# การจำแนกความน่าเชื่อถือของเนื้อหาในเว็บไซต์ภาษาไทย ด้านมะเร็งโดยใช้ CancerDic+

สุภาพร เกิดกิจ[1] องอาจ อุ่นอนันต์[2] และ พยุง มีสัจ[3]

[1,3]คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพฯ
[2]คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ กรุงเทพฯ
Emails: supaporn.k@kmutnb.ac.th[1], aongart.a@rmutsb.ac.th[2], pym@kmutnb.ac.th[3]

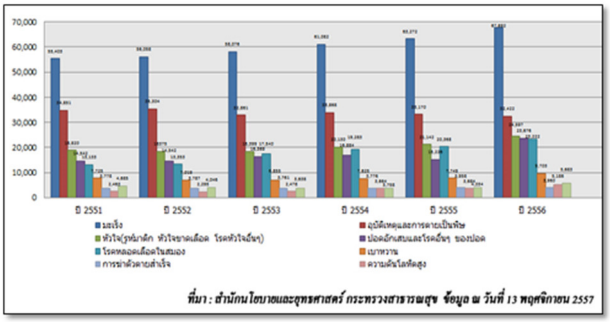
## บทคัดย่อ

ปัจจุบันจำนวนเว็บไซต์ที่ให้ความรู้ด้านมะเร็งมีอยู่เป็นจำนวนมาก ทำให้ผู้ใช้งานเข้าถึงข้อมูลได้อย่างสะดวกและมีปริมาณมากแต่จะทราบได้อย่างไรว่าเนื้อหาบนเว็บไซต์นั้นมีความน่าเชื่อถือหรือไม่ งานวิจัยนี้จึงมีวัตถุประสงค์ในการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็ง เพื่อแยกประเภทของเนื้อหาเว็บไซต์ที่มีความน่าเชื่อถือและไม่น่าเชื่อถือ ซึ่งงานวิจัยนี้นำเสนอ CancerDic+ เพื่อใช้ในการสกัดคำโดยมีการเพิ่มข้อมูลคำศัพย์เฉพาะด้านเกี่ยวกับมะเร็งและใช้เหมืองข้อมูล (Text Mining) ทำการจำแนกข้อมูล โดยมีการเปรียบเทียบค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall) ของการจำแนกความน่าเชื่อถือของเนื้อหาที่ผ่านเครื่องมือสกัดคำจาก Lexto, SWATH และ CancerDic+ ซึ่งผลการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์พบว่าการทำเหมืองข้อมูลโดยใช้ CancerDic+ สกัดคำให้ผลการจำแนกได้ดีที่สุด (Accuracy = 0.844, Precision = 0.838, Recall =0.845) ซึ่งสามารถนำไปประยุกต์ใช้งานอื่นได้อย่างมีประสิทธิภาพ

***คำสำคัญ—*** เหมืองข้อความ; การจำแนกข้อมูล; การสกัดคำ; ความน่าเชื่อถือของเว็บไซต์

## 1. บทนำ

ปัจจุบันการเติบของอินเทอร์เน็ตทำให้การค้นหาข้อมูลที่เกี่ยวข้องกับการแพทย์และการดูแลสุขภาพสามารถทำได้สะดวกและรวดเร็ว แต่ข้อมูลออนไลน์มีอยู่จำนวนมากนั้นมีทั้งข้อมูลที่มีความน่าเชื่อถือและไม่น่าเชื่อถือ จึงไม่สามารถนำข้อมูลเหล่านั้นมาใช้งานได้อย่างเหมาะสมและมีประสิทธิภาพได้ ซึ่งการดูแลสุขภาพต้องการข้อมูลที่มีความน่าเชื่อถือ ในงานวิจัยนี้จึงใช้เนื้อหาเกี่ยวกับโรคมะเร็งเนื่องจากเป็นโรคสำคัญหมายถึงโรคที่สามารถป้องกันได้แต่มีผู้เสียชีวิตเพิ่มขึ้นทุกปี ปัจจุบันมีอัตราการป่วยเพิ่มขึ้นอย่างต่อเนื่องและที่เป็นโรคที่มีอัตราการเสียชีวิตเป็นอันดับ 1 ตั้งแต่ปี พ.ศ. 2551-2556 จะเห็นได้จากรายงานสถิติข้อมูลของสำนักนโยบายและยุทธศาสตร์ กระทรวงสาธารณสุข แสดงดังรูปที่ 1 ดังนั้นหากมีข้อมูลออนไลน์ที่มีความน่าเชื่อถือก็จะสามารถให้ความรู้เพื่อช่วยลดอัตราการป่วยเป็นโรคมะเร็งได้




รูปที่ 1. จำนวนและอัตราผู้เสียชีวิตจากโรคสำคัญ ปี พ.ศ. 2551 – 2556

ดังนั้นการที่จะนำข้อมูลที่ได้จากเว็บไซต์มาใช้งาน จึงควรมีการจำแนกความน่าเชื่อถือของเว็บไซต์ ซึ่งในต่างประเทศมีการรับรองเนื้อหาภายในเว็บไซต์เกี่ยวกับทางการแพทย์และสุขภาพโดยมูลนิธิฮอน (Health On the Net Foundation: HON) ได้ดำเนินการส่งเสริมและแนะนำการนำข้อมูลออนไลน์ด้านสุขภาพที่มีประโยชน์เชื่อถือได้มาใช้งานได้เหมาะสมและมีประสิทธิภาพ โดยมูลนิธิ HON จะมีเกณฑ์การพิจารณาความน่าเชื่อถือของเว็บไซต์ 8 ข้อ ดังนี้ คุณสมบัติของผู้เขียน (Authoritative) ความสมบูรณ์ของบทความ (Complementarity) ความเป็นส่วนตัวของผู้ใช้งาน (Privacy) การแสดงแผล่งที่มา (Attribution) มีรองรับเรื่องการร้องเรียน (Justifiability) ความโปร่งใสของข้อมูล (Transparency) ระบุแหล่งเงินทุน (Financial Disclosure) และแยกส่วนเนื้อหาและโฆษณาอย่างชัดเจน (Advertising) เว็บไซต์ที่ได้รับการรับรองความน่าเชื่อถือจะได้รับสัญลักษณ์ภาพ HONcode เพื่อนำไปแสดงไว้ที่เว็บไซต์นั้น ๆ ซึ่งทำให้ผู้สืบค้นข้อมูลจากเว็บไซต์ที่ได้รับการรับรองความน่าเชื่อถือมีความมั่นใจในการนำข้อมูลเหล่านั้นมาใช้งานได้อย่างเหมาะสมและเกิดประโยชน์ [1] สำหรับประเทศไทยมีการรับรองเว็บไซต์ด้านพาณิชย์อิเล็กทรอนิกส์โดยกรมพัฒนาธุรกิจการค้า ซึ่งเว็บไซต์ที่ทำพาณิชย์อิเล็กทรอนิกส์ที่ผ่านการรับรองความน่าเชื่อถือจะได้รับสัญลักษณ์ภาพ DBD Verify แสดงที่หน้าเว็บไซต์ ทำให้ผู้ซื้อสินค้าออนไลน์เกิดความเชื่อมั่น แต่ในการรับรองเนื้อหาเว็บไซต์เกี่ยวกับ

สุขภาพด้านมะเร็งนั้นยังไม่มีการจำแนกหรือรับรองความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์

งานวิจัยนี้จึงมีวัตถุประสงค์ในการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็ง เพื่อแยกประเภทของเนื้อหาเว็บไซต์ที่มีความน่าเชื่อถือและไม่น่าเชื่อถือ โดยนำเสนอ CancerDic+ ซึ่งเป็นพจนานุกรมคำศัพท์เฉพาะด้านมะเร็งเพื่อตัดคำภาษาไทย โดยนำข้อมูลมาจากเว็บไซต์ ซึ่งประกอบไปด้วยข้อมูลที่มีความน่าเชื่อถือ เช่น ข้อมูลที่ได้จากเว็บไซต์หรือสถาบันทางการแพทย์ หรือบล็อกของแพทย์ ส่วนข้อมูลที่ไม่น่าเชื่อถือ เช่น ข้อมูลการขายประกันสุขภาพ ข้อมูลการขายอาหารเสริม ข้อมูลสมุนไพรหรือยาที่โฆษณาสรรพคุณเกินจริง จากนั้นจึงนำมาเข้ากระบวนการสกัดข้อความ (Text Extraction) เพื่อตัดคำ ในส่วนการให้ค่าน้ำหนักคำเพื่อที่จะนำไปเป็นตัวแทนเอกสารใช้วิธีการหาค่าน้ำหนัก (Term Weighting: TF) และการหาค่าความถี่ผกผัน (Inverse Document Frequency: IDF) แล้วจึงนำมาจำแนกประเภทความน่าเชื่อถือของเว็บไซต์ด้านมะเร็งและทำการประเมิน เพื่อเปรียบเทียบประสิทธิภาพการตัดคำและการจำแนกเนื้อหา

ในงานวิจัยได้แบ่งเนื้อหาออกเป็นส่วน ดังนี้ ส่วนที่ 2 ทฤษฎีที่เกี่ยวข้อง ส่วนที่ 3 วิธีดำเนินงาน ส่วนที่ 4 ผลการดำเนินงาน และส่วนที่ 5 สรุปผลและข้อเสนอแนะ

## 2. ทฤษฎีที่เกี่ยวข้อง

จากการศึกษาการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็งมีทฤษฎีที่เกี่ยวข้อง ดังนี้

### 2.1. ความน่าเชื่อถือของเว็บไซต์ (Website Credibility)

นิยามของความน่าเชื่อถือ (Credibility) [1] หมายถึง ความเชื่อถือได้ (Believability) ไม่ว่าจะเป็นบุคคลหรือวัตถุซึ่งมีลักษณะที่เชื่อถือได้ 2 ประการ คือ ความรู้สึกว่ามีคุณภาพผู้คนรับรู้ว่ามีคุณภาพ (Perceived) ซึ่งอาจไม่มีอยู่ในตัวบุคคลหรือวัตถุสารสนเทศจริง และความน่าเชื่อถือที่ได้จากการรับรู้ (Perception of Credibility) การประเมินความน่าเชื่อถือเป็นผลมาจากสมองที่จะประเมินปัจจัยที่สำคัญ ได้แก่ ความไว้เนื้อเชื่อใจได้ (Trustworthiness) สำหรับการประเมินสารสนเทศต่าง ๆ ผ่านเว็บและความเป็นผู้ชำนาญการ (Expertise) ซึ่งจะต้องมีความรู้ประสบการณ์และสมรรถนะ มีชื่อนักเขียนบทความและการอ้างอิงชัดเจน เป็นต้น ความไว้เนื้อเชื่อใจจะบอกถึงความดีและมีจรรยาบรรณของเว็บไซต์ ดังนั้นหากต้องการให้เว็บไซต์ที่มีความน่าเชื่อถือจะต้องทำให้ผู้มาเยี่ยมชมรับรู้ว่ามีความไว้เนื้อเชื่อใจได้ และความเป็นผู้ชำนาญการในระดับสูง นอกจากนี้ปัจจัย ที่ทำให้เว็บไซต์น่าเชื่อถือได้เพิ่มขึ้น เช่น การเพิ่มคุณค่าให้กับเว็บไซต์ โดยการปรับปรุงเนื้อหาให้ทันสมัย บทความต้องมีอ้างอิงหรือผู้แต่งเสมอไม่มีโฆษณามากเกินไป นามสกุลของเว็บไซต์ (Domain Name) ต้องเป็นขององค์กรที่จดทะเบียนอย่างถูกต้อง ทุกองค์ประกอบบนเว็บไซต์ทำงานได้ถูกต้องและชื่อเสียงในด้านดีขององค์กรก็จะส่งผลต่อเว็บไซต์ด้วย

### 2.2. การสกัดข้อความ (Text Extraction)

การสกัดข้อความนั้นเป็นกระบวนการของการทำเหมืองข้อความ (Text Mining) เพื่อใช้การวิเคราะห์คำออกจากเอกสาร ข่าวสาร ข้อความ และสารสนเทศต่าง ๆ ที่เป็นตัวอักษรโดยสามารถนำไปทำการแบ่งกลุ่ม (Clustering) การจำแนกข้อมูล (Classification) และการหาความสัมพันธ์ (Association) ซึ่งในการแบ่งกลุ่มเอกสาร (Document Clustering) [2] เป็นการวัดความคลายคลึงกันของข้อความในตัวเอกสาร โดยข้อมูลตัวอักษรจะถูกแปลงเป็นตัวเลขเพื่อทำขั้นตอน การแบ่งกลุ่มโดยใช้เทคนิคต่างๆ เช่น DBSCAN, K-mean, SOM และ Hierarchical ซึ่งก่อนการทำเหมืองข้อความจะต้องผ่านขั้นตอนการเตรียมข้อมูล (Preprocess) ก่อนซึ่งมีขั้นตอน [3][4][5] ดังนี้

#### *2.2.1. การตัดคำ (Word Segmentation)*

เป็นการแยกแต่ละคำจากเอกสารออกจากกัน โดยยังคงมีความหมายที่ถูกต้องสมบูรณ์อยู่ โดยการตัดคำนั้นใช้ฐานข้อมูลพจนานุกรมคำศัพท์ในการแบ่งคำออกมา [6]

#### *2.2.2. การกำจัดคำหยุด (Stop Word)*

เป็นการตัดคำที่ไม่มีความหมายออกจากเอกสารโดยการกำจัดคำหยุดนั้นใช้ฐานข้อมูลคำศัพท์ที่เป็นคำที่ไม่มีความหมาย ในการกำจัดคำออกเมื่อทำการตัดคำเรียบร้อยแล้วทำการเลือกคำที่ต้องการใช้ในการวิเคราะห์ (Feature Selection) [6]

#### *2.2.3. วิธีที่ใช้ในการตัดคำ (Word Segmentation Method)*

การตัดคำเป็นขั้นตอนที่มีความสำคัญอย่างยิ่งในการประมวลผลข้อความภาษาธรรมชาติ ซึ่งเป็นพื้นฐานในการที่จะนำข้อมูลที่ได้จากการตัดคำไปใช้งานในด้านอื่น ๆ ต่อไปซึ่งประสิทธิภาพการประมวลผลข้อความขึ้นอยู่กับประสิทธิภาพความถูกต้องของการตัดคำ สำหรับวิธีการที่นิยมใช้ในการตัดคำไทยสามารถแบ่งออกเป็น 3 วิธี ได้แก่ วิธีการใช้กฎ (Rule-based) วิธีการใช้พจนานุกรม (Dictionary-based) และวิธีการใช้คลังข้อความ (Corpus-based) [7]

#### *2.2.4. เทคนิคการตัดคำ (Word Segmentation Techniques)*

ในการตัดคำมีเทคนิคที่สามารถนำมาใช้โดยขึ้นอยู่กับลักษณะของคำที่จะนำมาตัดซึ่งมีเทคนิคที่นิยมใช้งานกันแบ่งออกเป็น 5 เทคนิค ดังนี้ เทคนิคการเทียบคำที่ยาวที่สุด (Longest Word Pattern Matching) เทคนิคการเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching) เทคนิคการตัดคำที่ใช้ความถี่ของคำหรือสถิติ (Probabilistic Word Segmentation) เทคนิคการย้อนรอยกลับ (Back Tracking) และเทคนิคการตัดคำแบบใช้คุณลักษณะ (Feature-based Approach) [7]

งานวิจัยเกี่ยวกับการตัดคำมีอยู่มากมาย ทางผู้วิจัยได้ทำการศึกษางานวิจัยของวิโรจน์ [8] เสนอการตัดคำด้วยวิธีไตรแกรมกับคลังข้อมูลเพื่อตัดคำระดับพยางค์และทดสอบกับพจนานุกรมแต่วิธี Maximum Collocation ไม่สามารถแบ่งคำได้ชัดเจนขึ้นอยู่กับคำที่มีอยู่ใน

พจนานุกรม และควรใช้วิธีการทางสถิติอื่นมาปรับปรุงประสิทธิภาพ ปโยธร [9] ใช้เทคนิควิธีการตรวจสอบย้อนกลับ (Back Tracking) และการเลือกคำที่ยาวที่สุด (Longest Matching) และควรให้โปรแกรมเพิ่มคำเฉพาะที่ไม่พบในพจนานุกรมในตอนแรกโดยอัตโนมัติเพื่อให้การตัดคำในเอกสารแบบเดียวกันเป็นไปได้ดีขึ้น สิทธิโชค [10] สร้างโมเดลวิธีการตัดคำภาษาไทยด้วยเทคนิคการเรียนรู้ของเครื่อง พบปัญหาการคำนวณความถูกต้องและแม่นยำของคอมพิวเตอร์ รวมทั้งการจัดขอบขวาในโปรแกรมประมวลผลคำเพื่อค้นหาคำ ส่วนของการแบ่งคำจะเสียเวลาในขั้นตอนการแปลงไฟล์หลายไฟล์เพื่อรวมเป็นไฟล์เดียวก่อนที่จะนำมาเป็นข้อมูลสำหรับสอน (Train Data) หรือข้อมูลสำหรับทดสอบ (Test Data) และมีข้อจำกัดคือสามารถตัดคำภาษาไทยที่มีข้อความเฉพาะอักขระเท่านั้น สำหรับงานวิจัยที่เกี่ยวข้องกับการตัดคำเพื่อแก้ปัญหาคำกำกวมโดย ชนินทร์ [11] ตัดคำด้วยวิธีพจนานุกรมคำกำกวมเพื่อแก้ปัญหาคำกำกวมและคำที่ไม่ปรากฏในพจนานุกรมด้วยโมเดล PTTSF (Parsing Thai Text with Syntax and Feature of Word) พบปัญหาไม่สามารถตัดคำที่เป็นคำกริยาระหว่างคำที่ไม่ปรากฏในพจนานุกรมได้ กานดา [2] ตัดคำในเอกสารภาษาไทยโดยการใช้กฎ (Rule-based) และพจนานุกรมแบบใหม่ร่วมกันมีการตัดคำในระดับพยางค์ ส่วนคำกำกวมสามารถแก้ปัญหาโดยการใช้วิธีตัดคำที่ยาวที่สุดร่วมกับวิธีการย้อนกลับแต่ยังไม่สามารถตัดคำได้ถูกต้องทุกครั้ง และควรมีการเพิ่มเติมปรับปรุงสำหรับการตัดคำประเภทนี้ด้วยการใช้ค่าสถิติของคำที่พบในเอกสารทั่วไปร่วมกับการใช้ความถี่ของคำที่พบในเอกสารที่นำมาตัดคำ ชูชาติ [12] นำเสนอกรอบการทำงานการเก็บคำที่ไม่รู้จักจากเว็บ โดยใช้การวิเคราะห์คำที่ไม่รู้จักร่วมกับพจนานุกรมเพื่อสกัดคำที่ไม่รู้จักโดยอัตโนมัติ ทำให้ผู้ใช้สามารถเพิ่มคำไทยที่ไม่รู้จักลงในพจนานุกรมคำไทยที่ไม่รู้จัก ผลการทดลองพบว่าสามารถวิเคราะห์คำที่ไม่รู้จักได้สูงสุดถึงร้อยละ 96

### 2.3. การจำแนกประเภทข้อมูล (Data Classification)

การทำเหมืองข้อความเป็นการสกัดเอาสิ่งที่มีประโยชน์ออกมาจากข้อความที่มีจำนวนมาก ซึ่งมีหลากหลายวิธี หนึ่งในนั้นคือการจำแนกข้อมูล (Classification) ซึ่งมีเทคนิคต่าง ๆ ที่นิยมใช้งานดังนี้

#### *2.3.1. ต้นไม้ตัดสินใจ (Decision Tree)*

เทคนิคต้นไม้ตัดสินใจ เป็นวิธีหนึ่งที่ใช้ในการจำแนกข้อมูล โดยมีลักษณะการทำงานเหมือนโครงสร้างของต้นไม้ รูปแบบของต้นไม้ตัดสินใจประกอบด้วยโหนดราก (Root Node) ซึ่งเป็นโหนดแรก และแตกสายย่อยเป็นโหนดลูก (Child Node) [13][14] โดยจะสามารถแปลงไปเป็นกฎที่ใช้ในการจำแนกข้อมูลหรือที่เรียกว่าฐานกฎ (Rule-based) เพื่อใช้ในการพยากรณ์ข้อมูล โดยกฎนั้นจะเป็นไปตามรูปแบบของต้นไม้ตัดสินใจ

#### *2.3.2. โครงข่ายประสาทเทียม (Artificial Neural Network: ANN)*

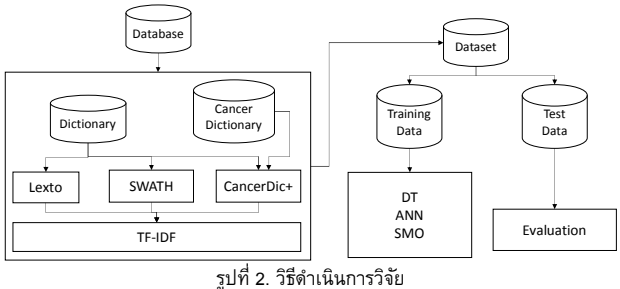
เป็นการจำแนกข้อมูลอีกวิธีหนึ่งที่นิยมใช้งาน มีการทำงานเลียนแบบหลักการทำงานของสมองมนุษย์ โดยมีหน่วยที่ใช้ในการประมวลผลเรียกว่า นิวรอน ซึ่งนิวรอนแต่ละนิวรอนสามารถรับค่าได้หลายอินพุตแต่มีเอาต์พุตได้เพียงเอาต์พุตเดียวเท่านั้น โดยทุกอินพุตมีค่าถ่วงน้ำหนัก (Weight) และในแต่ละนิวรอนนั้นจะมีค่าความเอนเอียงหรือไบแอส (Bias) อยู่หรือไม่ก็ได้ โดยเมื่อปรับค่าถ่วงน้ำหนักและค่าเอนเอียงที่เหมาะสมแล้วจะถูกส่งไปยังฟังก์ชันถ่ายโอน (Transfer Function) [15][16] เพื่อคำนวณค่าผลลัพธ์

#### *2.3.3. ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)*

เป็นเทคนิคที่มีการใช้สมการทางคณิตศาสตร์ในการจำแนกข้อมูล โดยพยายามหาจุดที่เส้นแบ่งกลุ่มของข้อมูลมีระยะความห่างระหว่างเส้นขอบเขต (Border Line) มากที่สุด ซึ่งวิธีนี้ทำให้มีข้อดี คือ สามารถรองรับจำนวนตัวแปรที่หลากหลายได้เป็นจำนวนมาก และค่อนข้างมีความถูกต้องสูง ซึ่งในการทำงานของตัวซัพพอร์ตเวกเตอร์แมชชีน มีฟังชั่นให้เลือกใช้อย่างหลากหลาย เช่น Linear Function, Polynomial Function และRadial Basis Function [17] แต่ต้องมีการเลือกใช้ฟังชั่นให้ตัวซัพพอร์ตเวกเตอร์แมชชีน จำแนกข้อมูลได้อย่างเหมาะสมด้วยเช่นกัน

## 3. วิธีดำเนินงาน

วิธีดำเนินงานวิจัยการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็งโดยใช้ CancerDic+ ได้เตรียมข้อมูลสำหรับการจำแนกโดยมีขั้นตอน แสดงดังรูปที่ 2




รูปที่ 2. วิธีดำเนินการวิจัย

### 3.1. การเก็บรวบรวมข้อมูล (Data Collection)

การเก็บข้อมูลโดยใช้ข้อมูลจากเว็บไซต์ที่สืบค้นด้วยคำสำคัญที่เกี่ยวกับมะเร็ง เช่น มะเร็ง อาการของมะเร็ง อาหารสำหรับผู้ป่วยมะเร็ง อาหารเสริมสำหรับมะเร็ง สมุนไพรรักษามะเร็ง เป็นต้น โดยมีจำนวน 484 แถว ซึ่งสามารถแบ่งเขตข้อมูลออกเป็น 6 ส่วน แสดงดังตารางที่ 1 ในส่วนของหลักเกณฑ์ที่ใช้ในการกำกับประเภทของเว็บไซต์ว่าเป็นเว็บที่น่าเชื่อถือหรือไม่ใช้จากแหล่งที่มาของข้อมูล ซึ่งข้อมูลที่น่าเชื่อถือมาจากเว็บไซต์โรงพยาบาล หน่วยงานสาธารณสุขของรัฐ บล็อคของแพทย์ ส่วนข้อมูลที่ไม่น่าเชื่อถือมาจากเว็บไซต์ขายประกัน ขายอาหารเสริมที่มีสรรพคุณเกินจริง

ตารางที่ 1. เขตและรูปแบบข้อมูล

| No | Name | Detail | Type |
|---|---|---|---|
| 1. | Text | เนื้อหาในเว็บไซต์ | Text |
| 2. | Lexto | ข้อความที่สกัดผ่าน Lexto | Text |
| 3. | SWATH | ข้อความที่สกัดผ่าน SWATH | Text |
| 4. | Dic | ข้อความที่สกัดผ่าน CancerDic+ | Text |
| 5. | Link | ลิงค์ที่มาของเนื้อหาในเว็บไซต์ | Text |
| 6. | Type | ประเภทของเนื้อหา(น่าเชื่อถือ,ไม่น่าเชื่อถือ) | Char(1) |

### 3.2. การสกัดคำ (Text Extraction)

การสกัดคำใช้เครื่องมือสำหรับการตัดคำได้แก่ เล็กซ์โต (Thai Lexeme Tokenizer: LexTo) SWATH (Smart Word Analysis for THai) และ CancerDic+ (Cancer Dictionary Plus) ที่ผู้วิจัยนำเสนอโดยผ่านขั้นตอนการตัดคำ (Word Segmentation) แล้วนำไปบันทึกลงฐานข้อมูลที่ผ่านการสกัดคำจากเครื่องมือดังกล่าวข้างต้น เป็นข้อมูลสำหรับการสอน (Train Data) และข้อมูลสำหรับการทดสอบ (Text Data)

### 3.3. การสร้างดัชนีเอกสาร (Document Indexing)

การสร้างดัชนีเอกสารเป็นขั้นตอนการแปลงเอกสารซึ่งเป็นภาษาธรรมชาติให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถประมวลผลได้ ซึ่งจะเป็นการสร้างตัวแทนเนื้อหาเอกสารให้อยู่ในรูปแบบเวกเตอร์ของน้ำหนักคำ (Term Weighting) เพื่อนำมาหาสร้างดัชนีโดยนิยมใช้รูปแบบของคำเดี่ยว [13] เริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกซ์ของเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมด ในกลุ่มจนกระทั่งกลายเป็นเมตริกซ์

ในส่วนของการคำนวณค่าน้ำหนักให้แก่ดัชนีใช้วิธีการ TF-IDF Weighting (Term Frequency-Inverse Document Frequency) แสดงดังสมการที่ (1)(2)(3)(4) [13]

$$TF - IDF = TF \times IDF \quad (1)$$

$$TF_t = \frac{n_t}{N} \quad (2)$$

$$IDF_t = 1 + \log\left(\frac{D}{d_t}\right) \quad (3)$$

$$TF - IDF_t = \frac{n_t}{N} \times [1 + \log\left(\frac{D}{d_t}\right)] \quad (4)$$

โดยที่

$n_t$ = จำนวนคำ $t$ ที่ปรากฏในเอกสาร

$N$ = จำนวนคำทั้งหมดที่ปรากฏในเอกสาร

$D$ = จำนวนเอกสารทั้งหมด

$d_t$ = จำนวนเอกสารที่มีคำ $t$ ปรากฏ

### 3.4. การจำแนกข้อมูลและการวัดประสิทธิภาพ

ในงานวิจัยนี้ผู้วิจัยได้ทำการเปรียบเทียบค่าประสิทธิภาพซึ่งพิจารณาจากค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และความครบถ้วน (Recall) จากโมเดล 3 รูปแบบ แสดงดังตารางที่ 2 โดยทุกโมเดลใช้วิธีการ 10-fold Cross Validation วัดความเที่ยงตรงของโมเดล

ตาราง 2. ตารางแสดงรายละเอียดของโมเดลที่ทดลอง

| โมเดล | ชื่อโมเดล | ชื่อย่อ |
|---|---|---|
| 1 | Decision Trees | DT |
| 2 | Artificial Neural Network (Back Propagation) | ANN |
| 3 | Support Vector Machine Optimization (Linear Fn.) | SMO |

การหาค่าประสิทธิภาพนั้นใช้การพิจารณาค่าประสิทธิภาพจากระดับค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall) แสดงดังสมการที่ (5)(6)(7) [8]

$$Precision(P) = \frac{TP}{TP+FP} \quad (5)$$

$$Recall(R) = \frac{TP}{TP+FN} \quad (6)$$

$$Accuracy(A) = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

โดยเมื่อ TP = True Positive, FP = False Positive, FT = True Negative และ FN = False Negative

## 4. ผลการดำเนินงาน

ผู้วิจัยได้ดำเนินการจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ด้านมะเร็ง โดยทำการสกัดคำ การคำนวณหาค่าดัชนีเอกสาร และเปรียบเทียบประสิทธิภาพการจำแนกประเภทความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ มีรายละเอียดดังนี้

### 4.1. ผลการสกัดคำและสร้างดัชนีเอกสาร

ในการดำเนินการสกัดคำจากข้อมูลเนื้อหาด้านมะเร็ง ด้วยเครื่องมือ LexTo และ SWATH กับฐานข้อมูลพจนานุกรมเล็กซิตรอน แต่ในส่วนของ CancerDic+ นั้นได้ใช้ฐานข้อมูลพจนานุกรมเล็กซิตรอนผสมกับฐานข้อมูลคำศัพย์เฉพาะทางด้านมะเร็ง โดยแสดงตัวอย่างการตัดคำด้วยโปรแกรมทั้ง 3 โปรแกรมข้างต้น แสดงดังรูปที่ 3

Full Text

สาเหตุทางกรรมพันธุ์ มะเร็งบางชนิด เช่น มะเร็งเต้านม มะเร็งรังไข่ และมะเร็งลำไส้ มีแนวโน้มเกิดขึ้นได้กับบุคคลภายในครอบครัวที่มีประวัติเป็นมะเร็งดังกล่าว

Lexto

สาเหตุ | ทางกรรมพันธุ์ | | มะเร็ง | บาง | ชนิด | | เช่น | | มะเร็ง | เต้านม | | มะเร็ง | รังไข่ | | และ | มะเร็ง | ลำไส้ | | มีแนวโน้ม | เกิดขึ้น | ได้ | กับ | บุคคลภายใน | ครอบครัว | ที่ | มีประวัติ | เป็น | มะเร็ง | ดังกล่าว |

SWATH

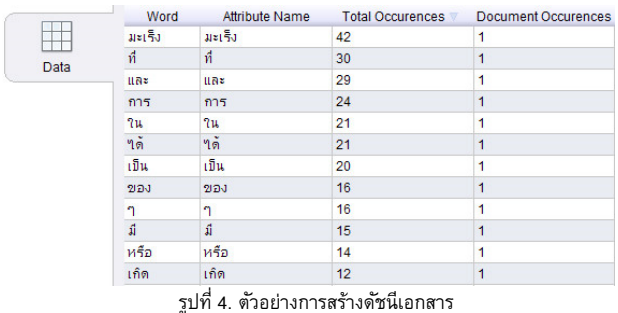
สาเหตุ|ทาง|กรรมพันธุ์| |มะ|เร็ง|บาง|ชนิด| |เช่น| |มะ|เร็ง|เต้า|นม| | มะ|เร็ง|รัง|ไข่| |และ|มะ|เร็ง|ลำ|ไส้| |มี|แนว|โน้ม|เกิด|ขึ้น|ได้|กับ| บุคคล|ภาย|ใน|ครอบ|ครัว|ที่|มี|ประ|วัติ|เป็น|มะ|เร็ง|ดัง|กล่าว|

CancerDic+

สาเหตุ|ทาง|กรรมพันธุ์| |มะเร็ง|บาง|ชนิด| |เช่น| |มะเร็ง|เต้านม| | มะเร็ง|รังไข่| |และ|มะเร็ง|ลำไส้| |มี|แนวโน้ม|เกิด|ขึ้น|ได้|กับ|บุคคล| ภาย|ใน|ครอบ|ครัว|ที่|มี|ประวัติ|เป็น|มะเร็ง|ดังกล่าว|

รูปที่ 3. ตัวอย่างการตัดคำ

ซึ่งจากการใช้เครื่องมือทั้ง 3 ทำการสกัดคำจะพบว่าโปรแกรม Lexto และ CancerDic+ มีการตัดคำที่เป็นคำสำคัญในด้านมะเร็งใกล้เคียงกันเช่น “มะเร็ง” ตัดคำออกมาเป็น |มะเร็ง| แต่ในส่วนของโปรแกรม SWATH มีการตัดคำสำคัญในด้านมะเร็งเช่น “มะเร็ง” ตัดคำออกมาเป็น |มะ|เร็ง| ซึ่งทำให้เกิดความแตกต่างกัน และในการสร้างดัชนีเอกสารได้ใช้โปรแกรม RapidMiner Studio ซึ่งมีตัวอย่างการสร้างดัชนีเอกสาร ซึ่งจะถูกนำไปใช้ในการจำแนกข้อมูลในขั้นตอนถัดไป แสดงดังรูปที่ 4


Data

| Word | Attribute Name | Total Occurences | Document Occurences |
|---|---|---|---|
| มะเร็ง | มะเร็ง | 42 | 1 |
| ที่ | ที่ | 30 | 1 |
| และ | และ | 29 | 1 |
| การ | การ | 24 | 1 |
| ใน | ใน | 21 | 1 |
| ได้ | ได้ | 21 | 1 |
| เป็น | เป็น | 20 | 1 |
| ของ | ของ | 16 | 1 |
| ๆ | ๆ | 16 | 1 |
| มี | มี | 15 | 1 |
| หรือ | หรือ | 14 | 1 |
| เกิด | เกิด | 12 | 1 |

รูปที่ 4. ตัวอย่างการสร้างดัชนีเอกสาร

### 4.2. ผลการหาค่าประสิทธิภาพโดยใช้ Lexto สกัดคำ

จากการทดลองจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ โดยผ่านการสกัดคำโดยใช้โปรแกรม Lexto ทำให้ได้ค่าประสิทธิภาพ ซึ่งประกอบไปด้วยค่าความถูกต้อง ค่าความแม่นยำ และค่าความครบถ้วน แสดงดังตารางที่ 3

ตาราง 3. แสดงเปรียบเทียบค่าประสิทธิภาพโดย Lexto สกัดคำ

| Model | 10-fold cross validation | | |
|---|---|---|---|
| | Accuracy | Precision | Recall |
| DT | 0.839 | 0.839 | 0.839 |
| ANN | 0.819 | 0.814 | 0.819 |
| SMO | 0.814 | 0.810 | 0.815 |

จากตารางพบว่าโมเดลการจำแนกข้อมูลที่มีค่าประสิทธิภาพ โดยเรียงจากมากไปน้อย ดังนี้ DT, ANN และ SMO ตามลำดับ

### 4.3. ผลการหาค่าประสิทธิภาพโดยใช้ SWATH สกัดคำ

จากการทดลองจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ โดยผ่านการสกัดคำโดยใช้โปรแกรม SWATH ทำให้ได้ค่าประสิทธิภาพ ซึ่งประกอบไปด้วยค่าความถูกต้อง ค่าความแม่นยำ และค่าความครบถ้วน แสดงดังตารางที่ 4

ตาราง 4. แสดงเปรียบเทียบค่าประสิทธิภาพโดยใช้ SWATH สกัดคำ

| Model | 10-fold cross validation | | |
|---|---|---|---|
| | Accuracy | Precision | Recall |
| DT | 0.777 | 0.775 | 0.783 |
| ANN | 0.769 | 0.765 | 0.773 |
| SMO | 0.764 | 0.762 | 0.769 |

จากตารางพบว่าโมเดลการจำแนกข้อมูลที่มีค่าประสิทธิภาพ โดยเรียงจากมากไปน้อย ดังนี้ DT, ANN และ SMO ตามลำดับ

### 4.4. ผลการหาค่าประสิทธิภาพโดยใช้ CancerDic+ สกัดคำ

จากการทดลองจำแนกความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ โดยผ่านการสกัดคำโดยใช้โปรแกรม CancerDic+ ทำให้ได้ค่าประสิทธิภาพ ซึ่งประกอบไปด้วยค่าความถูกต้อง ค่าความแม่นยำ และค่าความครบถ้วน แสดงดังตารางที่ 5

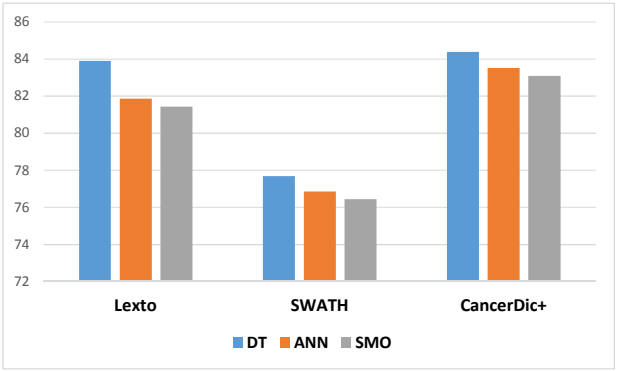
ตาราง 5. แสดงเปรียบเทียบค่าประสิทธิภาพโดยใช้ CancerDic+ สกัดคำ

| Model | 10-fold cross validation | | |
|---|---|---|---|
| | Accuracy | Precision | Recall |
| DT | 0.844 | 0.838 | 0.845 |
| ANN | 0.835 | 0.829 | 0.836 |
| SMO | 0.831 | 0.825 | 0.830 |

จากตารางพบว่าโมเดลการจำแนกข้อมูลที่มีค่าประสิทธิภาพ โดยเรียงจากมากไปน้อย ดังนี้ DT, ANN และ SMO ตามลำดับ

## 5. สรุปผลและข้อเสนอแนะ

จากการทดลองหาค่าประสิทธิภาพการจำแนกจากทั้ง 3 โมเดลจากการสกัดคำจากทั้ง 3 โปรแกรมทำให้ได้ค่าประสิทธิภาพของแต่ละโมเดลในแต่ละการทดลองจึงนำมาเปรียบเทียบกันทั้ง 3 รูปแบบการทดลอง เพื่อให้เห็นภาพได้อย่างชัดเจน โดยใช้ค่าความถูกต้อง (Accuracy) มาทำกราฟเปรียบเทียบ แสดงดังรูปที่ 5




รูปที่ 5: เปรียบเทียบค่าประสิทธิภาพ

จากรูปพบว่าผลการทดลองโมเดลที่มีค่าความถูกต้องมากที่สุด คือ โมเดลที่ผ่านการสกัดคำจาก CancerDic+ และใช้เทคนิคการจำแนกแบบ DT โดยมีค่าความถูกต้อง 0.844 ดีกว่าโมเดลที่ผ่านการสกัดคำจาก Lexto และ SWATH เนื่องจากมีการใช้คำศัพย์เฉพาะทางเกี่ยวกับโรคมะเร็งทำให้การสกัดคำในขั้นตอนการเตรียมข้อมูล ก่อนนำไปทำเหมืองข้อมูลมีความถูกต้องมากยิ่งขึ้น และจากการเก็บรวบรวมข้อมูลเนื้อหาจากเว็บไซต์ต่างๆ พบปัญหาที่ส่งผลต่อประสิทธิภาพการตัดคำ เช่น การสะกดคำผิด การใช้คำทับศัพท์ภาษาอังกฤษ อีกทั้งยังสามารถ

กำหนดเกณฑ์อื่น ๆ ที่ใช้ในการประเมินค่าความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ประกอบกับเนื้อหาได้ ซึ่งจะนำไปพัฒนาต่อไปในอนาคต

## เอกสารอ้างอิง


[1] "What makes a website credible?." (ออนไลน์). แหล่งที่มา http://captology.stanford.edu/resources/what-makes-a-website-credible.html

[2] นิเวศน์ จิระวิชิตชัย. "แบบจำลองการจำแนกเอกสารสำหรับภาษาไทยอัตโนมัติ," *The Journal of Industrial Technology* 2556, Vol. 2556. No. 1.

[3] X. Dai, Y. He, and Y. Sun, "A Two-layer Text Clustering Approach for Retrospective News Event Detection," *International Conference on Artificial Intelligence and Computational Intelligence (AICI)*, vol. 1, pp. 364–368, 2010.

[4] U. Gunasinghe, S. Matharage, and D. Alahakoon, "A sequence based dynamic SOM model for text clustering," *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2012.

[5] K.Norvag and R.Oyri, "New Item Extraction for Text Mining in Web Newspapers," *International Workshop on Challenges in Web Information Retrieval and Integration*, 2005.

[6] นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ. "การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ," *NIDA Development Journal*, Vol. 51, No. 3, หน้า 187-205, 2554.

[7] สายัณห์ เทพแดง. "*การปรับปรุงประสิทธิภาพของการตัดคำไทย ด้วยเทคนิคการจดจำนิพจน์ระบุนาม*," ปริญญานิพนธ์ วท.ม. สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์, 2553.

[8] Wirote Aroonmanakun. "Collocation and Thai Word Segmentation," *PROCEEDINGS OF SNLP-Oriental COCOSDA*, pp. 68-75, 2002.

[9] ปโยธร อุราธรรมกุล และ กานดา รุณนะพงศา. "การตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่," *JCSSE 2006*, vol. 2549, pp. 34-40, 2006.

[10] สิทธิโชค ทรัพย์ไพบูลย์กิจ และ สุพัฒนวรี ทิพย์เจริญ. "*การเพิ่มประสิทธิภาพการตัดคำภาษาไทยด้วยเทคนิคการเรียนรู้ด้วยเครื่อง*," ปริญญานิพนธ์ วท.ม. ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่, 2549.

[11] ชนินทร์ มหัทธนชัย. "การพัฒนาเทคนิคการตัดคำแบบอาศัยไวยกรณ์และบริบทคำรอบข้าง," *National Conference on Computer Information Technologies*, 2555.

[12] Choochart Haruechaiyasak. "A Collaborative Framework for Collecting Thai Unknown words from the web," *Proceeding COLING-ACL'06,* pp. 345-352, 2006.

[13] J. R. Quinlan. "Induction of Decision Trees," *in Machine Learning*, pp. 81–106, 1986.

[14] Zlatko J. Kovcic. "Early Prediction of Student Success: Mining Students Enrolment Data," *Proceeding of Informing Science & IT Education Conference (ImSITE)*, pp. 647-665, 2010.

[15] Edin Osmanbegovic, Mirza Suljic. "Data Mining Approach for Predicting Student Performance," *Journal of Economics and Business*, Vol. 5, No. 1, pp. 3-12, 2012.

[16] พยุง มีสัจ. **ระบบฟัซซีและโครงข่ายประสาทเทียม**. พิมพ์ครั้งที่ 1. กรุงเทพฯ : ศูนย์ผลิตตำราเรียน มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2555.

[17] ภัทร์พงศ์ พงศ์ภัทรกานต์. "การวิเคราะห์ปัจจัยที่ส่งผลต่อการพันสภาพของนักศึกษาระดับปริญญาตรี โดยใช้คอมมิตตีแมชชีน," *The 6th National Conference On Computing And Information Technology*, pp. 491-496, 2010.