เครื่องมือสำหรับการวิเคราะห์โครงสร้างและองค์ประกอบสำหรับบทคัดย่อของเอกสารทางวิชาการ

โดย

นายรัฐพล ชูพรม

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ แผน ข ระดับปริญญามหาบัณฑิต

ภาควิชาคอมพิวเตอร์

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2561

เครื่องมือสำหรับการวิเคราะห์โครงสร้างและองค์ประกอบสำหรับบทคัดย่อของเอกสารทางวิชาการ

โดย
นายรัฐพล ชูพรม

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ แผน ข ระดับปริญญามหาบัณฑิต
ภาควิชาคอมพิวเตอร์
บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร
ปีการศึกษา 2561

WARANUSMOVE : A MACHINE LEARNING TOOL FOR ANALYZING STRUCTURE OF ABSTRACT IN RESEARCH ARTICLE

By

MR. Ratthaphon CHOOPROM

A Independent Study Submitted in Partial Fulfillment of the Requirements

for Master of Science (INFORMATION TECHNOLOGY)

Department of COMPUTER SCIENCE

Graduate School, Silpakorn University

Academic Year 2018

หัวข้อ	เครื่องมือสำหรับการวิเคราะห์โครงสร้างและองค์ประกอบสำหรับบทคัดย่อของเอกสารทางวิชาการ
โดย	รัฐพล ชูพรม
สาขาวิชา	เทคโนโลยีสารสนเทศ แผน ข ระดับปริญญามหาบัณฑิต
อาจารย์ที่ปรึกษาหลัก	ผู้ช่วยศาสตราจารย์ ดร. ทัศนวรรณ ศูนย์กลาง

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

..... คณบดีบัณฑิตวิทยาลัย
(รองศาสตราจารย์ ดร.จุไรรัตน์ นันทานิช)

พิจารณาเห็นชอบโดย

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. วีณาวดี ม่วงอ้น)

..... อาจารย์ที่ปรึกษาหลัก
(ผู้ช่วยศาสตราจารย์ ดร. ทัศนวรรณ ศูนย์กลาง)

..... ผู้ทรงคุณวุฒิภายนอก
(รองศาสตราจารย์ ดร. จันทนา จันทราพรชัย)

57309302 : เทคโนโลยีสารสนเทศ แผน ข ระดับปริญญามหาบัณฑิต
คำสำคัญ : เหมืองข้อความ

นาย รัฐพล ชูพรม: เครื่องมือสำหรับการวิเคราะห์โครงสร้างและองค์ประกอบสำหรับบทคัดย่อของเอกสารทางวิชาการ อาจารย์ที่ปรึกษาวิทยานิพนธ์ : ผู้ช่วยศาสตราจารย์ ดร. ทัศนวรรณ ศูนย์กลาง

งานวิจัยนี้ได้พัฒนาเครื่องมือสำหรับการวิเคราะห์โครงสร้างและองค์ประกอบสำหรับบทคัดย่อของเอกสารทางวิชาการ โดยบทคัดย่อเป็นส่วนสรุปภาพรวมและรายละเอียดพื้นฐานของงานวิจัยและวิทยานิพนธ์ โดยปกติแล้วบทคัดย่อใช้เป็นส่วนช่วยให้ผู้อ่านสามารถค้นหาวัตถุประสงค์ของงานวิจัยได้อย่างรวดเร็ว งานวิจัยนี้มีวัตถุประสงค์ที่จะนำการเรียนรู้ของเครื่องมาใช้วิเคราะห์โครงสร้างของบทคัดย่อเพื่อพัฒนาเครื่องมือที่มีชื่อว่า WaranusMove ซึ่งเครื่องมือนี้สามารถวิเคราะห์โครงสร้างของบทคัดย่อที่มีอยู่ 5 มูฟประกอบด้วย background, purpose, method result และ discussion ซึ่งแสดงผลลัพธ์อยู่ในรูปแบบของแถบสีครอบคลุมแต่ละประโยคเพื่อบ่งบอกมูฟ นอกจากนั้นได้นำเอาโมเดลการเรียนรู้ของเครื่อง Support Vector Machine (SVM) และ Decision tree มาเปรียบเทียบประสิทธิภาพการจำแนกโครงสร้างของบทคัดย่อ

งานวิจัยนี้พัฒนาขึ้นในรูปแบบโปรแกรมประยุกต์บนเว็บไซต์ที่ใช้ภาษาโปรแกรม Python และ PHP ในการพัฒนาประกอบด้วย ส่วนการฝึกฝน ในส่วนนี้ระบบสามารถเรียนรู้บทคัดย่อได้ในหลากหลายสาขาวิชาและใช้โมเดลการเรียนรู้ของเครื่องได้ดังนี้ Decision tree, Naïve bays, SVM และ Random forest และสามารถเลือกกลุ่มของคุณลักษณะที่ใช้ในการวิเคราะห์ทั้ง 3 รูปแบบ ได้แก่ Lexical features, Grammatical and position features และ Lexical, Grammatical and position features ต่อมาคือส่วนการวิเคราะห์ ในส่วนนี้ผู้ใช้สามารถนำโมเดลที่เรียนรู้บทคัดย่อในสาขาวิชาต่างๆ จากส่วนการฝึกฝนมาใช้จำแนกโครงสร้างของประโยคในบทคัดย่อได้ นอกจากนี้ได้มีการประเมินประสิทธิภาพของการจำแนกประเภทโครงสร้างของบทคัดย่อด้วยวิธีการ 10-fold cross validation จากผลลัพธ์การวิจัยพบว่า Decision tree ให้ประสิทธิภาพการวิเคราะห์ที่ดีกว่า SVM สุดท้ายนี้เครื่องมือสำหรับการวิเคราะห์โครงสร้างของบทคัดย่อหรือ WaranusMove นั้นมีประโยชน์ต่อผู้เริ่มต้นการฝึกฝนเขียนบทคัดย่อในภาษาอังกฤษ และผู้วิจัยที่ต้องการสร้างโมเดลการจำแนกโครงสร้างของบทคัดย่ออีกด้วย

57309302 : Major (INFORMATION TECHNOLOGY)
Keyword : Text minig

MR. RATTHAPHON CHOOPROM : WARANUSMOVE : A MACHINE LEARNING TOOL FOR ANALYZING STRUCTURE OF ABSTRACT IN RESEARCH ARTICLE THESIS ADVISOR : ASSISTANT PROFESSOR DOCTOR TASANAWAN SOONKLANG

WaranusMove: A machine learning tool for analyzing structure of abstract in research article Abstract is a brief summary of a research article and thesis. It is usually used to help readers quickly discover the paper's purpose. This study aimed to develop a machine learning tool for analyzing abstract named WasarnusMove. The abstract analysis tool is able to analyze structure of abstract into five move, including background, purpose, method, result and discussion. Each move will be showed in separate blocks of colors. Moreover, Support Vector Machine (SVM) and decision tree were applied as classifiers to compare the performance.

The system was developed as a web-based application by using the Python and PHP language. Our tool comprises of training mode and analyzing mode. In training mode, abstracts in any field can be trained with various classifiers such as decision tree, naïve bayes, SVM, and random forest. The three group of features can be selected, which are lexical features, grammatical and position features, and lexical, grammatical and position features. In analyzing mode, many models from training mode can be deployed for user to categorize sentences in any abstract into each move.

We also evaluated the performance of two classifiers with 10-fold cross validation. The results suggested that decision tree performed better than SVM. Finally, WaranusMove tool is useful for beginner to practice writing abstract in English and for researcher to create the model for classifying abstracts.

กิตติกรรมประกาศ

การค้นคว้าอิสระฉบับนี้สำเร็จลุล่วงตามความหวังของผู้เขียนได้ เนื่องมาจากความช่วยเหลือและเมตตากรุณาของหลายท่านที่มีพระคุณยิ่ง ซึ่งให้การสนับสนุนผู้วิจัยตั้งแต่เริ่มต้นจนการค้นคว้าอิสระนี้เสร็จสมบูรณ์

ขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.ทัศนวรรณ ศูนย์กลาง อาจารย์ที่ปรึกษาการค้นคว้าอิสระ ที่กรุณารับเป็นที่ปรึกษาและเสียสละเวลาอันมีค่าในการให้คำปรึกษา ให้การติดตามถามถึงปัญหาตลอดระยะเวลาในการศึกษาอย่างต่อเนื่อง ซึ่งท่านได้ให้คำแนะนำ ข้อคิดเห็นต่างๆ อันเป็นประโยชน์อย่างยิ่งในการทำการค้นคว้าอิสระ รวมทั้งยังช่วยแก้ปัญหาแนะแนวทางของปัญหาที่เกิดขึ้นระหว่างการดำเนินงานอีกด้วย แม้ครั้งที่ตัวท่านเองมีปัญหาสุขภาพแต่ด้วยความมีจิตวิญญาณของความเป็นครู ท่านจึงไม่เคยที่จะทอดทิ้งผู้วิจัยจนประสบความสำเร็จ ตลอดจนผู้ช่วยศาสตราจารย์ ดร.สุนีย์ พงษ์พินิจภิญโญ ที่คอยสอบถามเสมอทั้งเรื่องงานวิจัย และเรื่องการงานทุกครั้งที่ได้พบกัน พร้อมทั้งให้ความช่วยเหลือสอบถามข้อสงสัยต่างๆ ในหลักสูตร รวมทั้งผู้ช่วยศาสตราจารย์ ดร.วีณาวดี ม่วงอ้น และ รศ. ดร. จันทนา จันทราพรชัย ที่เสียสละเวลาอันมีค่าเพื่อเป็นประธานและคณะกรรมการสอบ ทั้งนี้ขอขอบพระคุณศาสตราจารย์ ดร.บุษบา กนกศิลปธรรม และผศ.ดร.อรประภา ภุมมะกาญจนะ โรแบร์ ทุกครั้งที่พบกันก็คอยสอบถามถึงความคืบหน้าของการศึกษา อีกทั้งคุณประวิม เหลืองสมานกุล ที่คอยให้ความช่วยเหลือเรื่องของขั้นตอนวิธีการ ของเอกสารสอบต่างๆ เพื่อให้การสอบประสบความสำเร็จลุล่วง และคอยเตือนให้ลงทะเบียนเรียนในรายวิชาขณะที่ผู้วิจัยกำลังศึกษาเสมอมา รวมทั้งพี่เจ้าหน้าที่คณะวิทยาศาสตร์ที่ให้ความช่วยเหลือด้านเอกสาร

ท้ายนี้ขอน้อมรำลึกถึงอำนาจของคุณพระศรีรัตนตรัย บารมีแห่งองค์พระพิฆเณศวรเทพแห่งความสำเร็จ อันเป็นที่ยึดเหนี่ยวจิตใจให้ผู้วิจัยมีสติปัญญา ชี้ทางสว่างจนทำให้งานวิจัยสำเร็จลุล่วงไปได้ด้วยดี และขอมอบความกตัญญูกตเวทิตาคุณ แด่บิดา มารดา ญาติพี่น้อง และผู้มีพระคุณทุกท่านที่ให้การสนับสนุนจนทำให้งานวิจัยประสบความสำเร็จ สำหรับข้อบกพร่องต่างๆ ที่เกิดขึ้น ผู้วิจัยขอน้อมรับไว้แต่เพียงผู้เดียว

รัฐพล ชูพรม

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

ความเป็นมาของปัญหา

ปัจจุบันผลงานวิจัยจำนวนมากถูกตีพิมพ์สู่สาธารณะผ่านงานประชุมทางวิชาการที่จัดขึ้นในทั่วทุกมุมโลก งานวิจัยถูกเขียนขึ้นจากผู้เขียนที่มีความแตกต่างกันทั้งด้านวิธีการใช้ภาษา วิธีการเขียน หรือกระทั่งในบทความทางวิชาการประเภทเดียวกันอาจมีโครงสร้างที่แตกต่างออกไป การเขียนบทความทางวิชาการสำหรับนักวิจัย และผู้ที่กำลังศึกษาในระดับบัณฑิตศึกษานั้น จึงถือเป็นสิ่งสำคัญและต้องปฏิบัติ บทความทางวิชาการควรมีลักษณะบทความที่ดี เพื่อผู้อ่านที่เป็นผู้สนใจ หรือผู้ตรวจสอบบทความของผู้วิจัยเพื่อการตีพิมพ์ผลงานในงานประชุมทางวิชาการ สามารถอ่านบทความได้อย่างเข้าใจ และทราบถึงรายละเอียดของงานวิจัยได้

บทคัดย่อ (Abstract) เป็นเนื้อหาส่วนแรกของบทความทางวิชาการที่ผู้อ่านจะศึกษา โดยบทคัดย่อเป็นส่วนที่สรุปภาพรวมพร้อมทั้งรายละเอียดพื้นฐานของงานวิจัย วัตถุประสงค์ ขั้นตอนการดำเนินงาน ผลการทดลอง และผลสรุปของงานวิจัย อย่างสังเขป ดังนั้นการเขียนบทคัดย่อในบทความทางวิชาการจึงมีความสำคัญ และยังเป็นจุดเริ่มต้นที่ผู้อ่านบทความทางวิชาการสามารถทราบได้ว่ากำลังศึกษาในหัวข้อที่ตนสนใจอยู่หรือไม่ ปัจจุบันมีงานวิจัยจำนวนมากทางด้านภาษาศาสตร์ที่ศึกษาเกี่ยวกับการวิเคราะห์รูปแบบการเขียนตามแนวคิดการวิเคราะห์รูปแบบสัมพันธสาร (Discourse analysis) คือ อัตถภาควิเคราะห์ (Move analysis) ในงานเขียนประเภทต่างๆ อย่างกว้างขวาง ซึ่งการวิเคราะห์นี้เพื่อกำหนดว่าบทความที่วิเคราะห์นั้นประกอบด้วยโครงสร้างใดบ้าง ดังนั้นการวิเคราะห์และจำแนกโครงสร้างของบทคัดย่อจึงถูกนำมาวิเคราะห์ด้วยเช่นกัน

โดยทั่วไปลักษณะโครงสร้างของบทคัดย่อในงานแต่ละด้านที่เขียนขึ้นนั้น มีความคล้ายคลึงกัน อันประกอบไปด้วยส่วนต่างๆ หรือที่เรียกว่า มูฟ (Move) ดังนี้ พื้นฐานของงานวิจัย (Background) วัตถุประสงค์ (Purpose) กระบวนการทดลอง (Methodology) ผลการทดลอง (Result) สรุปผลการทดลอง (Conclusion) และอภิปรายผลการทดลอง (Discussion) ซึ่งในบางบทคัดย่อผู้เขียนอาจขึ้นเขียนโดยไม่มีองค์ประกอบดังกล่าวครบทุกส่วน แต่สำหรับผู้เริ่มต้นควรเขียนบทคัดย่อให้มีโครงสร้างครบถ้วน

ปัจจุบันมีผู้จัดทำงานวิจัยทางด้านการวิเคราะห์โครงสร้างของบทคัดย่อ ทั้งในรูปแบบโปรแกรมประยุกต์บนเว็บไซต์ ที่ผู้ใช้สามารถนำบทคัดย่อให้โปรแกรมช่วยในการตรวจสอบองค์ประกอบของบทคัดย่อ แสดงมูฟที่มีในบทคัดย่อพร้อมทั้งแสดงมูฟที่ขาดหายไปในบทความอีกด้วย ซึ่งสร้างความสะดวกสบายแก่สำหรับผู้ใช้ที่เริ่มต้นการเขียนบทคัดย่อ

ผู้วิจัยจึงเกิดแนวคิดพัฒนางานวิจัยในรูปแบบโปรแกรมประยุกต์บนเว็บไซต์ที่สามารถวิเคราะห์รูปแบบโครงสร้างของบทคัดย่อ และวิเคราะห์องค์ประกอบของบทคัดย่อหรือมูฟของเอกสารทางวิชาการได้ในหลายหลายสาขาวิชา โดยนำวิธีการเรียนรู้ของเครื่อง (Machine Learning) มาใช้ในการวิเคราะห์ในลักษณะของการทำเหมืองข้อความ และแสดงผลลัพธ์จากการวิเคราะห์โครงสร้างของบทคัดย่อ อีกทั้งมีส่วนการฝึกสอนที่สามารถเพิ่มคลังของบทความในสาขาวิชาต่างๆ เพิ่มวิธีการเรียนรู้ของเครื่องที่ต้องการให้โปรแกรมประยุกต์บนเว็บไซต์ใช้ในการวิเคราะห์โครงสร้างของบทคัดย่อ เพื่อผู้ใช้สามารถเลือกใช้วิธีการวิเคราะห์ได้เหมาะสมกับบทความของตนเองอีกด้วย

วัตถุประสงค์ของการวิจัย

1. เพื่อพัฒนาเครื่องมือวิเคราะห์โครงสร้าง และองค์ประกอบของบทความทางวิชาการในส่วนบทคัดย่อหรือมูฟ จากงานวิจัยเครื่องมือสำหรับการวิเคราะห์โครงสร้างและมูฟสำหรับบทคัดย่อของเอกสารทางวิทยาศาสตร์ กรณีศึกษาวิศวกรรมชีวเวช
2. เพื่อเพิ่มความหลากหลายในการวิเคราะห์โครงสร้าง และองค์ประกอบของบทความทางวิชาการในส่วนบทคัดย่อหรือมูฟ ในสาขาวิชาต่างๆ และเพิ่มความหลากหลายของวิธีการเรียนรู้ของเครื่องที่ใช้ในการวิเคราะห์โครงสร้างของบทคัดย่อ
3. เพื่อช่วยให้ผู้อ่านสามารถเข้าใจประเด็นสำคัญของเนื้อหาในบทคัดย่อได้ดีขึ้น

ผลที่คาดว่าจะได้รับ

1. เครื่องมือที่สามารถวิเคราะห์โครงสร้างและองค์ประกอบของบทคัดย่อในบทความทางวิชาการสาขาต่างๆ
2. ให้ผู้อ่านบทคัดย่อสามารถเข้าใจในเนื้อหา และประเด็นสำคัญของบทความได้ดียิ่งขึ้น
3. ช่วยแนะนำสำหรับผู้เริ่มต้นเขียนบทคัดย่อให้ได้รูปแบบที่สมบูรณ์

ขอบเขตการดำเนินงาน

1. การค้นคว้าอิสระนี้ได้พัฒนาต่อเนื่องจาก เครื่องมือสำหรับการวิเคราะห์โครงสร้างและมูฟสำหรับบทคัดย่อของเอกสารทางวิทยาศาสตร์ กรณีศึกษาวิศวกรรมชีวเวช (Tools for structure and move analysis of abstract for scientific article : Biomedical Engineering)

2. ส่วนวิเคราะห์โครงสร้าง ผู้ใช้งานสามารถเลือกสาขาวิชาของบทคัดย่อที่ต้องการให้วิเคราะห์ และเลือกวิธีการเรียนรู้ของเครื่องที่ต้องการใช้ในการวิเคราะห์บทคัดย่อ อีกทั้งสามารถรับเข้าบทคัดย่อของบทความทางวิชาการในสาขาวิชาต่างๆ จากผู้ใช้เพื่อทำการวิเคราะห์โครงสร้างและองค์ประกอบของบทคัดย่อ

3. บทคัดย่อที่ผ่านการวิเคราะห์โครงสร้างและองค์ประกอบแล้วจะแสดงผลการวิเคราะห์ออกเป็นมูฟต่าง ๆ ดังนี้

1) Background (B) เป็นส่วนที่บ่งบอกถึงความเป็นมาพื้นฐานของงานวิจัย

2) Purpose (P) เป็นส่วนที่บ่งบอกถึงวัตถุประสงค์ หรือเป้าหมายของงานวิจัย

3) Methodology (M) เป็นส่วนที่บ่งบอกถึงกระบวนการทำงานของงานวิจัย

4) Result (R) เป็นส่วนที่บ่งบอกผลลัพธ์ของงานวิจัย

5) Discussion (D) เป็นส่วนที่บอกถึงข้อสรุปของผลการวิจัย และการนำไปประยุกต์ใช้

4. ส่วนการฝึกสอน ผู้ใช้สามารถเพิ่มบทความทางวิชาการในสาขาต่างๆ เพื่อเป็นการเพิ่มสาขาวิชาที่ให้ผู้ใช้นำบทคัดย่อในสาขานั้นๆ มาวิเคราะห์ และสามารถเพิ่มวิธีการเรียนรู้ของเครื่องที่ใช้สำหรับการวิเคราะห์โครงงสร้างของบทคัดย่อ พร้อมทั้งแสดงสรุปรูปแบบโครงสร้างของบทคัดย่อในบทความทางวิชาการในแต่ละสาขาวิชาที่เพิ่มเข้าสู่ระบบอีกด้วย

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์เพื่อการพัฒนาเครื่องมือสำหรับการวิเคราะห์โครงสร้างและองค์ประกอบสำหรับบทคัดย่อของเอกสารทางวิชาการในรูปแบบโปรแกรมประยุกต์บนเว็บไซต์ เพื่อเป็นการศึกษากระบวนการวิเคราะห์โครงสร้างและองค์ประกอบของบทคัดย่ออีกด้วย จึงรวบรวมทฤษฎี หลักเกณฑ์ และงานวิจัยต่างๆ ที่มีส่วนเกี่ยวข้องกับงานวิจัย โดยแบ่งออกเป็นหัวข้อต่างๆ ดังนี้

ทฤษฎีที่เกี่ยวข้อง

ส่วนนี้จะแสดงรายละเอียดของทฤษฎีความรู้ที่เกี่ยวข้อง เพื่อเป็นแนวทางในการพัฒนาระบบ และใช้ในการศึกษาเพิ่มเติมให้กับงานวิจัยอีกด้วย

1. อัตภาควิเคราะห์ (Move Analysis)

บทความวิจัยทางวิทยาศาสตร์นั้นมีการแบ่งบทความวิจัยออกเป็น 4 ส่วนหรือภาค คือ ภาคบทนำ (Introduction) ภาควิธีวิจัย (Methods) ภาคผลวิจัย (Results) และภาคอภิปรายผลวิจัย (Discussion) หรือทางภาษาศาสตร์เรียกว่า IMRD นอกจากนี้แบทความทางวิชาการทางด้านวิทยาศาสตร์มีการเรียบเรียงโครงสร้างในแต่ละภาคอย่างเป็นแบบแผน (Conventional) ซึ่งสเวลส์เป็นผู้คิดค้นแนววิเคราะห์สัมพันธสาร (Discourse analysis) ขึ้นมาที่มีชื่อว่า Move analysis หรืออัตถภาควิเคราะห์

เนื่องจากไม่มีศัพท์บัญญัติสำหรับการเรียกคำที่เกี่ยวข้องกับอัตถภาควิเคราะห์ หรือ Move analysis ผู้วิจัยจึงใช้คำศัพท์ที่มีคำจำกัดความดังต่อไปนี้

อัตถภาควิเคราะห์ (Move Analysis) คือ แนวการวิเคราะห์สัมพันธสารที่สำคัญโดยคิดค้นขึ้นจากสเวลส์ (John M. Swales) และได้นำแนวอัตถภาควิเคราะห์มาใช้ในการวิเคราะห์โครงสร้างของประโยคในภาษาอังกฤษ ที่มีเป้าหมายเบื้องต้นของอัตถภาควิเคราะห์ คือ เพื่อแยกสัมพันธสารทั้งที่เป็นภาษาพูด และภาษาเขียนออกเป็นหน่วยย่อยตามหน้าที่ในการสื่อสารที่เรียกว่าอัตถภาค หรือมูฟ ดังปรากฏในคำนิยามที่สเวดลส์ให้ไว้ว่า

" A " move" in genre analysis is a discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse."

ดังนั้นอัตถภาค หรือมูฟจึงหมายถึงภาคของถ้อยคำ หรือข้อความที่แสดงหน้าที่ในการสื่อสารที่มีลักษณะจำเพาะเจาะจงของแต่ละอัตถภาคนั้น และยังแสดงความสัมพันธ์ต่อวัตถุประสงค์โดยรวมของตัวบทด้วยโดยอัตถภาคอาจมีขนาดเล็กคือเป็นเพียงวลี (Phrase) ประโยคย่อย หรืออนุพากย์ (Clause) ประโยค (Sentence) หรืออาจมีขนาดใหญ่เป็นถ้อยคำเรียง (Utterance) หรืออนุเฉท (Paragraph) (บุษบา กนกศิลปธรรม & สุดาพร ลักษณียนาวิน, 2549)

2. การสร้างคลังข้อมูลบทคัดย่อของเอกสารทางวิทยาศาสตร์สาขาวิศวกรรมชีวเวช

บุษา และสุดาพร (2549) กล่าวว่าการสร้างคลังข้อมูลสาขาวิศวกรรมชีวเวชภาษาอังกฤษเริ่มต้นจากการเลือกวารสารบทความทางวิชาการ 60 บทความ โดยบทความที่เลือกนั้นจะต้องคัดเลือกมาจากวารสารที่เป็นที่ยอมรับของนักวิชาการทั่วไป ผู้วิจัยสร้างคลังข้อมูลโดยยึดค่าดัชนีผลกระทบอ้างอิง (Impact factor) เป็นเกณฑ์สำคัญในการเลือกพิจารณาวารสาร โดยพิจารณาเห็นว่าเป็นเกณฑ์ที่เหมาะสมที่สุดที่จะสามารถแสดงว่าวารสารใดเป็นที่ยอมรับในสาขาวิศวกรรมชีวเวช

จากการพิจารณาค่าดัชนีผลกระทบอ้างอิงประจำปี ค.ศ. 2005 ผู้วิจัยเลือกวารสารนานาชาติ 5 รายชื่อ ที่มีค่าดัชนีผลกระทบอ้างอิงสูงสุด 5 อันดับของสาขาวิศวกรรมชีวเวช ดังตารางที่ 2.1 (ไม่รวมวารสารบรรณนิทัศน์หรือ Reviews บทบรรณาธิการ หรือ Editorials บทความพิเศษหรือ Special articles และบททบทวนพิเศษหรือ State of the art)

ตารางที่ 2.1 รายละเอียดวารสารนานาชาติที่มีค่าดัชนีผลกระทบอ้างอิงสูงสุด

ชื่อวารสาร	ค่าดัชนีผลกระทบอ้างอิง
1. IEEE Transactions on Medical Imaging (TMI)	3.757
2. Journal of Biomedical Materials Research (BMR)	2.497
3. IEEE Transactions on Biomedical Engineering (TBM)	2.302
4. Artificial Organs (AOR)	1.903
5. IEEE Transactions on Neutral Systems and Rehabilitation Engineering (TNS)	1.842

ที่มา: Citation Report 2005: Science Edition

บทความที่คัดเลือกเพื่อนำมาสร้างเป็นคลังข้อมูลจำนวน 60 บทความในสาขาวิศวกรรมชีวเวชโดยจำแนกตามอัตภาคดังตารางที่ 2.2 เป็นบทความที่ตีพิมพ์ในปี ค.ศ. 2006 เท่านั้น การที่ไม่ได้นำวารสารในช่วงระยะเวลาหลายปีก็เนื่องจากความเป็นไปได้ที่การเขียนบทความจะมีการเปลี่ยนแปลงพัฒนาการตามกาลเวลา

ตารางที่ 2.2 จำนวนตัวบทจำแนกตามอัตถภาค จำนวน 60 บทความ

	Background	Purpose	Method	Result	Discussion
วิศวกรรมชีวเวช	42	50	59	60	47

3. ไวยากรณ์

งานวิจัยนี้ต้องการสร้างเครื่องมือในการวิเคราะห์โครงสร้างและองค์ประกอบสำหรับบทคัดย่อของเอกสารทางวิชาการในรูปแบบโปรแกรมประยุกต์บนเว็บไซต์ โดยเอกสารทางวิชาการที่ใช้วิเคราะห์นั้นใช้ภาษาอังกฤษในการเขียน จึงมีความจำเป็นต้องทราบถึงโครงสร้างของไวยากรณ์ภาษาอังกฤษ (สุวรรณี เต็งอำนวย & อรสา เจนพนัส, 2557) ได้ให้รายละเอียดไวยกรณ์ของภาษาอังกฤษ ดังนี้

3.1 Articles

Articles คือ คำที่ใช้นำหน้าคำนาม มี 2 ชนิด คือ

1. Indefinite Article ได้แก่ A, An
2. Definite Articles ได้แก่ the

3.2 Pronoun

Pronoun (คำสรรพนาม) คือ คำที่ใช้แทนคำนาม เพื่อหลีกเลี่ยงการใช้คำนามซ้ำ คำสรรพนามจะทำหน้าที่เช่นเดียวกับกลุ่มคำนามที่สรรพนามนั้นไปแทนที่ จึงไม่มีคำมาขยายสรรพนามอีก แบ่งออกเป็น 5 รูปตามหน้าที่การใช้ทางไวยากรณ์ ดังตารางที่ 2.3

ตารางที่ 2.3 หน้าที่ของ Pronoun ทางไวยากรณ์

Nominative Case (ใช้เป็นประธาน)	Accusative Case (ใช้เป็นกรรม)	Possessive Adjective (คุณศัพท์เจ้าของ)	Possessive Pronoun (สรรพนามเจ้าของ)	Reflexive (เน้นย้ำความเป็นเจ้าของ)
I	Me	My	Mine	Myself
We	Us	Our	Ours	Ourselves
You	You	Your	Yours	Yourself
They	Them	Their	Theirs	Themselves
He	Him	His	His	Himself
She	Her	Her	Hers	Herself
It	It	It	Its	Itself

สำหรับการใช้ Pronoun ที่พบในแต่ละมูฟมักจะพบคำว่า We, Our ซึ่งจะพบมากในมูฟที่เป็น Purpose และได้นำคำคุณศัพท์เฉพาะ (Demonstrative adjective) ได้แก่ This, These แสดงความเป็นเจ้าของงานหรือการกล่าวถึงงานของตัวเอง

3.3 Preposition

Preposition (บุพบท) คือ คำที่ใช้แสดงความสัมพันธ์ระหว่างคำนามหรือสรรพนามกับคำอื่นๆ ในประโยค เช่น in, over, between, into, at, from, by, for, on, though, during เป็นต้น

3.4 Modal Verb

Modal Verb คือ กริยาช่วยได้แก่ can, could, may, might, will, would, shall, should เป็นต้น ใช้แสดงทั้งความเป็นไปได้ การคาดคะเนและแนวโน้มความเป็นไปได้

3.5 Infinitive with to

Infinitive with to คือ กริยาที่ไม่ได้ผันไปตามประธานหรือ Tense นำหน้าด้วย to โดยใช้ Infinitive with to เพื่อแสดงความปรารถนา แสดงวัตถุประสงค์หรือเหตุผล เป็นต้น

3.6 In order to

In order to มีความหมายว่า เพื่อที่จะ ... (เป็นการแสดงวัตถุประสงค์)

3.7 Tense

Tense (กาล) คือ รูปแบบของกริยาที่แสดงให้ทราบว่า การกระทำหรือเหตุการณ์นั้นเกิดขึ้นเมื่อไหร่แบ่งออกเป็น 3 ชนิด คือ

1. Present Tense ปัจจุบันกาล
2. Past Tense อดีตกาล
3. Future Tense อนาคตกาล

สำหรับงานนี้จะพิจารณาเพียงแค่ 3 Tense จะไม่พิจารณาในส่วนของ Aspect ซึ่งหมายถึงว่าสิ่งที่เกิดขึ้นนั้นจบลงแล้วหรือยังดำเนินการอยู่

3.8 Active voice and Passive voice

Active Voice คือ คำกล่าวหรือประโยคที่ประธานเป็นผู้กระทำหรือแสดงกริยาโดยตรง

เช่น **He** punished a boy. เขาทำโทษเด็กชายคนหนึ่ง
Mary eats a mango. แมรี่กินมะม่วง

Passive Voice หมายถึง ประโยคหรือข้อความที่ประธานเป็นผู้ถูกกระทำกริยานั้นโดยผู้อื่นหรือสิ่งอื่น

เช่น **A boy** was punished by him. เด็กถูกทำโทษโดยเขา
A mango is eaten by Mary. มะม่วงถูกกินโดยแมรี่

3.9 Extraposition “It” construction

Extraposition“It” construction คือ การกล่าวถึงในสิ่งที่คนรู้กันแล้วสำหรับรูปแบบที่นำมาใช้ คือ

It + verb to be + v.3 + that

เช่น

- It is concluded that specific multiwall carbonnanotube loadings can favorably improve the mechanical performance of bone cement.

- It is commonly accepted that locomotor-related neuronal circuitry resides in the lumbosacral spinal cord.

3.10 Nominalization

ภาษาอังกฤษมีโครงสร้าง หรือวิธีช่วยปรับระดับภาษาให้มีความเป็นทางการมากขึ้นโดยการทำ Nominalization คือ การทำให้เป็นคำนาม ซึ่งเกิดจากการเปลี่ยนประโยคทั้งหมดให้เป็นรูปคำนามทั้งหมด เช่น

- The noisy result of this simulation is de-noised by a three-dimensioned fitting of Gaussian basis functions.

- The effect of diffuse fibrosis and gap junction remodeling is simulated by reducing cellular coupling nonuniformly.

4. Regular Expressions

Regular expressions (RE) หรือเรียกว่า regex หรือ regexp คือ รูปแบบของข้อความที่สร้างขึ้นในรูปแบบของสตริง ที่สามารถจับคู่กับบางส่วน หรือทั้งหมดของข้อความที่เราสนใจ เพื่อแสดงรูปแบบใดๆ ที่ต้องการซึ่งมีประสิทธิภาพและยืดหยุ่น (López & Romero, 2014) โดยการใช้งาน regular expression ถูกนำไปใช้ได้ดีในหลากหลายสถานการณ์ต่างๆ ซึ่งทำให้การค้นหาสารสนเทศได้ง่ายมากยิ่งขึ้น อีกทั้งสามารถใช้ในการค้นหา และแทนที่ชื่อไฟล์ที่มีปริมาณมากได้ (Goyvaerts, 2007) ตัวอย่างการ regular expression ไปใช้แสดงดังต่อไปนี้ เช่น

- การตรวจสอบรูปแบบของการให้ข้อมูล เช่น ตรวจสอบรูปแบบของจดหมายอิเล็กทรอนิกส์
- เพื่อค้นหารูปแบบที่ปรากฏในบางส่วนของข้อความ เช่น ตรวจสอบการเขียนของคำว่า “color” หรือ “colour” ที่ปรากฏในเอกสาร
- เพื่อค้นหาส่วนใดส่วนหนึ่งของข้อความ เช่น ค้นหาจดหมายอิเล็อทรอนิกส์ หรือ ค้นหารหัสไปรษณีย์
- เพื่อแทนที่ข้อความบางส่วน เช่น เปลี่ยนคำที่ปรากฏ “color” หรือ “colour” เป็นคำว่า “red”
- แบ่งข้อความที่มีขนาดใหญ่ ออกเป็นข้อความที่มีขนาดเล็กลง เช่น แบ่งข้อความออกจากกันเมื่อพบกับสัญลักษณ์ต่างๆ dot, comma หรือ สัญลักษณ์ขึ้นบรรทัดใหม่

โดยโครงสร้างไวยกรณ์ของ regular expression ที่พบนั้นจะประกอบไปด้วยองค์ประกอบ 2 ชนิดด้วยกัน (Bird, Klein, & Loper, 2009) ดังนี้

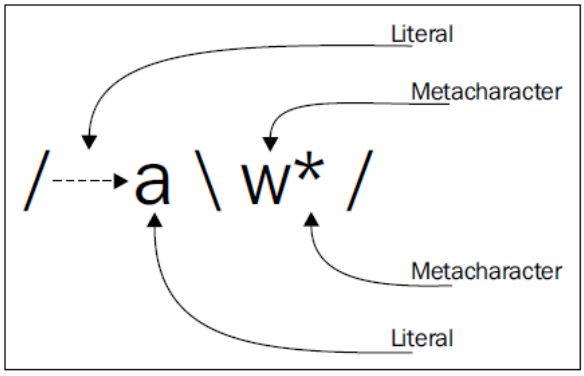
1. Literals คือ ตัวอักขระ หรือตัวอักษรโดยที่ทุกตัว เช่น ตัวอักษร a – z ตัวเลข 0 - 9
2. Metacharacters คือ อักขระพิเศษที่ใช้แทนเซตของสายอักขระ แสดงดังตัวอย่างในตารางที่ 2.4 ดังนี้

ตารางที่ 2.4 แสดงอักขระพิเศษที่ใช้แทนสายอักขระ

สัญลักษณ์	คำอธิบาย
.	ใช้บ่งบอกถึงอักขระในตำแหน่งนั้นของนิพจน์จะเป็นอักขระใดก็ได้ยกเว้นขึ้นบรรทัดใหม่
^acb	ใช้จับคู่ตำแหน่งเริ่มต้นของประโยค
abc$	ใช้จับคู่ตำแหน่งสุดท้ายของประโยค
[abc]	ใช้แทนชุดอักขระตัวหนึ่งที่อยู่ในขอบเขตซึ่งระบุในวงเล็บ
[A-Z0-9]	ใช้แทนช่วงชุดของอักขระที่ระบุอยู่ในขอบเขตซึ่งระบุในวงเล็บ
ed | ing | s	ใช้สำหรับสร้างทางเลือกที่จะใช้ค้นหานิพจน์
*	ใช้บ่งบอกว่ามีนิพจน์ก่อนหน้านี้จำนวน 0 นิพจน์หรือมากกว่า
+	ใช้บ่งบอกว่ามีนิพจน์ก่อนหน้านี้จำนวน 1 นิพจน์หรือมากกว่า
?	ใช้บ่งบอกว่ามีนิพจน์ก่อนหน้านี้จำนวน 0 หรือ 1 นิพจน์ (มีหรือไม่มีก็ได้)
{n}	ใช้บ่งบอกว่าต้องมีนิพจน์ก่อนหน้านี้จำนวน n นิพจน์
{n,}	ใช้ทำซ้ำอย่างน้อย n นิพจน์
{,n}	ใช้ทำซ้ำไม่มากกว่า n นิพจน์
{m,n}	ใช้ทำซ้ำอย่างน้อย m แต่ไม่มากกว่า n นิพจน์
a(b|c)+	วงเล็บเป็นตัวแบ่งขอบเขตการกระทำของเครื่องหมายต่างๆ

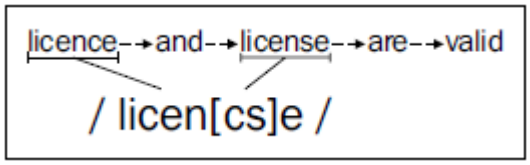
ในระบบปฏิบัติการคอมพิวเตอร์ มีการใช้เครื่องหมายดอกจัน (*) หรือเครื่องหมายคำถาม (?) เพื่อการค้นหาไฟล์ต่างๆ ในเครื่องคอมพิวเตอร์ เครื่องหมายดอกจันจะแสดงถึงตัวอักขระเดี่ยวกับค่าต่างๆ ที่อยู่บนชื่อไฟล์ เช่น file?.xml เป็นรูปแบบทที่จะค้นหาไฟล์ file1.xml, file2.xml และ file3.xml แต่ไม่สามารถจะจับคู่ไฟล์ file99.xml โดยรูปแบบดังกล่าวจะสามารถที่จะค้นหาไฟล์ใดๆ ที่เริ่มต้นด้วยคำว่า file และตามด้วยอักขระเดี่ยวใดๆ และตามด้วย .xml ส่วนเครื่องหมายดอกจัน เมื่อมีการใช้งานในลักษณะเดียวกัน คือ file*.xml จะค้นหาไฟล์ใดๆ ที่เริ่มต้นด้วย file และตามด้วยตัวเลขใดๆ ก็ได้ และตามด้วย .xml

รูปแบบของ regular expression ประกอบด้วยอักขระพื้นฐาน เช่น a ถึง z ตัวเลข 0 – 9 และอักขระพิเศษที่เรียกว่า metacharacters ที่กล่าวไว้ข้างต้น ตัวอย่างของ regular expression ที่จะจับคู่กับคำใดๆ ที่เริ่มต้นด้วยตัวอักษร a แสดงได้ในรูปที่ 2.1



รูปที่ 2.1 แสดงรูปแบบของ regular expression ที่ใช้ตัวอักขระและอักขระพิเศษที่ค้นหาคำใดๆ ที่เริ่มต้นด้วยตัวอักษร a

Character class คือ รูปแบบหนึ่งของ regular expression จะเป็นกลุ่มของอักขระ (character set) ที่นำอักขระพิเศษ หรือ metacharacter มาใช้ในการกำหนดรูปแบบของคำที่ต้องการค้นหาการ กำหนด character class นั้นเริ่มต้นจากวงเล็บเปิด ([) จากนั้นตามด้วยอักขระใดๆ ที่ต้องการ และสุดท้ายปิดด้วยเครื่องหมายวงเล็บปิด (]) ตัวอย่างเช่น การสร้างรูปแบบ regular expression ที่สามารถค้นหาคำว่า “license” ที่เขียนในรูปแบบของ British English และ American English ดังแสดงในรูปที่ 2.2



รูปที่ 2.2 แสดงรูปแบบ regular expression

ฟังก์ชันที่ใช้การค้นหารูปแบบของข้อความด้วย Regular expressions มีหลายฟังก์ชัน ดังแสดงตัวอย่างในตารางที่ 2.5

ตารางที่ 2.5 ฟังก์ชันที่ใช้ในการค้นหา Pattern

Method	Purpose
re.match()	ค้นหาข้อความจากจุดเริ่มต้นที่ตรงกันกับ RE ที่กำหนดไว้
re.search()	ค้นหาข้อความจากบริเวณ หรือตำแหน่งใดๆ ที่ตรงกันกับ RE ที่กำหนดไว้
re.findall()	ค้นหาส่วนของข้อความทั้งหมดที่ตรงกันกับ RE และคืนผลลัพธ์เป็น list
re.finditer()	ค้นหาส่วนของข้อความทั้งหมดที่ตรงกันกับ RE และคืนผลลัพธ์เป็น iterator

5. Natural Language Toolkit

Natural Language Toolkit หรือ NLTK (Bird et al., 2009) เป็นแพลตฟอร์มหรือชุดคำสั่งเพื่อการใช้งานที่เกี่ยวข้องกับทางด้านภาษาธรรมชาติ (Natural Language Processing) สำหรับการสร้างโปรแกรมภาษาไพธอน โดย NLTK ถูกพัฒนามาเพื่อช่วยแก้ปัญหาการเขียนโปรแกรมด้านการประมวลผลภาษา มีฟังก์ชันที่สามารถทำงานได้หลากหลาย เช่น Corpus readers, Tokenization, Parsers, Classifiers, Part-of-speech Tagging เป็นต้น มี Corpus ให้ใช้ทดสอบถึง 50 Corpus มีการแบ่งหมวดหมู่ เช่น คลังประโยคข่าว คลังบทความหนังสือที่ไม่มีลิขสิทธิ์ อาจแบ่งเป็นหมวดการเมือง กีฬา บันเทิง และอาชญากรรมซึ่งในคลังข้อมูลของ NLTK ได้ออกแบบมาให้มีความสมดุลและความหลากหลายของข้อมูลเพื่อตอบสนองตรงตามความต้องการของการนำไปใช้งาน

NLTK ยังได้รับการยกย่องว่าเป็นเครื่องมือที่ยอดเยี่ยมสำหรับการสอนและการทำงานด้านภาษาศาสตร์ สำหรับการค้นคว้าอิสระนี้ ได้นำฟังก์ชันใน NLTK มาใช้ร่วมในการวิเคราะห์บทความ มีดังนี้

5.1 Sentence Tokenize

การแบ่งประโยคในข้อความหรือบทความนั้น แบ่งแต่ละประโยคในข้อความหรือบทความเพื่อจำกัดการวิเคราะห์ในแต่ละประโยค โดยใช้ฟังก์ชันชื่อ sent_tokenize จาก NLTK เพื่อแบ่งประโยคจากบทคัดย่อที่ได้รับเข้ามาแล้วนำไปวิเคราะห์ระดับประโยคต่อไป ดังตัวอย่างการทำงานต่อไปนี้

“Tom always gets up at six o’clock. Then he eats breakfast. He usually eat bread and sausage before feeds the cat. He always eats lunch at twelve. After lunch time he always goes to the library until one o’clock.”

จากบทความเมื่อทำการ Sentence Tokenize จะได้ลิสต์ของประโยค ดังนี้

['Tom always gets up at six o’clock.', 'Then he eats breakfast.', 'He usually eat bread and sausage before feeds the cat.', 'He always eats lunch at twelve.', 'After lunch time he always goes to the library until one o\x92clock.']

5.2 Word Tokenize

การแบ่งคำในประโยคหรือข้อความในภาษาอังกฤษนั้นมักมีคำผู้เขียนทำการย่อในการเขียน เช่น do not เป็น don’t และ they will เป็น they’ll ในส่วนนี้จะใช้ฟังก์ชันชื่อ word_tokenize จาก NLTK เพื่อแบ่งคำจากประโยคเพื่อเพิ่มประสิทธิภาพความถูกต้อง ดังตัวอย่างการทำงานต่อไปนี้

“The students in the room know they'll be grilled on each day's case study.”

จากประโยคดังกล่าวเมื่อทำการ word tokenize จะได้คำที่ทำการแบ่งแล้ว ดังนี้

['The', 'students', 'in', 'the', 'room', 'know', 'they', "'ll", 'be', 'grilled', 'on', 'each', 'day', "'s", 'case', 'study', '.']

5.3 Part of Speech Tagger

Part of Speech Tagger (POS Tag) คือ การกำกับชนิดของคำตามการใช้งานของคำ ประเภทหรือชนิดของคำในภาษาอังกฤษซึ่งมีหน้าที่และตำแหน่งที่แตกต่างกันไปในประโยค สำหรับภาษาอังกฤษคำเดียวกัน ชนิดของคำอาจต่างกันขึ้นอยู่กับตำแหน่งและหน้าที่ของคำนั้นในประโยค ใน NLTK มีคลังประโยคจำนวนมากได้ทำการทดสอบและกำกับชนิดของคำไว้เรียบร้อยแล้ว หมวดหมู่ชนิดของคำที่ใช้กัน คือ

1. Noun คำนาม
2. Pronoun คำสรรพนาม
3. Verb กริยา
4. Adverb กริยาวิเศษณ์
5. Adjective คำคุณศัพท์
6. Prepositionบุพบท

7. Conjunction คำสันธาน

8. Interjection คำอุทาน

ตัวอย่าง การกำกับชนิดของคำในประโยค ดังตารางที่ 2.6

ตารางที่ 2.6 ตัวอย่างการกำกับชนิดของคำ

ประโยค	She	likes	big	snakes	but	I	hate	Them.
POS	pronoun	verb	adjective	noun	conjunction	pronoun	verb	pronoun
POS tag	PRP	VBZ	JJ	NNS	CC	PRP	VBP	PRP

การวัดประสิทธิภาพของ Part of Speech Tagger

การวัดประสิทธิภาพการทำงานของ Tagger นำคลังข้อมูลที่มีการกำกับชนิดของคำไว้ คือ Treebank Corpus ซึ่งเป็น Corpus ของ NLTK ที่มีทั้งลิสต์ของคำที่มีการกำกับชนิดของคำไว้แล้ว และลิสต์ของคำที่ยังไม่ได้กำกับชนิดของคำ โดยมีคำทั้งหมด 100,676 คำ

คลังต้นไม้ (Treebank) คือ คลังข้อความที่ในแต่ละประโยคได้กำกับโครงสร้างวากยสัมพันธ์ (Syntax) โครงสร้างวากยสัมพันธ์มักจะแทนด้วยโครงสร้างต้นไม้ ซึ่งเป็นที่มาของคำว่าคลังต้นไม้

คลังต้นไม้มักสร้างบนคลังประโยคที่ได้กำกับชนิดของคำไว้แล้ว ในลักษณะเดียวกันคลังต้นไม้ก็สามารถใช้เป็นฐานในการกำกับข้อมูลทางความหมายหรือข้อมูลทางภาษาศาสตร์อื่นๆ

ในการวัดประสิทธิภาพ ได้นำลิสต์ของคำที่ยังไม่ได้กำกับชนิดของคำมากำกับชนิดของคำทั้งหมดด้วย Part of Speech Tagger จากนั้นนำผลลัพธ์ที่ได้ไปเปรียบเทียบกับลิสต์ของคำที่กำกับชนิดของคำไว้แล้วว่าคำเดียวกันของทั้งสองลิสต์ มีการกำกับชนิดของคำเหมือนกันหรือไม่ พบว่ามีการกำกับชนิดของคำตรงกันทั้งหมด 100,242 คำ คิดเป็น 99%

5.4 Stop word

Stop word คือ คำที่ไม่มีความหมายในเนื้อหาเอกสาร ซึ่งเป็นคำที่สำคัญ การตัดคำที่ไม่สำคัญออกจะไม่มีผลทำให้ประสิทธิภาพการค้นหาข้อมูลต่ำลง และทำให้การค้นหาแต่ละครั้งเร็วขึ้น ตัวอย่างคำที่ไม่สำคัญในภาษาอังกฤษ เช่น a, an, the, in, out, before, up, down เป็นต้น เนื่องจากคำที่ไม่สำคัญในภาษาอังกฤษมีการระบุไว้อย่างชัดเจนว่ามีคำอะไรในรายการที่เรียกว่า Stop Word List ของ NLTK ที่สามารถเรียกใช้ได้ทันที

5.5 Stemming

การประมวลผลทางภาษา (Linguistic Processing) เพื่อเปลี่ยนรูปคำศัพท์จากรูปแบบเดิม ให้อยู่ในรูปแบบรากศัพท์ เช่น เปลี่ยนคำศัพท์ที่เป็นพหูพจน์ให้เป็นรูปแบบเอกพจน์ เปลี่ยนคำศัพท์ที่อยู่ในรูปแบบของ Tense ต่างๆ ให้เป็นคำรากศัพท์

```
>>> porter = nltk.PorterStemmer()
>>> lancaster = nltk.LancasterStemmer()
>>> [porter.stem(t) for t in tokens]
['DENNI', ':', 'Listen', ',', 'strang', 'women', 'lie', 'in', 'pond',
'distribut', 'sword', 'is', 'no', 'basi', 'for', 'a', 'system', 'of', 'govern',
'.', 'Suprem', 'execut', 'power', 'deriv', 'from', 'a', 'mandat', 'from',
'the', 'mass', ',', 'not', 'from', 'some', 'farcic', 'aquat', 'ceremoni', '.']
>>> [lancaster.stem(t) for t in tokens]
['den', ':', 'list', ',', 'strange', 'wom', 'lying', 'in', 'pond', 'distribut',
'sword', 'is', 'no', 'bas', 'for', 'a', 'system', 'of', 'govern', '.', 'suprem',
'execut', 'pow', 'der', 'from', 'a', 'mand', 'from', 'the', 'mass', ',', 'not',
'from', 'som', 'farc', 'aqu', 'ceremony', '.']
```

5.6 Lemmatization

การประมวลผลทางภาษา (Linguistic Processing) เพื่อเปลี่ยนรูปคำให้อยู่ในรูปแบบดั้งเดิมเช่น เปลี่ยนจากพหูพจน์ให้เป็นรูปแบบเอกพจน์ธรรมดา, Past tense ให้เป็น Tense ปกติเปลี่ยนตัวพิมพ์ใหญ่เป็นตัวพิมพ์เล็ก เปลี่ยนคำที่ลงท้ายด้วย –ing เป็นคำเดิม เช่น Running เป็น run เป็นต้น

5.7 Collocation and Bigrams

Collocation คือการจัดวางคำ หรือกลุ่มคำที่ต้องใช้ควบคู่กันในประโยค ซึ่งมีความสำคัญในการเขียน เพราะอาจจะทำให้ความหมายของวลี หรือประโยคมีความหมายพิเศษหรือผิดเพี้ยนไปจากเดิม เช่น red wine, big wind ซึ่งเราสามารถหา Collocation ได้

โดยการหา Bigrams ซึ่งเป็นคำคู่กันที่มีความถี่สูงโดยเรียกใช้ nltk.bigrams(*list_word*) โดยพารามิเตอร์เป็นลิสต์ของคำ ดังตัวอย่างการทำงานของ bigrams ต่อไปนี้

['Cells', 'interact', 'with', 'their', 'extracellular', 'matrix', 'through', 'cell', 'adhesion', 'contacts', '.']

จากลิสต์ของคำที่ได้จาก Word Tokenize จะได้ลิสต์ของ Bigrams ดังนี้

[('Cells', 'interact') , ('interact', 'with') , ('with', 'their') , ('their', 'extracellular'), ('extracellular', 'matrix'), ('matrix', 'through'), ('through', 'cell'), ('cell', 'adhesion'), ('adhesion', 'contacts'), ('contacts', '.')]

5.8 Frequency Distribution

การค้นพบคำที่มีการกล่าวถึงมากในเอกสาร ควรจะเป็นคำที่สื่อถึงหมวดหมู่ของเอกสารนั้น NLTK มีเครื่องมือที่สำหรับหา Frequency Distribution ที่ใช้ในการประมวลผลภาษาคือ FreqDist ซึ่งใช้ในส่วนของการนับความถี่ของคำศัพท์เพื่อสร้างไฟล์ความถี่ของมูฟแต่ละกลุ่มโดยความถี่ที่นับ คือ คำศัพท์ Bigrams และคำพร้อมแยกชนิดของคำเรียกใช้ nltk.FreqDist(*list_vocabulary*)

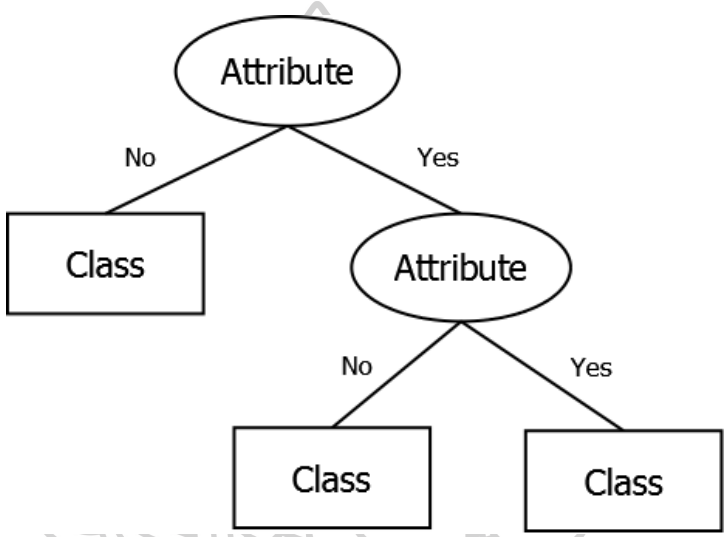
6. Decision Tree (เอกสิทธิ์, 2557)

ต้นไม้ตัดสินใจ (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557) คือ การเรียนรู้ของเครื่องที่นิยมใช้มากที่สุดรูปแบบหนึ่งเป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้ การเรียนรู้นี้เป็นการเรียนรู้โดยการจำแนก (Classification) ข้อมูลออกเป็นกลุ่ม (Class) ใช้คุณสมบัติ (Attribute) ของข้อมูลในการจำแนกข้อมูล ต้นไม้ตัดสินใจที่ได้เรียนรู้ทำให้ทราบว่าคุณสมบัติใดของข้อมูลที่เป็นตัวการการจำแนก และคุณสมบัติแต่ละตัวของข้อมูลมีความสำคัญมากน้อยต่างกัน ซึ่งเป็นประโยชน์ช่วยให้ผู้ใช้สามารถวิเคราะห์ข้อมูลและตัดสินใจได้ถูกต้องยิ่งขึ้น

ลักษณะโครงสร้างของต้นไม้ตัดสินใจประกอบไปด้วยโหนด (Node) ซึ่งแต่ละโหนดมีคุณลักษณะ (Attribute) ต่างๆ ของข้อมูลที่เป็นตัวทดสอบ โดยใช้คุณสมบัตินี้เป็นตัวตัดสินใจว่าข้อมูลนั้นไปในทิศทางใดกิ่งของต้นไม้ (Branch) แสดงให้เห็นถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบ และส่วนใบ (Leaf) ซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจ แสดงถึงกลุ่มของข้อมูล (Class) ที่เป็นผลลัพธ์ที่ได้จากการจำแนกข้อมูลโหนดที่อยู่บนสุดของต้นไม้เรียกว่า โหนดราก (Root node)

6.1 ลักษณะการเรียนรู้ของต้นไม้ตัดสินใจ

- ผลลัพธ์ของการเรียนรู้แสดงอยู่ในรูปที่สามารถเข้าใจได้ง่าย ทำให้สะดวกต่อการวิเคราะห์คุณสมบัติที่มีผลต่อการจำแนกกลุ่มต่างๆ
- เส้นทางจากโหนดรากไปถึงโหนดใบสามารถแสดงให้อยู่ในรูปกฎ IF-THEN
- ทนทานต่อข้อมูลมีสัญญาณรบกวน (Noisy data) เช่น คุณสมบัติที่ไม่เกี่ยวข้องและค่าคุณสมบัติที่ผิดพลาด หรือขาดหาย
- การเรียนรู้มีความรวดเร็วเมื่อเทียบกับอัลกอริทึมสำหรับการจำแนกชนิดอื่น



รูปที่ 2.3 แสดงโครงสร้างต้นไม้ตัดสินใจ

6.2 วิธีการเรียนรู้ของต้นไม้ตัดสินใจ (Decision Tree Learning)

การสร้างต้นไม้ตัดสินใจ มีลักษณะการสร้างแบบบนลงล่าง (Top-down) โดยเริ่มจากการสร้างรากของต้นไม้ แล้วจึงสร้างกิ่งต่อไปจนถึงใบ โดยแสดงขั้นตอนการสร้างต้นไม้ตัดสินใจได้ดังนี้

1. ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึกสอน (Training Set)
2. ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกัน ให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น
3. ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่ามาตรฐานเกน (Gain criterion) ของแต่ละคุณสมบัติเพื่อที่จะใช้เป็นเกณฑ์ในการคัดเลือกคุณสมบัติที่สามารถแบ่งแยก

ข้อมูลออกเป็นกลุ่มต่างๆ ได้ดีที่สุดโดยคุณสมบัติที่มีค่าเกนมากสุดจะถูกเลือกให้เป็นตัวทดสอบในการตัดสินใจ

4. กิ่งของต้นไม้ถูกสร้างขึ้นจากค่าต่างๆ ที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งที่สร้างขึ้น
5. วนซ้ำเพื่อหาคุณสมบัติที่มีค่าเกนมากที่สุด สำหรับข้อมูลที่แบ่งแยกออกมาในแต่ละกิ่ง เพื่อนำคุณสมบัตินี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่คุณสมบัติที่ถูกเลือกมาเป็นโหนดแล้วไม่สามารถถูกเลือกมาเป็นโหนดสำหรับระดับต่อๆ ไปอีกได้
6. ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อยๆ การวนซ้ำจะสิ้นสุดเมื่อเงื่อนไขข้อใดข้อหนึ่งเป็นจริง

6.3 การคำนวณค่ามาตรฐานเกน (Gain criterion)

การคำนวณค่ามาตรฐานเกน (Gain criterion) ใช้เกณฑ์ที่ใช้ช่วยประกอบการเลือกคุณสมบัติ คุณสมบัติใดที่ให้ค่าเกนสูงที่สุด จะแสดงว่าคุณสมบัตินั้นสามารถจำแนกกลุ่มข้อมูลได้ดีที่สุด การใช้ค่าเกนช่วยลดจำนวนครั้งของการทดสอบการแยกแยะข้อมูล อีกทั้งต้นไม้ตัดสินใจที่ได้ไม่มีความซับซ้อนมากเกินไปในการหาความสัมพันธ์ของคุณสมบัตินี้จะใช้ตัววัดที่เรียกว่า Information Gain (IG) ค่านี้คำนวณได้จากสมการ

$$IG_{(parent,\ child)} = entropy_{(parent)} - [p(c_1) \times entropy(c_1) + p(c_2) + \ldots$$

โดยที่ค่า entropy คือ ค่าที่ใช้วัดความมีระเบียบ และความไร้ระเบียบของข้อมูล โดยข้อมูลที่มีระเบียบจะให้ค่า entropy ที่มีค่าต่ำ ส่วนระบบที่ไร้ระเบียบจะให้ค่า entropy ที่มีค่าสูงซึ่งคำนวณได้จาก

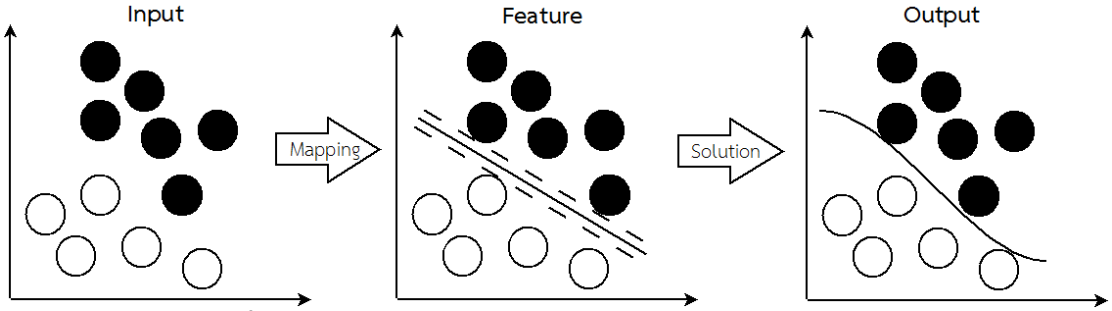
$$entropy(c_1) = -p(c_1)\ \log_p(c_1)$$

โดย p(c1) คือ ความน่าจะเป็นของ c1

เมื่อคำนวณค่า IG ของทุกคุณสมบัติแล้ว จึงเลือกคุณสมบัติที่มีค่า IG มากที่สุดขึ้นมาเป็นโหนดราก ถ้าข้อมูลที่อยู่ในโหนดมีคลาสเดียวกันทั้งหมด โหนดนี้ไม่ต้องทำการแตกกิ่งต่อไป แต่ถ้าไม่ใช่คลาสเดียวกันทั้งหมดต้องแตกกิ่งออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกัน

7. Support Vector Machine

Support Vector Machine (SVM) (นัศพ์ชาณัณ ชินปัญช์ธนะ, 2557) เป็นการเรียนรู้ด้วยเครื่องแบบมีผู้สอน (Supervise learning) เพื่อให้สามารถสร้างตัวจัดประเภทของข้อมูล (Classifier) ที่มีความหลากหลายมาก ดังนั้น SVM เป็นการเรียนรู้ด้วยเครื่องที่ใช้ในการแบ่งข้อมูล โดยสามารถทำงานได้ดีกับข้อมูลที่ไม่รู้จัก (Unknown dataset) ด้วยกระบวนการปรับเปลี่ยนรูปแบบข้อมูลให้อยู่ในรูปแบบที่มีมิติต่ำ (Low dimension dataset) บนพื้นที่ข้อมูลนำเข้า (Input space) ให้อยู่ในรูปแบบของข้อมูลที่มีมิติสูง (High dimension dataset) บนพื้นที่ข้อมูลคุณลักษณะ (Feature space) โดยใช้ฟังก์ชันในการปรับรูปแบบข้อมูลที่เรียกว่าฟังก์ชันเคอร์เนล (Kernel function) ซึ่งมีความสามารถดังกล่าว สามารถทำให้การสร้างตัวจำแนกประเภทข้อมูลด้วยสมการกำลังสอง (Quadratic equation) บนพื้นที่ข้อมูลคุณลักษณะเป็นไปได้ง่ายขึ้น และมีความชัดเจนในการจัดประเภทมากยิ่งขึ้น นอกจากนี้ตัวจำแนกประเภทข้อมูลที่ดีควรมีโครงสร้างแบบเส้นตรง (Linear Classifier) และสามารถสร้างพื้นที่ระยะห่างระหว่างตัวจำแนกประเภทข้อมูลเองกับค่าที่ใกล้ที่สุดของแต่ละกลุ่มข้อมูลได้มากที่สุด เพื่อประสิทธิภาพในการแบ่งประเภทของชุดข้อมูลแต่ละประเภทออกจากกันอย่างชัดเจน ซึ่งเส้นที่เหมาะสมจะถูกเรียกว่า ระนาบแบ่งเขตข้อมูลที่เหมาะสม (Optimal separating hyperplane) โดยหลักการในการทำงานเพื่อจำแนกประเภทข้อมูลของวิธี Support Vector Machine สามารถแสดงได้รูปที่ 2.4

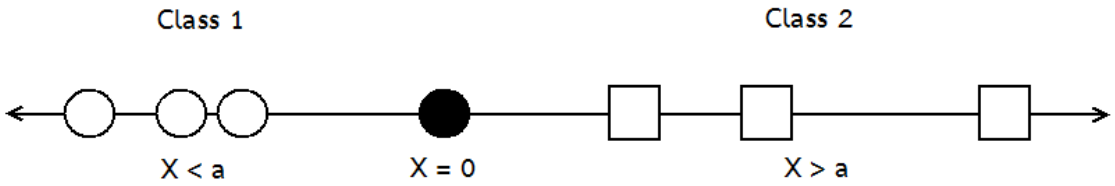


รูปที่ 2.4 แสดงหลักการทำงานของ Support Vector Machine

SVM จึงเป็นเครื่องมือที่ใช้ในการจำแนกประเภทที่มีความนิยมใช้ในงานหลากหลายด้าน เช่น การจัดหมวดหมู่เอกสาร การรู้จำใบหน้า โดยให้ประสิทธิภาพที่ดี ทั้งนี้เนื่องจาก SVM มีการจดจำข้อมูลที่อยู่บริเวณของแต่ละคลาสที่เรียกว่า เวกเตอร์ซัพพอร์ท (Support vector) อีกทั้งรองรับปริภูมิที่มีข้อมูลเป็นแบบเชิงเส้น และแบบไม่เชิงเส้นได้อีกด้วย โดยการอาศัยฟังก์ชันคอร์เนล (Kernel) ในการลดความซับซ้อนของข้อมูลโดยการส่ง (Mapping) ไปยังปริภูมิพิเศษที่มีความเป็นเชิงเส้น (Kernel trick) ภายในปริภูมิผลคูณภายใน (ปริญญา สงวนสัตย์, 2558)

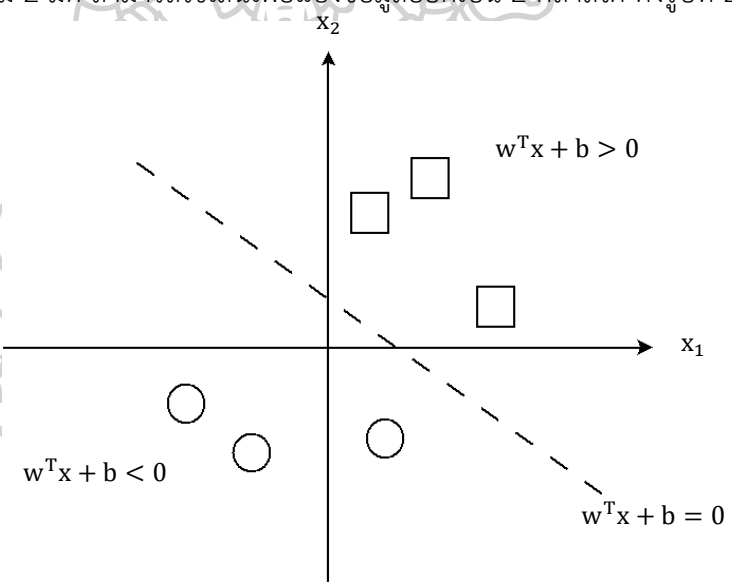
7.1 ระนาบเกิน (Hyperplane)

ระนาบเกิน หรือ Hyperplane ในบริภูมิต่างๆ แสดงในภาพ ดังต่อไปนี้ ในปริภูมิ 1 มิติ สามารถใช้จุดเพื่อแบ่งข้อมูลออกเป็น 2 คลาสได้ดังรูปที่ 2.5



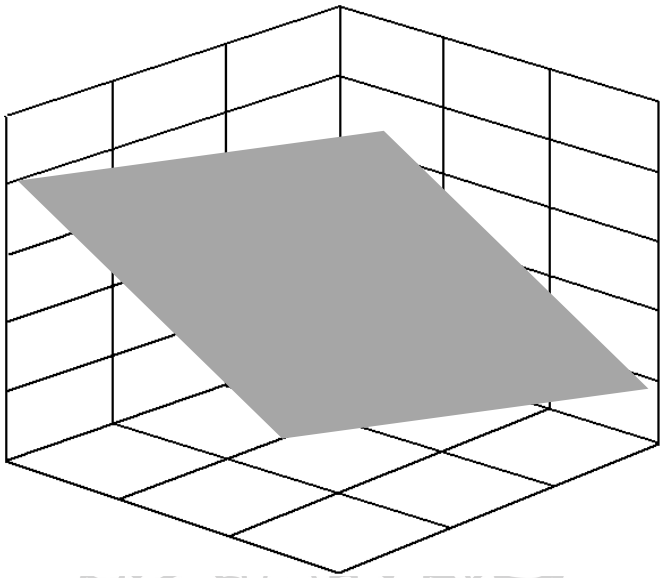
รูปที่ 2.5 แสดงตัวอย่างระนาบเกินในปริภูมิ 1 มิติ

ในปริภูมิ 2 มิติ สามารถใช้เส้นเพื่อแบ่งข้อมูลออกเป็น 2 คลาสได้ ดังรูปที่ 2.6



รูปที่ 2.6 แสดงตัวอย่างระนาบเกินในปริภูมิ 2 มิติ

ในปริภูมิ 3 มิติ สามารถใช้ระนาบเพื่อแบ่งข้อมูลออกเป็น 2 คลาสได้ ดังรูปที่ 2.7



รูปที่ 2.7 แสดงตัวอย่างระนาบเกินในปริภูมิ 3 มิติ

จากตัวอย่างพบว่าระนาบเกินไม่ว่าอยู่ในปริภูมิกี่มิติก็ตาม สามารถเขียนสมการของระนาบเกินในรูปเวกเตอร์ได้เหมือนกันคือ

$$w^{T}x + b = 0$$

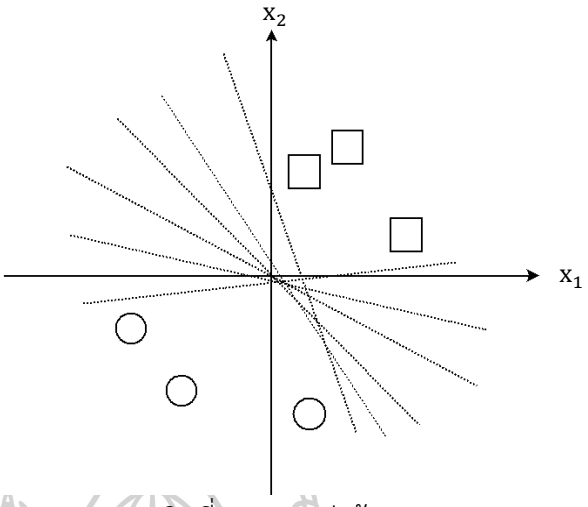
b คือ ค่าไบแอส หรือระยะเลื่อนขนานจากจุดกำเนิด

w คือ เวกเตอร์ปกติ (Normal vector) ซึ่งเวกเตอร์นี้จะตั้งฉากกันกับระนาบเกิน และมีมิติเท่ากับมิติของปริภูมินั้นๆ

x คือ เวกเตอร์นำเข้า (Input Vector)

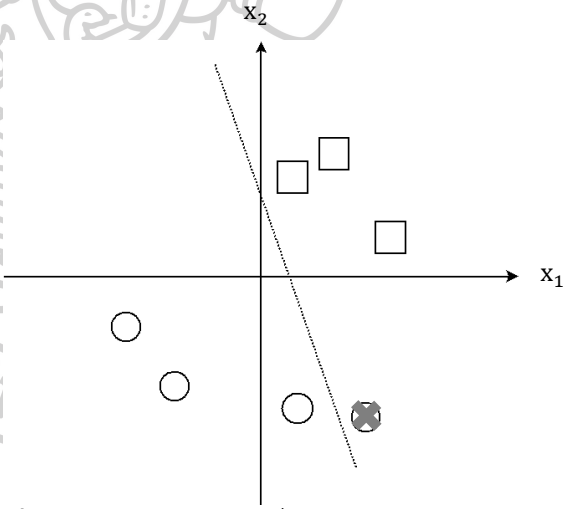
7.2 ระนาบเกินแบบบัญญัติ (Canonical Hyperplane)

ระนาบเกินที่ใช้ในการแบ่งแยกข้อมูล 2 คลาสออกจากกัน ระนาบเกินแบบใดเป็นระนาบที่ดีที่สุด ดังตัวอย่างในรูปที่ 2.8



รูปที่ 2.8 แสดงระนาบเกินที่สามารถแข่งข้อมูล 2 คลาสออกจากกัน

พบว่าทุกระนาบเกินในรูปที่ 2.8 สามารถใช้แบ่งข้อมูลชุดนี้ได้ถูกต้องทั้งหมด แต่หากมีข้อมูลค่าใหม่เกิดขึ้นมาในกรณีดังรูปที่ 2.9

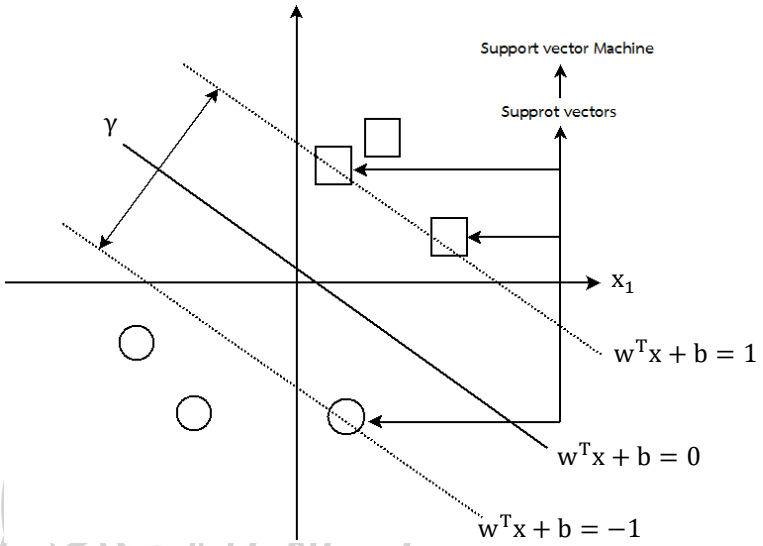


รูปที่ 2.9 แสดงระนาบเกินที่เกิดปัญหา Over-fitting

พบว่าข้อมูลค่าใหม่ถูกตัดสินผิดพลาด ซึ่งเรียกระนาบเกินนี้ว่าเกิดปัญหา Over-fitting ดังนั้น ระนาบเกินลักษณะนี้จึงไม่ควรเลือกใช้งาน โดยระนาบเกินที่ดีควรมีลักษณะที่สามารถใช้งานได้ทั่วไป (Generalization)

การแก้ไข้ปัญหา Over-fitting นั้นทำได้โดย หากทราบชุดข้อมูลทั้งหมดทุกกรณี เนื่องจากความเป็นจริงไม่สามารถทราบถึงข้อมูลทั้งหมดได้ ดังนั้นสามารถทำได้เพียงการหลีกเลี่ยงปัญหา Over-fitting เท่านั้นโดยส่วนใหญ่จะใช้หลักการการทำให้ใช้ทั่วไปได้ (Generalization)

ใน SVM ระยะขอบ (Margin) ที่มากที่สุดระหว่างคลาสหนึ่งกับอีกคลาสหนึ่งถูกใช้เป็นบรรทัดฐานในการตัดสินเลือกระนาบเกินที่เหมาะสมที่สุด (Optimal Hyperplane) และเวกเตอร์ข้อมูลที่อยู่บริเวณขอบของแต่ละคลาสจะถูกเรียกว่า เวกเตอร์ซัพพอร์ท (Support vector) แสดงดังในรูปที่ 2.10



รูปที่ 2.10 แสดงเวกเตอร์ซัพพอร์ท

โดย γ คือ ระยะของ (Margin) และข้อมูลบริเวณของขอบของแต่ละคลาสที่มีเครื่องหมายลูกศรชี้อยู่ทั้งหมด คือ เวกเตอร์ซัพพอร์ท ซึ่งเวกเตอร์ซัพพอร์ทเหล่านี้ถูกบันทึกไว้ในตัวจำแนกประเภท (Classifier) ซึ่งจำนวนเวกเตอร์ซัพพอร์ทยิ่งมาก ขนาดหน่วยความจำที่ใช้เก็บตัวจำแนกประเภทก็ยิ่งมากตามไปด้วย หากจำนวนเวกเตอร์ซัพพอร์ทมีสัดส่วนมากเกินไปเมื่อเทียบกับจำนวนข้อมูลที่นำไปใช้ฝึกฝนทั้งหมดจะบอกเราได้ถึงปัญหา Over-fitting ที่เกิดขึ้น

เมื่อทราบเกณฑ์ที่ใช้ในการเลือกระนาบเกินที่เหมาะสม ซึ่งเกี่ยวข้องกับระยะขอบ ดังนั้นความสัมพันธ์ระหว่างระยะขอบกับระนาบเกินต้องถูกนำมาพิจารณา โดยเริ่มต้นจากเวกเตอร์ซัพพอร์ทของแต่ละคลาส อย่างละเวกเตอร์ (x_+ , x_-) ซึ่งสอดคล้องกับสมการระนาบเกินในรูปที่ 2.7 ดังนี้

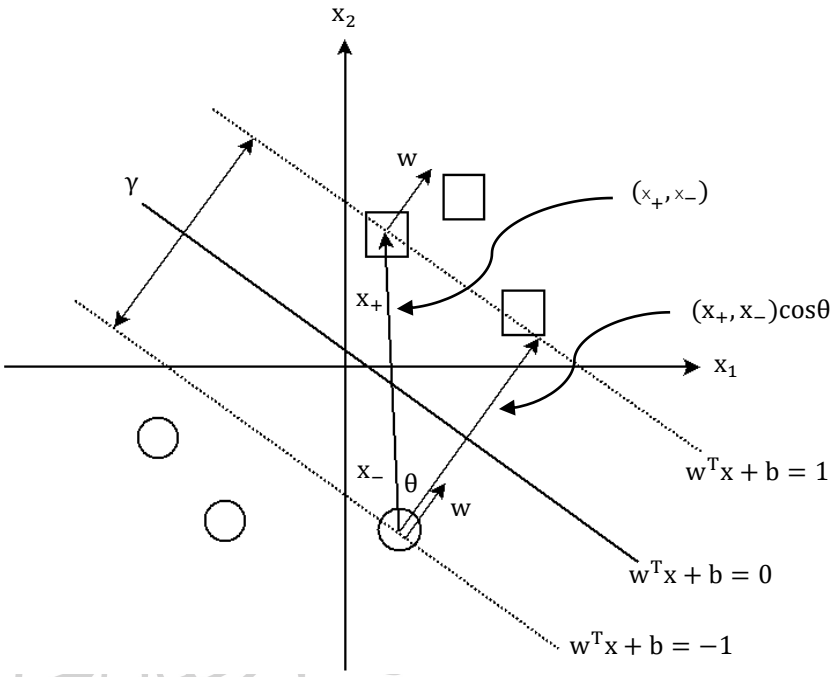
$$w^T x_+ + b = 1 \qquad (1)$$

$$w^T x_- + b = -1 \qquad (2)$$

นำสมการที่ 1 ลบ สมการที่ 2 จะได้

$$w^T (x_+ , x_-) = 2 \qquad (3)$$

โดยจะปรากฏเทอม (x_+ , x_-) ซึ่งเป็นเวกเตอร์ลัพท์ที่เริ่มจากระนาบเกินคลาสหนึ่งไปอีกคลาสหนึ่ง และกระทำมุม θ กับเวกเตอร์ปกติ w ดังรูปที่ 2.11



รูปที่ 2.11 แสดงความสัมพันธ์ระหว่างระยะขอบกับเวกเตอร์ปกติของระนาบเกิน

จากรูปที่ 2.11 เวกเตอร์ที่มีทิศทางเดียวกับเวกเตอร์ปกติ w และมีขนาดเท่ากับระยะขอบ γ สามารถหาได้จาก

$$((x_+ - x_-)\cos\theta = (x_+ - x_-)\left(\frac{w^\top (x_+ - x_-)}{\|w\|\|x_+ - x_-\|}\right) \qquad (4)$$

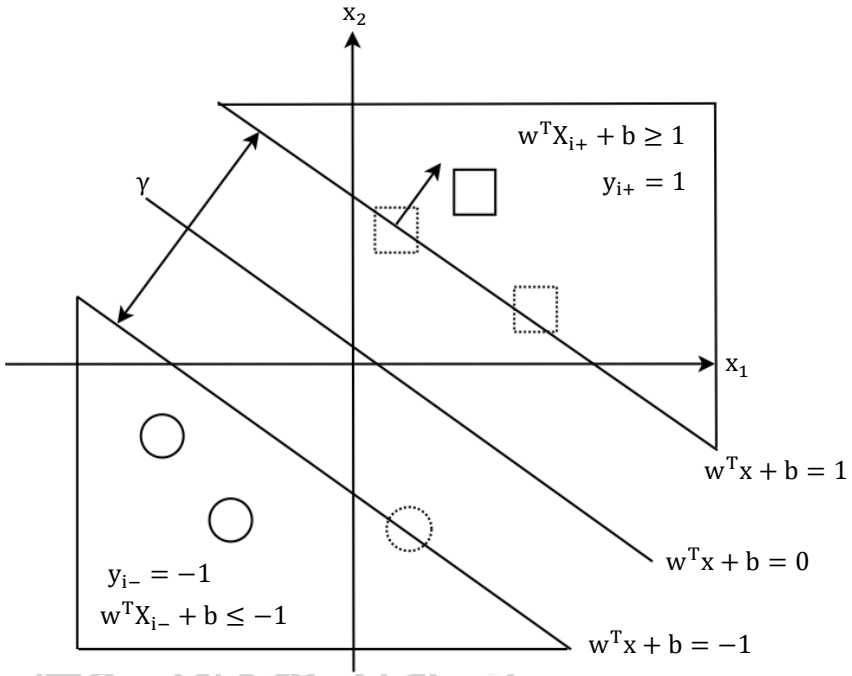
แทนค่าจากสมการที่ 3 ลงในสมการที่ 4 จะได้

$$((x_+ - x_-)\frac{w^\top (x_+ - x_-)}{\|w\|\|x_+ - x_-\|} = \frac{2}{\|w\|}\left(\frac{(x_+ - x_-)}{\|x_+ - x_-\|}\right) \qquad (5)$$

โดยเทอม $\frac{(x_+ - x_-)}{\|x_+ - x_-\|}$ คือ เวกเตอร์หนึ่งหน่วย (Unit vector) ที่มีทิศทางเดียวกับเวกเตอร์ปกติ w ดังนั้นระยะขอบ γ จึงมีค่าเท่ากับขนาดของเวกเตอร์ในสมการที่ 5 กล่าวคือ

$$\gamma = \frac{2}{\|w\|} \qquad (6)$$

เมื่อได้ความสัมพันธ์ระหว่างระยะขอบ และเวกเตอร์ปกติของระนาบเกินแล้วจึงสามารถสร้างปัญหาการหาค่าเหมาะสมที่สุดจากวัตถุประสงค์ที่ต้องการให้ระยะขอบมีความกว้างมากที่สุดในกรณีการจำแนกประเภทแบบ 2 คลาส (Binary Classification) อันได้แก่ คลาสที่มีค่าเป็นบวก (+) และคลาสที่มีค่าเป็นลบ (-) โดยการพิจารณาเงื่อนไขเป็นพื้นที่ในบริเวณของแต่ละคลาส ดังรูปที่ 2.12



รูปที่ 2.12 พื้นที่แสดงอาณาเขตของแต่ละคลาสในกรณีไบนารี

โดยพื้นที่ในบริเวณทั้งสองสามารถนำมารวมเป็นอสมการเดียวกันได้โดยการนำค่าของคำตอบ y ของข้อมูลแต่ข้อมูลไปคูณกับอสมการทั้งสองบริเวณ ดังนี้

$$y_i(w^T x_i + b) \geq 1, \forall_i \qquad (7)$$

จากความสัมพันธ์ของระยะขอบกับระนาบเกินในสมการที่ 6 ร่วมกับเงื่อนไขในอสมาการดังนั้น ปัญหาเพื่อการหาค่าเหมาะสมที่สุดจะเป็นดังนี้

Maximize: $\gamma = \frac{2}{\|w\|}$	(ให้ระยะขอบมีค่ากว้างที่สุด)
Subject to: $y_i(w^T x_i + b) \geq 1, \forall_i$	(ให้ข้อมูลอยู่ในบริเวณของคลาสที่ถูกต้อง)

เพื่อความสะดวกในการแก้ปัญหาซึ่งมักนิยมหาค่าน้อยที่สุดแทนที่จะหาค่ามากที่สุด ซึ่งปัญหานี้สามารถพิจารณาความสัมพันธ์ต่อไปนี้

$$\gamma = \frac{2}{\|w\|} = \frac{2}{\sqrt{w^T w}} \alpha \frac{2}{w^T w} \qquad (8)$$

ดังนั้นปัญหาการหาค่าที่เหมาะสมที่สุดนี้จะสามารถเขียนได้ใหม่เป็น

Maximize: $\frac{1}{2} w^T w$	(ให้ส่วนกลับของระยะขอบมีค่ากว้างที่สุด)
Subject to: $y_i(w^T x_i + b) \geq 1, \forall_i$	(ให้ข้อมูลอยู่ในบริเวณของคลาสที่ถูกต้อง)

จะพบว่าปัญหานี้อยู่ในรูปแบบกำหนดการกำลังสอง (Quadratic Programming, QP) ซึ่งมีลักษณะรูปแบบทั่วไป

Maximize: $J(w) = \frac{1}{2} w^T Q w + c^T w$

Subject to: $Aw \leq b, Ew = d, l \leq w \leq u$

ซึ่งไม่สามารถหาผลเฉลยรูปแบบปิด (Closed-from solution) ได้แต่สามารถหาผลเฉลยแบบทำซ้ำ (Iterative solution) ได้ด้วยขั้นตอนวิธีต่างๆ เช่น Trust regions หรือ Active set เป็นต้น

7.3 ซัพพอร์ทเวกเตอร์แมทชีนแบบหลายคลาส (Multi-class SVM)

SVM เป็นตัวจำแนกประเภทแบบไบนารี (2 คลาส) แต่การประยุกต์ในงานจริงนั้น หลายครั้งมีจำนวนคลาสที่มากกว่า 2 คลาส การทำให้ SVM สามารถนำไปใช้กับปัญหาหลายคลาส (Multi-class) ได้นั้นมี 2 แนวทาง คือ การรวม SVM แบบไบนารีหลายตัวเข้าด้วยกัน หรือ การพิจารณาทุกคลาสพร้อมกัน โดยกล่าวถึงวิธีการ 3 วิธีการ ได้แก่ 1-v-a, 1-v-1 และ DAG-SVM ดังรายละเอียดต่อไปนี้

7.3.1 One-against-all (1-v-a) เป็นการพิจารณาคลาสใดคลาสหนึ่งเทียบกับคลาสอื่นๆ ที่เหลือทั้งหมด หรือบางครั้งอาจเรียกว่า 1-v-r หรือ one-versus-rest โดยการกำหนดค่าคำตอบ y ดังนี้

$$y_i = \begin{cases} +1, x_i \in C_k \\ -1, x_i \notin C_k \end{cases}$$

เมื่อ C_k คือ คลาสใดคลาสหนึ่งที่ถูกพิจารณาในแต่ละครั้ง โดย $1 \leq k \leq K$ และจะถูกพิจารณาทั้งหมดทุกคลาส จำนวน K ครั้ง ตัวอย่างเช่น ในกรณีมีจำนวนคลาสทั้งหมด คือ 3 (K=3) เราจะต้องสร้าง SVM จำนวนทั้งสิ้น K ดังนี้

1. SVM ตัวที่ 1 เกิดจากการกำหนดให้ตัวอย่างทั้งหมดในคลาสที่ 1 มีค่าคำตอบเป็น +1 และสมาชิกที่เหลือทั้งหมดนอกจากในคลาสที่ 1 มีค่าคำตอบเป็น -1

2. SVM ตัวที่ 2 เกิดจากการกำหนดให้ตัวอย่างทั้งหมดในคลาสที่ 2 มีค่าคำตอบป็น +1 และสมาชิกที่เหลือทั้งหมดนอกจากในคลาสที่ 2 มีค่าคำตอบเป็น -1

3. SVM ตัวที่ 3 เกิดจากการกำหนดให้ตัวอย่างทั้งหมดในคลาสที่ 3 มีค่าคำตอบเป็น +1 และสมาชิกที่เหลือทั้งหมดนอกจากในคลาสที่ 3 มีค่าคำตอบเป็น -1

7.3.2 One-against-one เป็นการพิจารณาคลาสในคลาสหนึ่งเทียบกับคลาสอื่นๆ ที่เหลืออีกหนึ่งคลาส หรือบางครั้งอาจเรียกว่า 1-v-1, one-versus-one หรือ Pairwise SVM โดยการกำหนดค่าคำตอบ y ของ SVM แบบไบนารีที่พิจารณาเฉพาะคลาส C_j และคลาส C_k ดังนี้

$$y_i = \begin{cases} +1, x_i \in C_j \\ -1, x_i \in C_k \end{cases}$$

เมื่อ C_j คือคลาสใดคลาสหนึ่งที่ถูกพิจารณาเทียบกับคลาสอื่น คือ C_k ในแต่ละครั้งโดย $j \neq k$ และจะถูกพิจารณาทั้งหมดทุกคลาสในลักษณะการจัดหมู่จำนวน $K(K-1)/2$ ครั้ง จากนั้นใช้วิธีการแมกซ์วิน (Max Wins) ในการทดสอบตัวอย่าง ตัวอย่างเช่น ในกรณีที่มีจำนวนคลาสทั้งหมด คือ 3 (K=3) เราจะต้องสร้าง SVM แบบไบนารีจำนวนทั้งสิ้น 3 ตัว ดังนี้

1. SVM ตัวที่ 1 เกิดจากการกำหนดให้ตัวอย่างทั้งหมดในคลาสที่ 1 มีค่าคำตอบเป็น +1 และตัวอย่างทั้งหมดในคลาสที่ 2 มีค่าคำตอบเป็น -1

2. SVM ตัวที่ 2 เกิดจากการกำหนดให้ตัวอย่างทั้งหมดในคลาสที่ 1 มีค่าคำตอบเป็น +1 และตัวอย่างทั้งหมดในคลาสที่ 3 มีค่าคำตอบเป็น -1

3. SVM ตัวที่ 3 เกิดจากการกำหนดให้ตัวอย่างทั้งหมดในคลาสที่ 2 มีค่าคำตอบเป็น +1 และตัวอย่างทั้งหมดในคลาสที่ 3 มีค่าคำตอบเป็น -1

เนื่องจากจำนวนคลาสเป็น 3 ทั้งตัวอย่างวิธี 1-v-a และ 1-v-1 ทำให้ได้จำนวน SVM เท่ากัน แต่ในกรณีมากว่า เช่น จำนวนคลาสเป็น 10 ขั้นตอนวิธี 1-v-a จะมีจำนวน SVM เท่ากับจำนวนคลาส คือ 10 แต่ขั้นตอนวิธี 1-v-1 จะมีจำนวน SVM มากกว่าคือ 45 แต่ในการหาค่าที่เหมาะสมที่สุดในแต่ละครั้งจะลู่เข้าเร็วกว่า เพราะมีจำนวนตัวอย่างที่ใช้ในการพิจารณาในแต่ละครั้งน้อยกว่า

8. ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล

เอกสิทธิ์ (2557) กล่าวว่า การนำตัวจำแนกประเภทของข้อมูลไปใช้งานนั้นผู้ใช้จำเป็นต้องทราบถึงประสิทธิภาพของตัวจำแนกประเภทของข้อมูล ซึ่งงานวิจัยต่างๆ นิยมใช้ค่าวัดประสิทธิภาพต่างๆ ดังนี้ คือ

1. Precision เป็นการวัดความแม่นยำโดยพิจรณาแยกทีละคลาส
2. Recall เป็นการวัดคงามถูกต้องโดยพิจารณาแยกทีละคลาส
3. F-measure เป็นการวัดค่า Precision และ Recall พร้อมกันโดยพิจารณาแยกทีละคลาส
4. Accuracy เป็นการวัดความถูกต้องโดยพิจารณารวมทุกคลาส

การวัดประสิทธิภาพอาศัยตัววัดประสิทธิภาพของตัวจำแนกประเภทข้อมูล (Confusion Metrix) โดย Confusion Metrix คือ ตารางแบบจัตุรัสโดยมีจำนวนแถวเท่ากับจำนวนคอลัมน์ และเท่ากับจำนวนคลาส เช่น ข้อมูลมีคลาสคำตอบอยู่ 2 ค่า คือ yes และ no ฉะนั้นตาราง Confusion Metrix จะสร้างได้เป็นตารางขนาด 2x2 ซึ่งข้อมูลด้านคอลัมน์ คือ คลาสที่อยู่ในข้อมูล training data (actual) และข้อมูลในแนวแถว คือ คลาสที่ตัวจำแนกประเภทของข้อมูลทำนายได้มา (predicted) ดังตารางที่ 2.7

ตารางที่ 2.7 แสดงตาราง Confusion Metrix ของข้อมูลซึ่งมี 2 คลาส

Predicted / Actual	yes	no
yes	TP	FP
no	FN	TN

จากในตารางที่ 2.7 ค่าที่แสดงในช่องต่างๆ ของตารางประกอบไปด้วย

- True Positive (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งกำลังสนใจอยู่
- True Negative (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งไม่ได้สนใจอยู่
- False Positive (FP) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งกำลังสนใจอยู่
- False Negative (FN) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งไม่ได้สนใจอยู่

หลังจากที่สร้างตาราง Confusion Matrix สามารถที่จะคำนวณค่า Precision, Recall, F-measure, และ Accuracy ได้ดังต่อไปนี้

- Precision เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส

$$Precision = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Positive}$$

- Recall เป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาส

$$Recall = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Negative}$$

- F-measure เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- Accuracy เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส คือ จำนวน True Positive ของทุกคลาสรวมกัน

การจำแนกประเภทของข้อมูลต้องแบ่งข้อมูลออกเป็น 2 ชุดด้วยกันคือ ชุดข้อมูลสอน (Training set) เพื่อใช้สร้างโมเดลจำแนกประเภทข้อมูล และข้อมูลชุดทดสอบ (Testing set) เพื่อให้โมเดลจำแนกประเภทข้อมูลทำนายค่าคลาสคำตอบ มีวิธีการ 3 วิธีดังนี้

1. Self consistency test

Self consistency test หรือ Use training set เป็นวิธีการที่ง่าย ข้อมูลที่ใช้ในการสร้างโมเดลจำแนกประภทข้อมูล และข้อมูลที่ใช้ในการทดสอบเป็นข้อมูลชุดเดียวกัน เริ่มต้นสร้างโมเดลจำแนกประเภทข้อมูลด้วย Training data จากนั้นนำโมเดลที่สร้างได้มาทำนายด้วย Training data ชุดเดิม การวัดประสิทธิภาพวิธีนี้ให้ผลการวัดประสิทธิภาพที่มีค่าสูงมาก เนื่องจากเป็นข้อมูลชุดเดิมระบบได้ทำการเรียนรู้มาแล้ว แต่ผลการวัดประสิทธิภาพวิธีการนี้เหมาะสำหรับใช้ในการทดสอบประสิทธิภาพเพื่อดูแนวโน้มของโมเดลที่สร้างขึ้น ถ้าได้ผลการวัดประสิทธิภาพที่น้อยแสดงว่าโมเดลไม่เหมาะสมกับข้อมูล จึงไม่ควรจะนำไปทดสอบด้วยวิธีการแบ่งข้อมูลแบบต่างๆ

2. Split test

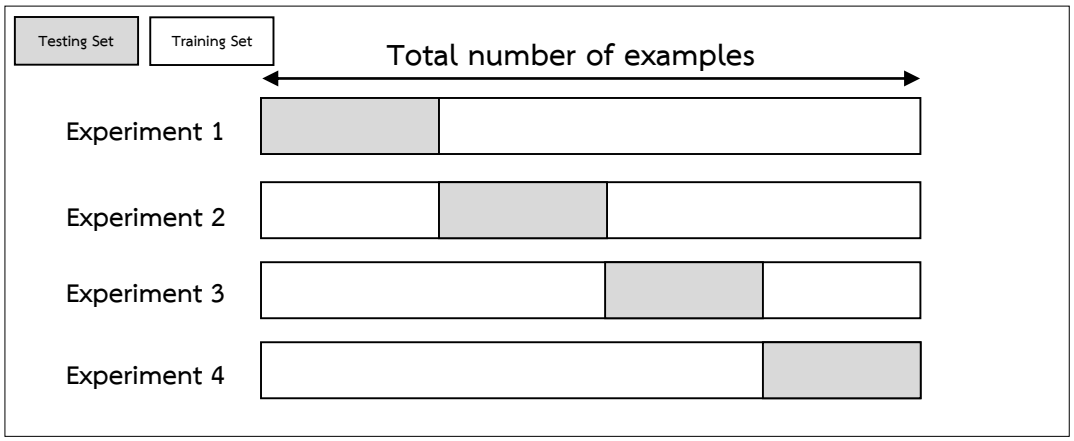
Split test เป็นการแบ่งข้อมูลด้วยวิธีการสุ่มออกเป็น 2 ส่วนเช่น 70% ต่อ 30% หรือ 80% ต่อ 20% โดยข้อมูลส่วนที่หนึ่ง (70% และ 80%) ใช้ในการสร้างโมเดลจำแนกประเภทข้อมูล และข้อมูลส่วนที่สอง (30% และ 20%) ใช้ในการทดสอบประสิทธิภาพของโมเดลจำแนกประเภทข้อมูล การทดสอบนี้ทำการสุ่มข้อมูลเพียงครั้งเดียวซึ่งในบางครั้งถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะคล้ายกับข้อมูลที่ใช้ในการสร้างโมเดลทำให้ผลการวัดประสิทธิภาพที่ดี ในทางตรงข้ามถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะแตกต่างกับข้อมูลที่ใช้สร้างโมเดลจำแนกประเภทข้อมูล จะให้ผลการวัดประสิทธิภาพที่ไม่ดี ดังนั้นจึงควรใช้วิธี Split test คือ ทำการสุ่มหลายๆ ครั้ง โดยข้อเสีย คือ ใช้เวลาน้อยในการสร้างโมเดลจำแนกประเภทข้อมูล ซึ่งเหมาะกับชุดข้อมูลที่มีขนาดใหญ่มากกว่า

3. Cross-validation test

Cross-validation test เป็นที่นิยมใช้ในการทดสอบประสิทธิภาพของโมเดลจำแนกประเภทของข้อมูล การวัดประสิทธิภาพด้วยวิธีนี้ทำการแบ่งข้อมูลออกเป็นหลายส่วน เช่น 5-fold cross-validation คือ ทำการแบ่งข้อมูลทั้งหมดออกเป็น 5 ส่วน โดยแต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดลของโมเดลจำแนกประเภทของข้อมูล ทำวนเช่นนี้จนครบจำนวนที่แบ่งไว้ โดยมีหลักการเลือกข้อมูล 2 วิธี คือ เลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม (K-Fold Cross Validation) และเลือกสุ่มข้อมูลแบบ Leave one out มีรายละเอียดดังนี้

1. การเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม (K-Fold cross validtion)

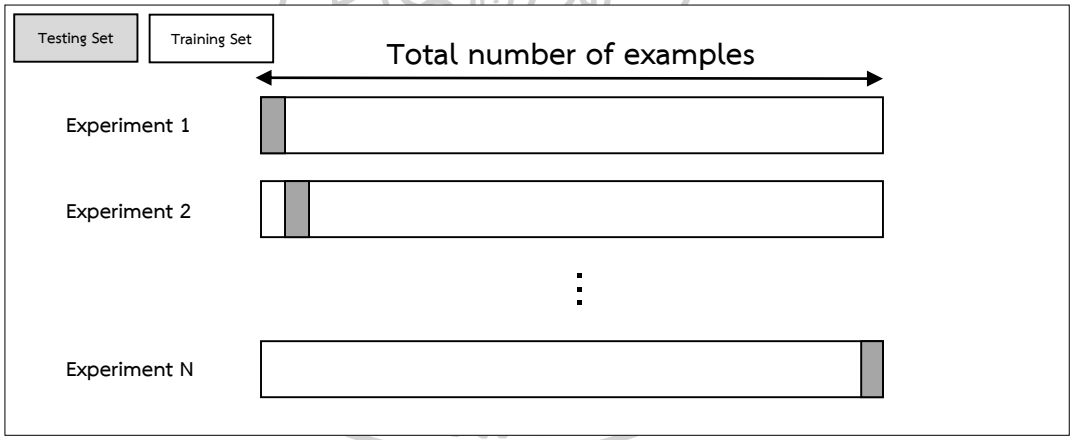
วิธีการนี้แบ่งข้อมูลออกเป็นกลุ่มจำนวน K กลุ่ม โดยเลือกข้อมูลกลุ่มที่ 1 เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลชุดสอน นำข้อมูลไปจำแนกประเภท จากนั้นจะสลับข้อมูลกลุ่มที่ 2 มาเป็นชุดทดสอบและข้อมูลกลุ่มอื่นที่เหลือเป็นชุดสอน โดยสลับอย่างนี้ไปเรื่อยๆ จนครบ K กลุ่ม ในขั้นตอนสุดท้ายหาค่าเฉลี่ยของค่าความถูกต้อง วิธีการนี้ข้อมูลทุกตัวอย่างจะได้เป็นทั้งชุดทดสอบและชุดสอนดังรูปที่ 2.13



รูปที่ 2.13 แสดงการเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม เมื่อ K=4

2. การเลือกสุ่มข้อมูลแบบ Leave one out

การเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม เมื่อกำหนดให้ K มีค่าเท่ากับจำนวนแถวข้อมูลทั้งหมด (N) ดังรูปที่ 2.14



รูปที่ 2.14 แสดงการเลือกสุ่มข้อมูลแบบ Leave one out

งานวิจัยที่เกี่ยวข้อง

การศึกษางานวิจัยในเรื่องที่เกี่ยวข้องกับกระบวนการวิเคราะห์โครงสร้าง และมูฟของเอกสารทางวิชาการในส่วนบทคัดย่อ ทำให้ทราบว่ามีผู้วิจัยนำเอาแนวคิดของกระบวนการการวิเคราะห์โครงสร้าง และมูฟของเอากสารทางวิชาการในส่วนบทคัดย่อนำมาประยุกต์ใช้ในการพัฒนาโปรแกรมประยุกต์เพื่อใช้ในงานต่างๆ โดยแนวคิดที่เกี่ยวข้องกับกับงานวิจัยได้นำเสนอไว้ดังนี้

บุษบา กนกศิลปธรรม และสุดาพร ลักษณียนาวิน (2549) ได้ศึกษาเกี่ยวกับอัตถภาควิเคราะห์ (move analysis) นับเป็นการวิเคราะห์สัมพันธสารที่สำคัญแนวหนึ่งโดยมีจุดมุ่งหมายเพื่อกำหนดว่าตัวบทที่วิเคราะห์นั้นประกอบด้วยโครงสร้างของอัตถภาคและอนุวัจน์อะไรบ้างโดยจุดประสงค์หลักของงานวิจัย เพื่อกำหนดว่าแต่ละภาค มีโครงสร้างทั้งในระดับอัตถภาค (move) และอนุวัจน์ (step) อะไร และบทความวิจัยทั้งฉบับซึ่งประกอบด้วย 4 ภาคหลักมีการเรียงตัวของอัตถภาคและอนุวัจน์อย่างไร ขั้นตอนหลักของการวิจัยมีดังนี้ 1) การสร้างคลังข้อมูลบทความวิจัยสาขาจุลชีววิทยา 2) การวิเคราะห์คลังข้อมูลด้วยอัตถภาควิเคราะห์ และ 3) การวิเคราะห์หาความเที่ยงในการกำหนดอัตถภาค จากนั้นประมวลผลที่ได้จากการทำอัตถภาควิเคราะห์เพื่อกำหนดเป็นโครงสร้างบทความวิจัยทั้งฉบับ ผู้วิจัยวิเคราะห์คลังข้อมูลตามแนวอัตถภาควิเคราะห์ทีละภาคตามลำดับคือ ภาคบทนำ ภาควิธีวิจัย ภาคผลวิจัย และภาคอภิปรายผลวิจัย เพื่อกำหนดว่าแต่ละภาคประกอบด้วยอัตถภาคใดบ้างและแต่ละอัตถภาคประกอบด้วยอนุวัจน์ใดบ้าง งานวิจัยนี้แสดงถึงโครงสร้างต้นแบบที่นักวิทยาศาสตร์ทั่วไป และนักจุลชีววิทยาสามารถใช้เป็นแนวทางประกอบการอ่านและการเขียนบทความเพื่อติดตามความรู้ ความก้าวหน้าทางวิทยาศาสตร์ได้อย่างมีประสิทธิภาพต่อไป

การนำเสนอแนวคิดเกี่ยวกับอัตถภาควิเคราะห์ (move analysis) ของบทความ เพื่อทราบถึงองค์ประกอบต่างๆ และโครงสร้างของบทความ จึงมีผู้วิจัยพยายามนำแนวคิดมาประยุกต์ใช้ในการสร้างงานวิจัยโดยสร้างเป็นโปรแกรมประยุกต์บนเว็บไซต์ เพื่ออำนวยความสะดวกแก่ผู้ใช้งานในด้านต่างๆ โดยมีงานวิจัยที่นำเอาแนวคิดดังกล่าวมาประยุกต์ใช้ ดังต่อนี้

ผู้วิจัยได้ทำงานวิจัยเพื่อนำเสนอซอฟต์แวร์คอมพิวเตอร์ที่เรียกว่า Mover (Anthony & Lashkia, 2003) ที่สามารถช่วยผู้ที่อ่านงานวิจัย ได้เข้าใจในโครงสร้างของเอกสารทางวิชาการที่มีลักษณะเฉพาะทาง โดยการมีระบุโครงสร้างของเอกสารทางวิชาการที่มีความแตกต่างกันในแต่ละด้านให้อย่างอัตโนมัติ โดยตัวระบบนั้นถูกทดสอบด้วยบทคัดย่อ ของบทความทางวิชาการ อีกทั้งระบบยังมีการทำงานที่รวดเร็ว ถูกต้อง และสามารถใช้งานได้จริงในขั้นตอนการอ่าน และการเขียน โดยผู้วิจัยได้นำงานวิจัยของ Swales ที่ศึกษาเกี่ยวกับโครงสร้างของบทความทางวิชาการ หรือการวิเคราะห์มูฟมาใช้ อีกทั้งเพื่อเป็นการแก้ปัญหาของการระบุมูฟ ของโครงสร้างบทความต่างๆ ที่มีโครงสร้างแตกต่างกันออกไป ที่ต้องใช้ผู้เชี่ยวชาญในการศึกษาถึงโครงสร้างของบทความนั้นๆ ดังนั้น ซอฟต์แวร์คอมพิวเตอร์ที่สร้างขึ้นนี้มีการนำเอาวิธีการเรียนรู้ของเครื่อง (Machine Learning) ที่มีลักษณะเป็น Supervised Learning มาใช้ โดยมีหลายหลายอัลกอลิทึมมาใช้ในการทดลอง

การทำงานของซอฟต์แวร์คอมพิวเตอร์นั้น เริ่มต้นจากการให้ระบบนั้นเรียนรู้โครงสร้างของบทความนั้นๆ โดยให้บทความจำนวนหนึ่งแก่ระบบเพื่อเรียนรู้เรียนโครงสร้าง และนำข้อความอีกจำนวนหนึ่งเข้าสู่ระบบ เพื่อให้ระบบทำการระบุโครงสร้างของบทความให้อย่างอัตโนมัติ และแสดง

ผลลัพธ์แก่ผู้ใช้ จากงานวิจัยนี้ให้ความสำคัญกี่ยวกับการระบุโครงสร้างของบทความทางวิชาการสาขาเทคโนโลยีสารสนเทศอย่างอัตโนมัติของระบบ จึงต้องออกแบบระบบการเรียนรู้ ดังนี้

1. ระบุการเรียนรู้ของระบบว่าต้องการเรียนรู้อะไร ผลลัพธ์ที่ได้เป็นอย่างไร
2. เลือก Training data ที่ต้องการให้ระบบเรียนรู้โดยใช้วิธีการเรียนรู้ของเครื่องในลักษณะ Supervised Learning ซึ่งเป็นที่นิยม โดยเลือกบทคัดย่อในสาขาเทคโนโลยีสารสนเทศจำนวน 100 บทความซึ่งนำมาจาก IEEE Transaction on parallel and distributed system ในปี 1998
3. แบ่ง Training data เข้าเป็นกลุ่มและทำการแบ่งมูฟให้เรียบร้อย
4. กำหนดรูปแบบการแสดงออกของความรู้ที่ต้องการที่ได้จากการเรียนรู้ฝึกฝนโมเดล
5. เลือกวิธีการเรียนรู้ หรือโมเดลที่ใช้จำแนกบทคัดย่อที่สามารถให้ผลลัพธ์ได้ตรงกับสมมติฐานที่ตั้งไว้
6. นำเอาโมเดลนั้นไปใช้กับบทคัดย่อที่เป็น Testing data
7. นำลัพธ์จากการวิเคราะห์ของระบบมาเปรียบเทียบกับผลลัพธ์ที่ได้จาการวิเคราะห์ของผู้เชียวชาญที่เป็นมนุษย์

โดยตัวแบบที่ใช้ในการแบ่งโครงสร้างของบทคัดย่อนั้น เรียกว่า CARS (Modified Create a Research Space) ของ Anthony และมีพื้นฐานมาจาก Swales

หลังจากการสร้างระบบเรียบร้อยแล้วจากนั้น ถึงขั้นตอนการทดสอบประสิทธิภาพของระบบกับบทคัดย่อทั่วไป เริ่มต้นการทดสอบจะต้องเตรียมบทคัดย่อที่มีการแบ่งมูฟไว้แล้วจำนวน 2 ชุด โดยชุดแรกจะใช้ในการฝึกฝนระบบจำนวน 554 ชุด และชุดที่สองจะใช้เป็นชุดข้อมูล unseen ใช้สำหรับการประเมินผลระบบจำนวน 138 ชุด โดยความถูกต้องของระบบสามารถวัดได้จากจำนวนที่ทำการแบ่งข้อมูลของชุดข้อมูลทดสอบได้อย่างถูกต้อง ผลลัพธ์ที่ได้เกิดจาการสุ่มการแบ่งข้อมูลต้นฉบับในคลังข้อมูลสู่ Training set และ Testing set จำนวน 5 ครั้ง หรือที่เรียกว่า 5-Fold cross validation และทำการวัดความถูกต้องซ้ำในแต่ละครั้งในการทดสอบ ดังตารางที่ 2.8

ตารางที่ 2.8 แสดงค่าความถูกต้องของระบบโดยใช้วิธี 5-Fold cross validation

Trial Set	Accuracy
1	70%
2	68%
3	65%
4	66%
5	70%
Average	68% $\pm$ 2.2%

รายละเอียดการทดสอบของข้อมูลได้จากตาราง Confusion matrix ซึ่งจะแสดงการแบ่งข้อมูลของระบบได้ จากตาราง Confusion matrix ผลลัพธ์จากการทดสอบครั้งที่ 1 แสดงให้เห็นว่าผลที่ได้นั้นมีทั้งที่มีประสิทธิที่ดี ถึง 92% และประสิทธิภาพต่ำ 17% อันเนื่องมาจากมีข้อมูลที่ใช้ในการ Training มีจำนวนน้อย จากคุณลักษณะสำคัญอย่างหนึ่งของตัวจำแนกประเภทข้อมูล Naïve Bayes คือ ความสามารถในการจัดลำดับการตัดสินใจจากความน่าจะเป็นมาที่สุดไปน้อยที่สุด ดังนั้นผู้วิจัยจึงให้ความสนใจในเรื่องของผลลัพธ์ที่ออกมากจากระบบที่มีมากกว่าหนึ่งตัวเลือกที่ใช้ในการตัดสินใจ ให้ผู้ใช้งานระบบสามารถที่จะเลือกผลลัพธ์ที่คิดว่าถูกต้องมากกว่าได้ให้แก่ระบบซึ่งเป็นการตรวจสอบระบบอีกครั้ง สำหรับระบบนี้ใช้ผลลัพธ์จากการจำแนกข้อมูลสูงสุดสองอันดับ โดยจะทำให้ผลจากการจำแนกประเภทของข้อมูลนั้นดีขึ้น อีกทั้งยังเป็นการเพิ่มความเชื่อมั่นของระบบด้วย ดังตารางที่ 2.9

ตารางที่ 2.9 แสดงประสิทธิภาพของการใช้ผลลัพธ์สองอันดับสูงสุด

Trial set	Accuracy
1	88%
2	86%
3	84%
4	86%
5	86%
Average	86% $\pm$ 1.4%

ผู้วิจัยได้นำเสนอวิธีการวิเคราะห์มูฟ และโครงสร้างของบทคัดย่อในเอกสารทางวิชาการอย่างอัตโนมัติ ด้วยวิธีการทางคอมพิวเตอร์ (Wu, Chang, Liou, & Chang, 2006) โดยเสนอในรูปแบบเครื่องมือที่เป็นซอฟแวร์คอมพิวเตอร์ต้นแบบผ่านทางอินเทอร์เน็ต (Web application) คือ CARE (Concordancer for Academic wRiting in English) ให้ผู้ใช้งานสามารถเรียนรู้ตัวอย่างประโยคที่ถูกกำหนดมูฟโดยระบบที่สร้างขึ้นประกอบด้วย 3 ช่องข้อความ เพื่อให้ผู้เรียนใส่ข้อความภาษาอังกฤษต่อไปนี้ Single word query, Multi-word query เช่น ค้นหาคำว่า “The result show” โดยจะค้นหาบทความที่ประกอบด้วยคำสามคำ และ Corpus selection โดยผู้ใช้สามารถเลือก specific domain ของ corpus ได้ และระบบจะแสดงผลลัพธ์ออกมาซึ่งประกอบด้วยประโยคที่ถูกกำหนดมูฟ ซึ่งแสดงในรูปที่ 2.15

รูปที่ 2.15 แสดงตัวอย่างการค้นหาวลี “the result show”

ประโยคที่อยู่ในบทคัดย่อทั้งหมดจะถูกวิเคราะห์ และถูกแยกออกเป็นมูฟที่แตกต่างกันไป โดยใช้วิธีการที่เกี่ยวข้องกับการรวบรวมบทคัดย่อจำนวนมากจากเว็บไซต์ search engine คือ Citeseer และสร้างโมเดลภาษาของมูฟจากบทคัดย่อดังกล่าว โดยในงานวิจัยนี้ผู้วิจัยไม่ได้กล่าวถึงสาขาของทบความทางวิชาการที่เก็บรวบรวมเพื่อสร้าง corpus ซึ่งซอฟแวร์ดังกล่าวจะทำให้เกิดแนวโน้มในการศึกษาด้านการเขียนผ่านเว็บไซต์โดยมีคอมพิวเตอร์เป็นเครื่องมือช่วยเหลือ

จากงานวิจัยนี้ระบบจะทำการวิเคราะห์มูฟของบทคัดย่อ ประกอบไปด้วย Background, Purpose, Method, Result และ Conclusion เริ่มต้นการเรียนรู้โครงสร้างของบทคัดย่อด้วยการเตรียมบทคัดย่อ ดังนี้

1. เก็บรวบรวมบทคัดย่อจากเว็บไซต์ใช้สำหรับการฝึกฝนโมเดลด้วยวิธีอัตโนมัติจาก search engine เพื่อสร้าง corpus A ของบทคัดย่อโดยมีผู้เชี่ยวชาญทางด้านวิทยาการคอมพิวเตอร์ทำการ tagged บทคัดย่อ

2. ทำการจัดกลุ่ม (Label) ให้กับประโยคแต่ละประโยคด้วยตัวผู้วิจัยเองให้อยู่ในบทคัดย่อกลุ่มเล็กๆ
3. สะกัดคำที่เกิดขึ้นร่วมกัน (Collocation) จากบทคัดย่อทั้งหมดด้วยวิธีอัตโนมัติ เพื่อทำให้ move-tagged-collocation เกิดความสมดุลและกระจายตัว โดยใช้ one-move-per-collocation โดยมีขั้นตอนดังต่อไปนี้

   Step 1: คำที่เกิดขึ้นร่วมกัน เช่น “paper address” ซึ่งสะกัดจาก corpus A โดยกำหนด label ด้วยมูฟ “P” จากนั้นจะใช้ “P” ในการกำหนด label ให้กับประโยคอื่นๆ (untagged sentences: US) ประโยคที่ถูกกำหนด label แล้วจะถูกเรียกว่า tagged sentences: TS

   (1) This **paper addresses** the state explosion problem in automata based ltl model checking. //P//

   (2) This **paper addresses** the problem of fitting mixture densities to multivariate binned and truncated data. //P//

   Step 2: หลังจากนั้นจะทำการหาคุณลักษณะของคำที่เกิดขึ้นร่วมกันอื่นๆ เช่น “address problem” ซึ่งได้มาจาก tagged sentences ใน corpus A เพื่อใช้ในกำหนด label ให้กับประโยค untagged sentences อื่นๆ เช่น

   (3) This paper **addresses** the state explosion **problem** in automata based ltl model checking.

   (4) This paper **addresses** the **problem** of fitting mixture densities to multivariate binned and truncated data.

   Step 3: ต่อจากนั้นคุณลักษณะ “address problem” สามารถที่จะนำไปใช้ประโยชน์ในการ tagged ประโยคด้วยมูฟ “P” แต่จะไม่รวมถึง collocation “paper address” ดังนั้นจึงค่อยๆ ขยายขอบเขตของคำอธิบาย ตัวอย่างเช่น

   (5) In this paper we **address** the **problem** of query answering using views for non-recursive data log queries embedded in a Description Logics knowledge base. //P//

   (6) We **address** the **problem** of learning robust plans for robot navigation by observing particular robot behaviors. //P//

จากตัวอย่างที่ 5 และ 6 เราสามารถที่จะกำหนดมูฟ “P” ให้กับคุณลักษณะอื่นๆได้ เช่น “we address” โดยขั้นตอนสามารถที่จะทำซ้ำในขณะที่ไม่มีคุณลักษณะใหม่ที่มีความถี่มากพอซึ่งแสดงในตารางที่ 2.10

ตารางที่ 2.10 แสดงตัวอย่างของ collocation ที่เพิ่มเติม

Type	Collocation	Move	Count of collocation with m_i	Total of collocation occurrences
NV	We present	P	3,441	3,668
NV	We show	R	1,985	2,069
NV	We propose	P	1,722	1,787
NV	We describe	P	1,505	1,583
...	...	...	...	...

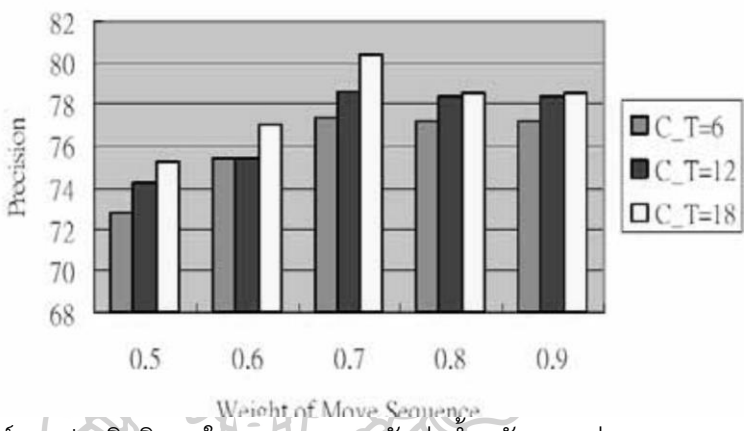
4. ทำการกำหนด (Label) มูฟให้แต่ละคำที่เกิดร่วมกันที่แตกต่างกันด้วยตัวผู้วิจัยเอง
5. ทำการขยายคำที่เกิดขึ้นร่วมกันที่แสดงในแต่ละมูฟด้วยวิธีอัตโนมัติ
6. สร้าง Markov model เพื่อใช้สำหรับการจำแนกมูฟของบทคัดย่อ โดยในช่วงเริ่มต้นการฝึกฝนตัวโมเดลเราจะใช้ประโยคที่มนุษย์ทำการกำหนดมูฟมาเป็นชุดข้อมูลฝึกฝนตัวโมเดล

บทคัดย่อที่นำมาใช้เป็น Training data ได้รับมาจาก search engine คือ Citeseer โดยในงานวิจัยนี้ผู้วิจัยไม่ได้กล่าวถึงสาขาของทบความทางวิชาการที่เก็บรวบรวมมาสร้าง corpus ซึ่ง Corpus มีขนาด 20,306 บทคัดย่อ (95,960 ประโยค) โดยมีบทคัดย่อจำนวน 106 บทคัดย่อ ที่ประกอบขึ้นจาก 709 ประโยคได้ถูกกำหนดมูฟจากผู้เชี่ยวชาญ 4 ท่าน จากนั้นทำการสกัดคำที่เกิดขึ้นร่วมกัน (collocation) จำนวน 72,708 ชนิด และทำการกำหนดมูฟให้กับคำที่เกิดขึ้นร่วมกันเหล่านั้น ในขณะเดียวกันเราได้ทำการสะกัดเอาคำที่เกิดขึ้นร่วมกัน (collocation) จำนวน 72,708 ชนิด และคำการกำหนดมูฟให้กับคำที่เกิดขึ้นร่วมกันจำนวน 317 collocations

ในขณะที่โปรแกรมทำงานใช้บทคัดย่อจำนวน 115 บทคัดย่อ ที่ประกอบไปด้วย 684 ประโยคเป็นตัวฝึกฝนตัวโมเดลของโมเดลจำแนกประเภทของข้อมูล จากนั้นจะใช้ HMM ที่ได้ฝึกฝนแล้วมาทำการทดลองกับค่าพารามิเตอร์ที่แตกต่างกันไป คือ ความถี่ของชนิด collocation จำนวนประโยคที่กับ collocation ในแต่ละบทคัดย่อ และลำดับคะแนนของมูฟและ collocation

การทำให้ประสิทธิภาพของ HMM ที่ทำการกำหนดมูฟอย่างอัตมัติให้มีประสิทธิภาพที่ดีภายใต้การกำหนดต่าพารามิเตอร์ที่แตกต่างกัน เช่น ค่าน้ำหนักของฟังก์ชันความน่าจะเป็น จำนวนประโยคในบทคัดย่อ และจำนวนที่น้อยที่สุดของตัวอย่าง collocation จากการกำหนดพารามิเตอร์ต่างๆ ทำให้ได้ผลของค่า precision ที่ดีที่สุด 80.54% เมื่อมีจำนวนประโยค 627

ประโยคที่มีคุณสมบัติกับชุดของพารามิเตอร์ที่หลากหลาย ประกอบไปด้วย ค่าน้ำหนักของฟังก์ชันความน่าจะเป็น คือ 0.7, ความถี่ของค่า threshold 18 ของ collocation ที่เหมาะสม และสองประโยคที่น้อยที่สุดที่ประกอบไปด้วย collocation โดยผลลัพธ์ของประสิทธิภาพในการ tagged กับค่าน้ำหนัก และค่า threshold ที่แตกต่างแสดงได้ในรูปที่ 2.16



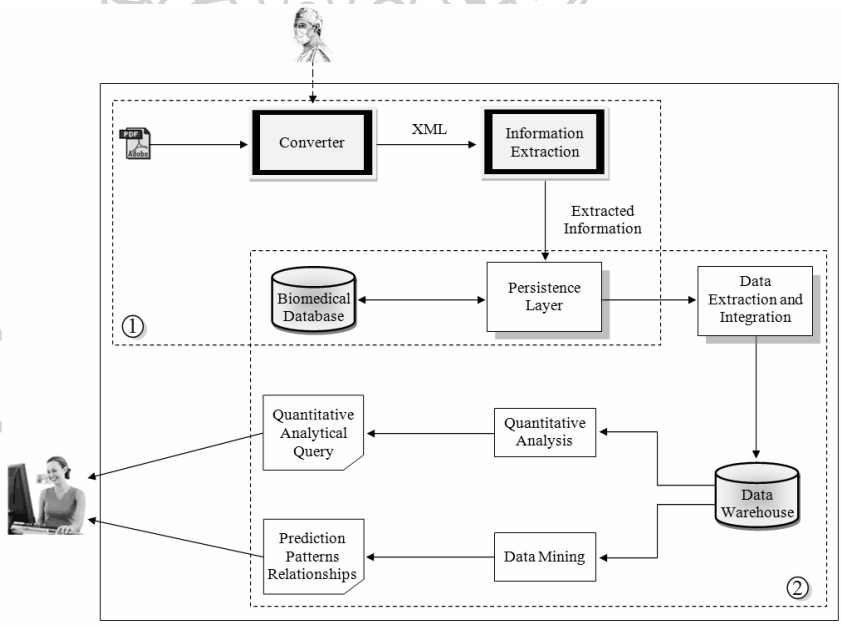
รูปที่ 2.16 แสดงผลลัพธ์ของประสิทธิภาพในการ tagged กับค่าน้ำหนัก และค่า threshold ที่แตกต่าง

จากงานวิจัยที่กล่าวมาข้างต้นจะพบว่า ผู้วิจัยได้นำแนวคิดเกี่ยวกับอัตถภาควิเคราะห์ (move analysis) ของบทความ นำมาใช้วิเคราะห์โครงสร้างของบทคัดย่อของบทความทางวิชาการ ทั้งสองงานวิจัยเพื่อใช้ในการศึกษาและเรียนรู้ของผู้ใช้งาน แสดงให้เห็นถึงโครงสร้างของบทคัดย่อในความทางวิชาการว่ามีโครงสร้างที่แบบแผน แต่ในความเป็นจริงมีบทความต่างๆ มากมายที่ปรากฏขึ้น จึงมีผู้วิจัยเล็งเห็นถึงการนำเอาแนวคิดในการวิเราะห์โครงสร้างประโยคข้างต้นนี้มาใช้ในการวิเคราะห์โครงสร้างของบทความชนิดอื่นๆ นอกเหนือจากบทคัดย่อของบทความทางวิชาการ ซึ่งแสดงให้เห็นได้จากงานวิจัยดังต่อไปนี้

อีกทั้งยังมีผู้วิจัยได้ศึกษาเกี่ยวกับปัญหาการประมวลผล และการสกัดข้อมูลจากเอกสารอิเล็กทรอนิกส์ที่มีลักษณะไม่มีโครงสร้างเป็นรูปแบบข้อความธรรมดา ในสาขาที่เกี่ยวข้องกับชีวการแพทย์ (Matos et al., 2012) ที่มีอยู่เป็นจำนวนมากรายงานเกี่ยวกับชนิดของผู้ป่วย อธิบายการรักษา จำนวนผู้ป่วยที่เข้ารับการรักษา อาการป่วยและปัจจัยเสี่ยงที่มีความเกี่ยวข้องกับโรค และผลการรักษาที่เป็นด้านลบ และด้านบวกของสุขภาพผู้ป่วยจะอยู่ในรูปแบบเอกสารจะเป็นเอกสารทางวิชาการด้านวิทยาศาสตร์ ที่ปรากฏอยู่ในวารสารทางวิชาการ เช่น American Journal of

Hematology, Blood และ Haematologica โดยนำเสนอในรูปแบบของระบบสนับสนุนการตัดสินใจเพื่อช่วยผู้เชี่ยวชาญในการตัดสินใจเรียกว่า IEDSS-Bio (An environment for Information Extraction and Decision Support System in Biomedical domain) จากกรณีศึกษานี้นำเอาวิธีการเรียนรู้ของเครื่อง (machine learning) เข้ามาทำหน้าที่ในการระบุชนิดของข้อความในแต่ละย่อหน้าว่าเป็นข้อความที่เกี่ยวข้องกับเรื่องใด เช่น อาการป่วย ผลกระทบจากการรักษา และจำนวนข้อมูลของผู้ป่วยที่ปรากฏอยู่ในบทความทางวิชาการเรื่อง Sickle Cell Anemia papers

งานวิจัยนี้มุ่งศึกษาประเด็นของกระบวนการสกัดข้อมูล และการระบุชนิดของข้อความแต่ละย่อหน้าหรือประโยค โดยวิธีการสกัดข้อความที่ไม่มีโครงสร้างนั้นจะนำวิธีการ Text mining มาใช้ในการวิเคราะห์ข้อมูลและองค์ประกอบของระบบ IEDSS-Bio ซึ่งแสดงอยู่ในรูปที่ 2.17 นั้นแบ่งออกเป็น 2 องค์ประกอบด้วยกัน คือ



รูปที่ 2.17 แสดงสภาพแวดล้อมของการวิเคราะห์ข้อมูล

1. Conversion and Extraction component ส่วนมีจุดมุ่งหมายที่จะแปลงเอกสารที่มีรูปแบบและชนิดที่แตกต่างกันให้อยู่ในรูปแบบเดียวกัน และสกัดข้อมูลที่ถูกต้องตรงตามความต้องการ จากเอกสารทางวิชาการด้านวิทยาศาสตร์ที่อยู่ในฐานข้อมูลชีวการแพทย์ จากเอกสารบทความทางวิชาการส่วนมากจะอยู่ในรูปแบบไฟล์ PDF จะถูก Convert module จะทำการแปลงไฟล์ PDF ให้อยู่ในไฟล์รูปแบบXML โดยข้อมูลภายในไฟล์นั้นจะเป็นลักษณะเป็นลำดับชั้นดังต่อไปนี้ เช่น section > page > paragraph > sentence เพื่อทำให้การสกัดนั้นง่ายขึ้น และผู้ใช้ยัง

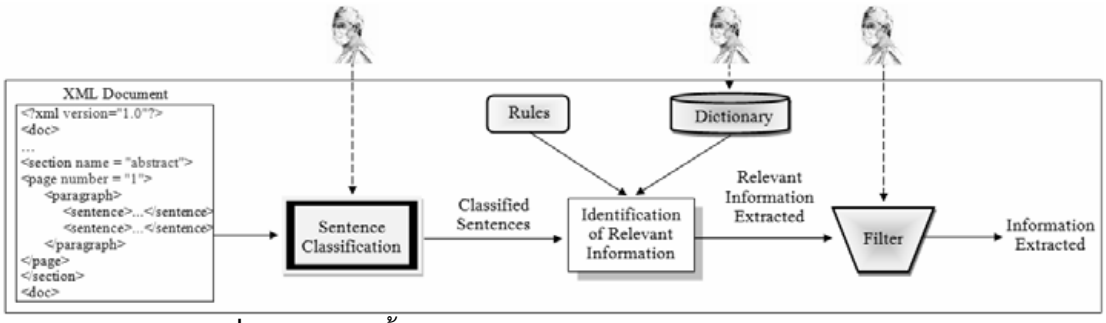
สามารถทำการตรวจสอบคุณภาพเอาสารที่สร้างขึ้นได้อีกด้วย ต่อมาจะเข้าสู่กระบวนการ Information extraction module โดยประมวลผลเอกสาร XML เพื่อสกัดข้อมูลที่ต้องการจากประโยค จากนั้นข้อมูลที่สกัดมาจะถูกเก็บในฐานข้อมูล biomedical โดย Persistence Layer module และส่วน Data Extraction and Integration (DEI) module จะทำการสกัดข้อมูลจากฐานข้อมูล biomedical ผ่าน Persistence Layer module เป็นเก็บไว้ในส่วน Data warehouse

2. Data Analysis component คือ ส่วนสำคัญในการระบุรูปแบบของเอกสารทางวิชาการชีวการแพทย์ โดยการนำเอา Data warehouse และเทคนิค Data mining มาใช้ โดยข้อมูลที่เก็บใน Data warehouse นั้นถูกนำมาใช้ในขั้นตอนการประมวลผลข้อมูล เพื่อช่วยในการสนับสนุนการวิเคราะห์ข้อมูล ดังต่อไปนี้

1) นำข้อมูลไปใช้สำหรับประมวลผลเพื่อเป็นที่ปรึกษาทางด้านการวิเคราะห์เชิงปริมาณ เช่น การตอบคำถามที่เป็นไปได้ “How many patients had clinical improvement and were treated with the hydroxyurea drug?”

2) นำข้อมูลมาใช้ในการประมวลผลด้วยวิธีการ Data mining เพื่อการระบุรูปแบบของข้อมูลที่สนใจ เช่น ต้องการค้นหารูปแบบที่เป็นไปได้ “A significant amount of patients under treatment with the hydroxyurea drug tend to have marrow depression.”

จากองค์ประกอบของระบบ IEDSS-Bio ทั้งสององค์ประกอบที่กล่าวไปข้างต้น ภายในองค์ประกอบที่ 1 มีส่วนที่เรียกว่า Information Extraction module ซึ่งใช้สำหรับการสกัดข้อมูลในเรื่องของชีวการแพทย์ (Biomedical domain) โดยวิธีการสกัดข้อมูลทางด้านชีวการแพทย์ดังแสดงในรูปที่ 2.18 นั้นเกิดการรวมกันของ 3 วิธีการด้วยกัน คือ

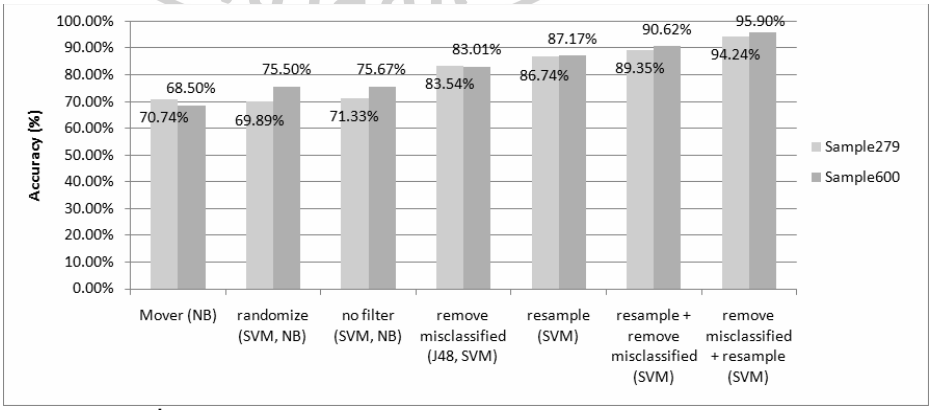


รูปที่ 2.18 แสดงขั้นตอนของ Information extraction module

1. Machine Learning ใช้สำหรับการจำแนกประโยค
2. Rule-based เป็นวิธีที่ใช้สำหรับการระบุรูปแบบของข้อความโดยใช้วิธีการ regular expression (rule) ใช้ในการสกัดข้อมูลจากประโยคที่ถูกจำแนกประเภทแล้วในขั้นตอนที่ผ่านมา
3. Dictionary based เป็นการใช้ดิกชันนารี่ที่ประกอบด้วยคำศัพท์ต่างๆ ที่เกี่ยวข้องกับทางวิจัยทางด้านวิทยาศาสตร์ ที่จะเป็นผู้ช่วยในการสร้างรูปแบบของ regular expression เก็บข้อมูลที่มีความเกี่ยวข้องกับเรื่องของชีวการแพทย์ ช่วยในการสร้างกฎ และใช้ในการเพิ่มค่าความถูกต้อง คือ Precision และค่า Recall ในการระบุรูปแบบของข้อมูล โดยผู้เชี่ยวชาญอาจมีการประเมินความถูกต้องของข้อมูลที่ถูกสกัดออกมาแล้วในขั้นตอนของการ filter แล้วเก็บลงในฐานข้อมูล อีกทั้งยังสามารถประเมินความถูกต้องในการจัดกลุ่มของประโยค และเพิ่มข้อมูลลงในดิกชันนารีได้อีกด้วย

การทดลองนั้นจะใช้วิธีการเรียนรู้ของเครื่องแบบ classification จำนวน 6 ชนิดดังนี้ Support Vector Machine (SVM), Naïve Bayes (NB), ID3, J48, Prism และ OneR ซึ่งจะนำผลลัพธ์ของแต่ละวิธีมาเปรียบเทียบกันในทุกๆ วิธีการจำแนกกลุ่มใช้ซอฟต์แวร์ Weka เป็นเครื่องมือในการทำเหมืองข้อมูล วิธีการทดลองการจำแนกกลุ่มของประโยคจะใช้ข้อมูลที่สนใจ คือ effect และ patients เลือกจำนวนตัวอย่างละชุด คือ Sample279 และ Sample600 สำหรับประโยค effect และ Sample204 และ Sample659 สำหรับประโยค patients

ผลลัพธ์ของการวัดค่าความถูกต้องของการจำแนกประเภทของข้อมูลโดยใช้การเรียนรู้ของเครื่องทั้ง 6 ชนิด โดยใช้วิธีการ 10-fold cross-validation แสดงได้จากรูปที่ 2.19



รูปที่ 2.19 แสดงผลลัพธ์จากการวัดความถูกต้องใน effect class

จากผลลัพธ์แสดงให้เห็นว่าวิธีการจำแนกข้อมูลโดยการใช้ SVM ปกติให้ผลลัพธ์ที่ดีกว่าวิธีการอื่นๆ โดยผลลัพธ์ที่ดีที่สุดเป็นผลของการใช้ filter Remove Misclassification (RM) และจากการทดลองการจำแนกประเภทของประโยค ซึ่งให้ผลลพธ์สูงถึง 95.9% สำหรับเอกสารทางวิชาการเกี่ยวกับเรื่อง Sickle Cell Anemia

จากงานวิจัยที่ได้ศึกษาผ่านมาข้างต้น จะพบว่าการวิเคราะห์โครงสร้างของบทความเพื่อทราบถึงลักษณะต่างๆ ของบทความนั้น จำเป็นที่ต้องใช้วิธีการเรียนรู้ของเครื่อง (Machine learning) ในการฝึกฝน และเรียนรู้โครงสร้างของบทความเพื่อที่จะทำนายและบ่งบอกโครงสร้างของบทความอื่นได้ วิธีการเรียนรู้ของเครื่องนั้นมีอยู่หลากหลายชนิด เช่น Support Vector Machine, Decision tree เป็นต้น โดยแต่ละวิธีการเรียนรู้นั้นจะให้ประสิทธิภาพในการเรียนรู้ที่แตกต่างกันไป โดยทั้งนี้ประสิทธิภาพขึ้นอยู่กับหลายปัจจัยด้วยกัน เช่น ลักษณะของข้อมูลที่จำแนก จำนวนคลาสที่จำแนก เป็นต้น ทั้งนี้ยังมีวิธีการเรียนรู้วิธีการอื่นที่เป็นที่นิยม และยังให้ประสิทธิภาพในการจำแนกประเภทของข้อมูลที่ดีกว่า หนึ่งในนั้นคือ การเรียนรู้ของเครื่องที่มีชื่อว่า Conditional Random Fields (CRFs) จึงมีงานวิจัยจำนวนมากได้นำ CRFs มาใช้ภายในงานวิจัย และให้ผลลัพธ์จากการจำแนกประเภทข้อมูลที่ดี ดังนั้นจึงได้นำเสนองานวิจัยที่ใช้ CRFs ดังต่อไปนี้

จากการศึกษาพบว่ามีการใช้ใช้งานเว็บ search engine ในการค้นหาเอกสารทางวิชาการเพิ่มมากขึ้น (Peng & McCallum, 2006) ผู้วิจัยถึงมองเห็นถึงคุณภาพของเอกสารทางวิชาการที่ถูกค้นคือมาได้นั้น จะต้องมีความถูกต้องตรงกันนั้นมีความสำคัญจึงทำการวิจัยเพื่อนำเสนอวิธีการสกัด meta-data ของเอกสารทางวิชาการ เช่น title, author, institution เป็นต้น โดยใช้วิธีการเรียนรู้ของเครื่อง คือ Conditional Random Field (CRFs) และทำการค้นหาวิธีการปฏิบัติที่หลากหลายที่จะนำเอา CRFs มาใช้กับการสกัดสารสนเทศทั่วไป วิธีการของ CRFs นั้นจะนำเอาข้อดีของ Hidden Markov models (HMM) และการจำแนกของ Support Vector Machine (SVM) การทดลองนั้นใช้ชุดข้อมูลเป็นบทความทางวิชาการทั้งหมดจำนวน 2 ชุด ซึ่งประกอบไปด้วยส่วนหัวเรื่องของบทความทางวิชาการ และส่วนของการอ้างอิง โดยข้อมูลทั้งหมดนี้เป็นข้อมูลพื้นฐานที่ใช้เป็นตัวกลางในการทดสอบจากหลายๆ จากวิจัยก่อนหน้านี้

การประเมินประสิทธิภาพของการสกัดข้อมูลนั้น มีวิธีการดังนี้

1. Measuring field-specific performance: คือการวัดความถูกต้องแบบเฉพาะเจาะจงในแต่ละสาขา ซึ่งประกอบไปด้วย

Word Accuracy ทำการกำหนดให้ A เป็นจำนวนคำของ true positive, B เป็นจำนวนคำของ false negative, C เป็นจำนวนคำของ false positive, D เป็นจำนวนคำของ true negative

และ A+B+C+D เป็นจำนวนของคำทั้งหมด ดังนั้นความถูกต้องของคำทั้งหมดคิดได้จาก $\frac{A+D}{A+B+C+D}$

F1-measure คำนวณได้จากค่า $Precision = \frac{A}{A+C}$, ค่า $Recall = \frac{A}{A+B}$ ค่า $F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$

2. Measuring overall performance: คือการวัดความถูกต้องโดยรวมทั้งหมดซึ่งประกอบไปด้วย

Overall word accuracy performance คือ เปอร์เซ็นต์ความถูกต้องของคำที่ทำนายเท่ากับคำที่คำที่ถูกกำหนดไว้ถูกต้องแล้ว โดยความถูกต้องของคำในแต่ละสาขากับจำนวนของคำทั้งหมด เช่น บทคัดย่อ

Averaged F-measure เป็นค่าที่คิดได้จากการเฉลี่ยของค่า F1-measure ทั้งหมดของบทความทุกสาขา Averaged F-measure กับจำนวนของคำที่เล็กน้อยที่เป็นองค์ประกอบของความถูกต้องของคำ

จากผลลัพธ์ที่ได้จากการทดลองได้นำเสนอผลการทดลองโดยรวม โดยการเปรียบเทียบระหว่าง CRFs, HMMs และ SVMs โดยใช้ชุดข้อมูล H แสดงให้เห็นว่า CRFs นั้นได้แสดงประสิทธิภาพที่ดีกว่า HMMs อีกทั้งยังให้ผลลัพธ์ที่ดีกว่าวิธีการ SVM อีกด้วย อีกทั้งยังสามารถทำงานได้อย่างมีประสิทธิภาพในบทความทุกๆ สาขา

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou and Mitsuru Ishizuka (2008) ได้ศึกษาเกี่ยวกับโครงสร้างของบทคัดย่อของบทความทางวิชาการ ซึ่งจะต้องใช้วิธีการของ Text mining เพื่อที่จะนำมาประยุกต์ใช้ในงานการสกัดสารสนเทศ การค้นคืนข้อมูล และการสรุปใจความสำคัญของข้อมูล เป็นต้น โดยผู้วิจัยได้นำเสนอวิธีการจัดหมวดหมู่ของประโยคในบทคัดย่อทางด้านวิทยาศาสตร์ให้อยู่ใน 4 องค์ประกอบด้วยกันคือ objective, method, result และ conclusion วิธีการการที่นำมาใช้ในการจำแนกประเภทของประโยค คือ การนำวิธีการเรียนรู้ของเครื่องแบบ supervised learning มาใช้

วิธีการจำแนกประเภทของประโยคนั้นใช้วิธีการแก้ปัญหาแบบ sequential labeling โดยการนำเอา Conditional Random Fields (CRFs) เพื่อกำหนดชนิดของประโยค (labeling) เพื่อจัดจำแนกหมวดหมู่นำมาใช้ในการระบุประโยคต่างๆ ของบทคัดย่อ ซึ่ง Corpus ที่นำมาใช้ในการ training model นั้นเป็นบทคัดย่อทางด้านการแพทย์ อีกทั้งงานวิจัยนี้ได้ทำการออกแบบคุณลักษณะไว้ 3 คุณลักษณะที่นำเสนอในแต่ละประโยคของบทคัดย่อสำหรับ CRFs โดยให้

y_i คือ label ที่ต้องการกำหนด

X_i คือ ประโยคที่ถูก label y_i กำหนดให้

อีกทั้งยังเพิ่มประโยครอบข้างคือ X_{i-1} , X_{i+1} และ unigram ในประโยค X_{i-1}

Content (n-grams) คือ คุณลักษณะนี้วิเคราะห์ลักษณะเฉพาะของรูปแบบประโยค เช่น to determine ...," และ "aim at ..." เป็นต้น ซึ่งใช้สำหรับการเริ่มต้นวัตถุประสงค์ของการศึกษางานวิจัย ดังนั้นเราจึงใช้คุณลักษณะสำหรับประโยคที่นำเสนอโดย i) words, ii) word bigrams และ iii) ซึ่งเป็นส่วนผสมของคำ และ bigrams คำจะถูกทำให้อยู่ในรูปแบบดั้งเดิมโดย GENIA tagger ซึ่งก็คือ part-of-speech tagger ที่ใช้สำหรับในสาขาชีวการแพทย์

Relative sentence location คือ การอธิบายถึงตำแหน่งของประโยคของบทคัดย่อ ซึ่งประโยคที่แสดงถึงวัถุประสงค์ของการศึกษาจะเขียนเริ่มต้นของย่อหน้า และส่วนสุดท้ายของย่อหน้าจะเป็นข้อสรุปของงานวิจัย โดยตำแหน่งของประโยคจะเป็นเบาะแสที่ดีในการบ่งบอกตำแหน่งของประโยคในบทคัดย่อผู้วิจัยจึงทำการกำหนด 5 คุณลักษณะในการกำหนดความสัมพันธ์ของตำแหน่งของประโยคใน 5 สเกลด้วยกัน

Features from previous/next sentences คือ การให้คุณสมบัติแก่ประโยคที่อยู่ด้านหน้า และด้านหลัง

จากการทดลองด้วยวิธีดังที่กล่าวไปข้างต้น ได้ให้ค่าความถูกต้องจากการจำแนกประเภทในระดับประโยค 95.5% และให้คาความถูกต้องจากการจำแนกประเภทในระดับบทคัดย่อ 68.8% ซึ่ง CRFs นั้นเหมาะสมต่อการวิเคราะห์มูฟของบทคัดย่อทางด้านวิทยาศาสตร์และยังให้ค่าความถูกต้องในการวิเคราะห์ระดับบทคัดย่อได้ดีกว่า SVM

มีผู้วิจัยได้ทำการศึกษางานวิจัยต่างๆ พบว่าโครงสร้าง หรือรูปแบบของย่อหน้าของบทความที่เขียนขึ้นส่วนมากเริ่มต้นจากประโยคที่บ่งบอกถึง background ซึ่งประกอบไปด้วยข้อมูลพื้นฐาน หรือเบื้องหลังของงานวิจัย (M.A., Cranefield, & Stanger, 2010) และแนะนำงานวิจัยที่เกี่ยวข้องโดยใช้วิธีการอ้างอิงเป็นตัวบ่งบอกถึงลักษณะเฉพาะของงานวิจัยดังกล่าวตามที่ผู้ทำงานวิจัยสนใจ ในประโยคที่อยู่สุดท้ายอธิบายหรืออ้างอิงถึงผลลัพธ์ของการทดลอง จึงนำเสนอวิธีการระบุบริบทต่างๆ ของประโยคที่มีความเกี่ยวข้องกันในบทความทางวิชาการ ซึ่งการระบุบริบทต่างๆ ของประโยคนั้นเป็นปัญหาแบบ sequential classification problem โดยในงานวิจัยนี้ได้นำเอาวิธีการ Conditional Random Fields (CRFs) มาใช้ในการจำแนกประโยคให้อยู่ในหมวดหมู่ที่ได้กำหนดไว้ และนำเสนอถึงความสำคัญของการอ้างอิง ซึ่งเป็นคุณลักษณะที่สำคัญที่ใช้ในการจำแนกประโยค

จากการศึกษารูปแบบของโครงสร้างของบทคัดย่อจากงานวิจัย ทางผู้วิจัยจึงทำการกำหนดแบบแผนสำหรับการจำแนกกลุ่มขึ้นมา โดยจำแนกออกเป็น 10 คุณลักษณะดังต่อไปนี้

1. Background Terms (BGT) คือ การอธิบายองค์ประกอบที่เป็นเบื้องหลังของงานวิจัย เช่น Several researchers และ Recent studies เป็นต้น

2. Subject of Inquiry Terms (SOI) คือ บอกถึงคำกริยาเพื่ออธิบายอะไรบางอย่าง เช่น examine, propose และ demonstrate เป็นต้น
3. Outcome Terms (OCT) คือ บอกถึงคำกริยาที่อ้างถึงผลลัพธ์ เช่น develop และ show เป็นต้น
4. Strength Terms (STH) คือ บอกถึงคำกริยาที่แสดงถึงจุดแข็งของงาน เช่น Improve เป็นต้น
5. Shortcomings Terms (SCT) คือ บอกถึงช่องว่างของงานที่เกี่ยวข้อง เช่น Notwithstanding, does not และ Despite this improvement เป็นต้น
6. Subjective Pronouns (SPN) คือ คำที่บ่งบอกถึงบุคคลของงานที่เกี่ยวข้อง เช่น They, These และ The authors เป็นต้น
7. Words of Stress (WOS) คือ เป็นการแสดงถึงการเน้น เช่น Hence, Furthermore, Therefore และ In addition เป็นต้น
8. Alternate Approach Terms (AAT) คือ บ่งบอกถึงวิธีการทางเลือกอื่นๆ เช่น different, instead และ alternative เป็นต้น
9. Result Terms (RES) คือ การอ้างอิงถึงผลลัพธ์ของงานปัจจุบัน เช่น We show และ our work shows เป็นต้น
10. Contrasting Terms (CON) คือ คำที่ใช้สำหรับการสร้างความแตกต่างให้กับประโยค เช่น In Contrast และ differs from that of เป็นต้น

จากการจำแนกกลุ่มของประโยคข้างต้น ผู้วิจัยได้ทำการกำหนด simple features และ citation feature สำหรับแต่ละประโยค ดังนี้

Citation features

1. sentHasCitation: ประโยคที่มีการอ้างอิง
2. prevSentHasCitation: ประโยคก่อนหน้าที่มีการอ้างอิง

Sentence Features

3. sentHasTerm=BGT: ประโยคที่ประกอบด้วย Background
4. sentHasTerm=OCT: ประโยคที่แสดงผลลัพธ์
5. sentHasTerm=STH: ประโยคที่แสดงจุดแข็ง
6. sentHasTerm=SCT: ประโยคที่มีข้อบกพร่อง
7. sentHasTerm=SPN: ประโยคที่มี subjective pronoun
8. sentHasTerm=WOS: ประโยคที่มีคำพูดที่มีการเน้นหนัก
9. sentHasTerm=AAT: ประโยคที่มีวิธีการทางเลือกอื่นๆ

10. sentHasTerm=RES: ประโยคที่อ้างอิงถึงผลลัพธ์ของงานปัจจุบัน

11. sentHasTerm=CON: ประโยคที่แสดงถึงความขัดแย้ง

Compound features

12. sentHasTerm=SOI_CIT: ประโยคที่มีการอ้างอิงและรายละเอียดเพิ่มเติม

13. sentHasTerm=SOI_NO_CIT: ประโยคที่ไม่มีการอ้างอิงและรายละเอียดเพิ่มเติม

ในการทดลองนั้นในนำบทความทางวิชาการจำนวน 50 บทความทางวิชาการ ทำการสุ่มเลือกมาจากฐานข้อมูล Lecture Notes in Computer Science (LNCS) นำมาจาก www.springerlink.com โดยเป็น Training set ทั้งหมด 50 บทความทางวิชาการประกอบไปด้วย 200 ย่อหน้า แต่ละย่อหน้าประกอบไปด้วยประโยคมากกว่า 1,063 ประโยค แต่ละย่อหน้านำเสนอลำดับของคุณลักษณะของแต่ละประโยค ชุดข้อมูลที่แตกต่างกัน 2 ชุด เตรียมจาก training set ชุดข้อมูลแรกสร้างขึ้นจากคุณลักษณะของประโยค ข้อมูลอีกชุดใช้ทั้งคุณลักษณะของการอ้างอิง และคุณลักษณะของประโยค

กระบวนการระบุและกำหนดประโยคอย่างอัตโนมัติ นั้นสร้างขึ้นด้วยภาษา Python เริ่มต้น

ขั้นตอนที่ 1 คือ ขั้นตอนการเตรียมบทความทางวิชาการ

- แปลงบทความทางวิชาการที่อยู่ในรูปแบบของ PDF จะถูกแปลงเป็นไฟล์ text
- เข้าสู่กระบวนการ cleansing data ทำการลบส่วนอื่นที่ไม่เกี่ยวข้อง และเก็บส่วนที่เกี่ยวข้องเอาไว้ ตัดคำและส่วนที่ไม่สื่อความหมายออก ลบช่องว่างออก

ขั้นตอนที่ 2 คือ ตัดประโยค

- ใช้ sentence tokenizer จาก Natural Language Toolkit (NLTK) ในการตัดแบ่งประโยค

ขั้นตอนที่ 3 คือ กำหนด regular expressions สำหรับการระบุคุณลักษณะที่ต้องการตามที่ได้จำแนกไว้เป็นหมวดหมู่ 10 หวดหมู่ข้างต้น

ขั้นตอนที่ 4 คือ ระบุคุณลักษณะให้กับประโยค ประโยคที่ตัดและในขั้นตอนที่ 2 จับคู่ได้กับ regular expression ที่กำหนดไว้ในขั้นตอนที่ 3

ขั้นตอนที่ 5 คือ ระบุคุณลักษณะในประโยคก่อนหน้าโดยการตรวจสอบการอ้างอิงในประโยคก่อนหน้า และเพิ่มคุณลักษณะ prevSentHasCitation

ผลลัพธ์ที่ได้แสดงได้จากค่า Precision, Recall และ F-score ของแต่ละการจำแนกประเภทแต่ละชนิด การจำแนกข้อมูลให้ค่าความถูกต้องสูงถึง 96.51% เมื่อฝึกฝนโมเดลด้วยสองคุณลักษณะ คือ คุณลักษณะอ้างอิง และคุณลักษณะของประโยค ถ้าให้การฝึกฝนด้วยคุณลักษณะเพียงอย่างเดียวจะให้ค่าประสิทธิภาพที่ต่ำ คือ 93.22%

ผู้วิจัยได้นำเสนอวิธีการที่จะอำนวยความสะดวกในการเข้าถึงบทความทางวิชาการอย่างอัตโนมัติ ด้วยวิธีการจำแนกบทประโยคเข้าสู่หมวดหมู่ทั้ง 11 หมวดหมู่ ในงานวิจัยนี้เรียกเครื่องมือนี้ว่า Core Scientific Concepts (CoreSCs) (Liakata, Saha, Dobnik, Batchelor, & Rebholz-Schuhmann, 2012) ซึ่งประกอบไปด้วย Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result และ Conclusion โดย CoreSCs ให้โครงสร้าง และบริบทต่างๆของบทความ และความสัมพันธ์กับบทความ การรู้จำโดยอัตโนมัติสามารถอำนวยความสะดวกในการสกัดข้อมูลชีวการแพทย์ จากลักษณะที่แตกต่างกันของข้อเท็จจริง สมมติฐานและหลักฐาน ที่ปรากฏอยู่ในเอกสารทางวิชาการที่เผยแพร่อยู่

วิธีการทดลองเริ่มต้นจากการนำเอา Training data และ Testing data นำเข้าสู่ Machine Learning ที่ทำการจำแนกบทความที่ประกอบไปด้วย 265 บทความเรื่องชีวเคมี และเคมี และทำการจำแนกในระดับประโยคโดยผู้เชี่ยวชาญใช้ CoreSC และได้ทำการฝึกฝน และเปรียบเทียบ Machine Learning ที่ทำการจำแนกประเภทของข้อมูล โดย machine learning ที่ใช้ คือ Support vector machine (SVM) และ conditional random fields (CRFs) บนคลังข้อมูลทั้ง 265 บทความการประเมินผลลัพธ์ของการจำแนกข้อมูลอย่างอัตโนมัติด้วยวิธีการมาตรฐาน ค่าความถูกต้องของ Experiment, Background และ Model จะมีค่า F1-score ที่สูงคือ 76%, 62% และ 53% ตามลำดับ เครื่องมือที่สร้างขึ้นนั้นอยู่ในรูปแบบของ Web-based จะเป็นเครื่องมือที่ทำงานโดยอัตโนมัติ การเขียนบทความทางวิชาการนั้น มักจะมีวิธีการเขียนที่มีรูปแบบดังต่อไปนี้ BKG มักจะตามด้วย PROB และ RES จะตามด้วย CON

อีกทั้งยังมีผู้วิจัยได้ศึกษาเทคนิคที่ใช้ในการวิเคราะห์โครงสร้างของข้อมูลอย่างอัตโนมัติของบทความทางวิชาการ (Guo, Silins, Stenius, & Korhonen, 2013) ซึ่งมีประโยชน์อย่างมากสำหรับใช้ในการเข้าถึงข้อมูลงานวิจัยเรื่องชีวการแพทย์ อย่างไรก็ตามวิธีการที่ปรากฏอยู่ในปัจจุบันจะใช้วิธีการของ supervised machine learning (ML) ในงานวิจัยนี้ได้นำเสนอคลังข้อมูลที่ประกอบไปด้วย 50 บทความทางชีวการแพทย์ (ประกอบด้วย 8171 ประโยค และ 234 619 คำ) โดยใช้วิธีการของ Argumentative Zoning (AZ) และทำการตรวจสอบวิธีการเรียนรู้ที่เป็นที่นิยมในวงกว้างคือ Support Vector Machines (SVM) อีกทั้งยังนำเสนอโปรแกรมประยุกต์ที่ใช้ AZ ในการตอบคำถามและสรุปความ

เริ่มต้นขั้นตอนแรกจากการสร้างคลังข้อมูล corpus จากบทความทางวิชาการ 50 บทความ จากวารสารทางวิชาการของบทความทางชีวการแพทย์ เช่น Carcinogenesis, Toxicological Sciences และ Journal of Biological Chemistry เป็นต้น และทำการ label

บทความตามแบบแผนของ AZ ดังนี้ Background (BKG) คือ พื้นฐานของการเรียนรู้,ข Problem (PROB) คือ ปัญหาของงานวิจัย Method (METH) คือ วิธีการที่ใช้ Result (RES) คือ ผลลัพธ์ที่ได้ Conclusion (CON) คือ ผลสรุป Future work (FUT) คือ แนวทางในอนาคต Connection (CN) และ Difference (DIFF) คือ งานที่เกี่ยวข้องทั้งทางตรง และทางอ้อม จากนั้นผู้วิจัยจึงสร้างเครื่องมือที่อนุญาตให้ผู้ใช้งานเปิดบทความทางวิชาการในโปรแกรม Web browser Firefox และทำการอธิบายทั้งหมด 50 บทความ ประโยคต่อประโยค

จากการทดลองได้แสดงผลลัพธ์จากการเรียนรู้ของ SVM กับข้อมูล 500 ประโยคซึ่งเป็น 6% ของ Corpus ได้ผลลัพธ์ที่เป็นที่น่าพอใจโดยมีค่าความถูกต้อง 82% ในงานด้านการตอบคำถามผู้ทำงานวิจัยด้านชีวการแพทย์ค้นหาข้อมูลที่เป็นที่ต้องการได้รวดเร็วจากใช้ AZ ในงานด้านการสรุปความ การสกัดประโยคนั้นยังให้ค่าความถูกต้องที่ดี

มีผู้วิจัยได้เสนองานวิจัยที่เกี่ยวข้องกับการนำเอาวิธีการในการระบุบริบทที่เกี่ยวข้องกับประโยคในบทความทางวิชาการ และมีการใช้ข้อมูลให้กับระบบการบริการข้อมูลสำหรับกลุ่มของผู้ทำงานวิจัย (Angrosh, Cranefield, & Stanger, 2013) อันเนื่องมาจากผู้สร้างงานวิจัยได้เผยแพร่งานของตนเองสู่สาธารณะในรูปแบบของฐานข้อมูลที่เป็นข้อความซึ่งมีความสำคัญสำหรับผู้ทำงานวิจัย อีกทั้งยังเป็นแหล่งข้อมูลที่มีคุณค่าสำหรับการนำเอามาใช้กับงานสกัดความรู้ทางวิทยาศาสตร์อย่างอัตโนมัติอีกด้วย เมื่อไม่นานมานี้ได้มีผู้สร้าง Web API เพื่อให้ใช้งานสำหรับการเข้าถึงฐานข้อมูลของบทความ อีกทั้งเพื่อให้สามารถสร้าง Web Application

ในงานวิจัยนี้ถึงมีวัตถุประสงค์ที่จะสร้างแบบแผนการกำกับประโยคในบทความทางวิชาการ และอธิบายการทดลองที่ทำการทดลองโดยการใช้ conditional random fields (CRFs) เพื่อใช้การจำแนกประเภทของประโยค อีกทั้งยังมีการนำเสนอโปรแกรมประยุกต์ที่มีชื่อว่า CitContExt (citation context extraction application) ที่สร้างขึ้นมาจากเทคนิคที่ได้กล่าวไว้ข้างต้น

การกำหนดชนิดขององค์ประกอบของประโยคกับการอ้างอิง และประโยคที่อยู่รอบๆ ประโยคนั้นๆ ทางผู้วิจัยทำการเลือกด้วยวิธีการสุ่มบทความทางวิชาการในเรื่องที่แตกต่างกันจำนวน 20 บทความจาก Lecture Notes in Computer Science (LNCS) ที่เผยเพร่จาก Springerlink.com โดยลักษณะของบทความทางวิชาการจาห LNCS นั้นจะมีรูปแบบการอ้างอิงที่เป็นลักษณะของตัวเลข จึงเป็นการอำนวยความสะดวกในการระบุประโยคกับการอ้างอิง เริ่มต้นจากการฝึกฝน data set จาก 20 บทความจะมี 246 ย่อหน้า และเรียนรู้ย่อหน้ากับการอ้างอิง ในแต่ละย่อหน้าคือพื้นที่ชุดของการอ้างอิง โดยที่พื้นที่การอ้างอิงถูกกำหนดเป็นข้อความที่ประกอบไปด้วยประโยคที่มีการอ้างอิง และประโยครอบๆ และกำหนดบริบทที่มีความเกี่ยวข้องกับประโยคที่เป็นไปได้

การวิเคราะห์นี้ส่งผลต่อการกำหนดชนิดของบริบทของประโยคที่ได้กำหนดไว้ ดังนี้

Context Types for Non-Citation Sentences เป็นบริบทที่กำหนดให้กับประโยคโดยไม่มีการอ้างอิงในบทความทางวิชาการ

Background (BGR) กำหนดให้กับประโยคที่เป็นประเด็นเบื้องหลัง

Issues (ISSUE) กำหนดให้ประโยคที่อ้างอิงถึง หรือถูกเขียนโดยผู้ทำงานวิจัย

Gaps (GAPS) กำหนดให้ประโยคที่เป็นประโยคทั่วไปที่บอกถึงหัวข้อที่ทำในปัจจุบัน

Description (DES) เป็นประโยคที่แสดงถึงการอธิบาย อธิบายงานวิจัยก่อนหน้านี้ วิธีการทำงาน ประเด็นของหัวข้อที่ทำงานวิจัย หรือเบื้องหลังของงานวิจัย

Current Work Outcome (CWO) ประโยคที่บอกถึงการอ้างอิงผลลัพธ์ของการทำงานวิจัย

Future Work (FW) เป็นประโยคที่บอกถึงจุดประสงค์ของงานที่จะทำในอนาคตในหัวข้อเดิมที่กำลังทำยู่

Context Types for Citation Sentences เป็นบริบทที่มีความเกี่ยวข้องกับประโยคการอ้างอิงที่ถูกกำหนดอยู่บนพื้นฐานของเหตุผล สำหรับใช้ในการทำงานที่อ้างอิงถึง ในบทความทางวิชาการปัจจุบัน โดยแยกแยะระหว่างบริบทต่อไปนี้สำหรับการอ้างอิง ดังนี้

Cited Work Identifies Gaps (CWIG) เป็นประโยคที่ใช้ในอ้างอิงงานที่ทำอยู่

Cited Work Overcomes Gaps (CWOG) เป็นประโยคที่บอกถึงการที่จะก้าวข้ามผ่านประเด็นของงานวิจัย

Uses Outputs from Cited Works (UOCW) ประโยคที่อ้างอิงถึงการผลลัพธ์ของงานที่รายงานออกมาของบทความทางวิชาการนั้นๆ

Results with Cited Work (RWCW) ประโยคที่มีความเกี่ยวข้องกับผลลัพธ์ขแงบทความในงานที่กำลังทำอยู่

Results with Cited Work (RWCW) เป็นประโยคที่เปรียบเทียบความแตกต่างของงานวิจัย

Shortcomings in Cited Work (SCCW) ประโยคที่อ้างอิงถึง ขอ้จำกัดของงานที่ทำอยู่

Issue Related Cited Work (IRCW) เป็นประโยคที่อ้างอิงงานอื่นๆ สำหรับประเด็นภายในหัวข้องานวิจัย และหัวข้อที่อภิปรายในบทความทางวิชาการ

จากนั้นจึงทำการสร้าง Framework สำหรับการกำหนดชนิดของบริบทของ CRFs ขึ้นมา โดยจะดำเนินการทดลองกับ 1000 ย่อหน้ากับการสกัดการอ้างอิง จาก 70 บทความทางวิชาการที่

เลือกจาก LNCS โดยใช้ 40 บทความเป็น training data set และทำการทดสอบตัวโมเดลกับอีก 30 บทความทางวิชาการที่เหลือ ในการ training โมเดลทั้งหมด 40 บทความจะใช้บทความจำนวน 20 บทความแรกเป็นชุดข้อมูลพัฒนา และใช้บทความทั้งหมด 40 บทความเป็นชุดข้อมูลประเมินผลของตัวจำแนกประเภทบทความ โดยมีขั้นตอนดังนี้

1. Feature Definition เริ่มต้นด้วยการกำหนดคุณลักษณะโดยขึ้นอยู่กับรูปแบบการให้คำอธิบายประกอบ และทำการกำหนดเข้าสู่ class ใดๆ

2. Feature Selection ได้ทำการทดลองกับข้อมูลชุดแรกเพื่อการวิเคราะห์การเลือกคุณลักษณะ

3. Developing the Classifier Model หลังจากระบุชุดคุณลักษณะที่ดีที่สุดแล้ว จะทำการระบุคุณลักษณะของประโยคด้วยตัวเองใน 20 บทความทางวิชาการ ผลลัพธ์ในการ training 40 บทความจะใช้ในการสร้างตัวจำแนกประเภท

4. Testing เป็นการประเมินผลการทำงานของตัวจำแนกประเภทด้วยการทดสอบชุดข้อมูลจำนวน 30 บทความทางวิชาการ

สำหรับการ Training CRFs จะใช้ MALLET เป็น Java-based package มี algorithm สำหรับการทำงานแบบ sequential data.

ผลของการจำแนกประเภทของ CRFs บนชุดข้อมูลทดสอบ สามารถที่จะให้ความถูกต้อง 92.08%, 92.92% และ 90.01% กับชุดข้มูลทดสอบทั้งสามชุด โดยค่าเฉลี่ยจากการจำแนกประเภทข้อมูลจากข้อมูลทดสอบ 30 บทความคือ 91.67%

บทที่ 3
วิธีการดำเนินการวิจัย

การพัฒนาเครื่องมือวิเคราะห์โครงสร้าง และองค์ประกอบหรือมูฟของบทคัดย่อในบทความทางวิชาการสาขาต่างๆ นั้นเป็นงานวิจัยที่เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติ ซึ่งเป็นสาขาหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence หรือ AI) ผู้วิจัยเลือกวิธีการเรียนรู้ของเครื่อง (Machine Learning) มาใช้ในการวิเคราะห์บทคัดย่อในบทความทางวิชาการ โดยทำในลักษณะของการทำเหมืองข้อความ (Text Mining)

จากการศึกษาการประมวลผลเพื่อวิเคราะห์รูปแบบ และโครงสร้างของเอกสารทางวิชาการนั้น ลักษณะของงานเป็นการจำแนกกลุ่มของประโยคในบทคัดย่อ (Classification) ให้อยู่ในมูฟต่างๆ สิ่งที่จำเป็นต้องให้ความสำคัญ คือ ความถูกต้องของการวิเคราะห์ ทั้งนี้มีหลายปัจจัยที่ส่งผลต่อความถูกต้องของการวิเคราะห์ ซึ่งการเรียนรู้ของเครื่องแบบการจำแนกประเภทข้อมูลนั้นส่งผลต่อความถูกต้องของการวิเคราะห์ การจำแนกประเภทข้อมูลมีหลากหลายขั้นตอนวิธี (Algorithm) ที่ใช้ในการจำแนกกลุ่มประโยคให้อยู่ในมูฟต่างๆ เช่น Decision tree, Naïve Bayes, K-Nearest Neighbors (KNN) และ Neural Network เป็นต้น

จากงานวิจัยเดิมการวิเคราะห์โครงสร้างของบทคัดย่อ เลือกใช้วิธีการจำแนกข้อมูลแบบ Decision tree เป็นตัวจำแนกกลุ่มของประโยค ซึ่งให้ประสิทธิภาพในการจำแนกข้อมูลที่ดี ผู้วิจัยเห็นว่าเพื่อเป็นการศึกษาในเรื่องการที่จะทำให้ประสิทธิภาพในการจำแนกข้อมูลที่ดีขึ้นนั้น จึงควรที่ศึกษาวิธีการจำแนกข้อมูลอื่นๆ นำมาทดสอบกับการทำเหมืองข้อมูล เพื่อใช้เปรียบเทียบประสิทธิภาพในการจำแนกประเภทข้อมูลของอัลกอริทึม จากงานวิจัยอื่นๆ พบว่ายังมีวิธีการที่ใช้ในการจำแนกข้อมูลอื่นๆ ที่ยังไม่ถูกนำมาใช้ หรือไม่เป็นที่นิยมที่จะนำมาใช้กับงานเหมืองข้อมูล ดังนั้นงานวิจัยนี้จึงเลือกใช้อัลกอริทึม Support Vector Machine (SVM) ในการจำแนกประเภทของข้อมูล

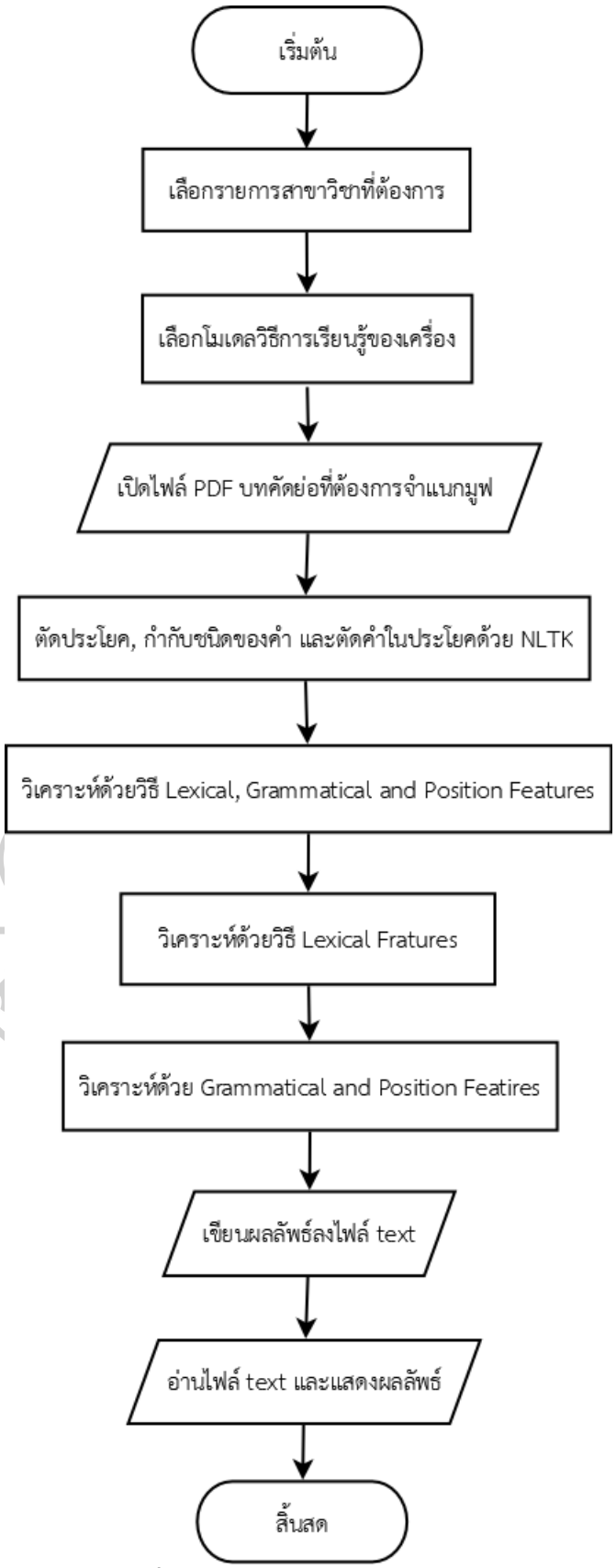
3.1 ออกแบบวิธีการทำงานของระบบ

งานวิจัยนี้เป็นการสร้างโปรแกรมประยุกต์ในรูปแบบเว็บไซต์ ใช้สำหรับการวิเคราะห์รูปแบบโครงสร้างของบทคัดย่อ และวิเคราะห์องค์ประกอบของบทคัดย่อ หรือมูฟของเอกสารทางวิชาการได้ในหลายหลายสาขาวิชาสาขา โดยใช้วิธีการวิเคราะห์โครงสร้าง และมูฟ 3 วิธี คือ

- การวิเคราะห์ด้วย Lexical Features
- การวิเคราะห์ด้วย Grammatical and Position Features
- การวิเคราะห์ด้วย Lexical, Grammatical and Position Features

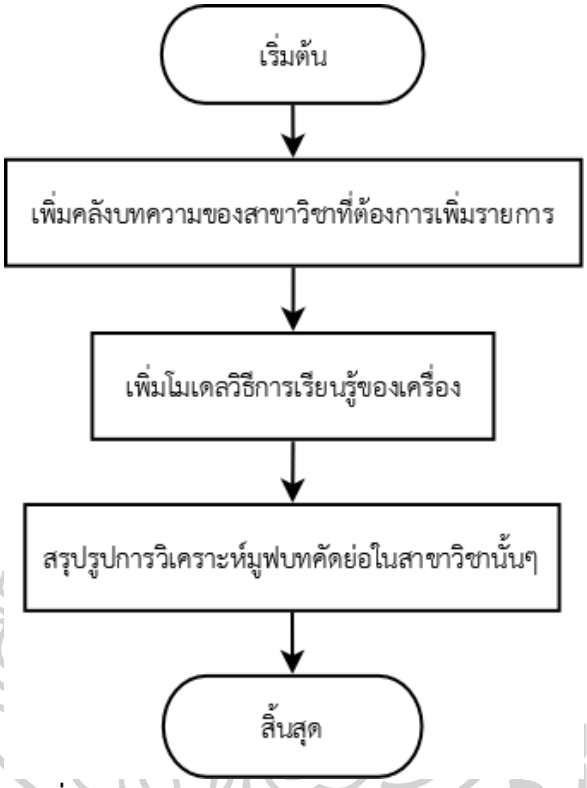
การทำงานของโปรแกรมนั้นแบ่งการทำงานออกเป็น 2 ส่วนด้วยกัน เริ่มต้นการทำงานของโปรแกรมผู้ใช้จะต้องสร้างรายการของสาขาวิชาต่างๆ ที่ผู้ใช้ต้องการและนำเข้าไฟล์บทคัดย่อในรูปแบบไฟล์ Text จากเอกสารทางวิชาการในสาขานั้นๆ เพื่อเป็นการฝึกฝนตัวโมเดลการจำแนกประเภทของมูฟ โดยมีรายละเอียดต่างๆ ดังนี้

ส่วนที่ 1 คือ ส่วนของผู้ใช้งาน โดยผู้ใช้ต้องเลือกรายการสาขาวิชาของบทคัดย่อที่ต้องการให้ระบบนำแนกโครงสร้าง จากนั้นเลือกโมเดลวิธีการเรียนรู้บทคัดย่อที่ต้องการใช้วิเคราะห์โครงสร้าง จากนั้นผู้ใช้จะต้องนำเข้าไฟล์บทคัดย่อของเอกสารทางวิชาการที่อยู่ในรูปแบบของไฟล์ PDF เมื่อรับบทคัดย่อเข้าสู่โปรแกรมแล้ว บทคัดย่อจะผ่านเข้าสู่กระบวนการวิเคราะห์ แสดงผลลัพธ์จากการวิเคราะห์ออกมา โดยมีการแสดงข้อความเพื่อแบ่งประโยคในบทคัดย่อออกจากกัน แสดงการทำงานของระบบได้จากรูปที่ 3.1



รูปที่ 3.1 แสดงการทำงานของระบบในส่วนผู้ใช้งาน

ส่วนที่ 2 คือ ส่วนฝึกสอน ผู้ใช้งานสามารถเพิ่มคลังบทความของสาขาวิชาที่ต้องการ เพิ่มโมเดลวิธีการเรียนรู้ของเครื่อง และเพิ่มคลังคำศัพท์ที่ใช้ในการวิเคราะห์โครงสร้างของบทคัดย่อ และระบบจะแสดงสรุปรูปแบบมูฟของบทคัดย่อในสาขาวิชานั้นๆ แสดงการทำงานของระบบได้จากรูปที่ 3.2



รูปที่ 3.2 แสดงการทำงานของระบบในส่วนฝึกสอน

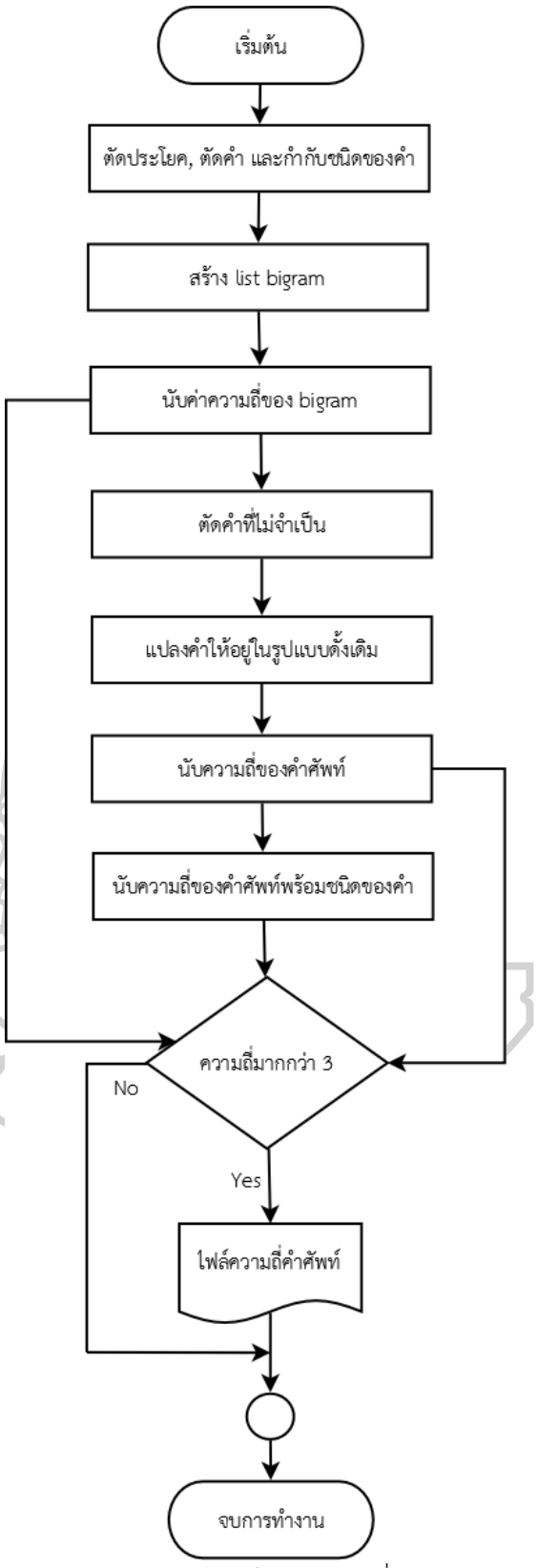
3.2 การวิเคราะห์ด้วย Lexical Features

จากรูปที่ 3.3 หลังจากที่โปรแกรมรับบทคัดย่อที่นำเข้าโดยผู้ใช้งานแล้ว จะเข้าสู่ขั้นตอนของการตัดประโยค, ตัดคำ และการกำกับชนิดของคำนั้น ในประโยค เพื่อทราบถึงหน้าที่ของคำในประโยค จากนั้นก็จะเข้าสู่กระบวนการวิเคราะห์โครงสร้าง และองค์ประกอบหรือมูฟด้วย Lexical Features ซึ่งมีวิธีการ 2 วิธีการด้วยกันดังนี้

1. การวิเคราะห์ด้วยวิธีดูจากค่าความถี่ของคำศัพท์ที่เกิดขึ้นในแต่ละมูฟของบทคัดย่อ โดยตัดทำที่ไม่สื่อความหมาย และคำที่ไม่ส่งผลกระทบต่อประโยคออก จากนั้นแปลงคำศัพท์ให้อยู่ในรูปดั้งเดิมแล้วนำไปเข้าสู่กระบวนการวิเคราะห์ โดยมีการแบ่งคุณลักษณะที่นำมาใช้ในการวิเคราะห์ดังนี้

- ความถี่ของคำศัพท์
- ความถี่ของ bigram (คำที่เขียนเรียงกันและเกิดขึ้นร่วมกันในประโยคโดยเก็บคำที่ติดกันในประโยคนั้น)
- ความถี่ของคำศัพท์พร้อมชนิดของคำ

ค่าความถี่ที่นำมาเปรียบเทียบได้จากการให้ระบบเรียนรู้จากฐานข้อมูลต้นแบบที่ได้รับการแบ่งมูฟจากนักภาษาศาสตร์ โดยกระบวนการเก็บค่าความถี่โดยให้ระบบเรียนรู้จากฐานข้อมูลต้นแบบนั้น จะทำการนับค่าความถี่ โดยเริ่มต้นจะทำการตัดประโยค ตัดคำจากบทความต้นแบบ มีการแบ่งประเภทของการเก็บค่าความถี่ ดังนี้ เก็บค่าความถี่ของคำศัพท์ที่เกิดขึ้น เก็บค่าความถี่ของ Bigrams และเก็บค่าความถี่ของคำศัพท์พร้อมชนิดของคำที่เกิดขึ้น โดยการเก็บค่าความถี่ของคำศัพท์และเก็บค่าความถี่ของคำศัพท์พร้อมชนิดของคำที่เกิดขึ้น จะทำการตัดคำศัพท์ที่ไม่สื่อความหมายและคำที่ไม่ส่งผลต่อประโยค แปลงคำศัพท์ให้อยู่ในรูปดั้งเดิม นับค่าความถี่ แล้วทำการเก็บข้อมูลที่พบค่าความถี่ตั้งแต่ 3 ขึ้นไป จะเก็บค่าความถี่โดยแยกในแต่ละมูฟ ดังแสดงในรูปที่ 3.3



รูปที่ 3.3 แสดง Flow Chart การเก็บค่าความถี่ของคำศัพท์โดยให้ระบบเรียนรู้จากฐานข้อมูลต้นแบบ

2. วิเคราะห์โดยเปรียบเทียบกับคลังศัพท์ โดยทำการตัดคำศัพท์ที่ไม่สื่อความหมาย และคำไม่ส่งผลต่อประโยค แปลงรูปของคำศัพท์ให้อยู่ในรูปดั้งเดิมแล้วนำไปวิเคราะห์ โดยวิเคราะห์ และเปรียบเทียบกับคลังศัพท์ว่าอยู่ในมูฟประเภทใด โดยวิเคราะห์จากคุณสมบัติดังนี้ วิเคราะห์และเปรียบเทียบคำ วิเคราะห์และเปรียบเทียบ N-gram (กลุ่มคำที่มีคำเกิดขึ้นร่วมกันจำนวน N คำ) วิเคราะห์และเปรียบเทียบคำที่ปรากฏร่วมกันในประโยค วิเคราะห์และเปรียบเทียบรูปแบบประโยค จากวิธีการวิเคราะห์ทั้งสองแบบทำให้เกิดคุณลักษณะทั้งหมดของ Lexical Features ดังต่อไปนี้

- **วิเคราะห์โดยเปรียบเทียบกับค่าความถี่ของคำศัพท์** ที่เกิดขึ้นในแต่ละมูฟของบทคัดย่อ โดยทำการตัดคำศัพท์ที่ไม่สื่อความหมาย และคำที่ไม่ส่งผลต่อประโยค แปลงคำศัพท์ให้อยู่ในรูปดั้งเดิมแล้วทำการวิเคราะห์

- **วิเคราะห์โดยเปรียบเทียบกับค่าความถี่ของ Bigram** (คำที่เขียนเรียงกันและเกิดขึ้นร่วมกันในประโยคโดยเก็บค่าทีละ 2 คำ ซึ่งเป็นคำที่ติดกันในประโยคนั้น) ที่เกิดขึ้นในแต่ละมูฟของบทคัดย่อแล้วทำการวิเคราะห์

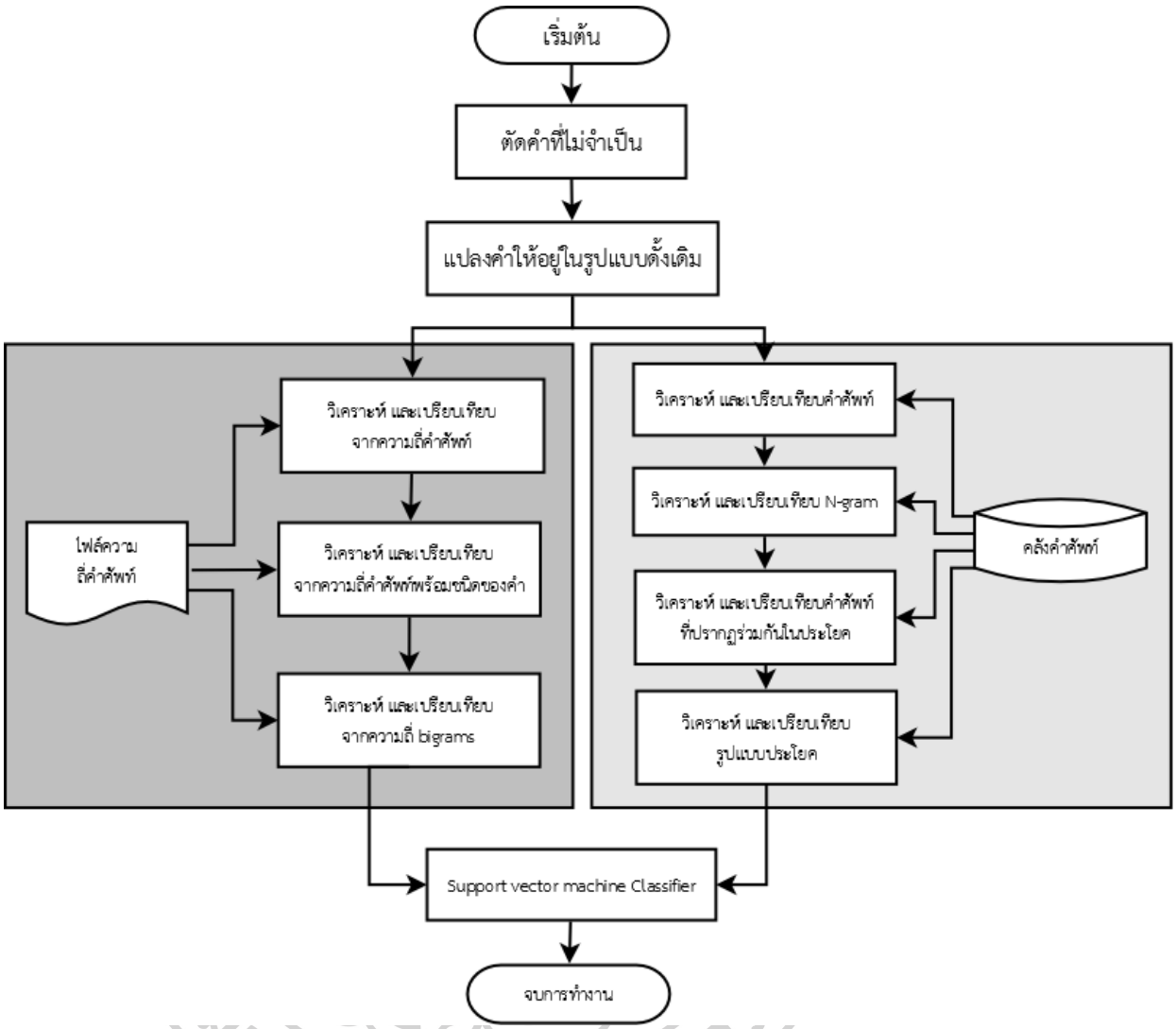
- **วิเคราะห์โดยเปรียบเทียบกับค่าความถี่ของคำศัพท์พร้อมชนิดของคำ** ที่เกิดขึ้นในแต่ละมูฟของบทคัดย่อ โดยนำลิสต์ของประโยคที่ทำการกำกับชนิดของคำเรียบร้อยแล้วมาทำการวิเคราะห์โดยนำลิสต์มาตัดคำศัพท์ที่ไม่สื่อความหมาย และคำที่ไม่ส่งผลต่อประโยค แปลงคำศัพท์ให้อยู่ในรูปดั้งเดิมแล้วทำการวิเคราะห์

- **วิเคราะห์และเปรียบเทียบคำกับคลังศัพท์** โดยทำการตัดคำศัพท์ที่ไม่สื่อความหมายและคำไม่ส่งผลต่อประโยค แปลงรูปของคำศัพท์ให้อยู่ในรูปดั้งเดิมแล้วนำไปวิเคราะห์ โดยวิเคราะห์และเปรียบเทียบกับคลังศัพท์ว่าพบอยู่ในมูฟประเภทใด

- **วิเคราะห์และเปรียบเทียบ N-gram กับคลังศัพท์** คำที่เขียนเรียงกันและเกิดขึ้นร่วมกันในประโยคโดยเก็บค่าทีละ N คำ ซึ่งเป็นคำที่ติดกันในประโยคนั้น ไปเปรียบเทียบกับคลังศัพท์โดยวิเคราะห์และเปรียบเทียบกับคลังศัพท์ว่าพบอยู่ในมูฟประเภทใด

- **วิเคราะห์และเปรียบเทียบคำที่ปรากฏร่วมกันในประโยคกับคลังศัพท์** ที่ปรากฏร่วมกันในประโยคกับคลังศัพท์ ซึ่งคำศัพท์ที่ปรากฏร่วมกันนั้นจะส่งผลต่อความหมายของประโยค โดยวิเคราะห์และเปรียบเทียบกับคลังศัพท์ว่าพบอยู่ในมูฟประเภทใด

- **วิเคราะห์และเปรียบเทียบรูปแบบประโยคกับคลังศัพท์** วิเคราะห์และเปรียบเทียบรูปแบบการขึ้นต้นของประโยค การขึ้นต้นประโยคในแต่ละมูฟจะมีรูปแบบที่แตกต่างกัน โดยทำการวิเคราะห์และเปรียบเทียบกับคลังศัพท์ว่าพบอยู่ในมูฟประเภทใด



รูปที่ 3.4 แสดง Flow Chart วิเคราะห์ด้วย Lexical Features

3.3 การวิเคราะห์ด้วย Grammatical and Position Features

เมื่อระบบรับบทคัดย่อที่นำเข้าโดยผู้ใช้งานแล้ว ระบบจะทำการทำการตัดประโยค ตัดคำ และกำกับชนิดของคำให้กับคำในประโยค ในการวิเคราะห์ด้วย Grammatical and Position Features นั้น ได้กำหนดคุณลักษณะด้วยการใช้ Regular Expression เพื่อใช้เป็นรูปแบบในการฝึกสอนระบบจากฐานข้อมูลต้นแบบทั้ง 60 บทความ คุณลักษณะของ Grammatical Features มีดังนี้

- Tense จะตรวจสอบโดยการดูจาก Verb และโครงสร้างของแต่ละ Tense แบ่งเป็น Present tense, Past tense, Future tense

- Voice จะตรวจสอบโดยการดูจาก Verb และโครงสร้างของแต่ละ Tense ในแบบที่ประธานเป็นผู้กระทำ (Active voice) และแบบที่ประธานเป็นผู้ถูกกระทำ (Passive voice)

- Pronoun จะตรวจสอบว่า หากพบคำว่า We, Our, This, These ปรากฏอยู่ในประโยค จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ Pronoun

- Preposition จะตรวจสอบว่า หากพบคำว่า in, over, between, into, at, from, by, for, on, during, among ปรากฏอยู่ในประโยค จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ Preposition

- Modal จะตรวจสอบว่า หากพบคำว่า will, shall, can, could, may, might, should, would ปรากฏอยู่ในประโยค จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ Modal

- To infinitive จะตรวจสอบว่า หากพบคำว่า To ที่ขึ้นต้นประโยค หรือ to ภายในประโยค แล้วตามด้วยคำที่มีการกำกับชนิดของคำด้วยกริยาช่องที่ 1 จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ To infinitive

- In order to จะตรวจสอบว่า หากพบคำว่า In order to ปรากฏอยู่ในประโยค จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ In order to

- Whether จะตรวจสอบว่า หากพบคำว่า Whether ปรากฏอยู่ในประโยค จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ Whether

- By + V.ing จะตรวจสอบว่า หากพบคำว่า By แล้วตามด้วยคำที่มีการกำกับชนิดของคำด้วย V.ing จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ By + V.ing

- Article จะตรวจสอบว่า หากพบคำว่า A, An ปรากฏอยู่ในประโยค จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ A, An

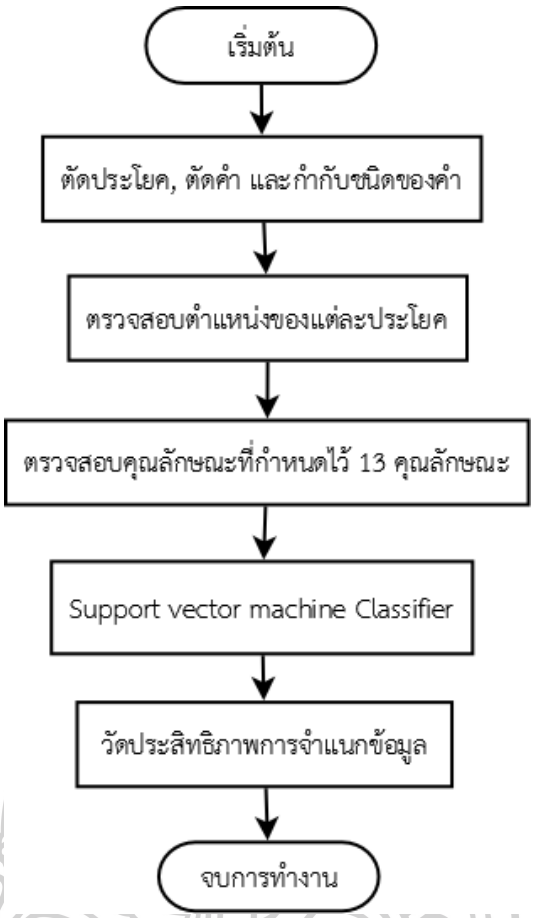
- Determiner จะตรวจสอบว่า หากพบคำว่า The ปรากฏอยู่ในประโยค จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ The

- Extraposition จะตรวจสอบว่า หากพบคำว่า It ตามด้วยคำที่เป็น Verb to be (is, am, are, was, were) และตามด้วยคำว่า that จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ Extraposition

- Nominalization จะตรวจสอบว่า หากพบคำว่า The ตามด้วยคำที่มีการกำกับชนิดของคำเป็นคำนามและตามด้วยคำว่า of จะกำหนดให้ประโยคนั้นๆ มีคุณลักษณะ Nominalization

การตรวจสอบตำแหน่งของแต่ละประโยคในบทความ โดยจะแบ่งบทความออกเป็น 3 ส่วน คือ

- ส่วนต้น กำหนดให้เป็นสองประโยคแรกที่พบในบทความ
- ส่วนท้าย กำหนดให้เป็นสองประโยคสุดท้ายที่พบในบทความ
- ส่วนที่จะไม่พิจารณา กำหนดให้เป็นประโยคอื่นๆ ที่นอกเหนือจากประโยคในส่วนต้นและส่วนท้าย



รูปที่ 3.5 แสดง Flow Chart การวิเคราะห์ด้วย Grammatical and Position Features

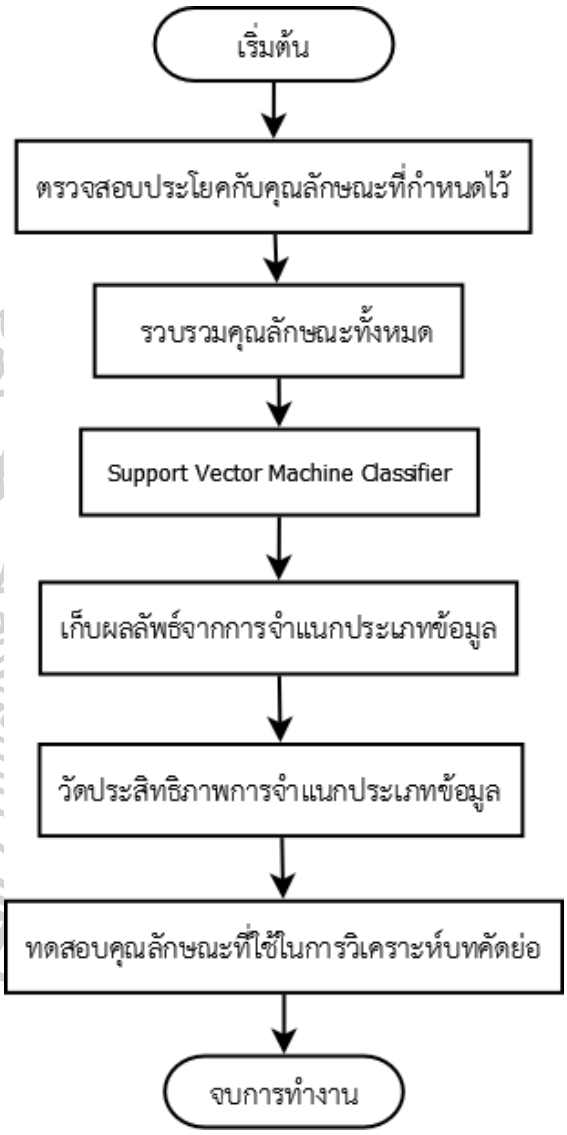
3.4 การวิเคราะห์ด้วย Lexical, Grammatical and Position Features

การวิเคราะห์ด้วย Lexical, Grammatical and Position Features จากรูปที่ 3.6 หลังจากที่รับบทความมาแล้ว ทำการตัดประโยค ตัดคำ และกำกับชนิดของคำให้กับประโยคเพื่อต้องการทราบหน้าที่ของคำในประโยค จากนั้นได้กำหนดคุณลักษณะด้วยการใช้ Regular Expression คือ

- To
- We
- This

โดยเมื่อใช้ Regular Expression แล้วจะนำไปวิเคราะห์กับไฟล์คำศัพท์ที่แบ่งเป็น 5 กลุ่มตามมูฟ คือ Background, Purpose, Method, Result และ Discussion อีกทั้งยังมี Position Features โดยจะเก็บผลลัพธ์การ Classify ของประโยคไว้ เพื่อใช้ในการวิเคราะห์ของประโยคถัดไป จากนั้นได้รวบรวมคุณลักษณะของวิธีการวิเคราะห์ด้วย Lexical Features และวิเคราะห์ด้วย Grammatical and Position Features มารวมกัน แล้วใช้คุณลักษณะทั้งหมดฝึกสอนระบบจาก

ฐานข้อมูลต้นแบบทั้ง 60 บทความ และได้ทำการทดสอบหาคุณลักษณะ โดยการลดคุณลักษณะลงครั้งละ 1 คุณลักษณะแล้ววัดประสิทธิภาพการจำแนกประเภทข้อมูล เพื่อพิจารณาว่าแต่ละคุณลักษณะมีผลต่อเปอร์เซ็นต์ความถูกต้องมากน้อยเพียงใด



รูปที่ 3.6 Flow Chart การวิเคราะห์ด้วย Lexical, Grammatical and Position Features

บทที่ 4
ผลการดำเนินการวิจัย

งานวิจัยนี้ได้ออกแบบขั้นตอนวิธีการทำงาน และพัฒนาเครื่องมือวิเคราะห์โครงสร้างและองค์ประกอบของบทความทางวิชาการในส่วนบทคัดย่อหรือมูฟในรูปแบบโปรแกรมประยุกต์บนเว็บไซต์ขึ้นมา โดยระบบประกอบด้วย 2 ส่วนหลัก คือ ส่วนการวิเคราะห์ (Analysis mode) และส่วนการฝึกฝน (Training mode) โดยมีการทดสอบกระบวนการทำงานของโปรแกรมประยุกต์บนเว็บไซต์ประกอบด้วย 2 การทดลอง คือ การวิเคราะห์โครงสร้างระดับประโยค และการทดสอบการจำแนกประเภทข้อมูลด้วย Support vector machine และ Decision tree classify ซึ่งใช้คุณลักษณะทั้ง 3 คุณลักษณะดังนี้ Lexical features, Grammar and position features และ Lexical, Grammar position features ในการทดลอง

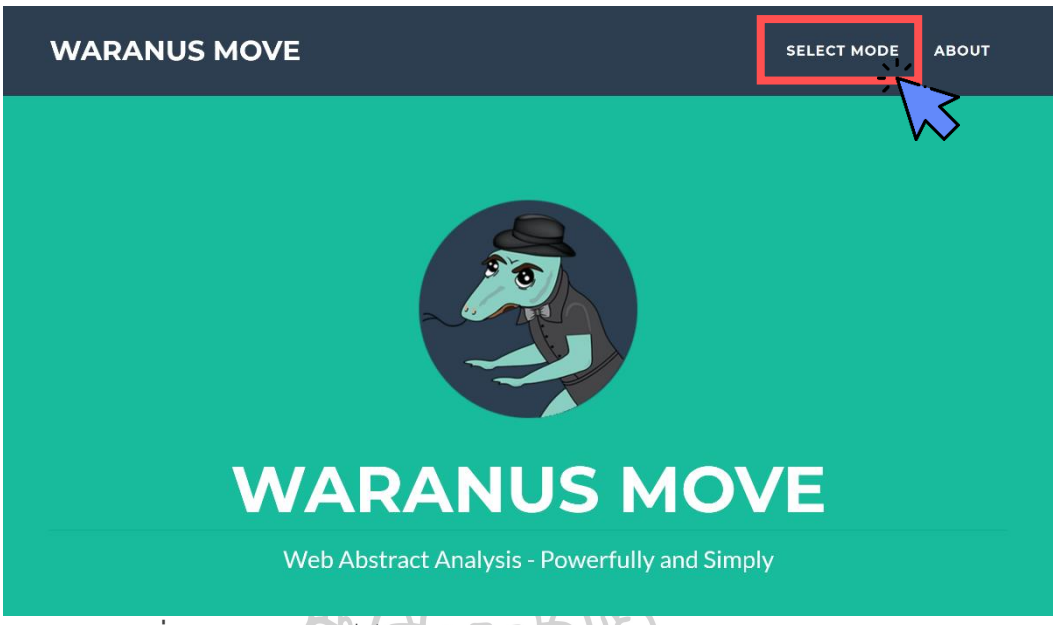
4.1 ส่วนประกอบของระบบ

4.1.1 ส่วนการวิเคราะห์ (Analysis mode)

ส่วนการวิเคราะห์ (Analysis mode) เป็นส่วนที่ผู้ใช้งานสามารถนำเข้าบทคัดย่อในรูปแบบของไฟล์เอกสาร PDF เพื่อทำการวิเคราะห์โครงสร้างของบทคัดย่อนั้นๆ โดยแสดงผลลัพธ์ออกมาเป็นมูฟ (Move) ดังนี้ Background, Purpose, Method, Result และ Discussion พร้อมทั้งแสดงกราฟสรุปจำนวนมูฟ อีกทั้งยังแสดงมูฟที่ปรากฏ และไม่ปรากฏอยู่ในบทคัดย่ออีกด้วย โดยผู้ใช้ต้องเลือกรายการสาขาวิชาของบทคัดย่อที่ต้องการให้ระบบจำแนกโครงสร้าง จากนั้นเลือกโมเดลการจำแนกโครงสร้างบทคัดย่อที่ต้องการใช้วิเคราะห์โครงสร้างบทคัดย่อ จากนั้นผู้ใช้ต้องนำเข้าไฟล์บทคัดย่อในรูปแบบไฟล์ PDF บทคัดย่อจะเข้าสู่กระบวนการวิเคราะห์โครงสร้าง โดยใช้การวิเคราะห์ทั้ง 3 รูปแบบต่อไปนี้ Lexical features, Grammatical and position features และ Lexical, Grammatical and position features โดยมีขั้นตอนและกระบวนการทำงานของระบบส่วนการวิเคราะห์ดังต่อไปนี้

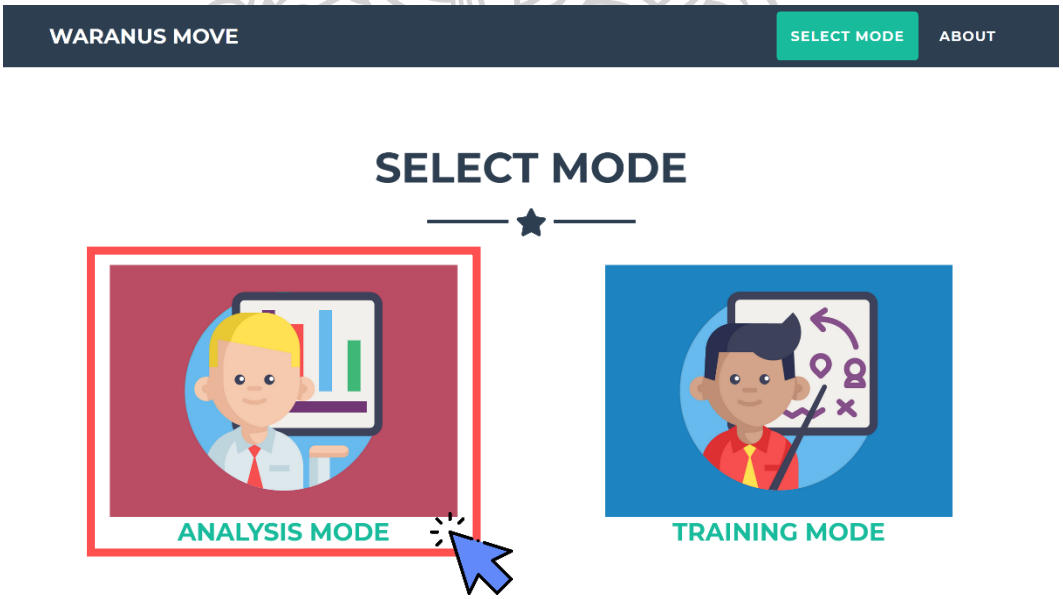
ขั้นตอนที่ 1 เลือกส่วนของการวิเคราะห์ (Analysis mode)

วิธีการเข้าสู่ขั้นตอนของกระบวนการวิเคราะห์บทคัดย่อ เมื่อเข้าสู่หน้าแรกของเว็บไซต์แสดงดังในรูปที่ 4.1 ผู้ใช้สามารถเลือกเมนูการวิเคราะห์ได้จากแถบเมนูด้านบนสุดของหน้าจอ โดยเลือกเมนู Select mode



รูปที่ 4.1 แสดงหน้าเว็บไซต์หน้าแรกของระบบวิเคราะห์โครงสร้างบทคัดย่อ

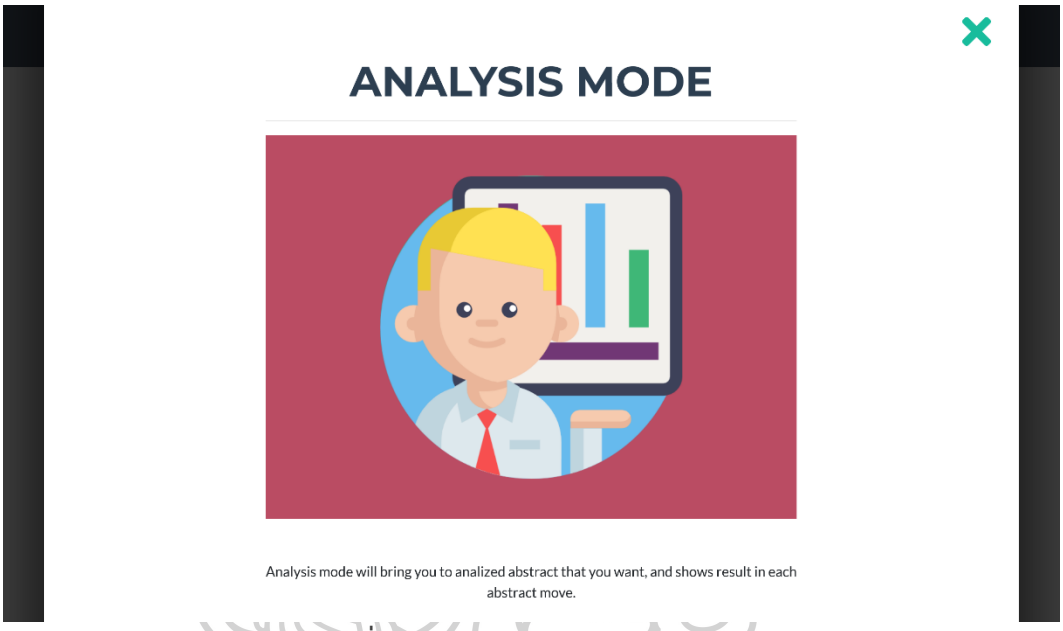
เมื่อผู้ใช้เลือกเมนู Select mode แล้วระบบจะแสดงเมนู 2 เมนู.ให้ผู้ใช้เลือกดังนี้ คือ Analysis mode และ Training mode ดังในรูปที่ 4.2 ให้ผู้ใช้เลือกเมนู Analysis mode



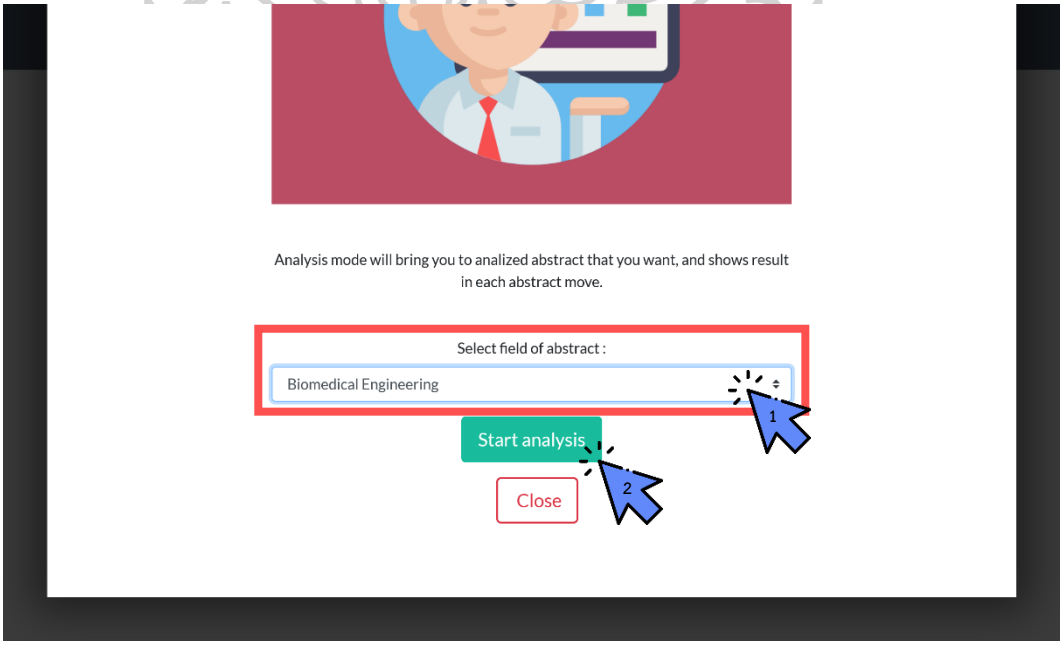
รูปที่ 4.2 แสดงเมนู Select mode

ขั้นตอนที่ 2 เลือกสาขาวิชาของบทคัดย่อ

เมื่อผู้ใช้เลือกเมนู Analysis mode ในขั้นตอนที่ 1 แล้ว จะเข้าสู่ส่วนการวิเคราะห์ในรูปที่ 4.3 โดยระบบแสดงรายการสาขาวิชาของบทคัดย่อ จากนั้นให้ผู้ใช้เลือกสาขาวิชาที่ต้องการ มีขั้นตอนดังตัวอย่างในรูปที่ 4.4 คือ 1) เลือกสาขาวิชา Biomedical Engineering เพื่อเลือกโมเดลการจำแนกโครงสร้างของบทคัดย่อในสาขาวิชา Biomedical Engineering และ 2) เลือกปุ่ม Start analysis เพื่อเข้าสู่ขั้นตอนการเลือกโมเดลจำแนกประเภทข้อมูลต่อไป



รูปที่ 4.3 แสดงส่วน Analysis mode



รูปที่ 4.4 แสดงส่วนรายการสาขาวิชาของบทคัดย่อที่ต้องการให้ระบบจำแนกโครงสร้าง

ขั้นตอนที่ 3 เลือกโมเดลจำแนกประเภทของข้อมูลและไฟล์บทคัดย่อ

เมื่อผู้ใช้เลือกปุ่ม Start analysis ในขั้นตอนที่ 2 แล้วระบบจะเข้าสู่ส่วนการวิเคราะห์โดยมีแสดงรายละเอียดในรูปที่ 4.5 ดังนี้

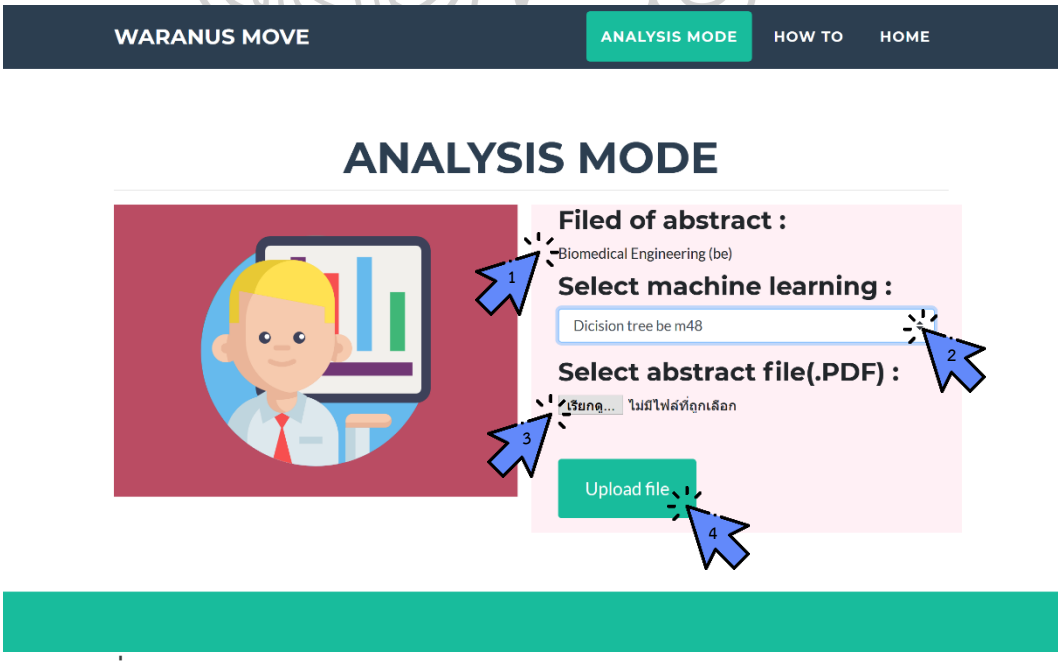
1. Filed of abstract แสดงชื่อและชื่อย่อของสาขาวิชาที่ผู้ใช้เลือกจากขั้นตอนที่ 2 จากตัวอย่างคือสาขาวิชา Biomedical Engineering (be)

2. Select machine learning แสดงรายชื่อโดเมลจำแนกประเภทของบทคัดย่อที่ได้ส่วนการฝึกฝน ให้ผู้ใช้เลือกโดเมลจำแนกปรเภทข้อมูลที่ต้องการ จากตัวอย่างคือ “Dicision tree be m48” โดยที่

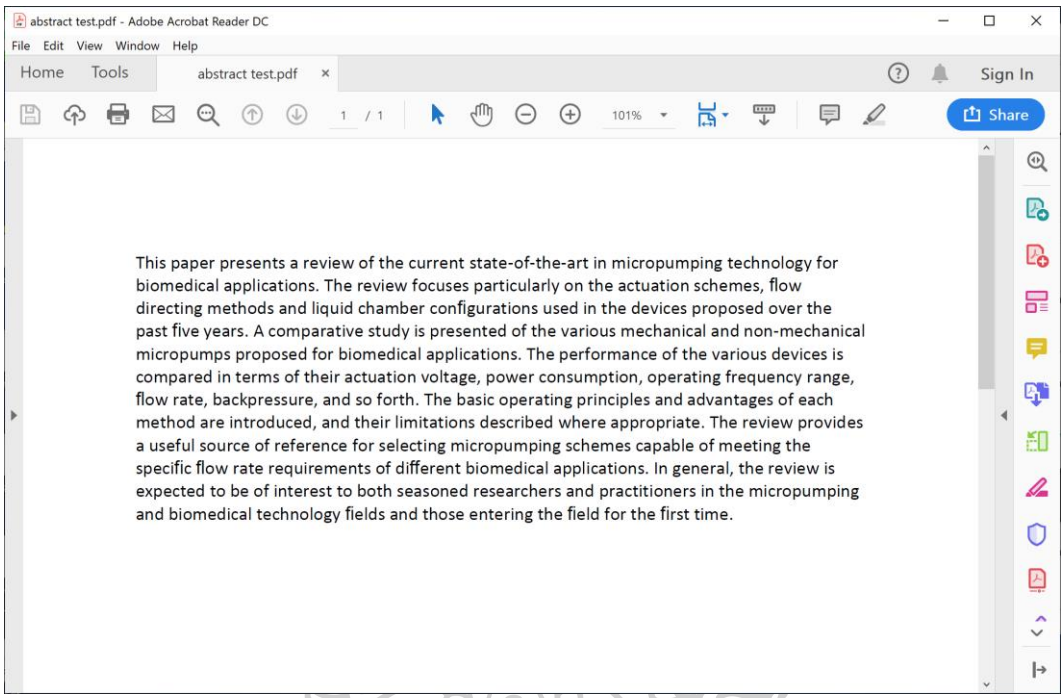
Dicision tree	คือ	Machine learning ต้นไม้ตัดสินใจ
be	คือ	ชื่อย่อของสาขาวิชาของบทคัดย่อ
m48	คือ	รหัสประจำตัวของโมเดลจำแนกประเภทข้อมูล

3. Select abstract file (.PDF) แสดงส่วนให้ผู้ใช้เลือกไฟล์บทคัดย่อที่ผู้ใช้ต้องการให้ระบบวิเคราะห์โครงสร้างในรูปแบบไฟล์ PDF โดยการกดปุ่ม “เรียกดู...” เพื่อเลือกไฟล์จากตัวอย่างไฟล์ PDF มีลักษณะดังรูปที่ 4.6

4. Upload file คือ ปุ่มในขั้นตอนสุดท้ายให้ผู้ใช้เลือกเพื่อให้ระบบดำเนินการเข้าสู่ขั้นตอนต่อไป หลังจากผู้ใช้เลือกโมเดลจำแนกประเภทข้อมูลและเลือกไฟล์บทคัดย่อแล้ว



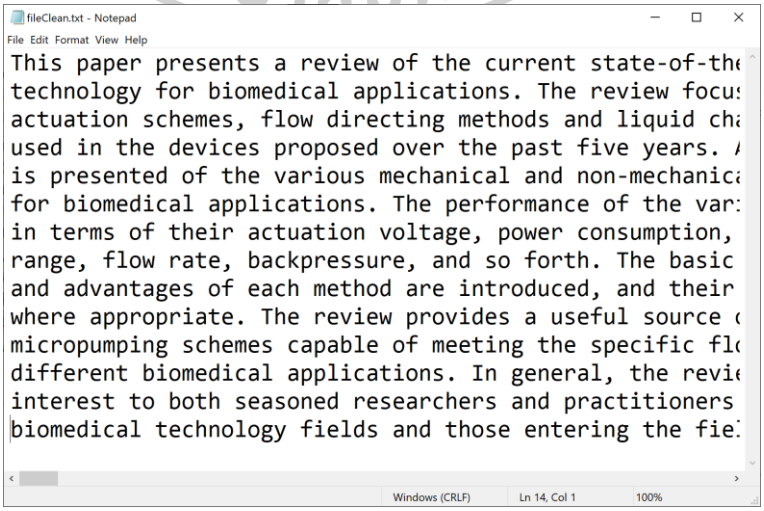
รูปที่ 4.5 แสดงส่วนการเลือกโมเดลสำหรับการจำแนกและวิเคราะห์โครงสร้าง และส่วนการนำเข้าไฟล์เอกสาร PDF



รูปที่ 4.6 แสดงตัวอย่างไฟล์บทคัดย่อในรูปแบบไฟล์ PDF

ขั้นตอนที่ 4 เลือกรูปแบบการวิเคราะห์โครงสร้างของบทคัดย่อ

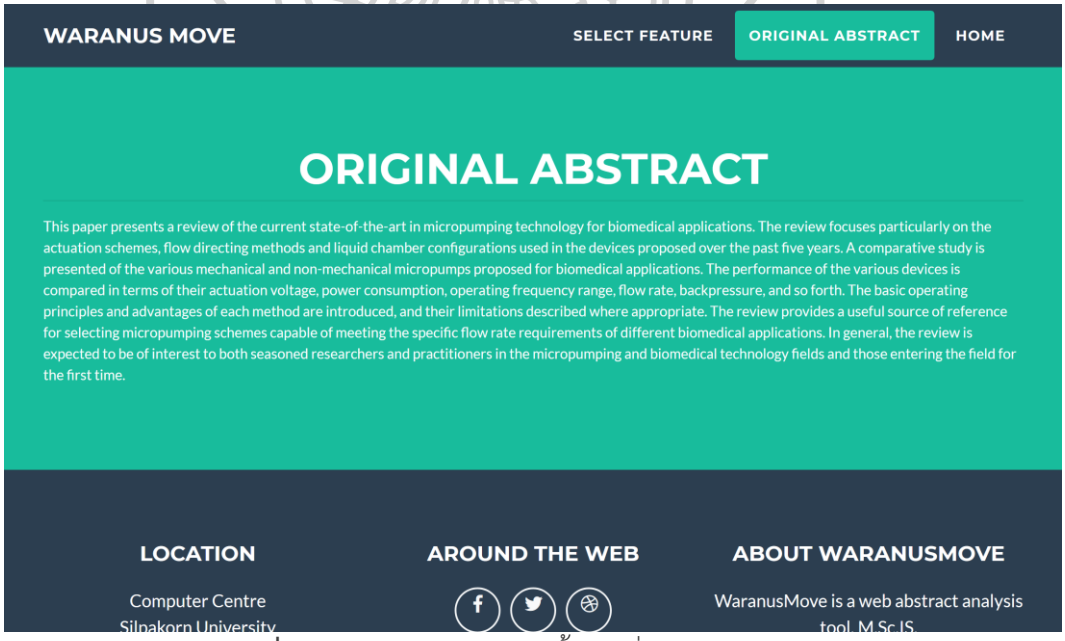
หลังจากขั้นตอนการเลือกโมเดลจำแนกประเภทของข้อมูลและไฟล์บทคัดย่อแล้วนั้น ระบบจะนำไฟล์เอกสาร PDF ที่ผู้ใช้เลือกเข้าสู่กระบวนการ Preprocessing ประกอบด้วยการแปลงไฟล์บทคัดย่อให้อยู่ในรูปแบบของไฟล์ Text และเข้าสู่กระบวนการ Cleaning text เพื่อเตรียมเอกสารให้พร้อมก่อนเข้าสู่การเลือกกลุ่มของคุณลักษณะที่ใช้ในการวิเคราะห์ทั้ง 3 รูปแบบ ได้แก่ Lexical features, Grammatical and position features และ Lexical, Grammatical and position features ตัวอย่างไฟล์ Text ที่ผ่านกระบวนการ Preprocessing แล้วดังรูปที่ 4.7



รูปที่ 4.7 แสดงไฟล์ Text ที่ผ่านกระบานการ Preprocessing แล้ว

จากนั้นระบบแสดงเมนูให้ผู้ใช้สามารถเลือกกลุ่มของคุณลักษณะที่ใช้ในการวิเคราะห์ทั้ง 3 รูปแบบได้แก่ Lexical features, Grammatical and position features และ Lexical, Grammatical and position features ดังรูปที่ 4.8 ส่วนต่อมาคือของบทคัดย่อต้นฉบับดั้งเดิม (Original abstract) ที่ผ่านกระบวนการ Preprocessing แล้วแสดงให้กับผู้ใช้อีกด้วยดังรูปที่ 4.9

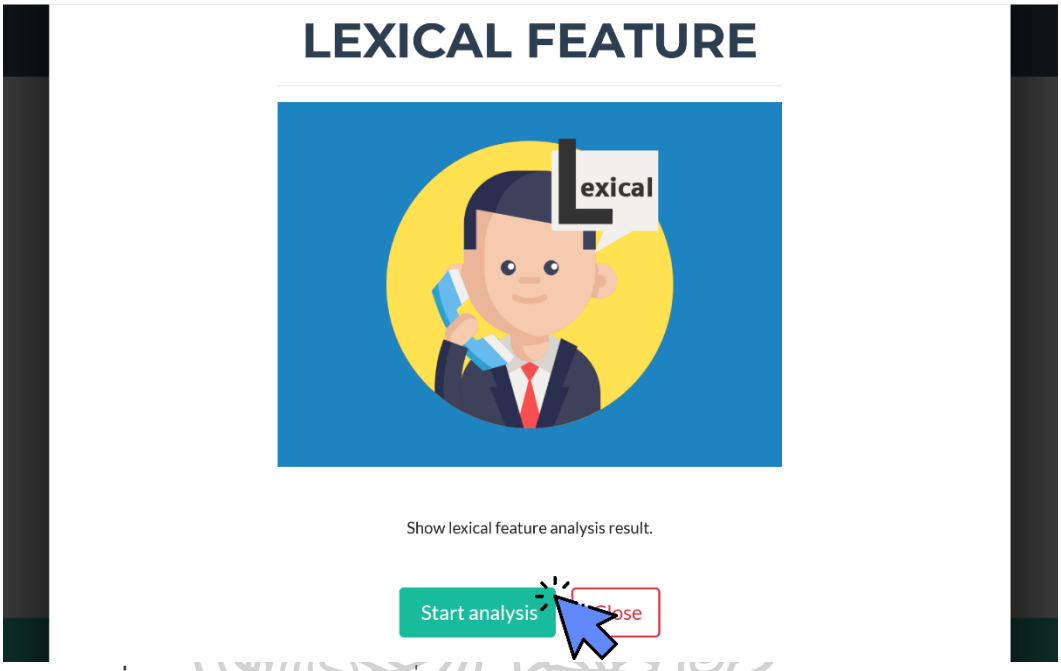
รูปที่ 4.8 แสดงเมนูกลุ่มของคุณลักษณะที่ใช้ในการวิเคราะห์โครงสร้างบทคัดย่อ



รูปที่ 4.9 แสดงส่วนบทคัดย่อดั้งเดิมที่ผู้ใช้นำเข้าสู่ระบบ

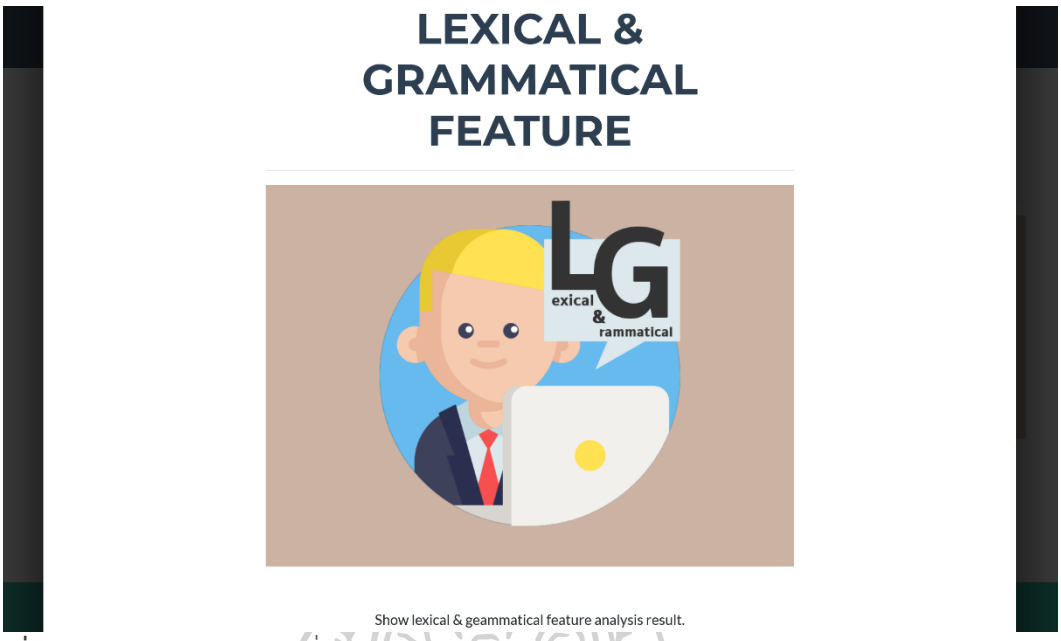
ขั้นตอนที่ 5 ยืนยันการเข้าสู่กระบวนการวิเคราะห์โครงสร้าง

จากรูปที่ 4.8 เมื่อผู้ใช้เลือกกลุ่มของคุณลักษณะที่ใช้ในการวิเคราะห์โครงสร้างของบทคัดย่อแบบใดแบบหนึ่งแล้วระบบจะแสดงหน้าจอยืนยันการวิเคราะห์โครงสร้าง ดังรูปที่ 4.10 – 4.12 ให้ผู้ใช้เลือกปุ่ม Start analysis เพื่อเข้าสู่กระบวนการวิเคราะห์โครงสร้างและแสดงผลการวิเคราะห์ต่อไป



รูปที่ 4.10 แสดงภาพหน้าจอเริ่มการวิเคราะห์โครงสร้างรูปแบบ Lexical features

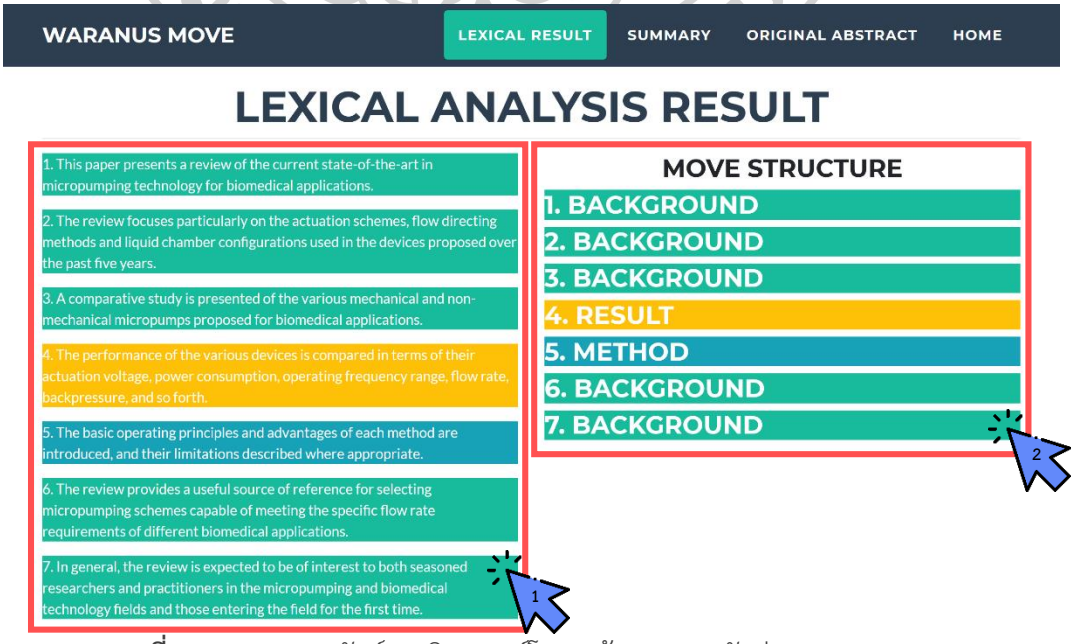
รูปที่ 4.11 แสดงภาพหน้าจอเริ่มการวิเคราะห์โครงสร้างรูปแบบ Grammatical and position features



รูปที่ 4.12 แสดงภาพหน้าจอเริ่มการวิเคราะห์โครงสร้างรูปแบบ Lexical, Grammatical and position features

ขั้นตอนที่ 6 แสดงผลการวิเคราะห์โครงสร้างของบทคัดย่อ

เมื่อผู้ใช้เลือกปุ่ม Start analysis ในขั้นตอนที่ 5 แล้วระบบจะเข้าสู่กระบวนกาวิเคราะห์โครงสร้างของบทคัดย่อที่ใช้โมเดลการจำแนกโครงสร้างของบทคัดย่อที่ผู้ใช้เลือก โดยผลการวิเคราะห์จากตัวอย่างเมื่อผู้ใช้เลือกรูปแบบการวิเคราะห์ Lexical features จะแสดงผลการวิเคราะห์ออกเป็น 2 ส่วนด้วยกันคือ ส่วนที่1 Lexical analysis result ส่วนที่2 Summary ส่วนที่3 Oirginal abstract



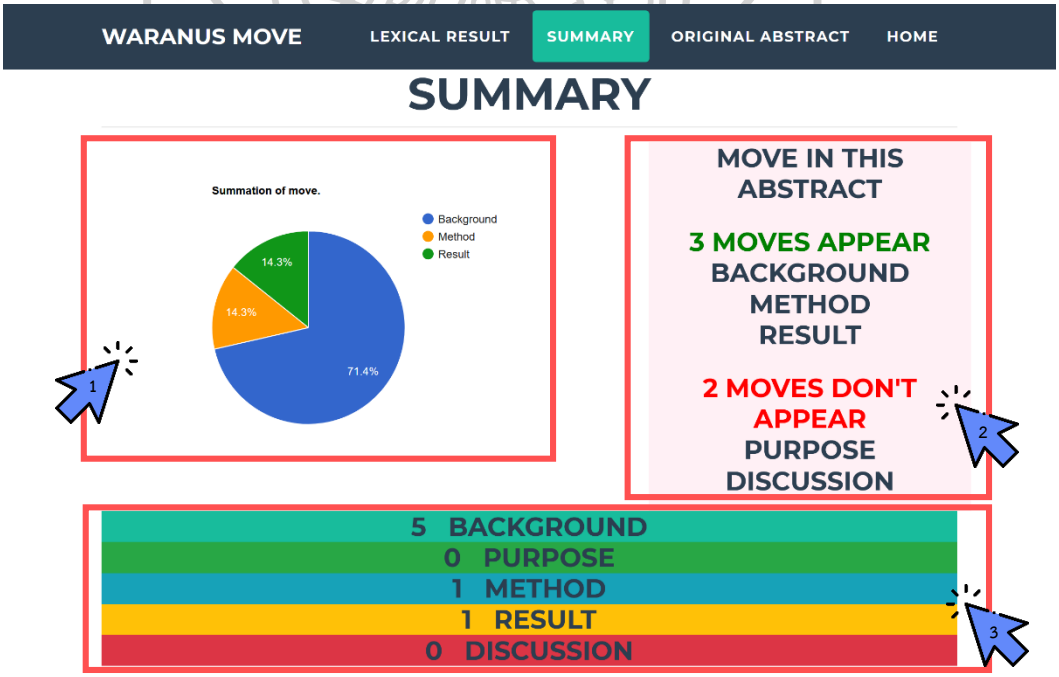
รูปที่ 4.13 แสดงผลลัพธ์การวิเคราะห์โครงสร้างของบทคัดย่อของ Lexical features

ส่วนที่ 1 Lexical analysis result ดังรูปที่ 4.13 จะแสดงผลการวิเคราะห์โครงสร้างออกเป็นสองส่วนส่วนแรกทางด้านซ้ายจะแสดงประโยคของบทคัดย่อ ซึ่งแต่ละประโยคจะแสดงแถบสีครอบคลุมประโยคที่แตกต่างกันไปแต่ละมูฟ และอีกส่วนหนึ่งทางด้านขวาคือส่วน Move structure จะแสดงโครงสร้างการเรียงลำดับของบทคัดย่อพร้อมแสดงแถบสีอีกด้วย

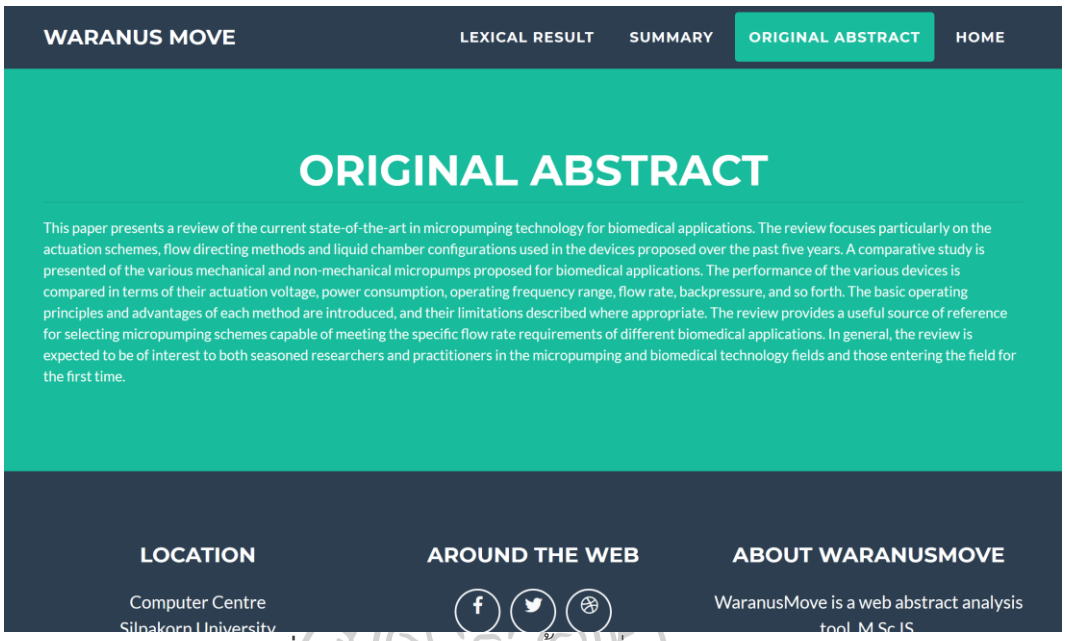
ส่วนที่ 2 Summary ดังรูปที่ 4.14 จะแสดงส่วนของการสรุปผลการวิเคราะห์โครงสร้างแบ่งออกเป็นสามส่วนด้วยกันดังนี้ ส่วนแรกทางด้านซ้ายแสดงกราฟสรุปจำนวนประโยคในแต่ละมูฟที่ปรากฏ พร้อมแสดงสัดส่วนของมูฟในรูปแบบเปอร์เซ็นต์ และเมื่อผู้ใช้นำตัวชี้ตำแหน่งวางไว้ที่ส่วนของกราฟจะแสดงจำนวนประโยคของแต่ละมูฟที่กรากฏในบทคัดย่อ ส่วนที่สองทางด้านขาวแสดงมูฟจำนวนมูฟที่ปรากฏและไม่ปรากฏอยู่ในบทคัดย่อ ส่วนสุดท้ายแสดงจำนวนของมูฟโดยมีแถบสีแยกตามแต่ละมูฟ

ส่วนที่ 3 Original abstract ดังรูปที่ 4.15 จะแสดงบทคัดย่อต้นฉบับดั้งเดิมแสดงให้ผู้ใช้ได้เปรียบเทียบกับกับผลลัพธ์ที่ได้จากการวิเคราะห์อีกด้วย

สรุปผลลัพธ์การวิเคราะห์โครงสร้างของบทคัดย่อมีการเรียงลำดับของมูฟดังนี้ Background>Background>Background>Result>Method>Background>Background ซึ่งมีมูฟที่ปรากฏ ได้แก่ Background จำนวน 5 ประโยคคิดเป็น 71.4% Method จำนวน 1 ประโยค คิดเป็น 14.3% และ Result จำนวน 1 ประโยคคิดเป็น 14.3% โดยไม่ปรากฏมูฟ Purpose และ Discussion ในบทคัดย่อ

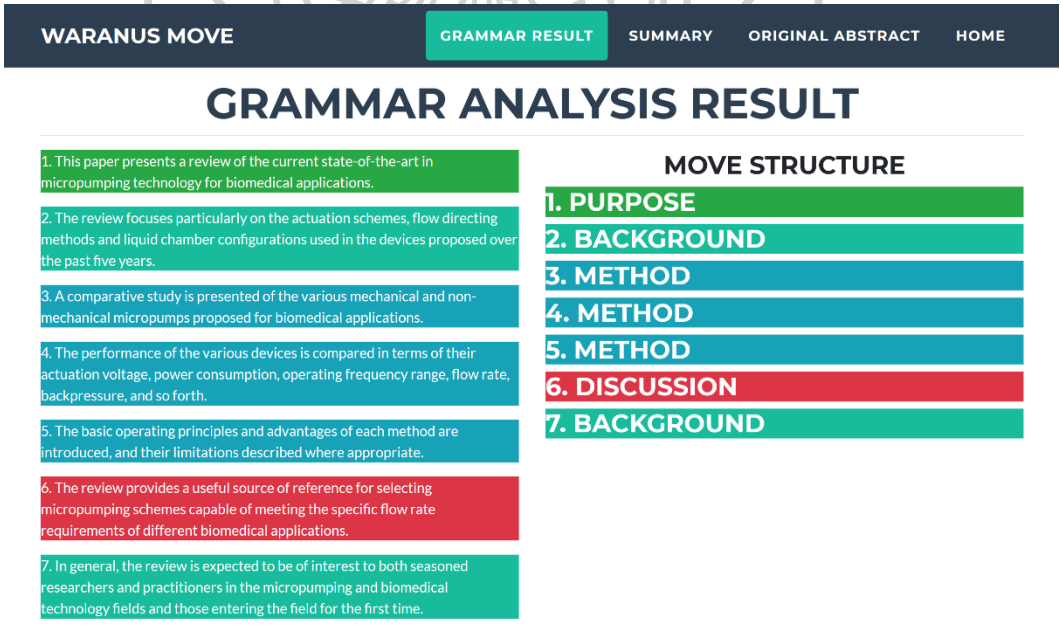


รูปที่ 4.14 แสดงผลลัพธ์ส่วนสรุปการวิเคราะห์โครงสร้างของบทคัดย่อของ Lexical features

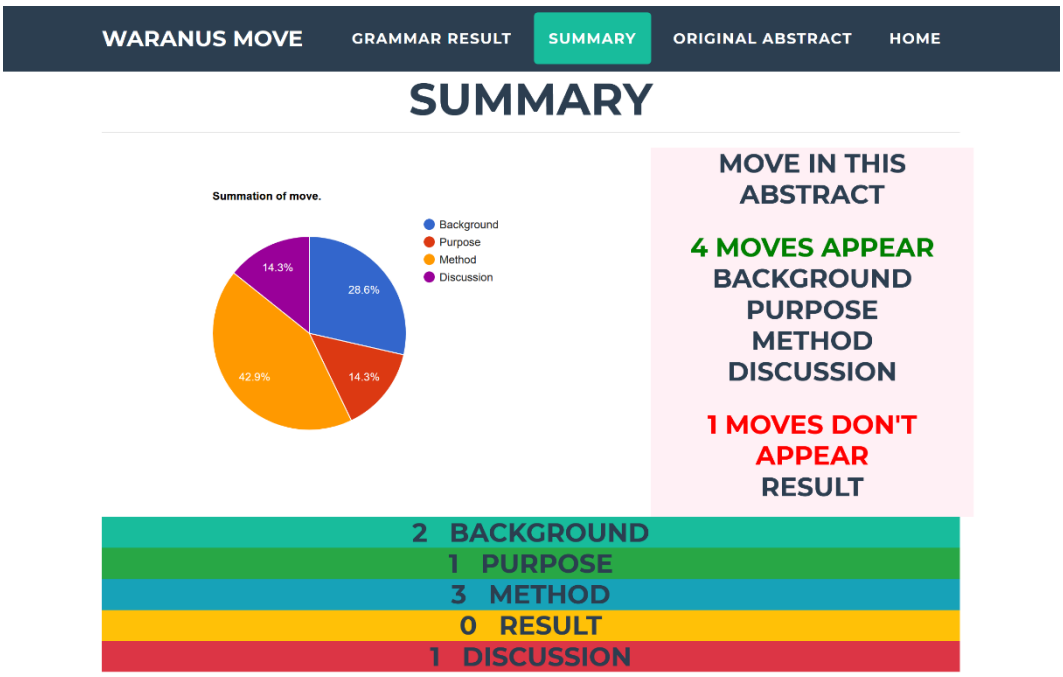


รูปที่ 4.15 แสดงบทคัดย่อดั้งเดิมที่ผู้ใช้นำเข้าสู่ระบบ

นอกจากการวิเคราะห์โครงสร้างของบทคัดย่อโดยใช้กลุ่มของคุณลักษณะแบบ Lexical features ดังตัวอย่างที่กล่าวมาข้างต้นแล้วนั้น ผู้ใช้งานยังสามารถเลือกการวิเคราะห์โครงสร้างของบทคัดย่อโดยใช้กลุ่มของคุณลักษณะแบบ Grammatical and position features และ Lexical, Grammatical and position features เพื่อดูผลลัพธ์ แสดงดังตัวอย่างรูปที่ 4.16 – 4.19

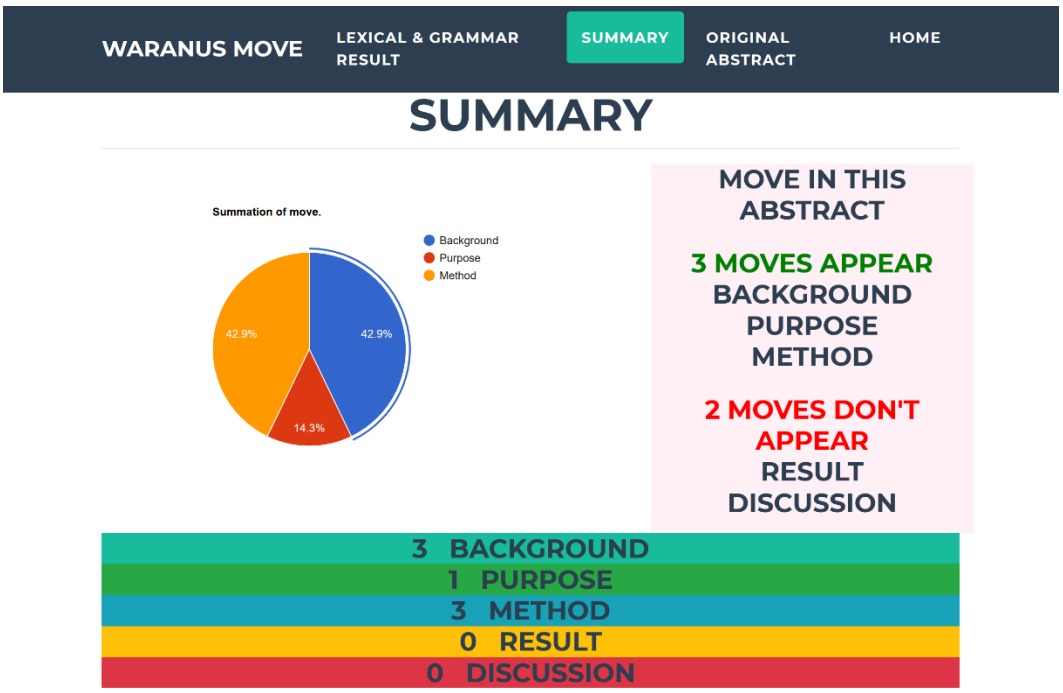


รูปที่ 4.16 แสดงผลลัพธ์การวิเคราะห์โครงสร้างของบทคัดย่อของ Grammatical and position features

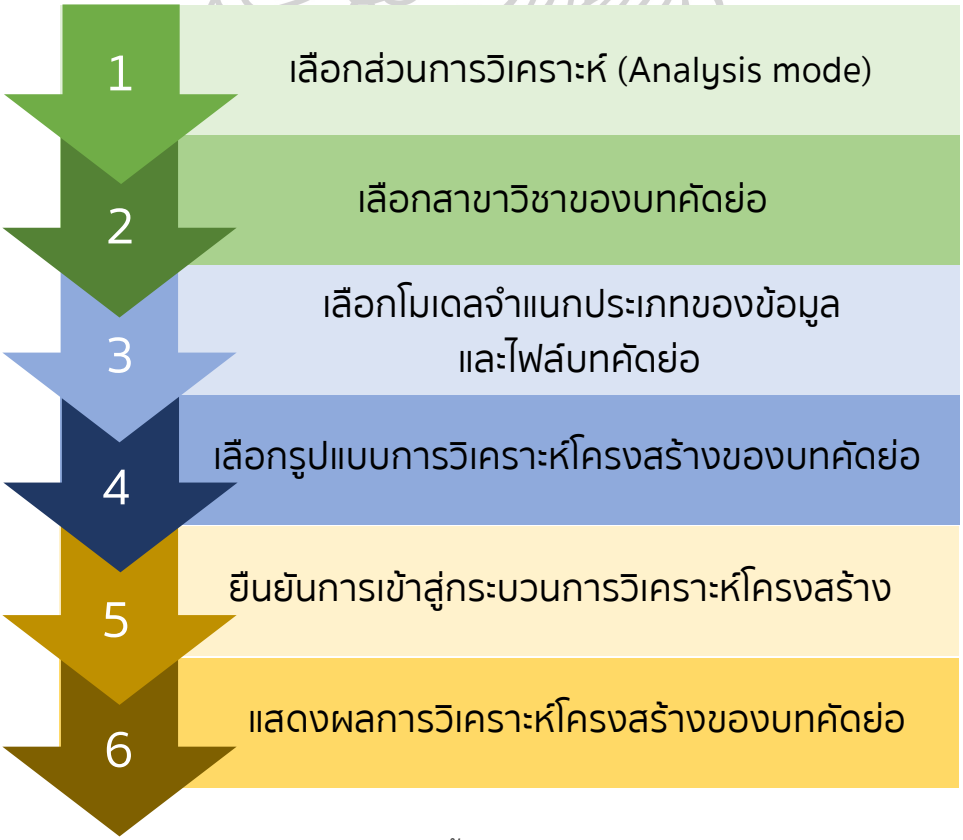


รูปที่ 4.17 แสดงผลลัพธ์ส่วนสรุปการวิเคราะห์โครงสร้างของบทคัดย่อของ Grammatical and position features

รูปที่ 4.18 แสดงผลลัพธ์การวิเคราะห์โครงสร้างของบทคัดย่อของ Lexical, Grammatical and position features



รูปที่ 4.19 แสดงผลลัพธ์ส่วนสรุปการวิเคราะห์โครงสร้างของบทคัดย่อของ Lexical, Grammatical and position features



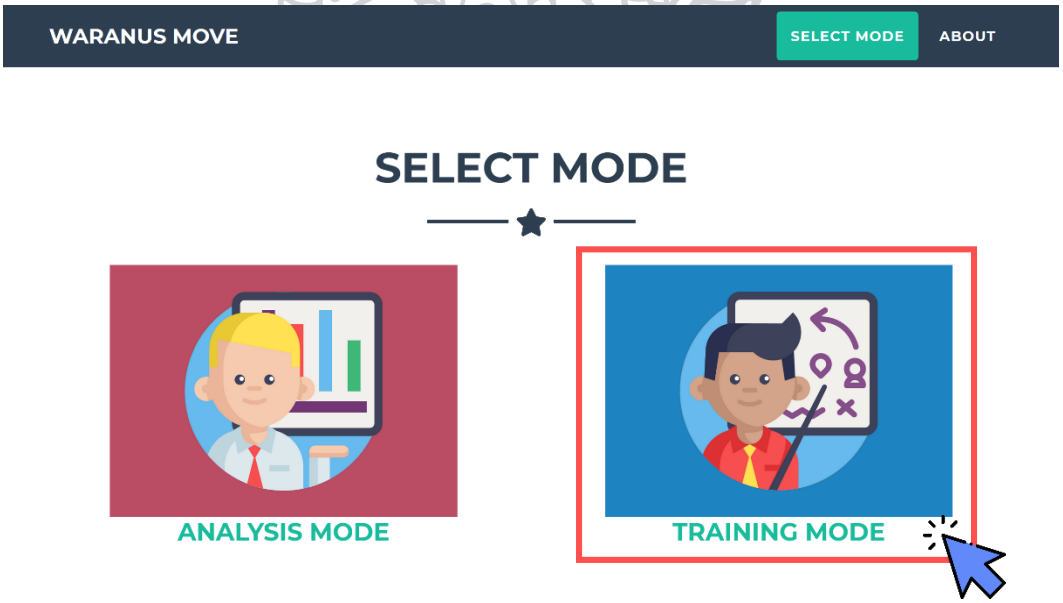
รูปที่ 4.20 สรุปขั้นตอนการทำงานส่วนการวิเคราะห์

4.1.2 ส่วนการฝึกฝน (Training mode)

ส่วนการฝึกฝน (Training mode) เป็นส่วนที่ผู้ใช้สามารถนำบทคัดย่อในสาขาวิชาต่างๆ เข้าสู่ระบบเพื่อฝึกฝนให้ระบบเรียนรู้โครงสร้างหรือมูฟของบทคัดย่อในสาขาวิชานั้นๆ ในรูปแบบของ Lexical features, Grammatical and position features และ Lexical, Grammatical and position features โดยใช้วิธีการเลือกสุ่มข้อมูลแบบ 10-Fold cross validation และสร้างโมเดลที่ใช้จำแนกโครงสร้างมูฟในสาขาวิชานั้นๆ ออกมาเพื่อใช้ในการวิเคราะห์โครงสร้างของบทคัดย่อต่อไป และแสดงผลลัพธ์ของการฝึกฝนของระบบในรูปแบบของกราฟที่แสดงค่าเฉลี่ยของประสิทธิภาพการฝึกฝน (Accuracy) โดยมีขั้นตอนกระบวนการทำงานของระบบส่วนการฝึกฝนดังต่อไปนี้

ขั้นตอนที่ 1 เลือกส่วนการฝึกฝน (Training mode)

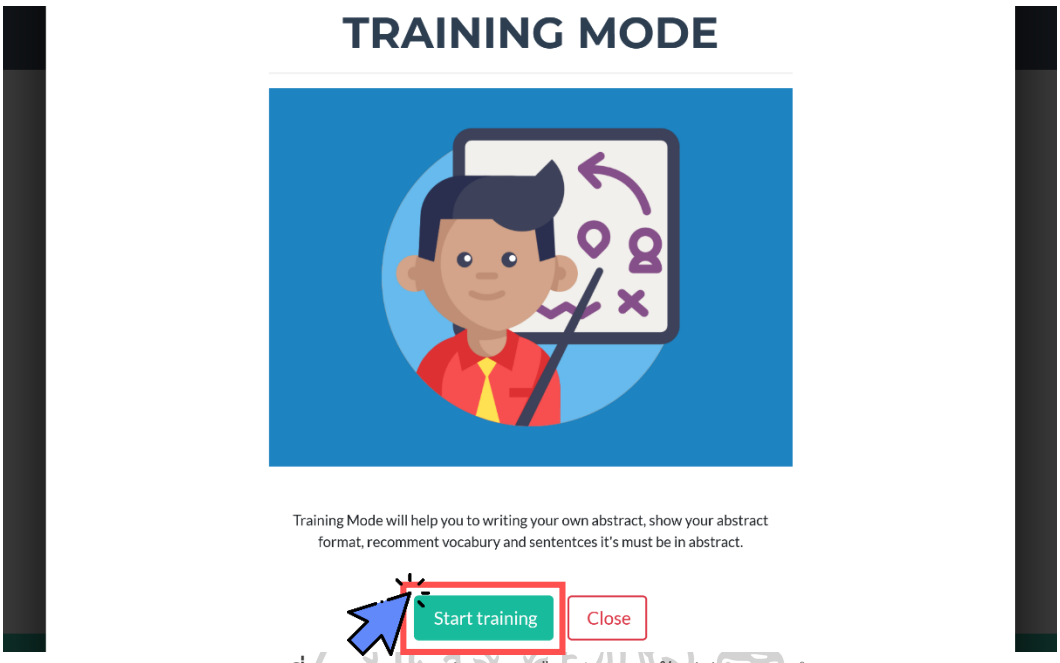
วิธีการเข้าสู่ขั้นตอนของการฝึกฝนเรียนรู้บทคัดย่อจากรูปที่ 4.21 ในเมนู Select mode เพื่อเข้าสู่ส่วนการฝึกฝนให้ผู้ใช้เลือก Training Mode



รูปที่ 4.21 แสดงการเลือกเมนูเพื่อเข้าสู่กระบวนส่วนการฝึกฝน

ขั้นตอนที่ 2 ยืนยันการเข้าสู่ส่วนการฝึกฝน

เมื่อผู้ใช้เลือกเมนู Training mode จากรูปที่ 4.21 แล้วระบบจะแสดงส่วนการยืนยันการเข้าสู่ส่วนการฝึกฝนดังรูปที่ 4.22 จากนั้นให้ผู้ใช้เลือกปุ่ม Start training เพื่อเข้าสู่ขั้นตอนการกรอกรายละเอียดที่จำเป็นในการฝึกฝนโมเดลการเรียนรู้โครงสร้างของบทคัดย่อต่อไป



รูปที่ 4.22 แสดงส่วนการยืนยันการเข้าสู่ส่วนการฝึกฝน

ขั้นตอนที่ 3 กรอกรายละเอียดการฝึกฝนโมเดลการจำแนกโครงสร้างของบทคัดย่อ

ในขั้นขั้นตอนที่ 3 นี้ระบบจะแสดงส่วนที่ผู้ใช้งานต้องกรอกรายละเอียดต่างๆ ที่จำเป็นต้องใช้ในการฝึกฝนโมเดลการจำแนกโครงสร้างของบทคัดย่อ โดยแบ่งออกเป็น 2 ส่วน คือส่วนการกรอกรายละเอียดสำหรับใช้ฝึกฝนโมเดล (Fill detail of new model.) และส่วนการอัพโหลดไฟล์เอกสารบทคัดย่อ ดังตัวอย่างในรูปที่ 4.23 มีรายละเอียดดังต่อไปนี้

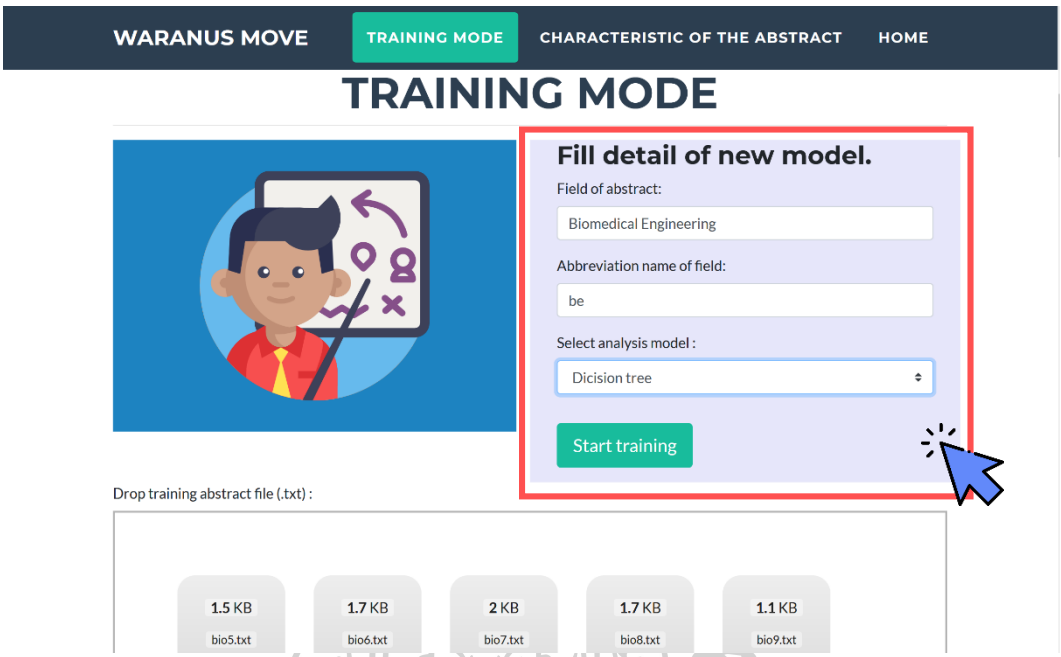
Field of abstract คือ ชื่อสาขาวิชาของบทคัดย่อ จากตัวอย่างคือ Biomedical Engineering

Abbreviation name of field คือ ชื่อย่อของสาขาวิชาของบทคัดย่อ จากตัวอย่างคือ be

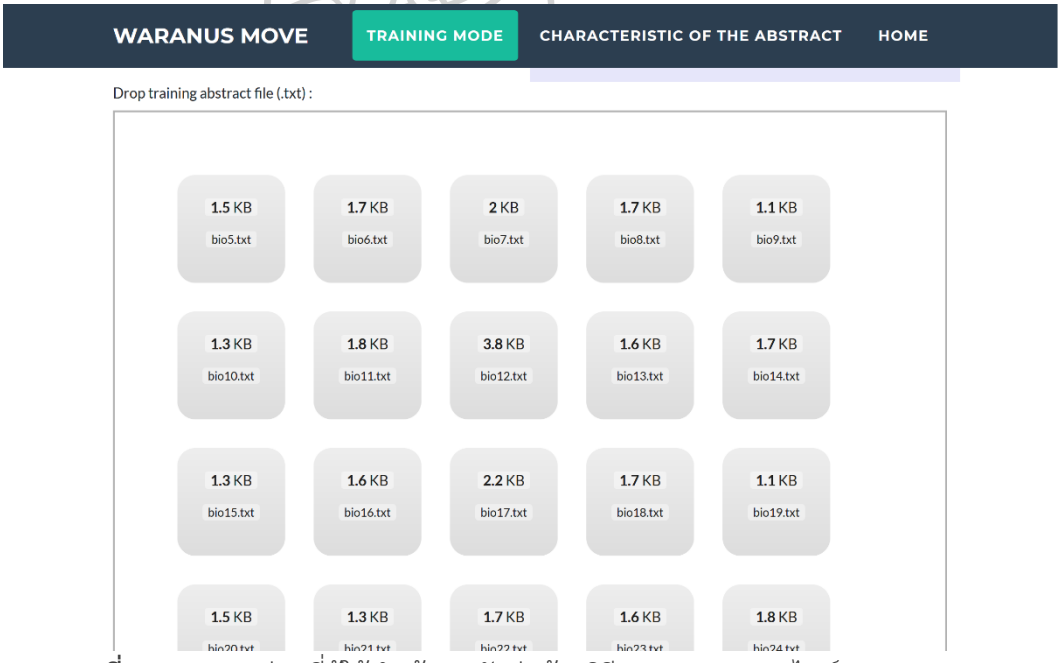
Select analysis model คือ เลือกการเรียนรู้ของเครื่องที่ต้องการใช้วิเคราะห์โครงสร้างบทคัดย่อ ได้แก่ Support Vector machine, Decision tree, Naïve bays และ Random forest จากตัวอย่างเลือก Decision tree

Start training คือ ปุ่มเริ่มต้นการวิเคราะห์โครงสร้างบทคัดย่อ

Drop training abstract file (.txt) คือ บริเวณสำหรับการลากและวางไฟล์บทคัดย่อ

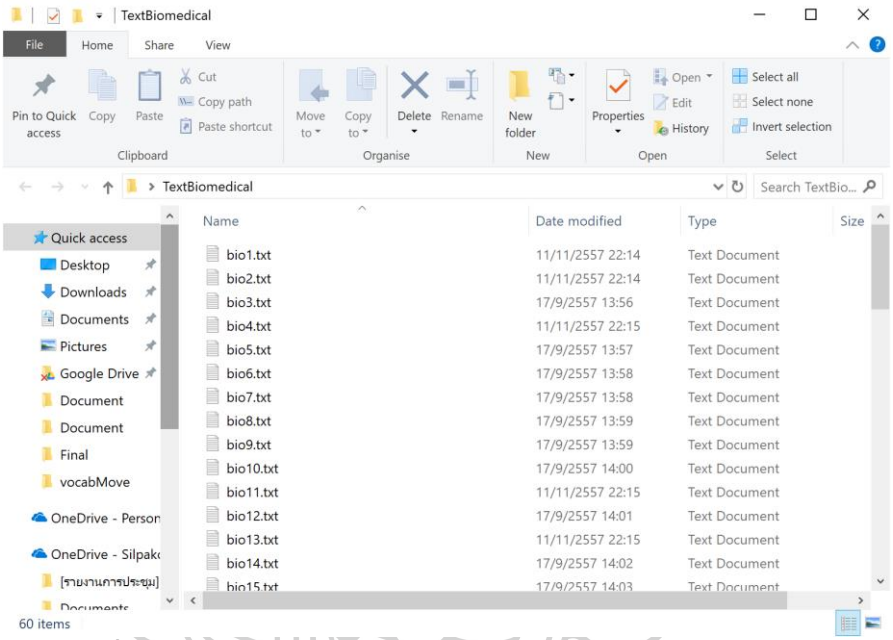


รูปที่ 4.23 แสดงส่วนของการกรอกรายละเอียดสำหรับการฝึกฝนโมเดลจำแนกโครงสร้างมูฟของบทคัดย่อ



รูปที่ 4.24 แสดงส่วนที่ผู้ใช้นำเข้าบทคัดย่อด้วยวิธีการลากและวางไฟล์เอกสาร Text

เมื่อผู้ใช้ลากและวางไฟล์บทคัดย่อดังในรูปที่ 4.24 แล้วไฟล์ที่ผู้ใช้นำเข้าเพื่อฝึกฝนการจำแนกโครงสร้างมูฟของบทคัดย่อทั้งหมดจะแทนด้วยรูปสี่เหลี่ยมสีเทา พร้อมชื่อไฟล์และขนาดของไฟล์ โดยที่ไฟล์เอกสารบทคัดย่อต้องเป็นไฟล์เอกสาร text ที่มีนามสกุลเป็น .txt ผู้ใช้ต้องตั้งชื่อไฟล์ดังนี้ คือ bio1.txt, bio2.txt, bio3.txt, bio4.txt, ... , bio60.txt ทั้งหมดจำนวน 60 ไฟล์ จากรูปที่ 4.25 คือไฟล์เอกสารบทคัดย่อ Text จำนวน 60 ไฟล์



รูปที่ 4.25 แสดงไฟล์เอกสารบทคัดย่อ Text สำหรับการฝึกฝนโมเดล

จากรูปที่ 4.26 แสดงลักษณะโครงสร้างของไฟล์เอกสารบทคัดย่อ Text ที่ใช้ในการฝึกฝนโมเดลจะต้องมีลักษณะดังนี้ แต่ละประโยคของบทคัดย่อต้องมีส่วนเฉลยของมูฟ เช่น [B] -> กำกับอยู่และตามด้วยประโยค ตัวอย่างประโยค Background คือ [B]-> The proposed method has been tested with both simulated and experimental data. โดยส่วนเฉลยของมูฟมีดังนี้

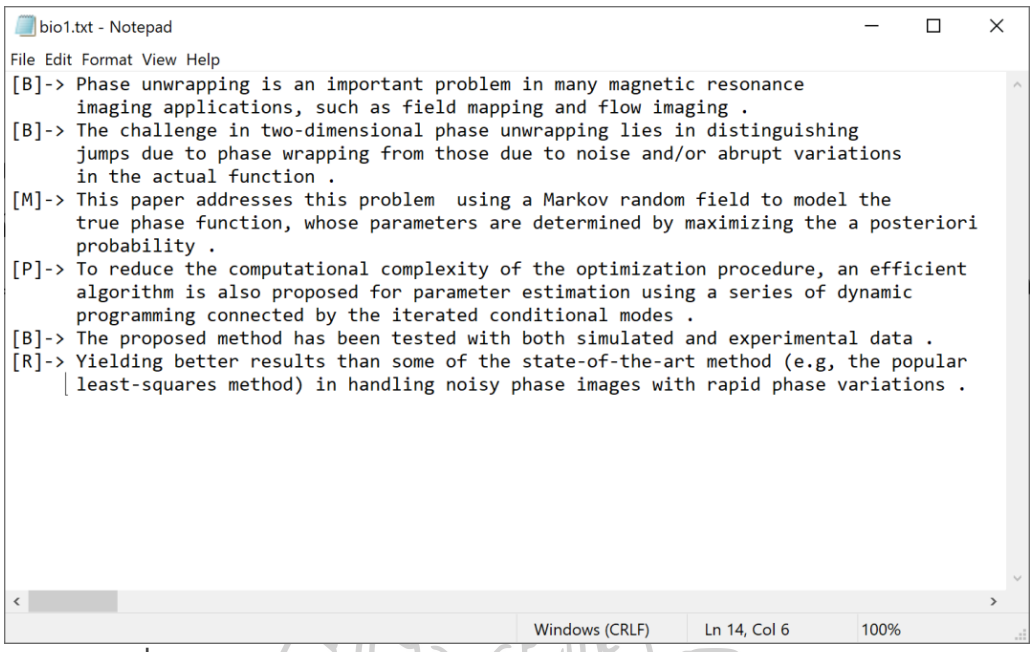
[B] -> คือ Background

[P] -> คือ Purpose

[M] -> คือ Method

[R] -> คือ Result

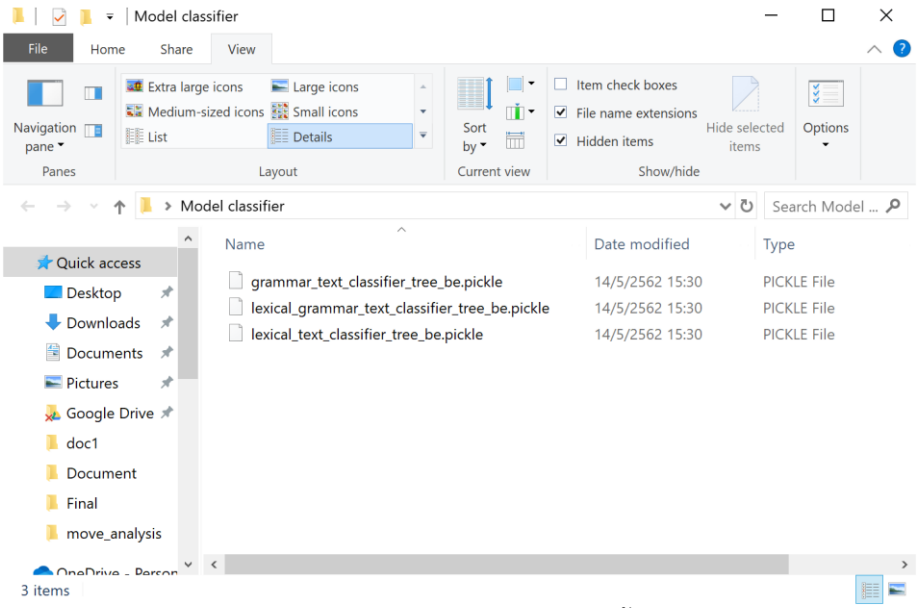
[D] -> คือ Discussion



รูปที่ 4.26 ลักษณะโครงสร้างของไฟล์เอกสาร Text สำหรับการฝึกฝนโมเดล

ขั้นตอนที่ 4 แสดงผลการฝึกฝนการจำแนกโครงสร้างของบทคัดย่อ

จากขั้นตอนที่ 3 เมื่อผู้ใช้งานกรอกรายละเอียด อัพโหลดไฟล์บทคัดย่อสำหรับการฝึกฝนโมเดลและเลือกปุ่ม Start training แล้วระบบจะเข้าสู่กระบวนการเรียนรู้โครงสร้างของบทคัดย่อและสร้างโมเดลจำแนกมูฟของบทคัดย่อของทั้ง 3 กลุ่มคุณลักษณะ คือ Lexical features, Grammatical and position features และ Lexical, Grammatical and position features ซึ่งโมเดลที่สร้างขึ้นจะนำไปใช้ในส่วนการวิเคราะห์โครงสร้างของบทคัดย่อ ดังรูปที่ 4.27



รูปที่ 4.27 โมเดลจำแนกมูฟของบทคัดย่อของทั้ง 3 กลุ่มคุณลักษณะ

โดยแสดงผลการเรียนรู้โครงสร้างของบทคัดย่อดังรูปที่ 4.28 แบ่งออกเป็น 2 ส่วน ดังนี้

ส่วน Trainging model was finished. แสดงผลลัพธ์การจำแนกโครงสร้างขอบทคัดย่อในรูปแบบของกราฟแท่งแสดงค่าเฉลี่ยประสิทธิภาพของการจำแนกโครงสร้างของบทคัดย่อทั้ง 3 กลุ่มของคุณลักษณะได้แก่ Lexical features, Grammatical and position features และ Lexical, Grammatical and position features โดยมีค่าเฉลี่ยประสิทธิภาพ 0.778, 0.555 และ 0.835 ตามลำดับซึ่งคุณลักษณะ Lexical, Grammatical and position features ให้ประสิทธิที่ดีที่สุด

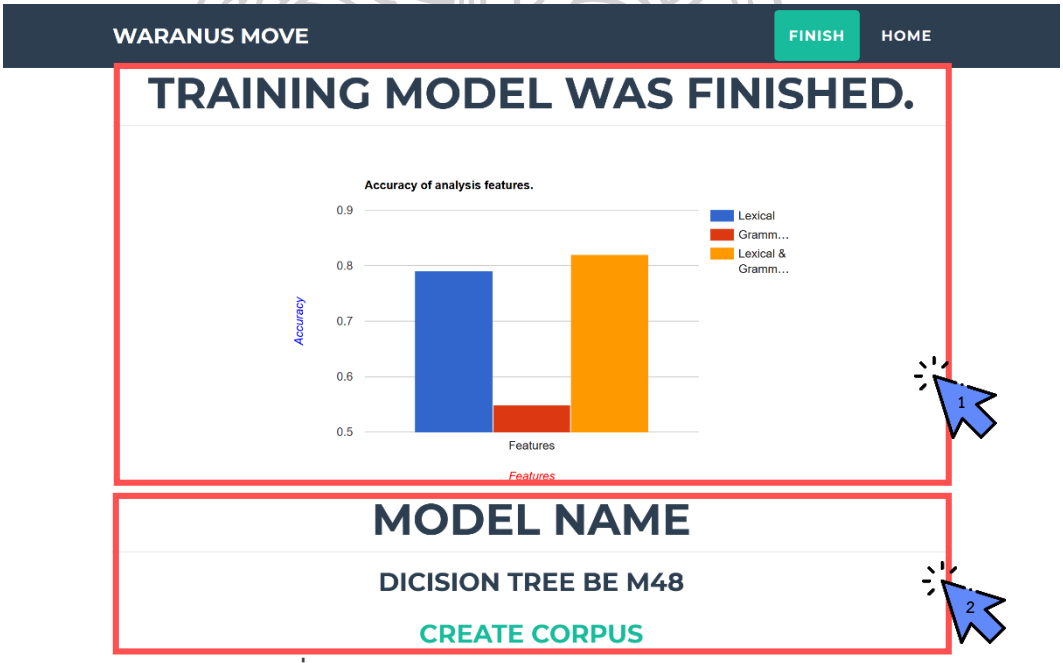
ส่วน Model name แสดงชื่อโมเดลที่ระบบสร้างขึ้น เพื่อนำไปใช้ในส่วนการวิเคราะห์โครงสร้างของบทคัดย่อ จากตัวอย่างโมเดลที่ระบบสร้างขึ้นมีชื่อว่า Decision tree be m48 โดยที่

Decision tree คือ ชื่อประเภทการเรียนรู้ของเครื่องที่ใช้เรียนรู้โครงสร้างของบทคัดย่อ จากตัวอย่างคือ Decision tree

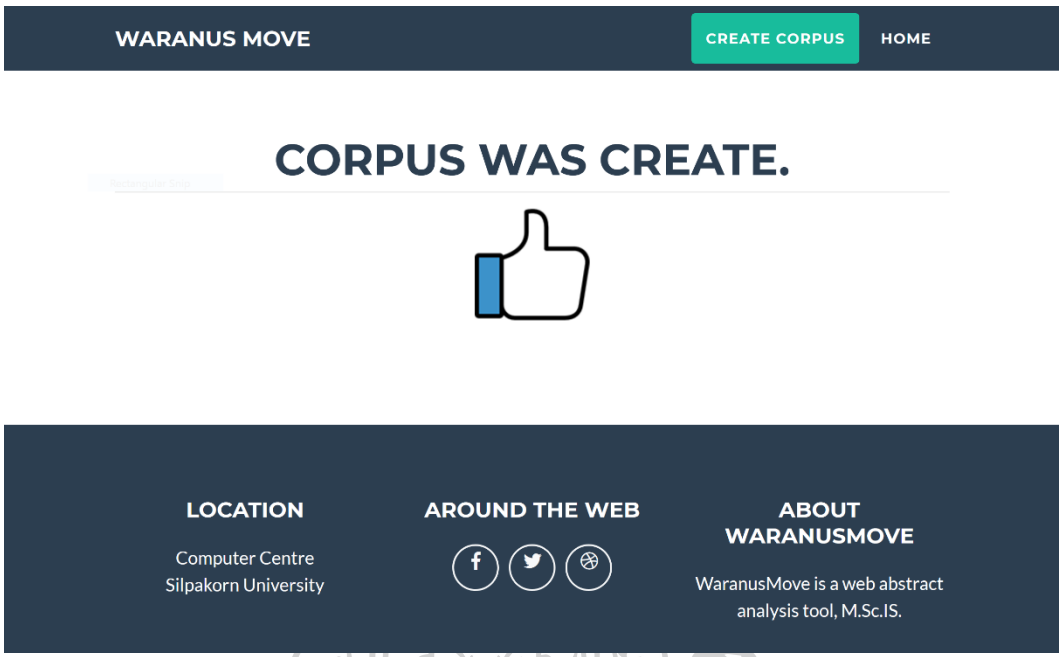
be คือ ชื่อย่อสาขาวิชาของบทคัดย่อ

m48 คือ รหัสประจำตัวของโมเดลจำแนกประเภทข้อมูล

อีกทั้งผู้ใช้ยังสามารถเพิ่มคลังคำของผู้ใช้เองได้จากบทคัดย่อที่ผู้ใช้นำเข้าสู่ระบบไว้ในขั้นตอนที่ 3 โดยเลือกจากเมนู Create corpus ระบบจะทำการสร้างคลังคำเก็บไว้ภายในระบบ เมื่อระบบสร้างคลังคำเรียบร้อยแล้วจะแสดงข้อความแจ้งเตือนผู้ใช้ ดังรูปที่ 4.28



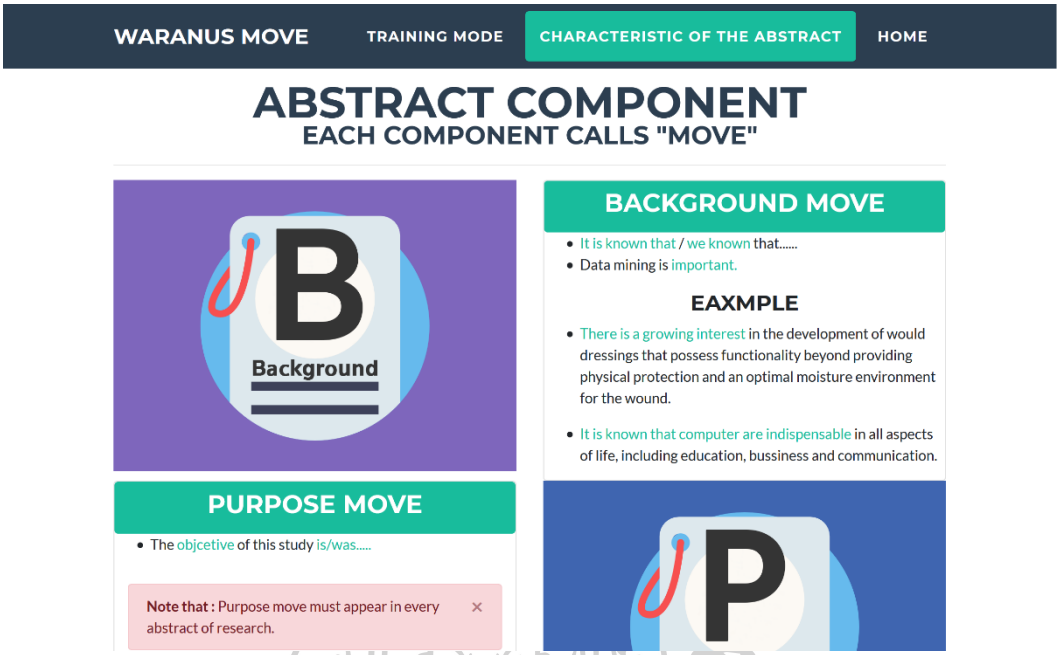
รูปที่ 4.28 แสดงผลลัพธ์ของการจำแนกโครงสร้างของบทคัดย่อ



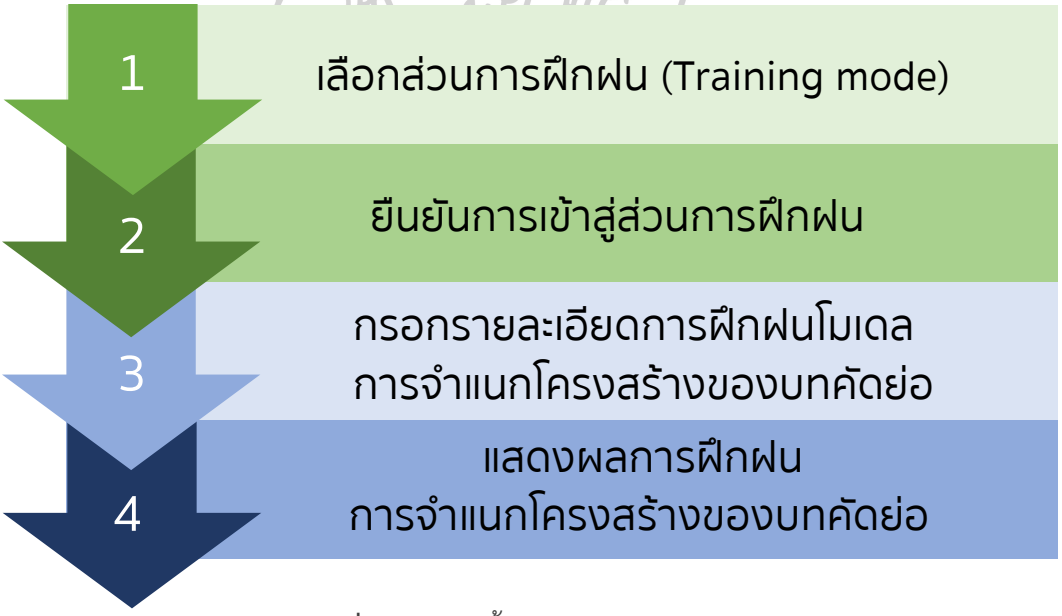
รูปที่ 4.29 แสดงระบบสร้างคลังคำเรียบร้อยแล้ว

ในส่วนการฝึกฝนยังมีส่วนที่อธิบายรายละเอียดโครงสร้างของบทคัดย่อ และมูฟต่างๆ เพื่อให้ผู้ใช้ได้ศึกษา ดังรูปที่ 4.30 – 4.31

รูปที่ 4.30 แสดงส่วนอธิบายรายละเอียดโครงสร้างของบทคัดย่อ



รูปที่ .4.32 แสดงส่วนอธิบายรายละเอียดโครงสร้างของมูฟ

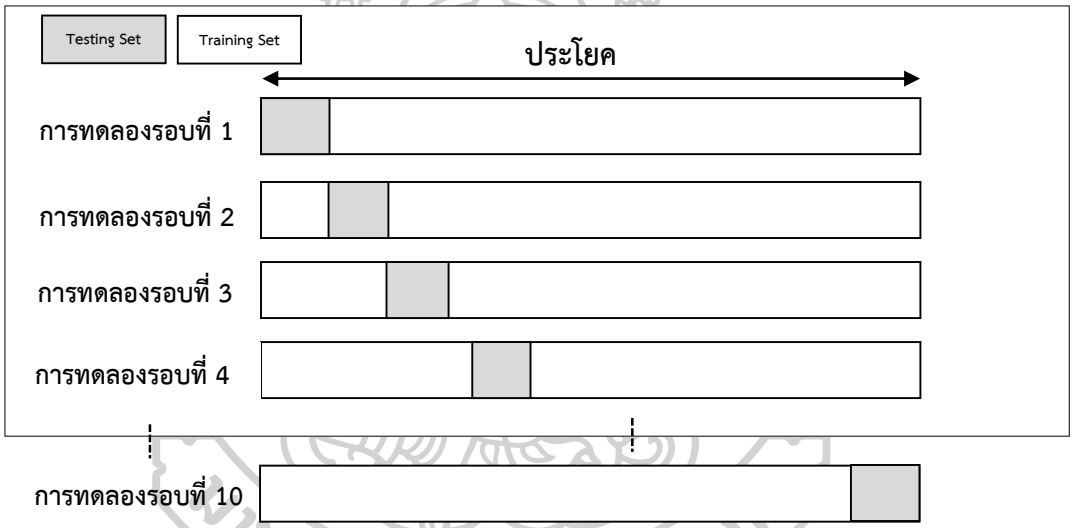


รูปที่ 4.33 สรุปขั้นตอนการทำงานของส่วนการฝึกฝน

4.2 การทดลอง

4.2.1 การวิเคราะห์โครงสร้างระดับประโยค

การวิเคราะห์โครงสร้างระดับประโยค เลือกใช้วิธีการทดลอง**เลือกสุ่มข้อมูล**แบบ 10-Fold cross validation โดยใช้บทความต้นแบบจำนวน 60 บทความ ซึ่งมีจำนวนประโยคทั้งหมด 528 ประโยค โดยจะทำการแบ่งข้อมูลออกเป็นส่วนๆ โดยแบ่งข้อมูลส่วนหนึ่งเป็นข้อมูลชุดฝึกสอน และข้อมูลส่วนที่หลือนำมาเป็นข้อมูลชุดทดสอบ โดยผลลัพธ์จากการทดลองที่ได้จากการสุ่มประโยคโดยไม่สนใจว่าแต่ละประโยคจะอยู่บทความเดียวกันหรือไม่ การทดลอง**เลือกสุ่มข้อมูล**แบบ 10-Fold cross validation ข้อมูลทุกข้อมูลจะถูกนำมาใช้เป็นข้อมูลชุดฝึกสอน และข้อมูลชุดทดสอบซึ่งมีขั้นตอนดังภาพที่ 4.26



รูปที่ 4.26 แสดงตัวอย่างวิธีการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation

- **รอบที่ 1** ประโยคที่ 1 – 53 เป็นข้อมูลชุดทดสอบ และประโยคที่ 54 – 528 เป็นข้อมูลชุดฝึกสอน
- **รอบที่ 2** ประโยคที่ 54 – 106 เป็นข้อมูลชุดทดสอบ และประโยคที่ 1 – 53, 107 – 528 เป็นข้อมูลชุดฝึกสอน
- **รอบที่ 3** ประโยคที่ 107 - 159 เป็นข้อมูลชุดทดสอบ และประโยคที่ 1 – 106, 160 – 528 เป็นข้อมูลชุดฝึกสอน
- **รอบที่ 4** ประโยคที่ 160 - 212 เป็นข้อมูลชุดทดสอบ และประโยคที่ 1 – 159, 213 – 528 เป็นข้อมูลชุดฝึกสอน

และทำต่อไปจนกระทั่งครบ 10 รอบ

4.2.2 การทดสอบการจำแนกประเภทข้อมูลด้วย Support vector machine และ Decision tree classify

การทดสอบการจำแนกประเภทข้อมูลด้วย Support vector machine และ Decision tree classify เป็นการนำเทคนิคการเรียนรู้ด้วยเครื่อง (Machine learning) ทั้งสองแบบ ซึ่งเป็นไลบรารี่ของ Scikit-learn มาทำการวิเคราะห์เพื่อเปรียบเทียบประสิทธภาพในการจำแนกประเภทของข้อมูล โดยใช้**วิธีการทดลองเลือกสุ่มข้อมูลแบบ** 10-Fold cross validation จากบทความต้นแบบจำนวน 60 บทความ จำนวน 528 ประโยค เพื่อทดสอบว่าการจำแนกประเภทของข้อมูลด้วยการเรียนรู้ด้วยเครื่องแบบใดที่ให้ประสิทธิภาพในการจำแนกประเภทของข้อมูลได้ดีกว่า

4.3 Lexical features

การวิเคราะห์ด้วย Lexical features แบ่งการทำงานออกเป็น 2 ส่วน คือ การวิเคราะห์โดยให้ระบบเรียนรู้จากบทความต้นแบบสาขาวิศวกรรมชีวเวชที่ได้รับการแบ่งมูฟจากนักภาษาศาสตร์แล้ว และการวิเคราะห์โดยเปรียบเทียบกับคลังคำศัพท์

การวิเคราะห์โดยให้ระบบเรียนรู้จากบทความต้นแบบสาขาวิศวกรรมชีวเวชเป็นการนำบทความที่ได้รับการแบ่งมูฟจากนักภาษาศาสตร์เข้าสู่ระบบ เพื่อวิเคราะห์คำ กลุ่มคำ และศึกษาโครงสร้างของปแต่ละมูฟ จึงได้นำเทคนิคการประมวลผลภาษาธรรมชาติมาสร้างเป็นกฎสำหรับการวิเคราะห์คำ ชนิดของคำ และกลุ่มคำ โดยเปรียบเทียบจากค่าความถี่ของคำ ชนิดคำ และคำที่มักเกิดร่วมกันในประโยคของแต่ละมูฟจากบมความต้นแบบ

การวิเคราะห์โดยเปรียบเทียบกับคลังคำศัพท์ ได้นำเทคนิคการประมวลผลภาษาธรรมชาติมาใช้ในการแปลงคำศัพท์ให้อยู่ในรูปดั้งเดิมก่อนนำมาวิเคราะห์ โดยสร้างกฎในการวิเคราะห์คำ กลุ่มคำ และรูปแบบของประโยค โดยเปรียบเทียบจากคลังคำศัพท์ที่มีการแยกประเภทของมูฟแล้ว

4.3.1 ผลการทดลองการวิเคราะห์โครงสร้างระดับประโยค

การวิเคราะห์ระดับประโยคด้วย Support vector machine ของ Lexical features

ตารางที่ 4.1 แสดงผลการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ด้วย Support vector machine ของ Lexical features

K	Background		Purpose		Method		Result		Discussion		All true	All sent	Accuracy (%)
	sent	all	sent	all	sent	all	sent	all	sent	all			
1	3	8	3	4	11	14	9	20	7	7	33	53	62.26
2	11	17	0	0	10	14	10	19	1	3	32	53	60.38
3	6	17	1	2	7	9	19	22	3	3	36	53	67.92
4	3	10	1	1	12	16	15	19	4	7	35	53	66.04
5	9	18	2	2	8	10	11	19	3	4	33	53	62.26
6	10	13	0	2	8	14	13	23	1	1	32	53	75.47
7	8	14	2	2	16	20	13	13	1	4	40	53	75.47
8	12	18	2	2	13	16	9	15	1	2	37	53	69.81
9	9	17	4	4	8	9	11	14	1	2	33	53	63.46
10	10	16	4	4	9	12	12	18	2	2	37	53	71.15

จากตารางที่ 4.1 วิธีการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ในการแบ่งข้อมูลทดสอบ จะพบว่าในช่วงค่า K เท่ากับ 7 มีค่าความถูกต้องที่ 75.47% ซึ่งมีค่าสูงที่สุด จึงนำเอาโมเดลการจำแนกประเภทของข้อมูล ที่ได้จากช่วงค่า K นี้เป็นตัวทดสอบจำแนกประเภทของข้อมูลบทคัดย่อต้นฉบับจำนวน 60 บทความ ผลลัพธ์แสดงได้ดังนี้

ตารางที่ 4.2 แสดงการวิเคราะห์การจำแนกโครงสร้างของบทคัดย่อต้นฉบับจากโมเดลที่ได้รับการฝึกสอนด้วย Support vector machine ที่ค่า K เท่ากับ 7 ของ Lexical features

Actual / Predict	Background	Purpose	Method	Result	Discussion
Background	**84**	0	5	7	1
Purpose	16	**23**	10	3	0
Method	20	0	**117**	11	0
Result	17	0	5	**136**	3
Discussion	11	1	2	20	**36**
All sent	148	24	139	177	40
Accuracy (%)	56.75	95.84	84.17	76.83	90

จากตารางที่ 4.2 จะพบว่าค่าความถูกต้องของการวิเคราะห์โครงสร้างระดับประโยคอยู่ที่ 75% โดยที่ค่าความถูกต้องของแต่ละมูฟมีค่าดังนี้

- Background 56.75%
- Purpose 95.84%
- Method 84.17%
- Result 76.83%
- Discussion 90%

การวิเคราะห์ระดับประโยคด้วย Decision tree classifier ของ Lexical features

ตารางที่ 4.3 แสดงผลการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ด้วย Decision tree classifier ของ Lexical features

K	Background		Purpose		Method		Result		Discussion		All true	All sent	Accuracy (%)
	sent	all	sent	all	sent	all	sent	all	sent	all			
1	7	9	3	4	15	19	16	17	3	4	44	53	83.02
2	8	9	3	3	12	15	16	18	7	8	47	53	88.68
3	7	14	3	5	13	16	8	12	4	6	35	53	66.04
4	12	17	3	7	9	11	11	12	5	6	40	53	75.47
5	4	4	4	7	14	19	14	19	4	4	40	53	75.47
6	9	10	5	8	13	16	13	17	2	2	42	53	79.25
7	7	8	2	2	13	15	17	20	5	8	44	53	83.02
8	13	15	1	1	13	17	15	18	2	2	44	53	83.02
9	8	9	3	4	9	13	14	21	4	5	38	53	73.08
10	1	3	3	5	12	17	9	13	11	14	36	53	69.23

จากตารางที่ 4.3 **วิธีการทดลองเลือกสุ่มข้อมูลแบบ** 10-Fold cross validation ในการแบ่งข้อมูลทดสอบ จะพบว่าในช่วงค่า K เท่ากับ 2 มีค่าความถูกต้องที่ 88.68% ซึ่งมีค่าสูงที่สุด จึงนำเอาโมเดลการจำแนกประเภทของข้อมูลที่ได้จากช่วงค่า K นี้เป็นตัวทดสอบจำแนกประเภทของข้อมูลบทคัดย่อต้นฉบับจำนวน 60 บทความ ผลลัพธ์แสดงได้ดังนี้

ตารางที่ 4.4 แสดงการวิเคราะห์การจำแนกโครงสร้างของบทคัดย่อต้นฉบับจากโมเดลที่ได้รับการฝึกสอนด้วย Decision tree classifier ที่ค่า K เท่ากับ 2 ของ Lexical features

Actual / Predict	Background	Purpose	Method	Result	Discussion
Background	**89**	0	6	2	0
Purpose	2	**43**	4	3	0
Method	2	1	**139**	6	0
Result	3	0	7	**150**	1
Discussion	2	2	2	6	**58**
All sent	98	46	158	167	59
Accuracy (%)	90.81	93.47	87.97	89.82	98.3

จากตารางที่ 4.4 จะพบว่าค่าความถูกต้องของการวิเคราะห์โครงสร้างระดับประโยคอยู่ที่ 90.71% โดยที่ค่าความถูกต้องของแต่ละมูฟมีค่าดังนี้

- Background 90.81%
- Purpose 93.47%
- Method 87.97%
- Result 89.82%
- Discussion 98.3%

สรุปการทดลองด้วยวิธีการเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ของคุณลักษณะ Lexical features ด้วย Decistion tree classifier จะให้ค่าความถูกต้องเท่ากับ 88.68% ซึ่งให้ค่าที่ดีกว่าการใช้ SVM ในการจำแนกมูฟ โดย SVM ให้ค่าความถูกต้องเท่ากับ 75.47%

4.3.2 การเปรียบเทียบการจำแนกประเภทของข้อมูลด้วย Support vector machine และ Decision tree classifier ของ Lexical features

ตารางที่ 4.5 แสดงผลการจำแนกประเภทของข้อมูลด้วย Support vector machine และ Decision tree classifier

Actual / Predict	Background		Purpose		Method		Result		Discussion	
	SVM	DTs	SVM	DTs	SVM	DTs	SVM	DTs	SVM	DTs
Background	**84**	**89**	0	0	5	6	7	2	1	0
Purpose	16	2	**23**	**43**	10	4	3	3	0	0
Method	20	2	0	1	**117**	**139**	11	6	0	0
Result	17	3	0	0	5	7	**136**	**150**	3	1
Discussion	11	2	1	2	2	2	20	6	**36**	**58**
All sent	148	98	24	46	139	158	177	167	40	59
Accuracy (%)	56.75	90.81	95.84	93.47	84.17	87.97	76.83	89.82	90	98.3

จากตารางที่ 4.5 จะพบว่าการจำแนกประเภทของข้อมูลด้วย Support vector machine ให้ค่าความถูกต้องเท่ากับ 75% และการจำแนกประเภทของข้อมูลด้วย Decision tree classifier ให้ค่าความถูกต้องเท่ากับ 90.71%

4.4 Grammatical and position features

การวิเคราะห์ด้วย Grammatical and position features ได้จากการเรียนรู้โครงสร้างของไวยกรณ์จากบทความต้นแบบสาขาวิศวกรรมชีวเวช และนำมาสร้างเป็นรูปแบบ (Pattern) ด้วย Regular expression รวมถึงการใช้ Position features มาตรวจสอบตำแหน่งของแต่ละประโยคในบทความ

4.4.1 ผลการทดลองการวิเคราะห์โครงสร้างระดับประโยค

การวิเคราะห์ระดับประโยคด้วย Support vector machine ของ Grammatical and position features

ตารางที่ 4.6 แสดงผลการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ด้วย Support vector machine ของ Grammatical and position features

K	Background		Purpose		Method		Result		Discussion		All true	All sent	Accuracy (%)
	sent	all	sent	all	sent	all	sent	all	sent	all			
1	6	9	2	2	13	22	7	16	4	4	32	53	60.38
2	3	5	0	5	8	12	9	26	3	5	23	53	43.40
3	6	9	2	5	13	20	9	17	2	2	32	53	60.38
4	4	8	2	6	7	13	13	18	5	8	31	53	58.49
5	5	5	4	6	6	21	10	19	2	2	27	53	50.94
6	6	11	2	4	7	18	13	16	3	4	31	53	58.49
7	9	12	2	3	8	19	7	12	7	7	33	53	62.26
8	11	16	0	0	10	20	5	12	4	5	30	53	56.60
9	5	6	2	7	5	16	9	20	3	3	24	53	46.15
10	7	8	3	5	7	17	11	18	1	4	29	53	55.77

จากตารางที่ 4.6 วิธีการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ในการแบ่งข้อมูลทดสอบ จะพบว่าในช่วงค่า K เท่ากับ 7 มีค่าความถูกต้องที่ 62.26% ซึ่งมีค่าสูงที่สุด จึงนำเอาโมเดลการจำแนกประเภทของข้อมูลที่ได้จากช่วงค่า K นี้เป็นตัวทดสอบจำแนกประเภทของข้อมูลบทคัดย่อต้นฉบับจำนวน 60 บทความ ผลลัพธ์แสดงได้ดังนี้

ตารางที่ 4.7 แสดงการวิเคราะห์การจำแนกโครงสร้างของบทคัดย่อต้นฉบับจากโมเดลที่ได้รับการฝึกสอนด้วย Support vector machine ที่ค่า K เท่ากับ 7 ของ Grammatical and position features

Actual / Predict	Background	Purpose	Method	Result	Discussion
Background	**63**	11	13	10	0
Purpose	4	**27**	12	9	0
Method	14	5	**95**	33	1
Result	0	0	61	**93**	7
Discussion	0	5	12	13	**40**
All sent	81	48	193	158	48
Accuracy (%)	77.78	56.25	49.22	58.86	83.34

จากตารางที่ 4.7 พบว่าค่าความถูกต้องของการวิเคราะห์โครงสร้างระดับประโยคด้วย Grammatical and position features อยู่ที่ 60.22% โดยที่ค่าความถูกต้องของแต่ละมูฟมีค่าดังนี้

- Background 77.78%
- Purpose 56.25%
- Method 49.22%
- Result 58.86%
- Discussion 83.34%

การวิเคราะห์โครงสร้างระดับประโยคด้วย Decision tree classifier ของ Grammatical and position features

ตารางที่ 4.8 แสดงผลการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ด้วย Decision tree classifier ของ Grammatical and position features

K	Background		Purpose		Method		Result		Discussion		All true	All sent	Accuracy (%)
	sent	all	sent	all	sent	all	sent	all	sent	all			
1	8	10	0	6	7	17	9	12	6	11	30	53	56.60
2	5	13	2	4	14	19	8	14	2	3	31	53	58.49
3	3	6	3	9	10	15	8	15	5	8	29	53	54.72
4	6	6	3	9	7	15	5	14	6	9	27	53	50.94
5	5	9	1	3	7	11	16	25	3	5	35	53	60.38
6	9	13	2	2	8	12	8	21	3	5	30	53	56.60
7	5	7	3	7	5	24	7	13	7	13	27	53	39.62
8	9	10	4	9	6	12	7	14	4	8	30	53	56.60
9	5	12	1	4	4	11	6	16	3	9	19	53	36.54
10	6	8	3	4	12	17	9	17	5	6	35	53	67.31

จากตารางที่ 4.8 **วิธีการทดลองเลือกสุ่มข้อมูลแบบ** 10-Fold cross validation ในการแบ่งข้อมูลทดสอบ จะพบว่าในช่วงค่า K เท่ากับ 10 มีค่าความถูกต้องที่ 67.31% ซึ่งมีค่าสูงที่สุด จึงนำเอาโมเดลการจำแนกประเภทของข้อมูลที่ได้จากช่วงค่า K นี้เป็นตัวทดสอบจำแนกประเภทของข้อมูลบทคัดย่อต้นฉบับจำนวน 60 บทความ ผลลัพธ์แสดงได้ดังนี้

ตารางที่ 4.9 แสดงการวิเคราะห์การจำแนกโครงสร้างของบทคัดย่อต้นฉบับจากโมเดลที่ได้รับการฝึกสอนด้วย Decision tree classifier ของ Grammatical and position features ที่ค่า K เท่ากับ 10 ของ Grammatical and position features

Actual / Predict	Background	Purpose	Method	Result	Discussion
Background	**86**	4	3	4	0
Purpose	7	**42**	2	1	0
Method	2	4	**109**	30	3
Result	2	4	21	**131**	3
Discussion	1	2	3	13	**51**
All sent	98	56	138	179	57
Accuracy (%)	87.75	75.00	78.98	73.18	89.47

จากตารางที่ 4.9 พบว่าค่าความถูกต้องของการวิเคราะห์โครงสร้างระดับประโยคด้วย Grammatical and position features อยู่ที่ 79.35% โดยที่ค่าความถูกต้องของแต่ละมูฟมีค่าดังนี้

- Background 87.75%
- Purpose 75.00%
- Method 78.98%
- Result 73.18%
- Discussion 89.47%

สรุปการทดลองด้วยวิธีการเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ของคุณลักษณะ Grammatical and position features ด้วย Decistion tree classifier จะให้ค่าความถูกต้องเท่ากับ 67.31% ซึ่งให้ค่าที่ดีกว่าการใช้ SVM ในการจำแนกมูฟ โดย SVM ให้ค่าความถูกต้องเท่ากับ 62.26%

4.4.2 การเปรียบเทียบการจำแนกประเภทของข้อมูลด้วย Support vector machine และ Decision tree classifier ของ Grammatical and position features

ตารางที่ 4.10 แสดงผลการจำแนกประเภทของข้อมูลด้วย Support vector machine และ Decision tree classifier ของ Grammatical and position features

Actual / Predict	Background		Purpose		Method		Result		Discussion	
	SVM	DTs	SVM	DTs	SVM	DTs	SVM	DTs	SVM	DTs
Background	**63**	**86**	11	4	13	3	10	4	0	0
Purpose	4	7	**27**	**42**	12	2	9	1	0	0
Method	14	2	5	4	**95**	**109**	33	30	1	3
Result	0	2	0	4	61	21	**93**	**131**	7	3
Discussion	0	1	5	2	12	3	13	13	**40**	**51**
All sent	81	98	48	56	193	138	158	179	48	57
Accuracy (%)	77.78	87.75	56.25	75.00	49.22	78.98	58.86	73.18	83.34	89.47

จากตารางที่ 4.10 จะพบว่าการจำแนกประเภทของข้อมูลด้วย Support vector machine ให้ค่าความถูกต้องเท่ากับ 60.22%6 และการจำแนกประเภทของข้อมูลด้วย Decision tree classifier ให้ค่าความถูกต้องเท่ากับ 79.35%

4.5 Lexical, Grammatical and position features

การวิเคราะห์ด้วย Lexical, Grammatical and position features ทำการสร้างรูปแบบ (Pattern) ในการตรวจสอบประโยคด้วย Regular expression โดยนำคุณลักษณะของการวิเคราะห์ด้วย Lexical features และคุณลักษณะที่วิเคราะห์ด้วย Grammatical and position มาใช้วิเคราะห์บทคัดย่อร่วมกัน

4.5.1 ผลการทดลองการวิเคราะห์โครงสร้างระดับประโยค

การวิเคราะห์ระดับประโยคด้วย Support vector machine ของ Lexical, Grammatical and position features

ตารางที่ 4.11 แสดงผลการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ด้วย Support vector machine ของ Lexical, Grammatical and position features

K	Background		Purpose		Method		Result		Discussion		All true	All sent	Accuracy (%)
	sent	all	sent	all	sent	all	sent	all	sent	all			
1	6	7	3	5	1	16	16	22	3	3	41	53	77.36
2	7	7	3	5	14	22	16	17	1	2	41	53	77.36
3	7	8	2	4	10	13	13	23	5	5	37	53	69.81
4	7	9	2	4	11	14	15	19	5	7	40	53	75.47
5	9	11	3	3	16	23	11	16	0	0	39	53	73.58
6	15	18	5	6	9	11	10	14	4	4	43	53	81.13
7	10	11	3	4	14	16	10	19	3	3	40	53	75.47
8	9	12	3	4	10	15	16	19	3	3	41	53	77.36
9	10	12	3	3	15	17	16	19	1	1	45	53	86.54
10	10	13	2	2	12	15	12	20	2	2	38	53	73.08

จากตารางที่ 4.11 วิธีการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ในการแบ่งข้อมูลทดสอบ จะพบว่าในช่วงค่า K เท่ากับ 9 มีค่าความถูกต้องที่ 86.54% ซึ่งมีค่าสูงที่สุด จึงนำเอาโมเดลการจำแนกประเภทของข้อมูลที่ได้จากช่วงค่า K นี้เป็นตัวทดสอบจำแนกประเภทของข้อมูลบทคัดย่อต้นฉบับจำนวน 60 บทความ ผลลัพธ์แสดงได้ดังนี้

ตารางที่ 4.12 แสดงการวิเคราะห์การจำแนกโครงสร้างของบทคัดย่อต้นฉบับจากโมเดลที่ได้รับการฝึกสอนด้วย Support vector machine ที่ค่า K เท่ากับ 9 ของ Lexical, Grammatical and position features

Actual / Predict	Background	Purpose	Method	Result	Discussion
Background	**92**	3	2	0	0
Purpose	4	**39**	9	0	0
Method	6	4	**132**	6	0
Result	1	0	16	**143**	1
Discussion	0	0	2	22	**46**
All sent	103	46	161	171	47
Accuracy (%)	89.32	84.78	81.98	83.62	97.87

จากตารางที่ 4.12 พบว่าค่าความถูกต้องของการวิเคราะห์โครงสร้างระดับประโยคด้วย Lexical, Grammatical and position features อยู่ที่ 85.60% โดยที่ค่าความถูกต้องของแต่ละมูฟมีค่าดังนี้

- Background 89.32%
- Purpose 84.78%
- Method 81.98%
- Result 83.62%
- Discussion 97.87%

การวิเคราะห์โครงสร้างระดับประโยคด้วย Decision tree classifier ของ Grammatical and position features

ตารางที่ 4.13 แสดงผลการทดลองเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ด้วย Decision tree classifier ของ Lexical, Grammatical and position features

K	Background		Purpose		Method		Result		Discussion		All true	All sent	Accuracy (%)
	sent	all	sent	all	sent	all	sent	all	sent	all			
1	7	8	4	7	12	12	17	21	3	4	44	53	83.02
2	6	6	6	9	15	18	15	15	4	5	46	53	86.79
3	8	8	6	10	10	15	12	15	5	5	41	53	77.36
4	11	13	1	2	8	10	15	22	5	6	40	53	75.47
5	15	18	0	0	7	11	14	16	7	8	43	53	81.13
6	7	9	5	6	15	16	11	16	4	6	42	53	79.25
7	8	9	6	6	12	16	12	13	8	9	46	53	86.79
8	5	6	4	6	12	15	14	18	6	8	41	53	77.36
9	7	9	5	5	16	16	16	16	6	6	50	53	96.15
10	11	11	3	4	12	14	13	15	5	5	44	53	84.62

จากตารางที่ 4.13 **วิธีการทดลองเลือกสุ่มข้อมูลแบบ** 10-Fold cross validation ในการแบ่งข้อมูลทดสอบ จะพบว่าในช่วงค่า K เท่ากับ 9 มีค่าความถูกต้องที่ 96.15% ซึ่งมีค่าสูงที่สุด จึงนำเอาโมเดลการจำแนกประเภทของข้อมูลที่ได้จากช่วงค่า K นี้เป็นตัวทดสอบจำแนกประเภทของข้อมูลบทคัดย่อต้นฉบับจำนวน 60 บทความ ผลลัพธ์แสดงได้ดังนี้

ตารางที่ 4.14 แสดงการวิเคราะห์การจำแนกโครงสร้างของบทคัดย่อต้นฉบับจากโมเดลที่ได้รับการฝึกสอนด้วย Decision tree classifier ที่ค่า K เท่ากับ 9 ของ Lexical, Grammatical and position features

Actual / Predict	Background	Purpose	Method	Result	Discussion
Background	**97**	0	0	0	0
Purpose	0	**52**	0	0	0
Method	2	0	**146**	0	0
Result	0	0	0	**161**	0
Discussion	0	0	0	0	**70**
All sent	99	52	146	161	70
Accuracy (%)	97.98	100	100	100	100

จากตารางที่ 4.14 พบว่าค่าความถูกต้องของการวิเคราะห์โครงสร้างระดับประโยคด้วย Lexical, Grammatical and position features อยู่ที่ 99.62% โดยที่ค่าความถูกต้องของแต่ละมูฟมีค่าดังนี้

- Background 97.98%
- Purpose 100%
- Method 100%
- Result 100%
- Discussion 100%

สรุปการทดลองด้วยวิธีการเลือกสุ่มข้อมูลแบบ 10-Fold cross validation ของคุณลักษณะ Lexical, Grammatical and position features ด้วย Decistion tree classifier จะให้ค่าความถูกต้องเท่ากับ 96.15% ซึ่งให้ค่าที่ดีกว่าการใช้ SVM ในการจำแนกมูฟ โดย SVM ให้ค่าความถูกต้องเท่ากับ 86.54%

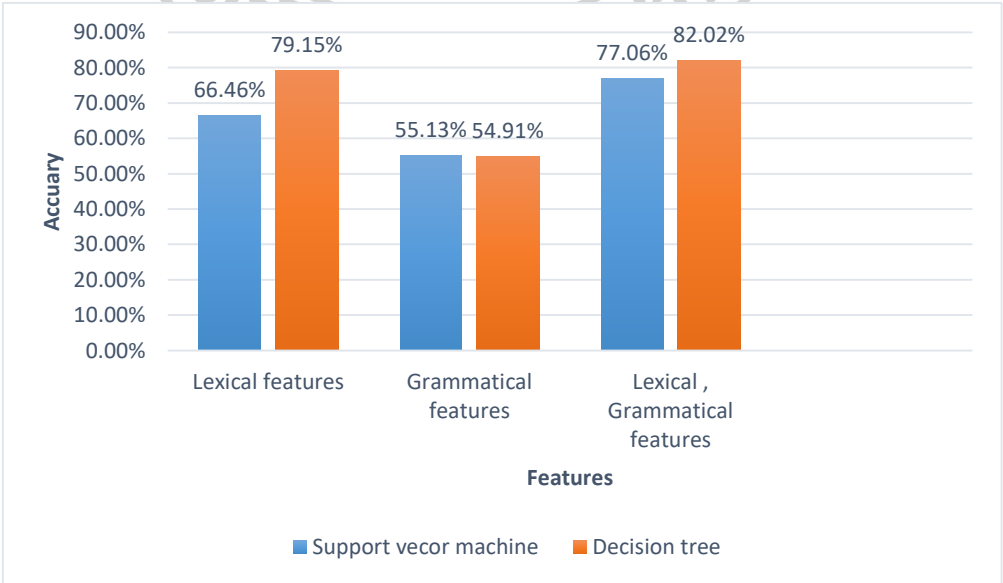
4.5.2 การเปรียบเทียบการจำแนกประเภทของข้อมูลด้วย Support vector machine และ Decision tree classifier ของ Lexical, Grammatical and position features

ตารางที่ 4.15 แสดงผลการจำแนกประเภทของข้อมูลด้วย Support vector machine และ Decision tree classifier ของ Lexical, Grammatical and position features

Actual / Predict	Background		Purpose		Method		Result		Discussion	
	SVM	DTs	SVM	DTs	SVM	DTs	SVM	DTs	SVM	DTs
Background	**92**	**97**	3	0	2	0	0	0	0	0
Purpose	4	0	**39**	**52**	9	0	0	0	0	0
Method	6	2	4	0	**132**	**146**	6	0	0	0
Result	1	0	0	0	16	0	**143**	**161**	1	0
Discussion	0	0	0	0	2	0	22	0	**46**	**70**
All sent	103	99	46	52	161	146	171	161	47	70
Accuracy (%)	89.32	97.98	84.78	100	81.98	100	83.62	100	97.87	100

จากตารางที่ 4.10 จะพบว่าการจำแนกประเภทของข้อมูลด้วย Support vector machine ให้ค่าความถูกต้องเท่ากับ 85.60% และการจำแนกประเภทของข้อมูลด้วย Decision tree classifier ให้ค่าความถูกต้องเท่ากับ 99.62%

4.6 เปรียบเทียบค่าเฉลี่ยประสิทธิภาพของการวิเคราะห์โครงสร้างระดับประโยค



รูปที่ 4.27 แสดงกราฟเปรียบเทียบค่าเฉลี่ยประสิทธิภาพของการวิเคราะห์โครงสร้างระดับประโยค

จากรูปที่ 4.27 แสดงกราฟเปรียบเทียบค่าเฉลี่ยประสิทธิภาพของการวิเคราะห์โครงสร้างระดับประโยคของ Lexical features, Grammatical and position features และ Lexical, Grammatical and position features โดยเปรียบเทียบการวิเคราะห์ของโมเดลจำแนกโครงสร้างของมูฟระหว่าง Support vector machine (SVM) กับ Decision tree classifier (ID3)

จากกราฟค่าเฉลี่ยประสิทธิภาพการวิเคราะห์โครงสร้าง Decision tree classifier จะมีประสิทธิภาพที่ดีกว่า Support vector machine ในรูปแบบการวิเคราะห์ของ Lexical features และ Lexical, Grammatical and position features แต่ในรูปแบบการวิเคราะห์แบบ Grammatical and position features SVM จะให้ค่าประสิทธิภาพที่ดีกว่า โดยที่การวิเคราะห์แบบ Lexical, Grammatical and position features นั้นจะให้ค่าประสิทธิภาพที่ดีกว่าอีกทั้ง 2 Features จึงเป็นรูปแบบที่เหมาะสมที่ใช้ในการวิเคราะห์

โดยรวมค่าประสิทธิภาพที่ได้จากการวิเคราะห์ทั้ง 3 รูปแบบ ค่าที่ได้ไม่แตกต่างกันมาก จากการตั้งข้อสันนิฐานของผู้วิจัยว่า Support vector machine จะให้ค่าประสิทธิภาพที่ดีกว่า Decision tree classifier แต่ผลที่ได้มีค่าประสิทธิภาพที่น้อยกว่าเล็กน้อยจึงเป็นที่น่าสังเกตว่าข้อมูลในการฝึกสอนนั้นอาจมีผลต่อการฝึกสอนของโมเดล ผู้วิจัยจึงได้เพิ่มข้อมูลในการฝึกสอน (Training set) ในแต่ละรอบฝึกฝนการเลือกสุ่มข้อมูลแบบ 10-Fold cross validation โดยใช้รูปแบบ Lexical, Grammatical and position features ซึ่งให้ประสิทธิภาพที่ดี ซึ่งผลลัพธ์ที่ได้จากการวิเคราะห์ โดยค่าประสิทธิภาพในแต่ละรอบการวิเคราะห์นั้นมีค่าที่เพิ่มขึ้นอย่างชัดเจน

บทที่ 5
สรุปผลการวิจัย

จากการศึกษาเกี่ยวกับโครงสร้าง และรูปแบบของการเขียนบทคัดย่อตามแนวคิดการวิเคราะห์รูปแบบสัมพันธสาร (Discourse analysis) คือ อัตถภาควิเคราะห์ (Move analysis) พบว่าลักษณะโครงสร้างของบทคัดย่อในงานแต่ละด้านที่เขียนขึ้นนั้น มีความคล้ายคลึงกัน อันประกอบไปด้วยส่วนต่างๆ หรือที่เรียกว่า มูฟ (Move) ดังนี้ พื้นฐานของงานวิจัย (Background) วัตถุประสงค์ (Purpose) กระบวนการทดลอง (Methodology) ผลการทดลอง (Result) สรุปผลการทดลอง (Conclusion) อภิปรายผลการทดลอง (Discussion) ซึ่งในบางบทคัดย่อผู้เขียน อาจเขียนโดยไม่มีองค์ประกอบดังกล่าวครบทุกส่วน และการปรับปรุง พัฒนางานวิจัยทางด้านของการวิเคราะห์โครงสร้างของบทคัดย่อ ในรูปแบบของโปรแกรมประยุกต์บนเว็บไซต์ที่สามารถวิเคราะห์รูปแบบโครงสร้างของบทคัดย่อ และวิเคราะห์องค์ประกอบของบทคัดย่อ หรือมูฟของเอกสารทางวิชาการได้ในหลายหลายสาขาวิชา โดยนำวิธีการเรียนรู้ของเครื่อง (Machine Learning) คือ Support vector machine มาใช้ในการวิเคราะห์ในลักษณะของการทำเหมืองข้อความตามคุณลักษณะต่างๆ คือ Lexical features, Grammatical and position features และ Lexical, Grammatical and position features โดยแสดงผลลัพธ์จากการวิเคราะห์โครงสร้างของบทคัดย่อ อีกทั้งมีส่วนของผู้ดูแลระบบที่สามารถเพิ่มคลังของบทความในสาขาวิชาต่างๆ เพิ่มโมเดลวิธีการเรียนรู้ที่ต้องการให้แก่โปรแกรมประยุกต์บนเว็บไซต์ที่ใช้ในการวิเคราะห์โครงสร้างของบทคัดย่อ พร้อมทั้งแสดงรูปแบบของมูฟในสาขานั้นๆ

ส่วนแรก คือส่วนของผู้ใช้งานหรือ โหมดวิเคราะห์ (Analysis mode) เป็นส่วนสำหรับผู้ที่ต้องการนำเข้าบทคัดย่อเพื่อให้ระบบวิเคราะห์โครงสร้าง โดยผู้ใช้งานสามารถเลือกสาขาวิชาของบทคัดย่อที่ต้องการให้ระบบวิเคราะห์ เลือกวิธีการเรียนรู้ของเครื่องที่ต้องการใช้ในการวิเคราะห์ และเลือกบทคัดย่อในรูปแบบของไฟล์ PDF สู่ระบบเพื่อทำการวิเคราะห์โครงสร้าง เมื่อระบบวิเคราะห์แล้วระบบจะแสดงสรุปผลลัพธ์ของโครงสร้างบทคัดย่อแยกตามสีของแต่ละบทคัดย่อ โดยเรียงลำดับของประโยคที่รับเข้าไป แสดงมูฟที่บทคัดย่อมีและขาดหายไป และแสดงกราฟสรุปจำนวนมูฟในบทคัดย่อในบทความที่มีอีกด้วย

ส่วนที่สอง คือส่วนฝึกสอนระบบหรือ โหมดฝึกสอน (Training mode) เป็นส่วนสำหรับการเพิ่มคลังบทความในสาขาวิชาต่างๆ อีกทั้งเพิ่มการเรียนรู้ด้วยเครื่อง หรือโมเดลที่ใช้ในการวิเคราะห์บทความสาขานั้นๆ ตามที่ระบบมีให้เลือกใช้งานได้ ผู้ใช้งานสามารถเพิ่มสาขาวิชาของบทคัดย่อ และชื่อย่อของสาขาวิชานั้น จากนั้นเลือกการเรียนรู้ด้วยเครื่องหรือโมเดลที่ต้องการใช้ในการฝึกสอนระบบให้เรียนรู้โครงสร้างของบทความตามที่ระบบมีให้เลือก จากนั้นผู้ใช้งานสามารถต้อง

นำเข้าบทคัดย่อที่ต้องการฝึกสอนระบบในรูปแบบไฟล์ Text พร้อมผลเฉลยในแต่ละประโยคว่าประโยคนั้นๆมีมูฟเป็นมูฟอะไร จำนวนทั้งหมด 60 ไฟล์ ไฟล์ละ 1 บทคัดย่อ โดยสามารถลากและวางไฟล์เพื่อเข้าสู่ระบบได้เลย เมื่อระบบทำการวิเคราะห์และสร้างโมเดลจำแนกประเภทของข้อมูลแล้วจะแสดงผลสรุปของการเรียนรู้แสดงผลลัพธ์ของการฝึกฝนในรูปแบบของกราฟแสดงค่าเฉลี่ยประสิทธิภาพ ผู้ใช้สามารถนำเมาส์วางไว้บริเวณแท่งกราฟจะแสดงกล่องข้อความแสดงค่าเฉลี่ยประสิทธิภาพขึ้นมา แล้วยังแสดงชื่อของโมเดลจำแนกประเภทของข้อมูลที่ระบบสร้างขึ้น

ปัญหาและอุปสรรค

1. การเพิ่มการเรียนรู้ของเครื่องสามารถที่จะเพิ่มได้จากที่ระบบมีกำหนดให้เท่านั้น เนื่องจากในการเรียกใช้งานการเรียนรู้ของเครื่องในแต่ละแบบมีคำสั่งการทำงานที่แต่ต่างกันออกไปจึงต้องเป็นผู้สร้างระบบเท่านั้นที่เป็นผู้ที่กำหนดขึ้นมา

2. การใช้งานการเรียนรู้ของเครื่อง Support vector machine นั้นข้อมูลต้นแบบที่ใช้ในการฝึกสอนโมเดลจำแนกประเภทของข้อมูลมีจำนวนไม่มากพอจึงทำให้ประสิทธิภาพการวิเคราะห์โครงสร้างบทคัดย่อออกไม่เป็นตามที่ต้องเป้าหมายไว้

3. เนื่องจากเป็นการฝึกสอนโมเดลการเรียนรู้ของเครื่อง จึงทำให้ใช้ระยะเวลาในการประมวลผลที่นาน

ข้อเสนอแนะในการวิจัย

1. ข้อมูลที่นำมาใช้เป็นบทความต้นแบบมีจำนวนน้อยเกินไป เพื่อให้ผลการวิเคราะห์ของการเรียนรู้ของเครื่องมีประสิทธิภาพและความถูกต้องที่มากขึ้นควรรวบรวมบทความเพิ่มเติม

2. เพิ่มความรวดเร็วในการประมวลผลของระบบ

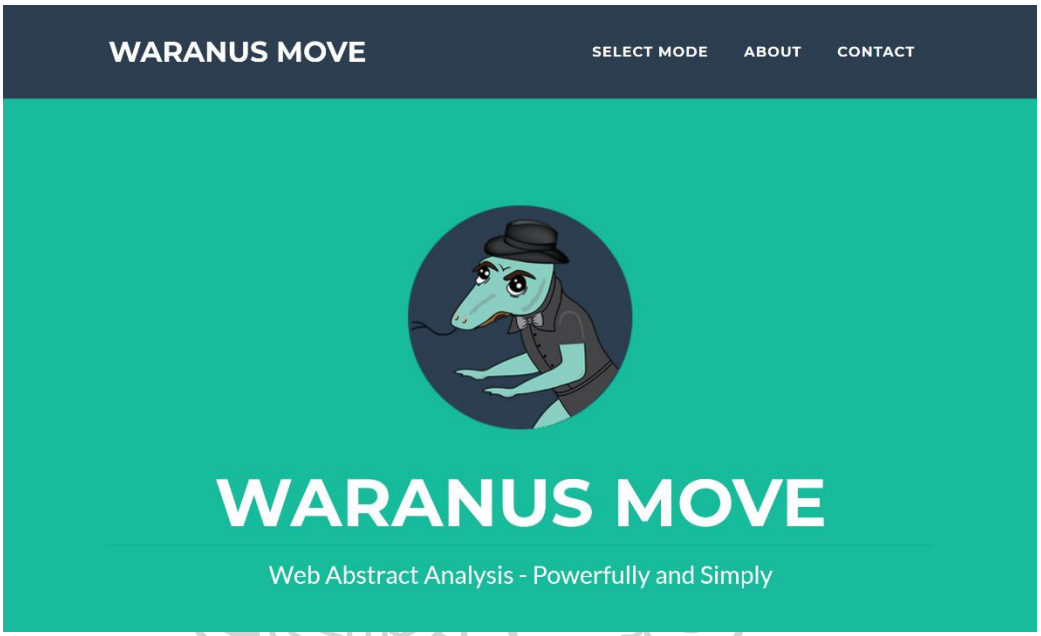
3. ควรปรับปรุงในส่วนของการเพิ่มโมเดลจำแนกประภทของข้อมูลการเรียนรู้เครื่องให้สามารถเลือกการเรียนรู้ของเครื่องได้มากว่านี้

ภาคผนวก

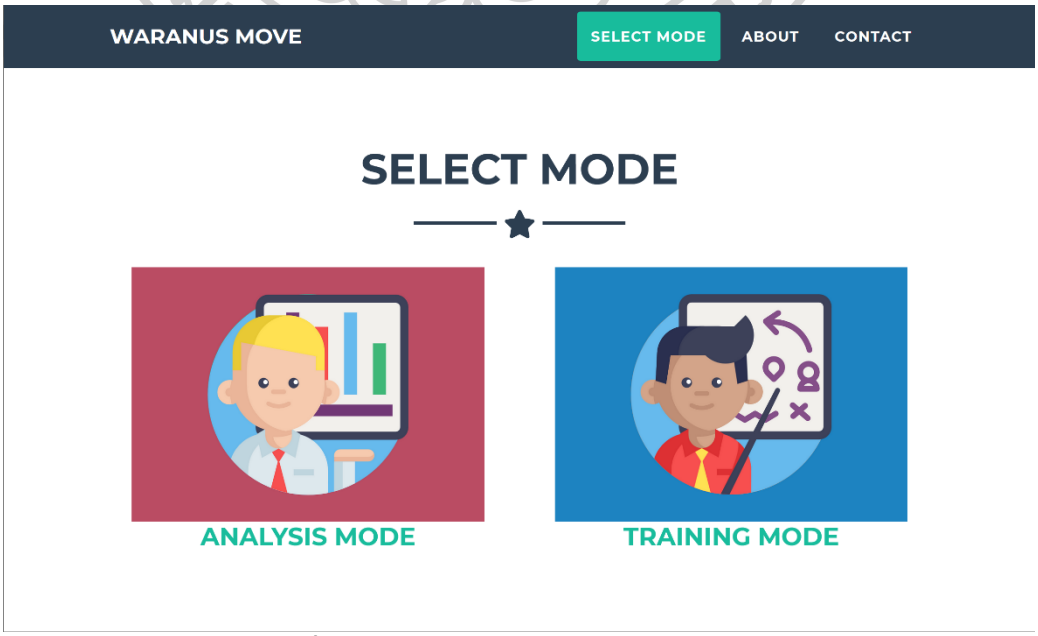
คู่มือการใช้งานระบบ

1. หน้าแรกของระบบ

การใช้งานเว็บไซต์เครื่องมือวิเคราะห์โครงสร้าง และองค์ประกอบของบทความทางวิชาการในส่วนบทคัดย่อหรือมูฟ



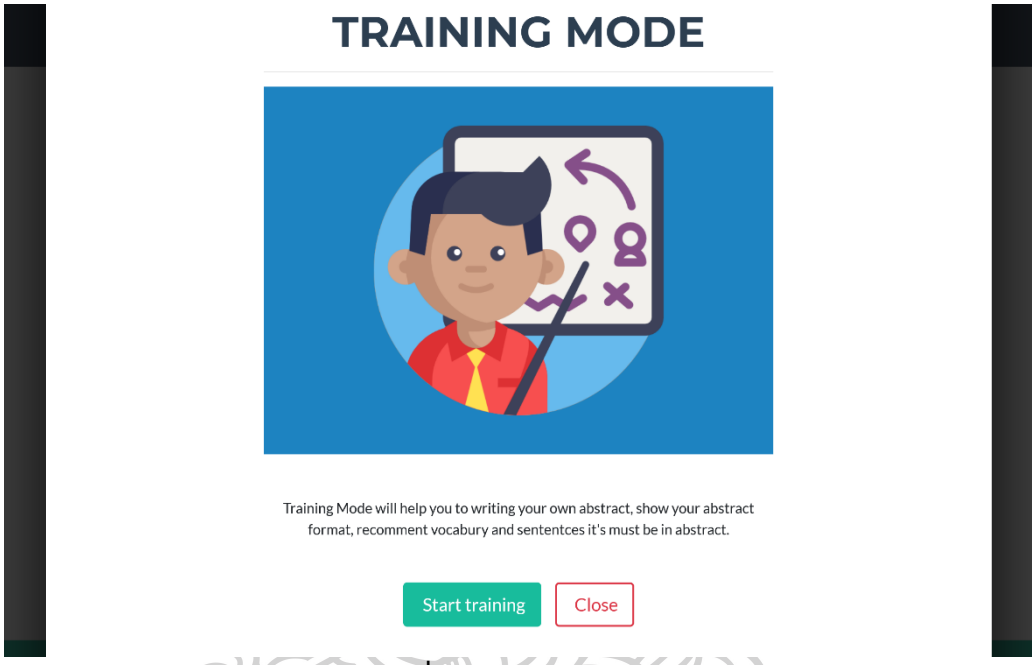
รูปที่ 1 แสดงหน้าจอของโปรแกรม



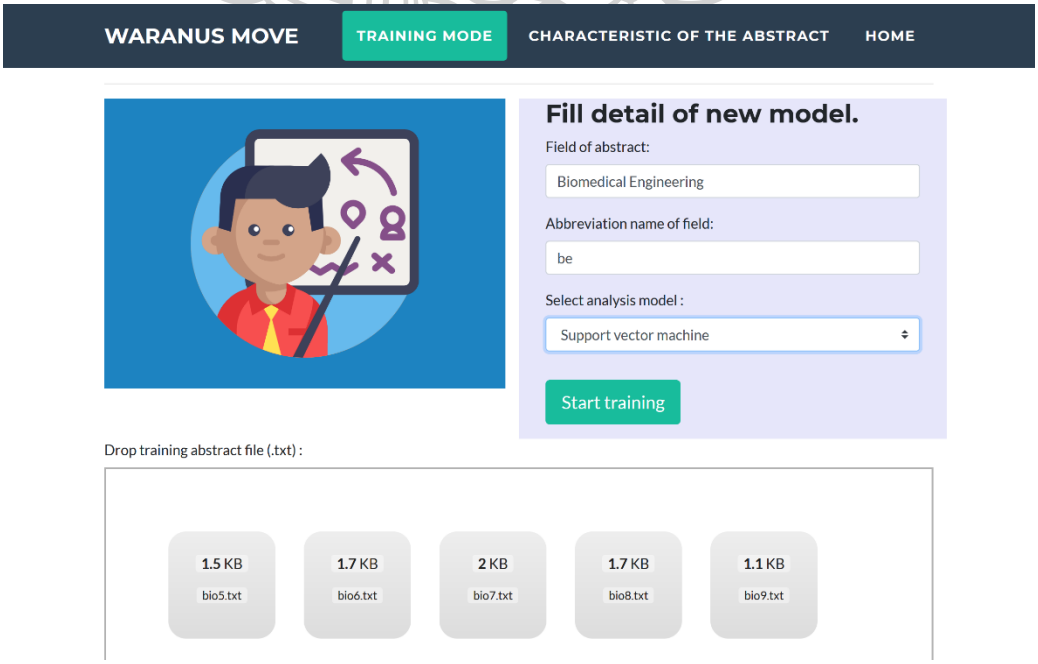
รูปที่ 2 แสดงโหมดการวิเคราะห์ของโปรแกรม

2. ส่วนการฝึกฝน (Training mode)

เป็นส่วนที่ผู้ใช้งานสามารถเพิ่มสาขาวิชาอื่นๆ ของบทความทางวิชาการในส่วนบทคัดย่อเข้าสู่ระบบ และสามารถเลือกเพิ่มการเรียนรู้ด้วยเครื่อง (Machine learning) สำหรับการเรียนรู้โครงสร้างของบทคัดย่อในสาขาวิชาใหม่ที่ผู้ใช้เพิ่มเข้าไป



รูปที่ 3 แสดงโหมดฝึกฝน



รูปที่ 4 แสดงหน้าจอหลักของส่วนการฝึกฝน

จากรูปที่ 4 ผู้ใช้งานต้องกรอกรายละเอียดของสาขาวิชาของบทคัดย่อที่ต้องเพิ่มเข้าสู่ระบบ คือ ชื่อของสาขาวิชา ชื่อย่อของสาขาวิชา และเลือกโมเดลการเรียนรู้ของเครื่อง อีกทั้งผู้ใช้สามารถลากและวางไฟล์เอกสารเพื่อเพิ่มบทคัดย่อในสาขาวิชาที่ต้องการได้อีกด้วย

การเพิ่มไฟล์เอกสารบทคัดย่อ ไฟล์เอกสารจะต้องมีผลเฉลยมูฟของแต่ละประโยค ลักษณะโครงสร้างของไฟล์เอกสารบทคัดย่อ Text จะต้องมีลักษณะคือ แต่ละประโยคของบทคัดย่อต้องมีส่วนเฉลยของมูฟ เช่น [B] -> กำกับอยู่และตามด้วยประโยค ตัวอย่างประโยค Background คือ [B]-> The proposed method has been tested with both simulated and experimental data. โดยส่วนเฉลยของมูฟมีดังนี้

[B] -> คือ Background

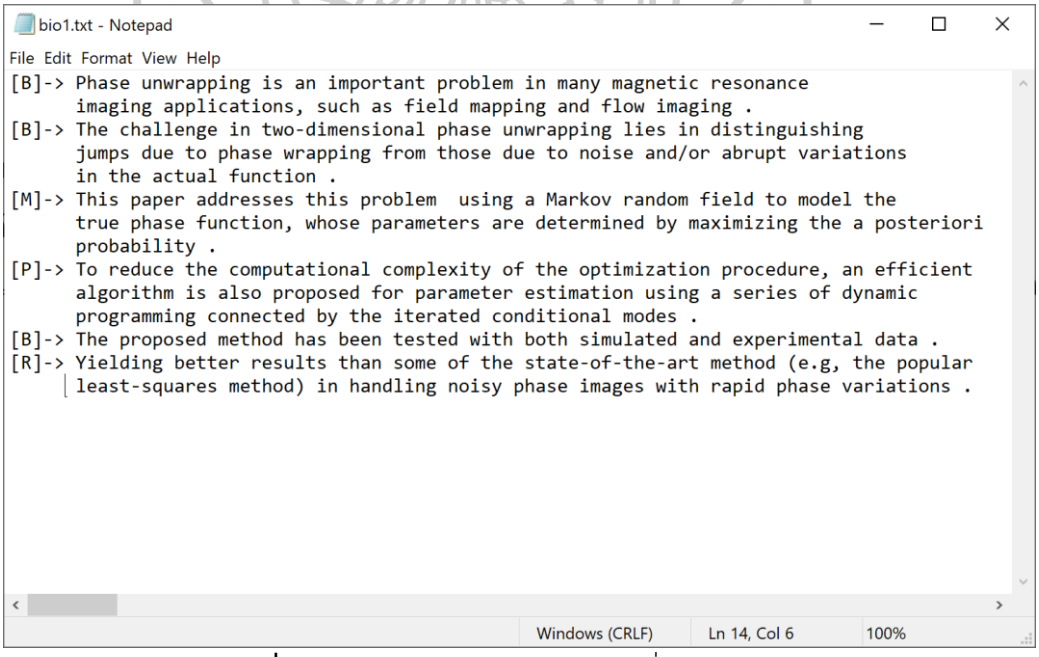
[P] -> คือ Purpose

[M] -> คือ Method

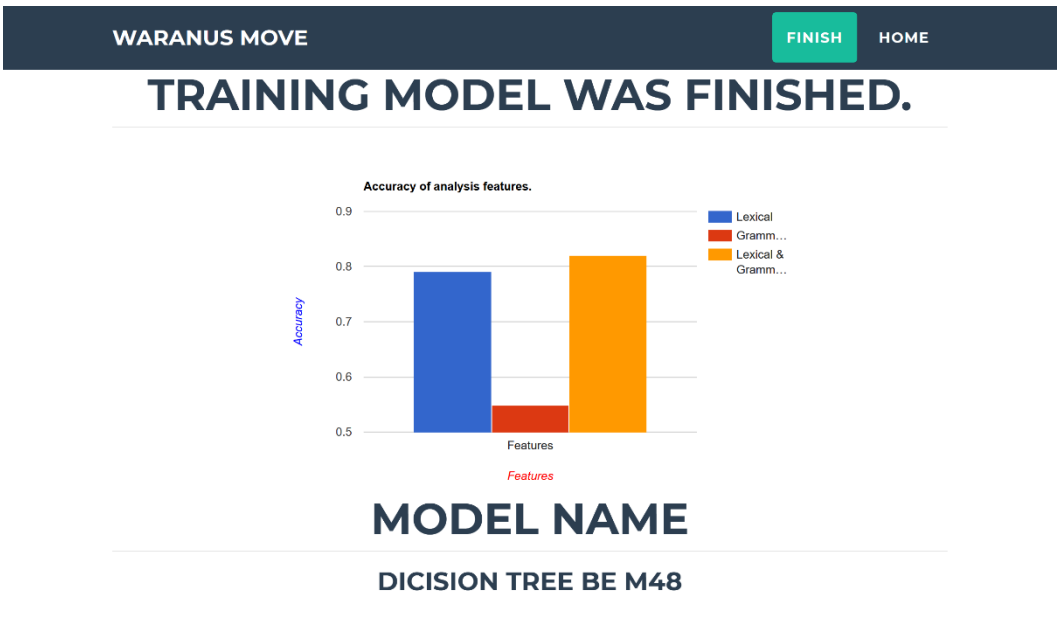
[R] -> คือ Result

[D] -> คือ Discussion

บทคัดย่อ ต้องเป็นไฟล์เอกสาร text ที่มีนามสกุลเป็น .txt ผู้ใช้ต้องตั้งชื่อไฟล์ดังนี้ คือ bio1.txt, bio2.txt, bio3.txt, bio4.txt, ... , bio60.txt ทั้งหมดจำนวน 60 ไฟล์ จากรูปที่ 5 คือไฟล์เอกสารบทคัดย่อ Text จำนวน 60 ไฟล์ อีกทั้งผู้ใช้สามารถเพิ่มคลังคำของผู้ใช้เองได้โดยเลือกจาก Create corpus เมื่อระบบสร้างคลังคำเรียบร้อยแล้วจะปรากฏ ดังรูปที่ 6



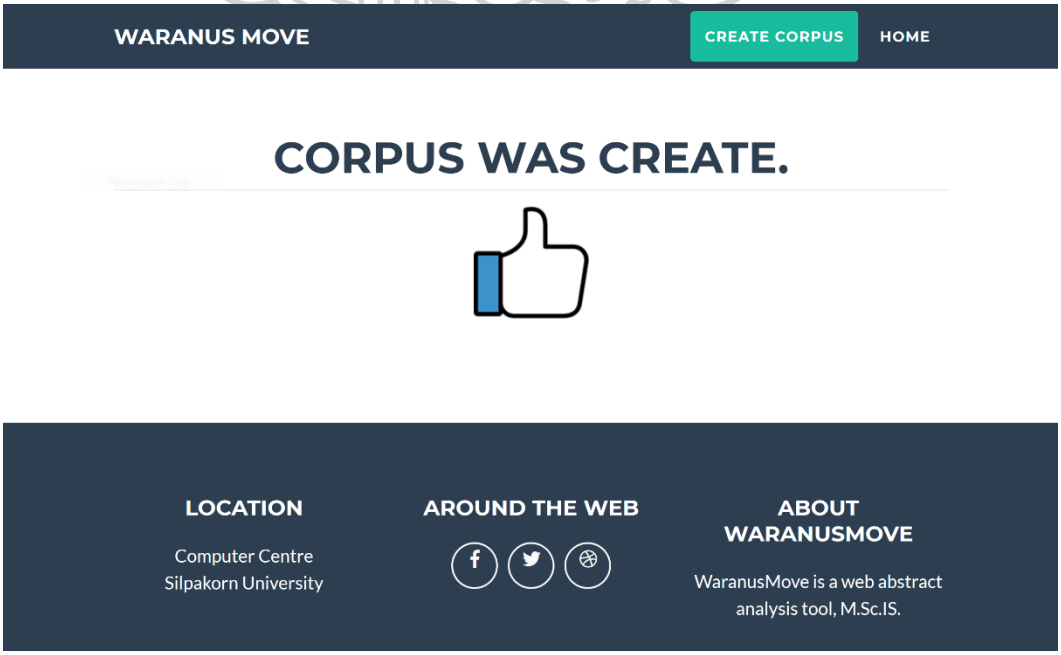
รูปที่ 5 แสดงตัวอย่างรูปแบบของไฟล์ที่รับเข้าสู่ระบบ



รูปที่ 6 แสดงหน้าจอของโปรแกรมเมื่อโปรแกรมเรียนรู้บทคัดย่อใหม่เรียบร้อยแล้ว

หลังจากที่ผู้ใช้ลากและวางไฟล์บทคัดย่อจำนวน 60 บท เข้าสู่ระบบแล้วจากนั้น ให้ผู้ใช้เลือก Start training โปรแกรมจะทำการเรียนรู้บทคัดย่อสาขาวิชาใหม่ และสร้างโมเดลการเรียนรู้ใหม่เพื่อนำไปใช้ในการวิเคราะห์บทคัดย่อ เมื่อโปรแกรมทำงานเสร็จจะแสดงหน้าจอดังรูปที่ 6

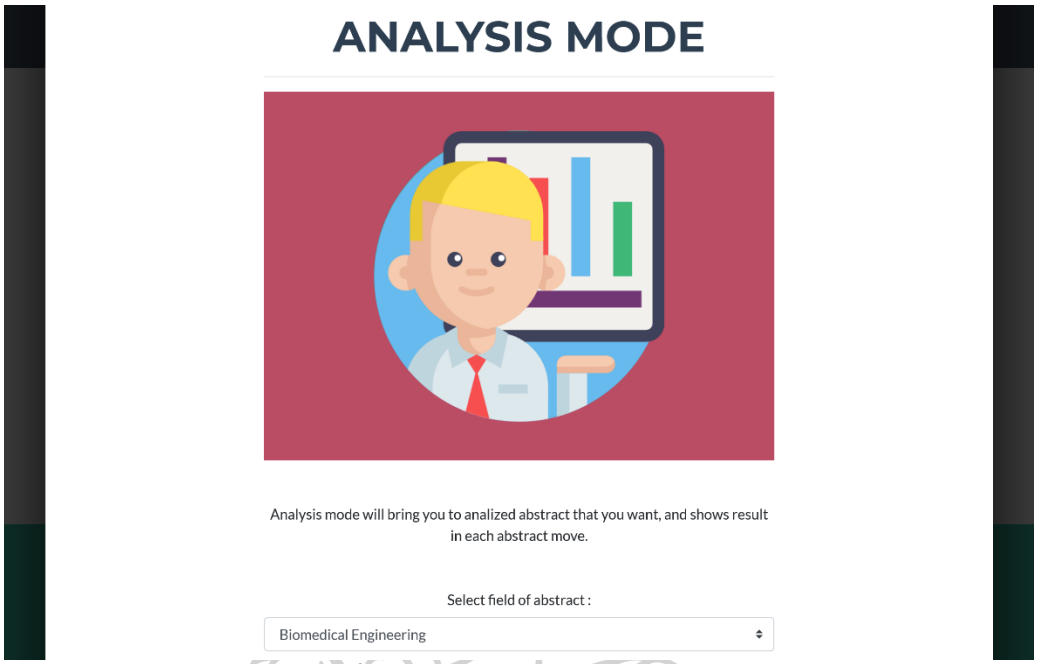
อีกทั้งผู้ใช้ยังสามารถเพิ่มคลังคำศัพท์ของผู้ใช้เองได้โดยเลือกจากเมนู Create corpus ในรูปที่ 6 เมื่อระบบสร้างคลังคำศัพท์เรียบร้อยแล้วจะแสดงข้อความแจ้งเตือนผู้ใช้ ดังรูปที่ 7



รูปที่ 7 แสดงระบบสร้างคลังคำเรียบร้อยแล้ว

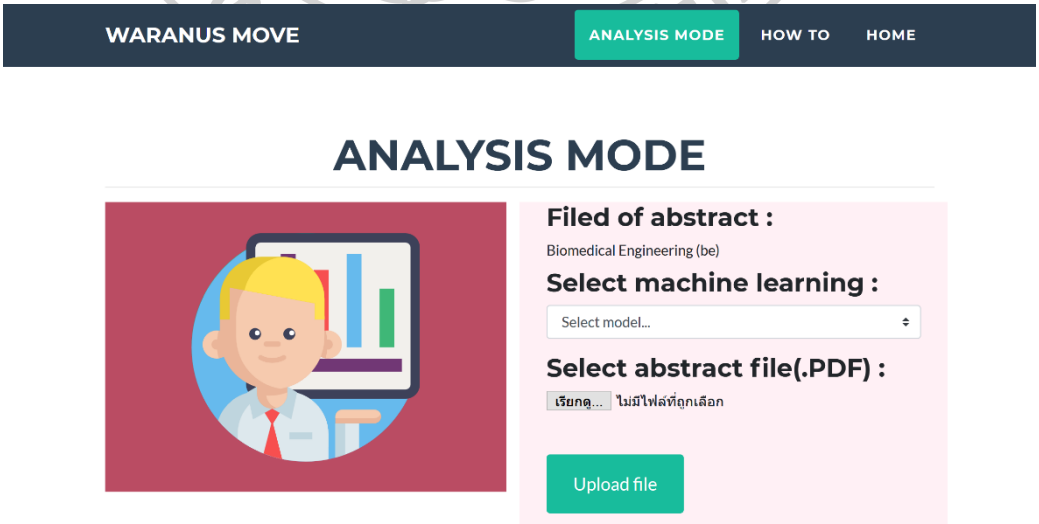
3. ส่วนการวิเคราะห์ (Analysis mode)

เป็นส่วนที่ผู้ใช้งานสามารถนำบทคัดย่อเข้าสู่โปรแกรม เพื่อให้โปรแกรมทำการวิเคราะห์โครงสร้างของบทคัดย่อหรือมูฟ



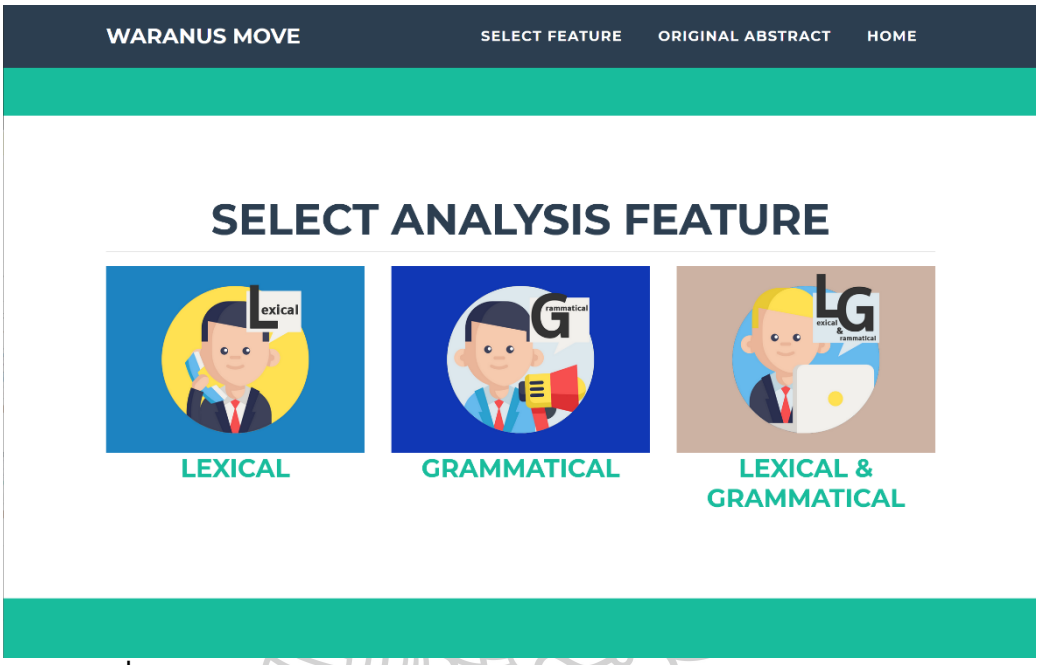
รูปที่ 8 แสดงโหมดวิเคราะห์บทคัดย่อ

จากรูปที่ 8 เมื่อเข้าสู่ส่วนการวิเคราะห์บทคัดย่อผู้ใช้งานจะต้องเลือกสาขาของบทความที่ผู้ใช้ต้องการวิเคราะห์



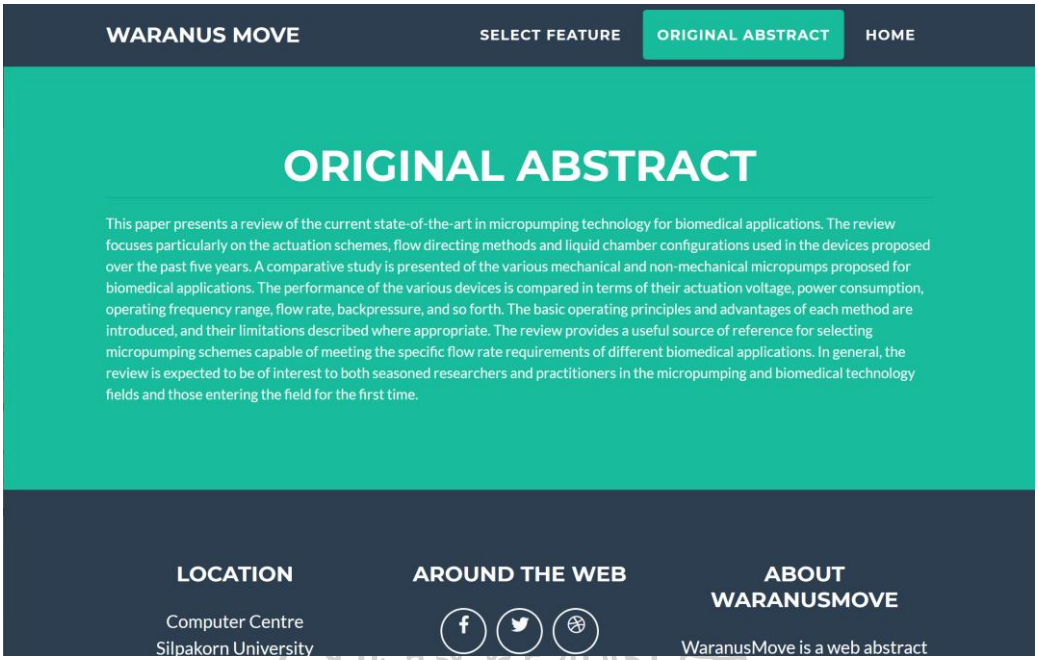
รูปที่ 9 แสดงหน้าจอหลักของส่วนการวิเคราะห์บทคัดย่อ

เมื่อผู้ใช้เข้าสู่ส่วนการวิเคราะห์บทคัดย่อจะเข้าสู่หน้าจอหลัก โดยในหน้าจอหลักจะแสดงสาขาวิชาที่ผู้ใช้เลือกมาก่อนหน้านี้ และผู้ใช้ต้องเลือกโมเดลที่ต้องการใช้ในการวิเคราะห์โครงสร้างของบทคัดย่อ ต่อมาผู้ใช้ต้องนำบทคัดย่อเข้าสู่ระบบโดยบทคัดย่อในรูปแบบของไฟล์ PDF และเลือก Upload file เพื่อทำการวิเคราะห์บทคัดย่อ จากนั้นผู้ใช้เลือก Upload file เพื่อให้ระบบทำการวิเคราะห์

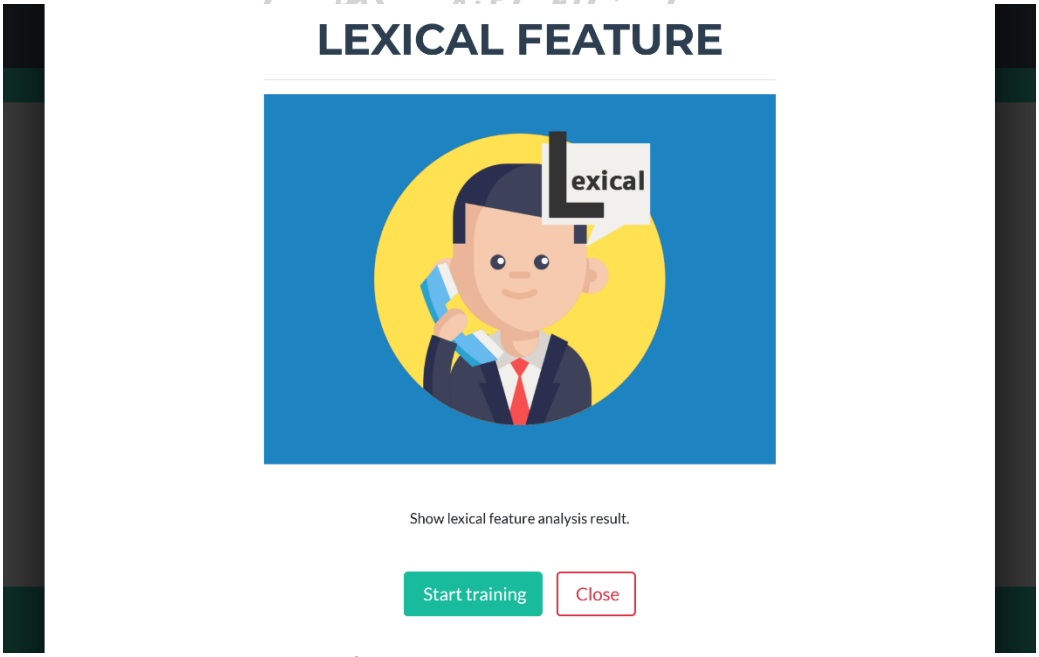


รูปที่ 10 แสดงหน้าจอการเลือกคุณลักษณะของการวิเคราะห์โครงสร้างบทคัดย่อ

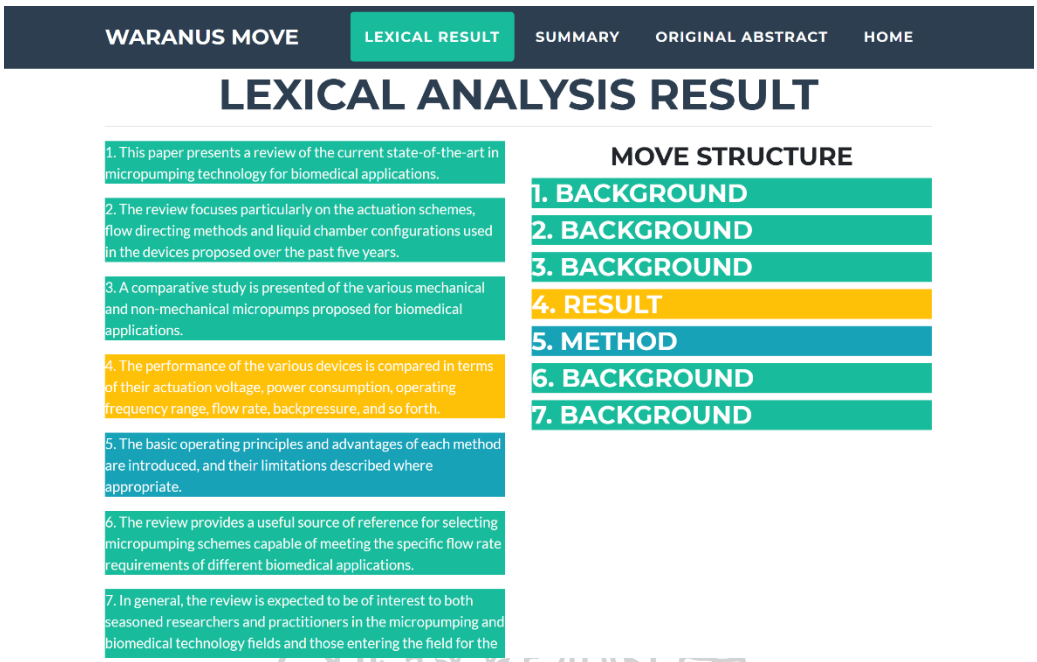
หลังจากผู้ใช้ Upload file บทคัดย่อเข้าสู่ระบบเพื่อทำการวิเคราะห์แล้วจะเข้าสู้หน้าจอการเลือกคุณลักษณะโดยจะมี 3 คุณลักษณะ คือ Lexical features Grammatical and position features และ Lexical, Grammatical and position features โดยผู้ใช้งานสามารถเลือกได้ และส่วนถัดมาโปรแกรมยังแสดงส่วน Original abstract ที่ผู้ใช้ Upload เข้าสู่ระบบ



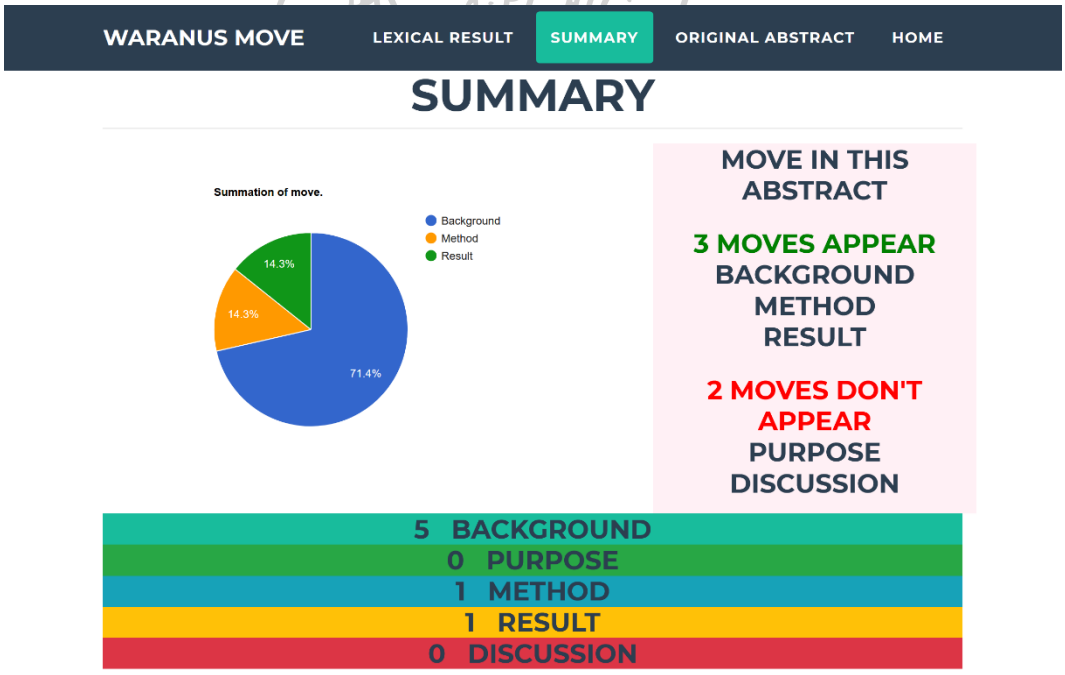
รูปที่ 11 แสดงส่วนของ Original abstract



รูปที่ 12 แสดงส่วน Lexical features

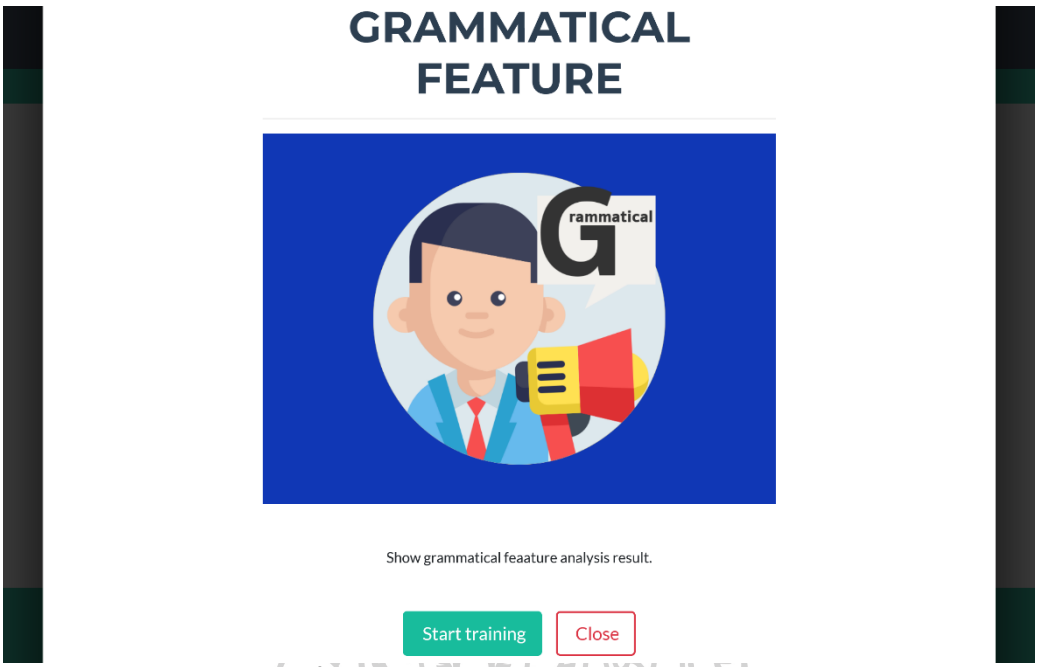


รูปที่ 13 แสดงผลการวิเคราะห์บทคัดย่อของ Lexical features

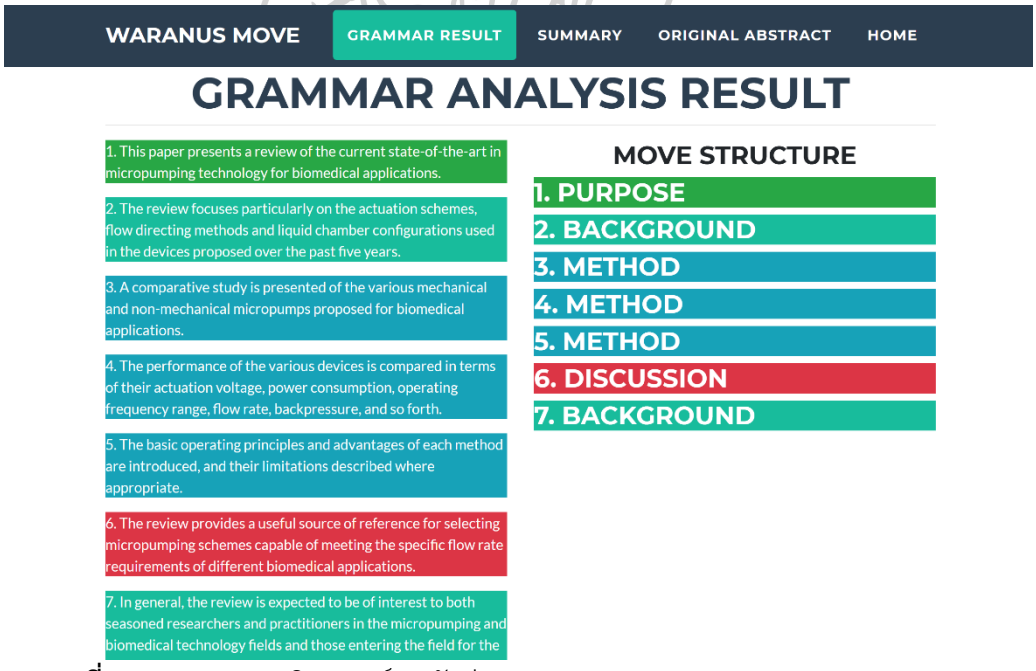


รูปที่ 14 แสดงผลสรุปของการวิเคราะห์บทคัดย่อ

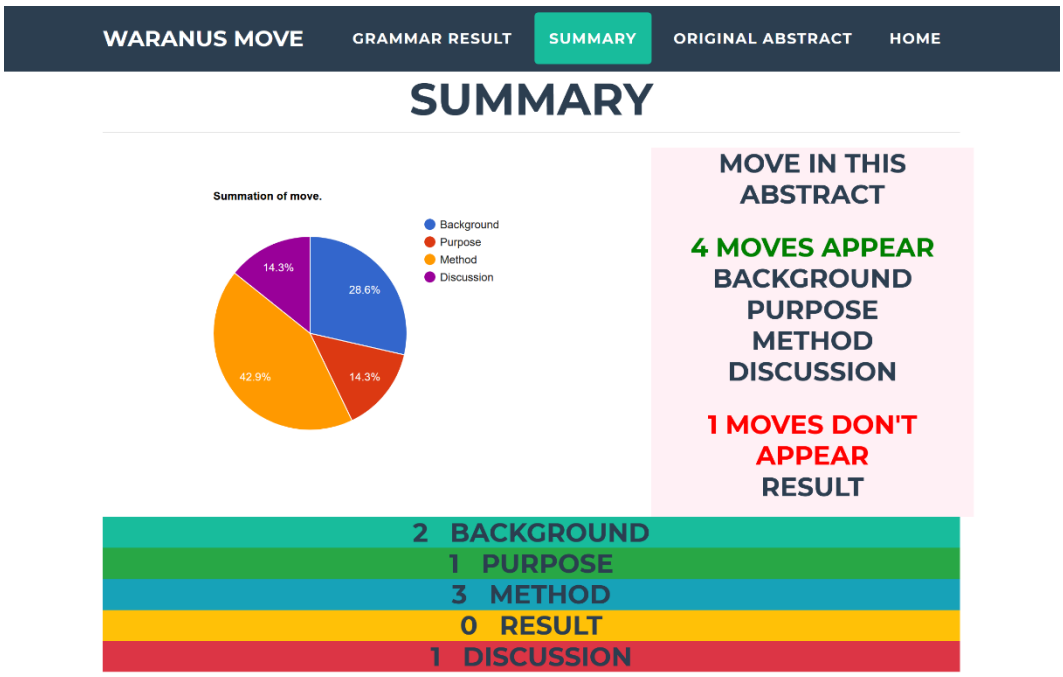
จากรูปที่ 13 และ 14 แสดงผลการวิเคราะห์บทคัดย่อโดยแยกแต่ละประโยคออกเป็นมูฟด้วยการไฮไลท์ประโยค แสดงการเรียงตัวของมูฟ ทั้งนี้ยังแสดงกราฟสรุปจำนวนมูฟในบทคัดย่อ มูฟที่ปรากฏและและไม่ปรากฏในบทคัดย่ออีกด้วย



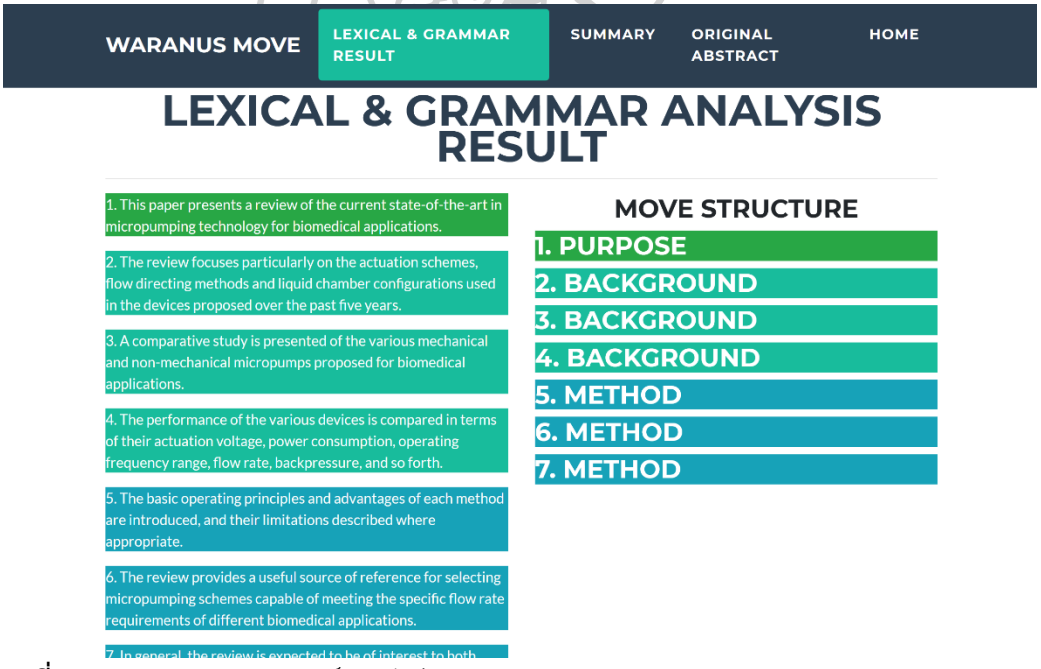
รูปที่ 15 แสดงส่วน Grammatical features



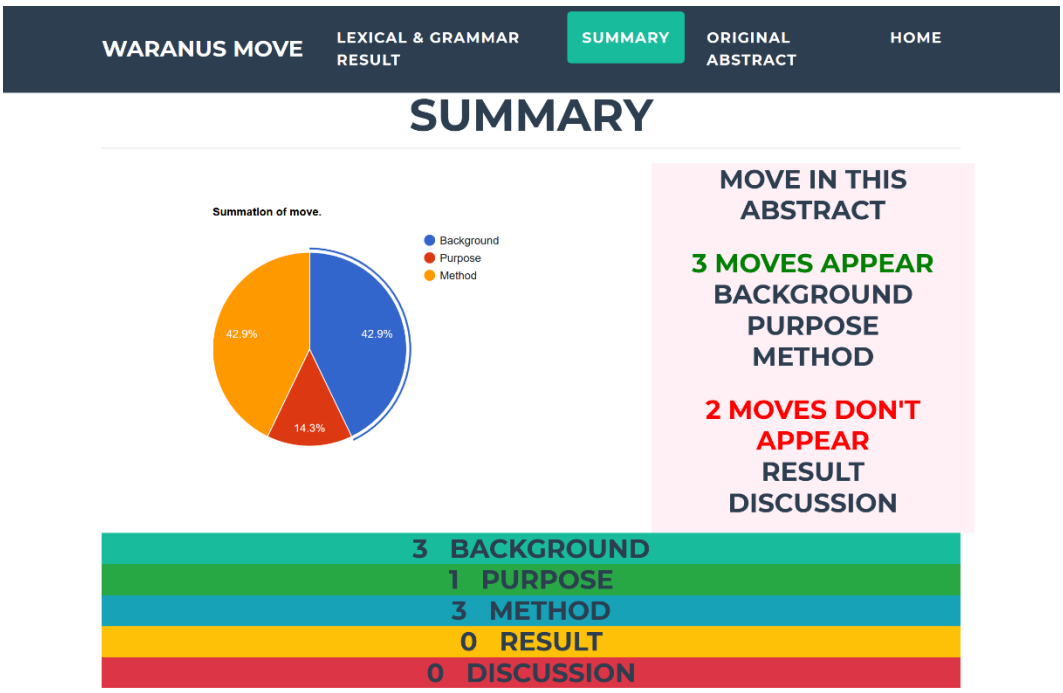
รูปที่ 16 แสดงผลการวิเคราะห์บทคัดย่อของ Grammatical and positon features



รูปที่ 17 แสดงผลสรุปของการวิเคราะห์บทคัดย่อ



รูปที่ 18 แสดงผลการวิเคราะห์บทคัดย่อของ Lexical, Grammatical and positon features



รูปที่ 19 แสดงผลสรุปของการวิเคราะห์บทคัดย่อ

รายการอ้างอิง

Angrosh, M. A., Cranefield, S., & Stanger, N. (2013). Conditional Random Field Based Sentence Context Identification: Enhancing Citation Services for the Research Community. *AWC '13 Proceedings of the First Australasian Web Conference,* 144, 59-68.

Anthony, L., & Lashkia, G. V. (2003). Mover: a machine learning tool to assist in the reading and writing of technical papers. *IEEE Transactions on Professional Communication,* 64(3), 185 - 193.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* (1st Edition ed.). United States of America: O'Reilly Media, Inc.

Goyvaerts, J. (2007). Regular Expressions: The Complete Tutorial. Retrieved from https://www.regular-expressions.info/print.html

Guo, Y., Silins, I., Stenius, U., & Korhonen, A. (2013). Active learning-based information structure analysis of fullscientific articles and two applications for biomedicalliterature review. *Bioinformatics,* 29(11), 1440-1447.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Journal Bioinformatics Volume 28 Issue 7,* 28(7), 991-1000.

López, F., & Romero, V. (2014). *Mastering Python Regular Expressions.* (1st Edition ed.). Birmingham B3 2PB, UK: Packt Publishing Ltd.

M.A., A., Cranefield, S., & Stanger, N. (2010). Context identification of sentences in related work sections using a conditional random field: Towards intelligent digital libraries. *JCDL'10,* 320-293.

Matos, P. F., Lombardi, L. O., Pardo, T. A. S., Ciferri, C. D. A., Vieira, M. T. P., & Ciferri, R. R. (2012). An environment for data analysis in biomedical domain: information extraction for decision support systems. *IEA/AIE'10,* 23, 306-316.

Peng, F., & McCallum, A. (2006). Information extraction from research papers using conditional random fields. *Information Processing and Management: an International Journal,* 42(4), 963-979.

Wu, J.-C., Chang, Y.-C., Liou, H.-C., & Chang, J. S. (2006). Computational Analysis of Move Structures in Academic Abstracts. *COLING-ACL '06,* 21, 41-44.

นัศพ์ชาณัณ ชินปัญช์ธนะ. (2557). การแปลความหมายภาพด้วยแนวคิดพื้นฐานความสัมพันธ์ของกราฟแบบลำดับชั้น.". Retrieved from http://www.dpu.ac.th/dpurc/research-394

บุษบา กนกศิลปธรรม, & สุดาพร ลักษณียนาวิน. (2549). การวิเคราะห์คลังข้อมูล: โครงสร้างบทความวิจัยจุลชีววิทยา. Retrieved from http://elibrary.trf.or.th/project_content.asp?PJID=MRG4780210

ปริญญา สงวนสัตย์. (2558). การเรียนรู้ของเครื่อง (*Machine Learning)*. นนทบุรี: คณะวิศวกรรมศาสตร์และเทคโนโลยี สถาบันการจัดการปัญญาภิวัฒน์.

สุวรรณี เต็งอำนวย, & อรสา เจนพนัส. (2557). *Successful English Grammar* ภาษาอังกฤษ. กรุงเทพฯ ภูมิบัณฑิต.

เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น (พิมพ์ครั้งที่ 2 ed.). ปทุมธานี: หสม. ดาต้า คิวบ์.

ประวัติผู้เขียน

ชื่อ-สกุล	รัฐพล ชูพรม
วัน เดือน ปี เกิด	29 พฤศจิกายน 2531
สถานที่เกิด	จังหวัดประจวบคีรีขันธ์
วุฒิการศึกษา	วิทยาศาสตรบัณฑิต เทคโนโลยีสารสนเทศ เกียรตินิยมอันดับ 2
ที่อยู่ปัจจุบัน	87 ถนนเลียบทางรถไฟไปคลองวาฬ ตำบลประจวบคีรีขันธ์ อำเภอเมือง จังหวัดประจวบคีรีขันธ์ 77000