นหาประเภทของยานพาหนะในที่เกิดเหตุจากส่วนของรายละเอียดแบบวนลูป เริ่มจากคำศัพท์ที่มีความยาวมากที่สุดไปจนถึงน้อยที่สุด

3.4) การโปรแกรมมิ่งภาษา PHP เพื่อค้นหาและระบุตำแหน่งของข้อความตามหลักการการรู้จำชื่อเอนทิตี้ ตามที่ระบุไว้ในคลังข้อมูลคำศัพท์ที่สร้างขึ้น เพื่อสกัดข้อมูลรายชื่อจังหวัดและประเภทของยานพาหนะในที่เกิดเหตุ อันจะนำไปกำหนดเป็นเงื่อนไขในการค้นหาจำนวนของยานพาหนะจากรายงานอุบัติเหตุ

3.5) ค้นหาและระบุตำแหน่งของข้อความตามหลักการการรู้จำชื่อเอนทิตี้อีกครั้ง เพื่อหาวันที่เกิดเหตุ จำนวนยานพาหนะในที่เกิดเหตุแต่ละชนิด จำนวนผู้บาดเจ็บ และจำนวนผู้เสียชีวิต จากนั้นจะกรองข้อมูลอีกครั้ง เพื่อให้ได้เฉพาะข้อมูลที่เป็นตัวเลข สำหรับนำไปคำนวณหาผลรวม

ผลลัพธ์ของการสกัดข้อมูลรายงานอุบัติเหตุทางถนนรายใหญ่ ด้วยการสร้างคลังข้อมูลคำศัพท์อุบัติเหตุทางถนน ร่วมกับการกำหนดรูปแบบอักขระ และการโปรแกรมมิ่งด้วยภาษา PHP ใน 6 ประเด็น ได้แก่ จังหวัด วันที่ จำนวนผู้เสียชีวิต จำนวนผู้บาดเจ็บ ประเภทและจำนวนของยานพาหนะที่เกิดเหตุ แล้วจะนำมาสรุปเป็นภาพรวมการเกิดอุบัติเหตุในแต่ละครั้ง ดังภาพที่ 6 ก่อนจะนำเสนอเป็นภาพข้อมูลให้ผู้ใช้สร้างเงื่อนไขในการค้นคืนสารสนเทศผ่านทางเว็บไซต์ในขั้นตอนต่อไป

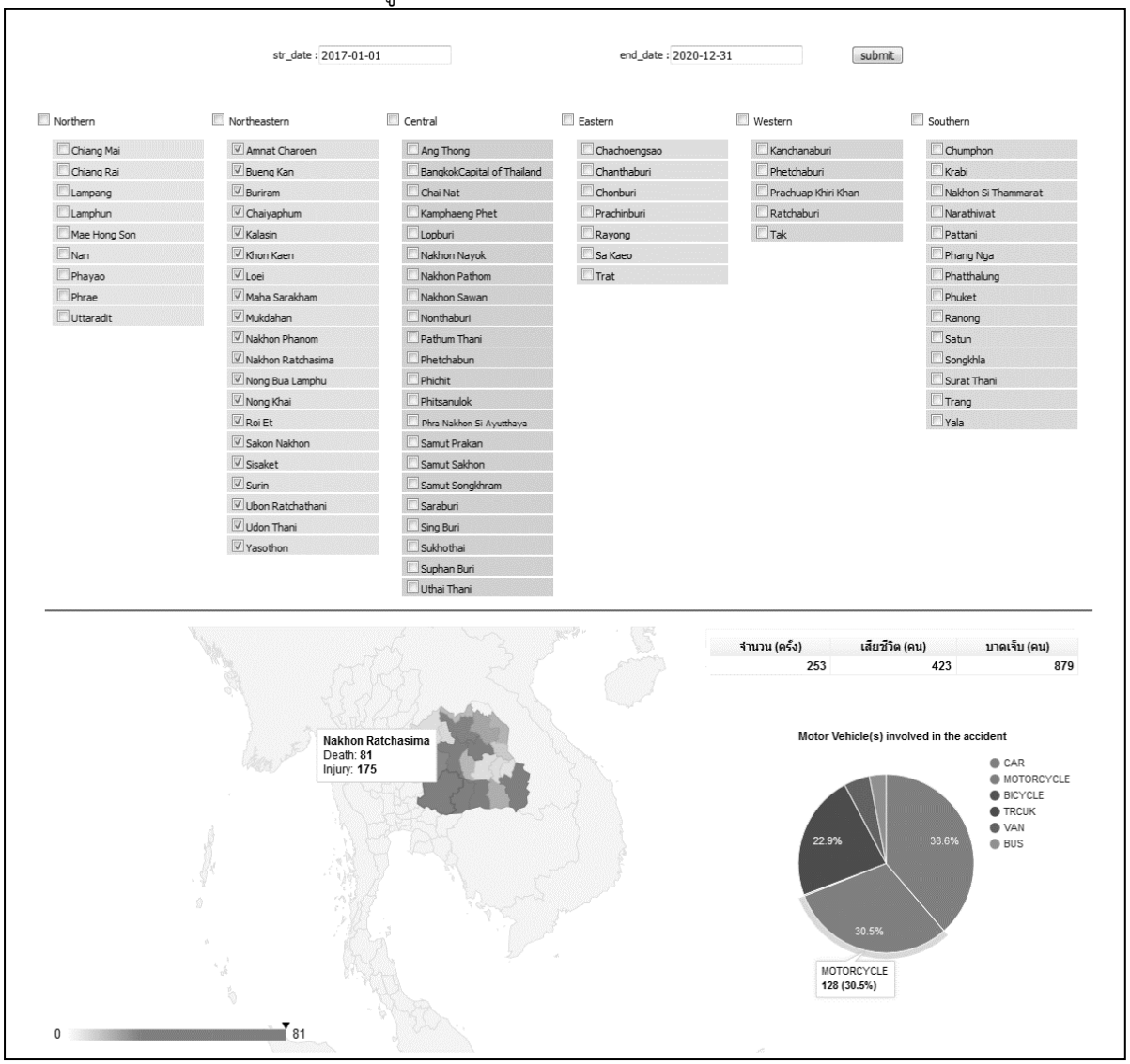
**proceed time :** 0.4369 วินาที
**memory usage :** 10.94 mb

| Code | #ContentPlace Holder1_lb_title | #section .clearfix | Extract Data |
|---|---|---|---|
| 003001 | Accident Sedan Car and Sedan Car Crossed the Traffic Lane Collided with Van, 4 Deaths, 7 Injuries, Nonsung District, Nakhon Ratchasima (06-03-2020, 05.10 p.m.) | Accident: Sedan Car and Sedan Car Crossed the Traffic Lane Collided with Van 4 Deaths 7 Injuries Nonsung District Nakhon Ratchasima (06032020 05.10 p.m.)Accident Notification CentreRoad Accident Victims Protection Co. Ltd. Department of Disaster Prevention and MitigationAccident ReportNotification No: 63/001/0003807 \| Source: Safety Radio Center. Date of Claim: Friday March 6th 2020 at about 07.02 p.m. Date of Accident: Friday March 6th 2020 at about 05.10 p.m.Accident Point: Mitrphap Road Than Prasat Nonsung District Nakhon Ratchasima. Accident Condition: 2lane traffic straightway Weather Condition: rain. Responsibility area: Nonsung Police Station Tel. 044379291Enquiry officer/Case owners: Tel. Accident scene: Accident: sedan car and sedan car crossed the traffic lane collided with van.Photo by: Motor Vehicle(s) involved in the accident1. Van white color license plate registration no. xxxxxx Bangkok.2. Sedan car Mitsubishi color license plate registration no. xxxxxxx Bangkok.3. Sedan car white color license plate registration no. xxxxxx Nakhon Ratchasima.List of deaths: 4 Deaths List of injured: 7 Injuries; taken to Pi Mai Hospital and Nonsung Hospital. Injuries Identification: Cause of the accident: (The real causes of the accident is being investigated.)Insurance Assistance and coordination1. License plate registration no. xxxxxx Bangkok insured under compulsory insurance Viriyah Insurance Public Company Limited policy No. 623020061110 Start Date: 20/10/2019 Expire Date: 20/10/2020. Voluntary insurance is NOT FOUND 2. License plate registration no. xxxxxxx Bangkok compulsory insurance is NOT FOUND 3. License plate registration no. xxxxxx Nakhon Ratchasima compulsory insurance is NOT FOUNDHotline(s) Nonsung Police Station (Tel. 044379291) Viriyah Insurance Public Company Limited (Tel. 1557) Mr.Precha Hook 31 staff Talad Kae point. (Tel. 0833726182) Department of Disaster Prevention and Mitigation (Call Center 1784) Road Accident Victims Protection Company Limited IOC and the Office of Insurance Commission (OIC)Road Accident Victims Protection Co. Ltd.44/1 Rungrojthanakul Building 11th Floor Ratchadapisek Rd. Huaykwang Bangkok 10310Call Center 1791 Tel: 021009191 Fax: 0264302934www.rvp.co.th www.thairsc.com | Province: Nakhon Ratchasima<br>Date: 2020-03-06<br>Motor Vehicle(s) : [2] CAR [1] VAN<br>Death(s) 4<br>Injured(s) 7 |

***ภาพที่ 6*** ผลลัพธ์จากการสกัดข้อมูลที่พัฒนาขึ้น

**การแสดงผลและค้นคืนสารสนเทศด้วยการนำเสนอภาพข้อมูล** คือกระบวนการนำข้อมูลดิบมานำเสนอในรูปแบบ เช่น กราฟเส้น แผนภูมิแท่ง หรือแผนภูมิวงกลม (Wike, 2019) โดยมีวัตถุประสงค์เพื่อถ่ายทอดข้อมูลเชิงปริมาณที่มีความซับซ้อนให้เกิดความน่าสนใจ เข้าใจง่าย เห็นภาพรวมได้ชัดเจน นิยมนำมาใช้ประกอบการรายงาน เพื่อวิเคราะห์ และสรุปผล ปัจจุบันมีหลากหลายเครื่องมือที่สามารถนำมาใช้ในการแสดงผลด้วยการนำเสนอภาพข้อมูล ในงานวิจัยชิ้นนี้ใช้เทคนิคการนำเสนอภาพข้อมูลด้วย Google Charts ในการสร้างนำเสนอภาพข้อมูล โดยจะส่งข้อมูลไปประมวลผลที่เครื่องแม่ข่าย และรับข้อมูลมาแสดงผลบนเว็บไซต์ผ่านส่วนต่อประสานโปรแกรมประยุกต์ (Application Program Interface : API) ของผู้ให้บริการ

การนำเสนอภาพข้อมูลในงานวิจัยชิ้นนี้จะเน้นการแสดงผลแบบจัดกลุ่ม (Classification) ด้วยแผนภูมิวงกลม (Pie Chart) เพื่อแสดงผลสถิติยานพาหนะที่เกิดอุบัติเหตุ ร่วมกับแผนที่ประเทศไทย (Geographical) เพื่อแสดงผลเปรียบเทียบสถิติการเกิดอุบัติเหตุ จำนวนผู้บาดเจ็บ และจำนวนผู้เสียชีวิตในภาพรวม (Chen, Härdle & Unwin (2008)) ดังภาพที่ 7




**ภาพที่ 7** การแสดงผล และค้นคืนสารสนเทศด้วยการนำเสนอภาพข้อมูล

จากภาพที่ 7 แสดงการค้นคืนสารสนเทศผ่านเงื่อนไขการค้นหาระหว่างวันเริ่มต้นและวันสิ้นสุดร่วมกับรายชื่อจังหวัดของประเทศไทย แล้วนำมารายงานสถิติการเกิดอุบัติเหตุในรูปแบบการนำเสนอภาพข้อมูล โดยการนำเสนอรูปแบบแผนที่เปรียบเทียบสถิติภาพรวมรายจังหวัดของประเทศไทย การแสดงผลจะใช้หลักสีตามลำดับ (Sequential Color) (Wilke, 2019) ดังนั้นในงานวิจัยชิ้นนี้จึงเลือกใช้สีขาว เหลือง ส้ม และแดงในการแสดงผลไล่ลำดับความเข้มของสีจากน้อยที่สุดไปยังมากที่สุดตามปริมาณการเกิดอุบัติเหตุ

**ผลการวิจัย**

การประเมินความสามารถในการค้นคืนสารสนเทศของระบบที่พัฒนาขึ้น โดยทำการวัดความสามารถในการค้นคืนสารสนเทศของระบบแบบประเมินประสิทธิภาพเอกสารที่ถูกเลือก แต่มิได้เรียงลำดับความคล้ายคลึง (Manning et al., 2008) ประกอบด้วย 2 ค่า คือ ค่าความแม่นยำ (Precision) ที่บอกถึงประสิทธิภาพของระบบในการค้นคืนเอกสารโดยดูจากอัตราส่วนของจำนวนเอกสารที่ถูกต้องจากเอกสารที่ถูกเลือกมาทั้งหมด ดังสมการที่ 1 ส่วนค่าความครบถ้วย (Recall) หมายถึง ประสิทธิภาพของการค้นคืนเอกสารโดยดูจากอัตราส่วนจำนวนเอกสารที่ถูกต้องที่เลือกมาต่อจำนวนเอกสารที่ถูกต้องทั้งหมดที่อยู่ในคอลเล็คชั่น ดังสมการที่ 2

$$\text{Precision} = \frac{X}{Y} \quad (1)$$

$$\text{Recall} = \frac{X}{Z} \quad (2)$$

โดยที่ X คือ จำนวนเอกสารที่ถูกต้องที่ถูกดึงมาเป็นผลลัพธ์

Y คือ จำนวนเอกสารทั้งหมดที่ถูกดึงมาเป็นผลลัพธ์

Z คือ จำนวนเอกสารที่ถูกต้องทั้งหมดที่อยู่ในคอลเล็คชั่น

อย่างไรก็ดีค่าความถูกต้องและค่าความครบถ้วนเป็นประโยชน์ต่อผู้ใช้ต่างกลุ่มกันไป โดยผู้ใช้จะให้ความสำคัญกับค่าความถูกต้องมากกว่าค่าความครบถ้วน เนื่องจากผู้ใช้ส่วนใหญ่ต้องการให้ข้อมูลที่เป็นผลลัพธ์ที่ตรงกับสิ่งที่ต้องการค้นหา แต่ในทางกลับกัน ผู้ที่พัฒนาระบบค้นหาข้อมูลจะเน้นไปที่ค่าความครบถ้วน เนื่องจากต้องการให้ระบบค้นหาและเลือกเอกสารที่ถูกต้องจากคอลเล็คชั่นให้ได้มากที่สุด ดังนั้นอีกค่าหนึ่งที่นำมาใช้ในการรวมค่าความถูกต้องและค่าความครบถ้วนเข้าด้วยกันโดยหาค่าเฉลี่ยที่ดีที่สุด (Weighted Harmonic Mean) เรียกว่า ค่าประสิทธิภาพโดยรวม (F-measure) หรือ F1 score ดังสมการที่ 3

$$\text{F-measure} = \frac{2 \text{ x } Precision \text{ x } Recall}{Precision + Recall} \quad (3)$$

ทั้ง 3 ค่าเป็นที่รู้จักและนิยมใช้มากที่สุดในการประเมินระบบค้นคืนสารสนเทศในปัจจุบัน ดังนั้นเพื่อให้ผลลัพธ์จากการค้นคืนสารสนเทศมีความถูกต้องมากยิ่งขึ้น ผู้วิจัยให้ความสำคัญกับการพิจารณาความคลาดเคลื่อน (Error) คือ ผลต่างระหว่างค่าที่วัดได้กับค่าที่แท้จริง โดยทั่วไปแสดงเป็นเปอร์เซ็นต์ ถ้าค่าที่วัดได้ใกล้เคียงกับค่าจริงมากแสดงว่าการวัดนั้นมีความแม่นยำหรือความถูกต้อง (Accuracy) สูง (นวภัทรา หนูนาค และทวีพล ซื่อสัตย์ (2555)) โดยการวัดทุกครั้งมักมีค่าความคลาดเคลื่อนเกิดขึ้นเสมอ แบ่งออกเป็น 3 ชนิด ได้แก่ 1) ความคลาดเคลื่อนที่เกิดจากผู้วัด (Human Error) อันเกิดจากความประมาทเลินเล่อในการวัด 2) ความคลาดเคลื่อนเชิงระบบ (Systematic Error) มักเกิดจากเครื่องมือวัดไม่ได้ประสิทธิภาพ และ 3) ความคลาดเคลื่อนเชิงสถิติ (Statistical Error) ซึ่งความคลาดเคลื่อนประเภทที่ 1 และ 2 สามารถจัดการได้ด้วยความรอบคอบ

ระมัดระวัง และการพัฒนาเครื่องมือวัดให้มีประสิทธิภาพ แต่สำหรับความคลาดเคลื่อนประเภทที่ 3 นั้น ไม่มีทางกำจัดให้หมดไปได้ เนื่องจากอยู่นอกเหนือการควบคุม จึงมักใช้การวัดซ้ำหลายๆ ครั้ง เนื่องจากจำนวนรอบมากขึ้นเท่าไร ความคลาดเคลื่อนก็จะลดน้อยลงเท่านั้น จนถึงในระดับที่ยอมรับได้

การวิจัยครั้งนี้แบ่งการวัดผลออกเป็น 2 ส่วน คือ ประเมินความสามารถในการค้นคืนสารสนเทศจากการสกัดข้อมูลด้วยเงื่อนไขจังหวัดและวันที่ โดยผู้วิจัยร่วมกับผู้เชี่ยวชาญด้านเทคโนโลยีสารสนเทศรวม 5 ท่าน ตามแนวคิดของ Nielsen (2000) ที่เสนอจำนวนผู้ประเมินผลการทดสอบความสามารถในการใช้งานไว้ว่า 5 คน จะพบปัญหาในระบบได้ 85% ทำการเลือกเงื่อนไขทีละครั้ง จำนวน 77 ครั้งตามจำนวนจังหวัดของประเทศไทย แล้วประเมินความถูกต้องของสารสนเทศแต่ละรายการว่าถูกต้องหรือไม่ แบ่งเป็นถูกต้องได้ 1 คะแนน หากไม่ถูกต้องได้ 0 คะแนน หากมีคะแนนมากกว่า 2 ใน 3 จึงจะถือว่าสารสนเทศรายการนั้นถูกต้อง เมื่อครบรอบจะทำการประเมินอีกครั้งจนครบ 3 รอบ แบบผลลัพธ์เป็นอิสระต่อกัน เพื่อนำผลการประเมินมาหาค่าเฉลี่ยจนค่าความคลาดเคลื่อนอยู่ในระดับที่ยอมรับได้ ผลการประเมินดังตาราง 1

เมื่อนำมาแปลผล 5 ระดับเพื่อให้ง่ายต่อการทำความเข้าใจตามหลักการแบ่งอันตรภาคชั้น (Class Interval) กำหนดคะแนนที่สูงที่สุด คือ 100 คะแนน และคะแนนที่ต่ำสุด คือ 0 คะแนน โดยการหากึ่งกลางพิสัย (บุญชม ศรีสะอาด, 2556) พบว่าแต่ละช่วงชั้นจะมีระยะห่างเท่ากันที่ร้อยละ 20 กำหนดระดับการแปลผลคุณภาพดังนี้

ต่ำกว่า 0.19 = ควรปรับปรุง
0.20 – 0.39 = พอใช้
0.40 – 0.59 = ปานกลาง
0.60 – 0.79 = ดี
0.80 ขึ้นไป = ดีมาก

จากตารางที่ 1 พบว่า ระบบที่พัฒนาขึ้นสามารถค้นหาเอกสารรายงานอุบัติเหตุได้ตรงตามเงื่อนไขการออกรายงาน โดยมีค่าความถูกต้องเท่ากับ 0.87 ค่าความครบถ้วน เท่ากับ 0.85 และเมื่อพิจารณาค่าประสิทธิภาพโดยรวมเท่ากับ 0.86 ซึ่งทั้ง 3 ค่ามีผลการประเมินอยู่ในระดับดีมาก

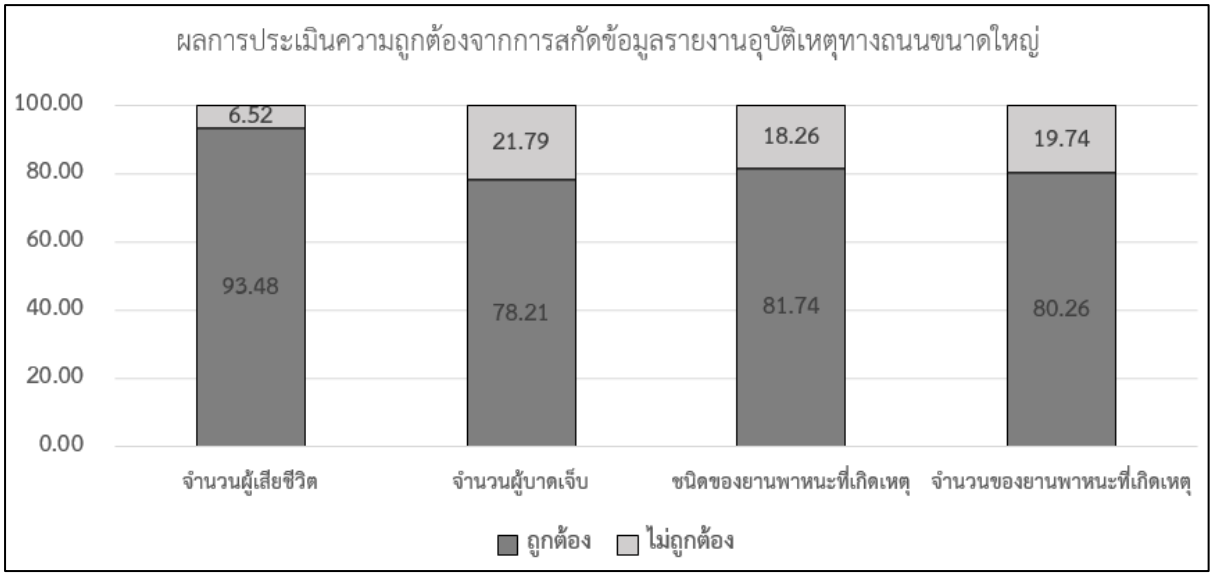
**ตารางที่ 1** ผลการวัดความถูกต้องและความสามารถของระบบ

| จำนวนเอกสาร | ผลการวัดความถูกต้องและความสามารถของระบบ | | | | | |
|---|---|---|---|---|---|---|
| | Precision | | Recall | | F-Measure | |
| 1,362 | 0.87 | ดีมาก | 0.85 | ดีมาก | 0.86 | ดีมาก |

จากนั้นทำการประเมินผลความถูกต้องในการสกัดข้อมูลอีกครั้งใน 4 ประเด็นคือ 1) จำนวนผู้เสียชีวิต 2) จำนวนผู้บาดเจ็บ 3) ประเภทของยานพาหนะที่เกิดเหตุ และ 4) จำนวนของยานพาหนะที่เกิดเหตุ จากการกำหนดขนาดของกลุ่มตัวอย่างโดย Nielsen (2000) ได้นำเสนอจำนวนผู้ประเมินผลการทดสอบความสามารถในการใช้งานไว้ว่า 5 คน จะพบปัญหาในระบบได้ 85% และถ้าผู้ทดสอบ 15 คนหรือมากกว่าจะพบปัญหาทั้งหมดภายในระบบ ในงานวิจัยชิ้นนี้ผู้วิจัยเลือกใช้วิธีการสุ่มตัวอย่างโดยไม่ใช้ความน่าจะเป็น (Non-probability Sampling) ในการเก็บรวบรวมข้อมูล เนื่องจากสามารถเก็บข้อมูลจากกลุ่มตัวอย่างจำนวนมากได้ในระยะเวลาจำกัด จากนั้นใช้วิธีการเลือกกลุ่มตัวอย่างแบบเฉพาะเจาะจง (Purposive Sampling) คือการเลือกแจกแบบสอบถามเฉพาะกลุ่มตัวอย่างที่มีประสบการณ์ในการใช้งานคอมพิวเตอร์และสมาร์ตโฟน 5 ปีขึ้นไป จำนวน 30 คน ที่ความเชื่อมั่น 100% ควบคุมให้ทดลองใช้ระบบค้นคืน

สารสนเทศผ่านห้องปฏิบัติการคอมพิวเตอร์ที่เชื่อมต่ออินเทอร์เน็ต และควบคุมการแสดงผลบน Mozilla Firefox เป็นเว็บบราวเซอร์ เริ่มจากให้ผู้ใช้กำหนดเงื่อนไขในการออกรายงาน ได้แก่ วันเริ่มต้นและวันสิ้นสุด และจังหวัดที่เกิดเหตุได้อย่างอิสระ จากนั้นจะนำไปสกัดข้อมูลตามกระบวนการที่ได้ออกแบบและพัฒนาขึ้น และแสดงผลลัพธ์ในรูปแบบการนำเสนอภาพข้อมูลผ่านทางเว็บไซต์

การวัดค่าความถูกต้องของการสกัดข้อมูลจากจำนวนเอกสารที่ถูกดึงออกมาตามเงื่อนไขที่กำหนดทีละรายการ โดยผู้ใช้สามารถค้นคืนสารสนเทศได้ไม่จำกัดจำนวนครั้ง ซึ่งผู้ใช้มองเห็นได้ชัดเจนว่าเป็นผลลัพธ์ที่ถูกต้องหรือไม่ถูกต้องใน 4 ประเด็นคือ 1) จำนวนผู้เสียชีวิต 2) จำนวนผู้บาดเจ็บ 3) ประเภทของยานพาหนะที่เกิดเหตุ และ 4) จำนวนของยานพาหนะที่เกิดเหตุ หากถูกต้องทั้งหมดคิดเป็น 1 คะแนน ในทางตรงกันข้าม หากมีผลลัพธ์ในประเด็นใดไม่ถูกต้อง ไม่ปรากฏ หรือในกรณีใดที่คำตอบไม่ชัดเจน ให้ถือว่าคำตอบเหล่านั้นเป็นคำตอบที่ไม่ถูกต้อง คิดเป็น 0 คะแนน กำหนดให้ผู้ใช้แต่ละคนทำการประเมินผล 3 ครั้ง แบบผลลัพธ์เป็นอิสระต่อกัน จากนั้นจะนำผลการประเมินของผู้ใช้แต่ละคนมารวมกันเพื่อหาค่าเฉลี่ย ผลลัพธ์จากการประเมินดังภาพที่ 8




**ภาพที่ 8** ผลการประเมินความถูกต้องจากการสกัดข้อมูลรายงานอุบัติเหตุรายใหญ่

จากภาพที่ 8 พบว่า ผลการประเมินค่าความถูกต้องจากการสกัดข้อมูลรายงานอุบัติเหตุรายใหญ่บนเว็บไซต์ใน 4 ประเด็น คือ คือ 1) จำนวนผู้เสียชีวิต 2) จำนวนผู้บาดเจ็บ 3) ประเภทของยานพาหนะ และ 4) จำนวนของยานพาหนะในที่เกิดเหตุ มีค่าความถูกต้องจากการประเมินของผู้ใช้คิดเป็นร้อยละ 93.48, 78.21, 81.74 และ 80.26 ตามลำดับ เป็นที่น่าสังเกตว่าผลการสกัดข้อมูลจำนวนผู้เสียชีวิต และผู้บาดเจ็บ ซึ่งเป็นข้อมูลเชิงปริมาณ และใช้กระบวนการสกัดข้อมูลด้วยหลักการการรู้จำชื่อเช่นเดียวกัน แต่กลับให้ผลลัพธ์ความถูกต้องที่แตกต่างกัน สวนทางกับผลลัพธ์จากการสกัดข้อมูลประเภทยานพาหนะที่ใช้กระบวนการสกัดข้อมูลด้วยการสร้างคลังคำศัพท์ ร่วมกับการใช้หลักการการรู้จำชื่อในการสกัดจำนวนของยานพาหนะในที่เกิดเหตุที่ให้ค่าความถูกต้องจากผลการประเมินใกล้เคียงกัน

ดังนั้นเมื่อนำผลการประเมินดังกล่าวมาวิเคราะห์ในรายละเอียดร่วมกับข้อคิดเห็นจากผู้ใช้และการสังเกตพฤติกรรมผู้ใช้ พบว่า ปัจจัยที่ทำให้ผลลัพธ์ของจำนวนผู้เสียชีวิตมีค่าความถูกต้องมากกว่าผลลัพธ์ของจำนวนผู้บาดเจ็บนั้น มีสาเหตุ

มาจากการกระบวนการสกัดข้อมูลที่มีเงื่อนไขในการทำงานตามลำดับก่อนหลัง เป็นที่สังเกตว่าหากรายงานอุบัติเหตุรายใหญ่รายการใดที่มีการระบุจำนวนผู้เสียชีวิตก่อนจำนวนผู้บาดเจ็บ จะทำให้ผลลัพธ์การสกัดข้อมูลถูกต้อง แม้ในรายการดังกล่าวไม่ระบุจำนวนผู้บาดเจ็บ แต่ในทางกลับกันหากรายการใดที่ไม่มีการระบุจำนวนผู้เสียชีวิต แต่ระบุจำนวนผู้บาดเจ็บ จะส่งผลทำให้ผลลัพธ์จากการสกัดข้อมูลผิดพลาดไปด้วย งานวิจัยในครั้งจึงต้องนำประเด็นปัญหาที่เกิดขึ้นมาปรับปรุงกระบวนการ เพื่อให้ผลลัพธ์มีความถูกต้องยิ่งขึ้น จึงได้ทำการปรับเปลี่ยนเงื่อนไขในการสกัดข้อมูลตามลำดับก่อนหลัง และทดสอบกับผู้ใช้กลุ่มเดิมในสภาพแวดล้อมเดียวกัน พบว่า ส่งผลให้ค่าความแม่นยำเพิ่มขึ้นจากเดิม 4.8% ค่าความครบถ้วนยังคงเดิม และค่าความครบถ้วนเปลี่ยนแปลงไปเล็กน้อย กล่าวได้ว่าเครื่องมือที่พัฒนาขึ้นยังคงมีค่าความเชื่อมั่นมากกว่า 0.70 (บุญชม ศรีสะอาด, 2556) ซึ่งอยู่ในระดับที่ยอมรับได้

**การอภิปรายผล**

ผลลัพธ์จากการสกัดข้อมูลชนิดและจำนวนของยานพาหนะในที่เกิดเหตุ จากการสร้างคลังคำศัพท์ร่วมกับการใช้หลักการการรู้จำชื่อ โดยผลการประเมินอยู่ในระดับที่น่าพอใจ และความถูกต้องของผลการประเมินนั้นมีความสัมพันธ์แบบแปรผันตรงกัน แสดงให้เห็นว่าการสกัดข้อมูลด้วยการสร้างคลังคำศัพท์นั้นจะมีประสิทธิภาพมากขึ้นเมื่อทำงานร่วมกับการรู้จำเอนทิตี้ แต่ปัญหาที่พบคือชื่อจังหวัดที่เขียนไม่ถูกต้องตาม และชื่อยานพาหนะที่ไม่ปรากฏในฐานข้อมูล เนื่องจากเป็นแบบไม่เป็นทางการหรือชื่อเรียกเฉพาะถิ่น เช่น E-tan หรือ Tuktuk ตลอดจนการให้ข้อมูลที่ไม่เป็นมาตรฐานทั้งจำนวนผู้บาดเจ็บ จำนวนผู้เสียชีวิต และจำนวนยานพาหนะในที่เกิดเหตุ ซึ่งทั้งหมดนี้เป็นองค์ประกอบสำคัญที่ทำให้ระบบสามารถสกัดความรู้ได้ถูกต้อง หากข้อมูลเหล่านี้มีไม่เพียงพอจะส่งผลทำให้ระบบไม่สามารถสกัดข้อมูลที่กำหนดได้ จึงทำให้การคำนวณค่ามาตรฐานในการประเมินต่ำกว่าที่ผู้ใช้กำหนด ดังนั้นจึงต้องนำคำศัพท์ดังกล่าวมาปรับปรุงคลังคำศัพท์ เพื่อให้ครอบคลุมและสอดคล้องกับบริบทการใช้งานมากขึ้น

**สรุปผลการวิจัย**

งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนาระบบสกัดข้อมูลกึ่งมีโครงสร้างรายงานอุบัติเหตุทางถนนรายใหญ่บนเว็บไซต์ตามกระบวนการอีแอลที และค้นคืนสารสนเทศด้วยการนำเสนอภาพข้อมูลปฏิสัมพันธ์กับผู้ใช้ด้วยการสร้างคลังข้อมูลคำศัพท์ร่วมกับการใช้หลักการการรู้จำชื่อ ด้วยการกำหนดรูปแบบอักขระ และโปรแกรมมิ่งด้วยภาษา PHP มาสกัดข้อมูลใน 6 ประเด็น ได้แก่ จังหวัด วันที่ จำนวนผู้เสียชีวิต จำนวนผู้บาดเจ็บ ประเภทและจำนวนของยานพาหนะที่เกิดเหตุ เพื่อสร้างระบบค้นคืนสารสนเทศแก่ผู้ใช้ด้วยการนำเสนอข้อมูลภาพผ่านเว็บไซต์

การประเมินความถูกต้องและความสามารถของระบบที่พัฒนาขึ้นในประเด็นจากการสกัดข้อมูลคือจังหวัด และวันที่ วัดผลโดยผู้เชี่ยวชาญ 5 คน ที่ความเชื่อมั่น 85% วัดผลทีละครั้ง จำนวน 77 ครั้ง ตามจำนวนจังหวัดในประเทศไทย ทำซ้ำเช่นเดิม 3 รอบแล้วนำมาหาค่าเฉลี่ยเพื่อป้องกันความคลาดเคลื่อน พบว่า ค่าความถูกต้อง ค่าความครบถ้วน และค่าประสิทธิภาพโดยรวม เท่ากับ 0.87, 0.85 และ 0.86 ตามลำดับ อยู่ในระดับดีมาก จากนั้นนำมาประเมินความถูกต้องจากการสกัดข้อมูลโดยผู้ใช้ใน 4 ประเด็นที่เหลือ คือ 1) จำนวนผู้เสียชีวิต 2) จำนวนผู้บาดเจ็บ 3) ชนิดของยานพาหนะในที่เกิดเหตุ และ 4) จำนวนของยานพาหนะในที่เกิดเหตุ เลือกใช้วิธีการสุ่มตัวอย่างโดยไม่ใช้ความน่าจะเป็น จำนวน 30 คน จากการเลือกกลุ่มตัวอย่างแบบเฉพาะเจาะจง ที่ความเชื่อมั่น 100% พบว่า ผลการประเมินจากการทำซ้ำ 3 ครั้ง คิดเป็นร้อยละ 93.48, 78.21, 81.74 และ 80.26 ตามลำดับ เมื่อวิเคราะห์ลงไปในรายละเอียด พบว่า กระบวนการ

สกัดข้อมูลที่มีเงื่อนไขในการทำงานตามลำดับก่อนหลัง ส่งผลต่อค่าความถูกต้องของจำนวนผู้บาดเจ็บและเสียชีวิต เมื่อเปลี่ยนลำดับเงื่อนไขการทำงาน พบว่า ค่าความแม่นยำสูงขึ้นเล็กน้อย และการสร้างคลังคำศัพท์ชนิดของยานพาหนะที่เกี่ยวข้องร่วมกับการใช้หลักการรู้จำชื่อส่งผลต่อความถูกต้องของประเภทและจำนวนของยานพาหนะในที่เกิดเหตุแบบแปรผันตรงต่อกัน

**ข้อเสนอแนะ**

ผลจากการสกัดข้อมูลในงานวิจัยชิ้นนี้ ไม่สามารถตรวจสอบความถูกต้องด้วยการเปรียบเทียบกับสถิติการเกิดอุบัติเหตุจากแหล่งอ้างอิงใดๆได้ เนื่องด้วยข้อจำกัดของข้อมูลตั้งต้นที่ยังมีจำนวนไม่สอดคล้องกับความเป็นจริง แต่งานวิจัยชิ้นนี้มุ่งนำเสนออีกแนวทางในการออกแบบและพัฒนาระบบสกัดข้อมูลประเภทกึ่งมีโครงสร้างสามารถนำไปใช้กับข้อมูลลักษณะเดียวกันในอนาคต อย่างไรก็ตาม หากระบบมีข้อมูลที่จำเป็นในการสกัดความรู้ทั้งหมดแล้ว สิ่งที่ควรพัฒนาถัดไปคือ การพัฒนาระบบให้มีการปรับปรุงคลังคำศัพท์อัตโนมัติเมื่อพบคำศัพท์ใหม่ๆที่เกี่ยวข้อง และการพัฒนาขั้นตอนวิธีที่ให้มีความถูกต้องมากขึ้น โดยการนำข้อมูลอุบัติเหตุเชิงคุณภาพจากหลายแหล่งมาวิเคราะห์และทดสอบตลอดจนพัฒนาระบบให้สามารถสกัดข้อมูลรายงานอุบัติเหตุที่เป็นภาษาไทยได้ ซึ่งจะดำเนินการต่อไปในอนาคต

**เอกสารอ้างอิง**


กานดา รุณนะพงศา และปโยธร อุราธรรมกุล. (2548). *การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่*. คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น.

นวภัทรา หนูนาค และทวีพล ซื่อสัตย์. (2555). *การวัดและเครื่องมือวัด ประยุกต์ใช้ในอุตสาหกรรมอาหาร, พิมพ์ครั้งที่ 1*. กรุงเทพมหานคร: คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บุญชม ศรีสะอาด. (2556). การวิจัยเบื้องต้นฉบับปรับปรุงใหม่พิมพ์ครั้งที่ 9. กรุงเทพฯ: สำนักพิมพ์สุวีริยาสาส์.

ศูนย์อำนวยการความปลอดภัยทางถนน. (2560). *นิยามข้อมูลอุบัติเหตุ*. กรุงเทพฯ: กรมป้องกันและบรรเทาสาธารณภัย กระทรวงมหาดไทย.

สถิตย์โชค โพธิ์สอาด และปิติภูมิ โพสาวัง. (2062). การจำแนกพฤติกรรมการขับขี่รถโดยสารสาธารณะโดยใช้วิธีการสกัดข้อความ และเทคนิคการเรียนรู้ของเครื่อง. *วารสารเทคโนโลยีสารสนเทศ*, *15*(1). 71-80.

สำนักงานนโยบายและแผนการขนส่งและจราจรกลุ่มพัฒนาความปลอดภัย. (2562). *รายงานการวิเคราะห์สถานการณ์อุบัติเหตุทางถนน ของกระทรวงคมนาคม พ.ศ. 2561*. กรุงเทพฯ สำนักแผนความปลอดภัย กระทรวงคมนาคม.

สำนักโรคไม่ติดต่อ กรมควบคุมโรค. (2015). *ข้อมูล และสถิติคนตายหายไปไหน*. [แผ่นพับ]. นนทบุรี:ผู้แต่ง.

สุพีร์พัฒน์ จองพานิช, รุ่ง สพสมัย และคณะ. (2550). *ชื่อจังหวัด อำเภอ/กิ่งอำเภอ ตำบล เขต และแขวงไทย-อังกฤษ ฉบับรับรองโดยราชบัณฑิตยสถาน*. กรุงเทพฯ : กรมการปกครอง กองวิชาการและแผนงาน, 2550.

สุรจิต ภูมิคง, ธรา อั่งสกุล, และจิติมนต์ อั่งสกุล. (2013). *วิธีการสกัดข่าวด้านความปลอดภัย*. วารสารเทคโนโลยีสุรนารี. *7*(2). 79-97.

Chen, C.H., Hardle, W. & Unwin, A. (2008). *Handbook of Data Visualization*. Springer.

Fry, B. (2007). *Visualizing Data*. O'Reilly Media, Inc.

Garcia-Molina, H., Ullman, J.D., & Widom, J. (*2009*). *Database systems: the complete book, 2e*. Prentice Hall.

Golffarelli, M., & Rizzi, S., (2009). *Data Warehouse Design: Modern Principles and Methodologies*. Mc Graw Hill.

Khalil, S. & Fakir, M. (2017). *RCrowler : An R package for parallel web crawling and scraping*. SoftwareX6(2017).

Kimball R., & Caserta J. (2004). *The Data Warehouse ETL Toolkit*. Wiley.

Manning, C.D., Raghavan, P., Schütze, H., (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mitchell, R. (2015). *Web Scraping with Python, 2nd Edition*. O'Reilly Media, Inc.

Ponniah, P., (2009). *Data warehousing fundamentals for it Professionals*. Wiley Publishing.

Rainardi, V., (2008). *Building a Data Warehouse With Examples in SQL ServerSQL Server*. Apress.

Wilke, C.O. (2019). *Fundamentals of Data Visualization A Primer on Making Informative and Compelling Figures*. O'Reilly Media, Inc.

World Health Organization. (2018). *Global status report on road safety 2018*. Geneva: World Health Organization.