JIT Mahasarakham University, Thailand
http://jit.it.msu.ac.th

Journal of Applied Informatics and Technology, 5(2), 86-99

Research Article

การวัดความคล้ายคลึงของคำในระบบถามตอบภาษาไทยสำหรับโรคเบาหวาน

Measurement of Word Similarity for Diabetes Question Answering System

ธนพล ชำนาญหาญ[1], เกสรา เพชรกระจ่าง[1,*], สันติ สถิตวรรธนะ[1], พงศกร เจริญเนตรกุล[1], ชัยสิทธิ์ ชูสงค์[1]
Tanapol Chamnanhan[1], Ketsara Phetkrachang[1,*], Santi Sathiwantana[1],
Pongsagorn Chalearnnetkul[1], Chaisit Choosong[1]

[1] สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย สงขลา 90000 ประเทศไทย
[1] Computer Engineering, Faculty of Computer Engineering, Rajamangala University of Technology Srivijaya, Songkhla 90000, Thailand
* Corresponding Author: Ketsara Phetkrachang, ketsara.p@rmutsv.ac.th

Received:
4 March 2023
Revised:
7 April 2023
Accepted:
29 May 2023

คำสำคัญ:

ความคล้ายคลึง, โรคเบาหวาน, ระบบถาม-ตอบ

Keywords:

Similarity, Diabetes, Question Answering System

บทคัดย่อ: โรคเบาหวานเป็นโรคเรื้อรังที่ไม่สามารถรักษาให้หายขาดได้ และเป็นปัญหาสำคัญต่อการสาธารณสุขของประเทศไทยจากปัญหาดังกล่าว กรมควบคุมโรคคาดการณ์ว่าในปี 2568 คนไทยเป็นเบาหวานกว่า 7.41 ล้านคน ดังนั้น ผู้ป่วยโรคเบาหวานจึงต้องดูแลตนเองอย่างต่อเนื่องซึ่งเป็นวิธีการหนึ่งที่ช่วยลดอุบัติการณ์ของภาวะแทรกซ้อนที่เกิดขึ้นกับระบบต่างๆ ของร่างกายซึ่งส่งผลกระทบต่อชีวิตของผู้ป่วยได้ งานวิจัยฉบับนี้จึงนำเสนอการวัดความคล้ายคลึงของคำในระบบถาม-ตอบภาษาไทยสำหรับผู้เป็นโรคเบาหวานด้วยวิธี Cosine, Dice และ Jackcard โดยมีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบประสิทธิภาพในการหาคำตอบเพื่อประโยชน์ของประชาชนที่ต้องการทราบอาการเบื้องต้นของโรคเบาหวานและการดูแลตนเองของผู้ป่วยเบาหวาน ผลการวิจัยเบื้องต้นจากการเปรียบเทียบประสิทธิภาพการหาคำตอบด้วยวิธีการวัดความคล้ายคลึงของคำถาม-คำตอบ พบว่าวิธี Cosine มีประสิทธิภาพสูงสุดในการหาคำตอบด้วยค่าความแม่นยำ 92.50% รองลงมาคือ วิธี Jaccard และวิธี Dice ซึ่งมีค่าความแม่นยำ 80.28% และ 52.50 % ตามลำดับ

Abstract: Diabetes is a chronic disease that cannot be cured and is a major problem for the public health of Thailand. The Department of Disease Control predicts that by 2025 there will be more than 7.41 million people in Thailand with diabetes. Continuous self-care for people with diabetes is one method that helps to reduce the incidences of complications arising in already

compromised body systems affecting the lives of patients. This research, therefore, presents a measure of the similarity of words in Thai question-answering systems for diabetes by using Cosine, Dice and Jaccard methods to compare the effectiveness of finding answers for the benefit of people who want to know about the initial symptoms of diabetes and self-care for people with diabetes. The preliminary results from the study comparing answer finding efficiency using the question-answer similarity measurement methods found that Cosine was the most effective in finding answers with a precision value of 92.50%, followed by Jaccard and Dice which had precision values of 80.28% and 52.50% respectively.

1. บทนำ

โรคเบาหวานเป็นโรคเรื้อรังที่เป็นปัญหาสำคัญทางสาธารณสุขของประเทศ และไม่สามารถที่จะรักษาให้หายขาดได้ องค์การอนามัยโลกคาดว่าประมาณปี พ.ศ. 2568 จะมีผู้ป่วยเป็นโรคเบาหวานทั่วโลกประมาณ 300 คน และประเทศไทย กรมควบคุมโรคคาดการณ์ว่าในปี พ.ศ. 2568 จะพบผู้ป่วยเป็นโรคเบาหวานมากกว่า 7.4 ล้านคน (Kitreerawutiwong & Tejativaddhana, 2013) ทั้งนี้โรคเบาหวานเป็นโรคที่ผู้ป่วยต้องการความดูแลในเรื่องการรักษาอย่างต่อเนื่อง เพื่อลดอัตราการเกิดภาวะแทรกซ้อนในหลายระบบของร่างกาย ส่งผลกระทบต่อการดำรงชีวิต ภาวะเศรษฐกิจของผู้ป่วยและครอบครัว รวมทั้งประเทศชาติ หัวใจสำคัญของการจัดการโรงเบาหวานคือ การค้นหาโรคตั้งแต่ระยะเริ่มแรก และการดูแลรักษา เพื่อชะลอการเกิดภาวะแทรกซ้อน ทั้งผู้ป่วยและครอบครัวควรได้รับความรู้ รวมทั้งข้อมูลที่เกี่ยวข้องอย่างเพียงพอ เพื่อให้เกิดการเรียนรู้ และมีการปรับเปลี่ยนพฤติกรรมสุขภาพที่เหมาะสม เพื่อควบคุมระดับ น้ำตาลในเลือดให้เป็นไปตามเป้าหมายการรักษา กรมการแพทย์ ซึ่งเป็นกรมวิชาการของกระทรวงสาธารณสุขมีภารกิจในการพัฒนาองค์ความรู้ และเทคโนโลยีทางการแพทย์เพื่อสนับสนุนต่อการพัฒนาคุณภาพการบริการแก่หน่วยงาน และสถานบริการสุขภาพ ดังนั้นจึงร่วมดำเนินการปรับปรุงแนวทางเวชปฏิบัติสำหรับโรคเบาหวาน และวารสารทางการแพทย์สำหรับโรคเบาหวานเพื่อให้แพทย์ และบุคลากรทางการแพทย์มีความรู้ความเข้าใจ และมีทักษะในการตรวจวินิจฉัย วางแผนการรักษา และฟื้นฟูสมรรถภาพเบื้องต้น รวมถึงการส่งต่อไปรับการรักษาที่ถูกต้องเป็นไปในแนวทางเดียวกันทั่วประเทศ อย่างเหมาะสมต่อไป จึงเป็นภาระทางด้านสาธารณสุขที่สำคัญของประเทศ

จากปัญหาดังกล่าว งานวิจัยนี้จึงนำเสนอการวัดความคล้ายคลึงของคำใน ระบบถาม-ตอบภาษาไทย สำหรับผู้เป็นเบาหวาน ในการถาม-ตอบจะใช้ภาษาธรรมชาติ โดยรับข้อมูลจาก ผู้ใช้ เช่น ผู้ป่วย ผู้ดูแลผู้ป่วย แพทย์ พยาบาล บุคลากรทางการแพทย์ เจ้าหน้าที่ภายในสถาบัน บุคคลทั่วไปสามารถที่จะตั้งคำถาม ข้อสงสัยเกี่ยวกับการดูแลรักษา และสอบถามข้อแนะนำต่างๆ ในการดูแลรักษาโรคเบาหวานเบื้องต้นได้ ซึ่งเป็นแนวทางหนึ่งที่ ช่วยบริการด้านการให้ความรู้ รวมถึงข้อมูลที่เกี่ยวข้องอย่างเพียงพอ เพื่อให้เกิดการเรียนรู้ และมีการปรับเปลี่ยนพฤติกรรมสุขภาพที่เหมาะสม เป็นการแบ่งเบาภาระทางด้านสาธารณสุขของประเทศ ระบบการดูแลรักษาผู้ป่วยโรคเบาหวานในลักษณะการถาม-การตอบจึงเป็นเครื่องมือส่งเสริมคุณภาพการบริการด้านสุขภาพแบบใหม่ที่เหมาะสม และเกิดประโยชน์สูงสุดต่อประชาชน เพื่อช่วยให้ผู้เป็นโรคเบาหวานมีคุณภาพชีวิตที่ดี รู้วิธีการรักษาได้ทันท่วงทีก่อนไปพบแพทย์ โดยแนวทางการถาม-การตอบ อาศัยองค์ความรู้ จากแนวทางเวชปฏิบัติสำหรับโรคเบาหวานของสมาคมต่อมไร้ท่อแห่งประเทศไทย ข้อมูลการดูแลรักษาโรคเบาหวานจากสำนักงานกองทุนสนับสนุนการสร้างเสริมสุขภาพ

(สสส.) ข้อมูลการดูแลรักษาโรคเบาหวานจากสมาคมโรคเบาหวานแห่งประเทศไทยในพระบรมราชูปถัมภ์ฯ สถาบันวิจัย และประเมินเทคโนโลยีทางการแพทย์ กรมการแพทย์ กระทรวงสาธารณสุข สำนักงานหลักประกันสุขภาพแห่งชาติ และข้อมูลอื่นๆ

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

Xie *et al.* (2018) ได้นำเสนอระบบถาม-ตอบและแจ้งเตือนล่วงหน้าบนมือถือเพื่อช่วยในการจัดการโรคเบาหวานแบบเรียลไทม์ผ่านอุปกรณ์ในการประเมินความเสี่ยงและแจ้งเตือนล่วงหน้าสำหรับการตรวจสอบสภาวะสุขภาพของผู้เป็นเบาหวานผ่านระบบ Dia-AID เพื่อช่วยเหลือผู้ป่วยที่มีความเสี่ยงสูง

Mutabazi *et al.* (2021) ได้นำเสนอ การทบทวนระบบการตอบคำถามที่เป็นข้อความทางการแพทย์ตามแนวทางการเรียนรู้ในรูปแบบการเรียนรู้เชิงลึก (Deep Learning) เพื่อเป็นแนวทางในการวัดประสิทธิภาพในการดึงข้อมูลสำคัญทางอินเตอร์เน็ตเป็นการตรวจสอบระบบถามตอบที่เป็นข้อความทางการแพทย์ เพื่อหาค่าความแม่นยำ ซึ่งเป็นการวิเคราะห์ตามแนวทางการเรียนรู้เชิงลึก

Wongsara, Homjun & Ketui (2021) ได้นำเสนอการวิจัยเรื่องการพัฒนาระบบตรวจข้อสอบอัตนัยภาษาไทยบนพื้นฐานการวัดค่าความคล้ายคลึง (Similarity) มีวัตถุประสงค์เพื่อ 1) พัฒนาระบบตรวจข้อสอบอัตนัยภาษาไทยบนพื้นฐานการวัดค่าความคล้ายคลึง และ 2) ประเมินผลการใช้ระบบตรวจข้อสอบอัตนัยภาษาไทยบนพื้นฐานการวัดค่าความคล้ายคลึง ผลการวิจัยพบว่าระบบมีความพึงพอใจอยู่ที่ระดับ มาก โดยมีค่าเฉลี่ยที่ 4.05 และมีค่าความเบี่ยงเบนมาตรฐาน (S.D.) ที่ 0.8

Ditcharoen & Techawiwatthanaboon (2018) ได้นำเสนองานวิจัยเรื่อง การเปรียบเทียบคำอธิบายรายวิชาเพื่อการเทียบการโอนหน่วยกิตของรายวิชาระดับมหาวิทยาลัยโดยใช้วิธีวัดความคล้ายคลึงของคำ และแบบจำลองเวกเตอร์สเปซ (Vector Space Model) โดยมีกระบวนการ 5 ขั้นตอน 1) การตัดคำไทยภาษาไทย 2) การกำจัดคำหยุด 3) การวิเคราะห์คำเชิงความหมาย 4) การหาค่าน้ำหนักของคำ และ 5) การวัดความคล้ายคลึงของรายวิชา ผลการวิจัย พบว่า วิธีการเปรียบเทียบคำอธิบายรายวิชาเพื่อการเทียบโอนหน่วยกิตที่พัฒนาขึ้น ได้ค่าความถูกต้อง เท่ากับ 82.61% ค่าความแม่นยำ (Precision) เท่ากับ100% ค่าความระลึก (Recall) 80.95 และค่าความถ่วงดุล (F-measure) เท่ากับ 89.47% ซึ่งสามารถนำไปประยุกต์ใช้ในการตรวจสอบความคล้ายคลึงของคำอธิบายรายวิชาเพื่อการเทียบโอนหน่วยกิตได้

Phetkrachang, Sathiwantanah, & Kongwan (2022) ได้นำเสนอ การพัฒนาระบบถาม-ตอบออนไลน์สำหรับเว็บบริการงานทะเบียนนักศึกษาของมหาวิทยาลัยด้วยเทคโนโลยีออนโทโลยี (Ontology) ซึ่งมีวัตถุประสงค์เพื่อสร้างต้นแบบระบบถามตอบ และใช้ค่าความแม่นยำ และค่าความระลึกในการประเมินประสิทธิภาพคำตอบ ผลการทดลองเบื้องต้นระบบมีความแม่นยำอยู่ที่ 90.91% ค่าความระลึก 83.33%

2.2 ทฤษฎีที่ใช้ที่ใช้ในการวิจัย

2.2.1 ระบบถาม-ตอบ

ระบบถาม-ตอบ เป็นระบบที่มีวัตถุประสงค์เพื่อถาม-ตอบในสิ่งที่ต้องการ หรือเป็นการประมวลผลข้อความที่รับประโยคคำถามจากผู้ใช้ โดยการอาศัยคำสำคัญที่อยู่ในประโยค เป็นการรับคำขอจากผู้ใช้ในรูปประโยคคำถามที่ เป็นภาษามนุษย์ และได้รับผลลัพธ์ที่เป็นคำตอบที่กระชับรวบรัด อาจเป็นคำตอบโดยตรง

กับเอกสาร หรือเป็นข้อความสั้นๆ ที่ไม่ใช่ข้อความทั้งเอกสาร (Radev *et al.*, 2001)

โดยทั่วไปสถาปัตยกรรมของระบบถาม-ตอบประกอบด้วย

1) Question Processing มีหน้าที่ในการทำความเข้าใจ กับคำถาม ว่าคำถามนั้นต้องการคำตอบอะไร โดยทำการสร้างคำขอ เพื่อใช้สำหรับการค้นหาเอกสารจากคลังข้อมูล โดยใช้คำสำคัญ (Keyword) ที่อยู่ในคำถาม

2) Document Processing เป็นขั้นตอนการประมวลผลเอกสารเพื่อเป็นคำตอบ หรือ แปลงข้อมูลให้อยู่ในรูปแบบใด รูปแบบหนึ่ง ก่อนทำการเปรียบเทียบระหว่างคำถาม และคำตอบของเอกสารที่อยู่ในคลังเอกสาร เพื่อให้การค้นคืนคำตอบมีประสิทธิภาพมากยิ่งขึ้น เช่น การทำดัชนี

3) Q-A Matching เป็นการเปรียบเทียบคำถาม กับคำตอบในคลังเอกสาร ซึ่งมีหลากหลายวิธีที่ใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างคำถาม และคำตอบที่เป็นไปได้ เช่น การวิเคราะห์เชิงภาษาศาสตร์ การวิเคราะห์ความคล้ายคลึง เป็นต้น

4) Answer Generation เป็นการสร้างคำตอบที่เป็นไปได้ทั้งหมด ซึ่งจะได้ผลลัพธ์เป็นข้อความ ในข้อความนั้นจะมีคำตอบที่แท้จริงอยู่ ซึ่งจะถูกสกัดออกมา และส่งผลลัพธ์ให้กับผู้ใช้

2.2.2 การตัดคำภาษาไทย (Thai Word Segmentation)

การตัดคำภาษาไทย เป็นการแบ่งข้อความหรือสายอักขระของตัวอักษรที่ต่อเนื่องกันออกเป็นหน่วยของคำ เพื่อหาขอบเขตของคำ ซึ่งลักษณะของภาษาไทยจะเขียนติดกันโดยไม่มีการใช้เครื่องหมายวรรคตอนใดๆ การตัดคำเพื่อให้ได้ขอบเขตของคำที่ถูกต้องโดยการให้คอมพิวเตอร์สามารถคำนวณหรือแบ่งอักขระไทยออกเป็นคำๆ วิธีการตัดคำภาษาไทยสามารถแบ่งได้เป็น 3 หลักการใหญ่ๆ คือ

1) หลักการตัดคำโดยใช้กฎ Rule Based Approach โดย Sornlertlamvanich *(1993)* กล่าวไว้ว่า การตัดคำโดยใช้กฎเป็นความพยายามในขั้นเริ่มต้นของการพัฒนาระบบตัดคำภาษาไทย โดยที่ได้เสนอวิธีการตรวจสอบกฎเกณฑ์ทางอักขระ วิธีที่กำหนดลักษณะของการประสมอักษร การเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการบ่งชี้ขอบเขตของคำ

2) หลักการตัดคำโดยใช้พจนานุกรม (Dictionary Approach) การตัดคำโดยใช้พจนานุกรมเป็นแนวคิดที่ได้รับการพัฒนาในยุคต่อมา โดยเก็บคำภาษาไทยไว้ในพจนานุกรม แล้วนำข้อความที่ป้อนเข้า (Input) ไปค้นหา และเทียบสายอักขระกับคำในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวโดยหาขอบเขตของคำ และประกอบด้วยคำใดบ้าง การนำพจนานุกรมมาใช้ในการตัดคำภาษาไทยจะมีการทำงานอยู่ 2 ขั้นตอน คือ ขั้นตอนแรกจะทำการตัดคำโดยเทียบระหว่างข้อความ กับคำในพจนานุกรม และขั้นตอนที่สองจะทำการเปรียบเทียบระหว่างคำที่ได้ในขั้นตอนแรก กับคำในพจนานุกรมอีกครั้งหนึ่ง เพื่อหาคำที่สามารถตัด วิธีการนี้จะสิ้นเปลืองทรัพยากรหน่วยความจำหลักค่อนข้างมาก

3) การตัดคำโดยใช้คลังข้อมูล เป็นการตัดคำโดยวิธีทางสถิติเข้ามาใช้ในการประมวลผลทางภาษา โดยใช้ฐานความรู้ในการตัดคำเพื่อแก้ปัญหาของคำที่ไม่มีในพจนานุกรม เช่น ชื่อเฉพาะ คำที่คลุมเครือ ซึ่งการตัดคำลักษณะนี้ อาจใช้วิธีการตัดคำโดยอาศัยความน่าจะเป็น หรือ การตัดคำโดยใช้คุณลักษณะของคำ

2.2.3 การกรองคำหยุด (Stop Word Elimination)

การกรองคำหยุด เป็นการนำคำที่ไม่มีนัยสำคัญออก โดยที่ไม่ทำให้ความหมายของคำในเอกสารเปลี่ยนแปลงไป คำที่ไม่มีนัยสำคัญในที่นี้หมายถึง

คำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลงไป ตัวอย่างประเภทของคำที่จัดว่าเป็นคำหยุดในภาษาไทย เช่น ประเภทคำบุพบท ได้แก่ ของ ใน แก่ แด่ ต่อ เมื่อ แค่ ฯลฯ ประเภทคำสันธาร ได้แก่ เพราะ และ แต่ ทั้ง ฯลฯ ประเภทคำสรรพนาม ได้แก่ คุณ ท่าน เธอ ฉัน เรา และประเภทคำวิเศษณ์ ได้แก่ มาก น้อย ใหญ่ เล็ก

2.2.4 การสกัดคุณลักษณะของเอกสารด้วย tf-idf

การสกัดคุณลักษณะของเอกสารด้วย tf-idf (Salton & Buckley, 1988) เป็นการรวมวิธีการของ tf กับ idf มารวมกันในการคำนวณ ดังสมการที่ 1

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t \quad (1)$$

วิธีการนี้เป็นการคำนวนค่า น้ำหนักของคำที่อยู่ในเอกสาร d ซึ่ง 1) ถ้า ปรากฏขึ้นบ่อยครั้งในเอกสารแสดงว่า เป็นคำที่เป็นตัวแทนเอกสารเหล่านั้นได้ ซึ่งค่า จะมี น้ำหนักสูง 2) ถ้าค่า t ปรากฏน้อยครั้งในเอกสาร แสดงว่า เป็นคำที่ไม่สำคัญไม่สามารถเป็นตัวแทนเอกสารได้ จะมี น้ำหนักต่ำ 3) ถ้า ปรากฏในทุกๆ เอกสาร จะทำให้ มีค่า น้ำหนักต่ำ สรุปได้ว่าเป็นคำที่ไม่สำคัญ และไม่สามารถเป็นตัวแทนเอกสารได้ ส่วนค่า มาจากสมการดังนี้

$$idf_t = log \frac{N}{df_t} \quad (2)$$

โดยที่

idf คือ ค่าส่วนกลับความถี่ของเอกสาร

N คือ จำนวนเอกสารในชุดเอกสารทั้งหมด

df คือ ค่าความถี่ของเอกสาร (Document Frequency) ของคำแต่ละคำ ถ้าคำใดปรากฏบนเอกสารทุกฉบับจะมีค่าเท่ากับ ซึ่งทำให้ค่าส่วนกลับความถี่ของเอกสารมีค่าเท่ากับศูนย์ในการให้ค่าน้ำหนักคำมีขั้นตอนดังต่อไปนี้

ขั้นตอนการคำนวณการให้ค่า น้ำหนักคำมีดังนี้

1) ให้ความสำคัญ กับความถี่ของคำที่ปรากฏอยู่ในเอกสาร และความถี่มีผลต่อการให้ น้ำหนักของคำในเอกสาร โดยที่ การนับความถี่ของคำ (Term Frequency : TF) คือ การใช้ความถี่ของคำ เช่น พบ 1 ครั้งเรียกว่า เทอม (Term) ทั้งนี้ขึ้นอยู่ กับจำนวนคำของเอกสาร โดยเทอมจะแทนคำศัพท์ของแต่ละคำ ค่า น้ำหนักคำ (Term Weight : W) ความถี่ของคำๆ หนึ่งที่พบในทุกๆ เอกสาร

2) การสร้างตารางความถี่ของคำเป็นขั้นตอนในการหาความถี่ของคำศัพท์ และกำหนดค่า น้ำหนักของคำศัพท์ในแต่ละเอกสารลงในตารางความถี่ของคำ

3) การนำเสนอข้อมูลในเชิงเวกเตอร์ ซึ่งคำศัพท์ในตารางความถี่ของคำจะถูกนำเสนอในเชิงเวกเตอร์ โดยจะถูกมองเป็นอาเรย์ของเวกเตอร์ เช่น (3, 1, 0, 0, 0)

4) การคำนวณค่าความเหมือนโดยคุณสมบัติของเวกเตอร์ทำให้สามารถคำนวณค่าความคล้ายคลึงระหว่างเวกเตอร์ได้จากค่าสัมประสิทธิ์โคไซน์ของมุมระหว่างคู่เวกเตอร์ ซึ่งจะมีค่าอยู่ระหว่าง 0 ถึง 1

5) สามารถจัดลำดับ (Ranking) ของเอกสารโดยใช้เกณฑ์ความสำคัญของคำ และการเข้ากันได้ของคำ

2.2.5 การวัดความคล้ายคลึงโดยใช้วิธี Cosine

เทคนิคการวัดความคล้ายคลึงแบบโคไซน์ (Cosine Similarity) (Senoussaoui *et al.*, 2014)

เป็นเทคนิคการหาความคล้ายคลึงเชิงมุม โดยใช้เวกเตอร์สเปซในการวัดค่าความคล้ายคลึง เป็นการหาความคล้ายคลึงกันจากค่าความต่างของมุมของวัตถุ 2 อย่างที่เกิดขึ้นบนเวกเตอร์ ซึ่งความคล้ายคลึงแบบโคไซน์จะมีค่าอยู่ระหว่าง 0 ถึง 1 เท่านั้น วิธีการนี้เป็นที่นิยม และมีประสิทธิภาพสูง ในการหาความคล้ายคลึงของวัตถุ 2 อย่าง เช่น ความคล้ายคลึงของคำที่อยู่คำถามและความคล้ายคลึงของคำที่อยู่ในเอกสารคำตอบถ้าเป็นระบบถาม-ตอบ ซึ่งตัวหารมีชื่อเรียกเฉพาะว่าระยะห่างยูคลีเดียน (Euclidean Distance) ซึ่งวิธีการนี้จะมีประสิทธิภาพในกรณีที่เอกสาร 2 เอกสารมีความยาวไม่เท่ากัน หรือ ทำให้มีความยุติธรรมต่อเอกสารที่สั้นกว่า

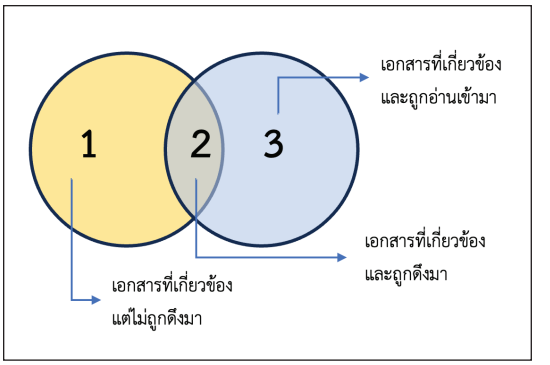
2.2.6 การวัดความคล้ายคลึงโดยใช้วิธี Jaccard

การวัดความคล้ายคลึงโดยใช้ Jaccard มักใช้กับข้อมูลที่เป็น Unary ซึ่งประกอบด้วยค่า 0 และ 1 ใช้เลขฐานสองแทนค่าดรรชนีของเอกสาร จะใช้กับชุดคำเฉพาะแต่ละประโยค ถูกกำหนดให้เป็นความแตกต่างระหว่างจุดตัดกันนำมาเปรียบเทียบกันโดยการพิจารณาตามขนาดข้อความของข้อมูลทั้งสองชุด ที่นำมาเปรียบเทียบกันมีผลให้ ดรรชนีตัวที่ i จะแทนด้วยเลข 0 หรือ 1 ในตำแหน่งที่ I ตามลำดับ ซึ่งค่า 1 นั้นจะแสดงถึงคำสำคัญอยู่ในเอกสาร สำหรับ 0 หมายถึงคำสำคัญไม่ได้อยู่ในเอกสารในตำแหน่งที่ระบุ (Jaccard, 1901) แสดงดังสมการที่ 3

$$d(a,b) = \frac{\sum_{i=1}^{n} a_i . b_i}{\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} b_i^2 - \sum_{i=1}^{n} a_i . b_i} \quad (3)$$

2.2.7 การวัดความคล้ายคลึงโดยใช้วิธี Dice coefficient

วิธีวัดความคล้ายคลึงของเอกสารแบบ Dice เป็นการวัดความคล้ายของเอกสาร 2 เอกสารโดยมีการแทนด้วยเวกเตอร์ของ น้ำหนักคำที่ปรากฏในเอกสาร



ภาพประกอบ 1 แสดงความสัมพันธ์ระหว่างกลุ่มของข้อมูลที่ใช้สืบค้น กับข้อสอบถาม

แล้วนำค่าของเวกเตอร์ทั้งสองเอกสารมาเปรียบเทียบกันโดยเพิ่มค่าเป็นสองเท่าแล้วหารด้วยผลบวกของผลรวมค่าของเวกเตอร์ทั้งสองเอกสาร (Kondrak, Marcu & Knight, 2003)

2.2.8 การวัดประสิทธิภาพของระบบด้วย Confusion Matrix

Frankes & Baeza-Yates (1992) กล่าวไว้ว่า การวัดประสิทธิภาพของระบบมีอยู่หลายวิธี แต่มีสองวิธีที่นิยมตามมาตรฐานของระบบค้นคืนเอกสารคือ การใช้การวัดค่าความแม่นยำ (Precision) ของข้อมูล และค่าความระลึก (Recall) ของข้อมูล

Kondrak, Marcu & Knight (2003) กล่าวไว้ว่า การวัดค่าความแม่นยำของข้อมูล และการวัดค่าความถูกต้องของข้อมูลนั้นเป็นวิธีการหนึ่งสำหรับการประเมินประสิทธิภาพการค้นคืนเอกสาร

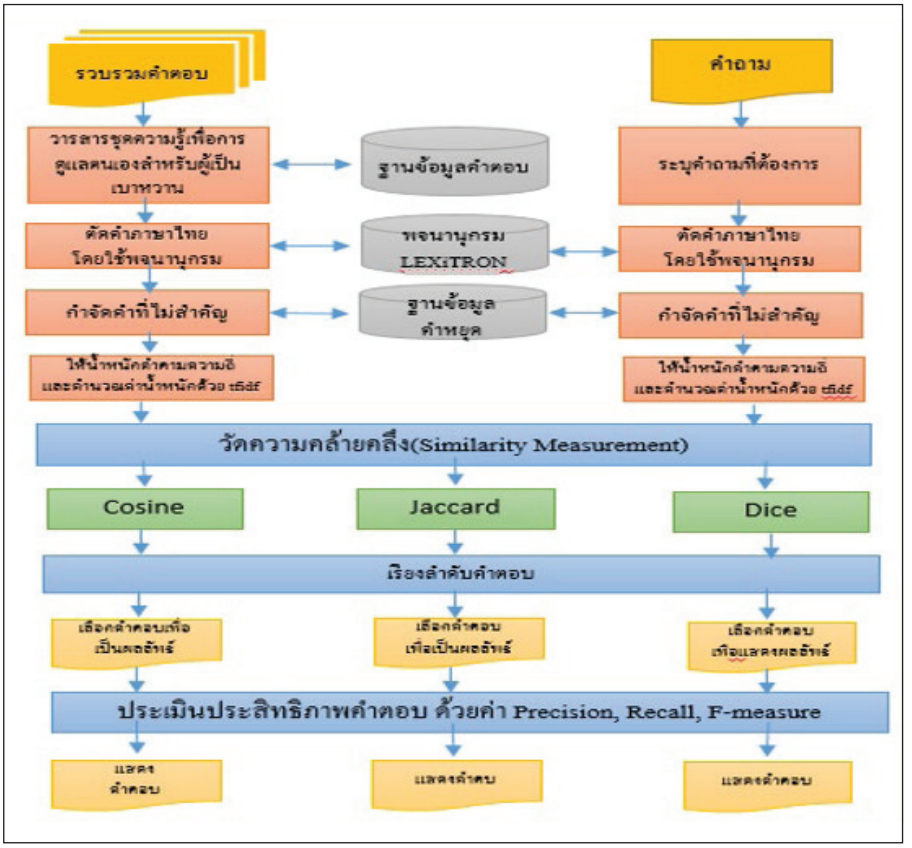
จากภาพประกอบ 1 จะแสดงความสัมพันธ์ระหว่างกลุ่มของข้อมูลที่ใช้สืบค้นกับข้อสอบถาม โดยที่

(ก) กลุ่มของเอกสารที่เกี่ยวข้องกับข้อคำถามแต่ไม่ได้เลือกขึ้นมา คือส่วนหมายเลข 1

(ข) กลุ่มของเอกสารที่เกี่ยวข้องกับข้อคำถามและถูกเลือกขึ้นมาคือส่วนหมายเลข 2

(ค) กลุ่มของเอกสารที่ไม่เกี่ยวข้องกับข้อคำถามแต่ถูกเลือกขึ้นมาคือส่วนหมายเลข 3

2.3 กรอบแนวคิดการทำงานวิจัย



ภาพประกอบ 2 กรอบแนวคิดการวิจัย

3. วิธีการดำเนินการวิจัย

งานวิจัยนี้มุ่งเน้นในการหาความคล้ายคลึงของคำที่อยู่ในคำถามเปรียบเทียบ กับข้อมูลที่อยู่ในคลังคำตอบ โดยอาศัยการวัดความคล้ายคลึง ซึ่งผู้วิจัยทำการทดลอง 3 วิธี ประกอบด้วย การวัดความคล้ายคลึงโดยใช้วิธี Cosine, Jaccard และ Dice ซึ่งผู้วิจัยได้นำเสนอกรอบแนวคิดดังภาพประกอบ 1 ประกอบด้วยขั้นตอนดังต่อไปนี้

3.1 การรวบรวมข้อมูลเพื่อเป็นคำตอบ

การวิจัยในครั้งนี้ ผู้วิจัยรวบรวมข้อมูล โดยอาศัยองค์ความรู้จากแนวทางเวชปฏิบัติสำหรับโรคเบาหวาน ของสมาคมต่อมไร้ท่อแห่งประเทศไทย (Wongsara, Homjun & Ketui, 2021) ข้อมูลการดูแลรักษาโรคเบาหวานจากสำนักงานกองทุนสนับสนุนการสร้างเสริมสุขภาพ (สสส.) ข้อมูลการดูแลรักษาโรคเบาหวานจากสมาคมโรคเบาหวานแห่งประเทศไทยในพระบรมราชูปถัมภ์ฯ สถาบันวิจัย และประเมินเทคโนโลยีทางการแพทย์ กรมการแพทย์ กระทรวงสาธารณสุข สำนักงานหลักประกันสุขภาพแห่งชาติ และข้อมูลอื่นๆ โดย แบ่งเอกสารออกเป็นชุดๆ ซึ่งแต่ละชุดจะทำการหาความสัมพันธ์ระหว่างคำ และหมายเลขเอกสาร ในการทดลองครั้งนี้ ผู้วิจัยจะใช้เอกสารจำนวน 300 ชุด (d1, d2, ..., d300) แสดงดังตาราง 1

ตาราง 1 ตัวอย่างข้อมูล

Id	Description
d1	อาการใจสั่นหวิวๆ มือสั่น หิว เป็นอาการของ น้ำตาลในเลือดต่ำ ดังนั้น จึงควรดื่ม น้ำหวาน หรือกินอาหาร ถ้าทำได้ แต่ถ้าไม่สะดวก เช่น เวลาเดินทาง อาจอมลูกกวาด น้ำตาลก้อน ควรเตรียมอาหารว่างติดตัวไปด้วย และควรมีบัตรประจำตัว ที่บอกว่าตัวเองเป็นเบาหวาน และบอกวิธีการช่วยเหลือเมื่อมีปัญหา น้ำตาลในเลือดต่ำ อย่างไรก็ตาม ถ้ามีอาการเช่นนี้ บ่อยๆ ควรปรึกษาหมอในกรปรับยาเบาหวานให้เหมาะสม
...	
...	
d300	

3.2 การรวบรวมข้อมูลคำถาม

ในการรวบรวมคำถาม ผู้วิจัยเก็บข้อมูลจากคำถามที่ถามบ่อย (Frequency Asked Question: FAQ) และเป็นคำถามที่ได้จากวารสารชุดความรู้เพื่อการดูแลตนเองสำหรับผู้เป็นเบาหวาน และแหล่งข้อมูลที่กล่าวถึงในหัวข้อ 3.1 โดยใช้จำนวน 30 คำถาม ในการทดสอบประสิทธิภาพของการตอบ ตัวอย่างคำถาม เช่น

คำถาม: ถ้ามีอาการใจสั่นหวิวๆ มือสั่น หิว ควรทำอย่างไร

3.3 นำข้อมูลในส่วนของคำถาม และคำตอบ ไปตัดคำภาษาไทย

การตัดคำภาษาไทย ใช้วิธีการตัดคำแบบพจนานุกรมโดยอาศัยคลังคำศัพท์ ทั้งนี้ผู้วิจัยเลือกใช้โปรแกรมทีเล็กส์ (TLex: Thai Lexeme Analyser) ซึ่งเป็นโปรแกรมตัดคำภาษาไทยที่พัฒนาด้วยเทคนิคการเรียนรู้ของเครื่อง พัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์ และคอมพิวเตอร์แห่งชาติ (NECTEC) (Sornlertlamvanich, 1993) จากนั้น ทำการกำจัดคำที่ไม่สำคัญออก หรืออาจเรียกว่า คำหยุด เช่น คำว่า และให้ การ เช่น ไว้ ฯลฯ ตัวอย่าง เช่น

คำถาม: ถ้ามีอาการใจสั่นหวิวๆ มือสั่น หิว ควรทำอย่างไร

ผลการตัดคำ: ถ้า|มี|อาการใจ|สั่น|หวิว| |ๆ| |มือ|สั่น| |หิว| |ควร|ทำ|อย่าง|ไร|

ผลการตัดคำในเอกสารคำตอบ: อาการใจ|สั่น|หวิว| |ๆ| |มือ|สั่น| |หิว| |เป็น|อาการ|ของ| น้ำตาล|ใน|เลือด|ต่ำ| |ดัง|นั้น| |จึง|ความ|ดื่ม| | น้ำ|หวาน|หรือ|กิน|อาหาร|ถ้า|ทำ|ได้| |แต่|ถ้า|ไม่|สะดวก| |เช่น| |เวลา|เดินทาง| |อาจอม|ลูกกวาด| | น้ำตาล|ก้อน| |ควร|เตรียม|อาหาร|ว่าง|ติด|ตัว|ไป|ด้วย| |และ|ควร|มี|บัตร|ประจำ|ตัว| |ที่|บอก|ว่า|ตัว|เอง|เป็น|เบาหวาน| |และ| |บอก|วิธี|การ|ช่วยเหลือ|เมื่อ|มี|ปัญหา| น้ำตาล|ใน|เลือด|ต่ำ| |อย่าง|ไร|ก็ตาม| |ถ้า|มี|อาการ|เช่น|นี้| |บ่อย| |ๆ| |ควร|ปรึกษา|หมอ|ใน|กร|ปรับ|ยา|เบาหวาน|ให้|เหมาะสม|

3.4 การคำนวณ น้ำหนักคำ

ในการคำนวณค่าน้ำหนักคำ *tf - idf* จะเป็นการใช้ค่า เป็นพื้นฐานโดยการรวมวิธีการของ *tf* และ *idf* เข้าด้วยกันเพื่อคำนวณค่า น้ำหนัก ซึ่ง

ใช้เป็นสมการของ Salton & Buckey (1988) ซึ่งแสดงดังสมการที่ 1

วิธีการนี้เป็นการคำนวนค่าน้ำหนักของคำ t ที่อยู่ในเอกสาร d ซึ่ง 1) ถ้า t ปรากฏขึ้นบ่อยครั้งในเอกสารแสดงว่า t เป็นคำที่เป็นตัวแทนเอกสารเหล่านั้นได้ ซึ่งค่า t จะมีน้ำหนักสูง 2) ถ้าค่า t ปรากฏน้อยครั้งในเอกสาร แสดงว่า t เป็นคำที่ไม่สำคัญไม่สามารถเป็นตัวแทนเอกสารได้ t จะมีน้ำหนักต่ำ 3) ถ้า t ปรากฏในทุกๆ เอกสาร จะทำให้ t มีค่าน้ำหนักต่ำ สรุปได้ว่าเป็นคำที่ไม่สำคัญ และไม่สามารถเป็นตัวแทนเอกสารได้ ส่วนค่า idf มาจากสมการที่ 2

โดยที่

idf คือ ค่าส่วนกลับความถี่ของเอกสาร

N คือ จำนวนเอกสารในชุดเอกสารทั้งหมด

df คือ ค่าความถี่ของเอกสาร (Document Frequency ของคำแต่ละคำ ถ้าคำใดปรากฏเอกสารทุกฉบับจะมีค่าเท่ากับ N ซึ่งทำให้ค่าส่วนกลับความถี่ของเอกสารมีค่าเท่ากับศูนย์ เมื่อรวมปัจจัยหรือกระบวนการทั้ง 2 อย่างนี้เข้าด้วยกัน สามารถนำไปคำนวณค่า น้ำหนักคำได้ดังสมการที่ 1

W_{td} คือ ค่า น้ำหนักของคำสำคัญที่ t ของเอกสาร d

tf_{td} คือ ความถี่ถ่วง น้ำหนักของคำสำคัญที่ t ของเอกสาร idf_j คือ ความถี่ผกผันของคำสำคัญ t

รูปแบบการทำงานกล่าวคือ

1) ให้ความสำคัญกับความถี่ของคำที่ปรากฏอยู่ในเอกสารและความถี่มีผลต่อการให้ น้ำหนักของคำในเอกสาร โดยที่

- การนับความถี่ของคำ (Term Frequency : TF) คือ การใช้ความถี่ของคำ เช่น พบ 1 ครั้งเรียกว่าเทอม (Term) ทั้งนี้ขึ้นอยู่กับจำนวนคำของเอกสาร โดยเทอมจะแทนคำศัพท์ของแต่ละคำ

- ค่า น้ำหนักคำ (Term Weight : W) ความถี่ของคำๆ หนึ่งที่พบในทุกๆ เอกสาร

2) การสร้างตารางความถี่ของคำเป็นขั้นตอนในการหาความถี่ของคำศัพท์ และกำหนดค่า น้ำหนักของคำศัพท์ในแต่ละเอกสารลงในตารางความถี่ของคำ

3) การนำเสนอข้อมูลในเชิงเวกเตอร์ ซึ่งคำศัพท์ในตารางความถี่ของคำจะถูกนำเสนอในเชิงเวกเตอร์ โดยจะถูกมองเป็นอาเรย์ของเวกเตอร์ เช่น (3, 1, 0, 0, 0)

4) การคำนวณค่าความคล้ายคลึงโดยคุณสมบัติของเวกเตอร์ทำให้สามารถคำนวณค่าความคล้ายคลึงระหว่างเวกเตอร์ได้จากค่า ซึ่งจะมีค่าอยู่ระหว่าง 0 ถึง 1

5) สามารถจัดลำดับ (Ranking) ของเอกสารโดยใช้เกณฑ์ความสำคัญของคำและการเข้ากันได้ของคำ

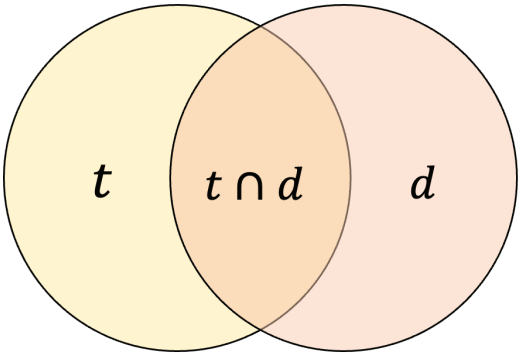
3.5 เปรียบเทียบประโยคคำถาม กับเอกสารคำตอบ

เป็นการประมวลผลคำถามเพื่อหาคำตอบโดยต้องความเข้าใจ กับคำถามว่าต้องการคำตอบอะไรซึ่งเป็นการสร้างคำขอเพื่อการค้นหาคำตอบจากเอกสารโดยใช้คำสำคัญ จากนั้นนำคำสำคัญที่ได้จากคำถามไปเปรียบเทียบ กับคำที่อยู่ในคลังคำตอบโดยผ่านคำนวณ น้ำหนักคำ

3.6 วัดความคล้ายคลึงของคำ

ในการวัดความคล้ายคลึงของคำในครั้งนี้ผู้วิจัยได้ทำการทดลอง 3 วิธี เพื่อวัดประสิทธิภาพคำตอบว่าวิธีการไหนได้ประสิทธิภาพคำตอบดีที่สุด โดยผู้วิจัยดำเนินการทดลอง วัดความคล้ายคลึงของคำ 3 วิธี ดังต่อไปนี้

3.6.1 การวัดความคล้ายคลึงโดยใช้ Cosine เป็นการวัดค่าความคล้ายคลึงโดยใช้



ภาพประกอบ 3 ภาพแสดงการ intersec ระหว่าง t และ d

เวกเตอร์สเปซในการวัดค่าความคล้ายคลึง เป็นการหาความคล้ายคลึงกันจากค่าความต่างของมุมของคำถาม และเอกสารคำตอบที่เกิดขึ้นบนเวกเตอร์ ซึ่งความคล้ายคลึงแบบ Cosine จะมีค่าอยู่ระหว่าง 0 ถึง 1 วิธีการนี้จะมีประสิทธิภาพในกรณีที่คำถาม กับเอกสารคำตอบมีความยาวไม่เท่ากัน หรือเป็นวิธีการทำให้มีความยุติธรรมต่อเอกสารที่สั้นกว่านั้นเอง ซึ่งแสดงดังสมการที่ 2

3.6.2 การหาความคล้ายคลึงโดยใช้ Jaccard ซึ่งจะขออธิบายโดยการใช้ภาพประกอบ 3

จากภาพประกอบ 3 อธิบายได้ว่า t คือคำที่อยู่ในคำถาม และ d คือคำที่อยู่ในเอกสารคำตอบ การหาค่าความคล้ายคลึงกันนั้น ก็คือ หากค่า t intersec d มีความหมายว่ามีความคล้ายคลึงกัน และค่าสูงสุดของ Jaccard คือ 1 ซึ่งจะเกิดขึ้นก็ต่อเมื่อ t intersec d มีค่าเท่ากับ t Union d ซึ่งแสดงดังสมการที่ 4

$$sim(t,d) = \frac{|t \cap d|}{|t \cup d|} = \frac{|A \cap d|}{|t| + |d| - |t \cap d|} \quad (4)$$

3.6.3 การวัดความคล้ายคลึงโดยใช้ Dice coefficient ซึ่งเป็นการคำนวณโดยการพิจารณาคุณลักษณะทั้งหมดของสองเวกเตอร์ แสดงดังสมการที่ 5

$$sim(t,d) = \frac{2\sum_{i=1}^{n} t_i \times d_i}{\sum_{i=1}^{n} t_i + \sum_{i=1}^{n} d_i} \quad (5)$$

t = เทอมของคำ, d = เอกสาร, i = ลำดับเอกสาร

3.7 การประเมินประสิทธิภาพคำตอบ

ในการประเมินผลการคำตอบ ผู้วิจัยเลือกใช้วิธีการประเมินประสิทธิภาพคำตอบโดยการใช้ค่าความแม่นยำ (Precision) แสดงดังสมการที่ 6 และค่าความระลึก (Recall) แสดงดังสมการที่ 7 เพื่อเปรียบเทียบจำนวนคำตอบทั้งหมด ซึ่งค่าความแม่นยำ หมายถึง สัดส่วนของจำนวนของคำตอบที่ระบบตอบถูกเปรียบเทียบกับจำนวนคำตอบทั้งหมดของระบบ ซึ่งเป็นการวัดค่าความสามารถของการขจัดคำตอบที่ไม่เกี่ยวข้องออกไป ส่วน ค่าความระลึกเป็นสัดส่วนของจำนวนคำตอบที่ระบบตอบถูกทั้งหมดเปรียบเทียบกับจำนวนคำตอบของผู้เชี่ยวชาญ หรือ เป็นการวัดความสามารถในการดึงคำตอบที่ถูกต้องจากเอกสารที่เกี่ยวข้องออกมา ส่วนค่าความถ่วงดุล (F-measure) หมายถึง ค่าเฉลี่ย ของความแม่นยำและค่าความระลึก ซึ่งเป็นค่าสัดส่วนระหว่างค่าความแม่นยำและค่าความระลึก แสดงดังสมการที่ 8

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FP} \quad (7)$$

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

TP (True Positive) หมายถึง สิ่งที่ตอบตรงกับสิ่งที่เกิดขึ้นจริง ในกรณี ตอบว่าจริง และสิ่งที่เกิดขึ้น ก็คือ จริง

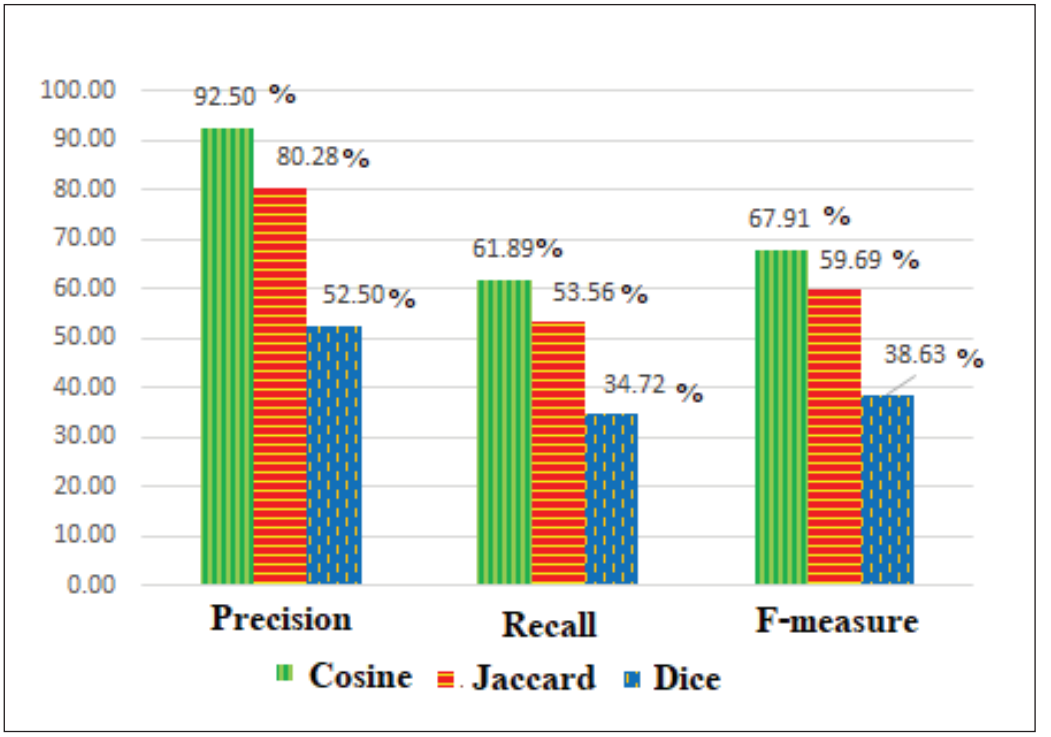
TN (True Negative) หมายถึง สิ่งที่ตอบตรงกับสิ่งที่เกิดขึ้น ในกรณี ตอบว่า ไม่จริง และสิ่งที่เกิดขึ้น ก็คือ ไม่จริง

FP (False Positive) หมายถึง สิ่งที่ตอบไม่ตรงกับสิ่งที่เกิดขึ้น คือตอบว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง

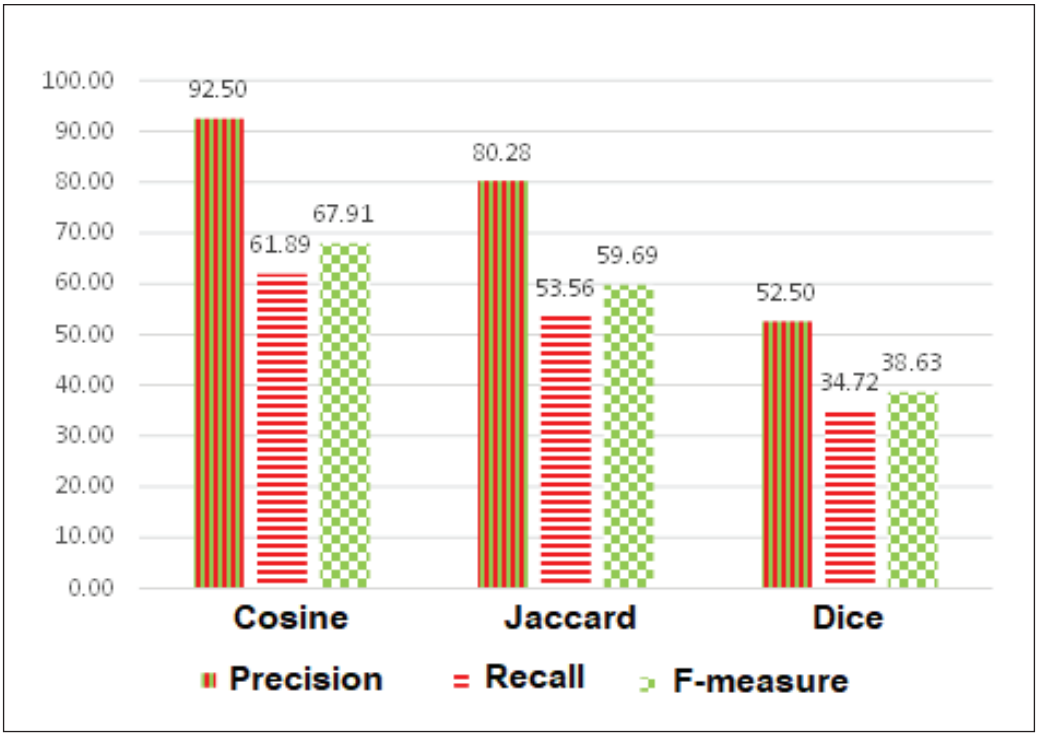
FN (False Negative) หมายถึง สิ่งที่ตอบไม่ตรงกับที่เกิดขึ้นจริง คือตอบว่าไม่จริง แต่สิ่งที่เกิดขึ้น คือ จริง

4. ผลการวิจัย

ข้อมูลที่นำมาทดลองประกอบด้วย คำถามจำนวน 30 คำถาม และเอกสารคำตอบจำนวน 300 ชุด ส่วนของคำถามคัดเลือกจากคำถามที่ถามบ่อย แนวทางการถาม-การตอบ อาศัยองค์ความรู้จาก แนวทางเวชปฏิบัติสำหรับโรคเบาหวาน ของสมาคมต่อมไร้ท่อแห่งประเทศไทย ข้อมูลการดูแลรักษาโรคเบาหวานจากสำนักงานกองทุนสนับสนุนการสร้างเสริมสุขภาพ (สสส.) ข้อมูลการดูแลรักษาโรคเบาหวานจาก สมาคมโรคเบาหวานแห่งประเทศไทยในพระบรมราชูปถัมภ์ฯ สถาบันวิจัย และประเมินเทคโนโลยีทางการแพทย์ กรมการแพทย์ กระทรวงสาธารณสุข สำนักงานหลักประกันสุขภาพแห่งชาติ และ ข้อมูลอื่นๆ การวิจัยในครั้งนี้ คณะผู้วิจัยได้ทำการประเมินความถูกต้องของการถามตอบของระบบถาม-ตอบภาษาไทยสำหรับผู้เป็นเบาหวาน โดยใช้ค่า Precision และ Recall ในการวัดค่าความสามารถในการตอบ เปรียบเทียบกับจำนวนคำตอบของผู้เชี่ยวชาญ ซึ่งในการทดลองครั้งนี้ ได้ทดลองด้วยวิธีการหาความคล้ายคลึงของคำที่อยู่ในคำถาม และเอกสารคำตอบ 3 วิธี ประกอบด้วย การหาความคล้ายคลึงของคำด้วยวิธี Cosine, Dice และ Jaccard เพื่อศึกษาเปรียบเทียบประสิทธิภาพการหาคำตอบ ผลการวิจัยเบื้องต้นพบว่าวิธี Cosine มีประสิทธิภาพในการหาคำตอบมากที่สุด และรองลงมาคือ Jaccard และ Dice ซึ่งมีค่าความถูกต้องอยู่ที่ 92.50%, 80.28% และ52.50% ตามลำดับ ดังภาพประกอบ 4 และ 5



ภาพประกอบ 4 แสดงประสิทธิภาพความถูกต้องของคำตอบด้วยวิธี Cosine, Jaccard และ Dice



ภาพประกอบ 5 การเปรียบเทียบประสิทธิภาพของการวัดค่าความคล้ายคลึงด้วยวิธี Cosine, Jaccard และ Dice

จากภาพประกอบ 4 และ 5 สามารถวิเคราะห์ได้ดังนี้

1.ค่า Precision หมายถึง ความสามารถในการดึงเอกสารคำตอบที่เกี่ยวข้อง และ ขจัดเอกสารที่ไม่เกี่ยวข้องออกไป วิธี Cosine จะได้ค่าประสิทธิภาพในการดึงคำตอบที่ถูกต้องมากที่สุด รองลงมาคือ Jaccard และ Dice ตามลำดับ คือ 92.50%, 80.28% และ52.50%

2. ค่า Recall หมายถึง ความสามารถในการดึงเอกสารที่เกี่ยวข้องเท่านั้น แต่ไม่กำจัดเอกสารที่ไม่เกี่ยวข้องออก วิธี Cosine จะได้ค่าประสิทธิภาพในการดึงคำตอบที่เกี่ยวข้อง มากที่สุด รองลงมาคือ Jaccard และ Dice ตามลำดับ คือ 61.89%, 53.56% และ34.72%

3. ค่า F-measure หรือค่าความถ่วงดุล หมายถึง ค่าเฉลี่ย หรือค่าความแม่นยำในการขจัดเอกสารที่ไม่เกี่ยวข้องและค่าความสามารถในการดึงเอกสารที่เกี่ยวข้อง

5. สรุปผลการวิจัย และข้อเสนอแนะ

ข้อมูลที่นำมาทดลองเบื้องต้น ประกอบด้วย ชุดเอกสารที่เป็นคำตอบจำนวน 300 ชุด และชุดเอกสารที่เป็นคำถามจำนวน 30 คำถาม จากวารสารชุดความรู้เพื่อการดูแลตนเองสำหรับผู้เป็นเบาหวาน จากนั้นนำข้อมูลทั้งส่วนของคำถามและเอกสารคำตอบมาผ่านกระบวนการตัดคำภาษาไทยโดยใช้พจนานุกรม กำจัดคำที่ไม่สำคัญออกโดยใช้ฐานข้อมูลคำหยุด จนได้เป็นตัวแทนของคำ นำตัวแทนของคำไปหาค่าความถี่เพื่อคำนวณเป็นค่า น้ำหนักคำต่อไป ในการวัดความคล้ายคลึงของคำผู้วิจัยเลือกใช้วิธีการวัดความคล้ายคลึง 3 แบบ ประกอบด้วยวิธี Cosine, Jaccard และ Dice เพื่อศึกษาเปรียบเทียบประสิทธิภาพการหาคำตอบผลการวิจัยเบื้องต้นจากการศึกษาเปรียบเทียบประสิทธิภาพการหาคำตอบด้วยวิธีการวัดค่าความคล้ายคลึงระหว่างคำถามและคำตอบ โดยพบว่าวิธี Cosine มีประสิทธิภาพในการหาคำตอบมากที่สุด และรองลงมาคือวิธี Jaccard และ

Dice โดยได้ค่าความถูกต้องอยู่ 92.50%, 80.28% และ 52.50% ตามลำดับ และในการทำงานวิจัยครั้งนี้สามารถสรุปได้ว่า ความคล้ายคลึงแบบ Cosine เหมาะสำหรับข้อมูลที่เป็นความถี่ของคำ คำอยู่ในรูปแบบกระจัดกระจาย และใช้กับข้อมูลที่มีความยาวของเวกเตอร์ วิธีการนี้จะมีประสิทธิภาพในกรณีที่คำถามกับเอกสารคำตอบมีความยาวไม่เท่ากัน ส่วนการความคล้ายคลึงแบบ Jaccard ควรใช้กับชุดคำเฉพาะที่เป็นประโยค/เอกสาร โดยไม่พิจารณาความถี่ของคำที่เกิดขึ้นความซ้ำของคำไม่ทำให้ความคล้ายคลึงลดลงและจะทำงานได้ดี กรณีไม่มีความถี่ของคำ และสุดท้ายความคล้ายคลึงแบบ Dice เหมาะกับข้อมูลที่ไม่พิจารณาความถี่ ซึ่งจะพิจารณาเป็นคำต่อคำ

กิตติกรรมประกาศ

งานวิจัยนี้เป็นงานวิจัยที่ได้รับการสนับสนุนจากสาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย และขอขอบคุณแหล่งข้อมูลที่ใช้ในการทดลอง อาศัยองค์ความรู้จากแนวทางเวชปฏิบัติสำหรับโรคเบาหวาน ของสมาคมต่อมไร้ท่อแห่งประเทศไทย ข้อมูลการดูแลรักษาโรคเบาหวานจากสำนักงานกองทุนสนับสนุนการสร้างเสริมสุขภาพ (สสส.)

เอกสารอ้างอิง

Ditcharoen, N. & Techawiwatthanaboon, S. (2018). An alternative approach to course description comparison for university credit transfer using word similarity measurement and vector space model. *Journal of Science & Technology MSU*, 37(4), 580-586. [In Thai]

Frankes, W. B. & Baeza-Yates, R. (1992). *Information retrieval: Data structure & algorithms*. NJ, United State : Prentice-Hall.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles,* 37(142), 547-579. *https://www.researchgate.net/publication/225035806_Etude_de_la_distribution_florale_dans_une_portion_des_Alpes_et_du_Jura*

Kitreerawutiwong, N. & Tejativaddhana, P. (2013). Validity and reliability of the Thai version of the experienced continuity of care for diabetes Mellitus (ECC-DM) questionnaire. *Journal of Public Health, Burapha University*, 8(1), 13-25. https://he02.tci-thaijo.org/index.php/phjbuu/article/view/45580/37720 [In Thai]

Kondrak, G, Marcu, D., & Knight, K. (2003). Cognates can improve statistical translation models. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers* (pp. 46–48).

Mutabazi, E., Ni, J., Tang, G., & Cao, W. (2021). A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12), 5456. https://doi.org/10.3390/app11125456

Phetkrachang, K., Sathiwantanah, S., & Kongwan, A. (2022). A development of online question answering system for student registration web service of university using ontology technology. *KKU Science Journal*, 50(1), 24-34. https://ph01.tci-thaijo.org/index.php/KKUSciJ/article/view/250295/169916 [In Thai]

Radev, D.R., Qi, H., Zheng, Z., Blair-Goldensohn, S., Zhang, Z., Fan, W., & Prager, J. (2001). Mining the web for answers to natural language questions. *International Conference on Information and Knowledge Management,* Atlanta, Georgia, USA., (pp.143-150). https://doi.org/10.1145/502585.502610

Richard, J. F., Godbout, P., & Grèhaigne, J. F. (2000). Students' precision and interobserver reliability of performance assessment in teamsports. *Research Quarterly for Exercise and Sport*, 71(1), 85-91. https://doi.org/10.1080/02701367.2000.10608885

Salton, G & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

Senoussaoui, M., Kenny, P., Stafylakis, T., & Dumouchel, P. (2014). A study of the cosine distance-based mean shift for telephone speech diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(1), 217-227. https://doi.org/10.1109/TASLP.2013.2285474

Sornlertlamvanich, V. *(1993). Word segmentation for Thai in machine translation system. Bangkok :* NECTEC. [In Thai]

Wongsara, R., Homjun, K, & Ketui, N. (2021). Development of Thai subjective scoring system based on cosine-similarity. *Journal of Applied Information Technology*, 7(2), 7-16. https://ph02.tci-thaijo.org/index.php/project-journal/article/view/245037/166606 [In Thai]

Xie, W., Ding, R., Yan, J., & Qu, Y. (2018). A mobile-based question-answering and early warning system for assisting diabetes management. *Wireless Communications and Mobile Computing*, 2018, 1-14. https://doi.org/10.1155/2018/9163160