การสรุปบทความแบบ Extractive Summarize โดยใช้ Text Rank และ BM25 กรณีศึกษาบทความสายพันธุ์ต่างๆ ของเมล็ดกัญชาเพศเมีย

BADS9000 การค้นคว้าอิสระ (Independent Study)

คณะสถิติประยุกต์ สาขาวิชาการวิเคราะห์ธุรกิจ และวิทยาการข้อมูล (Business Analytics and Data Science : BADS)

สถาบันบัณฑิตพัฒนบริหารศาสตร์ (National Institute of Development Administration : NIDA)

โดย *ณัฐวัฒน์ คงเมือง (รหัสประจำตัวนักศึกษา : 6310412004)*

ชื่อปริญญา *วิทยาศาสตรมหาบัณฑิต*

สาขาวิชา *การวิเคราะห์ธุรกิจ และวิทยาการข้อมูล*

ที่ปรึกษารายวิชา *รองศาสตราจารย์ ดร.โอม ศรนิล*

ปีการศึกษา *2564*

E-mail *km.nuttawat@gmail.com*

เบอร์โทรศัพท์ *065-259-9562*

บทคัดย่อ—กัญชา เป็นพืชสมุนไพรที่มีมาอย่างช้านานในโลก ในหลายประเทศได้นำมาใช้ในทางการแพทย์ และสันทนาการต่างๆ มากมาย แต่อย่างไรก็ตาม ก็ยังมีอีกหลายประเทศที่ยังไม่ได้ให้การสนับสนุน หรือไม่มีการออกกฎหมาย หรือแม้แต่การนำมาใช้อย่างทั่วไป หรือเป็นประโยชย์ในด้านต่างๆ อีกทั้งมุมมองในเรื่องของสารเสพติดอีกด้วย แต่ในปัจจุบันนี้หลายๆ ประเทศเริ่มให้ความสนใจ และเปิดเสรีเกี่ยวกับพืชชนิดนี้อย่างมากมาย ทั้งการนำมาใช้ในด้านอุปโภค บริโภคต่างๆ มีการสนับสนุนงานวิจัย การสันนาการต่างๆ การยอมรับในวงการแพทย์มากขึ้น จึงทำให้หลายประเทศเริ่มสนใจในพืชกัญชามากขึ้น ทั้งในเรื่องการเพาะปลูก หรือเรื่องอุตสาหกรรมต่างๆ ที่เกี่ยวกับกัญชากำลังเติบโตขึ้นอย่างต่อเนื่อง และรวดเร็ว

Dutch Passion (dutch-passion.com) เป็นเว็บไซต์หนึ่ง ที่มีการให้ความรู้ และขายสินค้าเมล็ดพันธุ์กัญชา จัดได้ว่าเป็นเว็บไซต์ที่เป็นธนาคารเมล็ดพันธุ์กัญชาของโลก มีเมล็ดพันธุ์กัญชามากมายหลากหลายชนิด โดยพืชกัญชานั้น เป็นพืชที่มีเมล็ดเพศผู้ และเมล็ดเพศเมีย และมีสารเคมีหลักที่ออกฤทธิ์ที่สำคัญ คือ Tetrahydrocannabinol (THC) โดยที่สาร THC นี้สามารถนำไปใช้ประโยชน์ในด้านต่างๆ ได้อย่างมากมาย ซึ่งในแต่ละสายพันธุ์ของกัญชา และแต่ละเพศของกัญชานี้ จะมีประสิทธิภาพ และมีความเข้มข้นของสารเคมีดังกล่าวแตกต่างกันออกไป กล่าวคือ เมล็ดพันธุ์กัญชาเพศผู้นั้นจะโดดเด่นในเรื่องของการขยาย และรักษาพันธุ์มากกว่าเพศเมีย แต่ในเมล็ดพันธุ์กัญชาเพศเมียนั้น จะมีความโดดเด่นในเรื่องของการออกดอก ซึ่งดอกของกัญชาจะมีสาร THC อยู่มากที่สุด ที่พบได้มากที่สุดในกัญชาเพศเมีย ดังนั้น ผู้วิจัยจึงมีความสนใจในการศึกษาในเรื่องของเมล็ดพันธุ์กัญชาเพศเมียนั้นเอง

ผู้วิจัยจึงได้ศึกษา และรวบรวมนำบทความของเมล็ดพันธุ์กัญชาเพศเมียต่างๆ ในเว็บไซต์ dutch-passion.com มาทำการสรุปข้อความ (Text Summarization) แบบ Extractive Summarization โดยทำเป็นขั้นตอนต่างๆ ดังต่อไปนี้ ประการที่หนึ่ง ผู้วิจัยได้ใช้ชุดเครื่องมือประมวลผลภาษาธรรมชาติ (Natural Language Toolkit : NLTK) มาใช้ในการเตรียมข้อมูลได้แก่การ Stopword Removal, Tokenization, POS Tagging และ Lemmatization ประการที่สอง ทำการคำนวณความคล้ายคลึงกันระหว่างคู่ประโยคแต่ละคู่โดยใช้อัลกอริธึม (Algorithm) TextRank และ BM25 ประการที่สาม คือ การสร้างกราฟ (Graph) โดยพิจารณาแต่ละประโยคเป็นโหนด (Node) และความคล้ายคลึงกันเป็นน้ำหนักขอบ และประการสุดท้าย คือ ทำการสรุปข้อความโดยการจัดอันดับแต่ละประโยคโดยใช้อัลกอริธึม (Algorithm) TextRank และเลือกประโยค 20% ของอันดับแรก เพื่อทำการสร้างสรุปข้อความ โดยผลที่ได้นั้น ข้อความในบทความของสายพันธุ์เมล็ดกัญชาต่างๆ ในเว็บไซต์ dutch-passion.com มีความกระชับขึ้น จำนวนคำ และจำนวนประโยคลดลง

ศัพท์ที่สำคัญ—เมล็ดพันธุ์กัญชาเพศเมีย, Algorithm, BM25, Extractive Summarization, Lemmatization, Node, POS Tagging, Stopword Removal, TextRank, Text Summarization และ Tokenization

บทที่ 1
บทนำ

1.1 ความเป็นมา และความสำคัญของปัญหา

กัญชา เป็นพืชในวงศ์ CANNABACEAE มีชื่อวิทยาศาสตร์ คือ Cannabissatival โดยลักษณะทางพฤกษศาสตร์ของต้นกัญชา เป็นพรรณไม้พืชล้มลุก ฤดูเดียว (ปีเดียว) จัดเป็นหนึ่งในสมุนไพรที่เก่าแก่ที่สุดในโลก ต้นมีความสูงได้ตั้งแต่ 1-3 เมตร ลำต้นมีลักษณะเป็นเหลี่ยมตั้งตรงมีขนาดเล็ก มีขนสีเขียวอมเทา ขยายพันธุ์โดยใช้เมล็ด ในประเทศไทยพบการปลูกมากเป็นแบบแยกเพศ ดอกเพศผู้ และดอกเพศเมีย จะแยกกันอยู่คนละต้น ออกดอกเป็นช่อที่ง่ามใบหรือปลายกิ่ง ดอกมีสีเหลือง หรือสีเขียว (สุรศักดิ์อิ่มเอี่ยม และคณะ,2562) กัญชาเป็นหนึ่งในพืชที่มนุษย์เพาะปลูกกันมาเป็นเวลานาน เพื่อใช้ประโยชน์จากเส้นใย และทำเป็นยา กัญชาได้รับการยอมรับจากการแพทย์แผนตะวันตกอย่างแพร่หลายในศตวรรษที่ 19 เนื่องจากมีการทดลองอย่างเป็นระบบ [1] โดยมีหลักฐานการใช้กัญชาทางด้านการแพทย์ที่พบในหลายประเทศ ทั้งประเทศจีน อินเดีย เปอร์เซีย และประเทศแถบยุโรป รวมถึงประเทศไทย ซึ่งได้รับอิทธิพลการใช้กัญชามาจากประเทศอินเดีย ในเรื่องของการนำมาประกอบเป็นเครื่องเทศปรุงอาหาร และการใช้เป็นยารักษาโรค (Woodbridge,2020) ซึ่งตามตำราสรรพคุณยาไทย ระบุว่ากัญชามีรสเมาเบื่อ มีสรรพคุณแตกต่างกันตามส่วนที่ใช้ เช่น ใบมีสรรพคุณแก้หอบหืด เจริญอาหาร ชูกำลัง แต่ทำให้จิตใจขลาดกลัว ตาลาย ประสาทหลอน ดอกมีสรรพคุณแก้โรคประสาท ทำให้นอนหลับ เจริญอาหาร ละลายเสมหะในลำคอ เป็นต้น และยังพบรายงานเกี่ยวกับการใช้ประโยชน์ทางการแพทย์ เช่น การรักษามะเร็ง เป็นยากันชัก ลดอาการปวด

ในต้นศตวรรษที่ 20 มีการผ่ากฎหมายยาเสพติดภาษีกัญชา ทำให้งานวิจัยเกี่ยวกับกัญชาในประเทศต่างๆ ได้หยุดชะงักลง แต่อย่างไรก็ตามในปี ค.ศ. 1964 มีการค้นพบสารออกฤทธิ์ในกัญชา Tetrahydrocannabinol (THC) ซึ่งออกฤทธิ์ต่อตัวรับ Cannabinoid 1 (CB_1) และ Cannabinoid 2 (CB_2) ซึ่งพบมากในช่อดอกของต้นกัญชาตัวเมียที่ยังไม่ได้ผสมพันธุ์ ที่มีอยู่ในระบบประสาท และระบบภูมิคุ้มกัน [2]ทำให้มีความสนใจทางด้านเภสัชวิทยาของกัญชาเพิ่มขึ้น ได้มีกระแสในการผลักดันกัญชาให้ถูกกฎหมาย เพื่อสามารถนำมาใช้ประโยชน์ทางการแพทย์และความบันเทิงได้ ซึ่งในงานวิจัยพบว่ากัญชามีประสิทธิภาพในการบรรเทาอาการปวดเรื้อรัง แก้คลื่นไส้อาเจียนจากยาเคมีบำบัด กระตุ้นความอยากอาหาร บรรเทาอาการเกร็งจากโรคปลอกประสาทเสื่อมแข็ง (Multiple Sclerosis) และรักษาโรคลมชักบางชนิด อย่างไรก็ตามการเปิดเสรีกัญชานั้นเพิ่มความเสี่ยงต่อสุขภาพกาย และสุขภาพจิตบางประการต่อสังคมได้ ใน

---

[1] นพดล หงษ์สุวรรณ, รัตนา อินทเกตุ, อาทิตยา โคตรสมบัติ และบุษราภรณ์ ทับสีแก้ว. **"เรื่อง ภูมิปัญญาการวิเคราะห์คุณลักษณะภายนอก เพื่อบ่งชี้เพศของต้นกัญชา ในพื้นที่เขตจังหวัดสกลนคร"**, วารสารวิทยาศาสตร์ และเทคโนโลยี มหาวิทยาลัยราชภัฏอุดรธานี ปีที่ 10 ฉบับที่ 1, 2565

[2] นพ.ธนน คงเจริญสมบัติ, สาขาวิชาโภชนาการคลินิก ภาควิชาอายุรศาสตร์ คณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย. **"เรื่อง กัญชา จากอดีตสู่ปัจจุบัน"**, วารสารโภชนบำบัด (Thai JPEN) ปีที่ 27 ฉบับที่ 2, กรกฎาคม - ธันวาคม 2562

ประเทศไทยนั้น ที่จังหวัดสกลนคร จังหวัดที่ได้ถูกจัดให้เป็นมหานครแห่งพฤกษเวช เป็นอีกหนึ่งพื้นที่ที่ขึ้นชื่อว่ามีกัญชาที่มีคุณภาพดีที่สุด ชื่อว่ากัญชาไทยสายพันธ์หางกระรอก (Thai Stick) ที่เป็นสายพันธุ์แท้ดั้งเดิมขึ้นอยู่ตามธรรมชาติบนเทือกเขาภูพาน จังหวัดสกลนคร มาตั้งแต่สมัยโบราณ ก่อนที่จะถูกกำหนดให้เป็นสิ่งเสพติดให้โทษประเภทที่ 5 จนมีกฎหมายนิรโทษกรรมสำหรับผู้ป่วยที่มีความจำเป็นต้องใช้กัญชาเพื่อรักษาโรคเฉพาะ จึงจะได้รับอนุญาตให้ปลูกกัญชาได้ และรวมถึงอนุญาตให้สามารถนำกัญชามาทำการวิจัย เพื่อทำเป็นยารักษาโรคได้ แต่ยังคงให้กัญชาเป็นยาเสพติดให้โทษ (ราชกิจจานุเบกษา 2562 : 7-8)

ประเทศไทย ได้มีการประกาศใช้กฎกระทรวงสาธารณสุข ลงวันที่ 30 สิงหาคม พ.ศ. 2562 เรื่อง ระบุชื่อยาเสพติดให้โทษในประเภทที่ 5 (ฉบับที่ 2) ขึ้น โดยสาระสำคัญของประกาศดังกล่าวนั้น คือการปรับปรุงนิยามของกัญชา โดยยกเว้นให้กัญชง สารสกัด Cannabidiol (CBD) เมล็ดกัญชง และน้ำมันจากเมล็ดกัญชง ตามเงื่อนไขที่กำหนด ไม่ตกอยู่ภายใต้นิยามของกัญชา ที่จัดว่าเป็นยาเสพติดให้โทษ กล่าวโดยสรุปคือในปัจจุบันกฎหมายไทยอนุญาตให้การใช้กัญชา ในการศึกษาวิจัย และการบำบัดรักษาโรคนั้น สามารถกระทำได้ ซึ่งทั้งนี้ต้องกระทำภายใต้หลักเกณฑ์ที่กฎหมายกำหนดเท่านั้น อย่างไรก็ดีการเสพ และครอบครองกัญชาเพื่อความสันทนาการส่วนบุคคลนั้น ยังคงเป็นความผิดตามพระราชบัญญัติยาเสพติดให้โทษ พ.ศ. 2522 เช่นเดิม และประเทศไทยในปัจจุบัน ประกาศกระทรวงสาธารณสุข เรื่อง ระบุชื่อยาเสพติดให้โทษในประเภทที่ 5 พ.ศ. 2565 มีผลบังคับใช้เมื่อวันที่ 9 มิถุนายน พ.ศ. 2565 ส่งผลให้ทุกส่วนของพืชกัญชา และกัญชงนั้น ไม่เป็นยาเสพติด ยกเว้นสารสกัดที่มีปริมาณสาร Tetrahydrocannabinol (THC) เกิน 0.2 % ที่ยังเป็นยาเสพติด ประกาศฉบับนี้ถูกประกาศลงราชกิจจานุเบกษาตั้งแต่วันที่ 9 กุมภาพันธ์ พ.ศ. 2565

แต่อย่างไรก็ตาม การร่างพระราชบัญญัติกัญชา และกัญชง ยังอยู่ระหว่างการพิจารณาของสภาผู้แทนราษฎร โดยผ่านวาระที่หนึ่ง คือ ขั้นรับหลักการ เมื่อวันที่ 8 มิถุนายน พ.ศ. 2565 ทำให้เกิดปัญหาในการคุ้มครองดูแลผู้บริโภค สาธารณสุข และหน่วยงานที่เกี่ยวข้องจึงต้องพลิกข้อกฎหมายอื่นๆ ขึ้นมาออกประกาศตามอำนาจของตน เพื่ออุดช่องว่าง ช่องโหว่ในระหว่างนี้ โดยบางฉบับออกมานานแล้ว แต่ประชาชนอาจยังไม่ได้รับข้อมูลข่าวสารมากนัก จึงปฏิบัติตัวไม่ถูกต้องในยุคของกัญชาเสรี ซึ่งขณะที่บางฉบับนั้น เพิ่งมีการออกมาหลังวันที่ 9 มิถุนายน พ.ศ. 2565

ตลาดกัญชานั้น ได้เติบโตขึ้นอย่างต่อเนื่องจากการผ่อนปรนทางกฎหมายของประเทศใหญ่ๆ ในตะวันตก เช่น แคนาดา และสหรัฐอเมริกาในบางรัฐ โดยมีการคาดการณ์ว่าตัวเลขของตลาดกัญชาอาจมีมูลค่าสูงถึง 7.5 หมื่นล้านดอลลาร์ และที่สำคัญอาจมีมูลค่าตลาดแซงหน้าอุตสาหกรรมน้ำอัดลมได้ในปี ค.ศ. 2030 อุตสาหกรรมกัญชากำลังได้รับความนิยมในหมู่ผู้ประกอบการ และนักลงทุนในหลายประเทศ ซึ่งหันมาจับธุรกิจ และลงทุนในธุรกิจกัญชาอันเป็นผลมาจากการปฏิวัติกฎหมายเกี่ยวกับกัญชาในหลายประเทศ ในปี ค.ศ. 2017

[3] งานศึกษาของเว็บไซต์หางานในอเมริกาพบว่า การหารับสมัครงานในวงการกัญชาสูงกว่าวงการไอทีเกือบเท่าตัว โดยปัจจัยหลักๆ คือ การทำให้กัญชาถูกกฎหมายในหลายรัฐ โดยเฉพาะรัฐ California ปัจจุบันคนทำงานในอุตสาหกรรมกัญชาในอเมริกาอย่างถูกกฎหมาย มีจำนวน 230,000 คน ส่วนในปี ค.ศ. 2021 เป็นปีที่หลายฝ่ายคาดการณ์ว่ากัญชาจะถูกต้องตามกฎหมายในทุกรัฐของสหรัฐอเมริกา และมูลค่าของอุตสาหกรรมนี้จะสูงขึ้นไปถึง 21 ล้านดอลลาร์ และมากกว่านั้นจะเกิดการจ้างงานกว่า 413,988 ตำแหน่งอีกด้วย ในยุคที่มีปัญญาประดิษฐ์ (AI) ที่มีหุ่นยนต์เข้ามาแย่งงานมนุษย์ทำ ด้วยหุ่นยนต์จะเข้ามาแทนที่การทำงานของมนุษย์อย่างแน่นอน ถึงแม้จะไม่ใช่ทุกสายงาน แต่กัญชากำลังจะเป็นอุตสาหกรรมที่ยิ่งใหญ่ในไม่ช้านี้ โดยที่สำคัญ คือ เกิดการจ้างงานคนจริงๆ ทั้งนี้อัตราการจ้างงานด้านการค้าโดยเฉลี่ย 21% ต่อปี ไปจนถึงปี พ.ศ. 2565 ไม่เพียงแค่จำนวนคนซื้อที่เพิ่มขึ้น แต่จะทำให้เรื่องเศรษฐกิจเติบโต แต่เมื่อมองตามหลักเศรษฐศาสตร์ ธุรกิจก็จะต้องมีการขยายตัว เกิดการจ้างงาน ซึ่งต้องเพิ่มตามตัวคิดเป็น 21% ต่อปี เมื่อเปรียบเทียบกับอุตสาหกรรมที่น่าจับตามองในอนาคตอย่างเช่นด้านสุขภาพ ที่ต่างก็มีมุมมองว่าแรงในยุคที่คนหันมาใส่ใจกับสุขภาพกลับพบว่ามีตัวเลขห่างกันมาก เพราะคาดคะเนว่าอุตสาหกรรมด้านสุขภาพจะเติบโตเฉลี่ยแค่ 2% ต่อปีเท่านั้น สำหรับพื้นที่กัญชาเสรีพืชชนิดนี้ สามารถยกคุณภาพชีวิตการสร้างงานให้กับคนอีกหลายแสนคน กระแสความนิยมในสาร Cannabidiol (CBD) ที่เพิ่มขึ้นในหลายประเทศทั่วโลก ความก้าวหน้าทางเทคโนโลยีในการแปรรูป รวมถึงภาพลักษณ์ของกัญชาที่เป็นพืชสมุนไพรโบราณชนิดหนึ่ง เป็นปัจจัยที่ทำตลาดกัญชาในอเมริกาขยายตัวเพิ่มขึ้นอีกในอนาคต โดยจะมีผลิตภัณฑ์ประเภทใหม่ๆ ออกสู่ตลาด เพิ่มขึ้น โดยเฉพาะผลิตภัณฑ์ที่มีภาพลักษณ์เป็นประโยชน์ต่อสุขภาพ และผลิตภัณฑ์ที่มีภาพลักษณ์ในเรื่องความยั่งยืน และเป็นมิตรกับสิ่งแวดล้อม ซึ่งประเทศชั้นนำทั่วโลกคาดการ์มูลค่าตลาดกัญชาทั่วโลกจะมีแนวโน้มเติบโตต่อเนื่อง หรือคิดเป็นมูลค่ากว่า 103.9 พันล้านดอลลาร์สหรัฐฯ ในปี พ.ศ. 2567 แบ่งเป็นตลาดกัญชาเพื่อการแพทย์มีสัดส่วนราวร้อยละ 60 ของมูลค่าตลาดกัญชาทั้งหมด และอีกร้อยละ 40 เป็นตลาดกัญชาเพื่อการสันทนาการ

อย่างไรก็ตามในหลายประเทศ กัญชายังคงเป็นพืชที่ถูกควบคุมอย่างเข้มงวดตามกฎหมาย Opium Act ปี ค.ศ. 2002 ซึ่งการปลูก และการแปรรูปมีเพียงหน่วยงานควบคุม คือ [4] หน่วยงาน Office of Medicinal Cannabis (OMC) ประเทศเนเธอร์แลนด์ ที่จะสามารถดำเนินการได้ ซึ่งเป็นหน่วยงานราชการ มีหน้าที่รับผิดชอบเกี่ยวกับการผลิตกัญชา เพื่อใช้ในทางการแพทย์ และการวิจัยทางวิทยาศาสตร์ รวมทั้งป้องกันการรั่วไหลกัญชา และยังมีนโยบาย Coffee Shop สามารถใช้กัญชา เพื่อนสันทนาการ ในบางพื้นที่ที่กำหนดไว้เท่านั้น มีการอนุญาตให้ปลูกกัญชาตามวัตถุประสงค์ คือ เพื่อศึกษาวิจัย หรือเพื่อผลิตผลิตภัณฑ์กัญชา ไม่กำหนดเรื่องขนาดของพื้นที่ มีระบบการรักษาความปลอดภัย ระบบควบคุม (Access Control) การเข้าถึงพื้นที่ปลูก และประตูทางเข้าพื้นที่เพาะปลูก ขั้นตอนควบคุมการปฏิบัติงานเกี่ยวกับการปลูกกัญชา เพื่อป้องกันการรั่วไหล

การปลูกกัญชา และการเลือกเมล็ดพันธุ์ ยังเป็นเรื่องใหม่กับบางประเทศ ผู้วิจัยจึงรวบรวม สืบค้นว่าเว็บไซต์ใด มีข้อมูลที่ครอบคลุมทุกสายพันธุ์ของกัญชา ซึ่งมีเว็บไซต์หนึ่งชื่อว่า Dutch Passion (dutch-passion.com) เป็นเว็บไซต์หนึ่งในธนาคารเมล็ดพันธุ์กัญชาที่เก่าแก่ที่สุดในโลก และนำเสนอพันธุ์กัญชาคลาสสิกดั้งเดิม และพันธุ์ใหม่ที่ดีที่สุดแก่ลูกค้า ก่อตั้งในปี ค.ศ. 1970 และได้รับการจัดตั้งขึ้นอย่างเป็นทางการในฐานะธนาคารเมล็ดพันธุ์ในปี ค.ศ. 1987 ผู้ทำวิจัยจึงได้ใช้เทคนิคการสรุปแบบแยกส่วน (Extractive summarization)โดยใช้ Text Rank และ BM25 เพื่อให้อ่านบทความได้ง่ายขึ้น เพราะเมล็ดสายพันธุ์กัญชานั้นมีอยู่มากมายหลากหลาย และมีข้อมูลทางคุณสมบัติ หรือเรื่องราว ที่แตกต่างกันออกไป จึงเป็นประโยชน์สำหรับผู้ที่ต้องการศึกษา หรือมีความสนใจ ให้ได้ศึกษา หรืออ่านบทความได้เข้าใจง่ายมากยิ่งขึ้น กระชับขึ้น และมีความสะดวกในการอ่านมากขึ้นอีกด้วย

1.2 วัตถุประสงค์ของการศึกษา

เพื่อศึกษาเรื่อง Text Summarization แบบ Extractive Summarization โดยใช้ข้อมูลจากเว็บไซต์ dutch-passion.com ในการศึกษาวิจัย เพื่อช่วยลดระยะเวลาในการอ่านบทความที่มีความยาว ทำให้ข้อความในบทความกระชับขึ้น เข้าใจขึ้น รวมถึงสามารถอ่านบทความได้ง่ายมากยิ่งขึ้น

1.3 ขอบเขตของการศึกษา

การศึกษาครั้งนี้ใช้ข้อมูลจากเว็บไซต์ dutch-passion.com โดยศึกษาจากเมล็ดพันธุ์กัญชาเพศเมียในเว็บไซต์ดังกล่าว

1.4 เครื่องมือที่ใช้ในการวิจัย

การวิจัยเชิงทดลอง (Experimental Research) เป็นการวิจัยที่ผู้วิจัยนั้น ได้ศึกษาหาข้อเท็จจริง จัดสร้างสถานการณ์ คือ การสรุปบทความแบบ Extractive Summarize เพื่อทำการสรุปข้อความที่ยาวให้มีความกระชับ เข้าใจง่าย และอ่านง่ายขึ้น โดยยังจะคงความหมายในบทความเดิมไว้ และรวบรวมบทความ และเงื่อนไขต่างๆ จากเว็บไซต์ที่เกี่ยวข้องกับสายพันธุ์เมล็ดกัญชาเพศเมียขึ้นมาทดลอง คือ เว็บไซต์ dutch-passion.com ด้วยการทดสอบภายใต้การควบคุมตัวแปรที่เกี่ยวข้องอย่างมีระเบียบแบบแผน คือการใช้ BM25 ร่วมกับ Text Rank และมีวัตถุประสงค์ที่แน่นอน เพื่อวัดผลการทดลองออกมา และสามารถกระทำซ้ำเพื่อพิสูจน์

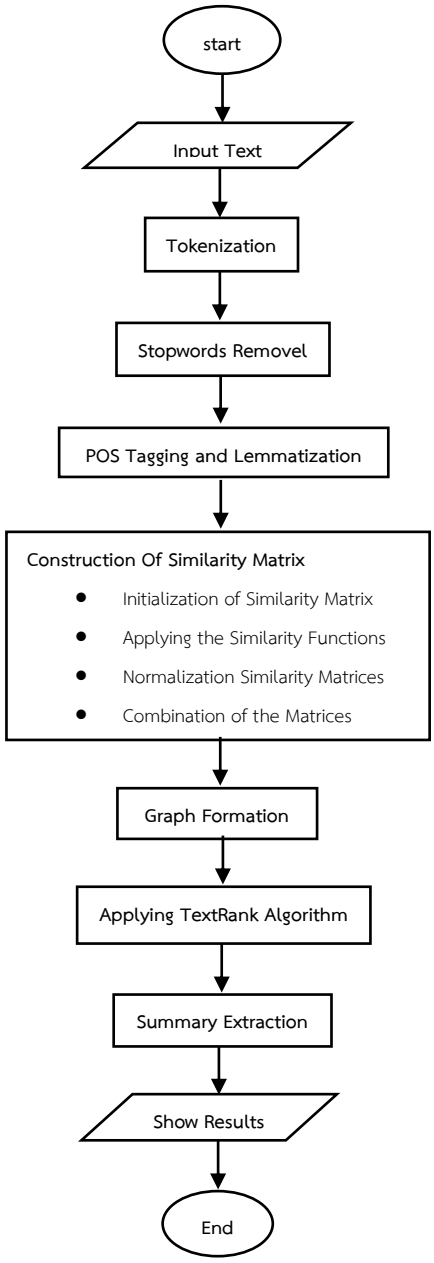
---

[3] อภิวัฒน์ จ่าตา. **“เรื่อง กัญชา : มิติพืชเศรษฐกิจ Cannabis dimensional plants”**

[4] สุนทร พุทธศรีจารุ. **“เรื่อง การพัฒนามาตรการทางกฎหมายควบคุมการใช้กัญชาทางการแพทย์ และการนำไปสู่การปฏิบัติ”**, วารสารอาหารและยา ฉบับเดือนพฤษภาคม-สิงหาคม 2562 (บทความวิจัย), สังกัดกองควบคุมวัตถุเสพติด สำนักงานคณะกรรมการอาหารและยา, 15 มีนาคม 2562

หรือทดสอบอีกได้ โดยรันโปรแกรมในภาษา Python ผ่าน Google Colab ซึ่ง Google Colab เป็นลิขสิทธิ์ที่บุคคลทั่วไปสามารถเข้าไปใช้งานได้

1.5 กรอบแนวคิด



ภาพที่ 1 : แผนผังแสดงขอบเขตการศึกษาวิจัย

บทที่ 2
แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

จากการศึกษาค้นคว้าข้อมูลเอกสาร พบว่าในการอ่านข้อความที่ยาว และมีจำนวนมาก ทำให้ใช้เวลาในการอ่านที่ยาวนาน ผู้วิจัยจึงใช้เทคนิคการสรุปข้อความแบบ Extractive Summarization โดยใช้ Text Rank และ BM25 ซึ่งผู้วิจัยได้ศึกษาแนวคิด ทฤษฎี รวมถึงเอกสารที่เกี่ยวข้อง ดังต่อไปนี้

1. การประมวลผลภาษาธรรมชาติ (NLP)
2. การสรุปข้อความอัตโนมัติ (Automatic Text Summarization)
3. ชุดเครื่องมือประมวลผลภาษาธรรมชาติ (Natural Language Toolkit : NLTK)
4. PageRank Algorithm
5. TextRank

2.1 การประมวลผลภาษาธรรมชาติ (NLP)

การประมวลผลภาษาธรรมชาติ (NLP) [5] เป็นการนำความรู้ทางด้านภาษาศาสตร์มาวิเคราะห์รูปแบบโครงสร้างของประโยค ตามหลักไวยากรณ์ และแปลความหมายของคำ เพื่อทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ แล้วนำข้อความนั้น มาเก็บไว้ในฐานความรู้ เพื่อให้คอมพิวเตอร์เกิดการเรียนรู้ และสามารถนำไปสร้างเป็นแบบจำลอง (Model) เพื่อนำไปใช้ประโยชน์ตามความต้องการ

2.2 การสรุปข้อความอัตโนมัติ (Automatic Text Summarization)

การสรุปข้อความอัตโนมัติ (Automatic Text Summarization) เป็นหนึ่งในปัญหาที่ท้าทาย และน่าสนใจที่สุด ในด้านการประมวลผลภาษาธรรมชาติ (NLP) เป็นกระบวนการสร้างสรุปข้อความที่กระชับ และมีความหมายจากแหล่งข้อมูลข้อความต่างๆ เช่น หนังสือ บทความข่าว บล็อกโพสต์ งานวิจัย อีเมล และทวีต โดยความต้องการระบบสรุปข้อความอัตโนมัติในทุกวันนี้ เพิ่มขึ้นอย่างรวดเร็ว เนื่องมาจากความพร้อมของข้อมูลที่เป็นข้อความจำนวนมาก ซึ่งการสรุปข้อความอัตโนมัตินั้น สามารถแบ่งออกเป็น 2 ประเภทกว้างๆ ดังต่อไปนี้

1. Extractive Summarization

วิธีการเหล่านี้อาศัยการแยกส่วนต่างๆ เช่น วลี และประโยค จากข้อความ และนำมารวมกัน เพื่อสร้างบทสรุป ดังนั้น การระบุประโยคที่เหมาะสมสำหรับการสรุปจึงมีความสำคัญสูงสุดในวิธีการแยก

2. Abstractive Summarization

วิธีการเหล่านี้ใช้เทคนิค NLP ขั้นสูง เพื่อสร้างบทสรุปใหม่ทั้งหมด บางส่วนของบทสรุปนี้อาจไม่ปรากฏในข้อความต้นฉบับด้วยซ้ำ

2.3 ชุดเครื่องมือประมวลผลภาษาธรรมชาติ (Natural Language Toolkit : NLTK)

Natural Language Toolkit (NLTK) หรือโดยทั่วไป [6]คือ ชุดของไลบรารี และโปรแกรมสำหรับการประมวลผลภาษาธรรมชาติเชิงสัญลักษณ์ และเชิงสถิติ (NLP) สำหรับภาษาอังกฤษที่เขียนด้วยภาษาโปรแกรม Python โดยมีเครื่องมือที่น่าสนใจดังต่อไปนี้

1. Tokenization

Tokenization เป็นกระบวนการที่มีข้อความจำนวนมากถูกแบ่งออกเป็นส่วนเล็กๆ ที่เรียกว่า Token ซึ่ง Token เหล่านี้มีประโยชน์มากสำหรับการค้นหารูปแบบ และถือเป็นขั้นตอนพื้นฐานสำหรับการแยกส่วน นอกจากนี้การย่อ Tokenization ยังช่วยแทนที่องค์ประกอบข้อมูลที่ละเอียดอ่อน ด้วยองค์ประกอบข้อมูลที่ไม่ละเอียดอ่อน

ส่วน [7] Sent Tokenize คือโมดูลย่อยที่พร้อมใช้งานสำหรับการ Tokenization ของประโยค (Tokenization of Sentences) หากต้องการนับคำเฉลี่ยต่อประโยค จะต้องใช้ทั้งตัวสร้างประโยค NLTK และตัวสร้าง Token คำ NLTK เพื่อคำนวณอัตราส่วน ผลลัพธ์ดังกล่าวทำหน้าที่เป็นคุณสมบัติที่สำคัญสำหรับการฝึกเครื่องจักร เนื่องจากคำตอบจะเป็นตัวเลข

---

[5] นงเยาว์ สอนจะโปะ. **"เรื่อง รูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ ของเครื่อง (Machine Learning ถ้วยเทคนิค Unsupervised Learning รวมกับการประมวลผล ภาษาธรรมชาติ (Natural Language Processing)"**, วารสารวิชาการศรีปทุมชลบุรี, ปีที่ 14 ฉบับที่ 4 เมษายน-มิถุนายน, 2561

[6] Mudda Prince. **"Stop Words and Tokenization with NLTK"**, medium.com, 1 ตุลาคม 2562

[7] Daniel Johnson. **"NLTK Tokenize : Words and Sentences Tokenizer with Example"**, guru99.com, 14 พฤษภาคม 2565

2. Stop Word

Stop Word เป็นคำที่ใช้บ่อย เช่น I, a, an, in เป็นต้น คำเหล่านี้ไม่ได้มีส่วนสำคัญต่อเนื้อหาข้อมูลของประโยค ดังนั้นจึงแนะนำให้ลบออกโดยการจัดเก็บรายการคำที่พิจารณาว่าเป็น [8]Stop Word โดย Library NLTK มีรายการดังกล่าวสำหรับ 16 ภาษาที่แตกต่างกันซึ่งสามารถอ้างถึงได้

3. NLTK Lemmatization

[9]NLTK Lemmatization เป็นกระบวนการของการจัดกลุ่มรูปแบบการผันคำของคำเพื่อวิเคราะห์คำเหล่านั้นเป็นคำเดียวหรือ [10] การพิจารณาบริบทและแปลงคำเป็นรูปแบบฐานที่มีความหมาย ซึ่งเรียกว่า การวิเคราะห์ทางสัณฐานวิทยา (Morphological Analysis)

ตัวอย่างการ Lemmatization

‘Caring’ -> Lemmatization -> ‘Care’

Wordnet เป็นฐานข้อมูลคำศัพท์ขนาดใหญ่ ที่เปิดเผยต่อสาธารณะสำหรับภาษาอังกฤษ โดยมีเป้าหมายเพื่อสร้างความสัมพันธ์ทางความหมายที่มีโครงสร้างระหว่างคำ และมีความสามารถเช่นเดียวกับการ Lemmatization ซึ่งเก่าที่สุด และใช้บ่อยที่สุด โดยต้องสร้างอินสแตนซ์ของ WordNetLemmatizer() และเรียกใช้ฟังก์ชัน lemmatize() ในคำเดียว

4. ส่วนของการติดแท็กคำพูด (Parts of Speech Tagging)

การติดแท็ก (POS Tagging : POS) เป็นส่วนหนึ่งของการแท็กคำพูดเป็นกระบวนการในการทำเครื่องหมายคำ ในรูปแบบข้อความ สำหรับส่วนใดส่วนหนึ่งของคำพูด ตามคำจำกัดความ และบริบท มีหน้าที่รับผิดชอบในการอ่านข้อความในภาษา และกำหนด Token เฉพาะ (Parts of Speech) ให้กับแต่ละคำ เรียกอีกอย่างว่าการติดแท็กทางไวยากรณ์

ตัวอย่างการ POS Tagging

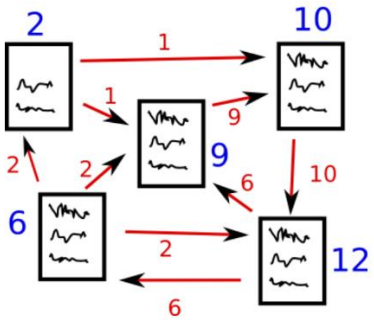
Input	Everything to permit us.
Output	[(‘Everything’, NN), (‘to’, TO), (‘permit’, VB), (‘us’, PRP)]

โดย NN = noun, singular (cat, tree)
TO = infinite marker (to)
VB = verb (ask)
PRP = personal pronoun
(hers, herself, him, himself)

5. Wordnet Lemmatizer และ POS Tagging

ในกระบวนการ POS Tagging นั้นอาจไม่สามารถระบุแท็ก POS ที่ถูกต้องสำหรับทุกคำ สำหรับข้อความขนาดใหญ่ได้ด้วยตนเอง ดังนั้น จึงต้องหาแท็ก POS ที่ถูกต้องสำหรับแต่ละคำแทน และจับคู่กับอักขระ Input ที่ถูกต้อง ที่ WordnetLemmatizer ยอมรับ และส่งผ่านเป็น Argument ที่สองไปยัง lemmatize() ต่อ ซึ่งแสดงให้เห็นถึงความสัมพันธ์ในส่วนของ Wordnet Lemmatizer กับ POS tag นั้นเอง

2.4 PageRank Algorithm

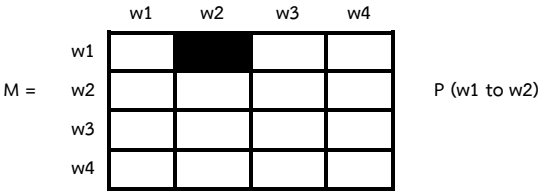


สมมติว่าเรามีหน้าเว็บ 4 หน้า — w1, w2, w3 และ w4 หน้าเหล่านี้มีลิงก์ที่ชี้ไปยังอีกหน้าหนึ่ง บางหน้าอาจไม่มีลิงก์ ซึ่งเรียกว่าหน้าห้อย

Webpage	Links
w1	[w4, w2]
w2	[w3, w1]
w3	[ ]
w4	[w1]

หน้าเว็บ w1 มีลิงก์ที่นำไปยัง w2 และ w4
หน้าเว็บ w2 มีลิงก์สำหรับ w3 และ w1
หน้าเว็บ w4 มีลิงค์สำหรับหน้าเว็บเท่านั้น w1
หน้าเว็บ w3 ไม่มีลิงก์ และด้วยเหตุนี้จึงเรียกว่าหน้าห้อย

ในการจับความน่าจะเป็นของผู้ใช้ที่นำทางจากหน้าหนึ่งไปยังอีกหน้าหนึ่ง เราจะสร้างเมทริกซ์สี่เหลี่ยมจัตุรัส M ซึ่งมี n แถวและ n คอลัมน์ โดยที่ n คือจำนวนหน้าเว็บ โดยแต่ละองค์ประกอบของเมทริกซ์นี้ แสดงถึงความน่าจะเป็นที่ผู้ใช้จะเปลี่ยนจากหน้าเว็บหนึ่งไปอีกหน้าเว็บหนึ่ง ตัวอย่างเช่น เซลล์ที่เน้นสีดำด้านล่างมีความน่าจะเป็นของการเปลี่ยนจาก w1 เป็น w2 นั้นเอง



การเริ่มต้นของความน่าจะเป็นได้อธิบายไว้ในขั้นตอน ดังต่อไปนี้

1) ความน่าจะเป็นที่จะไปจากหน้า i ไปยัง j เช่น M[i][j] เริ่มต้นด้วย 1/(จำนวนลิงก์ที่ไม่ซ้ำในหน้าเว็บ wi)
2) หากไม่มีลิงก์ระหว่างหน้า i และ j ความน่าจะเป็นจะถูกเริ่มต้นด้วย 0
3) หากผู้ใช้เข้ามาที่หน้าที่ห้อยนั้น ถือว่าเขามีแนวโน้มที่จะเปลี่ยนไปใช้หน้าใดก็ได้เท่าๆ กัน ดังนั้น M[i][j] จะเริ่มต้นด้วย 1/(จำนวนหน้าเว็บ)

[8] Michael Fuchs Python. **“NLP-Text Pre-Processing II (Tokenization and Stop Words)”**, michael-fuchs-python.netlify.app, 25 พฤษภาคม 2564

[9] Koray Tuğberk GÜBÜR. **“NLTK Lemmatization : How to Lemmatize Words with NLTK?”**, Python SEO, Holistic SEO, 7 ธันวาคม 2564

[10] Selva Prabhakaran. **“Lemmatization Approaches with Examples in Python”**, machinelearningplus.com, 2 ตุลาคม 2561

ดังนั้น ในกรณีนี้ เมทริกซ์ M จะเริ่มต้นได้ดังนี้

M =

	w1	w2	w3	w4
w1	0	0.5	0	0.5
w2	0.5	0	0.5	0
w3	0.25	0.25	0.25	0.25
w4	1	0	0	0

ดังนั้น [11]ค่าในเมทริกซ์นี้จะได้รับการอัปเดตซ้ำๆ เพื่อให้ได้อันดับของหน้าเว็บ

ซึ่ง PageRank Algorithm มีหลักการคล้ายๆ กับอัลกอริธึม TextRank โดยแทนหน้าเว็บด้วยประโยค และความคล้ายคลึงกันระหว่างสองประโยคใดๆ จะใช้เทียบเท่ากับความน่าจะเป็นในการเปลี่ยนหน้าเว็บและคะแนนความคล้ายคลึงกันจะถูกเก็บไว้ในเมทริกซ์สี่เหลี่ยมจัตุรัส คล้ายกับเมทริกซ์ M ที่ใช้สำหรับ PageRank

2.5 TextRank

[12]TextRank เป็นอัลกอริธึมที่ไม่มีผู้ดูแล (Unsupervised Algorithm) สำหรับการสรุปข้อความอัตโนมัติ (Automated Summarization) ซึ่งสามารถใช้เพื่อให้ได้คีย์เวิร์ดที่สำคัญที่สุดในเอกสาร ได้รับการแนะนำโดย Rada Mihalcea และ Paul Tarau [13]ซึ่งแนวคิดของอัลกอริธึม TextRank พิจารณาจากความสำคัญของคำ ขึ้นอยู่กับจำนวนโหวตที่ได้รับ และความสำคัญของคำอื่นๆ ที่โหวตให้ คล้ายกับแนวคิดของอัลกอริธึม PageRank โดยอัลกอริธึม TextRank ใช้รูปแบบ PageRank เหนือกราฟที่สร้างขึ้นเฉพาะ สำหรับงานสรุป ซึ่งสิ่งนี้ทำให้เกิดการจัดอันดับขององค์ประกอบในกราฟ โดยองค์ประกอบที่สำคัญที่สุด คือ องค์ประกอบที่อธิบายข้อความได้ดีกว่า วิธีนี้ช่วยให้ TextRank สร้างบทสรุปโดยไม่จำเป็นต้องใช้คลังข้อมูลการฝึกอบรม หรือการติดป้ายกำกับ และอนุญาตให้ใช้อัลกอริทึมในภาษาต่างๆ

บทที่ 3
วิธีการดำเนินงาน

3.1 แหล่งที่มาของข้อมูล

การรวบรวม และการศึกษาข้อมูลต่างๆ นั้น ผู้วิจัยได้ใช้การศึกษาการสรุปบทความแบบ Extractive Summarize โดยใช้ Text Rank และ BM25 ผู้วิจัยได้รวบรวมข้อมูลบทความจากเว็บไซต์ dutch-passion.com โดยศึกษาเฉพาะบทความเมล็ดพันธุ์กัญชาเพศเมียต่างๆ ซึ่งการรวบรวมข้อมูลบทความของแต่ละสายพันธุ์เมล็ดกัญชาเพศเมีย ในเว็บไซต์ดังกล่าวนั้น พบว่ามีบทความสายพันธุ์เมล็ดกัญชาเพศเมียทั้งหมด 49 สายพันธุ์

3.2 ขั้นตอนการดำเนินงานวิจัย

ขั้นตอนการดำเนินการวิจัย เพื่อศึกษาการสรุปบทความแบบ Extractive Summarize โดยใช้ Text Rank และ BM25 กรณีศึกษาบทความสายพันธุ์ต่างๆ ของเมล็ดกัญชาเพศเมีย มีรายละเอียดดังนี้

1. การเตรียมข้อมูล

การเตรียมข้อมูลเป็นกระบวนการสำคัญอย่างยิ่ง ที่ต้องใช้ความรู้ด้านการประมวลผลภาษาธรรมชาติ (NLP) โดยผู้วิจัยได้ใช้ชุดเครื่องมือประมวลผลภาษาธรรมชาติ (Natural Language Toolkit : NLTK) มาใช้ในการเตรียมข้อมูล เพื่อให้สามารถนำข้อมูลด้านภาษาไปวิเคราะห์ต่อไปได้ โดยในงานวิจัยมีขั้นตอนดังนี้

1) Tokenization

ผู้วิจัยได้ทำกระบวนการ Tokenization โดยใช้โมดูลย่อย Sent Tokenize ในการแบ่ง Token และทำการแปลงตัวอักษรให้เป็นตัวพิมพ์เล็กให้หมด

2) Stopword Removal

ผู้วิจัยได้ทำกระบวนการ Stop Word คำที่ไม่ได้มีส่วนสำคัญต่อเนื้อหาข้อมูลของประโยค

3) POS Tagging and Lemmatization

ผู้วิจัยได้ทำกระบวนการแปลงคำต่างๆ ให้อยู่ในรูปแบบพื้นฐานของคำนั้นๆ โดยการ Tag คำก่อน เป็นการอ้างอิงคำ เพื่อให้การแปลงคำนั้นสมบูรณ์

2. อัลกอริธึม TextRank (TextRank Algorithm)

เริ่มจากแปลงข้อความมาเป็นกราฟ โดยใช้ประโยคเป็นโหนดซึ่งจำเป็นต้องใช้ฟังก์ชันในการคำนวณความคล้ายคลึงของประโยค เพื่อสร้างขอบระหว่าง ซึ่งใช้เพื่อถ่วงน้ำหนักขอบกราฟ หากยิ่งมีความคล้ายคลึงกันระหว่างประโยคมากเท่าใด ขอบระหว่างทั้งสองก็จะมีความสำคัญมากขึ้นในกราฟ ซึ่งสามารถบอกได้ว่ามีแนวโน้มที่จะเปลี่ยนจากประโยคหนึ่งไปอีกประโยคหนึ่ง หากคล้ายกันมาก โดยมีดังต่อไปนี้

1) เมทริกซ์ความคล้ายคลึงกันแบบผสม
(Combination Similarity Matrix)

โดยการคำนวณความคล้ายคลึงของประโยคนั้น เกิดจากการรวมกันระหว่างการคำนวณความคล้ายคลึงกันแบบดั้งเดิม และ BM25 ดังต่อไปนี้

1.1) ฟังก์ชันความคล้ายคลึงกันแบบดั้งเดิม

กำหนดฟังก์ชันความสัมพันธ์ของความคล้ายคลึงกันระหว่างสองประโยค ซึ่งตามเนื้อหาประโยคที่ทั้งคู่แบ่งปัน โดยคำนวณจากจำนวน Token คำศัพท์ระหว่างประโยค หารด้วยความยาวของแต่ละรายการ เพื่อหลีกเลี่ยงประโยคที่ยาวขึ้น

กำหนดฟังก์ชันความคล้ายคลึงกันแบบดั้งเดิม :

$$Sim(s_i, s_j) = \frac{|\{w_n | w_n \in s_i \& w_n \in s_j\}|}{\log(|s_i|) + \log(|s_j|)}$$

โดย s_i, s_j เป็นสองประโยคที่แสดง

และให้ n เป็นชุดของคำใน

$$s_i = w_1^i + w_2^i + \cdots + w_n^i$$

1.2) BM25

BM25 / Okapi-BM25 เป็นฟังก์ชันการจัดอันดับที่ใช้กันอย่างแพร่หลาย ในฐานะเป็นเครื่องมือสำหรับงานดึงข้อมูล ซึ่ง BM25 เป็นรูปแบบของ TF-IDF โดยใช้แบบจำลองความน่าจะเป็น

---

[11] Prateek Joshi. **"An Introduction to Text Summarization using the TextRank Algorithm (With Python implementation)"**, www.analyticsvidhya.com, 1 ธันวาคม 2561

[12] Federico Barrios, Federico López, Luis Argerich และ **Rosa Wachenchauzer."Variations of the Similarity Function of TextRank for Automated Summarization"**, Cornell University, arxiv.org, 11 กุมภาพันธ์ 2559

[13] Wengen Li and Jiabao Zhao. **"TextRank Algorithm by Exploiting Wikipedia for Short Text Keywords Extraction"**, School of Management and Engineering, Nanjing University, 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), กรกฎาคม 2559

กำหนดให้ BM25 :

$$BM25(D,S) = \sum_{i=1}^{n} IDF(s_i) \cdot \frac{f(s_i,D) \cdot (k_1+1)}{f(s_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgDL})}$$

โดย D และ S เป็นสองประโยค
k และ b เป็นพารามิเตอร์

$f(s_i, D)$ คือ ความถี่ของเทอม s_i ในประโยค D

|D| คือ ความยาวของคำในประโยค D

และ avgDL คือ ความยาวเฉลี่ย
ของประโยคในคอลเลกชัน

และ $IDF(s_i)$ คือ สูตรการแก้ปัญหา
คำปรากฏในเอกสารมากกว่าครึ่ง
ของคอลเลกชัน

โดยกำหนดให้ :

$$IDF(s_i) = \ln\left(\frac{N - n(s_i) + 0.5}{n(s_i) + 0.5}\right) + 1$$

โดย avgIDF คือ IDF
(เฉลี่ยสำหรับเงื่อนไขทั้งหมด)

N คือ จำนวนเอกสารทั้งหมด
ในคอลเลกชัน

$n(s_i)$ คือ จำนวนเอกสารที่มี s_i

2) การประยุกต์อัลกอริทึม PageRank
(Applying the PageRank Algorithm)

ก่อนที่จะดำเนินการเลือกประโยคสรุป ให้แปลงค่าที่ได้จากเมทริกซ์ความคล้ายคลึงกันแบบผสมเป็นกราฟ บนกราฟนี้เราจะใช้อัลกอริทึม PageRank เพื่อคำนวณความสำคัญของแต่ละจุดยอด โดยโหนดของกราฟนี้จะเป็นตัวแทนของประโยค และขอบจะแสดงถึงคะแนนความคล้ายคลึงกันระหว่างประโยค ซึ่งประโยคที่มีความหมายมากที่สุดจะถูกเลือก และนำเสนอในลำดับเดียวกันกับที่ปรากฏในเอกสารเป็นบทสรุป

บทที่ 4
ผลการวิจัย

ผู้วิจัยได้ทำการทดลองเลือกประโยคจาก 10%, 20% และ 30% ของอันดับแรกที่ได้จากการทำ TextRank ซึ่งพบว่าการทดลองเลือกประโยคจาก 10%, 20% และ 30% นั้น ได้จากประโยคที่มีความสำคัญจากบทความสายพันธุ์ต่างๆ ของเมล็ดกัญชาเพศเมีย เพียง 10%, 20% และ 30% เรียงตามลำดับ และผู้วิจัยได้ทำการประเมินผลคะแนน ROUGE (ROUGE Score) โดยใช้ ROUGE-1 ในการคำนวณ โดยสรุปได้เป็น 3 ประเด็น ดังต่อไปนี้

1) การทดลองเลือกประโยคจาก 10% มี Precision=1.0 หมายความว่า 100% ของ n-grams ในการสรุปบทความที่สร้างขึ้นนั้น มีอยู่ในบทความต้นฉบับ และ recall=0.15450928381962864 หมายความว่า 15% ของ n-grams ในบทความต้นฉบับมีอยู่ในบทความที่สรุป
2) การทดลองเลือกประโยคจาก 20% มี precision=1.0 หมายความว่า 100% ของ n-grams ในการสรุปบทความที่สร้างขึ้นนั้น มีอยู่ในบทความต้นฉบับ และ recall=0.28050397877984085 หมายความว่า 28% ของ n-grams ในบทความต้นฉบับมีอยู่ในบทความที่สรุป
3) การทดลองเลือกประโยคจาก 30% มี precision=1.0 หมายความว่า 100% ของ n-grams ในการสรุปบทความที่สร้างขึ้นนั้น มีอยู่ในบทความต้นฉบับ และ recall=0. 35610079575596815 หมายความว่า 36% ของ n-grams ในบทความต้นฉบับมีอยู่ในบทความที่สรุป

บทที่ 5
สรุปผลการศึกษา อภิปรายผล และข้อเสนอแนะ

5.1 อภิปรายผลการศึกษา

การสรุปบทความแบบ Extractive Summarize โดยใช้ Text Rank และ BM25 กรณีศึกษาบทความสายพันธุ์ต่างๆ ของเมล็ดกัญชาเพศเมีย จากเว็บไซต์ dutch-passion.com นั้น เป็นเทคนิคที่เข้าใจง่ายโดยใช้ความรู้ด้านการประมวลผลภาษาธรรมชาติ (NLP) และชุดเครื่องมือประมวลผลภาษาธรรมชาติ (Natural Language Toolkit : NLTK) ในขั้นตอนการเตรียมข้อมูล และใช้อัลกอริธึม TextRank ร่วมกับฟังก์ชันการจัดอันดับการดึงข้อมูล BM25 ในการสรุปบทความ

โดยผู้วิจัยได้ใช้การเลือกประโยคจาก 20% อันดับแรกจากการทำ TextRank เพราะเป็นจำนวนประโยคที่มีความสำคัญ ซึ่งไม่เยอะ และไม่น้อยจนเกินไป ซึ่งผู้วิจัยได้ทำการประเมินคะแนน ROUGE (ROUGE Score) พบว่าการทดลองเลือกประโยคจาก 20% มี precision=1.0 หมายความว่า 100% ของ n-grams ในการสรุปบทความที่สร้างขึ้นนั้นมีอยู่ในบทความต้นฉบับ และ recall=0.28050397877984085 หมายความว่า 28% ของ n-grams ในบทความต้นฉบับมีอยู่ในบทความที่สรุปนั้นเอง

จากการศึกษาบทความนั้น ได้พบว่ามีจำนวนข้อความที่ลดลง กระชับขึ้น เนื้อหาอ่านเข้าใจ หมายความว่าอัลกอริธึม TextRank และฟังก์ชัน BM25 สามารถทำงานร่วมกันได้อย่างมีประสิทธิภาพ

5.2 ข้อเสนอแนะ

ในการศึกษาเรื่องที่เกี่ยวกับการประมวลผลภาษาธรรมชาติ (NLP) ควรให้ความสำคัญที่สุดในการเตรียมข้อมูล เพื่อให้เทคนิคที่ทำมีประสิทธิภาพมากที่สุด โดยสำหรับผู้สนใจในเรื่องการสรุปข้อความ หรือบทความแบบ Extractive Summarize โดยใช้ Text Rank ผู้วิจัยแนะนำให้ใช้ฟังก์ชัน BM25 และในการสรุปข้อความ สำหรับการประเมินผลคะแนน ROUGE (ROUGE Score) โดยใช้ ROUGE-1 มีประสิทธิภาพที่ดีในการประเมินผลคะแนนการสรุปข้อความ

บรรณานุกรม

นพดล หงษ์สุวรรณ, รัตนา อินทเกตุ, อาทิตยา โคตรสมบัติ และบุษราภรณ์ ทับสีแก้ว. **“เรื่อง ภูมิปัญญาการวิเคราะห์คุณลักษณะภายนอก เพื่อบ่งชี้เพศของต้นกัญชา ในพื้นที่เขตจังหวัดสกลนคร”**, วารสารวิทยาศาสตร์ และเทคโนโลยี มหาวิทยาลัยราชภัฏอุดรธานี ปีที่ 10 ฉบับที่ 1, 2565

นพ.ธนน คงเจริญสมบัติ, สาขาวิชาโภชนาการคลินิก ภาควิชาอายุรศาสตร์ คณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย. **“เรื่อง กัญชา จากอดีตสู่ปัจจุบัน”**, วารสารโภชนบำบัด (Thai JPEN) ปีที่ 27 ฉบับที่ 2, กรกฎาคม - ธันวาคม 2562

อภิวัฒน์ จ่าตา. **“เรื่อง กัญชา : มิติพืชเศรษฐกิจ Cannabis dimensional plants”**

สุนทร พุทธศรีจารุ. **“เรื่อง การพัฒนามาตรการทางกฎหมายควบคุมการใช้กัญชาทางการแพทย์ และการนำไปสู่การปฏิบัติ”**, วารสารอาหารและยา ฉบับเดือน

พฤษภาคม-สิงหาคม 2562 (บทความวิจัย), สังกัดกองควบคุมวัตถุเสพติด สำนักงานคณะกรรมการอาหารและยา, 15 มีนาคม 2562

นงเยาว์ สอนจะโปะ. **“เรื่อง รูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ ของเครื่อง (Machine Learning ถ้วยเทคนิค Unsupervised Learning รวมกับการประมวลผล ภาษาธรรมชาติ (Natural Language Processing)”**, วารสารวิชาการศรีปทุมชลบุรี, ปีที่ 14 ฉบับที่ 4 เมษายน-มิถุนายน, 2561

Mudda Prince. **“Stop Words and Tokenization with NLTK”**, medium.com, 1 ตุลาคม 2562

Prateek Joshi. **“An Introduction to Text Summarization using the TextRank Algorithm (With Python Implementation)”**, analyticsvidhya.com, 1 ธันวาคม 2561

Federico Barrios, Federico López, Luis Argerich และ Rosa Wachenchauzer**. “Variations of the Similarity Function of TextRank for Automated Summarization”**, Cornell University, arxiv.org, 11 กุมภาพันธ์ 2559

Wengen Li and Jiabao Zhao**. “TextRank Algorithm by Exploiting Wikipedia for Short Text Keywords Extraction”,** School of Management and Engineering, Nanjing University, 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), กรกฎาคม 2559

Koray Tuğberk GÜBÜR. **“NLTK Lemmatization : How to Lemmatize Words with NLTK?”**, Python SEO, Holistic SEO, 7 ธันวาคม 2564

Selva Prabhakaran. **“Lemmatization Approaches with Examples in Python”**, machinelearningplus.com, 2 ตุลาคม 2561

Michael Fuchs Python. **“NLP-Text Pre-Processing II (Tokenization and Stop Words)”**, michael-fuchs-python.netlify.app, 25 พฤษภาคม 2564

Daniel Johnson. **“NLTK Tokenize : Words and Sentences Tokenizer with Example”**, guru99.com, 14 พฤษภาคม 2565