รูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ของเครื่อง (Machine Learning) ด้วยเทคนิค Unsupervised Learning ร่วมกับการประมวลผลภาษาธรรมชาติ (Natural Language Processing)

THAI TEXT AUTOMATIC CLUSTERING MODEL USING MACHINE LEARNING BY UNSUPERVISED LEARNING WITH NATURAL LANGUAGE PROCESSING

นงเยาว์ สอนจะโปะ*

Nongyao Sornjapo

บทคัดย่อ

แหล่งข้อมูลขนาดใหญ่ที่อยู่บนอินเทอร์เน็ตล้วนเป็นข้อมูลที่มีประโยชน์ต่อการนำไปใช้งานทางด้านธุรกิจ การได้มาซึ่งข้อมูลที่สามารถนำมาใช้ประโยชน์ได้จริงจะต้องผ่านกระบวนการวิเคราะห์ข้อมูลทางสถิติและนำอัลกอริทึมเข้ามาช่วย ผ่านการประมวลผลด้วยเครื่องคอมพิวเตอร์ เพื่อเป็นเครื่องมือในการค้นพบองค์ความรู้ที่ซ่อนอยู่ในแหล่งข้อมูล การประมวลผลข้อความภาษาไทยจะมีความยุ่งยากกว่าภาษาอังกฤษ เพราะโครงสร้างประโยคและไวยากรณ์มีความซับซ้อนมากกว่า บทความนี้นำเสนอรูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ของเครื่อง (machine learning) ด้วยเทคนิค Unsupervised Learning ร่วมกับการประมวลผลภาษาธรรมชาติ (natural language processing) โดยรูปแบบดังกล่าวผ่านการสังเคราะห์และพัฒนากระบวนการทำงานแบ่งออกเป็น 3 มอดูล (modules) คือ 1) มอดูลการประมวลผลภาษาธรรมชาติ เป็นการวิเคราะห์โครงสร้างข้อความภาษาไทยให้อยู่ในรูปแบบที่เครื่องคอมพิวเตอร์สามารถนำไปประมวลได้ถูกต้อง 2) มอดูลการเรียนรู้ของเครื่องแบบไม่มีผู้สอน เป็นโมเดลการเรียนรู้สำหรับการจำแนกกลุ่มข้อความภาษาไทยให้เป็นไปอย่างอัตโนมัติ และ 3) มอดูลเหมืองความรู้ คือแหล่งจัดเก็บข้อมูลที่ผ่านการประมวลผลการเรียนรู้ของเครื่องแบบไม่มีผู้สอนในมอดูลที่ 2 ข้อมูลที่จัดเก็บไว้ในเหมืองความรู้จัดว่าเป็นแหล่งขุมทรัพย์ที่มีประโยชน์ต่อการนำไปใช้งานทางด้านธุรกิจ และด้านอื่น ๆ เพราะเป็นแหล่งข้อมูลที่ถูกรวบรวมมาจากทุกช่องทางบนเครือข่ายอินเทอร์เน็ต

คำสำคัญ: การจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ, การประมวลผลภาษาธรรมชาติ, การเรียนรู้ของเครื่อง, การเรียนรู้แบบไม่มีผู้สอน

* อาจารย์ประจำคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม วิทยาเขตชลบุรี

ABSTRACT

Resources available on the internet are crucial because they can be implemented in the business sector. Those really implemented need to be processed by statistical and algorithm analysis as well as computer processes. When searching for reliable sources from the database, it is found that the Thai language seems more problematic than English when the search is being processed. This might be that of the difference of language culture which the Thai language seems more complex. This paper introduces the automatic Thai language classification model using machine learning, unsupervised learning and natural language processing. The mentioned model has been analyzed and developed into three modules, as follows: 1) Natural language processing module is the analysis of Thai text structure in the form that the computer can be processed correctly, 2) Maching unsupervised learning module is the automatic clustering of Thai text, and 3) Knowledge mining module is the data source that has passed the learning process of a uniform, no instructor in module by data stored in knowledge mining. It is a source of valuable treasures for business and other uses as a source of information gathered from all channels on the Internet.

Keywords: automatic Thai text clustering, natural language processing, machine learning, unstructured learning.

บทนำ

ระบบคอมพิวเตอร์และเทคโนโลยีอินเทอร์เน็ตมีการขยายตัวเพิ่มสูงขึ้นอย่างรวดเร็วตลอดระยะเวลาหลายปีที่ผ่านมา และมีแนวโน้มเพิ่มสูงขึ้นเรื่อย ๆ ส่งผลให้เกิดการสร้างและจัดเก็บข้อมูลอิเล็กทรอนิกส์หลายชนิดบนอินเทอร์เน็ต เช่น เว็บไซต์ (website) จดหมายอิเล็กทรอนิกส์ (e-Mail) เอกสารข่าว (news) วิดีโอ ไฟล์เสียง รูปภาพ และไฟล์เอกสารต่าง ๆ (document) ปริมาณข้อมูลและเนื้อหามีความหลากหลายมากขึ้นทำให้ยากต่อการสืบค้นข้อมูล ซึ่งข้อมูลเหล่านี้สามารถนำไปใช้ประโยชน์ทางด้านธุรกิจ เช่น การรับรู้พฤติกรรมการเลือกใช้สินค้าและบริการจากผู้บริโภค รับรู้การแสดงออกทางความคิดเห็นในแง่บวกหรือแง่ลบต่อผลิตภัณฑ์ สามารถนำข้อมูลเชิงลึกมากำหนดกลยุทธ์ในการวางแผนด้านการตลาดหรือการสร้างผลิตภัณฑ์ใหม่ให้กับธุรกิจ การแสดงความคิดเห็นผ่านสื่อสังคมออนไลน์หรือบนเว็บบอร์ดส่วนใหญ่นิยมใช้ภาษาที่ไม่มีโครงสร้างของประโยคที่แน่นอน (unstructured data) หรือเป็นภาษาธรรมชาติ (natural language) ที่ไม่ถูกหลักไวยากรณ์ทางภาษา (กานดา แผ่วัฒนากุล และปราโมทย์ ลือนาม, 2556) ข้อความภาษาไทยจะมีโครงสร้าง

ที่ซับซ้อนกว่าภาษาอังกฤษ ทำให้ยากต่อการวิเคราะห์และการสืบค้น ดังนั้น วิธีการจัดการกับแหล่งข้อมูลขนาดใหญ่ที่อยู่บนอินเทอร์เน็ตเพื่อให้ได้องค์ความรู้ที่ซ่อนอยู่และสามารถนำไปใช้ประโยชน์ได้ตรงตามความต้องการ จะต้องคิดค้นพัฒนากระบวนการจำแนกกลุ่มข้อมูลจากแหล่งข้อมูลขนาดใหญ่ให้เป็นไปแบบอัตโนมัติ เพื่อสามารถจำแนกข้อมูลและนำไปใช้ประโยชน์ได้ตามความต้องการและมีประสิทธิภาพ

ปัจจุบันนักวิจัยให้ความสนใจเกี่ยวกับการจำแนกหมวดหมู่เอกสารภาษาไทยโดยใช้เครื่องคอมพิวเตอร์ประมวลผลแบบอัตโนมัติ การประมวลผลข้อความภาษาไทยให้มีประสิทธิภาพนั้นจะต้องคำนึงถึงลักษณะโครงสร้างของภาษา ปัญหาที่พบเบื้องต้น คือ การตัดคำในภาษาไทย เพราะการเขียนประโยคในภาษาไทยจะเขียนติดกันโดยไม่มีเครื่องหมายวรรคตอนที่แสดงถึงการแบ่งคำที่ชัดเจนเหมือนภาษาอังกฤษ (นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ, 2554) การวิเคราะห์ข้อความภาษาไทยมีขั้นตอนที่ยุ่งยากกว่าภาษาอังกฤษเนื่องจากมีรูปแบบการเขียนที่ไม่ตายตัว งานวิจัยด้านการวิเคราะห์ความคิดเห็นภาษาไทยผ่านสื่อสังคมออนไลน์ หรือการทำเหมืองความคิดเห็น (opinion mining) ได้นำวิธีการประมวลผลภาษาธรรมชาติมาประยุกต์ใช้ร่วมกับการวิเคราะห์เหมืองข้อความ (text mining) เพื่อใช้ในการวิเคราะห์ความคิดเห็นของผู้บริโภคในด้านบวกหรือด้านลบต่อการใช้ผลิตภัณฑ์ (กานดา แผ่วัฒนากุล และปราโมทย์ ลือนาม, 2556; มาสวีร์ มาศดิศรโชติ, 2557) นอกจากนี้มีการศึกษาเกี่ยวกับการนำวิธีการเรียนรู้ของเครื่อง (machine learning: ML) มาประยุกต์ใช้ร่วมกับวิธีการประมวลผลภาษาธรรมชาติเพื่อใช้ในการจัดกลุ่มเอกสาร โดยแบ่งออกเป็น 2 ลักษณะ คือ การจำแนกหมวดหมู่ (classification หรือ categorizarion) และการจัดกลุ่ม (clustering) การจัดหมวดหมู่ของเอกสาร คือ การจำแนกเอกสารออกตามเนื้อหาเอกสาร โดยมีการกำหนดหมวดหมู่ของเอกสารไว้ล่วงหน้า แล้วทำการเปรียบเทียบกับเอกสารต้นฉบับ เอกสารที่มีเนื้อหาตรงกันจะถูกจัดอยู่ในหมวดหมู่เดียวกัน ส่วนวิธีการจัดกลุ่มเอกสาร คือ การแบ่งกลุ่มของเอกสารตามเนื้อหา ซึ่งไม่มีการกำหนดหมวดหมู่หรือกลุ่มของเอกสารไว้ล่วงหน้า จะเป็นการแบ่งกลุ่มตามเนื้อหาและลักษณะของเอกสาร เอกสารที่มีเนื้อหาหรือลักษณะเหมือนกันจะอยู่ด้วยกัน โดยผลลัพธ์ที่ได้จากวิธีการเรียนรู้ของเครื่องนั้น เมื่อเทียบกับการทำด้วยฝีมือมนุษย์ในการจำแนกประเภทเอกสารหรือการจัดหมวดหมู่เอกสารมีความถูกต้องใกล้เคียงกัน ทำให้ประหยัดแรงงานมนุษย์และไม่จำเป็นต้องอาศัยผู้เชี่ยวชาญในการจำแนกเอกสารหรือจัดหมวดหมู่ (Marquez, 2000; Sebastiani, 2002; Galitsky, 2013)

เนื่องจากการจำแนกกลุ่มเอกสารส่วนใหญ่ยังมุ่งเน้นการจำแนกกลุ่มเอกสารหรือการจัดหมวดหมู่เอกสารภาษาอังกฤษ ยังขาดการพัฒนาการจำแนกกลุ่มข้อความภาษาไทย ดังนั้น บทความนี้จะนำเสนอแนวคิดที่สำคัญเกี่ยวกับการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ รวมทั้งทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้อง โดยแสดงวิธีการประมวลผลภาษาธรรมชาติ (NLP) และประเภท

การเรียนรู้ของเครื่อง (ML) เพื่อเป็นกรอบแนวคิดในการนำเสนอรูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ของเครื่อง (ML) ด้วยเทคนิค Unsupervised Learning ร่วมกับการประมวลผลภาษาธรรมชาติ (NLP)

โครงสร้างข้อความภาษาไทย

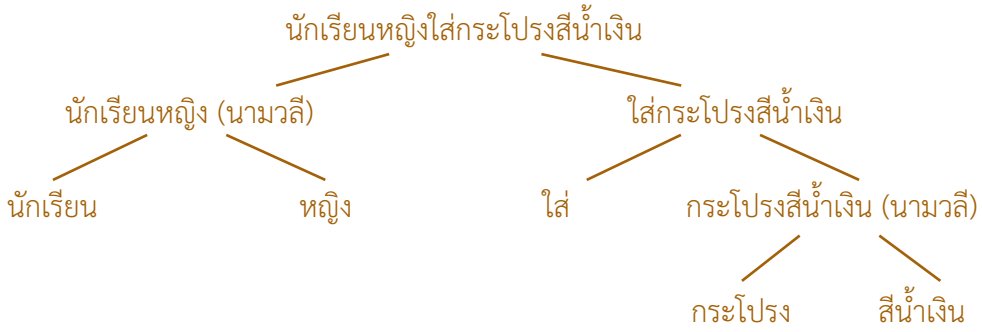
คำว่า “ภาษา” หมายถึง ถ้อยคำที่ใช้พูดหรือเขียน เพื่อสื่อความหมายของชนกลุ่มใดกลุ่มหนึ่ง เช่น ภาษาไทย ภาษาจีน หรือเพื่อสื่อความเฉพาะวงการ เช่น ภาษาราชการ ภาษากฎหมาย เป็นต้น (ราชบัณฑิตยสถาน, 2556) ภาษาไทยเป็นภาษาประจำชาติไทยที่ใช้เป็นเครื่องมือสื่อสารระหว่างกัน เพื่อให้เกิดความเข้าใจตรงกันทั้งผู้รับสารและผู้ส่งสาร ไม่ว่าจะเป็นการแสดงความคิดเห็น ความต้องการและความรู้สึก

คำในภาษาไทยประกอบด้วย เสียง รูปพยัญชนะ สระ วรรณยุกต์ และความหมาย ส่วนประโยคเป็นการเรียงคำตามเกณฑ์ของภาษา คือ ประธาน - กริยา - กรรม และข้อความประกอบด้วยประโยคหลายประโยคเรียงกัน ความสัมพันธ์ทางไวยากรณ์ของคำในหมวดต่าง ๆ ที่ประกอบเข้าเป็นประโยคมีอยู่ 2 ประการ คือ (โศรยา วิมลสถิตพงษ์, 2558)

1. ความสัมพันธ์ในด้านการเรียงลำดับ เป็นการเรียงลำดับคำในวลีหรือประโยค จะต้องเป็นไปตามระเบียบของภาษานั้น ๆ จึงจะเป็นประโยคที่ฟังรู้เรื่องและมีความหมาย เช่น ฉัน ดื่ม น้ำ เมื่อนำมารวมกันจะปรากฏรูปประโยคต่าง ๆ ได้แก่ “ฉันดื่มน้ำ” “น้ำดื่มฉัน” “ฉันน้ำดื่ม” “ดื่มฉันน้ำ” สังเกตว่า 2 ประโยคแรกมีการเรียบเรียงคำเป็นไปตามระเบียบของประโยคพื้นฐานของภาษาไทยแต่มีความหมายแตกต่างกัน กฎการเรียงคำในประโยคเป็นความสัมพันธ์ทางไวยากรณ์ของหมวดคำ เช่น ประธาน (ผู้กระทำ) กริยา และกรรม (ผู้ถูกกระทำ) ดังนั้น จากตัวอย่างข้างต้นแสดงให้เห็นว่า ระเบียบการเรียบเรียงประโยคพื้นฐานของภาษาไทย คือ ประธาน - กริยา - กรรม

2. ความสัมพันธ์ด้านการจับกลุ่ม หรือความสัมพันธ์เป็นลำดับชั้น เป็นการเรียงลำดับคำในประโยคที่มีความสัมพันธ์กัน นอกจากมีความสัมพันธ์กันยังสามารถจัดคำต่าง ๆ ในประโยคนั้นได้เป็นกลุ่มย่อย ๆ โดยที่คำในกลุ่มย่อยมีความสัมพันธ์บางอย่างต่อกัน เช่น “นักเรียนหญิงใส่กระโปรงสีน้ำเงิน” จากตัวอย่างสามารถแบ่งเป็น 2 กลุ่มใหญ่ คือ “นักเรียนหญิง” ซึ่งทำหน้าที่เป็นประธาน และ “ใส่กระโปรงสีน้ำเงิน” ทำหน้าที่เป็นภาคแสดง และสามารถแบ่งย่อยได้อีกเป็น “ใส่” และ “กระโปรงสีน้ำเงิน” กลุ่มของคำที่มีความสัมพันธ์ร่วมกันเป็นโครงสร้างของไวยากรณ์ที่ใหญ่ขึ้น (construction)

นอกจากนี้ “นักเรียนหญิง” ยังเป็นโครงสร้างระดับวลีที่ประกอบไปด้วยหน่วยประกอบย่อย 2 หน่วย คือ “นักเรียน” และ “หญิง” ส่วน “กระโปรงสีน้ำเงิน” เป็นโครงสร้างระดับวลีที่ประกอบด้วยหน่วยประกอบย่อย คือ “กระโปรง” และ “สีน้ำเงิน” สามารถเขียนแผนภูมิความสัมพันธ์ตามลำดับชั้นของหน่วยประกอบต่าง ๆ ได้ดังนี้



ภาพที่ 1 ความสัมพันธ์ตามลำดับชั้นของหน่วยประกอบของประโยค

การสื่อความหมายของประโยค จะต้องอาศัยการแสดงความสัมพันธ์ทางความหมายระหว่างประโยคต่าง ๆ เข้าด้วยกัน เรียกว่า “การเชื่อมโยง” โดยสามารถแสดงความสัมพันธ์ระหว่างประโยคออกเป็น 5 แบบ (วิจินตน์ ภาณุพงศ์ และคณะ, 2552) ดังนี้

1. การอ้างอิง (reference) คือ การแสดงความสัมพันธ์ที่เกิดจากรูปภาษาหนึ่งที่มีความหมายเป็นอิสระในตัวเอง แต่การตีความต้องอาศัยหรืออ้างถึงความหมายของรูปภาษาอื่น ซึ่งอาจอยู่ในรูปคำหรือวลีก็ได้

2. การแทนที่ (substitution) คือ การแสดงความสัมพันธ์ที่เกิดจากการแทนที่รูปภาษาหนึ่งด้วยภาษาอีกรูปหนึ่ง เช่น “ฉันชอบลูกแมว ที่เขาซื้อมาให้” เป็นความสัมพันธ์ระหว่างรูปภาษาซึ่งเป็นความสัมพันธ์ทางไวยากรณ์

3. การละ (ellipsis) คือ การแสดงความสัมพันธ์ที่เกิดจากการแทนที่รูปภาษาหนึ่งด้วยหน่วยทางภาษาที่ไม่ปรากฏรูป ซึ่งตีความหมายได้จากข้อความที่นำหน้า ข้อความที่อาจจะเป็นคำ วลี ประโยค หรืออนุพากย์ก็ได้

4. การใช้คำเชื่อม (conjunction) คือ การแสดงความสัมพันธ์ทางความหมายระหว่างประโยคหรือข้อความที่ปรากฏต่อเนื่องกัน เช่น คำว่า “อย่างไรก็ตาม” “เพราะ” “เพราะฉะนั้น” “เนื่องด้วย” “หลังจากนั้น” “ดังตัวอย่างต่อไปนี้” “ยกตัวอย่างเช่น” “อาทิ” “โดยทั่วไป” “นอกจากนี้” เป็นต้น

5. การเชื่อมโยงคำ (lexical cohesion) คือ การแสดงความสัมพันธ์โดยการใช้คำศัพท์ที่เกี่ยวข้องกัน มี 2 ประเภท คือ 1) การซ้ำ เป็นการใช้คำศัพท์ซ้ำรูปเดิมทั้งหมดหรือซ้ำเฉพาะส่วน และ 2) การใช้คำเข้าชุดกัน เป็นการใช้คำที่มีความหมายเกี่ยวพันกันทางความหมาย เช่น “ตำรวจจับผู้ร้ายที่ฆ่าเจ้าของร้านทองได้แล้วเมื่อวานนี้” ประโยคนี้จัดเป็นกลุ่มคำที่มีความหมายตรงกันข้าม “ฉันขายผลไม้หลายอย่าง ไม่ว่าจะเป็น กล้วย ส้ม ทุเรียน” ประโยคนี้จัดเป็นคำเข้าชุดอยู่ในกลุ่มเดียวกัน

ภาษาไทยเป็นเครื่องมือที่ใช้ในการติดต่อสื่อสารระหว่างกัน โดยโครงสร้างของประโยคที่ใช้จะต้องเป็นไปตามรูปแบบไวยากรณ์ของภาษาที่ประกอบด้วย ประธาน - กริยา - กรรม และมีการสื่อความหมายที่ตรงตามวัตถุประสงค์ ซึ่งบางครั้งรูปแบบประโยคมีความถูกต้องตามรูปแบบไวยากรณ์ของภาษา แต่มีการสื่อความหมายกำกวมหรือเป็นไปไม่ได้ ดังนั้น จะต้องอาศัยการแสดงความสัมพันธ์ทางความหมายระหว่างประโยค เพื่อให้รู้ความหมายที่แท้จริง

การประมวลผลภาษาธรรมชาติสำหรับจำแนกข้อความภาษาไทย

การประมวลผลภาษาธรรมชาติ (NLP) เป็นกระบวนการที่ทำให้คอมพิวเตอร์เข้าใจภาษาธรรมชาติของมนุษย์ที่ใช้สื่อสารกัน ให้อยู่ในรูปแบบโครงสร้างที่คอมพิวเตอร์สามารถเข้าใจและนำไปประมวลผลได้ วิธีการประมวลภาษาธรรมชาติได้ผ่านการสังเคราะห์ โดยแบ่งออกเป็น 5 ขั้นตอน ดังนี้ (บุญเสริม กิจศิริกุล, 2548; กรมวุฒิ นงนุช, อนุชา ซาเฮาะ และสุวุฒิ ตุ้มทอง, 2559; มาสวีร์ มาศดิศรโชติ, 2557; Yong-Yi & Yang, 2014; Larsson et al., 2017)

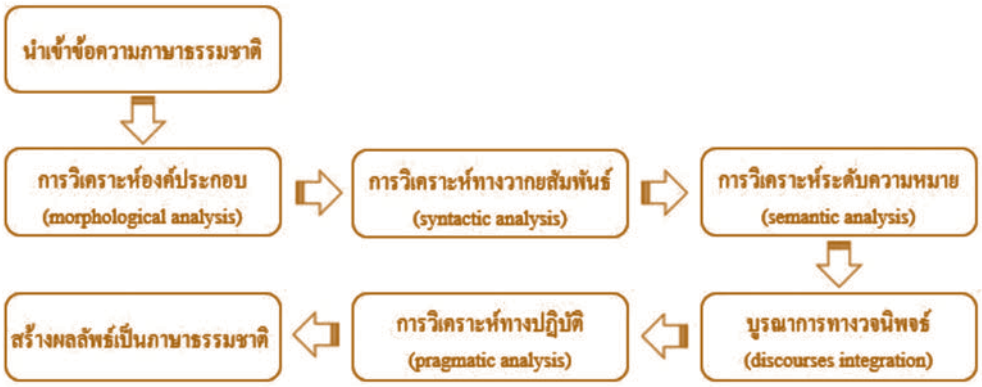
1. การวิเคราะห์ทางองค์ประกอบ (morphological analysis) เป็นการวิเคราะห์ในระดับของคำที่จะแยกย่อยเป็นอะไรได้บ้าง เช่น “งานประจำ” แยกได้ “งาน” และ “ประจำ”

2. การวิเคราะห์ทางวากยสัมพันธ์ (syntactic analysis) เป็นการวิเคราะห์ทางไวยากรณ์เพื่อดูโครงสร้างของประโยค และความเกี่ยวข้องของส่วนต่าง ๆ ในประโยคที่รับเข้ามาว่าคำไหนเป็นประธาน กริยา และกรรม หรือส่วนใดเป็นวลี เพื่อใช้แสดงความสัมพันธ์ของคำต่าง ๆ

3. การวิเคราะห์ระดับความหมาย (semantic analysis) เป็นการวิเคราะห์ความหมายของคำ หลังจากผ่านการวิเคราะห์ทางวากยสัมพันธ์มาแล้วจึงมากำหนดค่าของแต่ละคำว่าหมายถึงสิ่งใด ซึ่งบางคำเขียนถูกต้องตามหลักไวยากรณ์ แต่บางครั้งความหมายที่ได้เป็นความหมายที่กำกวม หรืออาจไร้ความหมาย หรือเป็นไปไม่ได้ เช่น “ฉันกินรถยนต์” ซึ่งถูกตามหลักไวยากรณ์ เพราะมีโครงสร้างประโยค ประธาน - กริยา - กรรม แต่มีความหมายเป็นไปไม่ได้ เป็นต้น

4. บูรณาการทางวจนิพนธ์ (discourse integration) เป็นการวิเคราะห์ความหมายของประโยคโดยพิจารณาจากประโยคข้างเคียง เนื่องจากคำบางคำในประโยคจะเข้าใจความหมายได้ต้องดูประโยคก่อนหน้าหรือประโยคหลังประกอบด้วย

5. การวิเคราะห์ในทางปฏิบัติ (pragmatic analysis) คือ การแปลความหมายหรือการตีความของประโยคใหม่อีกครั้งว่าผู้พูดตั้งใจจะสื่อความหมายอะไร หลังจากผ่านกระบวนการทั้ง 4 ขั้นตอนที่กล่าวมา



ภาพที่ 2 การประมวลผลภาษาธรรมชาติ

การประมวลผลภาษาธรรมชาติ (NLP) เป็นการนำความรู้ทางด้านภาษาศาสตร์มาวิเคราะห์รูปแบบโครงสร้างของประโยคตามหลักไวยากรณ์ และแปลความหมายของคำเพื่อทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ แล้วนำข้อความนั้นเก็บไว้ในฐานความรู้เพื่อให้คอมพิวเตอร์เกิดการเรียนรู้และสามารถนำไปสร้างเป็นแบบจำลอง (model) เพื่อนำไปใช้ประโยชน์ตามความต้องการ

การเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง (ML) เป็นศาสตร์แขนงหนึ่งของปัญญาประดิษฐ์ (artificial intelligence: AI) เป็นกระบวนการที่ทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้ด้วยตนเอง หลักการทำงานเกี่ยวข้องกับการศึกษาและการพัฒนาอัลกอริทึมที่สามารถเรียนรู้และปรับตัวตามข้อมูลที่ได้รับโดยอาศัยแบบจำลอง (model) ที่สร้างจากชุดข้อมูลที่มีความสัมพันธ์กัน และสามารถนำไปทำนายหรือใช้ในการตัดสินใจภายหลัง ประเภทการเรียนรู้ของเครื่องได้ผ่านการสังเคราะห์รูปแบบการเรียนรู้ออกเป็น 3 รูปแบบ ดังนี้ (บุญเสริม กิจศิริกุล, 2548; Mitchell, 1997; Smola & Vishwanathan, 2008; Alpaydin, 2014; Armstrong, 2015; Cates, Lawrence, Penedo & Samatova, 2017)

1. การเรียนรู้แบบมีผู้สอน (supervised learning) เป็นการเรียนรู้โดยอาศัยข้อมูลที่ป้อนเข้าไปเก็บไว้เป็นตัวอย่าง เพื่อให้คอมพิวเตอร์ใช้ในการเปรียบเทียบกับข้อมูลที่เข้ามาใหม่แล้วทำนายหรือจัดหมวดหมู่ที่มีความเหมือนกันมากที่สุดให้อยู่ด้วยกัน ผลลัพธ์ที่ได้คือ การจัดหมวดหมู่ (classification) และการวิเคราะห์การถดถอย (regression)

2. การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) เป็นการเรียนรู้ที่ไม่มีข้อมูลตัวอย่างที่ใช้บอกว่าข้อมูลนั้นคืออะไร แต่จะเรียนรู้จากการหาความสัมพันธ์จากข้อมูลนำเข้า (input) โดยพิจารณาจากรูปแบบ (patterns) หรือโครงสร้างของข้อมูล (data structure) แล้วนำมาจัดเป็นกลุ่ม (cluster) บนพื้นฐานของความเหมือน (similarities) และความแตกต่าง (differences) ระหว่างรูปแบบของข้อมูลนำเข้า (input patterns) ตัวอย่าง การหาโครงสร้างของข้อมูลที่ซ่อนอยู่

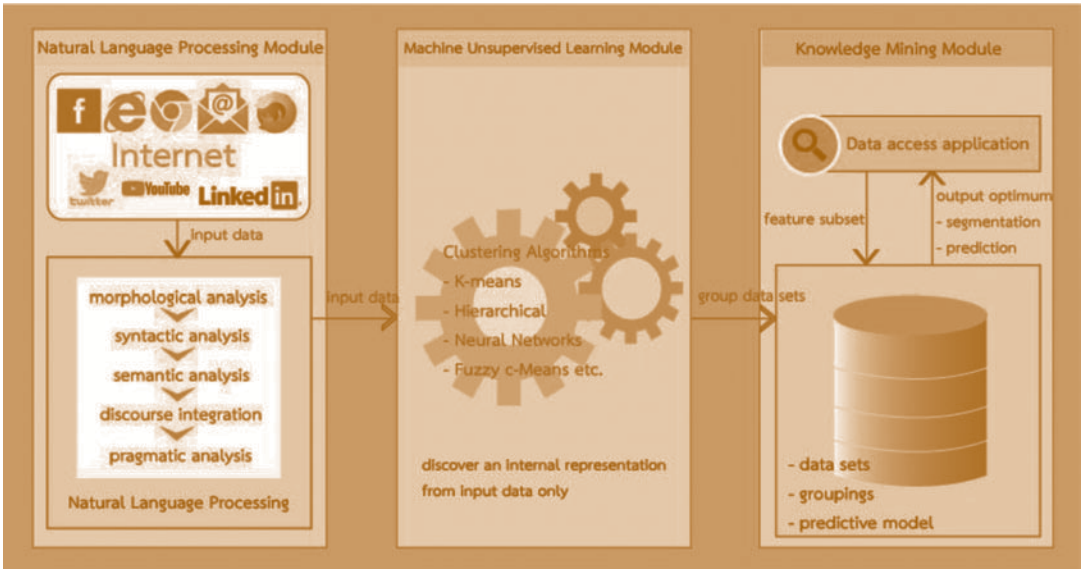
ประกอบด้วย การลดมิติข้อมูล (dimensionality reduction) การจัดกลุ่ม (clustering) และการเรียนรู้แบบซับซ้อน (manifold learning)

3. การเรียนรู้แบบเสริมกำลัง (reinforcement learning) เป็นการเรียนรู้แบบไม่มีผู้สอนแต่จะมีปฏิสัมพันธ์กับสิ่งแวดล้อม ทำให้เกิดการกระตุ้นในการสร้างแบบจำลองเพื่อตอบสนองว่าจะต้องทำอย่างไรต่อไป การกระทำจะเปลี่ยนไปตามสภาพแวดล้อมในขณะนั้น ตัวอย่างการนำไปประยุกต์ใช้งาน เช่น ควบคุมหุ่นยนต์ การเล่นเกม และการแนะนำเส้นทางการขับรถยนต์ เป็นต้น

การเรียนรู้ของเครื่อง (ML) เป็นกระบวนการทำงานที่ต้องการให้คอมพิวเตอร์มีความสามารถเรียนรู้สิ่งต่าง ๆ ด้วยตัวเอง หรือมีการทำงานเป็นไปอย่างอัตโนมัติ โดยมีรูปแบบการเรียนรู้แตกต่างกันไป การเลือกประเภทการเรียนรู้ของเครื่องจะขึ้นอยู่กับวัตถุประสงค์การนำไปใช้งาน

รูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ของเครื่อง (ML) โดยใช้เทคนิค Unsupervised Learning ร่วมกับการประมวลผลภาษาธรรมชาติ (NLP)

จากการศึกษาทฤษฎีและวรรณกรรมที่เกี่ยวข้องกับการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ ทำให้สามารถกำหนดกรอบแนวคิดในการสร้างแบบจำลองการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ของเครื่อง (ML) ด้วยเทคนิค Unsupervised learning ร่วมกับการประมวลผลภาษาธรรมชาติ (NLP) ซึ่งสามารถแบ่งการทำงานออกเป็น 3 มอดูล (modules) ดังภาพที่ 3



ภาพที่ 3 แบบจำลองการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ของเครื่อง (ML) ด้วยเทคนิค Unsupervised learning ร่วมกับการประมวลผลภาษาธรรมชาติ (NLP)

จากภาพที่ 3 สามารถอธิบายกระบวนการทำงานแบบจำลองการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ของเครื่อง (ML) ด้วยเทคนิค Unsupervised learning ร่วมกับการประมวลผลภาษาธรรมชาติ (NLP) ดังนี้

1. มอดูลการประมวลผลภาษาธรรมชาติ (natural language processing module: NLPM) เป็นการวิเคราะห์โครงสร้างข้อความภาษาไทยให้อยู่ในรูปแบบที่เครื่องคอมพิวเตอร์เข้าใจและสามารถนำไปประมวลผลได้ เป็นกระบวนการทำงานเมื่อรับข้อมูลนำเข้า (input data) คือ ข้อความภาษาไทยจากแหล่งข้อมูลต่าง ๆ ผ่านอินเทอร์เน็ต เช่น ข้อความจากเว็บไซต์ สื่อสังคมออนไลน์ เว็บบอร์ด จดหมายอิเล็กทรอนิกส์ และข้อความในเอกสารต่าง ๆ เป็นต้น โดยการวิเคราะห์โครงสร้างข้อความภาษาไทยด้วยวิธีการประมวลผลภาษาธรรมชาติ (NLP) มีขั้นตอนการวิเคราะห์แบ่งออกเป็น 5 ขั้นตอน ดังนี้

1.1 การวิเคราะห์ทางองค์ประกอบ (morphological analysis) เป็นขั้นตอนการนำข้อความภาษาไทยมาตัดเป็นคำ (word segmentation) เพื่อให้สามารถวิเคราะห์ในระดับของคำได้

1.2 การวิเคราะห์ทางวากยสัมพันธ์ (syntactic analysis) เป็นการวิเคราะห์ไวยากรณ์ของภาษาไทยเพื่อดูโครงสร้างของประโยค ได้แก่ ประธาน - กริยา - กรรม เพื่อใช้แสดงความสัมพันธ์ของประโยค

1.3 การวิเคราะห์ระดับความหมาย (semantic analysis) เป็นการวิเคราะห์ความหมายของคำในประโยคเพื่อสื่อความหมายที่ถูกต้อง เพราะบางประโยคมีโครงสร้าง ประธาน - กริยา - กรรม ถูกต้อง แต่สื่อความหมายที่เป็นไปไม่ได้ จึงมากำหนดค่าความหมายให้แต่ละคำว่าสื่อความหมายถึงอะไร

1.4 บูรณาการทางวจนิพนธ์ (discourse integration) หลังจากได้ความหมายของคำที่ชัดเจนในขั้นตอนที่ 3 แล้ว ในขั้นตอนนี้เป็นการวิเคราะห์ความหมายของประโยค โดยพิจารณาความสัมพันธ์ของประโยคข้างเคียงเพื่อให้ได้ความหมายที่แท้จริง

1.5 การวิเคราะห์ในทางปฏิบัติ (pragmatic analysis) เป็นขั้นตอนสุดท้ายคือ การแปลความหมายหรือตีความของประโยคใหม่อีกครั้ง เพื่อให้ได้ความหมายที่แท้จริง และได้ข้อความภาษาไทยที่มีรูปแบบโครงสร้างทางภาษาที่เครื่องคอมพิวเตอร์สามารถเข้าใจความหมายได้ถูกต้อง

2. มอดูลการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (machine unsupervised learning module: MULM) เป็นการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ เป็นขั้นตอนการวิเคราะห์ข้อมูลทางสถิติโดยใช้การเรียนรู้ของเครื่อง (ML) ด้วยเทคนิคการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) หลักการทำงานจะใช้อัลกอริทึมการจัดกลุ่มข้อมูล (clustering algorithms) ซึ่งรองรับการทำงานของอัลกอริทึม เช่น K-means, Hierarchical, Neural Networks, Fuzzy C-means เป็นต้น โดยจะเลือกใช้อัลกอริทึมที่มีผลลัพธ์ในการประมวลผลที่ดีที่สุด เมื่อได้รับข้อมูลนำเข้า (input data) ผ่านการประมวลผลข้อมูลจากมอดูลที่ 1 จะจำแนกกลุ่มข้อความภาษาไทยที่สื่อความหมายเหมือนกัน หรือสื่อความหมายในประเด็นเดียวกันอยู่ในกลุ่มเดียวกัน ซึ่งการแบ่งกลุ่มข้อมูลจะไม่มี

การกำหนดประเภทข้อมูลไว้ก่อน หรือไม่ทราบจำนวนกลุ่มล่วงหน้า โดยปริมาณข้อมูลและจำนวนกลุ่มข้อมูลจะมีปริมาณมากหรือน้อยขึ้นอยู่กับจำนวนข้อมูลนำเข้า เพราะข้อมูลที่เข้ามาใหม่ถ้าสื่อความหมายเหมือนกับกลุ่มข้อมูลที่มีอยู่แล้วก็จะอยู่ในกลุ่มนั้น ๆ แต่ถ้าสื่อความหมายที่แตกต่างกันจะแยกกลุ่มต่อไป ซึ่งลักษณะการทำงานจะเกิดการขยายกลุ่มหรือขยายขนาดของข้อมูลต่อไปเรื่อย ๆ ขึ้นอยู่กับปริมาณข้อมูลที่นำเข้ามา ซึ่งทำให้เกิดโมเดล (model) การเรียนรู้ของเครื่อง (ML) สำหรับการจำแนกกลุ่มข้อมูลภาษาไทยแบบอัตโนมัติ

3. มอดูลเหมืองความรู้ (knowledge mining module: KMM) คือแหล่งจัดเก็บข้อมูลที่ผ่านการประมวลผลการเรียนรู้ของเครื่องแบบไม่มีผู้สอนในมอดูลที่ 2 โดยข้อมูลที่จัดเก็บจะผ่านการจำแนกกลุ่มข้อมูลออกเป็นกลุ่ม ๆ ที่สื่อความหมายเหมือนกันและมีความสัมพันธ์กันจะจัดอยู่ในกลุ่มเดียวกัน ข้อมูลที่นำมาวิเคราะห์ผ่านกระบวนการทำงานของระบบจะมาจากแหล่งข้อมูลที่หลากหลายที่เผยแพร่อยู่บนอินเทอร์เน็ต เช่น เว็บไซต์ สื่อสังคมออนไลน์ จดหมายอิเล็กทรอนิกส์ เว็บบอร์ด และเอกสารต่าง ๆ ดังนั้น ข้อมูลที่จัดเก็บไว้จัดว่าเป็นแหล่งขุมทรัพย์ที่มีประโยชน์ต่อการนำไปใช้งานทางด้านธุรกิจ เช่น การรับรู้พฤติกรรมการใช้สินค้าหรือบริการของผู้บริโภค การรับรู้ทัศนคติของลูกค้าต่อผลิตภัณฑ์ การรับรู้ข้อมูลเชิงลึกเพื่อนำไปกำหนดกลยุทธ์ทางด้านการตลาดหรือนำเสนอผลิตภัณฑ์ใหม่ เป็นต้น นอกเหนือจากความรู้ที่ได้รับทางด้านธุรกิจ ยังสามารถรับรู้ข้อมูลที่เป็นประโยชน์ในด้านอื่น ๆ เนื่องจากเป็นแหล่งข้อมูลที่ถูกรวบรวมมาจากทุกช่องทางผ่านเครือข่ายอินเทอร์เน็ต

สรุป

บทความนี้นำเสนอรูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้กระบวนการเรียนรู้ของเครื่อง (ML) และเลือกใช้เทคนิค Unsupervised Learning สำหรับการแบ่งกลุ่มข้อความภาษาไทยแบบอัตโนมัติ การจำแนกกลุ่มข้อความจะเรียนรู้จากข้อมูลที่นำเข้าสู่ระบบซึ่งไม่มีการกำหนดหมวดหมู่ไว้ล่วงหน้า และด้วยโครงสร้างภาษาไทยมีความซับซ้อน การแบ่งประโยคหรือคำในภาษาไทยยังมีรูปแบบที่ไม่แน่นอน ทำให้กระบวนการทำงานมีความยุ่งยากมากกว่าภาษาอังกฤษ จึงนำหลักการประมวลผลภาษาธรรมชาติ (NLP) เข้ามาช่วยในการวิเคราะห์โครงสร้างและการสื่อความหมายที่ถูกต้อง เพื่อให้เครื่องคอมพิวเตอร์สามารถเข้าใจและนำไปประมวลผลได้ รูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติที่นำเสนอสามารถแบ่งกระบวนการทำงานออกเป็น 3 มอดูลหลัก คือ 1) มอดูลการประมวลผลภาษาธรรมชาติ (NLP) เป็นการวิเคราะห์โครงสร้างข้อความภาษาไทยให้เครื่องคอมพิวเตอร์สามารถเข้าใจความหมายแล้วนำไปประมวล 2) มอดูลการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (MULM) เป็นโมเดลการเรียนรู้สำหรับการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ และ 3) มอดูลเหมืองความรู้ (KMM) เป็นส่วนที่ใช้จัดเก็บข้อมูลที่ผ่านการจำแนกกลุ่มข้อความแล้ว โดยข้อมูลที่จัดเก็บจะมีความสัมพันธ์เกี่ยวข้องกันทั้งทางด้านความหมายและโครงสร้างของข้อมูล

ปัจจุบันภาคธุรกิจให้ความสำคัญกับแหล่งข้อมูลขนาดใหญ่มหาศาลบนอินเทอร์เน็ต ซึ่งข้อมูลเหล่านี้เปรียบได้กับแหล่งขุมทรัพย์ที่มีค่าต่อการดำเนินธุรกิจ เพราะเป็นข้อมูลที่เป็นข้อเท็จจริงที่เกิดขึ้นในปัจจุบัน ดังนั้น รูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ของเครื่อง (ML) ด้วยเทคนิค Unsupervised Learning ร่วมกับการประมวลผลภาษาธรรมชาติ (NLP) ที่นำเสนอจัดว่าเป็นกระบวนการทำงานหนึ่งที่สามารถนำไปใช้ในการรวบรวมข้อมูลจากแหล่งข้อมูลขนาดใหญ่บนอินเทอร์เน็ตมาจัดเก็บไว้เป็นเหมืองความรู้ (knowledge mining) และสามารถนำความรู้นี้ไปใช้ให้เกิดประโยชน์ในด้านต่าง ๆ เช่น วิเคราะห์การรับรู้พฤติกรรมการใช้สินค้าหรือบริการของผู้บริโภค การรับรู้ทัศนคติของลูกค้าต่อผลิตภัณฑ์ หรือการรับรู้การแสดงความคิดเห็นของผู้บริโภคในด้านบวกหรือด้านลบต่อการใช้ผลิตภัณฑ์ และการรับรู้ข้อมูลเชิงลึกเพื่อนำไปกำหนดกลยุทธ์ทางด้านการตลาดหรือนำเสนอผลิตภัณฑ์ใหม่ เป็นต้น โดยองค์ความรู้ที่จัดเก็บและนำไปใช้ประโยชน์จะเป็นไปตามแหล่งข้อมูลที่นำเข้าสู่ระบบ นอกจากนี้ รูปแบบดังกล่าวยังสามารถนำไปประยุกต์ใช้กับการจำแนกกลุ่มข้อความภาษาไทยในด้านอื่น ๆ เช่น นำไปประยุกต์ใช้ในการคัดเลือกบทความวิชาการที่เป็นภาษาไทยให้ตรงกับความเชี่ยวชาญของคณะกรรมการพิจารณาผลงาน โดยจำแนกกลุ่มตามเนื้อหาบทความที่นำเข้าสู่ระบบผ่านกระบวนการ 3 ขั้นตอน คือ 1) วิเคราะห์โครงสร้างและความหมายด้วยการประมวลผลภาษาธรรมชาติ (NLPM) 2) ส่งข้อมูลสู่กระบวนการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (MULM) เพื่อจำแนกกลุ่มบทความแบบอัตโนมัติ และ 3) จัดเก็บข้อมูลในเหมืองความรู้ (KMM) เพื่อนำเสนอบทความให้คณะกรรมการที่มีความเชี่ยวชาญตรงกับเนื้อหาบทความ

บรรณานุกรม

กรมวุฒิ นงนุช, อนุชา ซาเฮาะ และสุวุฒิ ตุ้มทอง. (2559). การวิเคราะห์บทความอัตโนมัติโดยใช้กระบวนการภาษาธรรมชาติ. ใน นภัทร วัจนเทพินทร์ (บรรณาธิการ), ***การประชุมวิชาการระดับชาติ มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ ครั้งที่ 1*** (หน้า 472-479). พระนครศรีอยุธยา: สถาบันวิจัยและพัฒนา มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ.

กานดา แผ่วัฒนากุล และปราโมทย์ ลือนาม. (2556). การวิเคราะห์เหมืองความคิดเห็นบนเครือข่ายสังคมออนไลน์. ***วารสารการจัดการสมัยใหม่, 11***(2), หน้า 11-20.

นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ. (2554). การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ. ***NIDA Development Journal, 51***(3), หน้า 187-205.

บุญเสริม กิจศิริกุล. (2548). ***ปัญญาประดิษฐ์ เอกสารคำสอนวิชา 2110654.*** กรุงเทพฯ: ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย.

มาสวีร์ มาศดิศรโชติ. (2557). การทำเหมืองความคิดเห็นภาษาไทย. ***วารสารศรีปทุมปริทัศน์ ฉบับวิทยาศาสตร์และเทคโนโลยี, 6***(1), หน้า 120-128.

ราชบัณฑิตยสถาน. (2556). ***พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554 เฉลิมพระเกียรติพระบาทสมเด็จพระเจ้าอยู่หัว เนื่องในโอกาสพระราชพิธีมหามงคลเฉลิมพระชนมพรรษา 7 รอบ 5 ธันวาคม 2554.*** กรุงเทพฯ: ราชบัณฑิตยสถาน.

วิจินตน์ ภาณุพงศ์ และคณะ. (2552). ***บรรทัดฐานภาษาไทย เล่ม 3: ชนิดของคำ วลี ประโยค และสัมพันธสาร.*** กรุงเทพฯ: สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ.

ศิริพร อ่วมมีเพียร และสันติพงษ์ ไทยประยูร. (2559). ระบบติดตามการคัดลอกเนื้อหาเว็บอัตโนมัติโดยใช้วิธีการเลือกข้อความสำคัญ. ***วารสารเทคโนโลยีสารสนเทศ, 12***(2), หน้า 1-9.

โศรยา วิมลสถิตพงษ์. (2558). ***การศึกษาภาษาไทยตามแนวภาษาศาสตร์.*** อุดรธานี: คณะมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยราชภัฏอุดรธานี.

Alpaydin, E. (2014). ***Introduction to machine learning*** (3rd ed.). Massachusetts, MA: MIT Press.

Armstrong, H. (2015). ***Machines Thai learn in the wild.*** London, UK: Nesta.

Cates, S., Lawrence, S., Penedo, C., & Samatova, V. (2017). A machine learning approach to research curation for investment process. ***Journal of Investment Management, 15***(1), pp. 39-49.

Galitsky, B. (2013). Machine learning of syntactic parse trees for search and classification of text. ***Journal Engineering Applications of Artificial Intelligence, 26***(3), pp. 1027-1091.

Larsson, K., et al. (2017). Text mining for improved exposure assessment. ***Journal Public Library of Science, 12***(3), pp. 1-21.

Marquez, L. (2000). ***Machine learning and natural language processing.*** Barcelona, Spain: Universitat Politechnica de Catalunya.

Mitchell, T. M. (1997). ***Machine learning.*** New York, NY: McGraw-Hill.

Sebastiani, F. (2002). Machine learning in automated text categorization. ***Journal ACM Computing Surveys (CSUR), 34***(1), pp. 1-47.

Smola, A., & Vishwanathan, S. V. N. (2008). ***Introduction to machine learning.*** London, UK: Cambridge University Press.

Yong-Yi, Fanjiang, & Yang, Syu. (2014). Semantic-based automatic service composition with functional and non-functional requirements in design time: A genetic algorithm approach. ***Journal Information and Software Technology, 56***(3), pp. 352-373.