

การใช้วิธีการเปรียบเทียบสายอักขระแบบประมาณในการตรวจสอบ ชื่อผลิตภัณฑ์เครื่องสำอางในแบบคำขอจดแจ้งเครื่องสำอาง

จิราภรณ์ วงษ์เลิศ¹, วีรยุทธ์ เลิศนที²

¹กลุ่มงานคุ้มครองผู้บริโภคและเภสัชสาธารณสุข สำนักงานสาธารณสุขจังหวัดพระนครศรีอยุธยา
²สาขาวิชาชีวการแพทย์และสารสนเทศศาสตร์ทางสุขภาพ คณะเภสัชศาสตร์ มหาวิทยาลัยศิลปากร

บทคัดย่อ

วัตถุประสงค์: เพื่อพัฒนาโปรแกรมสำหรับการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในแบบคำขอจดแจ้งเครื่องสำอางด้วยวิธีการเปรียบเทียบสายอักขระแบบประมาณโดยการหาระยะทางเลเวนชเตย์น และประเมินประสิทธิภาพของโปรแกรมโดยเปรียบเทียบผลการพิจารณาชื่อผลิตภัณฑ์เครื่องสำอางกับผลการพิจารณาโดยเจ้าหน้าที่ผู้เชี่ยวชาญ **วิธีการ:** กรอบแนวคิดในการพัฒนาโปรแกรมอยู่บนพื้นฐานของทฤษฎีวงจรการพัฒนาระบบ ปัญหาของการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางได้ถูกวิเคราะห์พบว่ามี 2 ปัญหา คือ ความล่าช้าและความคลาดเคลื่อนในการตรวจสอบ โปรแกรมประยุกต์บนเว็บสำหรับตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางได้ถูกพัฒนาขึ้น อัลกอริทึมหลักของโปรแกรมประยุกต์นี้คือการเปรียบเทียบสายอักขระแบบประมาณโดยการหาระยะทางเลเวนชเตย์น การประเมินประสิทธิภาพของโปรแกรมพิจารณาจากค่า F1 **ผลการวิจัย:** การใช้เทคนิค n-gram ในการกำหนดคุณลักษณะของคำเพื่อเพิ่มประสิทธิภาพในการเปรียบเทียบสายอักขระ พบว่า คุณลักษณะของคำที่เหมาะสมกับงานวิจัยนี้คือ รูปแบบ trigram กำหนดค่า threshold ของความคล้ายคลึงกันของสายอักขระที่ 80, 85 และ 90 การกำหนดค่า threshold 80 หมายถึง สายอักขระสองชุดซึ่งมีความคล้ายคลึงกันมากกว่าหรือเท่ากับร้อยละ 80 จะกำหนดว่าเป็นสายอักขระชุดเดียวกัน ในการเปรียบเทียบประสิทธิภาพของโปรแกรมโดยใช้เครื่องมือตัดคำภาษาไทยทั้ง 8 วิธี เครื่องมือตัดคำที่มีประสิทธิภาพสูงสุดเมื่อกำหนดค่า threshold เป็น 85 และ 90 คือ attacut และ tltk ตามลำดับ ไม่พบความแตกต่างกันอย่างมีนัยสำคัญทางสถิติเมื่อเปรียบเทียบผลการพิจารณาชื่อผลิตภัณฑ์เครื่องสำอางกับผลการพิจารณาโดยเจ้าหน้าที่ผู้เชี่ยวชาญ **สรุป:** โปรแกรมถูกใช้เพื่อการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในแบบคำขอจดแจ้งเครื่องสำอางด้วยวิธีการเปรียบเทียบสายอักขระแบบประมาณ ผลการพิจารณาชื่อผลิตภัณฑ์เครื่องสำอางมีความถูกต้องและเป็นไปในทิศทางเดียวกันกับการพิจารณาโดยผู้เชี่ยวชาญ จึงสามารถนำไปใช้เพื่อสนับสนุนเจ้าหน้าที่ในการตรวจสอบข้อมูลในแบบคำขอจดแจ้งเครื่องสำอางเบื้องต้น

คำสำคัญ: การตัดคำภาษาไทย เปรียบเทียบสายอักขระแบบประมาณ ระยะทางเลเวนชเตย์น ชื่อผลิตภัณฑ์เครื่องสำอาง

รับต้นฉบับ: 24 ส.ค. 2565, ได้รับบทความฉบับปรับปรุง: 7 ต.ค. 2565, ivalงตีพิมพ์: 12 ต.ค. 2565

ผู้ประสานงานบทความ: จิราภรณ์ วงษ์เลิศ กลุ่มงานคุ้มครองผู้บริโภคและเภสัชสาธารณสุข สำนักงานสาธารณสุขจังหวัดพระนครศรีอยุธยา อำเภอพระนครศรีอยุธยา จังหวัดพระนครศรีอยุธยา 13000 **E-mail:** jiraporn.wonglert@gmail.com

Using Approximate String Matching for Verifying Cosmetic Product Names in the Cosmetic Notification Request Form

Jiraporn Wonglert¹, Verayuth Lertnattee²

¹Consumer Protection and Public Health Pharmacy Department,
Phra Nakhon Si Ayutthaya Provincial Public Health Office

²Department of Biomedicine and Health Informatics, Faculty of Pharmacy, Silpakorn University

Abstract

Objectives: To develop a program for checking cosmetic product names in the cosmetic notification request form using approximate string matching by Levenshtein edit distance and to evaluate the program's performance by comparing its verifying result with that from experts. **Method:** The program development framework was based on the system development life cycle theory. The problems on verifying cosmetic product names were analyzed with two problems identified including being time consuming and inaccuracy in the verification of product names. A web-based application for verifying cosmetic product names was developed. The main algorithm in the application was an approximate string matching by Levenshtein edit distance. The performance of the application was assessed using F1. **Results:** The n-gram technique was applied to create a set of appropriate features for increasing the performance in string matching. The appropriate n-gram in this study was trigram with thresholds at 80, 85, and 90. The threshold at 80 meant that two strings with a text similarity greater than or equal to 80 percent were defined as the same string. In comparing the efficiency of the program using the 8 engines of Thai word segmentation, the best engines at thresholds set to 85 and 90, were the attacut and tltk, respectively. There were no significant differences in the results on verification of cosmetic product names of the program and those from experts. **Conclusion:** This program was developed for verifying the names of cosmetic products in the notification request form using approximate string matching. The results of verification on cosmetic product names are accurate and consistent to those of experts. It can be used to support authorities in preliminary verification of information in the cosmetic notification request form.

Keywords: Thai word segmentation, approximate string matching, Levenshtein edit distance, cosmetic product names

บทนำ

ในปัจจุบันคนส่วนใหญ่ให้ความสำคัญเกี่ยวกับการดูแลสุขภาพมากขึ้น รวมทั้งในด้านความงามและผิวพรรณ ส่งผลให้เกิดการวิจัยและพัฒนาผลิตภัณฑ์ประเภทใหม่ ๆ ทำให้อุตสาหกรรมเครื่องสำอางมีอัตราการเจริญเติบโตและขยายตัวอย่างต่อเนื่อง จากข้อมูลของ Euromonitor (1) อุตสาหกรรมเครื่องสำอางประเทศไทยในปี 2564 เติบโตร้อยละ 5 หรือมีมูลค่ารวม 1.447 แสนล้านบาท กลุ่มผลิตภัณฑ์ดูแลผิวครองส่วนแบ่งสูงสุดร้อยละ 57.5 รองลงมาคือ กลุ่มผลิตภัณฑ์ดูแลเส้นผม ร้อยละ 21 กลุ่มผลิตภัณฑ์ตกแต่งร้อยละ 15.5 และน้ำหอมร้อยละ 6 ตามลำดับ นอกจากนี้จากข้อมูลของกรมเจรจาการค้าระหว่างประเทศ (2) พบว่า ประเทศไทยเป็นผู้ส่งออกเครื่องสำอางอันดับที่ 2 ของอาเซียน รองจากสิงคโปร์ และเป็นอันดับที่ 10 ของโลก ด้วยเหตุผลข้างต้นจึงทำให้อุตสาหกรรมเครื่องสำอางได้รับความสนใจจากผู้ประกอบการทั้งรายใหญ่และรายย่อย พระราชบัญญัติเครื่องสำอาง พ.ศ. 2558 กำหนดให้ผู้ประกอบการที่ประสงค์จะผลิต นำเข้า หรือรับจ้างผลิตเครื่องสำอางจะต้องจดทะเบียนรายละเอียดของเครื่องสำอางต่อผู้รับจดทะเบียน และเมื่อผู้รับจดทะเบียนออกใบรับจดทะเบียนให้แล้ว จึงจะสามารถผลิตหรือนำเข้าเครื่องสำอางนั้นได้ (3)

ตั้งแต่วันที่ 1 สิงหาคม พ.ศ. 2559 สำนักงานคณะกรรมการอาหารและยา (อย.) ได้มีนโยบายให้ผู้ประกอบการยื่นคำขอจดทะเบียนเครื่องสำอางผ่านระบบจดทะเบียนเครื่องสำอางอัตโนมัติ (e-submission) เพื่อเป็นการอำนวยความสะดวกให้แก่ผู้ประกอบการ ส่งผลให้มีคำขอจดทะเบียนเครื่องสำอางถูกส่งเข้ามาในระบบเป็นจำนวนมาก จากการศึกษาของเกษรา จันทรวงศ์ไพศาล และชิตชนก เรือนก้อน (4) พบว่า ปริมาณคำขอจดทะเบียนเครื่องสำอาง ตั้งแต่เดือนตุลาคม 2558 จนถึงเดือนกันยายน 2562 มีจำนวนทั้งสิ้น 815,627 รายการ เฉลี่ยปีละประมาณ 203,907 รายการ โดยมีการอนุมัติเฉลี่ยต่อปีที่ 160,372 รายการ (ร้อยละ 78.65) ไม่อนุมัติเฉลี่ย 43,535 รายการ (ร้อยละ 21.35)

ตั้งแต่วันที่ 15 สิงหาคม พ.ศ. 2562 เป็นต้นมา อย. มอบหมายให้พนักงานเจ้าหน้าที่ในสำนักงานสาธารณสุขจังหวัดเป็นผู้รับจดทะเบียนเครื่องสำอางผ่านระบบ e-submission ในปี 2564 จังหวัดพระนครศรีอยุธยา มีรายการคำขอจดทะเบียนเครื่องสำอางที่พนักงานเจ้าหน้าที่ต้องพิจารณาประมาณ 1,200 รายการ หรือเฉลี่ย 100 คำขอต่อเดือน โดย

มีระยะเวลาในการพิจารณาแต่ละคำขอไม่เกิน 3 วันทำการ สำหรับสำนักงานสาธารณสุขจังหวัดพระนครศรีอยุธยามีพนักงานเจ้าหน้าที่ตรวจสอบและพิจารณาคำขอจดทะเบียนเครื่องสำอางรวมทั้งหมด 4 ท่าน

การยื่นคำขอผ่านระบบ e-submission แม้ว่าจะสามารถตรวจสอบข้อมูลเบื้องต้นในส่วนของการห้ามใช้ในชื่อผลิตภัณฑ์เครื่องสำอาง ทั้งชื่อการค้าและชื่อเครื่องสำอาง สามารถตรวจสอบสารห้ามใช้ในสูตรตำรับ รวมถึงตรวจสอบเงื่อนไขการใช้สารที่เป็นส่วนประกอบในสูตรได้ แต่ระบบดังกล่าวยังมีข้อจำกัดบางประการ เช่น กรณีชื่อการค้าหรือชื่อเครื่องสำอางพบคำที่มีความคล้ายคลึงแต่ไม่เหมือนกับคำห้ามใช้ (text similarity) ซึ่งอาจเกิดจากการสะกดผิดทั้งที่ไม่ได้ตั้งใจหรือเป็นความตั้งใจของผู้ประกอบการในการหลีกเลี่ยงคำห้ามใช้ หรือต้องการใช้ชื่อที่มีความเป็นเอกลักษณ์ ระบบ e-submission จะไม่สามารถตรวจสอบคำนั้นได้ นอกจากนี้ยังไม่สามารถตรวจสอบความสอดคล้องกันระหว่างชื่อผลิตภัณฑ์เครื่องสำอาง (ชื่อการค้าและชื่อเครื่องสำอาง) กับสารที่เป็นส่วนประกอบในตำรับได้ ในส่วนนี้จึงเป็นหน้าที่ของพนักงานเจ้าหน้าที่ในการตรวจสอบ การตรวจสอบความสอดคล้องกันระหว่างชื่อผลิตภัณฑ์เครื่องสำอางกับสารที่เป็นส่วนประกอบในตำรับ เช่น หากต้องการใช้ชื่อ vitamin c เป็นส่วนหนึ่งของชื่อการค้าหรือชื่อเครื่องสำอาง ในสูตรตำรับต้องมีวิตามินซีหรืออนุพันธ์ของวิตามินซี เช่น ascorbic acid, ascorbyl palmitate เป็นส่วนประกอบด้วย (5)

เนื่องจากกระบวนการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางโดยพนักงานเจ้าหน้าที่ยังมีข้อจำกัดในเรื่องของความรู้ ประสบการณ์ และดุลยพินิจของพนักงานเจ้าหน้าที่แต่ละบุคคล ซึ่งอาจทำให้เกิดความผิดพลาดในการตรวจสอบได้ รวมถึงข้อมูลหรือองค์ความรู้ที่มีอยู่ก็กระจัดกระจาย ไม่มีการจัดเก็บอย่างเป็นระบบ ทำให้ใช้เวลาในการตรวจสอบข้อมูลแต่ละคำรับค่อนข้างนาน ดังนั้นผู้วิจัยจึงได้นำวิธีการเปรียบเทียบสายอักขระแบบประมาณมาประยุกต์ใช้เพื่อเพิ่มประสิทธิภาพในการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในแบบคำขอจดทะเบียนเครื่องสำอาง ให้มีความถูกต้อง รวดเร็ว และเป็นไปในทิศทางเดียวกัน โดยการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในงานวิจัยนี้จะครอบคลุมการทำงานทั้งหมด 5 ขั้นตอน คือ 1) การตรวจสอบคำห้ามใช้ในชื่อการค้าภาษาไทย 2) การตรวจสอบคำห้ามใช้ในชื่อการค้าภาษาอังกฤษ 3) การตรวจสอบคำห้าม

ใช้ในชื่อเครื่องสำอางภาษาไทย 4) การตรวจสอบคำห้ามใช้ในชื่อเครื่องสำอางภาษาอังกฤษ และ 5) การตรวจสอบความสอดคล้องระหว่างชื่อผลิตภัณฑ์เครื่องสำอาง (ชื่อการค้าและชื่อเครื่องสำอาง) กับสารที่เป็นส่วนประกอบในตำรับ

การเปรียบเทียบสายอักขระแบบประมาณ (approximate string matching) หรือที่เรียกอีกชื่อหนึ่งว่า fuzzy string matching เป็นวิธีการหาความแตกต่างระหว่างคำหรือข้อความที่มีความคล้ายคลึงกันแต่ไม่เหมือนกันทุกประการ (6-7) จากการทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องพบว่า อัลกอริทึมที่นิยมใช้กันอย่างแพร่หลายในการวัดความคล้ายคลึงกันของคำหรือข้อความ คือ วิธีการหาระยะทางเลเวนชเตย์น (Levenshtein distance, LD หรือ Levenshtein edit distance) ซึ่งถูกพัฒนาโดย Vladimir Levenshtein ในปี 1996 (8) วิธีนี้หาค่าความต่างกันของสายอักขระสองชุด ระหว่างชุดแรกที่เป็นต้นแบบและชุดที่สองที่เป็นชุดเปรียบเทียบ โดยค่าความต่างกันจะวัดจากจำนวนของการที่จะต้องทำการตัดออก (deletions) การแทรก (insertions) และการแทนที่ (substitutions) อักขระในชุดที่นำมาเปรียบเทียบจนกระทั่งมีลักษณะเหมือนชุดอักขระที่เป็นต้นแบบทุกประการ

Yulianto, Arifudin และ Alamsyah (9) ประยุกต์ใช้อัลกอริทึมของเลเวนชเตย์นในการเติมข้อความอัตโนมัติและการตรวจสอบตัวสะกดเพื่อแก้ไขข้อความที่ผิดพลาดในการค้นหาข้อมูลในห้องสมุดพบว่าทำให้ข้อมูลมีความถูกต้องร้อยละ 86 ส่วนงานวิจัยของ Haldar และ Mukhopadhyay (10) นำอัลกอริทึมของเลเวนชเตย์นมาประยุกต์ใช้ร่วมกับการแบ่งกลุ่มอักขระตามความคล้ายคลึงกันเพื่อนำไปใช้ในการค้นหาคำในดัชนีนารี พบว่าทำให้การค้นหาคำในดัชนีนารีมีประสิทธิภาพมากยิ่งขึ้น สำหรับในประเทศไทยพบการนำอัลกอริทึมของเลเวนชเตย์นมาใช้ในการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ เพื่อติดแท็กจำแนกปัญหาในการให้บริการของรถโดยสารสาธารณะ โดยจักรินทร์ สันติรัตนภักดี และศุภกฤษฎี นวัตกรรมกุล (11) พบว่า การเปรียบเทียบความคล้ายคลึงกันของอักขระสองชุดด้วยเทคนิคการวัด LD ช่วยให้การเปรียบเทียบกับคำศัพท์ที่มีความถูกต้อง และจำแนกปัญหาการให้บริการของรถโดยสารสาธารณะได้อย่างมีประสิทธิภาพ นอกจากนี้ การศึกษาของฐิติมา ผลึกเพชร และวิรุทธิ์ เลิศนที ได้นำอัลกอริทึมเลเวนชเตย์นมาประยุกต์ใช้เพื่อหาชื่อพ้องรูปและ

ชื่อคล้ายของพืชสมุนไพร โดยผลจากการใช้อัลกอริทึมดังกล่าวสามารถจำแนกชื่อพืชสมุนไพรได้ 5 กลุ่มตามคำร้อยละความเหมือนระหว่างสายอักขระของชื่อพืชสมุนไพร (12)

การตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางทั้ง 5 ขั้นตอน จะเป็นการนำคำแต่ละคำไปเปรียบเทียบสายอักขระแบบประมาณกับคำสำคัญในฐานข้อมูล แต่เนื่องจากชื่อผลิตภัณฑ์เครื่องสำอางทั้งชื่อการค้าและชื่อเครื่องสำอางในส่วนที่เป็นภาษาไทยมีลักษณะเป็นข้อความที่ยาวต่อเนื่องกันโดยไม่มีการใช้ตัวอักษรหรือสัญลักษณ์ที่นำมาใช้คั่นระหว่างคำ ไม่มีการใช้เครื่องหมายวรรคตอนแสดงการแบ่งคำดังเช่นภาษาอังกฤษ นอกจากนี้ยังมีโครงสร้างของคำที่ไม่แน่นอน ซึ่งเป็นอุปสรรคในการนำภาษาไทยไปใช้ในการเปรียบเทียบสายอักขระ ดังนั้นในการพัฒนาโปรแกรมฯ ต้องมีการตัดคำ (word segmentation หรือ word separation) เพื่อเป็นการแบ่งคำแต่ละคำในประโยคออกจากกันก่อนที่จะนำไปใช้ในกระบวนการต่อไป สำหรับวิธีที่ใช้ตัดคำจะแบ่งออกเป็น 3 ประเภทใหญ่ ๆ ได้แก่ การใช้กฎ การใช้พจนานุกรม และการใช้คลังข้อความ เทคนิคที่ใช้ในการตัดคำที่นิยมใช้กันทั่วไปคือ วิธีการเทียบคำที่ยาวที่สุด (longest word pattern matching) วิธีการเทียบคำที่สั้นที่สุด (shortest word pattern matching) วิธีการตัดคำที่ใช้ความถี่ของคำ (word usage frequency) และวิธีการย้อนรอยกลับ (back tracking) (13)

เนื่องจากคำสำคัญในฐานข้อมูล ได้แก่ คำห้ามใช้และคำที่สามารถนำมาใช้เป็นชื่อผลิตภัณฑ์เครื่องสำอางได้ โดยต้องมีสารที่สอดคล้องในสูตรตำรับ มีลักษณะเป็นทั้งคำเดี่ยวและกลุ่มคำที่ประกอบด้วยคำตั้งแต่สองคำขึ้นไป แต่คำที่ได้จากขั้นตอนการตัดคำจะมีลักษณะเป็นคำเดี่ยวหลาย ๆ คำ เมื่อนำไปเปรียบเทียบสายอักขระแบบประมาณกับคำสำคัญในฐานข้อมูล อาจไม่สื่อความหมายหรือไม่ตรงกับคำที่ต้องการ ผู้วิจัยจึงได้นำเทคนิค n-gram เข้ามาใช้ในการกำหนดคุณลักษณะของคำ เพื่อเป็นการเพิ่มความถูกต้องของคำ และเพิ่มประสิทธิภาพในการเปรียบเทียบสายอักขระ

เอ็นแกรม (n-gram) คือคำหรือสายอักขระที่อยู่ติดกันจำนวน n ลำดับ ยกตัวอย่างเช่น bigram หมายถึงสายอักขระที่มีคำสองคำที่ติดกัน n-gram ถูกนำไปประยุกต์ใช้กับงานด้านการประมวลผลภาษาธรรมชาติ ซึ่งส่วนใหญ่จะนำไปใช้ในการแก้ไขข้อจำกัดของการตัดคำในภาษาไทย ช่วยเพิ่มประสิทธิภาพของระบบให้ดีขึ้น

งานวิจัยของอัครพล เอกวงค์อนันต์ (14) ศึกษาการระบุคำไทยและคำทับศัพท์ด้วยแบบจำลองเอ็นแกรม พบว่า การใช้ n-gram ในการระบุภาษา ทำให้ระบบสามารถระบุภาษาของคำไทย คำทับศัพท์ภาษาอังกฤษ และญี่ปุ่นได้ถูกต้องกว่าร้อยละ 90

การวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาโปรแกรมสำหรับใช้ในการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางด้วยวิธีการเปรียบเทียบสายอักขระแบบประมาณโดยการหา LD และประเมินประสิทธิภาพของโปรแกรมโดยเปรียบเทียบผลการพิจารณาชื่อผลิตภัณฑ์เครื่องสำอางกับการพิจารณาโดยเจ้าหน้าที่ผู้เชี่ยวชาญ

วิธีการวิจัย

งานวิจัยนี้เป็นการวิจัยเชิงทดลองที่ได้รับการยกเว้นการพิจารณาด้านจริยธรรม จากคณะกรรมการพิจารณาจริยธรรมการวิจัยในมนุษย์ มหาวิทยาลัยศิลปากร (เลขที่โครงการ REC 65.0222-037-0882)

ประชากรและกลุ่มตัวอย่าง

ประชากร คือ แบบคำขอจดแจ้งเครื่องสำอาง (แบบ จ.ค.1) ที่ถูกส่งเข้ามาในระบบจดแจ้งเครื่องสำอางอัตโนมัติ e-submission ของสำนักงานสาธารณสุขจังหวัดพระนครศรีอยุธยา ตั้งแต่วันที่ 1 ตุลาคม พ.ศ. 2559 ถึงวันที่ 30 พฤศจิกายน พ.ศ. 2564 จำนวน 3,100 รายการ แบบคำขอจดแจ้งเครื่องสำอางสามารถจำแนกตามผลการตรวจสอบได้เป็น 2 ประเภท คือ แบบคำขอจดแจ้งเครื่องสำอางที่พนักงานเจ้าหน้าที่ตรวจสอบและพิจารณาว่าผ่าน จำนวน 1,590 รายการ และแบบคำขอจดแจ้งเครื่องสำอางที่พนักงานเจ้าหน้าที่ตรวจสอบและพิจารณาว่าไม่ผ่าน จำนวน 1,510 รายการ

การคำนวณขนาดตัวอย่างอ้างอิงจากการวิจัยเบื้องต้นโดยตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในแบบคำขอจดแจ้งเครื่องสำอาง จำนวน 200 รายการด้วยวิธีการเปรียบเทียบสายอักขระแบบประมาณเทียบกับการตรวจสอบโดยเจ้าหน้าที่ผู้เชี่ยวชาญ ซึ่งเป็นการศึกษาเปรียบเทียบโดยใช้ข้อมูล 2 กลุ่มที่ไม่เป็นอิสระต่อกันที่อิงกับการทดสอบ McNemar's test (15) การทดสอบเป็นแบบสองทาง โดยกำหนดความคลาดเคลื่อนชนิดที่หนึ่งที่ 0.05 อำนาจการทดสอบที่ และ McNemar odds ratio เท่ากับ 3 การคำนวณได้ขนาดตัวอย่างเท่ากับ 389 ตัวอย่าง การวิจัยนี้จึงใช้ตัวอย่างแบบคำขอจดแจ้งฯ ในระบบ e-

submission ของสำนักงานสาธารณสุขจังหวัดพระนครศรีอยุธยา 400 รายการ ในการเลือกตัวอย่าง ผู้วิจัยใช้วิธีการสุ่มตัวอย่างแบบง่ายด้วยการใช้โปรแกรมคอมพิวเตอร์

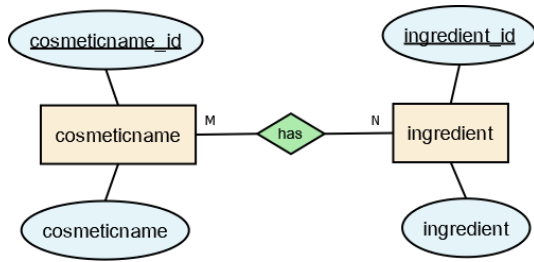
การพัฒนาโปรแกรมฯ ผู้วิจัยมุ่งเน้นที่กระบวนการตรวจสอบคำห้ามใช้ในชื่อผลิตภัณฑ์เครื่องสำอาง และคำที่สามารถนำมาใช้เป็นชื่อผลิตภัณฑ์เครื่องสำอางได้โดยต้องมีสารที่สอดคล้องในสูตรตำรับ ซึ่งได้กำหนดกรอบแนวคิดในการพัฒนาโปรแกรมโดยใช้ทฤษฎีวงจรการพัฒนากระบวนการ (system development life cycle: SDLC) ซึ่งประกอบด้วย 5 ขั้นตอน ได้แก่ การวิเคราะห์ การออกแบบ การพัฒนา การทดสอบ และการบำรุงรักษา (16)

ขั้นตอนที่ 1 การวิเคราะห์

ผู้วิจัยสัมภาษณ์พนักงานเจ้าหน้าที่ผู้ทำหน้าที่ตรวจสอบและพิจารณาคำขอจดแจ้งเครื่องสำอางรายบุคคล จำนวน 3 ท่านเพื่อวิเคราะห์กระบวนการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในปัจจุบัน โดยใช้คำถามหลัก คือ การตรวจสอบและพิจารณาคำขอจดแจ้งเครื่องสำอางในปัจจุบันพบปัญหาด้านใดบ้างที่ส่งต่อการพิจารณาคำขอของพนักงานเจ้าหน้าที่ ลักษณะของคำถามเป็นแบบปลายเปิดเพื่อให้ผู้ถูกสัมภาษณ์มีอิสระในการตอบคำถาม จากนั้นนำข้อมูลที่ได้อาวิเคราะห์และสรุปเป็นประเด็นปัญหา พร้อมกับศึกษาเอกสารต่าง ๆ ที่เกี่ยวข้องกับการตรวจสอบและพิจารณาชื่อผลิตภัณฑ์เครื่องสำอาง ได้แก่ หลักเกณฑ์การพิจารณาการจดแจ้งเครื่องสำอาง (ฉบับปรับปรุง พ.ศ. 2564) คู่มือจดแจ้งเครื่องสำอางระบบอัตโนมัติ และแบบคำขอจดแจ้งเครื่องสำอาง (แบบ จ.ค.1) ทั้งยังศึกษาแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องกับการพัฒนาโปรแกรมสำหรับใช้ในการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางด้วยวิธีการเปรียบเทียบสายอักขระแบบประมาณโดยการหา LD

ขั้นตอนที่ 2 การออกแบบ

ผู้วิจัยออกแบบฐานข้อมูล หน้าจอรับข้อมูล และหน้าจอแสดงผลข้อมูล ซึ่งฐานข้อมูลที่เกี่ยวข้องกับการเก็บคำสำคัญ ได้แก่ คำห้ามใช้ในชื่อผลิตภัณฑ์เครื่องสำอางและคำที่สามารถนำมาใช้เป็นชื่อผลิตภัณฑ์เครื่องสำอางได้โดยต้องมีสารที่สอดคล้องในสูตรตำรับ ซึ่งแสดงความสัมพันธ์ของข้อมูลดังแสดงในรูปที่ 1 การเขียนโครงสร้างฐานข้อมูลที่เกี่ยวข้องใช้ภาษา SQL ใช้ระบบจัดการฐานข้อมูล MySQL และใช้โปรแกรม phpMyAdmin เป็นเครื่องมือช่วยในการจัดการฐานข้อมูล



รูปที่ 1. ความสัมพันธ์ระหว่างข้อมูล (entity relationship diagram : ER-diagram) เส้นที่เชื่อมแสดงความสัมพันธ์ของชุดข้อมูลแบบ many to many (M:N)

ขั้นตอนที่ 3 การพัฒนา

ผู้วิจัยพัฒนาโปรแกรมโดยอาศัยความรู้ด้านการออกแบบและพัฒนาเว็บไซต์ด้วยภาษา PHP เป็นภาษาหลัก ซึ่งเป็นภาษาสคริปต์ (scripting language) ที่ทำงานแบบ server-side-script ภาษา PHP มีลักษณะเป็น open source ที่มีความสามารถเป็นแบบไดนามิก (dynamic language) สามารถใช้งานได้ทั้งบนระบบปฏิบัติการยูนิกซ์ (Unix) และวินโดวส์ (Windows) และยังสามารถติดต่อกับโปรแกรมฐานข้อมูล เช่น MySQL ได้ (17) โดยพัฒนาร่วมกับภาษา HTML ซึ่งเป็นภาษาที่ใช้ในการพัฒนาเว็บเพจ ช่วยให้โปรแกรมมีการตอบสนองกับผู้ใช้งานได้ดียิ่งขึ้น และสามารถเรียกใช้งานได้จากเว็บเบราว์เซอร์ (18) การพัฒนาโปรแกรมในแต่ละส่วน ดังนี้

การตัดคำภาษาไทย: การพัฒนาโปรแกรมในส่วนของการตัดคำภาษาไทย ผู้วิจัยใช้ Library PythaiNLP (19) ซึ่งเป็นไลบรารีประมวลผลภาษาไทยแบบ open source ในภาษาไพธอน (python) มีฟังก์ชันที่ถูกออกแบบมาให้รองรับการตัดคำหลากหลายอัลกอริทึม (algorithm) สามารถปรับเปลี่ยนเครื่องมือการตัดคำ (engine) ได้หลายรูปแบบ งานวิจัยนี้เปรียบเทียบเครื่องมือตัดคำ 8 วิธี ได้แก่ newmm, longest, multi_cut, deepcut, attacut, newmm-safe, titk และ sefr_cut เพื่อหาเครื่องมือตัดคำที่เหมาะสม

		ล	เล	ล	ช	อ	ร	อ
	0	1	2	3	4	5	6	7
ล	1							
ห	2							
ล	3							
จ	4							
อ	5							
ร	6							
อ	7							

		ล	เล	ล	ช	อ	ร	อ
	0	1	2	3	4	5	6	7
ล	1	0	1	2	3	4	5	6
ห	2	1						
ล	3	2						
จ	4	3						
อ	5	4						
ร	6	5						
อ	7	6						

		ล	เล	ล	ช	อ	ร	อ
	0	1	2	3	4	5	6	7
ล	1	0	1	2	3	4	5	6
ห	2	1	1	2	3	4	5	6
ล	3	2	2	1	2	3	4	5
จ	4	3	3	2	2	3	4	5
อ	5	4	4	3	3	2	3	4
ร	6	5	5	4	4	3	2	3
อ	7	6	6	5	5	4	3	2

รูปที่ 2. ตัวอย่างการหาค่า LD จากการเปรียบเทียบสายอักขระแบบประมาณ กรณีภาษาไทย

ตารางที่ 1. ตัวอย่างผลการตัดคำด้วย Library PyThaiNLP โดยใช้เครื่องมือตัดคำทั้ง 8 วิธี

ข้อความตัวอย่าง: สบู่ก้อนมะขามผสมน้ำผึ้ง		
ลำดับ	engine	ผลลัพธ์ที่ได้
1	newmm	สบู่, ก้อน, มะขาม, ผสม, น้ำผึ้ง
2	longest	สบู่, ก้อน, มะขาม, ผสม, น้ำผึ้ง
3	multi_cut	สบู่, ก้อนมะขาม, ผสม, น้ำผึ้ง
4	deepcut	สบู่ก้อน, มะขาม, ผสม, น้ำผึ้ง
5	attacut	สบู่, ก้อน, มะขามผสม, น้ำ, ผึ้ง
6	newmm-safe	สบู่, ก้อน, มะขาม, ผสม, น้ำผึ้ง
7	titk	สบู่, ก้อน, มะขาม, ผสม, น้ำผึ้ง
8	sefr_cut	สบู่ก้อน, มะขาม, ผสมน้ำผึ้ง

ที่สุด ที่ทำให้การเปรียบเทียบสายอักขระแบบประมาณโดยการหา LD มีประสิทธิภาพสูงสุด ตัวอย่างผลการตัดคำด้วย Library PyThaiNLP โดยใช้เครื่องมือตัดคำทั้ง 8 วิธีแสดงอยู่ในตารางที่ 1

การกำหนดคุณลักษณะของคำ: ผู้วิจัยใช้เทคนิค n-gram ในการกำหนดคุณลักษณะของคำ 3 รูปแบบ ได้แก่ 1-แกรม (unigram) 2-แกรม (bigram) และ 3-แกรม (trigram) ซึ่ง bigram จะเป็นการรวมกันของคำในรูปแบบ unigram และ bigram ส่วน trigram จะเป็นการรวมกันของคำในรูปแบบ unigram, bigram และ trigram

การเปรียบเทียบสายอักขระแบบประมาณโดยการหา LD : การหา LD เป็นการวัดหาค่าความต่างกันของสายอักขระสองชุด โดยค่าความต่างกันจะวัดจากจำนวนของการที่จะต้องทำการตัดออก การแทรก และการแทนที่ของอักขระหนึ่งตัว ระยะทางคำนวณจากจำนวนน้อยที่สุดของระยะทางที่แก้ไข ยกตัวอย่างเช่น หาความต่างกันของคำภาษาไทย “เนเจอร์” กับ “เลเซอร์” ได้ค่า LD เท่ากับ 2 (รูปที่ 2) และคำภาษาอังกฤษ “ALCOHOL” กับ “ALGOHOL” ได้ค่า LD เท่ากับ 1 (รูปที่ 3)

		A	L	G	O	H	O	L
	0	1	2	3	4	5	6	7
A	1							
L	2							
C	3							
O	4							
H	5							
O	6							
L	7							

		A	L	G	O	H	O	L
	0	1	2	3	4	5	6	7
A	1	0	1	2	3	4	5	6
L	2	1						
C	3	2						
O	4	3						
H	5	4						
O	6	5						
L	7	6						

		A	L	G	O	H	O	L
	0	1	2	3	4	5	6	7
A	1	0	1	2	3	4	5	6
L	2	1	0	1	2	3	4	5
C	3	2	1	1		3	4	5
O	4	3	2	2	1	2	3	4
H	5	4	3	3	2	1	2	3
O	6	5	4	4	3	2	1	2
L	7	6	5	5	4	3	2	1

รูปที่ 3. ตัวอย่างการหาค่า LD จากการเปรียบเทียบสายอักขระแบบประมาณ กรณีภาษาอังกฤษ

สำหรับการทำงานของโปรแกรมฯ จะมี 2 ส่วนหลัก คือ ส่วนของการตรวจสอบคำห้ามใช้ในชื่อผลิตภัณฑ์เครื่องสำอางทั้งภาษาไทยและภาษาอังกฤษ และส่วนของการตรวจสอบความสอดคล้องกันระหว่างชื่อผลิตภัณฑ์เครื่องสำอางกับส่วนประกอบในตำรับ ซึ่งใช้วิธีการเปรียบเทียบสายอักขระโดยการหา LD ที่พัฒนาโดยใช้ภาษา PHP ทำการเปรียบเทียบกับคำสำคัญในฐานข้อมูลที่ได้จัดทำไว้เมื่อได้ค่า LD โปรแกรมจะนำค่าที่ได้ไปคำนวณหาความคล้ายคลึงกันของข้อความ (text similarity) ค่าที่ได้จากการคำนวณแสดงถึงลักษณะความคล้ายซึ่งมีค่าระหว่าง 0 ถึง 1 หากมีค่าเท่ากับ 0 หมายถึงสายอักขระทั้งสองสายไม่มีความคล้ายกันเลย แต่หากมีค่าเท่ากับ 1 หมายถึงสายอักขระทั้งสองสายเหมือนกันทุกประการ (20) สูตรการคำนวณนำเสนอในสมการที่ 1

$$\text{Text Similarity } (t1, t2) = 1 - \frac{\text{Levenshtein Distance } (t1, t2)}{\text{Max}(n(t1), n(t2))} \quad (1)$$

โดย t1, t2 แทน สายอักขระชุดต้นแบบและสายอักขระชุดเปรียบเทียบ; Levenshtein Distance แทน ค่าความต่างกันระหว่างสายอักขระสองชุด โดยวัดจากจำนวนของการตัดออก แทรก แทนที่อักขระในชุดที่นำมาเปรียบเทียบจนกระทั่งมีลักษณะเหมือนชุดอักขระที่เป็นต้นแบบทุกประการ ซึ่งการตัดออก แทรก หรือแทนที่อักขระ 1 ครั้ง จะมีค่าน้ำหนักเท่ากับ 1; n แทน ความยาวของสายอักขระ; Max(n(t1), n(t2)) แทน ค่าความยาวสูงสุดของสายอักขระใดสายอักขระหนึ่งจากสายอักขระสองชุด

การกำหนดค่า threshold: threshold เป็นค่าที่ใช้ในการเปรียบเทียบความคล้ายคลึงกันของคำหรือข้อความโดยการหา LD ซึ่งการกำหนดค่า threshold 80 หมายถึง สายอักขระสองชุดซึ่งมีความคล้ายคลึงกัน (text similarity) มากกว่าหรือเท่ากับร้อยละ 80 จะกำหนดว่าเป็นสายอักขระชุดเดียวกัน สำหรับงานวิจัยนี้กำหนดค่า threshold ที่ 80, 85 และ 90

ขั้นตอนการทำงานของโปรแกรมแสดงอยู่ในรูปที่ 4

ขั้นตอนที่ 4 การทดสอบ

ผู้วิจัยทดสอบโปรแกรมเพื่อหาข้อบกพร่องหลังจากนั้นปรับปรุงให้ดีขึ้น ต่อมาทดสอบโปรแกรมกับตัวอย่าง และเปรียบเทียบผลการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางกับการตรวจสอบโดยเจ้าหน้าที่ผู้เชี่ยวชาญ และประเมินประสิทธิภาพของโปรแกรม

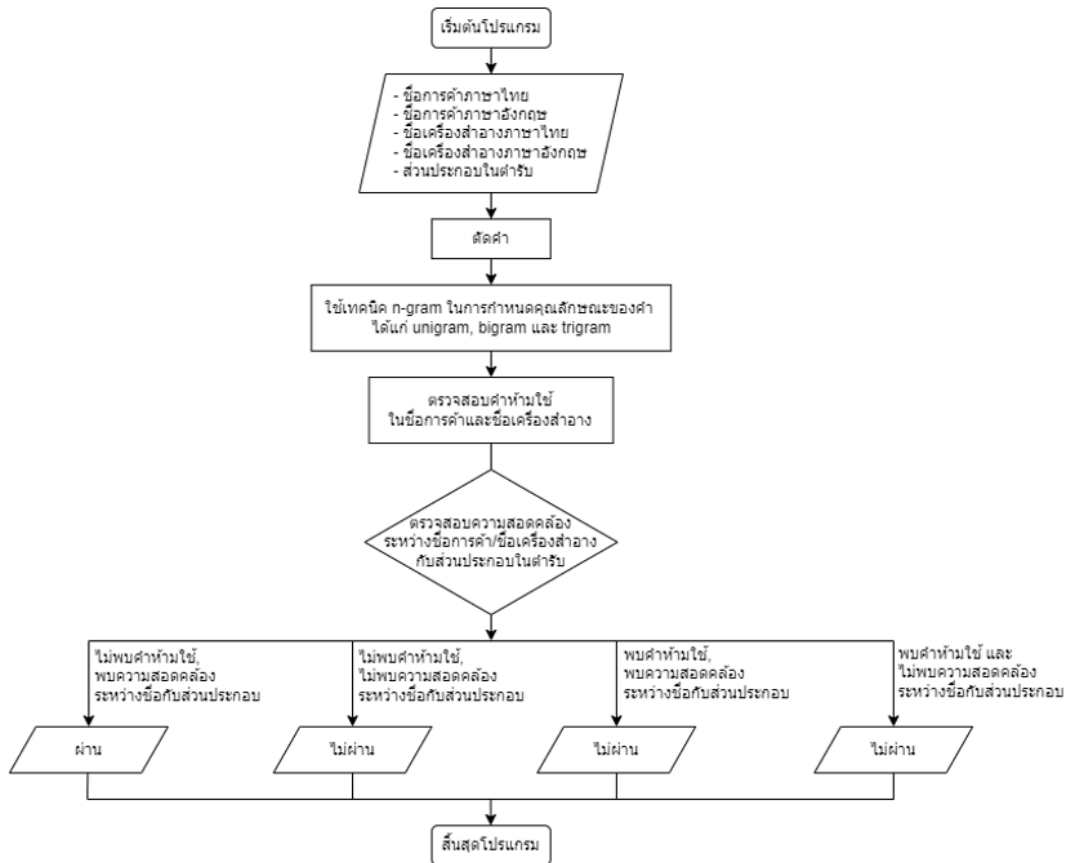
ขั้นตอนที่ 5 การบำรุงรักษา

เมื่อนำโปรแกรมไปใช้งานต้องมีการติดตามและบำรุงรักษา เพื่อให้โปรแกรมสามารถใช้งานได้อย่างมีประสิทธิภาพ รวมถึงการแก้ไขปรับปรุงโปรแกรมกรณีที่เกิดข้อผิดพลาดที่พบบ่อย และการเขียนโมดูลการทำงานเพิ่มหากผู้ใช้งานพบว่ามีปัญหาเกิดขึ้น โดยอาจจะมีการบันทึกปัญหาเหล่านั้นไว้และส่งต่อไปให้ผู้วิจัยเพื่อทำการแก้ไขโปรแกรมต่อไป

การประเมินประสิทธิภาพของโปรแกรม

ผู้วิจัยประเมินประสิทธิภาพของโปรแกรมโดยการเปรียบเทียบผลการตรวจสอบด้วยโปรแกรมกับผลการตรวจสอบโดยเจ้าหน้าที่ผู้เชี่ยวชาญ ซึ่งแบบคำขอจดแจ้งเครื่องสำอางที่ใช้ในการวิจัยนี้จะถูกตรวจสอบชื่อการค้า ชื่อเครื่องสำอาง และส่วนประกอบในตำรับเครื่องสำอาง และประเมินผลการตรวจสอบใหม่อีกครั้งโดยเจ้าหน้าที่ผู้เชี่ยวชาญจำนวน 3 ท่าน โดยประเมินอย่างเป็นอิสระต่อกัน หากผลการประเมินผ่านทั้ง 3 ท่าน จะถือว่าเป็นแบบคำขอจดแจ้งเครื่องสำอางที่เจ้าหน้าที่ผู้เชี่ยวชาญพิจารณาว่าผ่าน

การศึกษานำข้อมูลที่ได้มาวิเคราะห์ประสิทธิภาพการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางโดยใช้ confusion matrix คำนวณหาค่าความถูกต้อง (accuracy) ค่าความระลึก (recall) ค่าความเที่ยง (precision) และค่า F1 (21) ซึ่งเป็นค่าเฉลี่ยฮาร์มอนิก (harmonic mean) ระหว่างค่าความเที่ยงและค่าความระลึก สูตรการคำนวณนำเสนอในสมการที่ (2) - (5)



รูปที่ 4. ขั้นตอนการทำงานของโปรแกรม

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

โดย TP (true positive): ผลบวกจริง หมายถึง โปรแกรมแสดงผลการตรวจสอบเป็น “ผ่าน” เจ้าหน้าที่ผู้เชี่ยวชาญตรวจสอบและพิจารณาว่า “ผ่าน”; TN (true negative): ผลลบจริง หมายถึง โปรแกรมแสดงผลการตรวจสอบเป็น “ไม่ผ่าน” เจ้าหน้าที่ผู้เชี่ยวชาญตรวจสอบและพิจารณาว่า “ไม่ผ่าน”; FP (false positive): ผลบวกлож หมายถึง โปรแกรมแสดงผลการตรวจสอบเป็น “ผ่าน” แต่เจ้าหน้าที่ผู้เชี่ยวชาญตรวจสอบและพิจารณาว่า “ไม่ผ่าน”; FN (false negative): ผลลบлож หมายถึง โปรแกรมแสดงผลการตรวจสอบเป็น “ไม่ผ่าน” แต่เจ้าหน้าที่ผู้เชี่ยวชาญตรวจสอบและพิจารณาว่า “ผ่าน”

การวิเคราะห์ข้อมูล

การเปรียบเทียบผลการตรวจด้วยวิธีการเปรียบเทียบสายอักขระแบบประมาณ และผลการตรวจสอบโดย

เจ้าหน้าที่ผู้เชี่ยวชาญใช้สถิติ McNemar's Chi-Square โดยกำหนดระดับนัยสำคัญเท่ากับ 0.05 ($\alpha = 0.05$)

ผลการวิจัย

ผลการวิเคราะห์กระบวนการตรวจสอบ

ผลการวิเคราะห์กระบวนการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในปัจจุบัน พบปัญหาที่อาจทำให้การตรวจสอบและการพิจารณาคำขอจดแจ้งเครื่องสำอางล่าช้าหรือเกิดความคลาดเคลื่อน ได้แก่

1. ปัญหาความล่าช้าในการตรวจสอบและพิจารณาคำขอ เนื่องจากข้อมูลคำห้ามใช้เป็นชื่อผลิตภัณฑ์เครื่องสำอางและคำที่สามารถใช้เป็นชื่อผลิตภัณฑ์เครื่องสำอางได้โดยต้องมีสารที่สอดคล้องในสูตรตำรับเดิมอยู่ในลักษณะของคู่มือ ทำให้การค้นหาและตรวจสอบข้อมูลที่ต้องการทำได้ยากและใช้เวลานาน

2. การขาดความคล่องตัวในการใช้และเรียกดูข้อมูล เนื่องจากข้อมูลมีลักษณะเป็นคู่มือ อีกทั้งยังมีการจัดเก็บข้อมูลหลายแหล่งที่กระจัดกระจาย เช่น อยู่ในเว็บไซต์ของกลุ่มควบคุมเครื่องสำอางของสำนักงานคณะกรรมการ

อาหารและยา แอปพลิเคชันไลน์ (LINE) เอกสารประกอบ การประชุมหรืออบรม เป็นต้น ทำให้ไม่สามารถเรียกดูข้อมูล ทั้งหมดได้ทันที

3. เนื่องจากข้อมูลเกี่ยวกับชื่อผลิตภัณฑ์ เครื่องสำอางและสารที่เป็นส่วนประกอบในตำรับมีการ ปรับปรุงอยู่ตลอด ทำให้ข้อมูลที่มีอยู่อาจไม่เป็นปัจจุบัน ส่งผลให้การพิจารณาคำขอจดแจ้งเครื่องสำอางมีความ คลาดเคลื่อนได้

4. อาจเกิดความคลาดเคลื่อนในการพิจารณาจาก การตรวจสอบข้อมูลโดยพนักงานเจ้าหน้าที่ เนื่องจาก กระบวนการทำงานเดิมที่อาศัยความรู้ ประสบการณ์ และ ความจำเป็นในการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอาง

5. ถึงแม้ว่าจะมีระบบจดแจ้งเครื่องสำอางอัตโนมัติ e-submission ช่วยตรวจสอบคำห้ามใช้ในชื่อผลิตภัณฑ์ เครื่องสำอางได้ แต่การตรวจสอบดังกล่าวสามารถตรวจได้ เฉพาะคำห้ามใช้ที่มีการสะกดเหมือนกับคำห้ามใช้ใน ฐานข้อมูลทุกประการ หากมีการสะกดคำผิด ทั้งที่ไม่ได้ ตั้งใจ หรือเป็นความตั้งใจของผู้ประกอบการในการ หลีกเลี่ยงคำห้ามใช้ หรือต้องการใช้ชื่อที่มีความเป็น เอกสิทธิ์ ระบบ e-submission อาจไม่สามารถตรวจสอบ ได้ ส่งผลให้การตรวจสอบและพิจารณาชื่อผลิตภัณฑ์ เครื่องสำอางมีความคลาดเคลื่อนได้

การพัฒนาโปรแกรมฯ

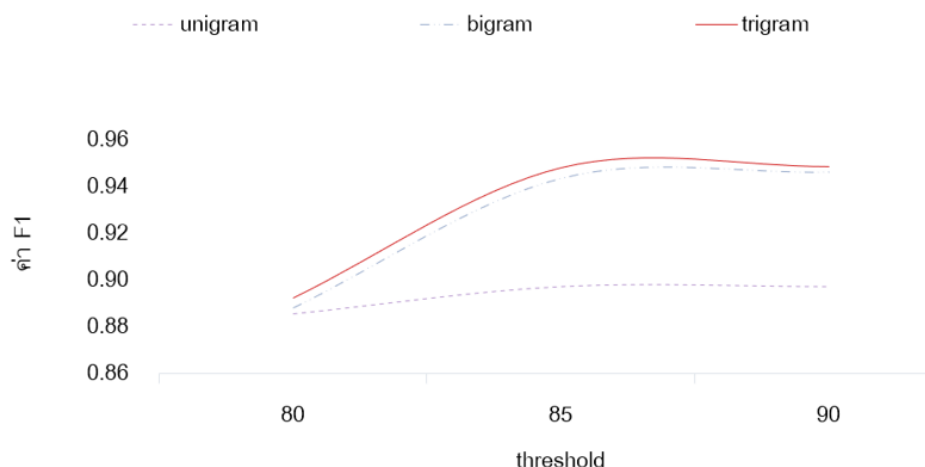
การทำงานของโปรแกรมฯ ประกอบด้วย 2 ส่วน หลัก ได้แก่ การตรวจสอบคำห้ามใช้และการตรวจสอบความ สอดคล้องกันระหว่างชื่อผลิตภัณฑ์เครื่องสำอางกับ ส่วนประกอบในตำรับ ในการใช้งานโปรแกรมฯ ผู้ใช้งาน

สามารถกรอกข้อมูลที่ต้องการตรวจสอบ ได้แก่ ชื่อการค้า ภาษาไทย ชื่อการค้าภาษาอังกฤษ ชื่อเครื่องสำอาง ภาษาไทย ชื่อเครื่องสำอางภาษาอังกฤษ และสารที่เป็น ส่วนประกอบในตำรับเครื่องสำอาง ผลลัพธ์จากการ ตรวจสอบข้อมูลจะแสดงผลการตรวจสอบคำขอจดแจ้ง เครื่องสำอางเป็น 4 กรณี ดังนี้ 1) ผ่าน (ไม่พบคำห้ามใช้ และมีความสอดคล้องกันระหว่างชื่อกับส่วนประกอบใน ตำรับเครื่องสำอาง) 2) ไม่ผ่าน (ไม่พบคำห้ามใช้ แต่ชื่อกับ ส่วนประกอบในตำรับเครื่องสำอางไม่สอดคล้องกัน) 3) ไม่ ผ่าน (พบคำห้ามใช้ แต่มีความสอดคล้องกันระหว่างชื่อกับ ส่วนประกอบในตำรับเครื่องสำอาง) และ 4) ไม่ผ่าน (พบคำ ห้ามใช้ และชื่อกับส่วนประกอบในตำรับเครื่องสำอางไม่ สอดคล้องกัน)

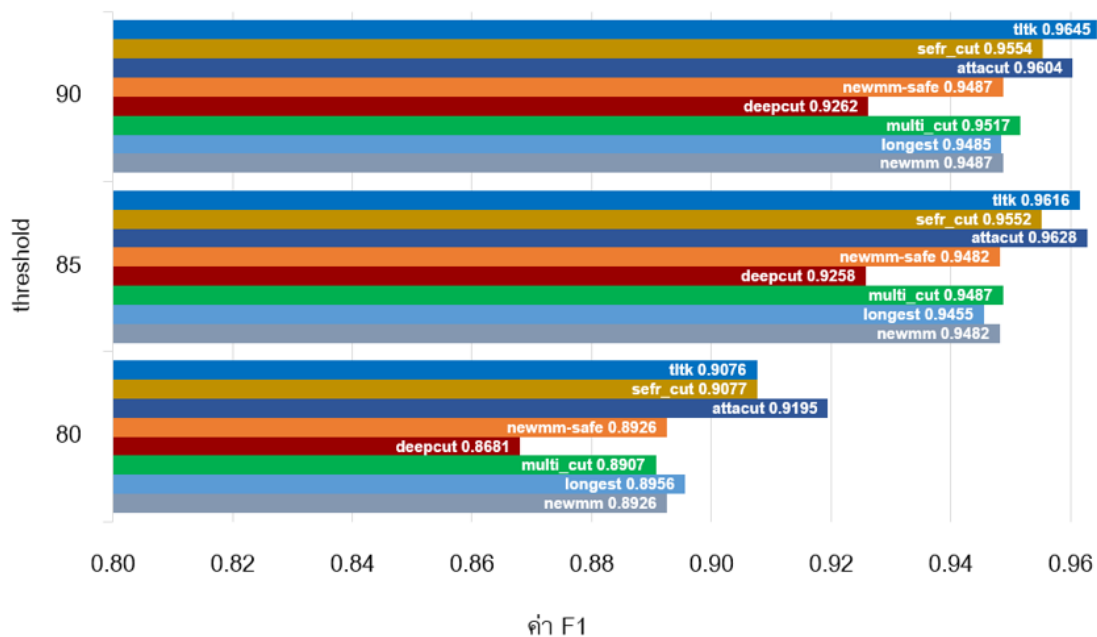
การประเมินประสิทธิภาพของโปรแกรม

ในการประเมินประสิทธิภาพของโปรแกรมฯ โดย กำหนดค่า threshold เริ่มต้นที่ 80 และใช้การเปรียบเทียบ คุณลักษณะของคำรูปแบบ unigram, bigram และ trigram พบว่า คุณลักษณะของคำรูปแบบ trigram มีค่า F1 สูงสุด เท่ากับ 0.8926 รองลงมาคือแบบ bigram และ unigram โดยมีค่า F1 เท่ากับ 0.8883 และ 0.8855 ตามลำดับ (รูปที่ 5)

เพื่อหาค่า threshold ที่เหมาะสมที่สุดที่ทำให้การ เปรียบเทียบสายอักขระให้มีประสิทธิภาพสูงสุด ผู้วิจัยได้ ปรับค่า threshold เป็น 85 และ 90 ผลการทดสอบพบว่า คุณลักษณะของคำรูปแบบ trigram ยังคงมีค่า F1 สูงสุด และมีค่าเพิ่มขึ้น เป็น 0.9482 และ 0.9487 ตามลำดับ (รูปที่ 5)



รูปที่ 5. ผลการเปรียบเทียบค่า F1 ของคำแต่ละคุณลักษณะที่ค่า threshold 80, 85 และ 90



รูปที่ 6. ผลการเปรียบเทียบประสิทธิภาพของโปรแกรมโดยพิจารณาจากค่า F1 เมื่อใช้เครื่องมือตัดคำ 8 วิธี

ผู้วิจัยจึงได้เลือกใช้คุณลักษณะคำรูปแบบ trigram เพื่อให้การเปรียบเทียบสายอักขระครอบคลุมคุณลักษณะของคำรูปแบบต่าง ๆ ทั้งคำเดี่ยวและกลุ่มคำตั้งแต่สองคำขึ้นไป เป็นการเพิ่มความถูกต้องของคำ และเพิ่มประสิทธิภาพในการเปรียบเทียบสายอักขระ

ในการเปรียบเทียบการตัดคำภาษาไทยโดยใช้เครื่องมือการตัด 8 วิธี ในการศึกษาเทียบกับการตรวจสอบโดยเจ้าหน้าที่ผู้เชี่ยวชาญ ผู้วิจัยใช้คุณลักษณะคำรูปแบบ trigram และกำหนดค่า threshold ความคล้ายคลึงกันของคำที่ 80, 85 และ 90 จากผลการศึกษาพบว่า ที่ค่า threshold 85 การตัดคำภาษาไทยด้วยวิธี attacut มีค่า F1 มากที่สุดอยู่ที่ 0.9628 และที่ค่า threshold 90 การตัดคำภาษาไทยด้วยวิธี titk มีค่า F1 มากที่สุดอยู่ที่ 0.9645 (รูปที่ 6)

ผลจากโปรแกรมฯ และผลจากผู้เชี่ยวชาญ

การเปรียบเทียบผลการประเมินโดยโปรแกรมฯ และเจ้าหน้าที่ผู้เชี่ยวชาญพบว่า ที่ค่า threshold 85 การใช้เครื่องมือ attacut ในการตัดคำภาษาไทย ทำให้การทำงานของโปรแกรมฯ มีประสิทธิภาพมากที่สุด และที่ค่า threshold 90 การใช้เครื่องมือ titk ในการตัดคำภาษาไทย ทำให้การทำงานของโปรแกรมฯ มีประสิทธิภาพมากที่สุด ผลการประเมินโดยโปรแกรมฯ และเจ้าหน้าที่ผู้เชี่ยวชาญในทั้งสองกรณีไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ($P=0.607$ และ 0.179 ตามลำดับ) (ตารางที่ 2)

confusion matrix และค่าประสิทธิภาพของโปรแกรม ที่ค่า threshold 85 และ 90 โดยใช้ attacut และ titk ในการตัดคำ แสดงอยู่ในตารางที่ 3

การอภิปรายผล

การตรวจสอบข้อผิดพลาดที่เครื่องสำอองด้วยวิธีการเปรียบเทียบสายอักขระแบบประมาณโดยการหา LD ร่วมกับการใช้เทคนิค n-gram และการกำหนดความคล้ายคลึงกันของคำที่ค่า threshold 80, 85 และ 90 ตามลำดับ พบว่า การใช้เทคนิค trigram ทำให้ผลลัพธ์ที่ได้

ตารางที่ 2. การเปรียบเทียบผลการตรวจสอบข้อผิดพลาดที่เครื่องสำอองโดยโปรแกรมฯ และโดยเจ้าหน้าที่ผู้เชี่ยวชาญ ($n = 400$ รายการ)

ผลการตรวจ โดยผู้เชี่ยวชาญ	การตรวจโดยโปรแกรมฯ		P ¹
	ผ่าน	ไม่ผ่าน	
threshold 85, engine = attacut			
ผ่าน	194	6	0.607
ไม่ผ่าน	9	191	
threshold 90, engine = titk			
ผ่าน	190	10	0.180
ไม่ผ่าน	4	196	

1: McNemar's chi-square, binomial distribution

ตารางที่ 3. confusion matrix และค่าประสิทธิภาพของโปรแกรม ที่ค่า threshold 85 และ 90 โดยใช้ attacut และ titk ในการตัดคำ

engine		confusion matrix		ค่าประสิทธิภาพของโปรแกรม			
				accuracy	precision	recall	F1
threshold 85	attacut	TP = 194	FN = 6	0.9525	0.9557	0.9700	0.9628
		FP = 9	TN = 191				
	titk	TP = 188	FN = 12	0.9625	0.9843	0.9400	0.9616
		FP = 3	TN = 197				
threshold 90	attacut	TP = 194	FN = 6	0.9600	0.9510	0.9700	0.9604
		FP = 10	TN = 190				
	titk	TP = 190	FN = 10	0.9650	0.9794	0.9500	0.9645
		FP = 4	TN = 196				

มีประสิทธิภาพมากขึ้น ดังนั้นในงานวิจัยนี้จึงได้เลือกใช้เทคนิค trigram ในการกำหนดคุณลักษณะของคำ ซึ่งสอดคล้องกับงานวิจัยของ แกมกาญจน์ สมประเสริฐศรี (22) ที่ศึกษาการสกัดวลีสำคัญภาษาไทยอัตโนมัติโดยใช้เทคนิคเอ็นแกรมร่วมกับการเรียนรู้ของเครื่องจักร พบว่าจำนวนแกรมที่เหมาะสมในการสกัดวลีที่คาดว่าเป็นวลีสำคัญ คือจำนวนแกรมตั้งแต่ 1-3 แกรม

การใช้โปรแกรมฯ ในการตรวจสอบเกิดผลลบ (FN) และผลบวก (FP) ผู้วิจัยตรวจสอบ FN และ FP ดังกล่าวอีกครั้งพบว่า FN และ FP ที่เกิดขึ้นนั้นเป็น TP และ TN สาเหตุการเกิด FN ได้แก่ 1) ผลการตัดคำภาษาไทยได้คำห้ามใช้ หรือคำที่ใกล้เคียงคำห้ามใช้ เช่น คำว่า “เลเซอร์” กับ “เลเซอร์” ซึ่งมีความคล้ายคลึงกันร้อยละ 87.5, คำว่า “ดีเอส” เป็นคำทับศัพท์ของการนำด้วยย่อภาษาอังกฤษมาใช้ เมื่อผ่านการตัดคำแล้วได้คำว่า “ดี” และ “เอส” ซึ่งคำว่า “ดี” เป็นคำห้ามใช้ในภาษาไทย 2) ผลการตัดคำภาษาไทยได้คำที่สามารถนำไปใช้ในชื่อการค้าและชื่อเครื่องสำอางได้โดยต้องมีสารที่สอดคล้องในสูตรตำรับ เช่น คำว่า “ซิเคนเจอร์” ผ่านขั้นตอนการตัดคำแล้วได้คำว่า “ซิ” และ “เนเจอร์” ซึ่งคำว่า “เนเจอร์” เป็นคำที่สามารถนำไปใช้ในชื่อการค้าและชื่อเครื่องสำอางได้โดยต้องมีสารที่สอดคล้องในสูตรตำรับ เมื่อไม่พบสารในตำรับ จึงทำให้ผลการตรวจสอบคลาดเคลื่อนจาก “ผ่าน” เป็น “ไม่ผ่าน” 3) คำภาษาอังกฤษที่เป็นคำพ้องรูปกับคำห้ามใช้ เช่น การใช้ข้อความ EAU DE PARFUM หรือ EAU DE TOILETTE เป็นชื่อการค้าหรือชื่อเครื่องสำอาง ซึ่งข้อความดังกล่าวสื่อ

ถึงประเภทของน้ำหอม พนักงานเจ้าหน้าที่ประเมินให้ข้อความดังกล่าวผ่านการตรวจสอบ แต่การตรวจสอบด้วยโปรแกรมในขั้นตอนการตัดคำจะได้คำว่า “DE” ซึ่งเป็นคำห้ามใช้ในภาษาอังกฤษ จึงทำให้โปรแกรมประเมินผลเป็น ไม่ผ่านการตรวจสอบ

สาเหตุการเกิดผลบวก (FP) ได้แก่ การใช้คำที่สามารถนำไปใช้ในชื่อการค้าและชื่อเครื่องสำอางได้โดยต้องมีสารที่สอดคล้องในสูตรตำรับ ซึ่งคำภาษาไทยบางคำเขียนติดกันโดยไม่มีเว้นวรรค เช่น “ฮันนี่โกลด์” เมื่อพนักงานเจ้าหน้าที่ตรวจสอบพบว่าข้อความดังกล่าวประกอบด้วยคำจำนวน 2 คำ คือ “ฮันนี่” และ “โกลด์” ซึ่งทั้งสองคำเป็นคำที่สามารถใช้เป็นชื่อการค้าหรือชื่อเครื่องสำอางได้ โดยต้องมีสารที่สอดคล้องในสูตรตำรับ แต่ในสูตรตำรับไม่พบสารที่มีความสอดคล้องกับคำข้างต้น พนักงานเจ้าหน้าที่จึงประเมินเป็น ไม่ผ่านการตรวจสอบ แต่ในการตรวจสอบด้วยโปรแกรมฯ การใช้เครื่องมือตัดคำภาษาไทยบางอย่างไม่สามารถตัดคำภาษาไทยดังกล่าวได้ และได้นำคำว่า “ฮันนี่โกลด์” ไปเปรียบเทียบกับคำที่มีในฐานข้อมูล แต่ไม่พบว่ามี ความคล้ายคลึงกับคำว่า “ฮันนี่” และ “โกลด์” เลย จึงไม่มีการไปเทียบหาสารที่เกี่ยวข้องในสูตรตำรับ ทำให้โปรแกรมแสดงผลการพิจารณาเป็นผ่านการตรวจสอบ

ในการกำหนดค่า threshold หรือความคล้ายคลึงกันของคำต้องเป็นค่าที่เหมาะสม หากกำหนดค่า threshold ที่สูงเกินไป ทำให้การเปรียบเทียบสายอักขระแบบประมาณ (approximate string matching) มีความใกล้เคียงกับการเปรียบเทียบสายอักขระที่เหมือนกันทุกประการ (exact

คำที่ใช้ในการเปรียบเทียบ	คำในฐานข้อมูล	จำนวนอักขระของ สายอักขระที่ยาวที่สุด	ค่าความคล้ายคลึงกัน ของคำ (%)
W H I T E R N I N G	W H I T E N I N G	10	90
แ อ ล ก อ ฮ ร ี	แ อ ล ก อ ฮ ล ี	9	89
C O L L A G E L	C O L L A G E N	8	88
A L G O H O L	A L C O H O L	7	86
L O V E	C L O V E	5	80

รูปที่ 7. ตัวอย่างการเปรียบเทียบสายอักขระและค่าความคล้ายคลึงกันของคำ

string matching) ส่งผลให้ไม่ได้ประโยชน์จากการเปรียบเทียบสายอักขระโดยการหา LD ยกตัวอย่าง กรณีชื่อผลิตภัณฑ์เครื่องสำอางมีคำว่า “ALGOHOL” เป็นส่วนหนึ่งของชื่อ เมื่อเปรียบเทียบกับคำในฐานข้อมูลพบว่า คล้ายคลึงกับคำว่า “ALCOHOL” ซึ่งเป็นคำที่สามารถใช้เป็นชื่อผลิตภัณฑ์เครื่องสำอางได้โดยต้องมีสารที่เกี่ยวข้องในสูตรตำรับ ทั้งสองคำต่างกันที่การสะกดคำด้วยตัวอักษร G แทน C ซึ่งอาจเป็นการสะกดผิดโดยที่ไม่ได้ตั้งใจ หรือเป็นความตั้งใจของผู้ประกอบการเพื่อให้ชื่อผลิตภัณฑ์เครื่องสำอางมีความเป็นเอกลักษณ์ โดดเด่น แตกต่างจากผลิตภัณฑ์อื่น จากการเปรียบเทียบสายอักขระพบว่า ทั้งสองคำมีความยาวสายอักขระเท่ากับ 7 เมื่อเปรียบเทียบสายอักขระโดยการหา LD แทนที่ G ด้วย C 1 ตำแหน่ง ทั้งสองคำจึงมีความคล้ายคลึงกันเท่ากับร้อยละ 86 ซึ่งหากกำหนดค่า threshold ที่ 85 จะพบว่าทั้งสองคำมีความคล้ายคลึงกัน จากนั้นโปรแกรมจะเทียบหาสารในตำรับที่เกี่ยวข้องกับชื่อ ALCOHOL ต่อไป แต่หากกำหนดค่า threshold ที่ 90 โปรแกรมจะไม่สามารถตรวจสอบได้ว่า ALGOHOL มีความคล้ายคลึงกับคำว่า ALCOHOL ในฐานข้อมูล ทำให้ไม่สามารถไปเทียบหาสารในตำรับที่เกี่ยวข้องกับชื่อ ALCOHOL ได้ ส่งผลให้เกิดความคลาดเคลื่อนของการตรวจสอบและพิจารณาชื่อผลิตภัณฑ์เครื่องสำอางได้

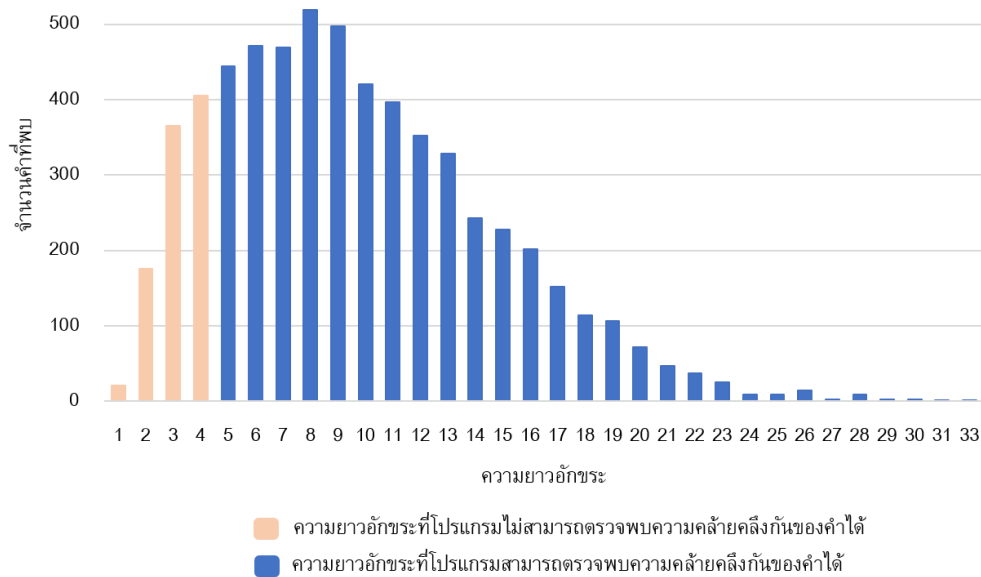
ส่วนการกำหนดค่า threshold ที่ต่ำเกินไป ทำให้การเปรียบเทียบสายอักขระตรวจพบคำที่มีความคล้ายคลึงกันกับคำในฐานข้อมูลจำนวนมาก หากมีความคล้ายคลึงกับคำห้ามใช้ จะทำให้แสดงผลการพิจารณาเป็นไม่ผ่านการตรวจสอบ หรือหากพบว่ามีค่าความคล้ายคลึงกับคำที่สามารถใช้เป็นชื่อผลิตภัณฑ์เครื่องสำอางได้ โดยต้องมีสารที่สอดคล้องในสูตรตำรับ โปรแกรมจะไปทำการเทียบหาสาร

ในตำรับที่เกี่ยวข้องกับชื่อนั้น ๆ หากไม่พบสารในตำรับ จะแสดงผลการพิจารณาเป็นไม่ผ่านการตรวจสอบ ซึ่งอาจทำให้ผลการตรวจสอบและพิจารณาชื่อผลิตภัณฑ์เครื่องสำอางคลาดเคลื่อนได้เช่นกัน

จากการเปรียบเทียบสายอักขระโดยการหา LD เมื่อทำการตัดออก แทรก หรือแทนที่อักขระ 1 ครั้ง จะมีค่าเท่ากับ 1 และเมื่อนำไปคำนวณหาความคล้ายคลึงกันของคำ ซึ่งจากสูตรการคำนวณจะเป็นการหารค่าระหว่าง LD ต่อความยาวสูงสุดของสายอักขระใดสายอักขระหนึ่งจากสายอักขระสองชุด ดังนั้นหากคำมีความยาวของสายอักขระที่สั้นเมื่อมีการเปลี่ยนแปลงเพียง 1 ตำแหน่ง ในการเปรียบเทียบสายอักขระจะมีผลทำให้ความคล้ายคลึงกันของคำลดลงมากกว่าคำที่มีสายอักขระยาวกว่า (รูปที่ 7)

ดังนั้น ปัจจัยที่มีผลต่อการเปรียบเทียบสายอักขระแบบประมาณอีกประการหนึ่งที่สำคัญ คือ ความยาวของสายอักขระ คำแต่ละคำมีความยาวของสายอักขระที่แตกต่างกัน คำที่พบในงานวิจัยนี้มีความยาวอักขระตั้งแต่ 1 จนถึง 33 อักขระ (รูปที่ 8) และจากข้อมูลที่น่าเสนอข้างต้นแสดงให้เห็นว่า การกำหนดค่า threshold ที่ 80 คำที่มีความยาวในช่วง 5-6 อักขระ เมื่อมีการเปลี่ยนแปลง 1 ตำแหน่ง โปรแกรมสามารถตรวจพบความคล้ายคลึงกันของคำได้ เช่นเดียวกับการกำหนดค่า threshold ที่ 85 และ 90 ส่งผลให้โปรแกรมสามารถตรวจพบความคล้ายคลึงกันของคำที่มีความยาวในช่วง 7-9 อักขระ และตั้งแต่ 10 อักขระขึ้นไปตามลำดับ ส่วนคำที่มีความยาวอักขระในช่วง 1-4 อักขระ โปรแกรมไม่สามารถตรวจพบความคล้ายคลึงกันของคำได้

ในงานวิจัยนี้ คำที่พบส่วนใหญ่เป็นคำที่มีความยาวอักขระ 7-9 คำและตั้งแต่ 10 อักขระขึ้นไป โดยมีจำนวน 1486 และ 2,777 คำ ตามลำดับ ดังนั้นผลการเปรียบเทียบประสิทธิภาพของโปรแกรมฯ จึงพบว่า ที่ค่า threshold 90



รูปที่ 8. ข้อมูลความยาวอักขระของคำที่พบในงานวิจัย

โปรแกรมมีประสิทธิภาพสูงสุดในการเปรียบเทียบสายอักขระ โดยมีค่า F1 สูงสุดเมื่อเทียบกับที่ค่า threshold 5 และ 80

แม้ว่าการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในแบบคำขอจดแจ้งเครื่องสำอางโดยใช้โปรแกรมฯ จะให้ผลลัพธ์ที่ดี แต่ยังไม่สามารถใช้ทดแทนเจ้าหน้าที่ได้อย่างสมบูรณ์ เนื่องจากอาจมีข้อมูลชื่อการค้า ชื่อเครื่องสำอาง หรือส่วนประกอบในตำรับที่เป็นข้อมูลใหม่ หรือข้อมูลที่มีความซับซ้อน เช่น คำพ้องรูป คำพ้องเสียง โปรแกรมอาจตรวจสอบคลาดเคลื่อนได้ ซึ่งยังคงต้องให้พนักงานเจ้าหน้าที่ต้องพิจารณาซ้ำอีกครั้งหนึ่ง แต่โปรแกรมฯ สามารถชี้เป้าแบบคำขอจดแจ้งที่มีความเสี่ยงในการใช้ชื่อการค้าหรือชื่อเครื่องสำอางที่ไม่เป็นไปตามหลักเกณฑ์การพิจารณาการจดแจ้งเครื่องสำอาง ช่วยให้พนักงานเจ้าหน้าที่ตรวจสอบได้อย่างถูกต้องและรวดเร็วมากยิ่งขึ้น

ข้อเสนอแนะ

คำห้ามใช้และคำที่สามารถใช้เป็นชื่อผลิตภัณฑ์เครื่องสำอางได้ในฐานข้อมูลส่วนหนึ่งมีรูปแบบเป็นคำทับศัพท์ภาษาอังกฤษ ซึ่งวิธีการสะกดคำทับศัพท์ภาษาอังกฤษไม่มีรูปแบบที่ตายตัว จึงควรมีนักภาษาศาสตร์เข้ามาช่วยในการตรวจสอบความถูกต้องของคำก่อนเก็บค่าลงในฐานข้อมูล เพื่อช่วยให้ผลการตรวจสอบและพิจารณาคำขอจดแจ้งเครื่องสำอางมีความสมบูรณ์มากยิ่งขึ้น

การศึกษานี้เก็บรวบรวมแบบคำขอจดแจ้งเครื่องสำอางที่ยื่นเข้าระบบ e-submission ของจังหวัด

พระนครศรีอยุธยาเพียงแห่งเดียว และมีขนาดตัวอย่างแบบคำขอจดแจ้งเครื่องสำอางน้อย ซึ่งอาจมีความหลากหลายไม่เพียงพอที่จะเป็นตัวแทนของประชากรแบบคำขอจดแจ้งเครื่องสำอางทั่วประเทศได้ ดังนั้นก่อนนำโปรแกรมไปใช้ในวงกว้างจึงควรมีการทดสอบในตัวอย่างแบบคำขอจดแจ้งเครื่องสำอางในจำนวนที่มากขึ้นและทดสอบในวงกว้าง ซึ่งอาจพบประเด็นที่ต้องปรับปรุงเพิ่มเติมหรือทำให้ทราบประสิทธิภาพที่แท้จริงของโปรแกรมได้

เนื่องจากความยาวของสายอักขระมีผลต่อการคำนวณค่าความคล้ายคลึงกันของคำและการเลือกค่า threshold ที่เหมาะสม ดังนั้นสำหรับการศึกษาวิจัยในอนาคต ผู้วิจัยควรเพิ่มการทำงานของโปรแกรมให้มีการตรวจสอบความยาวของสายอักขระก่อน จากนั้นจึงเลือกค่า threshold ที่เหมาะสมในการเปรียบเทียบความคล้ายคลึงกันของคำ เพื่อให้โปรแกรมทำงานได้อย่างมีประสิทธิภาพมากขึ้น ครอบคลุมการเปรียบเทียบสายอักขระของคำที่มีความยาวอักขระที่แตกต่างกัน

สรุป

โปรแกรมที่พัฒนาขึ้นสามารถตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในแบบคำขอจดแจ้งเครื่องสำอางได้ตามวัตถุประสงค์ จากผลการศึกษารูปได้ว่า คุณลักษณะของคำที่เหมาะสมกับงานวิจัยนี้ที่สุดคือ คุณลักษณะของแบบ trigram นอกจากนี้ยังพบว่า ความยาวอักขระมีผลต่อการกำหนดค่า threshold ของความคล้ายคลึงกันของคำ

โดยหากค่าส่วนใหญ่มีความยาวอักขระอยู่ในช่วง 7-9 อักขระ การกำหนดค่า threshold ที่ 85 จะทำให้โปรแกรมสามารถตรวจพบความคล้ายคลึงกันของคำได้ แต่ถ้าค่ามีความยาวสายอักขระเพิ่มขึ้น ตั้งแต่ 10 อักขระขึ้นไป การกำหนดค่า threshold ที่ 90 จะทำให้โปรแกรมสามารถตรวจพบความคล้ายคลึงกันของคำได้มากกว่า

สำหรับงานวิจัยนี้การกำหนดค่า threshold ที่ 90 ร่วมกับใช้เครื่องมือ tsk ในการตัดคำ จะเหมาะสมกับการใช้งาน ช่วยให้การตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางและการพิจารณาคำขอจดแจ้งเครื่องสำอางมีประสิทธิภาพมากที่สุด

การวิเคราะห์ข้อมูลประสิทธิภาพของโปรแกรมพบว่า การตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางในแบบคำขอจดแจ้งเครื่องสำอางด้วยวิธีการเปรียบเทียบสายอักขระแบบประมาณ และการตรวจสอบโดยเจ้าหน้าที่ผู้เชี่ยวชาญไม่แตกต่างกัน ผลการศึกษาชี้ให้เห็นว่า การตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางด้วยวิธีการที่นำเสนอ ทำให้ผลการตรวจสอบชื่อผลิตภัณฑ์เครื่องสำอางและการพิจารณาคำขอจดแจ้งเครื่องสำอางมีความถูกต้อง เป็นไปในทิศทางเดียวกัน สามารถนำไปใช้เพื่อช่วยเจ้าหน้าที่ในการตรวจสอบข้อมูลในแบบคำขอจดแจ้งเครื่องสำอางเบื้องต้นได้

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณ เกษัชกรสุชาติ ถนอมวรารักษ์ หัวหน้ากลุ่มงานคุ้มครองผู้บริโภคฯ ตลอดจนเจ้าหน้าที่กลุ่มงานผู้บริโภคทุกท่านที่ได้ให้ความช่วยเหลือจนงานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดี งานวิจัยนี้ได้รับทุนอุดหนุนการวิจัยประเภทนักศึกษาระดับบัณฑิตศึกษาจากกองทุนวิจัยและสร้างสรรค์ คณะเภสัชศาสตร์ มหาวิทยาลัยศิลปากร

เอกสารอ้างอิง

1. Marketeer team. Beauty market in 2022 with beauty tech: a case study of L'Oreal [online]. 2022 [cited Sep 21, 2022]. Available from: marketeeronline.co/archives/266777.
2. Department of Trade Negotiation, Ministry of Commerce. Thailand cosmetic market: plans for preference utilization in FTAs to increase exports

[online]. 2021 [cited Sep 21, 2022]. Available from: www.dtn.go.th/th/news/607e50d7ef4140b573032f40?cate=5cff753c1ac9ee073b7bd1c5.

3. Cosmetic Act B.E. 2558. Royal Gazette No. 132, Part 86A (Sep 8, 2015).
4. Juntarawongpaisarn K, Ruengorn C. Situation of cosmetic notification on e-Submission and monitoring of Thai Food and Drug Administration, fiscal year 2016-2019. Thai Food and Drug Journal 2021; 28: 60-71.
5. Cosmetic Control Group, Food and Drug Administration. Thailand cosmetic regulation guidelines (updated 2021) [online]. 2021 [cited Jan 9, 2022]. Available from: www.fda.moph.go.th/sites/Cosmetic/PublishingImages/SitePages/Permission/29-6-65-Update.pdf.
6. Navarro G. A guided tour to approximate string matching. ACM Comput Surv 2001; 33: 31-88.
7. Watcharapinchai N, Rujikietgumjorn S. Approximate license plate string matching for vehicle re-identification. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS); 2017 Aug 29 - Sep 1; Lecce, Italy: IEEE; 2017. p. 1-6.
8. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl 1966; 10: 707-10.
9. Yulianto MM, Arifudin R, Alamsyah A. Autocomplete and spell checking Levenshtein distance algorithm to getting text suggest error data searching in library. Sci J Inform 2018; 51: 67-75.
10. Haldar R, Mukhopadhyay D. Levenshtein distance technique in dictionary lookup methods: an improved approach [online]. 2011 [cited Dec 15, 2021]. Available from: arxiv.org/ftp/arxiv/papers/1101/1101.1232.pdf
11. Santirattaphakdi C, Niwattanakul S. The design and development of public bus complaint classification process for service problem tagging. Journal of Engineering and Digital Technology 2021; 9: 77-91.

12. Paluekpet T. Applying Levenshtein's algorithm to find homograph and look-alike herb names [master thesis]. Nakhon Pathom: Silpakorn University; 2017.
13. Urathamakun P, Runapongsa K. Improved rule-based and new dictionary for Thai word segmentation. The 3rd Joint Conference on Computer Science and Software Engineering; 2006 June 28-30; Bangkok, Thailand. 2006. p. 34-40.
14. Ekwonganan A. Identification of Thai and transliterated words by n-gram models [master thesis]. Bangkok: Chulalongkorn University; 2005.
15. Connor RJ. Sample size for testing differences in proportions for the paired-sample design. *Biometrics*. 1987; 43: 207-11.
16. Aiumsiriwong O. System analysis and design. Bangkok: Se-Education; 2012.
17. Supaartagorn C. Web application development with PHP and MySQL+AJAX jQuery. 2nd ed. Bangkok: Simplify Press; 2018.
18. Wichitbunyarak P. HMTL: a web language. *Executive Journal* 2011; 3: 203-7.
19. Phatthiyaphaibun W, Chaovavanich K, Polpanumas C, Suriyawongkul A, Lowphansirikul L, Chormai P. PyThaiNLP: Thai natural language processing in python [online]. 2016 [cited Jun 27, 2022]. Available from: doi.org/10.5281/zenodo.3519354.
20. Parikh N, Singh G, Sundaresan N. Query suggestion with large scale data. In: Govindaraju V, Rao CR, editors. *Handbook of statistics machine learning : theory and applications*. Vol. 31. Amsterdam: Elsevier; 2013. p. 493-518.
21. Kumar G. Evaluation metrics for intrusion detection systems-a study. *Int J Comput Sci Mob Appl* 2014; 2: 11-7.
22. Somprasertsri G. Automatic Thai keyphrase extraction using n-gram and machine learning approach. *Journal of Information Science* 2014; 32: 111-28.