การตรวจเทียบภายในหาการลักลอกงานวิชาการภาษาไทยโดยใช้แบบจำลอง
ซัพพอร์ตเวกเตอร์แมชชีน

นางสาวศิวพร ทวนไธสง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

AN INTRINSIC PLAGIARISM DETECTION OF THAI ACADEMIC TEXTS USING A SUPPORT VECTOR MACHINE MODEL

Miss Siwaporn Tuanthaisong

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Arts Program in Linguistics
Department of Linguistics
Faculty of Arts
Chulalongkorn University
Academic Year 2013

หัวข้อวิทยานิพนธ์	การตรวจเทียบภายในหาการลักลอกงานวิชาการภาษาไทยโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน
โดย	นางสาวศิวพร ทวนไธสง
สาขาวิชา	ภาษาศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะอักษรศาสตร์

(ผู้ช่วยศาสตราจารย์ ดร.ประพจน์ อัศววิรุฬหการ)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร.เทพชัย ทรัพย์นิธิ)

ศิวพร ทวนไธสง : การตรวจเทียบภายในหาการลักลอกงานวิชาการภาษาไทยโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน. (AN INTRINSIC PLAGIARISM DETECTION OF THAI ACADEMIC TEXTS USING A SUPPORT VECTOR MACHINE MODEL) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร.วิโรจน์ อรุณมานะกุล, 79 หน้า.

วิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์เพื่อพัฒนาระบบการตรวจเทียบภายในหาการลักลอกงานวิชาการในภาษาไทยด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (SVM.) โดยเปรียบเทียบประสิทธิภาพของระบบระหว่างแบบจำลองที่ใช้ข้อมูลรับเข้าเป็นคำกับแบบจำลองที่ใช้ข้อมูลรับเข้าเป็นตัวอักษร ประสิทธิภาพของลักษณ์ทางสถิติและลักษณ์ทางภาษาที่มีผลกับแบบจำลองและความแม่นยำของการหาคำตอบเมื่อพิจารณาจากความยาวของข้อความที่ลักลอก

งานวิจัยนี้ใช้คลังข้อมูลที่สร้างจากวิทยานิพนธ์ภาษาไทยระดับบัณฑิตศึกษา จุฬาลงกรณ์มหาวิทยาลัย จำนวน 300 เล่ม จำนวนคำทั้งสิ้น 5,155,589 คำ ใช้แบบจำลองทางสถิติซัพพอร์ตเวกเตอร์แมชชีน ในโปรแกรม weka เวอร์ชัน 3.7.10 ทดลองกับข้อมูลรับเข้าเป็นย่อหน้าแบบคำและแบบตัวอักษร ใช้การเรียนรู้ระบบแบบ supervised learning ให้คำตอบ 2 ประเภท คือ ใช่สำหรับย่อหน้าที่มีการลักลอก และไม่ใช่สำหรับย่อหน้าที่ไม่ได้ลักลอก ผลการทดลองกับลักษณ์ทางสถิติพบว่าชุดลักษณ์ที่ให้ผลดีที่สุดในการตรวจหาย่อหน้าลักลอก คือ ชุดลักษณ์ทางสถิติ จำนวน 7 ลักษณ์ จากข้อมูลรับเข้าแบบคำ สามารถตรวจจับย่อหน้าที่ลักลอกได้ถูกต้อง 318 ย่อหน้า จาก 735 ย่อหน้า มีค่าความครบถ้วนที่ 0.43 สำหรับ สำหรับการทดลองกับลักษณ์ทางภาษา ที่เปรียบเทียบค่าเฉลี่ยคำที่มีความถี่สูงสุด การเลือกใช้คำและชุดคำเขียนผิดพบว่า ลักษณ์ประเภทนี้ไม่สามารถแยกประเภทของย่อหน้าทั้ง 2 ประเภทได้ แม้จะพบการใช้ต่างกันจริงในข้อมูล ปัจจัยที่ทำให้แบบจำลองไม่ได้ผลเนื่องจากลักษณ์นั้นๆพบแบบไม่สม่ำเสมอในคลังข้อมูล สำหรับปัจจัยเรื่องความยาวของย่อหน้าลักลอกต่อการตรวจเทียบภายใน ผลจากการทดลองนี้ยังไม่สามารถระบุถึงความสัมพันธ์ของความยาวย่อหน้าที่มีต่อความแม่นยำในการตรวจจับได้ เพราะย่อหน้าลักลอกที่ตรวจจับได้ถูกต้องมากที่สุดในการทดลอง คือ ย่อหน้าลักลอกขนาดกลางและขนาดยาวซึ่งมีผลตรวจจับผิดพลาด 16.55% และ 36.67% ตามลำดับ ขณะที่ ไม่สามารถตรวจจับย่อหน้าขนาดสั้นได้เลย คือมีผลตรวจจับผิดพลาด 100%

ภาควิชา ภาษาศาสตร์ ลายมือชื่อนิสิต .....

สาขาวิชา ภาษาศาสตร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2556

5380173722 : MAJOR LINGUISTICS
KEYWORDS: INTRINSIC PLAGIARISM-THAI ACADEMIC WRITING-SUPPORT VECTOR MACHINE

SIWAPORN TUANTHAISONG: AN INTRINSIC PLAGIARISM DETECTION OF THAI ACADEMIC TEXTS USING A SUPPORT VECTOR MACHINE MODEL. ADVISOR: ASSOC. PROF. WIROTE AROONMANAKUN, Ph.D., 79 pp.

The main purpose of this study is to develop the intrinsic plagiarism detection in Thai academic writing system using Support Vector Machine model (SVM.) as well as comparing performance of two different kinds of input and feature and then analyzes whether the length of input has an effect on accuracy.

This study uses 300 pieces of master theses of undergraduate students from Chulalongkorn University consists of 5,155,589 words in total. Support Vector Machine model applied in the research is libsvm available in weka 3.7.10 software. To compare the performance of word-based and character-based inputs, both types of input are prepared from the same data and use the same set of statistic features in experiments. Supervised learning is applied to train the model with 2 answers, “yes” for plagiarized paragraph and “no” for non-plagiarized paragraph. Result from word-based input using the set of 7 statistic features shows the best recall score at 0.43 on testing data while 318 out of 735 plagiarized paragraphs are correctly classified. A demonstrative experiment in linguistics feature using spelling variation fails to correctly identify plagiarized paragraphs though those linguistic features are found in some plagiarized paragraphs. The reason why these linguistic features could not be used in the model is because they do not occur regularly in plagiarized paragraphs. To examine whether length of input has an effect on the model, the correct answers are grouped by their length, however, the analysis still could not shows any relation between the performance and length of data as a result of 16.55%, 36.67 % and 100% wrong prediction in middle length, long length and short length plagiarized paragraph respectively.

Department: Linguistics Student's Signature .....
Field of Study: Linguistics Advisor's Signature .....
Academic Year: 2013

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณ รองศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็น อย่างสูง ที่ได้ให้คำแนะนำและความช่วยเหลือในการทำวิจัย ตลอดจนปรับแก้วิทยานิพนธ์ฉบับนี้ จนสำเร็จลุล่วงไปด้วยดี และขอขอบคุณ รองศาสตราจารย์ดร. สมชาย ประสิทธิ์จูตระกูล และ ดร. เทพชัย ทรัพย์นิธิ กรรมการสอบวิทยานิพนธ์ที่ได้ให้ข้อชี้แนะและเสียสละเวลาในการตรวจแก้วิทยานิพนธ์ฉบับนี้ให้มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบคุณคณาจารย์ในภาควิชาภาษาศาสตร์ที่ให้ความรู้และช่วยเหลือผู้วิจัยตลอดระยะเวลาที่ศึกษาอยู่ที่ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ขอขอบคุณสมาชิกในครอบครัวของข้าพเจ้า พ่อ แม่ และพี่สาวทั้ง 2 คน ที่ให้กำลังใจและสนับสนุนการศึกษาของข้าพเจ้าอย่างเต็มความสามารถ

ขอบคุณเพื่อนๆ ชาวภาษาศาสตร์ จุฬาลงกรณ์ คุณนัชชา ถิระสาโรช และอีกหลายคนที่ข้าพเจ้าไม่สามารถเอ่ยชื่อได้หมด สำหรับความช่วยเหลือทั้งในและนอกชั้นเรียนภาษาศาสตร์ รวมถึงพี่ๆเจ้าหน้าที่ในภาควิชาทุกท่านที่ช่วยเหลือผู้วิจัยตลอดระยะเวลาการศึกษา

สุดท้ายนี้ขอขอบคุณบรรดานักวิจัยทุกท่านที่ข้าพเจ้าได้ศึกษาจากผลงานของท่าน เพื่อเป็นความรู้ และเป็นประโยชน์กับข้าพเจ้าในงานวิจัยนี้

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาของปัญหาและความสำคัญ

ปัญหาการลักลอกงานวิชาการ หรือ การนำผลงานผู้อื่นมาใช้โดยไม่อ้างอิง นอกจากจะเป็นการละเมิดจรรยาบรรณนักวิจัยซึ่งต้องมีคุณธรรมทางวิชาการ ไม่ลอกเลียนผลงานผู้อื่น พึงให้เกียรติและอ้างถึงบุคคลหรือแหล่งที่มาของข้อมูลที่ใช้ในการวิจัยแล้ว [1] พฤติกรรมดังกล่าวอาจเป็นเหตุให้ผู้กระทำขาดความน่าเชื่อถือ และได้รับบทลงโทษจากการกระทำดังกล่าว เช่น กรณีมหาวิทยาลัยแห่งหนึ่งในประเทศไทยเรียกคืนวุฒิการศึกษาระดับปริญญาดุษฎีบัณฑิต เนื่องจากตรวจพบว่าวิทยานิพนธ์ที่ได้รับอนุมัติเข้าข่ายลักลอกผลงาน [2] การให้ความรู้เกี่ยวกับการอ้างอิงผลงานอย่างถูกต้อง เป็นวิธีหนึ่งที่ช่วยแก้ปัญหาการลักลอกงานวิชาการโดยรู้เท่าไม่ถึงการณ์ อย่างไรก็ดีปัญหาการลักลอกผลงานผู้อื่นโดยเจตนา ได้นำไปสู่การพัฒนาเครื่องมือตรวจเทียบหาการลักลอกงานวิชาการด้วยโปรแกรมคอมพิวเตอร์

แนวคิดในการสร้างเครื่องมือตรวจเทียบหาการลักลอกงานวิชาการ แบ่งออกเป็น 2 ประเภท คือ 1. การตรวจเทียบภายนอกหาการลักลอกงานวิชาการ (External plagiarism detection or Extrinsic plagiarism detection) แนวคิดนี้จะนำผลงานที่ต้องสงสัยไปตรวจเทียบกับต้นฉบับที่เก็บไว้เป็นฐานข้อมูล ตัวอย่างโปรแกรมในประเภทนี้ เช่น Turnitin และ IThenticate 2. การตรวจเทียบภายในหาการลักลอกงานวิชาการ (Internal plagiarism detection or Intrinsic plagiarism detection) แนวคิดนี้จะนำผลงานที่ต้องสงสัยไปตรวจสอบลักษณะการเขียนที่เปลี่ยนแปลงไป โดยตั้งสมมติฐานว่า ส่วนที่มีลักษณะการเขียนต่างออกไปและไม่มีการอ้างอิงเป็นส่วนที่ลักลอกมาจากงานผู้อื่น โดยผู้วิจัยเห็นว่าการตรวจเทียบลักษณะนี้แม้จะไม่สามารถบ่งชี้ถึงที่มาของเอกสารต้นฉบับได้ แต่ก็สามารถตรวจพบการลักลอกงานวิชาการในกรณีที่ไม่มีต้นฉบับสำหรับใช้ตรวจเทียบหรือมีจำนวนจำกัด และอาจครอบคลุมถึงงานที่ผู้เขียนไม่ได้เขียนเองทั้งหมดแต่มีบางส่วนเขียนโดยนักเขียนเงา (ghost writer)

เนื่องจากข้อจำกัดของการตรวจเทียบภายนอกหาการลักลอกงาน คือ ต้องการฐานข้อมูลต้นฉบับแบบดิจิตัลเพื่อใช้เปรียบเทียบกับเอกสารที่ต้องสงสัย จึงมีความเป็นไปได้ที่เอกสารลักลอกมาจากแหล่งที่ไม่ได้เผยแพร่แบบดิจิตัล เช่น หนังสือต่างๆ ที่เป็นผลงานก่อนอินเทอร์เน็ตจะเข้ามามีบทบาทในการเก็บข้อมูลและสืบค้นอาจตรวจหาไม่พบการลักลอก ด้วยเหตุดังกล่าวในงานวิจัยนี้ผู้วิจัยจึงสนใจศึกษาการตรวจเทียบภายในหาการลักลอกงานวิชาการ (Internal or Intrinsic plagiarism)

แนวคิดหลักของการตรวจเทียบภายในเพื่อหาการลักลอกงานวิชาการ เกี่ยวข้องกับการมองหาข้อความที่มีลักษณะภาษาที่แตกต่างจากส่วนอื่นๆ ในเอกสารนั้น ซึ่งสะท้อนลักษณะการเขียนเฉพาะบุคคลได้ ดังนั้นการศึกษาลักษณะเฉพาะบุคคลของผู้เขียน จึงมีส่วนสำคัญเพราะทำให้สามารถนำลักษณะลักษณ์ (feature) ต่างๆ ของการเขียน เช่น การใช้เครื่องหมายวรรคตอน ลักษณะการใช้ช่องว่าง การเลือกใช้คำศัพท์ ฯลฯ มาใช้ร่วมกับหลักการทางคณิตศาสตร์เพื่อนำมาประกอบการตรวจเทียบภายในหาการลักลอกงานวิชาการได้ ทั้งนี้ก่อนที่จะนำข้อมูลที่ต้องสงสัยไปตรวจเทียบภายในหาการลักลอกนั้น ข้อมูลรับเข้าจะต้องระบุขอบเขตของคำให้ได้ก่อนจึงสามารถนำไปทดลองและประมวลผลต่อไป ด้วยเหตุนี้ในภาษาซึ่งไม่มีการระบุขอบเขตคำในงานเขียนอย่างภาษาไทยหรือภาษาจีนจึงจำเป็นต้องนำข้อมูลรับเข้าไปผ่านกระบวนการตัดคำก่อน (word segmentation) ในงานวิจัยบางงาน จึงมีแนวคิดใช้ข้อมูลรับเข้าเป็นตัวอักษร (character) เลย เพราะจะมีความเป็นอิสระและสามารถใช้ได้กับทุกภาษา โดยเฉพาะอย่างยิ่งเหมาะกับภาษาตะวันออก อย่างเช่นภาษาไทยและจีน เนื่องจากเป็นภาษาที่ไม่มีขอบเขตของคำและประโยค อีกทั้งยังไม่ต้องการการกำกับข้อมูลใดๆ ก่อน [3]

อย่างไรก็ตาม ผู้วิจัยยังไม่พบว่ามีการศึกษาวิจัยการตรวจเทียบภายในหาการลักลอกงานวิชาการในภาษาไทยมาก่อน ผู้วิจัยจึงสนใจว่าการตรวจเทียบภายในหาการลักลอกผลงานวิชาการภาษาไทยแบบที่ไม่อิงความรู้ทางภาษา คือแบบที่มองข้อมูลรับเข้าเป็นตัวอักษรกับแบบที่มีการอ้างอิงคำหรือแบบที่ข้อมูลรับเข้าเป็นคำนั้นจะมีประสิทธิภาพแตกต่างกันหรือไม่ นอกจากนี้ผู้วิจัยยังสนใจศึกษาหาลักษณะลักษณ์ (feature) ที่เหมาะสมต่อการใช้ตรวจเทียบภายในสำหรับภาษาไทย เช่น การเขียนคำผิด การเลือกใช้คำที่มีความหมายเดียวกัน ลักษณะการใช้ช่องว่าง ความยาวเฉลี่ยของข้อความ เป็นต้น

การตรวจเทียบภายในหาการลักลอกทางวิชาการนั้น จุดสำคัญอยู่ที่การหาลักษณ์ (feature) เพื่อใช้บ่งบอกลักษณะเฉพาะของผู้เขียนผลงานนั้นๆ ในงานวิจัยที่ผ่านมามีการใช้แบบจำลองการเรียนรู้ด้วยเครื่องหลายประเภท (learning machine) เพื่อช่วยในการสกัดลักษณ์เฉพาะบุคคลของผู้เขียน [4] เช่น ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) โครงข่ายประสาทเทียม (Neural Networks) และเบย์เซียน แคสสิไฟเออร์ (Bayesian Classifiers) [5] ในงานวิจัยนี้ ผู้วิจัยสนใจที่จะใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) ในการสกัดคุณสมบัติเพื่อใช้ในการตรวจเทียบภายในหาการลักลอกงานวิชาการ ซึ่งแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนได้รับความสนใจอย่างกว้างขวาง และถูกใช้เพื่อการจดจำรูปแบบและแก้ปัญหาการจัดกลุ่มข้อมูลทั้งในงานวิจัยด้านวิทยาศาสตร์และมนุษยศาสตร์ เช่น การทดลองที่ใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนเพื่อรู้จำลายมือพยัญชนะไทย และสามารถจำแนกกลุ่มพยัญชนะได้ความถูกต้องมากกว่า 86 เปอร์เซ็นต์ [6] นอกจากนี้ ยังมีการใช้แบบจำลอง 3 ประเภทเพื่อเปรียบเทียบการทำนาย

โรคพาร์กินสัน พบว่าแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนมีความแม่นยำ 92.19 เปอร์เซ็นต์ ในขณะที่ตัวแบบทำนายที่ใช้โครงข่ายประสาทเทียม (Neural Network) และเทคนิคต้นไม้ตัดสินใจ (Decision Tree) มีความแม่นยำ 90.10 เปอร์เซ็นต์ และ 88.02 เปอร์เซ็นต์ ตามลำดับ [7]

1.2 วัตถุประสงค์

1. พัฒนาระบบการตรวจเทียบภายในหาการลักลอกงานวิชาการในภาษาไทยโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน
2. เปรียบเทียบประสิทธิภาพระบบการตรวจเทียบภายในหาการลักลอกงานวิชาการในภาษาไทย ระหว่างแบบจำลองที่ใช้ข้อมูลรับเข้าเป็นคำกับแบบจำลองที่ใช้ข้อมูลรับเข้าเป็นตัวอักษร
3. ทดสอบผลของขนาดความยาวของสายคำที่ลักลอกมาจากต้นฉบับที่มีต่อค่าความแม่นยำในการตรวจจับ

1.3 สมมติฐาน

1. แบบจำลองที่รับข้อมูลเข้าเป็นสายคำได้ผลดีกว่าแบบที่รับข้อมูลเข้าเป็นสายอักขระ
2. ลักษณ์ที่ใช้ลักษณะทางภาษาศาสตร์โดยตรง เช่น การใช้คำไวยากรณ์ คำสรรพนาม จะให้ผลการทดสอบที่ดีกว่า ลักษณ์ที่ที่ไม่ได้ใช้ลักษณะทางภาษาศาสตร์โดยตรง เช่น ค่าเฉลี่ยความยาวของคำ ค่าเฉลี่ยจำนวนพยางค์ในคำ ความยาวเฉลี่ยของข้อความ
3. ข้อความที่ลักลอกมายิ่งมีความยาวมากจะยิ่งตรวจจับได้แม่นยำมากขึ้น

1.4 ขอบเขตของการวิจัย

เลือกสร้างคลังข้อมูลฝึกฝนและทดสอบเฉพาะลักษณะการลักลอกแบบไม่มีการเปลี่ยนแปลงจากต้นฉบับ (cut and paste) ประเภทเอกสารทางวิชาการเท่านั้น โดยใช้ข้อมูลวิทยานิพนธ์ภาษาไทยที่เป็นลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย จำนวน 300 เล่ม เป็นวิทยานิพนธ์ระดับบัณฑิตศึกษาของสาขามนุษยศาสตร์ จำนวน 150 เล่ม และวิทยานิพนธ์ระดับบัณฑิตศึกษาของสาขาวิทยาศาสตร์การแพทย์ จำนวน 150 เล่ม ซึ่งทั้งหมดมีเผยแพรที่คลังปัญญาจุฬาเพื่อประเทศไทย (CUIR) โดยกำหนดข้อมูลที่ลอกมาจากชิ้นงานต้นฉบับเป็นแบบสายคำ ที่มีความยาวแตกต่างกันใน 3 ขนาด คือ ขนาดสั้น (50-100 คำ) ขนาดปานกลาง (101-200 คำ) ขนาดยาว (มากกว่า 200 คำ) ทั้งนี้ส่วนที่นำมาทดลองในงานวิจัยนี้เป็นประเภทตัวบทเท่านั้น ไม่รวมส่วนของตารางและกราฟแสดงผล

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. เป็นแนวทางการตรวจเทียบภายในหาการลักลอกงานวิชาการในภาษาอื่นๆ
2. เป็นประโยชน์ต่อการศึกษาวัจนลีลา

1.6 วิธีการดำเนินการวิจัย

1. ทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องเกี่ยวกับการตรวจเทียบภายในหาการลักลอกงานวิชาการ
2. เก็บรวบรวมข้อมูลและสร้างคลังข้อมูลโดยแบ่งข้อมูลออกเป็น 2 ส่วนคือข้อมูลฝึกฝน 90 เปอร์เซ็นต์ และข้อมูลทดสอบ 10 เปอร์เซ็นต์
3. แบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลที่ผ่านกระบวนการตัดคำ และชุดข้อมูลที่เป็นสายอักขระโดยไม่มีการตัดคำ
4. กำหนดคุณสมบัติที่จะใช้ในการฝึกฝนและทดสอบ
5. พัฒนาระบบการตรวจเทียบภายในหาการลักลอกงานวิชาการในภาษาไทยโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน
6. ทดสอบระบบการตรวจเทียบภายในหาการลักลอกงานวิชาการในภาษาไทย
7. ประเมินผล วิเคราะห์ และสรุปผลการวิจัย

1.7 เครื่องมือที่ใช้ในการวิจัย

1. โปรแกรมภาษา Perl จาก www.perl.org/docs.html
2. นำเข้าแบบจำลองทางสถิติซัพพอร์ตเวกเตอร์แมชชีน libsvm และทดลองผ่านโปรแกรม Weka 3.7.10 จาก www.cs.waikato.ac.nz/ml/weka
3. โปรแกรมตัดคำ CUThai Segmentation version 2.01 ของภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
4. โปรแกรม PDFill PDF tool 10

บทที่ 2
ทบทวนวรรณกรรม

ในส่วนนี้จะกล่าวถึงความหมายของการลักลอกผลงานโดยทั่วๆไป และความหมายเฉพาะของการลักลอกงานทางวิชาการที่มีผู้ให้ความหมายไว้ นอกจากนี้จะกล่าวถึงลักษณ์ (feature) ประเภทต่างๆ ที่ใช้ในการระบุลักษณะการเขียนเฉพาะของบุคคล (authorship attribution) งานวิจัยที่ผ่านมา และแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนเพื่อเป็นแนวทางในการทำวิจัยครั้งนี้

2.1 ความหมายของการลักลอกผลงาน (Plagiarism)

การลักลอกผลงาน คือ การที่นำความคิดของคนอื่นมาใช้โดยไม่อ้างอิงเหมือนกับความคิดหรือข้อความนั้นเป็นข้อความของผู้เขียนเองทั้งแบบที่ลักลอกจากภาษาเดียวกันและการลักลอกแบบข้ามภาษา โดยการลักลอบแบบข้ามภาษา คือ การแปลจากต้นฉบับภาษาหนึ่งแล้วนำมาเขียนเป็นอีกภาษาหนึ่ง รวมถึงมีการอ้างอิงแหล่งที่มาที่ถูกต้องแต่เนื้อหาส่วนใหญ่ในงานนั้นเป็นส่วนที่อ้างอิงมามากกว่าความคิดของผู้เขียนเอง ซึ่งก็ถือเป็นการลักลอกผลงานอย่างหนึ่ง แม้กระทั่งเจ้าของผลงานเองเมื่อนำผลงานของตนเองมาเขียนซ้ำโดยไม่อ้างอิงที่มา ก็ถือเป็นการลักลอกที่เรียกว่า self plagiarism ทั้งนี้สาเหตุที่ทำให้เกิดการลักลอกทางวิชาการนั้นเกิดขึ้นได้แบบไม่เจตนา เช่น อ้างอิงที่มาอย่างไม่ถูกวิธี และแบบเจตนา เช่น ตั้งใจคัดลอกมาโดยไม่อ้างอิง [8] สำหรับการลักลอกผลงานทางวิชาการนั้นมีการให้ความหมายเพิ่มเติมดังนี้

การลักลอกงานวิชาการ (Academic Plagiarism) เกิดขึ้นเมื่อผู้เขียนอ้างอิง หรือกล่าวซ้ำข้อมูลจากอีกแหล่งหนึ่งมากเกินกว่าสี่คำโดยปราศการการอ้างอิงแหล่งที่มา ทั้งนี้พฤติกรรมดังกล่าวถือว่าผู้เขียนตั้งใจที่จะเสนอผลงานดังกล่าวว่าเป็นผลงานของผู้เขียนเอง [9]

สำหรับพฤติกรรมการลักลอกงานประเภทวิชาการของนักเรียนนั้น [10] ได้รวบรวมไว้ว่ามีหลักๆ อยู่ 4 แบบ คือ 1. ขโมยผลงานของผู้อื่นและนำมาเสนอว่าเป็นผลงานของตนเอง เช่น ซื้อผลงานมาจากผู้ที่รับจ้างเขียนงานแทน เช่น เวบไซค์ www.cheathouse.com หรือ www.schoolsucks.com รวมถึงการลักลอกผลงานมาแบบไม่อ้างอิง 2. ส่งผลงานที่เขียนโดยผู้อื่น เช่น เพื่อนหรือญาติ 3. อ้างอิงที่มาของผลงานอย่างถูกต้องแต่ไม่ใส่เครื่องหมาย”-“ ทำให้เกิดความเข้าใจผิดว่าเป็นข้อความที่ผู้วิจัยนำมาเขียนใหม่มากกว่าเป็นการอ้างอิงมาโดยตรง 4. นำข้อมูลมาเขียนใหม่โดยไม่อ้างอิงที่มาอย่างเหมาะสม

2.2 ลักษณะการลักลอกงานเขียนที่พบ

การลักลอกงานเขียนจากผลงานของผู้อื่นโดยไม่อ้างอิงแหล่งที่มานั้น จำแนกได้จากหลายเกณฑ์ ในที่นี้ผู้วิจัยจะกล่าวถึง หลักเกณฑ์ที่น่าสนใจ 2 อย่างคือ แบ่งตามวิธีการที่เปลี่ยนแปลงต้นฉบับ และแบ่งตามวิธีการที่เปลี่ยนแปลงต้นฉบับและปริมาณข้อความต้นฉบับที่ถูกลักลอก

2.2.1 พิจารณาวิธีการเปลี่ยนแปลงจากต้นฉบับ จะแบ่งได้เป็น 3 ลักษณะกว้างๆ ดังนี้ [11]

2.2.1.1 การลอกข้อความจากต้นฉบับมาแบบไม่เปลี่ยนแปลงอะไรเลย (copy and paste) ซึ่งการลักลอกลักษณะนี้จะตรวจพบได้ง่ายที่สุด

2.2.1.2 การลอกข้อความจากต้นฉบับมาและมีการเปลี่ยนแปลงเล็กน้อย เช่น แทนที บางคำด้วยคำพ้องความหมาย ตัดหรือเพิ่มบางคำ เพื่อให้ต่างจากต้นฉบับ

2.2.1.3 การลอกข้อความจากต้นฉบับโดยเปลี่ยนแปลงให้ต่างจากต้นฉบับ โดยการแต่งประโยคใหม่ หรือใช้โครงสร้างประโยคให้ต่างจากเดิม

2.2.2 พิจารณาจากวิธีการเปลี่ยนแปลงจากต้นฉบับและปริมาณข้อความที่ถูกลักลอกมาจากต้นฉบับ โดยไม่อ้างอิงซึ่งแบ่งได้เป็น 5 ระดับ [12] ดังนี้

2.2.2.1 ระดับที่หนึ่ง ลักลอกแบบทุกตัวอักษรมาจากต้นฉบับปริมาณมากกว่าครึ่งถึงทั้งหมดของเอกสารต้นฉบับ

2.2.2.2 ระดับที่สอง ลักลอกแบบทุกตัวอักษรมาจากต้นฉบับปริมาณน้อยกว่าครึ่งหนึ่งของเอกสารต้นฉบับ

2.2.2.3 ระดับที่สาม ลักลอกแบบทุกตัวอักษรมาจากบางส่วนของต้นฉบับ เช่น ย่อหน้า ประโยค รูปภาพ หรือตาราง

2.2.2.4 ระดับที่สี่ ลักลอกจากย่อหน้าและบางหน้าของต้นฉบับ แล้วนำมาเขียนใหม่

2.2.2.5 ระดับที่ห้า ลักลอกแบบทุกตัวอักษรมาจากต้นฉบับโดยอ้างอิงผิดวิธี

2.3 ประเภทของการตรวจเทียบหาการลักลอกงาน

งานวิจัยที่เกี่ยวกับการตรวจเทียบหาการลักลอกงาน ในปัจจุบัน แบ่งออกเป็น 2 ประเภท ดังนี้

2.3.1 การตรวจเทียบภายนอกหาการลักลอกงานวิชาการ (External plagiarism detection or Extrinsic plagiarism detection) ใช้การตรวจเทียบระหว่างงานเขียนที่น่าสงสัยกับฐานข้อมูลต้นฉบับที่เก็บไว้ และเทียบหาความคล้ายกันของเอกสารที่น่าสงสัยและเอกสารต้นฉบับ

2.3.2. การตรวจเทียบภายในหาการลักลอกงานวิชาการ (Internal plagiarism detection or Intrinsic plagiarism detection) ประเภทนี้จะนำผลงานที่ต้องสงสัยไปตรวจเทียบ

ภายในหาลักษณะทางภาษาที่เปลี่ยนไป เช่น การเปลี่ยนสรรพนาม การใช้คำไวยากรณ์ ความสั้นยาวของประโยค ซึ่งมีแนวคิดว่า โดยทั่วไปผู้อ่านจะสามารถจับลักษณะทางภาษาที่เปลี่ยนไปได้แม้ไม่ต้องอาศัยโปรแกรมตรวจเทียบใดๆ [13] [14] [15] แนวคิดนี้จึงถูกนำมาพัฒนาโดยนำเอาลักษณ์ทางภาษามาใช้ควบคู่กับหลักการทางคณิตศาสตร์ (Stylometry) เพื่อตรวจเทียบการลักลอกประเภทนี้ ตัวอย่างของการตรวจเทียบภายในดังตัวอย่างต่อไปนี้

*"**Our** goal is to identify files that came from **the same source** or contain parts that came from **the same source**. **We** say that two files are similar if they contain a significant number of common substrings that are not too small. **We** would like to find enough common substrings to rule out chance, without requiring too many so that **we** can detect similarity even if significant parts of the files are different. However, **my** interest in plagiarism lies within academic institutions, so the document domain will be local research articles. The limited scope of domain will make it easier to determine if it is **same source** or not."* [8]

ในชิ้นงานนี้มีลักษณะที่ภาษาเปลี่ยนไป 2 ส่วน คือ สรรพนาม จากเดิมที่ใช้ our และ we เปลี่ยนเป็น my และการใช้คำไวยากรณ์ the กับ คำว่า same source ไม่สม่ำเสมอ กล่าวคือ ใช้เพียง 2 ใน 3 จากคำที่พบทั้งหมด

แนวคิดเกี่ยวกับลักษณะเฉพาะบุคคลในงานเขียน ถูกนำไปประยุกต์ใช้ในงานทางภาษาธรรมชาติหลายอย่าง เช่น ในการระบุตัวตนเจ้าของผลงาน (Author verification) โดยนำงานเขียนมาพิจารณาว่า เขียนโดยบุคคลนั้นๆหรือไม่ ระบุข้อมูล เกี่ยวกับ อายุ การศึกษา เพศ ของเจ้าของผลงานจากงานเขียน (Author profiling or characterization) ใช้เป็นเครื่องช่วยในการสืบสวนหาตัวอาชญากร เช่น ใช้ลักษณะการเขียนเพื่อระบุตัวผู้เขียนจดหมายเรียกค่าไถ่และจดหมายข่มขู่ [16] รวมถึงในงานการตรวจเทียบภายในหาการลักลอกงานวิชาการที่ผู้วิจัยสนใจศึกษา

2.4 ระดับชั้นของข้อความ (Text layers)

ก่อนที่จะกล่าวถึงเกี่ยวกับลักษณ์ (feature) ที่ใช้ในงานทางภาษาธรรมชาติ [17] ได้อ้างถึงระดับชั้นของข้อความและโครงสร้างลึกที่ซ่อนอยู่ ซึ่งมีความสำคัญต่อการเข้าใจถึงแนวคิดเรื่องลักษณ์ (feature) ที่ใช้ในการตรวจเทียบลักษณะงานเขียนของบุคคล เช่น ตัวอย่างข้อความ This is a text.

Grapheme layer: This is a text.
Symbol layer: T h i s i s a t e x t .

Character bigram : (Th) (hi) (is) (si) (is) (sa] (at) (te) (ex) (xt)

Character trigram: (Thi) (his) (isi) (sis) (isa) (sat) (ext) (tex) (ext)

Token layer: (This) (is) (a) (text.)

Part-Of-Speech layer: This/DT is/VBZ a/DT text/NN ./.

Constituent layer: (This (is (a (text)))).

ตัวอย่างดัดแปลงจาก [17]

ตัวอย่างของประโยคข้างต้น แสดงถึงความเชื่อมโยงของแนวคิดเรื่องระดับชั้นต่างๆ ของข้อความกับลักษณ์ (feature) ซึ่งลักษณ์ (feature) ต่างๆ ที่นำมาใช้ในงานทางภาษาธรรมชาตินั้น พบได้เมื่อมองข้อความหนึ่งๆ เป็นระดับชั้นที่ต่างกัน ทั้งนี้ระดับชั้นของข้อความต่างก็มีโครงสร้างภายในที่ประกอบกันเป็นส่วนๆ และส่วนที่ประกอบกันนั้นล้วนมีความสัมพันธ์ต่อกัน แตกต่างกันไปในแต่ละระดับ ประเภทของส่วนย่อยที่ประกอบกันในระดับชั้นหนึ่งๆ คือตัวแทนของลักษณ์ (feature) ที่นำมาใช้ประโยชน์ในงานทางภาษาธรรมชาติ เช่น เมื่อมองประโยคตัวอย่างในระดับของคำ พบว่า This is a text. ประกอบไปด้วย 4 คำ ที่มีความถี่ (word frequency) คำละหนึ่งครั้ง เมื่อนำข้อความนี้ไปพิจารณาถึงลักษณะงานเขียนเฉพาะบุคคล ลักษณ์ที่ปรากฏอาจแสดงผลดังนี้ ผู้เขียนมักใช้ประโยคที่ไม่ซับซ้อน (simple sentence) และมีความยาวของประโยคไม่มาก (average word per sentence) ผู้เขียนมักใช้คำที่สั้น (average syllable per word) ในงานการตรวจเทียบภายในหาการลักลอก เมื่อพบลักษณะภาษาที่ไม่สม่ำเสมอหรือแปลกกว่าส่วนอื่นๆ เช่น ใช้ประโยคยาวขึ้น ใช้คำเขียนผิดและถูกปนกัน หรือใช้คำศัพท์หลายพยางค์ขึ้นก็จะสันนิษฐานว่าเป็นส่วนที่ลักลอกมา เป็นต้น

2.5 ประเภทของลักษณ์ในการตรวจเทียบลักษณะงานเขียนบุคคล (Stylometric Features)

จุดประสงค์ของการใช้ลักษณ์ (feature) เพื่อศึกษาวัจนลีลาในการเขียน ในทางภาษาศาสตร์คอมพิวเตอร์ จะมุ่งเน้นที่การคัดเลือกลักษณ์ที่สามารถตรวจวัดค่าได้ เพื่อนำผลจากการวัดค่านั้นมาตอบคำถามในงานด้านต่างๆ โดยมองข้อความเป็นคำหรือตัวอักษรที่เรียงต่อกัน ทั้งนี้ลักษณ์ (feature) ที่นำมาใช้ทั้งหมดนั้น โดยทั่วไปจะไม่ใช่คำเนื้อหาเพื่อหลีกเลี่ยงคำเฉพาะที่ขึ้นอยู่กับหัวเรื่อง คำไวยากรณ์จึงถูกนำมาใช้มากในการบ่งชี้ลักษณะการเขียนเฉพาะบุคคลประเภทของลักษณ์ (feature) ที่ใช้ในงานวิจัยที่ผ่านมา มีดังนี้

2.5.1 ลักษณ์ด้านศัพท์ (Lexical feature) แบ่งออกเป็น คุณสมบัติด้านศัพท์แบบอิงตัวอักษร (Lexical feature-character based) และคุณสมบัติด้านศัพท์แบบอิงคำ (Lexical feature-word based) ลักษณ์ประเภทย่อยที่ใช้ เช่น ความถี่ของตัวอักษร (character frequency) ความถี่ของสายตัวอักษรที่ความยาวเป็น n (character n-gram frequency) ค่าเฉลี่ยความยาวของ

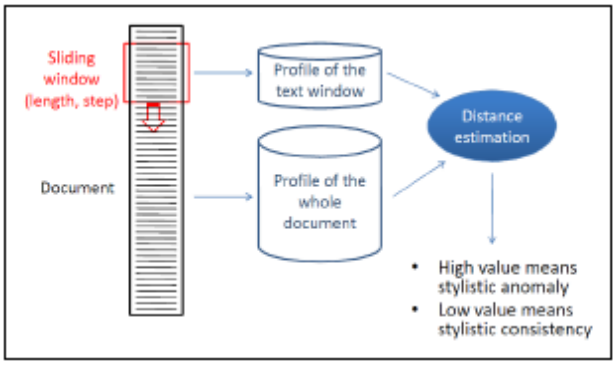
คำ (average word length) ค่าเฉลี่ยจำนวนพยางค์ต่อคำ (average number of syllables per word) ค่าเฉลี่ยความยาวของประโยค (average sentence length)

2.5.2 ลักษณ์ด้านโครงสร้างประโยค (syntactic feature) เช่น การกำกับหมวดคำ (POS) ความถี่ของสายหมวดคำที่มีความยาวเป็น n (POS n-gram frequency) ความถี่ของการใช้เครื่องหมายวรรคตอน (Frequency of punctuations) ความถี่ของการใช้คำไวยากรณ์ (Frequency of function words)

2.5.3 ลักษณ์ด้านรูปแบบ (structural feature) เช่น ค่าเฉลี่ยความยาวของย่อหน้า (Average paragraph length) การย่อหน้า (Indentation) การใช้คำขึ้นต้นและลงท้าย (Use of greetings and farewells) การใช้ลายเซ็นต์ (Use of signature)

นอกจากลักษณ์ (feature) ที่กล่าวถึงดังข้างต้นซึ่งเป็นที่นิยมในงานวิจัยสาขานี้แล้ว ประเภทของลักษณ์ที่นำมาใช้ อาจมีที่แตกต่างไปบ้างตามแต่จะคิดว่าอะไรบ้างที่จะสามารถนำมาใช้เป็นลักษณ์ได้ เช่น Biber ใช้ลักษณ์รายการคำศัพท์เพื่อระบุงานเขียนที่แสดงการโต้แย้งในภาษาอังกฤษ เช่น Almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly และ somewhat [18]

ตัวอย่างงานวิจัยที่ผ่านมาของงานตรวจเทียบภายในหาการลักลอกงานที่น่าสนใจ ถูกเสนอโดย [3] โดยใช้ข้อมูลรับเข้าแบบเป็นสายอักขระ (n gram) กำหนด จำนวน n เป็น 3 ตัวอักษร โดยวิธีการนี้เรียกว่า character n-gram profiles (CNP) ซึ่ง CNP จะมีสมาชิกย่อยๆเป็นข้อมูลของสายอักขระแบบต่างๆ ที่ขนาดความยาว (n) นำมา normalize ความถี่ด้วยความยาวตัวบท แนวคิดหลักกำหนดขนาดของวินโดว์และนำวินโดว์ไปเปรียบเทียบกับวินโดว์อื่นๆ ในชิ้นงาน ซึ่งกำหนดค่าพารามิเตอร์ที่ใช้ในการศึกษา โดยการประมาณค่าจากข้อมูลในคลังข้อมูลจำนวน 200 ชิ้น



ภาพที่ 2.1 แสดงกระบวนการ character n-gram profile

จากนั้นวัดค่าความแตกต่าง (dissimilarity measure) เพื่อเทียบหาลักษณะการเขียนที่ผิดปกติภายในชิ้นงานต้องสงสัย ดังสมการ

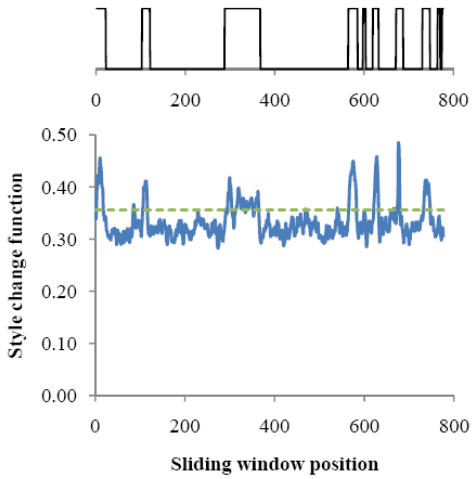
$$nd_1(A,B) = \frac{\sum_{g \in P(A)} \left(\frac{2(f_A(g) - f_B(g))}{f_A(g) + f_B(g)} \right)^2}{4|P(A)|}$$

เมื่อกำหนดให้ A และ B เป็นเอกสาร 2 ชิ้น และ P(A), P(B) เป็นโปรไฟล์ ของ A และ B ตามลำดับ ซึ่งคือเวคเตอร์ของค่าความถี่ที่ถูก normalize ของของสายอักขระ n-gram ที่ปรากฏอย่างน้อยหนึ่งครั้งในชิ้นงาน โดย $f_A(g)$ และ $f_B(g)$ เป็นค่าความถี่ที่ normalized แล้วของ n-gram g ที่พบใน A และ B ตามลำดับ ส่วน |P(A)| แสดงขนาดของโปรไฟล์ของข้อความ A ค่าของฟังชั่น nd จะแสดงความแตกต่าง อยู่ระหว่าง 0 และ 1 โดย 0 แสดงความเหมือนที่มากที่สุด การวัดค่านี้ว่า normalized nd_1 ซึ่งเมื่อคำนวณค่า nd นี้ในขอบเขตวินโดว์ต่างๆ ที่ขยับไปเรื่อยๆ โดยมี w เป็น sliding window ที่มีขนาด l และ S เป็นขนาดวินโดว์ที่ทับซ้อนกับวินโดว์ถัดไปทางขวา ฟังชั่นของลักษณะการเขียนที่เปลี่ยนแปลงไป (style change) ในแต่ละช่วงวินโดว์สามารถคำนวณโดยสมการ

$$sc(i,D) = nd_1(w_i, D),\ i = 1 \ldots |w|$$

เมื่อ |W| เป็นจำนวนวินโดว์ทั้งหมด แตกต่างกันไปตามขนาดความยาวของ text เมื่อ X แทนความยาวทั้งหมดของข้อความ จำนวน |W| คำนวนโดยสมการ

$$|w| = \lfloor 1 + \frac{x - l}{s} \rfloor$$



ภาพที่ 2.2 แสดงผลการตรวจเทียบลักลอกโดย character n-gram profile

รูปข้างบนแสดงผลการตรวจเทียบลักลอกภายในเมื่อใช้ฟังชันลักษณะการเขียนที่เปลี่ยนแปลงไป จากนั้นตรวจเทียบหาการลักลอกว่าเอกสารใดมีการลักลอกหรือไม่โดยเปรียบเทียบค่าความ

เบี่ยงเบนมาตรฐานของ sc ทั้งหมดในเอกสารนั้น ถ้าค่า S น้อยกว่าค่า threshold ที่ 0.2 เอกสารนั้นถือว่าไม่มีการลักลอก

Plagiarism-free criterion: $S<t_1$

ในการเตรียมข้อมูลก่อนนำมาทดสอบ จะเปลี่ยนตัวอักษรทั้งหมดเป็นตัวเล็ก และตัดกลุ่มของตัวอักษร 3 ตัว (tri-gram) ที่ไม่มีตัวอักษรอยู่เลยออก จากนั้นแบ่งข้อมูลเป็นวินโดว์ตามขนาดความยาวของตัวอักษรที่กำหนด ด้วยเหตุนี้จากความยาววินโดว์ ที่กำหนดไว้ 1,000 ตัวอักษร เมื่อเลือกข้อมูลจริงขนาดของวินโดว์จริงจึงมากกว่าค่าที่กำหนดนี้ อย่างไรก็ตามในงานได้กำหนดค่า threshold ของขนาดวินโดว์จริง คือ 1,500 ตัวอักษร หากข้อมูลวินโดว์ใดเกินค่านี้ก็จะทิ้งไป ทั้งนี้เพื่อลดการตรวจจับแบบ false negative เช่น ข้อมูลหน้าสารบัญออกไป

เมื่อตัดสินว่าเอกสารใดมีการลักลอกแล้ว ในการที่จะตัดสินว่าส่วนใดในเอกสารนั้นเป็นข้อความที่คัดลอกมานั้นจะใช้สมการ sc(I',D) > M' + a*S' โดยที่ M' และ S' เป็นค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานที่มีการปรับแล้ว ค่า M' และ S' คำนวณจากวินโดว์ที่ไม่มีลักษณะผิดปกติ เพราะคาดว่าวินโดว์ที่มีลักษณะผิดปกติจะเป็นส่วนที่ลักลอกมาจึงไม่ควรนำมาคำนวณค่าเฉลี่ย และส่วนเลี่ยงเบนมาตรฐานด้วย การตัดสินว่าวินโดว์ใดมีลักษณะผิดปกติคำนวณได้จากการหาค่า M และ S ซึ่งเป็นค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของวินโดว์ทั้งหมดก่อน แล้ววัดวินโดว์ที่มีค่า sc เกินกว่าค่า M+S ออก

วิธีการตรวจจับแบบ CNP นี้ชนะเป็นอันดับหนึ่งในการแข่งขันตรวจเทียบภายในหาการลักลอกงานในปี 2009 (PAN 2009) โดยได้ค่า F-measure ที่ 0.3086 อย่างไรก็ตามยังพบว่าความยาวของสายตัวอักษรที่น้อยกว่า 3,000 ตัวอักษร มีผลต่อค่าความแม่นยำในการตรวจเทียบภายในโดยวิธีการนี้ และเมื่อลองเปลี่ยนความยาวของตัวอักษรเป็น 4 และ 5 ตัวอักษร ก็ยังได้ค่าความแม่นยำที่ไม่ต่างจากสายตัวอักษรขนาด 3 ตัวอักษร

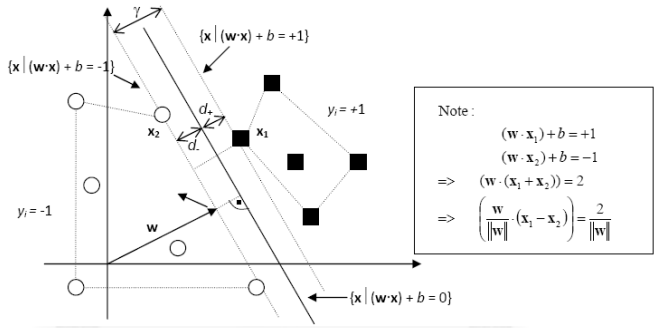
2.6 แนวคิดเกี่ยวกับการใช้แบบจำลองทางสถิติ (Machine learning)

การใช้แบบจำลองทางสถิติเพื่อการเรียนรู้ (Machine learning) เป็นศาสตร์ที่ศึกษาถึงอัลกอลิทึมของคอมพิวเตอร์ เพื่อสอนให้คอมพิวเตอร์เรียนรู้และปรับเปลี่ยนพฤติกรรมการตอบสนองต่อข้อมูลต่างๆให้เป็นไปอย่างอัตโนมัติในการสกัดหาและเลือกลักษณ์ (feature selection and extraction) การใช้อัลกอรึทึมการจำแนกประเภท (classification algorithms) เช่น ซัพพอร์ตเวกเตอร์แมชชีนจะช่วยจัดการกับข้อมูลที่มีมิติของข้อมูลสูง อีกทั้งจะช่วยจัดการเลือกคุณสมบัติ และจำกัดจำนวนของรายการลักษณ์ (feature) และเลือกใช้ลักษณ์ที่เหมาะสมที่สุด เป็นผลให้ค่าความแม่นยำของการจัดกลุ่มข้อมูลเพิ่มขึ้น ดังตัวอย่างการทดลองได้ผลค่าความแม่นยำจาก 97.85 เปอร์เซ็นต์ เป็น 99.01 เปอร์เซ็นต์ เมื่อจำกัดจำนวนลักษณ์ลงและเลือกใช้เพียง 134 ลักษณ์ที่

เหมาะสมที่สุด [19] โดยในงานวิจัยนี้ ผู้วิจัยจึงสนใจใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน เพื่อใช้ลักษณ์ทางสิถิติและลักษณ์ทางภาษาศาสตร์เพื่อตรวจเทียบภายในหาการลักลอกงานวิชาการ

2.6.1 แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

แบบจำลองนี้ถูกเสนอโดย Vapnik ในปี 1999 โดยใช้การเรียนรู้บนพื้นฐานทางสถิติ ซึ่งการทำงานหลักของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน จะสร้างแบบจำลอง (model) ที่ได้มาจากเรียนรู้จากกลุ่มข้อมูลตัวอย่าง (ข้อมูลฝึกฝน training data) และนำแบบจำลองนั้นไปประมาณการณ์ (predict) กลุ่มตัวอย่างในอนาคต ซัพพอร์ตเวกเตอร์แมชชีนจะสร้างไฮเปอร์เพลนที่เหมาะสมในระนาบของข้อมูลตัวอย่าง (training data) โดยพยายามสร้างระยะห่างระหว่างข้อมูล 2 ประเภทให้มากที่สุด โดยนิยามระยะห่างของจุดทั้งสองส่วนที่อยู่ใกล้เส้นไฮเปอร์เพลนมากที่สุด เป็น (d_+) และ(d_-) ระยะขอบหรือมาร์จิ้น (margin) เกิดจากระยะที่ (d_+) +(d_-) ไฮเปอร์เพลนที่เหมาะสมที่สุดคือระยะที่มาร์จิ้น γ (margin) กว้างที่สุด โดยเรียกข้อมูลบนขอบมาร์จิ้นว่า ซัพพอร์ตเวกเตอร์ (support vector) ลักษณะการจัดกลุ่มข้อมูลดังรูปภาพที่ 2.3



ภาพที่ 2.3 แสดงผลการจัดกลุ่มข้อมูลแบบซัพพอร์ตเวกเตอร์แมชชีน

เมื่อกำหนดให้ คลังข้อมูลฝึกฝนประกอบด้วยตัวอย่าง จำนวน l แสดงในรูป $\{x_k, y_k\}, k = 1,\ldots,l$ และ $x_k \in \Re^n$, $y_k \in \{-1,+1\}$ โดย Xk แสดงอินพุทเว็คเตอร์ และ Yk เป็นชนิดของข้อมูลไฮเปอร์เพลนที่เหมาะสมกับระนาบข้อมูลจะถูกกำหนดด้วยพารามิเตอร์ (W,b) เมื่อ W เป็นเวกเตอร์ที่ตั้งฉากกับไฮเปอร์เพลน และ b เป็นค่าคงที่ที่สัมพันธ์กับตำแหน่งดั้งเดิมของข้อมูลก่อนการแปลงเป็นข้อมูลมิติสูงขึ้น สมการของไฮเปอร์เพลนเชิงเส้นจะกำหนดด้วยสมการ (W*X)+b=1 เพื่อลดปัญหาในเรื่องสเกล W และ b ก็ถูกกำหนดด้วยสมการ | (W*X)+b=1| สำหรับจุดที่อยู่ใกล้ไฮเปอร์เพลนที่มากที่สุด สมการของไฮเปอร์เพลนเพื่อแสดงกลุ่มข้อมูลเชิงเส้นแสดงได้ ดังสมการ

$$y_i\left[(\mathrm{w}\cdot \mathrm{x}_i)+b\right] \geq 1 \qquad \forall i$$

แต่เดิมแบบจำลองนี้ถูกนำมาใช้กับข้อมูลเชิงเส้น แต่ลักษณะข้อมูลจริงนั้นมักเป็นข้อมูลแบบไม่เป็นเชิงเส้น (nonlinear dataset) จึงใช้เคอร์เนลมาใช้แก้ปัญหาโดยใช้ฟังชั่นเพื่อย้ายข้อมูลจาก input space ไปยัง feature space โดยทำให้ข้อมูลเรียงตัวในมิติที่สูงขึ้น เรียกว่าพื้นที่มิติสูง (Higher Dimensional Space) สมการแสดงข้อมูลเชิงไม่เส้นเพื่อใช้แปลงข้อมูลไปสู่มิติที่สูงขึ้น

- Maximize

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

- Subject to (1) $\sum_{i=1}^{l} \alpha_i y_i = 0$, and

(2) $0 \le \alpha_i \le C \qquad \forall i$.

เมื่อตัวแปร $\alpha_i \ge 0$ เรียกว่า Positive Lagrange Multiplier และ K(xi,xj) คือ ฟังก์ชันเคอร์เนล และ C เป็นค่าคงที่ที่ใช้ปรับหรือชดเชยค่าความผิดพลาดในข้อมูลฝึกฝน และความซับซ้อนของแบบจำลองโดยเคอร์เนลที่นิยมใช้มีอยู่ 3 ชนิดด้วยกัน คือ

โพลิโนเมียล (Polynomial)

$$K(\mathbf{x}, \mathbf{y}) = (\gamma(\mathbf{x} \cdot \mathbf{y}) + \beta)^d$$

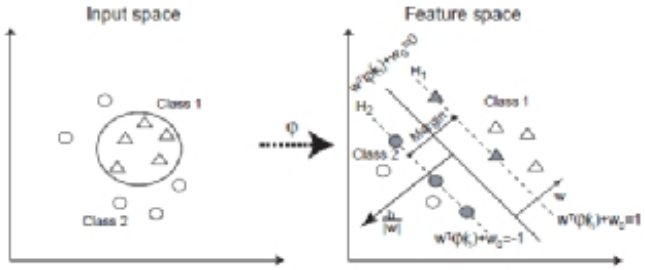
เรเดียลเบสิสฟังก์ชัน (Radial Basis Function-RBF)

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$$

และซิกมอยด์ (Sigmoid)

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\gamma(\mathbf{x} \cdot \mathbf{y}) + \beta)$$

เมื่อ γ, β และ d คือพารามิเตอร์ของเคอร์เนล



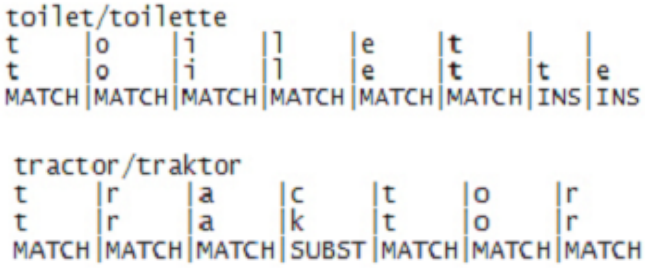
ภาพที่ 2.4 แสดงรูปแบบการจัดข้อมูลแบบ input space ใหม่ เป็นข้อมูล feature space ที่เรียงตัวในมิติสูงขึ้น

2.6.2 แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนกับงานด้านภาษาธรรมชาติ

ในงานทางด้านภาษาธรรมชาติ (NLP) มีการใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (SVM) ในหลายสาขา เช่น ใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการรู้จำชื่อเฉพาะ (Named Entity) เช่น ชื่อสถานที่ ชื่อบุคคล รวมถึง จำนวนตัวเลข ในภาษาทางตอนใต้ของอินเดียชื่อ Mayalalam ซึ่งมีผู้พูดราว 35.9 ล้านคนทั่วโลก ลักษณะของภาษานี้คือ จะเรียงคำในตำแหน่งใดก็ได้ และเป็นภาษาที่ใช้ prefix และ suffix มาก อีกทั้งไม่มีการขึ้นต้นคำด้วยอักษรตัวใหญ่เช่นในภาษาอังกฤษ นอกจากลักษณะของภาษาที่กล่าวมาแล้ว ยังมีข้อจำกัดเรื่องคลังข้อมูลภาษาที่ผ่านการกำกับหมวดคำ ที่ต้องสร้างขึ้นเอง ในงานครั้งนี้ได้ใช้ข้อมูลจากข่าวและนิตยสาร ที่มีหัวเรื่องใหญ่ๆ แตกต่างกันจำนวน 5 ประเภท คือ การเมือง สุขภาพ กีฬา วิทยาศาสตร์ และเกษตรศาสตร์ ใช้ข้อมูลฝึกฝนจำนวน 8,000 ประโยค โดยใช้ prefix และ suffix เป็นลักษณ์เพื่อใช้แยกประเภทของคำที่เป็นชื่อเฉพาะ นอกจากนั้นยังใช้ลักษณ์ของวันที่ เวลา ปี และอายุ เพื่อชนิดย่อยของชื่อเฉพาะ จากนั้นใช้ข้อมูลทดสอบจำนวน 2,000 ประโยค โดยผลของงานด้านการรู้จำชื่อเฉพาะมีประโยชน์ต่องานในหลายแขนง เช่น ในระบบตอบคำถามอัตโนมัติ การสกัดข้อมูล และสรุปความได้ผลการทดสอบคือ ชื่อเฉพาะประเภท สถานที่ (location) ได้ค่าความครบถ้วนที่มากที่สุด คือ 96.21 เปอร์เซ็นต์ ค่าความแม่นยำที่ 95.30 เปอร์เซ็นต์ และค่าเอฟสกอร์ที่ 95.75 เปอร์เซ็นต์ ค่าเฉลี่ยความครบถ้วน ค่าความแม่นยำ และค่าเอฟสกอร์ ทั้งหมดอยู่ที่ 89.12 เปอร์เซ็นต์ 89.15เปอร์เซ็นต์ และ 89.13 เปอร์เซ็นต์ ตามลำดับ [20]

ตัวอย่างของงานอีกแขนงหนึ่งคือ การใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการพยากรณ์รายการคำร่วมเชื้อสายในภาษาเป้าหมาย [21] แนวคิดเรื่องการพยากรณ์คำเชื้อสายนั้น เริ่มมาจากงานของ Levinshtein ในปี 1965 โดยพิจารณาจากคำร่วมเชื้อสายที่มีตัวสะกดคล้ายกัน โดย Levinshtein ได้คำนวนการเปลี่ยนแปลงของคำร่วมเชื้อสาย ผ่าน 4 กระบวนการ คือ การจับคู่ การแทนที่ การเพิ่มเข้า และการตัดออก ซึ่งรู้จักกันในชื่อ Edit distance (ED) ประโยชน์ของรายการคำร่วมเชื้อสายเป็นประโยชน์ต่อการแปลข้ามภาษา งานวิจัยชิ้นนี้ Mulloni ได้ใช้ข้อมูลฝึกฝนเป็นรายการ

รูปเขียนคำศัพท์คู่คำร่วมเชื้อสายระหว่างภาษาอังกฤษและภาษาเยอรมันจำนวน 1,683 ชุดคำ เพื่อพยากรณ์รูปเขียนคำคู่เชื้อสายคำอื่นๆ ในภาษาเยอรมันที่เป็นข้อมูลทดสอบจำนวน 422 คำ



ภาพที่ 2.5 ตัวอย่างรายการคำคู่ร่วมเชื่อสายในคลังข้อมูลฝึกฝนของ Mulloni

เมื่อใช้คำร่วมเชื้อสายทั้งสองภาษาที่มีรูปเขียนเหมือนกันเป็นเบสไลน์ ผลของการพยากรณ์แบ่งออกเป็น 3 แบบ คือ ถูก ผิด และใกล้เคียงมาก ผลปรากฏว่าการใช้แบบจำลองจดจำรูปแบบ ให้ค่าความแม่นยำเพิ่มจากเบสไลน์ 50.58 เปอร์เซ็นต์ แต่ค่าความแม่นยำโดยรวมได้ผลเพียง 30.33 เปอร์เซ็นต์ โดยผู้วิจัยได้ให้ข้อเสนอแนะไว้ว่า คำที่ทายออกมาใกล้เคียงมากสามารถแก้ได้โดยการเพิ่มตัวอย่างในข้อมูลฝึกฝน แม้ค่าความแม่นยำโดยรวมจะน้อยแต่เมื่อพิจารณาว่าใช้ลักษณ์ตัวอักษรเพียงอย่างเดียวในการรู้จำผลที่ได้ก็ยังน่าสนใจ

Original EN	Original DE	Output DE
majestically	majestatetisch	majestisch
setting	setzend	settend
machineries	maschinerien	machinerien
naked	nakkt	nackt
southwest	suedwestlich	suedwest
dancing	tanzend	danzend

ภาพที่ 2.6 ตัวอย่างคำตอบที่ผลการทดสอบพยากรณ์โดยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนของ Mulloni

2.7 โปรแกรม weka

ในส่วนนี้ผู้วิจัยจะกล่าวถึง โปรแกรม weka เวอร์ชัน 3.7.10 ที่ผู้วิจัยเลือกใช้ในการวิจัยครั้งนี้ โปรแกรม weka เริ่มพัฒนาในปี 1999 โดยมหาวิทยาลัย Waikato ประเทศนิวซีแลนด์ เป็นโปรแกรมที่รวบรวมอัลกอลิทึมที่มีประโยชน์ต่อการทำเหมืองข้อมูล โดย weka สามารถใช้ในการจำแนกข้อมูล (classification) และการจัดกลุ่ม (cluster) ได้ โดยมีตัวช่วยจัดการทดสอบ (test option) ให้เลือกใช้ได้แบ่งเป็น 4 ประเภท คือ 1. แบบใช้ข้อมูลเรียนรู้เป็นข้อมูลทดสอบด้วย (use training set) 2. แบบใช้คลังข้อมูลเรียนรู้และฝึกฝนแยกออกจากกันเป็น 2 ไฟล์ (supplied test set) 3. แบบที่ใช้

คลังข้อมูลทุกส่วนเป็นทั้งข้อมูลฝึกฝนและทดสอบ (cross-validation) 4. แบบที่เลือกแบ่งชุดข้อมูลฝึกฝนและทดสอบเป็นสัดส่วนเปอร์เซ็นต์ (percentage split) ในการวิจัยครั้งนี้ ผู้วิจัยเลือกใช้การทดสอบแบบที่ 2 คือแยกคลังข้อมูลเรียนรู้และทดสอบออกจากกัน โดยใช้ข้อมูลฝึกฝนทั้งหมด 90 เปอร์เซ็นต์ คือวิทยานิพนธ์ 270 เล่ม และใช้ข้อมูลที่เหลือ 10 เปอร์เซ็นต์ คือวิทยานิพนธ์ 30 เล่ม เป็นข้อมูลทดสอบ

สำหรับแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ผู้วิจัยเลือกใช้ ได้แก่ libsvm นั้น สามารถนำเข้าจากตัวช่วยติดตั้งในโปรแกรมมาทดลองในโปรแกรม weka ได้ ส่วนผลการทำนายนั้นสามารถเลือกให้ weka เลือกแสดงผลเป็นค่าความครบถ้วน ค่าความแม่นยำ ค่า F-measure และสรุปรวมคำตอบของการทำนายแยกตามประเภทของคำตอบ นอกจากนั้น ยังสามารถเลือกให้ weka แสดงผลการทำนายแต่ละหน่วยของชุดข้อมูลทดสอบตามลำดับด้วย เช่น การเลือกแสดงผลเป็นแบบ csv ดังตัวอย่าง

inst#,actual,predicted,error,prediction
1,2:yes,1:no,+,1
2,1:no,1:no,,1
3,1:no,2:yes,+,1

2.8 การออกแบบคลังข้อมูลภาษาเพื่อตรวจเทียบการลักลอกงาน

ในการพัฒนาโปรแกรมตรวจเทียบการลักลอกงาน คลังข้อมูลภาษาเป็นส่วนหนึ่งที่สำคัญทั้งในแง่ที่เป็นฐานข้อมูลต้นฉบับ และในแง่ที่ใช้เป็นคลังข้อมูลภาษาเพื่อทดสอบความแม่นยำในการตรวจเทียบการลักลอกงาน ในงานประชุม PAN 2010 เพื่อการพัฒนาโปรแกรมสำหรับงานตรวจหาการลักลอกงานทั้งแบบตรวจภายในและตรวจภายนอก ได้แบ่งประเภทของคลังข้อมูลภาษาเพื่อใช้ประโยชน์ในการพัฒนาโปรแกรมตรวจเทียบ เป็น 3 ประเภท ดังนี้ [22]

2.7.1 คลังข้อมูลภาษาแบบ real plagiarism สร้างโดยใช้ข้อมูลชิ้นงานที่ตรวจพบการลักลอกในสถานการณ์จริงและนำมาสร้างเป็นคลังข้อมูล อย่างไรก็ตามการสร้างคลังข้อมูลประเภทนี้มีอุปสรรคหลายอย่าง เนื่องจากข้อจำกัด เช่น เข้าถึงข้อมูลได้ยากเนื่องจากการลักลอกงานที่พบมักถูกปกปิด และต้องได้รับการอนุญาตให้เข้าถึงข้อมูลจากทั้งเจ้าของผลงานต้นฉบับและเจ้าของผลงานที่พบการลักลอก นอกจากนี้ การเผยแพร่ข้อมูลการลักลอกงานยังมีเรื่องของจริยธรรมและกฎหมายเข้ามาเกี่ยวข้องอีกด้วย

2.7.2 คลังข้อมูลภาษาแบบ simulated plagiarism สร้างโดยใช้ผู้ลักลอกงานที่คัดเลือกมาเพื่อสมมติพฤติกรรมการลักลอกงาน ทั้งนี้การสร้างข้อมูลประเภทนี้ ยังมีข้อโต้แย้งว่า เมื่อมองใน

ด้าน จิตวิทยาแล้ว ผู้เขียนที่ตั้งใจลักลอกงานในสถานการณ์จริงและผู้เขียนที่ลักลอกงานในสถานการณ์สมมติมีแรงจูงใจที่ต่างกัน ดังนั้นพฤติกรรมการลักลอกงานที่สมมติขึ้นจะสามารถใช้เป็นตัวแทนของลักษณะข้อมูลที่ผ่านการลักลอกโดยผู้เขียนที่ตั้งใจลักลอกงานหรือไม่ อย่างไรก็ตามจากมุมมองทางภาษาศาสตร์ ยังไม่มีผลสรุปที่ชัดเจนว่าพฤติกรรมการลักลอกงานในสถานการณ์จริงแตกต่างกับพฤติกรรมการลักลอกงานในสถานการณ์สมมติ ตัวอย่าง ของโปรแกรมเชิงพาณิชย์ที่ให้บริการสร้างคลังข้อมูลประเภทนี้คือ Amazon's Mechanical Turk, AMT จากโครงการนำร่องการสร้างข้อมูลโดยใช้โปรแกรม AMT สามารถให้รายละเอียด เกี่ยวกับกลุ่มผู้เขียนสมมติ เช่น เพศ อายุ ระดับการศึกษา ภาษาที่พูดเป็นภาษาแม่ เวลาเฉลี่ยในการแก้ไขต้นฉบับ ในโครงการนำร่องนี้เสียค่าใช้จ่ายจำนวนครึ่งดอลล่าร์สหรัฐต่อจำนวนงานเขียนประมาณ 500 คำ และใช้เวลาครึ่งชั่วโมงต่องานเขียนหนึ่งชิ้น ผลของการทดลองพบว่าระยะเวลาการทำจะเปลี่ยนแปลงไป ขึ้นอยู่กับจำนวนค่าจ้างแต่ไม่ส่งผลต่อคุณภาพผลงาน

2.7.3 คลังข้อมูลภาษาแบบ artificial plagiarism สร้างชิ้นงานที่บิดเบือนจากต้นฉบับโดย 3 กรรมวิธี คือ โดยการสุ่ม (random text operations) คือนำเอกสารต้นฉบับมาตัดบางคำหรือประโยคออก เพิ่มคำ สลับคำ โดยการเปลี่ยนคำ (semantic word variation) คือ เปลี่ยนคำแต่ความหมายคงเดิมเหมือนกัน หรือ โดยการสลับตำแหน่งคำ (POS-preserving words shuffling) คือ การสลับตำแหน่งของคำที่มีหมวดคำเดียวกันในข้อความ ดังตัวอย่างการสร้างคลังข้อมูลดังภาพที่ 2.7

Original Text	The quick brown fox jumps over the lazy dog. det. adj. adj. n. v. p. det. adj. n.
Random text operations	over jumps quick brown fox The lazy. The brown jumps the. Quick dog The lazy fox over
Semantic word variation	The quick brown dodger leaps over the lazy canine. The quick brown canine jumps over the lazy canine.
POS-preserving word shuffling	The lazy quick dog jumps over the brown fox. The brown lazy dog jumps over the quick fox.

ภาพที่ 2.7 แสดงตัวอย่างการสร้างคลังข้อมูลแบบ artificial plagiarism

จากชนิดของการสร้างคลังข้อมูลภาษาทั้ง 3 แบบ โดยแบบที่ 2 และ 3 เป็นแบบที่นิยมใช้กันมาก ทั้งนี้ประเภทของการตรวจเทียบลักลอกงาน มีส่วนสำคัญมากในการพิจารณาสร้างคลังข้อมูล เนื่องจากในงานวิจัยนี้ ผู้วิจัยต้องการทดสอบประสิทธิภาพของระบบการตรวจเทียบภายในแบบต่างๆ ซึ่งควรจะทำงานได้ดีหากมีการเปลี่ยนแปลงรูปแบบการเขียนภายในเอกสาร ดังนั้นผู้วิจัยจะสร้าง

คลังข้อมูลภาษา โดยคัดลอกส่วนหนึ่งของชิ้นงานต้นฉบับมาแทรกในชิ้นงานที่ลักลอกโดยไม่เปลี่ยนแปลงส่วนใดเลย ซึ่งการสร้างคลังข้อมูลเปรียบเทียบโดยวิธีนี้ เป็นแบบเดียวกับที่ใช้ในงานแข่งขันโปรแกรมตรวจเทียบภายในหาการลักลอก ในปี 2010 (PAN 2010)

นอกจากชนิดของการสร้างคลังข้อมูลแล้ว ขนาดความยาวของข้อความที่ลักลอก และความยาวของชิ้นงาน ในคลังข้อมูลก็เป็นสิ่งที่ต้องคำนึงถึง ในการแข่งขัน PAN 2010 มีขนาดความยาวของชิ้นงานในคลังข้อมูลทั้งหมด 3 ประเภท คือ แบบสั้น 50 เปอร์เซ็นต์ (ความยาว 1-10 หน้า) แบบปานกลาง 35 เปอร์เซ็นต์ (ความยาว 10-100 หน้า) แบบยาว 15 เปอร์เซ็นต์ (ความยาว 100-500 หน้า) ส่วนความยาวของข้อมูลที่ลักลอกจากต้นฉบับก็มีทั้งหมด 3 ประเภทเช่นกัน คือ แบบสั้น 34 เปอร์เซ็นต์ (ความยาว 50-150 คำ) แบบปานกลาง 33 เปอร์เซ็นต์ (ความยาว 300-500 คำ) และแบบยาว 33 เปอร์เซ็นต์ (ความยาว 3,000 – 5,000 คำ) โดยการสร้างคลังข้อมูลเพื่อฝึกฝนและทดสอบการตรวจเทียบภายในหาการลักลอกงานวิชาการ ในงานวิจัยนี้ ผู้วิจัยกำหนดให้คลังข้อมูลต้นฉบับก่อนการลักลอกมี 3 ขนาด คือ ขนาดสั้น 7-15 หน้า (50 เปอร์เซ็นต์) ขนาดปานกลาง 16-50 หน้า (35 เปอร์เซ็นต์) และขนาดยาว > 50 หน้า (15 เปอร์เซ็นต์) ผู้วิจัยได้แบ่งขนาดของข้อมูลต้นฉบับออกเป็น 3 ขนาดดังกล่าว เปรียบเหมือนเป็นตัวแทนของงานเขียนวิชาการประเภทต่างๆ ได้แก่ ขนาดสั้นแทนประเภทบทความวิชาการ ขนาดปานกลาง และยาวแทนประเภทของวิทยานิพนธ์ ส่วนความยาวของข้อมูลที่ลักลอกจากต้นฉบับมีทั้งหมด 3 ประเภทเช่นกัน คือ แบบสั้นมีความยาว 50-100 คำ แบบปานกลางมีความยาว 101-200 คำ และแบบยาวมีความยาวมากกว่า 200 คำ ทั้งนี้รายละเอียดของการสร้างคลังข้อมูลและรายละเอียดต่างๆ จะกล่าวต่อไปในบทที่ 3 ในหัวข้อการสร้างคลังข้อมูล

บทที่ 3
คลังข้อมูลและลักษณ์ที่ใช้

3.1 คลังข้อมูลต้นฉบับ

สร้างคลังข้อมูลที่ใช้ในการวิจัยครั้งนี้ สร้างขึ้นจากวิทยานิพนธ์ภาษาไทยระดับบัณฑิตศึกษา จุฬาลงกรณ์มหาวิทยาลัย รวม 300 เล่ม จาก 2 สาขาวิชา คือ สาขาวิชาวิทยาศาสตร์ จำนวน 150 เล่ม และสาขาวิชามนุษยศาสตร์ จำนวน 150 เล่ม เพื่อเปรียบเทียบผลของข้อมูลต้นฉบับที่มีต่อการตรวจเทียบหาการลักลอก จึงกำหนดความยาวชุดข้อมูลต้นฉบับก่อนการลักลอกเป็นสามขนาด คือ ขนาดสั้น (7-15 หน้า) จำนวน 106 ชิ้น ขนาดกลาง (16-50 หน้า) จำนวน 150 ชิ้น และขนาดยาว (มากกว่า 50 หน้า) จำนวน 44 ชิ้น เนื่องจากต้นฉบับข้อมูลทั้งหมดจัดเก็บเป็นไฟล์ข้อมูลแบบเข้ารหัสชนิด pdf ผู้วิจัยแปลงข้อมูลต้นฉบับทั้งหมดเป็นไฟล์อักษร (text) ด้วยโปรแกรม PDFill PDF Tools 10.0 เมื่อต้นฉบับทั้งหมดเป็นไฟล์อักษรแล้ว ผู้วิจัยใช้จำนวนบรรทัดเป็นตัวกำหนดจำนวนหน้า โดยกำหนดหน่วย 30 บรรทัดเป็น 1 หน้าวิทยานิพนธ์ เหตุผลที่ไม่สามารถใช้จำนวนหน้าจากขนาดเล่มจริงมาอ้างอิงได้เลยนั้น เพราะส่วนที่นำมาสร้างคลังข้อมูลจะใช้เฉพาะตัวบทไม่รวมส่วนชื่อบท ตาราง กราฟแสดงผล และในวิทยานิพนธ์เล่มจริง จำนวนบรรทัดต่อหน้าในแต่ละเล่มมีความคลาดเคลื่อนกันอยู่ระหว่าง 27-32 บรรทัด

การออกแบบคลังข้อมูลกำหนดวิธีการให้ใกล้เคียงกับสถานการณ์ลักลอกจริงที่สุด โดยจัดกลุ่มชุดข้อมูลออกเป็นกลุ่มย่อยๆ ตามสาขาวิชา และกำหนดให้มีคำสำคัญของวิทยานิพนธ์ร่วมกัน ในกลุ่ม เช่น กลุ่มย่อยที่ 1 คือ วิทยานิพนธ์สายวิทยาศาสตร์ จากคณะพยาบาลศาสตร์ มีคำสำคัญร่วมกัน คือ “ปัจจัยคัดสรร” เป็นต้น เมื่อสร้างคลังข้อมูลแล้วเสร็จ มีกลุ่มย่อยในชุดข้อมูลวิทยานิพนธ์รวม 19 กลุ่ม แบ่งเป็นสายวิทยาศาสตร์ จำนวน 9 กลุ่ม และของสายมนุษยศาสตร์ จำนวน 10 กลุ่ม จากนั้นจำลองการลักลอกงานโดยด้วยการคัดลอกเนื้อหาจากวิทยานิพนธ์เล่นอื่นๆไปรวมกับวิทยานิพนธ์เล่มต้นฉบับตามสัดส่วนที่ออกแบบไว้ ทั้งนี้กำหนดให้ข้อมูลจากวิทยานิพนธ์เล่มต้นฉบับและวิทยานิพนธ์เล่มที่เป็นข้อมูลลักลอกมาจากกลุ่มย่อยเดียวกันทั้งหมด เนื่องจากเงื่อนไขที่ใช้ในการเก็บข้อมูลที่กล่าวไปข้างต้น ทำให้ชุดข้อมูลในกลุ่มย่อยที่เป็นข้อมูลต้นฉบับมีจำนวนเล่มและความยาวเล่มตามขนาด สั้น กลาง ยาว ไม่เท่ากัน ทั้งนี้เมื่อเก็บข้อมูลต้นฉบับแล้วเสร็จ มีสัดส่วนคลังข้อมูลต้นฉบับ ดังที่แสดงในตารางที่ 3.1

เลขที่กลุ่ม	ที่มา (กลุ่ม – คณะ)	คำสำคัญ	ความยาวของเล่ม			รวม (เล่ม)
			สั้น	กลาง	ยาว	
11	วิทยาศาสตร์ – พยาบาล	ปัจจัยคัดสรร	3	31	7	41
12	วิทยาศาสตร์ – พยาบาล	ปัจจัยทำนาย	0	11	1	12
13	วิทยาศาสตร์ – แพทย์	ภาวะซึมเศร้า	8	10	2	20
14	วิทยาศาสตร์ – แพทย์	ความเครียด	10	7	2	19
15	วิทยาศาสตร์ – แพทย์	ความรู้ ทัศนคติ พฤติกรรม	5	1	2	8
16	วิทยาศาสตร์ – แพทย์	ความชุก ปัจจัยที่เกี่ยวข้อง	10	0	2	12
17	วิทยาศาสตร์ – จิตวิทยา	ประสบการณ์	3	8	1	12
18	วิทยาศาสตร์ – จิตวิทยา	ความเครียด	0	4	2	6
19	วิทยาศาสตร์ – พยาบาล	ความสัมพันธ์ระหว่างปัจจัยส่วนบุคคล	14	3	3	20
21	มนุษยศาสตร์ – นิเทศน์	การสื่อสาร	25	26	5	56
22	มนุษยศาสตร์ – นิเทศน์	กระบวนการสื่อสาร	6	9	3	18
23	มนุษยศาสตร์ – นิเทศน์	ทัศนคติ	5	7	6	18
24	มนุษยศาสตร์ – นิเทศน์	กลยุทธ์การประชาสัมพันธ์	1	7	1	9
25	มนุษยศาสตร์ – นิเทศน์	ภาพลักษณ์	2	8	4	14
26	มนุษยศาสตร์ – นิเทศน์	วาทกรรม	2	5	0	7
27	มนุษยศาสตร์ – นิเทศน์	อุดมการณ์	4	1	0	5
28	มนุษยศาสตร์ – นิเทศน์	กลยุทธ์การสื่อสาร	4	5	1	10
29	มนุษยศาสตร์ – นิเทศน์	การสื่อความหมาย	3	4	0	7
30	มนุษยศาสตร์ – นิเทศน์	การเปิดรับข่าวสาร	1	3	2	6
		รวม	-106- (35%)	-150- (50%)	-44- (15%)	-300- (100%)

ตารางที่ 3.1 แสดงรายละเอียดของคลังข้อมูลต้นฉบับ

งานวิจัยชิ้นนี้ มุ่งหวังจะหาลักษณ์ที่ใช้บ่งชี้การลักลอกงานวิชาการในย่อหน้านั้นๆแบบไม่อ้างอิงฐานข้อมูลต้นฉบับ ข้อมูลที่ใช้ทั้งหมดนำมากำกับขอบเขตของย่อหน้าโดยเทียบกับเอกสารต้นฉบับด้วยสัญลักษณ์ที่ผู้วิจัยกำหนดขึ้นเอง คือ ^ ที่ขอบเขตของย่อหน้า และบรรทัดว่างระหว่างย่อหน้า สาเหตุที่ต้องนำข้อมูลมากำกับย่อหน้าอีกครั้งเนื่องจากการแปลงข้อมูลจากต้นฉบับเป็นไฟล์อักษร ทำให้ไฟล์ได้สูญเสียตัวบ่งชี้ย่อหน้าทั้งหมด คือ บรรทัดว่างก่อนขึ้นย่อหน้าใหม่และการใช้แท็บ

ตัวอย่างข้อมูลก่อนการกำกับขอบเขตย่อหน้า

นอกจากนี้ คำว่า การประชาสัมพันธ์ ยังมีความหมายกว้างขวางมาก นักวิชาการหลายท่าน ทั้งชาวไทยและชาวต่างชาติ ต่างให้คำจำกัดความของคำว่า การประชาสัมพันธ์ ในแง่มุมต่างๆ ดังนี้ จอห์น อี. มาร์สตัน (John E. Marston) นักวิชาการชาวอเมริกันที่มีชื่อเสียงมากผู้หนึ่งกล่าวว่า "การประชาสัมพันธ์นั้นเป็นการสื่อสารที่โน้มน้าวใจ โดยมีการวางแผนเพื่อให้เกิดอิทธิพลต่อกลุ่มประชาชนที่สำคัญ" (John E. Marston อ้างใน พรทิพย์ พิมลสินธุ์, 2539 : 5)

ตัวอย่างข้อมูลหลังการกำกับขอบเขตย่อหน้า

^ นอกจากนี้ คำว่า การประชาสัมพันธ์ ยังมีความหมายกว้างขวางมาก นักวิชาการหลายท่าน ทั้งชาวไทยและชาวต่างชาติ ต่างให้คำจำกัดความของคำว่า การประชาสัมพันธ์ ในแง่มุมต่างๆ ดังนี้ ^

^ จอห์น อี. มาร์สตัน (John E. Marston) นักวิชาการชาวอเมริกันที่มีชื่อเสียงมากผู้หนึ่ง กล่าวว่า "การประชาสัมพันธ์นั้นเป็นการสื่อสารที่โน้มน้าวใจ โดยมีการวางแผนเพื่อให้เกิดอิทธิพลต่อกลุ่มประชาชนที่สำคัญ" (John E. Marston อ้างใน พรทิพย์ พิมลสินธุ์, 2539 : 5) ^

เมื่อกำกับขอบเขตของย่อหน้าทั้งหมดของข้อมูล 300 เล่มแล้ว แบ่งย่อหน้าทั้งหมดออกจากกันเป็นไฟล์ย่อย บรรจุข้อความยาวเพียง 1 ย่อหน้า ระบุชื่อไฟล์ทั้งหมดในรูปแบบเดียวกันทั้งหมด ประกอบด้วย ตัวเลข 5 หลัก ตามด้วย สัญลักษณ์ _ และตามด้วยตัวเลขอีก 1-3 หลัก โดยตัวเลข 2 หลักแรก บอกถึงที่มาของวิทยานิพนธ์ว่ามาจากสาขาไหนและมีคำสำคัญอะไร ตัวเลขหลักที่ 3 ถึง 5 บอกถึงเลขที่ของเล่มในกลุ่มย่อยนั้นๆ ส่วนตัวเลข 1-3 หลักหลังสัญลักษณ์ _ บอกถึงลำดับที่ของย่อหน้านั้นๆในเล่ม เช่น 11001_0 เลข 11 แสดงวิทยานิพนธ์สายวิทยาศาสตร์ คณะพยาบาลศาสตร์ คำสำคัญว่า "ปัจจัยคัดสรร" เลขที่เล่มในกลุ่ม 001 และ เป็นย่อหน้าที่ 1 ในเล่ม (ย่อหน้าที่ 1 จะเริ่มด้วยเลข 0) เป็นต้น เมื่อแล้วเสร็จมีจำนวนย่อหน้าทั้งหมดในข้อมูลก่อนที่จะแทรกย่อหน้าที่ลักลอกเข้าไป มีสัดส่วนดังตารางที่ 3.2

ขนาดความยาวของย่อหน้าจากต้นฉบับวิทยานิพนธ์				รวม
<50 คำ	50-100 คำ	101-200 คำ	>201 คำ	
33,982 ย่อหน้า	14,212 ย่อหน้า	9,864 ย่อหน้า	4,029 ย่อหน้า	62,087 ย่อหน้า

ตารางที่ 3.2 แสดงความยาวย่อหน้าของข้อมูลก่อนแทรกย่อหน้าที่ลักลอก

3.2 คลังข้อมูลส่วนที่ลักลอก

ใช้ข้อมูลชุดเดียวกันกับชุดข้อมูลต้นฉบับแต่มีเงื่อนไข คือ

1. ข้อมูลลักลอกและข้อมูลต้นฉบับมาจากข้อมูลในกลุ่มย่อยเดียวกันเท่านั้น
2. ข้อมูลลักลอกและข้อมูลต้นฉบับต้องไม่มาจากผู้เขียนคนเดียวกัน
3. แทรกข้อมูลที่ลักลอกปนในต้นฉบับเป็นย่อหน้าใหม่เสมอและนำมาแทรกทั้งย่อหน้า
4. เพื่อวิเคราะห์ผลที่ได้จากจากตรวจจับในเล่มเปรียบเทียบกับความยาวของข้อความที่ลักลอก ในวิทยานิพนธ์หนึ่งเล่ม ย่อหน้าที่ลักลอกมาปนกับต้นฉบับต้องมีย่อหน้าขนาดสั้นปานกลางและยาวอย่างน้อยอย่างละ 1 ย่อหน้า
5. เมื่อแทรกข้อมูลที่ลักลอกลงในข้อมูลต้นฉบับแล้ว กำหนดให้มีข้อมูลต้นฉบับมากกว่าข้อมูลที่ลักลอกเสมอ

เมื่อกำหนดข้อปฏิบัติเกี่ยวกับการแทรกข้อมูลลักลอกลงในคลังข้อมูลต้นฉบับ ดังรายละเอียดข้างต้นแล้ว ผู้วิจัยได้ออกแบบคลังข้อมูลที่มีการลักลอกเพื่อนำมาศึกษา ดังนี้

1. นำข้อมูลต้นฉบับก่อนแทรกส่วนที่ลักลอก 300 เล่ม มาแบ่งกลุ่มตามจำนวนหน้าในเล่ม โดยคิด 30 บรรทัดเท่ากับ 1 หน้า ได้กลุ่มย่อยของข้อมูลทั้งหมด จำนวน 3 กลุ่ม คือ เล่มขนาดสั้น 106 เล่ม (35 เปอร์เซ็นต์) เล่มขนาดปานกลาง 150 เล่ม (50 เปอร์เซ็นต์) และเล่มขนาดยาว 44 เล่ม (15 เปอร์เซ็นต์) ดังรายละเอียดในตารางที่ 3.1
2. กำหนดประเภทของปริมาณที่ลักลอกรวมในเล่ม เป็น 3 ประเภท คือ ประเภทลักลอกมาน้อย เท่ากับ 5-15 เปอร์เซ็นต์ ของปริมาณเล่มต้นฉบับ ประเภทลักลอกมาปานกลาง เท่ากับ 16-30 เปอร์เซ็นต์ ของปริมาณเล่มต้นฉบับ และประเภทลักลอกมามาก เท่ากับ 31-40 เปอร์เซ็นต์ ของปริมาณเล่มต้นฉบับ
3. กำหนดสัดส่วนของจำนวนเล่มวิทยานิพนธ์ให้มีปริมาณการลักลอกตามข้อ 2 ดังนี้ ประเภทลักลอกมาน้อย 150 เล่ม (50 เปอร์เซ็นต์) ประเภทลักลอกมาปานกลาง 76 เล่ม (25 เปอร์เซ็นต์) และประเภทลักลอกมามาก 74 เล่ม (25 เปอร์เซ็นต์) ที่ผู้วิจัยออกแบบให้มีสัดส่วนต่างกัน เนื่องจาก ผู้วิจัยเชื่อว่าพฤติกรรมการลักลอกรวมในเล่มแบบน้อย น่าจะใกล้เคียงกับสถานการณ์ลักลอกจริงมากที่สุด จึงออกแบบให้ปริมาณการลักลอกมาน้อยในเล่ม มีสัดส่วนมากที่สุด ทั้งนี้ การคำนวณปริมาณการลักลอกต่อเล่มใช้หน่วยคำในการคำนวณ เช่น ไฟล์เลขที่ 21047 เป็นวิทยานิพนธ์ขนาดยาว 61.80 หน้า มีคำก่อนการลักลอกเท่ากับ 27,551 คำ แทรกย่อหน้าที่ลักลอกรวม 3,638 คำ เท่ากับ 13.20 เปอร์เซ็นต์ จากข้อมูลต้นฉบับ จัดเป็นประเภทที่ลักลอกมาน้อย

4. จากสัดส่วนของปริมาณที่ลักลอกรวมในเล่มทั้ง 3 แบบ กำหนดให้สมาชิกในกลุ่มมาจากข้อมูลต้นฉบับขนาดสั้น กลาง และยาว ในสัดส่วนเท่าๆกัน ตามสัดส่วนวิทยานิพนธ์ต้นฉบับในข้อ 1. คือ จากวิทยานิพนธ์ที่มีความยาวเล่มแบบสั้น 35 เปอร์เซ็นต์ วิทยานิพนธ์ที่มีความยาวเล่มแบบปานกลาง 50 เปอร์เซ็นต์ และวิทยานิพนธ์ที่มีความยาวเล่มมาก 15 เปอร์เซ็นต์ ดังรายละเอียดดังนี้
   - 4.1 กลุ่มที่ลักลอกมาน้อยรวม 150 เล่ม แบ่งประเภทได้ดังนี้
     - 4.1.1 วิทยานิพนธ์ที่มีความยาวเล่มแบบสั้นและลักลอกมาน้อย 54 เล่ม
     - 4.1.2 วิทยานิพนธ์ที่มีความยาวเล่มแบบปานกลางและลักลอกมาน้อย 72 เล่ม
     - 4.1.3 วิทยานิพนธ์ที่มีความยาวเล่มแบบยาวและลักลอกมาน้อย 24 เล่ม
   - 4.2 กลุ่มที่ลักลอกมาปานกลางรวม 76 เล่ม แบ่งประเภทได้ดังนี้
     - 4.2.1 วิทยานิพนธ์ที่มีความยาวเล่มแบบสั้นและลักลอกมาปานกลาง 26 เล่ม
     - 4.2.2 วิทยานิพนธ์ที่มีความยาวเล่มแบบปานกลางและลักลอกมาปานกลาง 39 เล่ม
     - 4.2.3 วิทยานิพนธ์ที่มีความยาวเล่มแบบยาวและลักลอกมาปานกลาง 11 เล่ม
   - 4.3 กลุ่มที่ลักลอกมามากรวม 74 เล่ม แบ่งประเภทได้ดังนี้
     - 4.3.1 วิทยานิพนธ์ที่มีความยาวเล่มแบบสั้นและลักลอกมามาก 26 เล่ม
     - 4.3.2 วิทยานิพนธ์ที่มีความยาวเลมแบบปานกลางและลักลอกมามาก 39 เล่ม
     - 4.3.3 วิทยานิพนธ์ที่มีความยาวเล่มแบบยาวและลักลอกมามาก 9 เล่ม
5. กำหนดสัดส่วนของย่อหน้าที่ลักลอก แล้วปนในข้อมูลต้นแบบ เพื่อวิเคราะห์ปัจจัยเรื่องความยาวของข้อความที่ลักลอกต่อความแม่นยำในการตรวจจับด้วยแบบจำลอง แต่เดิมผู้วิจัยได้ออกแบบให้จำนวนของย่อหน้าที่ลักลอกแบ่งตามประเภทความยาวของข้อความ มีสัดส่วนที่เท่ากัน คือ 33 เปอร์เซ็นต์ ทั้งย่อหน้าที่ลักลอก แบบสั้น ปานกลาง และยาว แต่ในคลังข้อมูลจริง พบว่าความยาวของย่อหน้าที่พบนั้นไม่ได้มีสัดส่วนเท่ากันอย่างที่ออกแบบไว้ เมื่อมีข้อจำกัดเรื่องข้อความลักลอกและต้นฉบับที่ต้องมาจากกลุ่มย่อยเดียวกัน และหลีกเลี่ยงการใช้ข้อความที่ลักลอกซ้ำกันหลายครั้ง จึงออกแบบคลังข้อมูลสำหรับลักลอกใหม่ เป็นย่อหน้าลักลอกขนาดสั้น 50 เปอร์เซ็นต์ ย่อหน้าลักลอกขนาดปานกลาง 35 เปอร์เซ็นต์ และย่อหน้าลักลอกขนาดยาว 15 เปอร์เซ็นต์ ซึ่งใกล้เคียงกับจำนวนและประเภทของย่อหน้าจริงในคลังข้อมูลลักลอก ดังตาราง ที่ 3.3

ขนาดความยาวของย่อหน้าที่ใช้เป็นคลังข้อมูลลักลอก			รวม
50-100 คำ	101-200 คำ	>200 คำ	
14,212 ย่อหน้า (50.55%)	9,864 ย่อหน้า (35.10%)	4,029 ย่อหน้า (14.35%)	28,105 ย่อหน้า

ตารางที่ 3.3 แสดงความยาวย่อหน้าของคลังข้อมูลลักลอก

6. ปนย่อหน้าที่ลักลอกในข้อมูลต้นฉบับตามที่ได้ออกแบบไว้ ทั้งนี้ สัดส่วนของย่อหน้าที่ลักลอกทั้งหมด ที่ขนาดสั้น 50 เปอร์เซ็นต์ ขนาดปานกลาง 35 เปอร์เซ็นต์ และขนาดยาว 15 เปอร์เซ็นต์นั้น เป็นสัดส่วนโดยรวมทั้งหมดของย่อหน้าในคลังข้อมูลลักลอก ไม่ใช่สัดส่วนการปนย่อหน้าในแต่ละเล่ม แต่ผู้วิจัยได้ออกแบบให้ ย่อหน้าที่ลักลอกในเล่ม ต้องประกอบไปด้วยย่อหน้าลักลอกขนาดสั้น กลาง และยาว อย่างน้อยชนิดละ 1 ย่อหน้า ดังตัวอย่างการปนส่วนที่ลักลอกในข้อมูลต้นฉบับที่แสดงในตาราง 3.4

ไฟล์เลขที่	ความยาว	จำนวนคำ	จำนวนคำที่ลักลอก	ความยาวของย่อหน้าที่ลักลอก			ประเภท
				50-100 คำ	101-200 คำ	>200 คำ	
13017	13.20 หน้า	5948 คำ	776 คำ (13.05%)	1 ย่อหน้า	4 ย่อหน้า	1 ย่อหน้า	ประเภท 4.1.1
11035	27.53 หน้า	13694 คำ	1058 คำ (7.73%)	7 ย่อหน้า	2 ย่อหน้า	1 ย่อหน้า	ประเภท 4.1.2
22017	54.73 หน้า	25171 คำ	1422 คำ (5.65%)	5 ย่อหน้า	3 ย่อหน้า	2 ย่อหน้า	ประเภท 4.1.3
23008	10.53 หน้า	4541 คำ	1085 คำ (23.89%)	2 ย่อหน้า	3 ย่อหน้า	2 ย่อหน้า	ประเภท 4.2.1
22001	39.67 หน้า	20369 คำ	5726 คำ (28.11%)	18 ย่อหน้า	12 ย่อหน้า	8 ย่อหน้า	ประเภท 4.2.2
30006	51.90 หน้า	24955 คำ	6931 คำ (27.77%)	41 ย่อหน้า	14 ย่อหน้า	5 ย่อหน้า	ประเภท 4.2.3

ตารางที่ 3.4 แสดงตัวอย่างการสร้างคลังข้อมูลอ้างอิงจากประเภทในข้อ 4.3

ไฟล์เลขที่	ความยาว	จำนวนคำ	จำนวนคำที่ลักลอก	ความยาวของย่อหน้าที่ลักลอก			ประเภท
				50-100 คำ	101-200 คำ	>200 คำ	
21018	13.93 หน้า	6838 คำ	2287 คำ (33.45%)	2 ย่อหน้า	7 ย่อหน้า	3 ย่อหน้า	ประเภท 4.3.1
29004	21.07 หน้า	10427 คำ	40.65 คำ (38.99%)	21 ย่อหน้า	9 ย่อหน้า	4 ย่อหน้า	ประเภท 4.3.2
23011	56.87 หน้า	27210 คำ	8877 คำ (32.62%)	30 ย่อหน้า	18 ย่อหน้า	11 ย่อหน้า	ประเภท 4.3.3

ตารางที่ 3.4 (ต่อ) แสดงตัวอย่างการสร้างคลังข้อมูลอ้างอิงจากประเภทในข้อ 4.3

เมื่อทำคลังข้อมูลเอกสารวิทยานิพนธ์ที่ผ่านการลักลอกแล้วเสร็จ ผู้วิจัยได้แบ่งข้อมูลทั้งหมดออกเป็น 2 ชุด คือ ชุดข้อมูลฝึกฝน จำนวน 90 เปอร์เซ็นต์ ประกอบด้วยวิทยานิพนธ์ 270 เล่ม และชุดข้อมูลทดสอบ 10 เปอร์เซ็นต์ ประกอบด้วยวิทยานิพนธ์ 30 เล่ม จัดแบ่งกลุ่มในสองลักษณะ คือ ตามปริมาณการลักลอกภายในเล่มว่ามีน้อย ปานกลาง หรือมาก และความยาวของย่อหน้าที่ลักลอกมาว่ามีขนาดสั้น ปานกลาง หรือยาว โดยคลังข้อมูลฝึกฝนและทดสอบมาจากข้อมูลวิทยานิพนธ์สายวิทยาศาสตร์และมนุษยศาสตร์อย่างละเท่าๆกัน อีกทั้งในคลังข้อมูลยังมีจำนวนปริมาณการลักลอกต่อเล่ม และจำนวนข้อความที่ลักลอกมาในปริมาณที่ใกล้เคียงกัน เมื่อแล้วเสร็จมีย่อหน้าที่ใช้ฝึกฝน 59,777 ย่อหน้า และ ย่อหน้าที่ใช้ทดสอบ 8,336 ย่อหน้า ดังรายละเอียดของข้อมูลฝึกฝนและข้อมูลทดสอบในตารางที่ 3.5

ประเภทของชุดข้อมูล	จำนวนทั้งหมด (ย่อหน้า)	จำนวนทั้งหมด (เล่ม)	ปริมาณการลักลอกต่อเล่ม			จำนวนข้อความที่ลักลอกโดยรวม (ย่อหน้า)		
			น้อย 5-15%	ปานกลาง 16-30%	มาก 31-40%	สั้น (50-100 คำ)	ปานกลาง (101-200 คำ)	ยาว (> 200 คำ)
ข้อมูลฝึกฝนวิทยาศาสตร์	28,547 ย่อหน้า	135 เล่ม	70 เล่ม	32 เล่ม	33 เล่ม	1340 ย่อหน้า	918 ย่อหน้า	400 ย่อหน้า
ข้อมูลฝึกฝนมนุษยศาสตร์	31,230 ย่อหน้า	135 เล่ม	68 เล่ม	35 เล่ม	32 เล่ม	1350 ย่อหน้า	900 ย่อหน้า	383 ย่อหน้า

ตารางที่ 3.5 รายละเอียดของคลังข้อมูลฝึกฝนและทดสอบ

ประเภทของชุดข้อมูล	จำนวนทั้งหมด (ย่อหน้า)	จำนวนทั้งหมด (เล่ม)	ปริมาณการลักลอกต่อเล่ม			จำนวนข้อความที่ลักลอกโดยรวม (ย่อหน้า)		
			น้อย 5-15%	ปานกลาง 16-30%	มาก 31-40%	สั้น (50-100 คำ)	ปานกลาง (101-200 คำ)	ยาว (> 200 คำ)
ข้อมูลทดสอบวิทยาศาสตร์	3,550 ย่อหน้า	15 เล่ม	5 เล่ม	6 เล่ม	4 เล่ม	166 ย่อหน้า	133 ย่อหน้า	51 ย่อหน้า
ข้อมูลทดสอบมนุษยศาสตร์	4,786 ย่อหน้า	15 เล่ม	7 เล่ม	3 เล่ม	5 เล่ม	159 ย่อหน้า	157 ย่อหน้า	69 ย่อหน้า
รวม	68,113 ย่อหน้า	300 เล่ม	150 เล่ม	76 เล่ม	74 เล่ม	3015 ย่อหน้า	2108 ย่อหน้า	903 ย่อหน้า

ตารางที่ 3.5 (ต่อ) รายละเอียดของคลังข้อมูลฝึกฝนและทดสอบ

3.3 ข้อมูลรับเข้า

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองต่อข้อมูลรับเข้า 2 ประเภท คือ แบบรับเข้าเป็นคำและรับเข้าเป็นตัวอักษร ทั้งนี้ ผู้วิจัยจะใช้ข้อมูลชุดเดียวกันสำหรับข้อมูลรับเข้าทั้ง 2 ประเภท ต่างกันเพียงข้อมูลชุดที่รับเข้าเป็นคำจะผ่านการตัดคำโดยอัตโนมัติด้วยโปรแกรมตัดคำ CUThai Segmentation version 2.01 ของภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยก่อน โดย 1 หน่วยข้อมูลรับเข้ามีขนาด 1 ย่อหน้า

3.3.1 ตัวอย่างชุดข้อมูลรับเข้าเป็นคำ

<p><w>^<w>การ<w>แสดง<w>บัลเล่ต์<w>ใน<w>ยุค<w>แรก<w><s><w>ๆ<w><s><w>ถือ<w>เป็น<w>เครื่องมือ<w>ทาง<w>ศิลปะ<w>อัน<w>หนึ่ง<w>ที่<w>สื่อ<w>ถึง<w>อำนาจ<w><s>และ<w>ความ<w>

<p><w>ร่ำรวย<w>ของ<w>ตระกูล<w>ใหญ่<w>ที่อยู่<w>ใน<w>ประเทศ<w>อิตาลี<w><s>พัฒนา<w>การ<w>ของ<w>การ<w>แสดง<w>บัลเล่ต์<w>เริ่ม<w>เปลี่ยน<w>ไป<w>เมื่อ<w>มี<w>

<p>การ<w>ก่อตั้ง<w>โรงเรียน<w>สอน<w>เต้นรำ<w>ขึ้น<w>เป็น<w>ครั้ง<w>แรก
<w><s><w>เมื่อ<w>ปี<w>คริสต์ศักราช<w><s><w>1661<w><s>ใน<w>ชื่อ<w><s><w>
รอยัล<w><s>อะ<w>คา<w>เด<w>มี<w><s><w>เดอ<w>
<p>ลา<w>ดองส์
<w><s><w>(<w>Royal<w><s><w>Academie<w><s><w>de<w><s><w>la<w><s><
w>Danse<w>)<w><s>ซึ่ง<w>ก่อตั้ง<w>โดย<w><s>พระเจ้า<w>หลุยส์<w>ที่
<w><s><w>14<w><s><w>แห่ง<w>ประเทศ<w>ฝรั่งเศส<w>
<p>ทำให้<w>วงการ<w>ศิลปะ<w>การ<w>เต้น<w>บัลเล่ต์<w>เกิด<w>นัก<w>เต้น<w>
บัลเล่ต์<w>อาชีพ<w>ขึ้น<w>เป็น<w>ครั้ง<w>แรก<w><s>เพื่อ<w>ตอบสนอง<w>ความ<w>
<p><w>บันเทิง<w>ให้<w>แก่<w>ชนชั้น<w>สูง<w>ใน<w>สังคม<w>ฝรั่งเศส<w><s>และ
<w>มี<w>การ<w>สร้างสรรค์<w>ผลงาน<w>การ<w>แสดง<w>บัลเล่ต์<w>มาก<w>ขึ้น<w>
<p><w>ตาม<w>ลำดับ<w>^<w>

โดย สัญลักษณ์ในข้อมูลรับเข้าแบบคำ แสดงความหมาย ดังนี้

<p>	แทนการขึ้นบรรทัดใหม่	<w>	แสดงขอบเขตของคำ
<s>	แสดงช่องว่าง	^	แสดงขอบเขตของย่อหน้า

3.3.2 ตัวอย่างชุดข้อมูลรับเข้าเป็นตัวอักษร

^ การแสดงบัลเล่ต์ในยุคแรก ๆ ถือเป็นเครื่องมือทางศิลปะอันหนึ่งที่สื่อถึงอำนาจ และความร่ำรวยของตระกูลใหญ่ที่อยู่ในประเทศอิตาลี พัฒนาการของการแสดงบัลเล่ต์เริ่มเปลี่ยนไปเมื่อมีการก่อตั้งโรงเรียนสอนเต้นรำขึ้นเป็นครั้งแรก เมื่อปีคริสต์ศักราช 1661 ในชื่อ รอยัล อะคาเดมี เดอ ลาดองส์ (Royal Academie de la Danse) ซึ่งก่อตั้งโดย พระเจ้าหลุยส์ที่ 14 แห่งประเทศฝรั่งเศส ทำให้วงการศิลปะการเต้นบัลเล่ต์เกิดนักเต้นบัลเล่ต์อาชีพขึ้นเป็นครั้งแรก เพื่อตอบสนองความบันเทิงให้แก่ชนชั้นสูงในสังคมฝรั่งเศส และมีการสร้างสรรค์ผลงานการแสดงบัลเล่ต์มากขึ้นตามลำดับ ^

3.3.3 ตัวอย่างย่อหน้าที่ลักลอก

Þs^2.1.3. การดูแลช่วยเหลืออย่างต่อเนื่อง การกระทำหน้าที่ของผู้ประกอบวิชาชีพการพยาบาลต้องครอบคลุมการดูแลช่วยเหลือบุคคล ครอบครัว และชุมชนทั้งในภาวะเจ็บป่วยคือ ช่วยให้สภาพการเจ็บป่วยหายไป (Restoration) ด้วยการช่วยเหลือ ดูแล บำบัด รักษา/บรรเทาอาการของโรค การฟื้นฟูสภาพ (Rehabilitation) เพื่อกระตุ้นหรือรักษาระดับของการฟื้นหาย (Recover or

Healing) การส่งเสริมสุขภาพ (Health Promotion) การป้องกันโรค(Disease Prevention) การป้องกันสุขภาพ (Health Protection)^฿s

ในชนิดย่อหน้าที่ลักลอกทั้งที่ใช้กับข้อมูลรับเข้าแบบคำและข้อมูลรับเข้าแบบตัวอักษร จะมีขนาดสั้น กลาง ยาว แบ่งตามจำนวนคำเท่ากันกับข้อมูลชนิดต้นฉบับ แต่มีจุดสังเกตคือย่อหน้าที่ลักลอกจะนำหน้าด้วยสัญลักษณ์บอกประเภทย่อหน้าลักลอกแบ่งตามขนาดความยาวของย่อหน้า คือ ฿l แสดงย่อหน้าลักลอกแบบยาว ฿m แสดงย่อหน้าแบบปานกลาง ฿s แสดงย่อหน้าแบบสั้น

3.4 ลักษณ์ที่ใช้เรียนรู้และทดสอบ

ผู้วิจัยกำหนดลักษณ์ต่างๆ ที่ใช้ในการเรียนรู้และทดสอบ ทั้งลักษณ์ที่ใช้ค่าคำนวณทางคณิตศาสตร์เพียงอย่างเดียว และลักษณ์ที่ใช้ความรู้ทางภาษาร่วมกับการคำนวณทางคณิตศาสตร์ โดยหน่วยในชุดข้อมูลที่นำมาคำนวณทางคณิตศาสตร์ มีดังนี้

1. คำ คือ หน่วยที่อยู่ระหว่างเครื่องหมายแสดงขอบเขตของคำ <w>....<w> ทั้งที่เป็นอักษรภาษาไทย ภาษาอังกฤษ สัญลักษณ์ ตัวเลข แต่ไม่รวมช่องว่าง <s> สัญลักษณ์พิเศษ ^ ที่ใช้แสดงขอบเขตย่อหน้า สัญลักษณ์ <p>แสดงการขึ้นบรรทัดใหม่ และสัญลักษณ์ ฿l,฿m,฿s ที่ใช้แสดงย่อหน้าที่มีการลักลอกแบบมาก ปานกลาง และน้อย
2. ตัวอักษร คือ หน่วยใดๆที่ปรากฏในชุดข้อมูล ทั้งที่เป็นอักษรภาษาไทย ภาษาอังกฤษ สัญลักษณ์ ตัวเลข แต่ไม่รวมช่องว่าง <s> สัญลักษณ์พิเศษ ^ ที่ใช้แสดงขอบเขตย่อหน้า สัญลักษณ์ และสัญลักษณ์ ฿l,฿m,฿s ที่ใช้แสดงย่อหน้าที่มีการลักลอกแบบมาก ปานกลาง และน้อย
3. ช่องว่าง หรือ การเขียนเว้นวรรค คือ หน่วยที่แสดงด้วยสัญลักษณ์ <s> ในชุดข้อมูล

เพื่อให้แบบจำลองทำนายคำตอบว่าย่อหน้านั้นๆมีการลักลอกงานหรือไม่ ลักษณ์ที่ใช้จะมีลักษณะสัมพัทธ์เพื่อเทียบความต่างของย่อหน้าปัจจุบันกับส่วนอื่นๆ ที่เหลือในเล่ม โดยจะดูลักษณะพื้นฐานทั่วไปก่อนในเรื่องของจำนวนคำ จำนวนการเว้นวรรค ความยาวย่อหน้า ความยาวข้อความ เป็นต้น และเนื่องจากเป็นการเปรียบเทียบการข้อมูลรับเข้าที่เป็นสายคำและสายอักขระ ลักษณ์ที่ใช้จึงเป็นสองกลุ่มในลักษณะเดียวกันสำหรับการเปรียบเทียบประสิทธิภาพของข้อมูลรับเข้าแบบคำกับแบบตัวอักษร จากนั้นจึงพิจารณาใช้ลักษณ์ที่เป็นเรื่องเฉพาะทางภาษาประกอบ เช่น คำที่พบใช้บ่อยในเล่ม คำที่เขียนผิดหรือต่างจากคำเดียวกันในเล่ม ลักษณ์ที่ใช้จึงจัดได้เป็นสามกลุ่ม ดังนี้

3.5 ลักษณ์ทางสถิติที่ใช้กับแบบจำลองรับเข้าแบบคำ

1. ค่าเฉลี่ยจำนวนคำต่อย่อหน้าในเล่ม

หาค่าเฉลี่ยของจำนวนคำต่อย่อหน้าในเล่มเพื่อเปรียบเทียบกับย่อหน้าปัจจุบัน ใช้ลักษณ์นี้เพื่อดูลักษณะความยาวเฉลี่ยในการเขียนย่อหน้าหนึ่งๆ มีวิธีคำนวณค่าดังนี้

ค่าเฉลี่ยจำนวนคำต่อย่อหน้าในเล่ม = (จำนวนคำทั้งหมดในเล่ม – จำนวนคำของย่อหน้าปัจจุบัน)/(จำนวนย่อหน้าทั้งหมดในเล่ม -1)

ตัวอย่างการคำนวณ เมื่อให้วิทยานิพนธ์เล่มที่ 1 มี 433 ย่อหน้า จำนวนคำทั้งหมด 53,000 คำ และย่อหน้าปัจจุบัน มีจำนวน 252 คำ = (53,000 คำ – 252 คำ) / (433-1) = 122.10 คำ

ผลที่ได้จะเป็นค่าเฉลี่ยจำนวนคำต่อย่อหน้าของวิทยานิพนธ์เล่มนั้นๆ ซึ่งไม่รวมย่อหน้าปัจจุบัน เช่น จากการคำนวณข้างต้น วิทยานิพนธ์เล่มที่ 1 มีค่าเฉลี่ยจำนวนคำต่อย่อหน้าในเล่มที่ 122.10 คำ

2. จำนวนคำต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม

หาจำนวนคำที่แตกต่างระหว่างย่อหน้าปัจจุบันกับค่าเฉลี่ยจำนวนคำต่อย่อหน้าในเล่ม ใช้วิธีคำนวณดังนี้

จำนวนคำต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม = ค่าเฉลี่ยจำนวนคำต่อย่อหน้าในเล่ม–จำนวนคำของย่อหน้าปัจจุบัน

3. ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่ม

หาค่าเฉลี่ยของจำนวนช่องว่างต่อย่อหน้าในเล่ม ใช้ลักษณ์นี้เพื่อดูลักษณะการเขียนว่าในหนึ่งย่อหน้า มีการเขียนเว้นวรรคมากน้อยเพียงใด วิธีคำนวณค่าดังนี้

ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่ม = (จำนวนช่องว่างทั้งหมดในเล่ม–จำนวนช่องว่างของย่อหน้าปัจจุบัน)/ (จำนวนย่อหน้าทั้งหมดในเล่ม -1)

ตัวอย่างการคำนวณ เมื่อให้วิทยานิพนธ์เล่มที่ 1 มี 433 ย่อหน้า จำนวนช่องว่างทั้งหมด 5,976 ครั้ง และย่อหน้าปัจจุบัน มีช่องว่างจำนวน 18 ครั้ง = (5,976 ครั้ง – 18 ครั้ง) / (433-1) = 13.79 ครั้ง

ผลที่ได้จะเป็นค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าของวิทยานิพนธ์เล่มนั้นๆ ซึ่งไม่รวมย่อหน้าปัจจุบัน เช่น จากการคำนวณข้างต้น วิทยานิพนธ์เล่มที่ 1 มีค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่มที่ 13.79 ครั้ง เนื่องจาก ข้อมูลรับเข้าแบบคำและแบบตัวอักษรเป็นชุดเดียวกัน ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่มของทั้ง 2 ชุดข้อมูลรับเข้าจึงเป็นจำนวนเดียวกัน

4. จำนวนช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่ม

หาจำนวนการใช้ช่องว่างที่แตกต่างกันของย่อหน้าปัจจุบันกับค่าเฉลี่ยของจำนวนช่องว่างต่อย่อหน้าในเล่ม ใช้วิธีคำนวณดังนี้

จำนวนช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่ม = ค่าเฉลี่ยของช่องว่างต่อย่อหน้าในเล่ม– จำนวนช่องว่างของย่อหน้าปัจจุบัน

เนื่องจากข้อมูลรับเข้าแบบคำและแบบตัวอักษรเป็นข้อมูลชุดเดียวกัน จำนวนช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่มของทั้ง 2 ชุดข้อมูลรับเข้าจึงเป็นจำนวนเดียวกัน

5. ค่าเฉลี่ยจำนวนคำต่อช่องว่างในเล่ม

หาค่าเฉลี่ยจำนวนคำต่อช่องว่างในเล่ม โดยไม่รวมย่อหน้าปัจจุบัน ใช้ลักษณ์นี้เพื่อดูความถี่ของจำนวนคำต่อการเว้นวรรคหนึ่งครั้งในเล่ม มีวิธีคำนวณค่าดังนี้

ค่าเฉลี่ยจำนวนคำต่อช่องว่างในเล่ม = (จำนวนคำทั้งหมดในเล่ม-จำนวนคำของย่อหน้าปัจจุบัน)/(จำนวนช่องว่างทั้งหมดในเล่ม – จำนวนช่องว่างของย่อหน้าปัจจุบัน)

ตัวอย่างการคำนวณ เมื่อให้วิทยานิพนธ์เล่มที่ 1 มี จำนวนคำทั้งหมด 25,087 คำ มีช่องว่าง 5,976 ครั้ง จำนวนคำในย่อหน้าปัจจุบัน 83 คำ และมีช่องว่าง 18 ครั้ง = (25,087 คำ – 83 คำ) / (5,976 ครั้ง – 18 ครั้ง) = 4.20 คำ/ช่องว่าง

ผลที่ได้จะเป็นค่าเฉลี่ยของจำนวนคำต่อการเว้นวรรคในเล่มวิทยานิพนธ์นั้นๆ ซึ่งไม่รวมย่อหน้าปัจจุบัน เช่น จากการคำนวณข้างต้น วิทยานิพนธ์เล่มที่ 1 มีค่าเฉลี่ยจำนวนของจำนวนคำต่อการเว้นวรรคที่ 4.20 คำต่อการเว้นวรรคหนึ่งครั้ง

6. ค่าเฉลี่ยจำนวนคำต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม

หาจำนวนที่ต่างกันของค่าเฉลี่ยจำนวนคำต่อช่องว่างของย่อหน้าปัจจุบันกับค่าเฉลี่ยจำนวนคำต่อช่องว่างในเล่ม ใช้วิธีคำนวณดังนี้

จำนวนค่าเฉลี่ยของจำนวนคำต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม = ค่าเฉลี่ยจำนวนคำต่อช่องว่างในย่อหน้าปัจจุบัน–ค่าเฉลี่ยจำนวนคำต่อช่องว่างในเล่ม

7. สัดส่วนของคำในย่อหน้าปัจจุบันต่อทั้งเล่ม

หาสัดส่วนของย่อหน้าปัจจุบันจากจำนวนคำ เมื่อพิจารณาร่วมกันทุกย่อหน้าในเล่ม ใช้วิธีคำนวณดังนี้

สัดส่วนของคำในย่อหน้าปัจจุบันต่อทั้งเล่ม = (จำนวนคำในย่อหน้าปัจจุบัน/จำนวนทั้งหมดในเล่ม)* 100

ตัวอย่างการคำนวณ เมื่อให้วิทยานิพนธ์เล่มที่ 1 มี จำนวนคำทั้งหมด 25,087 คำ จำนวนคำในย่อหน้าปัจจุบัน 83 = (83 คำ/25,087 คำ) *100 = 0.33 ผลที่ได้คือ ย่อหน้าปัจจุบันมีสัดส่วนจำนวนคำต่อทั้งเล่มอยู่ที่ 0.33

3.6 ลักษณ์ทางสถิติที่ใช้กับแบบจำลองรับเข้าแบบตัวอักษร

1. ค่าเฉลี่ยจำนวนตัวอักษรต่อย่อหน้าในเล่ม

ใช้วิธีหาค่าเฉลี่ยเหมือนกับวิธีการหาค่าเฉลี่ยจำนวนคำต่อย่อหน้าในเล่ม แต่หน่วยที่นำมาคำนวณคือตัวอักษร

2. จำนวนตัวอักษรต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม

ใช้วิธีหาค่าเฉลี่ยเหมือนกับวิธีการหาจำนวนคำต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม แต่หน่วยที่นำมาคำนวณคือตัวอักษร

3. ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่ม

เนื่องจาก ข้อมูลรับเข้าแบบคำและแบบตัวอักษรเป็นชุดเดียวกัน ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่มของข้อมูลรับเข้าแบบตัวอักษรจึงเท่ากับของข้อมูลรับเข้าแบบคำ

4. จำนวนช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่ม

เนื่องจากข้อมูลรับเข้าแบบคำและแบบตัวอักษรเป็นข้อมูลชุดเดียวกัน จำนวนช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่มของของข้อมูลรับเข้าแบบตัวอักษรจึงเท่ากับของข้อมูลรับเข้าแบบคำ

5. ค่าเฉลี่ยจำนวนอักษรต่อช่องว่างในเล่ม

หาค่าเฉลี่ยของจำนวนอักษรต่อช่องว่างในเล่มโดยไม่รวมย่อหน้าปัจจุบัน ใช้วิธีคำนวณเหมือนการคำนวณค่าเฉลี่ยจำนวนคำต่อช่องว่างในเล่ม แต่ใช้หน่วยที่นำมาคำนวณเป็นตัวอักษร

6. ค่าเฉลี่ยจำนวนตัวอักษรต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม

หาจำนวนที่ต่างกันของค่าเฉลี่ยจำนวนตัวอักษรต่อช่องว่างโดยไม่รวมย่อหน้าปัจจุบัน ใช้วิธีคำนวณเหมือนการคำนวณค่าเฉลี่ยจำนวนคำต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม แต่ใช้หน่วยที่นำมาคำนวณเป็นตัวอักษร

7. สัดส่วนของตัวอักษรในย่อหน้าปัจจุบันต่อทั้งเล่ม

หาสัดส่วนของตัวอักษรในย่อหน้าปัจจุบันต่อทั้งเล่ม ใช้วิธีคำนวณเหมือนการคำนวณสัดส่วนของคำในย่อหน้าปัจจุบันต่อทั้งเล่ม แต่ใช้หน่วยที่นำมาคำนวณเป็นตัวอักษร

3.7 ลักษณ์ทางภาษา

1. ค่าต่างของชุดคำที่มีความถี่การใช้สูงที่สุดในเล่ม

โดยลักษณ์นี้ ผู้วิจัยได้พัฒนามาจากแนวคิดที่ใช้ชุดคำและความถี่เพื่อแสดงถึงลักษณะการเขียนเฉพาะบุคคล ซึ่งถูกนำมาใช้ในงานวิจัยเกี่ยวกับภาษาธรรมชาติ เช่น Argamon และ Leviton ได้ทดลองใช้คำที่มีความถี่สูงสุดจำนวน 200 คำแรกในงานเขียนนวนิยายจำนวน 20 เล่ม เพื่อแสดงถึงลักษณะการเลือกใช้คำศัพท์ของผู้เขียนที่แตกต่างกันในแต่ละคน จากนั้นจำแนกชื่อผู้เขียนที่แตกต่างกันจำนวน 8 คน ด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน พบว่าลักษณ์ดังกล่าว สามารถจำแนกผู้เขียนโดยได้ค่าความถูกต้องที่ 99 เปอร์เซ็นต์ [24]

จากแนวคิด ข้างต้น ซึ่งให้ความสำคัญกับชุดคำที่ผู้เขียนเลือกใช้มากในงานเขียนของตนเอง ผู้วิจัยจึงสันนิษฐานว่า ในงานเขียนชิ้นหนึ่งๆ ผู้เขียนจะใช้ชุดคำที่มีความถี่สูงอย่างสม่ำเสมอ ดังนั้น ย่อหน้าที่มีการใช้คำความถี่สูงสุดต่างจากค่าเฉลี่ยของย่อหน้าอื่นๆในเล่ม จึงน่าจะเป็นย่อหน้าที่ลักลอกมา ทั้งนี้ ใช้วิธีวัดค่าความต่างของย่อหน้าปัจจุบันกับทั้งเล่ม ดังนี้

ค่าต่างของชุดคำที่ความถี่สูงสุดในเล่ม = (ความถี่รวมของชุดคำที่มีความถี่สูงสุด อันดับที่ k/ (จำนวนย่อหน้าทั้งหมดในเล่ม -1)) – ความถี่รวมของชุดคำที่มีความถี่สูงสุด อันดับที่ k ในย่อหน้านั้นๆ

โดย อันดับที่ k ดังกล่าวผู้วิจัยได้เริ่มทดลองที่ ความถี่สูงสุด 100 ลำดับแรก เพราะจากคลังข้อมูลฝึกฝนและทดสอบทั้งหมด ความถี่รวมที่อันดับ 100 มีขนาดประมาณ ครึ่งหนึ่งของคลังข้อมูลทั้งหมด

2. คำเขียนผิด

คำเขียนผิด เป็นลักษณะอีกอย่างหนึ่ง ซึ่งใช้ในการระบุตัวผู้เขียน เช่น ในงานหาตัวอาชญากรจากจดหมายข่มขู่ หรือ เรียกค่าไถ่ โดยเชื่อว่า การใช้คำเขียนผิดจะพบอย่างสม่ำเสมอในผู้เขียนคนนั้นๆ ในงานวิจัยนี้ ผู้วิจัยเชื่อว่า หากพบคำที่เขียนผิดและเขียนถูกปนกันในงานชิ้นนั้นๆ ส่วนที่มีลักษณะที่พบน้อยกว่า น่าจะเป็นส่วนที่ลักลอกมา ทั้งนี้ คำเขียนผิดที่จะนำมาพิจารณาจะพิจารณาจากรูปผิวเท่านั้น ไม่ได้พิจารณาในส่วนของการใช้ผิดความหมายด้วย

ชุดคำเขียนผิดที่นำมาพิจารณา คัดเลือกจากคำที่เขียนผิดและถูกที่มีความถี่รวมกันทั้งหมดเกิน 50 ครั้ง และความถี่ของคำที่ผิดมีมากกว่า 10 เปอร์เซ็นต์ของความถี่คำที่เขียนถูก ทั้งนี้ ชุดลักษณ์คำเขียนผิด 1 ลักษณ์ประกอบด้วยสมาชิก 2 คำซึ่งคำทั้ง 2 มีรูปผิวที่ต่างกันเพียงเล็กน้อยและใช้แทนกันได้ในทุกบริบท โดยการตัดสินว่าคำใดอยู่ในกลุ่มคำเขียนผิดนั้น จะอ้างอิงจากพจนานุกรม

ภาษาไทย ฉบับราชบัณฑิตยสถาน พ.ศ. 2542 ชุดคำที่เข้าเกณฑ์ดังกล่าว มีดังนี้ (กฎ, กฏ) (ปรากฏ, ปรากฎ) (ศีรษะ, ศรีษะ) (กะทันหัน, กระทันหัน) (เกม, เกมส์) (พ.ศ., พศ.)

โดยให้ค่าของลักษณ์นี้เป็นตัวเลข 3 กลุ่ม จากการเปรียบเทียบย่อหน้าปัจจุบันกับย่อหน้าอื่นๆ ในเล่ม ดังนี้

ให้ค่า -1 เมื่อพบการใช้ชุดคำที่ผิดหรือถูกต่างจากย่อหน้าอื่นๆ หรือ ใช้ปนกันทั้งสองคำในย่อหน้าปัจจุบัน

ให้ค่า 0 เมื่อไม่พบการใช้ชุดคำที่ผิดหรือถูกในย่อหน้านั้นๆ

ให้ค่า 1 เมื่อพบการใช้ชุดคำที่ผิดหรือถูกในย่อหน้านั้นๆ และเหมือนกับย่อหน้าอื่นๆในเล่ม

เช่น ไฟล์ 11005_372 มีค่า เท่ากับ 1

*<w>เมื่อ<w>พิจารณา<w>ดู<w>ว่า<w>สัม<w>ประสิทธิ์<w>ถดถอย<w>ของ<w>ตัว<w>พยากรณ์<w>ทั้งหมด<w>ใน<w>รูป<w>คะแนน<w>มาตรฐาน<w><s><w>**ปรากฏ<**w>ว่า*

ไฟล์ 11005_390 มีค่า เท่ากับ -1

*<w>ใน<w>รูป<w>คะแนน<w>มาตรฐาน<w><s><w>**ปรากฏ**<w>ว่า<w><s><w>ตัว<w>พยากรณ์<w>ที่<w>สามารถ<w>พยากรณ์<w>ความสามารถ<w>ของ<w><p><w>ผู้ดูแล<w>ใน<w>การ<w>ดูแล<w>เด็ก<w>วัย<w>ก่อน<w>เรียน<w>*

ไฟล์ 11005_391 มีค่า เท่ากับ 0

<w>ความสามารถ<w>ของ<w>ผู้ดูแล<w>ใน<w>การ<w>ดูแล<w>เด็ก<w>วัย<w>ก่อน<w>เรียน<w>ที่<w>ติด<w>เชื้อ<w>เอช<w>ไอวี<w><s><w>ตัว<w>พยากรณ์<w>ทุก<w>ตัว<w>มี<w>ความ<w><p>สัมพันธ์<w>ทาง<w>บวก<w>

3. การเลือกใช้คำและรูปแบบที่แตกต่างกัน

ลักษณ์ประเภทนี้ ผู้วิจัยจะพิจารณาถึงชุดคำและรูปแบบที่พบแตกต่างกันในคลังข้อมูล โดยมีข้อสันนิษฐานเดียวกับลักษณ์ประเภทคำเขียนผิดว่า การเลือกใช้คำที่มีความหมายเหมือนกัน หรือรูปแบบการเขียนที่ไม่สม่ำเสมอในงานเขียนชิ้นหนึ่งๆ ส่วนย่อหน้าที่พบการเลือกใช้คำและรูปแบบการเขียนที่เป็นส่วนน้อยกว่าเมื่อเทียบกับย่อหน้าอื่นๆในเล่ม น่าจะเป็นส่วนที่ลักลอกมา

ชุดคำและรูปแบบการเขียนที่นำมาพิจารณา คัดเลือกการใช้รวมกันทั้งหมดเกิน 50 ครั้ง และความถี่ของการใช้ที่ต่างกันมีมากกว่า 10 เปอร์เซ็นต์เหมือนกับชุดคำเขียนผิด เพียงแต่คำที่ใช้เป็นลักษณ์ในชุดนี้จะประกอบด้วยคำทั้งหมด 3 ประเภท คือ คำที่เขียนได้หลายแบบและยังไม่มีการบรรจุไว้ในพจนานุกรมภาษาไทย ฉบับราชบัณฑิตยสภาน ปี พ.ศ. 2542 เช่น ศัพท์บัญญัติที่มาจาก

ภาษาต่างประเทศ 2. คำที่มีรูปผิวต่างกันแต่มีความหมายเหมือนกัน 3. รูปแบบการใช้คำย่อ ชุดคำที่เข้าเกณฑ์ดังกล่าวและนำมาใช้เป็นหนึ่งลักษณ์ มีดังนี้ (อินเทอร์เน็ต, อินเตอร์เน็ต) (พ.ศ.<-s>, พ.ศ.<+s>) (เปอร์เซ็นต์ , เครื่องหมาย %) (พ่อ, บิดา) (แม่, มารดา) (ร้อยละ, เปอร์เซ็นต์, เครื่องหมาย %)

โดย ให้ค่าของลักษณ์เป็นตัวเลข 3 กลุ่ม แบบเดียวกับลักษณ์ประเภทคำเขียนผิด กับ ชุดคำดังนี้ (อินเทอร์เน็ต, อินเตอร์เน็ต) (พ.ศ.<-s>, พ.ศ.<+s>) (เปอร์เซ็นต์ , เครื่องหมาย %) (พ่อ, บิดา) (แม่, มารดา) และ ให้ค่าของลักษณ์เป็น 2 คำตอบ คือ yes และ no สำหรับ ชุดคำ (ร้อยละ, เปอร์เซ็นต์, เครื่องหมาย %) โดยให้ค่า yes สำหรับย่อหน้าที่เลือกใช้คำในชุดคำต่างออกไปเมื่อเทียบกับย่อหน้าอื่นๆในเล่ม และ no สำหรับย่อหน้าที่ไม่ต่างกับย่อหน้าอื่นๆ รวมถึงย่อหน้าที่ไม่พบคำในชุดนี้ด้วย

ซึ่งการใช้ที่ปนกันดังกล่าว ในชุดคำที่มีความหมายเดียวกันคือ (พ่อ, บิดา) (แม่, มารดา) ผู้วิจัยได้นำคำที่ปรากฏร่วม (collocation) มาพิจารณาร่วมด้วย ซึ่งถ้าพบคำปรากฏร่วมที่ใช้ในบริบทที่ต่างกันจากชุดความหมายที่ต้องการ ความถี่ของคำในบริบทนั้นๆจะถูกสะกัดออกและไม่นำมาประมวลผล ได้แก่ “แม่” ที่ประกอบเป็นชื่อเฉพาะ เช่น **แม่**ฟ้าหลวง **แม่**ฮ่องสอน **แม่**กลอง หรือ “แม่” ที่ไม่ได้หมายถึงผู้ให้กำเนิด เช่น **แม่**น้ำ **แม่**ชี และ พ่อ เช่น **พ่อ**ขุน หลวง**พ่อ** ทั้งนี้ แม้ชุดคำดังกล่าวจะใช้หมายถึงความหมายที่ต้องการ แต่ไม่พบการใช้ร่วมกับคำปรากฏร่วม (collocation) นั้นๆเลย ก็จะถูกสกัดออกเช่นกัน เช่น **พ่อ**เลี้ยงเดี่ยว **แม่**เลี้ยงเดี่ยว **พ่อ**พันธุ์ **แม่**พันธุ์ ซึ่ง ไม่เคยพบการใช้ **บิดา**เลี้ยงเดียว **มารดา**เลี้ยงเดี่ยว **บิดา**พันธุ์ และ **มารดา**พันธุ์

3.8 การให้คำตอบ

การให้คำตอบของชุดข้อมูล มี 2 คำตอบ

คือ yes สำหรับย่อหน้าที่มีการลักลอก
no สำหรับย่อหน้าที่ไม่มีการลักลอก

3.9 การเตรียมข้อมูลเพื่อใช้กับแบบจำลอง

ในงานวิจัยนี้ ผู้วิจัยได้เลือกใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนผ่านโปรแกรม weka 3.7.10 ซึ่งแบบจำลองนี้กำหนดข้อมูลรับเข้าได้ทั้ง 2 ประเภทคือ ไฟล์ชนิด CSV และ ไฟล์ชนิด ARFF เนื่องจากลักษณ์เกือบทั้งหมดที่ใช้ มีค่าเป็นตัวเลข (numeric) ผู้วิจัยจึงเตรียมข้อมูลเป็นชนิด CSV ก่อน และแปลงเป็นชนิด ARFF ด้วยตัวช่วยหนึ่งในโปรแกรม weka ในภายหลัง

ลำดับ	เลขที่ไฟล์	ลักษณ์ที่ใช้							คำตอบ
		1	2	3	4	5	6	7	
1	11001_7	57.9	-83.1	13.49	-34.51	4.29	-1.35	0.55	no
2	11001_8	58.03	-28.97	13.56	-2.44	4.28	1.16	0.34	yes
3	11001_9	57.99	-46.01	13.52	-17.48	4.29	-0.94	0.41	no

ตารางที่ 3.6 ตัวอย่างข้อมูลรับเข้าแบบคำและค่าลักษ์ทางสถิติแบบไฟล์ CSV

ลำดับ	เลขที่ไฟล์	ลักษณ์ที่ใช้							คำตอบ
		1	2	3	4	5	6	7	
1	11001_7	235.3	-315.7	13.49	-34.51	17.45	-5.97	0.53	no
2	11001_8	235.78	-108.22	13.56	-2.44	17.39	4.11	0.33	yes
3	11001_9	235.59	-191.41	13.52	-17.48	17.42	-3.65	0.41	no

ตารางที่ 3.7 ตัวอย่างข้อมูลรับเข้าแบบตัวอักษรและค่าลักษ์ทางสถิติแบบไฟล์ CSV

ตัวอย่างข้อมูลรับเข้าแบบคำที่แปลงจากไฟล์ชนิด CSV เป็น ARFF โดยตัวช่วยของ weka อ้างอิงจากข้อมูลในตารางที่ 3.6

@relation trainwordbase ------>แสดงชื่อไฟล์ต้นฉบับก่อนที่จะแปลงเป็นไฟล์ชนิด ARFF

@attribute ‘1’ numeric ------>ลักษณ์ลำดับที่ 1 ชื่อลักษณ์ ‘1’ประเภท ตัวเลข
@attribute '2’ numeric ------>ลักษณ์ลำดับที่ 2 ชื่อลักษณ์ ‘2’ประเภท ตัวเลข
@attribute '3' numeric ------>ลักษณ์ลำดับที่ 3 ชื่อลักษณ์ ‘3’ประเภท ตัวเลข
@attribute '4' numeric ------>ลักษณ์ลำดับที่ 4 ชื่อลักษณ์ ‘4’ประเภท ตัวเลข
@attribute '5' numeric ------>ลักษณ์ลำดับที่ 5 ชื่อลักษณ์ ‘5’ประเภท ตัวเลข
@attribute '6’ numeric ------>ลักษณ์ลำดับที่ 6 ชื่อลักษณ์ ‘6’ประเภท ตัวเลข
@attribute '7' numeric ------>ลักษณ์ลำดับที่ 7 ชื่อลักษณ์ ‘7’ประเภท ตัวเลข
@attribute Class {no,yes} ------>แสดงประเภทคำตอบ 2 คำตอบ no, yes
@data
57.9,-83.1,13.49,-34.51,4.29,-1.35,0.55,no
58.03,-28.97,13.56,-2.44,4.28,1.16,0.34,yes
57.99,-46.01,13.52,-17.48,4.29,-0.94,0.41,no

ข้อมูลรับเข้าแต่ละไฟล์จะแยกคนละบรรทัดเรียงตามลำดับก่อนหลังแบบเดียวกับต้นฉบับที่เป็นไฟล์ CSV แต่จะไม่รวมเลขที่ไฟล์ เนื่องจาก weka จะมองทุกอย่างเป็นลักษณ์ของแบบจำลอง เลขที่ไฟล์จึงถูกลบออกก่อนแปลงไปไฟล์ชนิด ARFF โดยค่าของลักษณ์ต่างๆจะเรียงลำดับตามลำดับของลักษณ์ในส่วนของ @ attribute แต่ละลักษณ์แยกออกจากกันด้วยสัญญลักษณ์ ”,” โดยคำตอบจะอยู่ตำแหน่งท้ายสุด

บทที่ 4
วิธีการทดลองและผลการทดลอง

ในส่วนนี้ ผู้วิจัยจะกล่าวถึงขั้นตอนต่างๆ ในการพัฒนาการตรวจเทียบภายในหาการลักลอกงานวิชาการภาษาไทย วิธีการประเมินประสิทธิภาพแบบจำลอง รวมถึงการทดลองเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลรับเข้าต่างกัน ระหว่างข้อมูลรับเข้าแบบคำและข้อมูลรับเข้าแบบตัวอักษร จากนั้นเปรียบเทียบประสิทธิภาพของลักษณ์ทางสถิติและลักษณ์ทางภาษาที่ใช้กับแบบจำลอง โดยกำหนดขั้นตอนการทดลองดังนี้

1. ฝึกฝนและทดสอบข้อมูลรับเข้าแบบคำและแบบตัวอักษรกับลักษณ์ทางสถิติที่มาจากพื้นฐานเดียวกัน และมีจำนวนเท่ากัน จากนั้นเปรียบเทียบประสิทธิภาพของข้อมูลรับเข้าทั้งสองประเภทว่าแบบไหนที่ให้ผลที่ดีกว่ากัน

2. ฝึกฝนและทดสอบลักษณ์ทางภาษา ซึ่งจะทดลองลักษณ์ทางภาษาทั้งในแบบ global และ local คือเปรียบเทียบความต่างของลักษณ์ทางภาษาในย่อหน้านั้นกับทั้งเล่ม และความต่างของลักษณ์ทางภาษาในย่อหน้านั้นกับย่อหน้าก่อนหน้าจำนวนหนึ่ง จากนั้นเปรียบเทียบประสิทธิภาพของผลที่ได้กับลักษณ์ทางสถิติ โดยใช้ลักษณ์ทางสถิติที่ได้ผลดีที่สุดในการทดลองข้อ 1

3. ฝึกฝนและทดสอบลักษณ์ทางสถิติร่วมกับลักษณ์ทางภาษาที่ดีที่สุดที่ได้จากการทดลองในข้อ 2 เพื่อดูว่าจะช่วยเพิ่มประสิทธิภาพของแบบจำลองหรือไม่

4. นำชุดลักษณ์ที่ทำนายคำตอบได้ดีที่สุดจากการทดลองข้อ 1, 2 และ 3 มาวิเคราะห์ว่าปัจจัยเรื่องความยาวของข้อความลักลอกมีผลอย่างไรกับความแม่นยำในการตรวจจับการลักลอก

4.1 การประเมินประสิทธิภาพของแบบจำลอง

การประเมินผลประสิทธิภาพของแบบจำลองจะใช้ผลที่ได้จากการทำนายข้อมูลทดสอบจำนวน 10 เปอร์เซ็นต์ของข้อมูลทั้งหมด โดยใช้แบบจำลอง (model) ที่สร้างจากข้อมูลฝึกฝน จำนวน 90 เปอร์เซ็นต์ของข้อมูลทั้งหมด

การประเมินประสิทธิภาพของแบบจำลองจะแสดงด้วยค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) และค่า F-measure

ค่าความแม่นยำ คือ ค่าที่แสดงให้เห็นว่าระบบสามารถตรวจจับย่อหน้าที่มีการลักลอกได้แม่นยำมากน้อยขนาดไหนเมื่อเทียบจากจำนวนย่อหน้าที่มีการลักลอกทั้งหมดที่แบบจำลองตรวจจับออกมา สามารถคำนวณได้จากสูตร

$$P = \frac{\text{จำนวนย่อหน้าที่ลักลอกตรวจจับได้ถูกต้อง} * 100}{\text{จำนวนย่อหน้าลักลอกทั้งหมดที่ตรวจจับออกมา}}$$

ค่าความครบถ้วน คือ ค่าที่แสดงให้เห็นว่าระบบสามารถตรวจจับย่อหน้าที่มีการลักลอกได้ครบถ้วนขนาดไหน เมื่อเทียบกับย่อหน้าทั้งหมดที่มีการลักลอกในเอกสารทั้งหมด สามารถคำนวณได้จากสูตร

$$R = \frac{\text{จำนวนย่อหน้าที่ลักลอกที่ตรวจจับได้ถูกต้อง} * 100}{\text{จำนวนย่อหน้าที่ลักลอกทั้งหมดที่มีในเอกสาร}}$$

ค่า F-measure คือ ค่าความถูกต้องโดยรวม เป็นค่าเฉลี่ยของค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall) สามารถคำนวณได้จากสูตร

$$\text{F-measure} = \frac{2 * P * R}{P + R}$$

ทั้งนี้ การประเมินประสิทธิภาพของแบบจำลอง แม้จะพิจารณาจากค่าความครบถ้วนของย่อหน้าที่มีการลักลอกเป็นหลัก แต่จะพิจารณาประสิทธิภาพของการตรวจจับย่อหน้าที่ไม่มีการลักลอกประกอบด้วย เช่น กรณีที่ค่าความครบถ้วนของลักษณ์ประเภทหนึ่งสามารถตรวจจับย่อหน้าที่มีการลักลอกได้ค่าความครบถ้วน 100 เปอร์เซ็นต์ แต่ไม่สามารถตรวจจับย่อหน้าที่ไม่มีการลักลอกได้เลย ลักษณ์ประเภทนั้นไม่ถือเป็นลักษณ์ที่ได้ผลกับแบบจำลอง

4.2 การทดลองประสิทธิภาพของข้อมูลรับเข้าแบบคำและแบบตัวอักษร

เพื่อทดสอบประสิทธิภาพของชนิดข้อมูลรับเข้าทั้ง 2 ประเภท จึงกำหนดให้ทดลองโดยใช้ลักษณ์ทางสถิติเพียงอย่างเดียวก่อน ทั้งนี้ลักษณ์ทางสถิติที่นำมาใช้ทดลองกำหนดให้มีจำนวนเท่ากันและเป็นลักษณ์ที่คำนวณจากพื้นฐานเดียวกันทั้งหมด แตกต่างกันเพียงข้อมูลรับเข้าแบบคำจะใช้จำนวนคำในการคำนวณ ส่วนข้อมูลรับเข้าแบบตัวอักษรจะใช้จำนวนตัวอักษรในการคำนวณ โดยลักษณ์ทั้งหมดจะมีค่าเป็นตัวเลข (numeric) รายการลักษณ์ทางสถิติทั้งหมดที่ใช้ในการทดลอง แยกตามประเภทข้อมูลรับเข้ามีดังนี้

ลักษณ์สำหรับข้อมูลรับเข้าแบบคำ

1. ค่าเฉลี่ยจำนวนคำต่อย่อหน้าในเล่ม
2. จำนวนคำต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม
3. ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่ม

4. จำนวนช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่ม
5. ค่าเฉลี่ยจำนวนคำต่อช่องว่างในเล่ม
6. ค่าเฉลี่ยจำนวนคำต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม
7. สัดส่วนของคำในย่อหน้าปัจจุบันต่อทั้งเล่ม

ลักษณ์สำหรับข้อมูลรับเข้าแบบตัวอักษร

1. ค่าเฉลี่ยจำนวนตัวอักษรต่อย่อหน้าในเล่ม
2. จำนวนตัวอักษรต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม
3. ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่ม
4. จำนวนช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่ม
5. ค่าเฉลี่ยจำนวนตัวอักษรต่อช่องว่างในเล่ม
6. ค่าเฉลี่ยจำนวนตัวอักษรต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม
7. สัดส่วนของตัวอักษรในย่อหน้าปัจจุบันต่อทั้งเล่ม

จากนั้น ทดลองประสิทธิภาพของข้อมูลรับเข้าทั้งสองประเภท โดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ในโปรแกรม weka 3.7.10 ใช้เคอเนลแบบโพลีโมเนียลและประเมินผลประสิทธิภาพของแบบจำลองทั้ง 2 ประเภท ทั้งนี้คำตอบที่ถูกต้องของข้อมูลทดสอบ เท่ากับ 735 ย่อหน้าที่ลักลอก และ 7,601 ย่อหน้าที่ไม่ลักลอก

เพื่อดูประสิทธิภาพของลักษณ์กลุ่มต่างๆ ที่มีผลกับแบบจำลอง ผู้วิจัยได้แบ่งลักษณ์ทั้งหมด 7 ลักษณ์ข้างต้นออกเป็นกลุ่มย่อย 3 กลุ่ม ตามที่มาของค่าทางสถิติ คือ กลุ่มที่มาจากค่าทางสถิติของคำหรือตัวอักษร กลุ่มที่มาจากค่าทางสถิติของคำหรือตัวอักษรและช่องว่าง และกลุ่มที่มาจากค่าต่างทางสถิติของคำหรือตัวอักษรและช่องว่าง จากนั้นออกแบบการทดลองออกเป็น 4 ครั้ง คือทดลองแต่ละครั้งแยกตามที่มาของค่าทางสถิติทั้ง 3 กลุ่ม และทดสอบลักษณ์ทั้งหมดรวมกัน ดังรายละเอียด ดังนี้

1. ทดลองใช้ลักษณ์ที่มาจากค่าทางสถิติของคำหรือตัวอักษร จำนวน 3 ลักษณ์ คือ ค่าเฉลี่ยจำนวนคำหรือตัวอักษรต่อย่อหน้าในเล่ม จำนวนคำหรือตัวอักษรต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม และสัดส่วนของคำหรือตัวอักษรในย่อหน้าปัจจุบันต่อทั้งเล่ม เหตุผลที่ผู้วิจัยเลือกนำเอาลักษณ์ทางสถิติที่มาจากคำหรือตัวอักษรมาทดลองก่อน เนื่องจากสนใจว่า เมื่อใช้ลักษณ์ที่มาจากคำและตัวอักษรเพียงอย่างเดียว มีผลอย่างไรต่อแบบจำลอง อีกทั้งยังเป็นการทดสอบว่าย่อหน้าที่มาจากต้นฉบับเดียวกัน แต่แตกต่างกันด้วยหน่วยคำที่ผ่านการตัดคำอัตโนมัติและแบบตัวอักษรที่ไม่ต้องผ่านการตัดคำจะให้ผลเหมือนกันหรือไม่อย่างไร

2. ทดลองใช้ลักษณ์ที่มาจากค่าทางสถิติของคำหรือตัวอักษรและช่องว่างจำนวน 4 ลักษณ์

คือ ค่าเฉลี่ยจำนวนคำหรืออักษรต่อช่องว่างในเล่ม จำนวนคำหรือตัวอักษรต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่ม จำนวนช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่ม ในการทดลองขั้นนี้ ผู้วิจัยสนใจว่า เมื่อใช้ลักษณ์เกี่ยวกับค่าเฉลี่ยช่องว่างที่มีค่าเหมือนกันทั้ง 2 แบบจำลอง ร่วมกับลักษณ์ที่แสดงการใช้ช่องว่าง (การเว้นวรรค)ต่อจำนวนคำหรือตัวอักษร ให้ผลดีขึ้นกว่าการใช้ลักษณ์ที่ได้จากคำหรือตัวอักษรอย่างเดียวหรือไม่อย่างไร

3. ทดลองใช้ลักษณ์ที่มาจากค่าต่างทางสถิติเมื่อเปรียบเทียบย่อหน้าปัจจุบันกับค่าทางสถิติภายในเล่ม จำนวน 3 ลักษณ์ คือ จำนวนคำหรือตัวอักษรที่ต่างจากค่าเฉลี่ยในเล่ม ค่าเฉลี่ยจำนวนคำหรือตัวอักษรต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม และจำนวนช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม เพื่อทดสอบว่าลักษณ์ที่แสดงเพียงค่าต่างระหว่างย่อหน้าปัจจุบันต่อทั้งเล่ม ซึ่งผู้วิจัยคิดว่าน่าจะเป็นลักษณ์ที่ให้ผลดีที่สุด จะให้ผล

4. ทดลองใช้ลักษณ์ทางสถิติทั้งหมด จากการทดลองที่ 1-3 จำนวน 7 ลักษณ์

4.2.1 การทดลองลักษณ์ทางสถิติครั้งที่ 1 – จำนวนลักษณ์ 3 ลักษณ์

ผลการทดลองการเปรียบเทียบประสิทธิภาพของแบบจำลองรับเข้าแบบคำและแบบตัวอักษร เมื่อใช้ลักษณ์ทางสถิติที่มาจากคำหรือตัวอักษรเพียงอย่างเดียว ทดสอบแล้วได้ผล ดังต่อไปนี้

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
รับเข้าแบบตัวอักษร	1,2,7	ลักลอก	0.08	0.91	0.15
	1,2,7	ไม่ลักลอก	0.66	0.01	0.03
รับเข้าแบบคำ	1,2,7	ลักลอก	0.23	0.29	0.26
	1,2,7	ไม่ลักลอก	0.93	0.91	0.92

ตารางที่ 4.1 แสดงผลการเปรียบเทียบแบบจำลองทั้ง 2 ประเภทกับลักษณ์ทางสถิติที่มาจากคำหรือตัวอักษรรวม 3 ลักษณ์

เมื่อใช้ลักษณ์ทางสถิติที่มาจากจำนวนคำหรือตัวอักษรเพียงอย่างเดียว ได้แก่ ค่าเฉลี่ยจำนวนคำหรือตัวอักษรต่อย่อหน้าในเล่ม จำนวนคำหรือตัวอักษรต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม และสัดส่วนของคำหรือตัวอักษรในย่อหน้าปัจจุบันต่อทั้งเล่ม พบว่าแบบจำลองที่รับเข้าแบบคำมีค่าความครบถ้วนของการตรวจหาย่อหน้าที่ลักลอกที่ 0.29 อย่างไรก็ดี แม้ในแบบจำลองที่ข้อมูลรับเข้าเป็นตัวอักษรมีค่าความครบถ้วนในการตรวจหาย่อหน้าที่ลักลอกที่มากกว่าคือ 0.91 แต่เมื่อนำค่าความแม่นยำของย่อหน้าที่ลักลอกรวมถึงค่าความครบถ้วนของย่อหน้าที่ไม่ได้ลักลอกมาพิจารณาด้วย กลับพบว่า ลักษณ์จากแบบจำลองที่รับเข้าแบบตัวอักษร ได้ทำนายข้อมูลทดสอบแบบเลือกตอบเป็นคำตอบเดียว คือ เป็นย่อหน้าลักลอกเกือบทั้งหมด ดังผลการทำนาย ย่อหน้าที่ลักลอก 8,160 ย่อหน้า

และ ย่อหน้าที่ไม่ลักลอก 176 ย่อหน้า ดังนั้น ลักษณ์ทางสถิติที่มาจากข้อมูลรับเข้าแบบตัวอักษร จึงไม่มีประสิทธิภาพในการตรวจเทียบภายในหาการลักลอก เนื่องจาก ไม่สามารถแยกประเภทย่อหน้าที่ไม่ลักลอกออกจากย่อหน้าที่ลักลอกได้ ดังนั้น ลักษณ์ทางสถิติที่มาจากข้อมูลรับเข้าแบบคำจึงมีประสิทธิภาพเหนือกว่าลักษณ์ทางสถิติที่มาจากข้อมูลรับเข้าแบบตัวอักษร แม้จะมาจากข้อมูลต้นฉบับเดียวกัน

4.2.2 การทดลองลักษณ์ทางสถิติครั้งที่ 2 – จำนวนลักษณ์ 4 ลักษณ์

ผลการทดลองการเปรียบเทียบประสิทธิภาพของแบบจำลองรับเข้าแบบคำและแบบตัวอักษร เมื่อใช้ลักษณ์ทางสถิติที่มาจากคำและช่องว่างภายในเล่มจำนวน 4 ลักษณ์ ทดสอบแล้วได้ผลดังต่อไปนี้

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
รับเข้าแบบตัวอักษร	3,4,5,6	ลักลอก	0.09	0.99	0.17
	3,4,5,6	ไม่ลักลอก	0.99	0.04	.07
รับเข้าแบบคำ	3,4,5,6	ลักลอก	0.08	1	0.16
	3,4,5,6	ไม่ลักลอก	1	0.02	0.002

ตารางที่ 4. 2 แสดงผลการเปรียบเทียบแบบจำลองทั้ง 2 ประเภทกับลักษณ์ทางสถิติที่มาจากคำหรือตัวอักษรและช่องว่าง รวม 4 ลักษณ์

เมื่อใช้ลักษณ์ทางสถิติที่มาจากคำหรือตัวอักษรและช่องว่าง ได้แก่ ค่าเฉลี่ยจำนวนคำหรือตัวอักษรต่อช่องว่างในเล่ม ค่าเฉลี่ยจำนวนคำหรือตัวอักษรต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่ม และจำนวนช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่ม พบว่าประสิทธิภาพของลักษณ์ที่มาจากแบบจำลองทั้ง 2 ประเภทต่อย่อหน้าที่ลักลอกได้ผลที่ไม่ต่างกันนัก คือ 0.99 และ 1 สำหรับข้อมูลรับเข้าแบบตัวอักษรและคำ ตามลำดับ อย่างไรก็ดี เมื่อนำค่าความแม่นยำในการทำนายและค่าความครบถ้วนของย่อหน้าชนิดที่ไม่ลักลอกมาร่วมพิจารณาด้วย กลับพบว่า ลักษณ์จากแบบจำลองทั้ง 2 ประเภท ได้ทำนายข้อมูลแบบเลือกตอบเป็นคำตอบเดียว คือ ย่อหน้าลักลอก 8,068 ย่อหน้า และย่อหน้าไม่ลักลอก 268 ย่อหน้า สำหรับแบบจำลองแบบตัวอักษร และ ย่อหน้าลักลอก 8,324 ย่อหน้า และ ย่อหน้าที่ไม่ลักลอก 12 ย่อหน้า สำหรับแบบจำลองแบบคำ ดังนั้น ลักษณ์ทางสถิติ จำนวน 4 ลักษณ์ข้างต้น ที่มาจากคำหรือตัวอักษรและช่องว่าง จึงไม่มีประสิทธิภาพในการตรวจเทียบภายในหาการลักลอก เนื่องจาก ไม่สามารถแยกประเภทย่อหน้าที่ไม่ลักลอกออกจากย่อหน้าที่ลักลอกได้

4.2.3 การทดลองลักษณ์ทางสถิติครั้งที่ 3 – จำนวนลักษณ์ 3 ลักษณ์

ผลการทดลองการเปรียบเทียบประสิทธิภาพของแบบจำลองรับเข้าแบบคำและแบบตัวอักษร เมื่อใช้ลักษณ์ทางสถิติที่มาจากการเปรียบเทียบค่าต่างทางสถิติในย่อหน้าปัจจุบันกับค่าทางสถิติภายในเล่มจำนวน 3 ลักษณ์ ทดสอบแล้วตัวเลขมีความแตกต่างกันมาก ดังต่อไปนี้

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
รับเข้าแบบตัวอักษร	2,4,6	ลักลอก	0.08	1	0.16
	2,4,6	ไม่ลักลอก	1	0.01	0.02
รับเข้าแบบคำ	2,4,6	ลักลอก	0.11	0.001	0.003
	2,4,6	ไม่ลักลอก	0.91	0.99	0.003

ตารางที่ 4. 3 แสดงผลการเปรียบทียบแบบจำลองทั้ง 2 ประเภทกับลักษณ์ทางสถิติภายในย่อหน้าปัจจุบันและ ภายในเล่มรวม 3 ลักษณ์

เมื่อใช้ลักษณ์ทางสถิติที่มาจากค่าต่างภายในย่อหน้าปัจจุบัน เปรียบเทียบกับค่าเฉลี่ยทางสถิติในเล่ม ได้แก่ จำนวนคำหรือตัวอักษรต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม ค่าเฉลี่ยจำนวนคำหรือตัวอักษรต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม และจำนวนช่องว่างต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม พบว่าค่าความครบถ้วนของลักษณ์จากแบบจำลองแบบตัวอักษรอยู่ที่ 1 มากกว่าลักษณ์จากแบบจำลองแบบคำที่มีค่าเพียง 0.001 อย่างไรก็ดี เมื่อพิจารณาถึงค่าความแม่นยำและค่าความครบถ้วนของย่อหน้าที่ไม่ลักลอกแล้วพบว่า แบบจำลองทั้ง 2 แบบ ได้ทำนายผลออกมาเป็นแบบเลือกตอบเป็นคำตอบเดียว โดยแบบตัวอักษร ทำนายคำตอบเกือบทั้งหมดเป็นย่อหน้าที่ลักลอก 8,248 ย่อหน้า ในขณะที่ แบบคำทำนายคำตอบเกือบทั้งหมดเป็นย่อหน้าไม่ลักลอก 8,327 ย่อหน้า ดังนั้น แม้ลักษณ์จากข้อมูลรับเข้าปบบตัวอักษรมีค่าความครบถ้วนในการทำนายย่อหน้าที่ลักลอกมาก แต่กลับไม่มีประสิทธิภาพการตรวจเทียบภายในย่อหน้าลักลอก เนื่องจาก ไม่สามารถแยกย่อหน้าที่ไม่ลักลอกออกจากย่อหน้าที่ลักลอกได้

4.2.4 การทดลองลักษณ์ทางสถิติครั้งที่ 4 – จำนวนลักษณ์ 7 ลักษณ์

ผลการทดลองการเปรียบเทียบประสิทธิภาพของแบบจำลองรับเข้าแบบคำและแบบตัวอักษร เมื่อใช้ลักษณ์ทางสถิติที่ทั้งหมด 7 ลักษณ์ ทดสอบแล้วได้ผล ดังต่อไปนี้

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
รับเข้าแบบตัวอักษร	1-7	ลักลอก	0.16	0.05	0.08
	1-7	ไม่ลักลอก	0.91	0.97	0.94
รับเข้าแบบคำ	1-7	ลักลอก	0.19	0.43	0.26
	1-7	ไม่ลักลอก	0.94	0.82	0.87

ตารางที่ 4.4 แสดงผลการเปรียบเทียบแบบจำลองทั้ง 2 ประเภทกับลักษณ์ทางสถิติรวม 7 ลักษณ์

เมื่อใช้ลักษณ์ทางสถิติที่มาจากคำหรือตัวอักษรและช่องว่างในย่อหน้าปัจจุบันกับย่อหน้าอื่นๆ ในเล่ม รวม 7 ลักษณ์ พบว่า ลักษณ์ที่มาจากข้อมูลรับเข้าแบบคำ มีค่าความครบถ้วนในการตรวจหาย่อหน้าลักลอกที่ 0.43 มากกว่าลักษณ์ที่มาจากข้อมูลรับเข้าแบบตัวอักษณที่มีค่าความครบถ้วนเพียง 0.05 เมื่อนำค่าความแม่นยำและค่าความครบถ้วนของย่อหน้าที่ไม่ลักลอกมาร่วมด้วย พบว่าแบบจำลองทั้งสองประเภททำนายย่อหน้าที่ไม่มีการลักลอกได้ค่าความครบถ้วนและแม่นยำได้ไม่ต่างกันมากนัก ดังนั้น เมื่อใช้ค่าทางสถิติทั้ง 7 ลักษณ์ ลักษณ์ที่มาจากแบบจำลองแบบคำมีประสิทธิภาพที่ดีกว่าแบบตัวอักษร เพราะสามารถทำนายย่อหน้าที่ลักลอกได้ครบถ้วนและแม่นยำกว่าแบบตัวอักษร คือ ถูกต้อง 318 ย่อหน้า จาก 735 ย่อหน้า ในขณะที่ แบบตัวอักษรทำนายได้ ถูกต้องเพียง 40 ย่อหน้า

สำหรับการทดสอบลักษณ์ทางสถิติทั้ง 4 ครั้งข้างต้นกับข้อมูลรับเข้าทั้งสองประเภท ได้ผลดังที่รายงานไปแล้วข้างต้น ชุดของลักษณ์ทางสถิติที่ทดลองแล้วได้ผลที่ดีที่สุดคือแบบจำลองที่ทดลองใช้กับลักษณ์ทั้ง 7 ลักษณ์ จากแบบจำลองรับเข้าแบบคำ ดังผลการทดลองที่ 4.2.4 ทั้งนี้ ในขั้นต่อไปผู้วิจัยจะทดลองนำลักษณ์ทางสถิติที่ให้ผลดีที่สุดดังกล่าวไปทดลองร่วมกับลักษณ์ทางภาษา เพื่อทดสอบประสิทธิภาพผลของลักษณ์ทั้งสองประเภทที่มีต่อการทำนาย

4.3 การทดลองประสิทธิภาพแบบจำลองกับลักษณ์ทางภาษา

ในการทดลองขั้นนี้ ผู้วิจัยจะทดลองผลของลักษณ์ทางภาษาที่มีต่อแบบจำลอง เพื่อนำไปเปรียบเทียบกับผลการทดลองที่ดีที่สุดกับลักษณ์ทางสถิติในการทดลองที่ 4.2.4 ทั้งนี้ ผู้วิจัยได้เลือกทดลองกับลักษณ์ทางภาษา ใน 3 ลักษณะ คือ

1. ลักษณ์ที่เปรียบเทียบความต่างของชุดคำที่มีความถี่การใช้สูงที่สุดในเล่ม
2. ลักษณ์ที่เปรียบเทียบคำเขียนผิด จำนวน 6 ชุด คือ (กฎ, กฏ) (กะทันหัน, กระทันหัน) (เกม, เกมส์) (ปรากฏ, ปรากฎ) (พ.ศ., พศ.) (ศีรษะ, ศรีษะ)
3. ชุดลักษณ์ที่เปรียบเทียบการเลือกใช้คำและรูปแบบการเขียน จำนวน 6 ชุด คือ

(พ.ศ.<-s>, พ.ศ.<+s>) (เปอร์เซ็นต์, เครื่องหมาย%) (พ่อ, บิดา) (แม่, มารดา)(ร้อยละ, เปอร์เซ็นต์, เครื่องหมาย %) (อินเทอร์เน็ต, อินเตอร์เน็ต)

โดยชุดลักษณ์ทางภาษาข้างต้น ผู้วิจัยจะทดลองใน 2 ลักษณะ ได้แก่ ทดลองจากลักษณ์ที่มาจากค่า global คือ เปรียบเทียบย่อปัจจุบันกับย่อหน้าอื่นๆ ในเล่มแบบรวมกันทั้งหมดในครั้งเดียว และ ทดลองจากลักษณ์ที่มาจากค่า local คือ เปรียบเทียบย่อหน้าปัจจุบันกับย่อหน้าอื่นๆที่อยู่ติดกันภายในเล่มแบบเป็นช่วงๆ ซึ่งการทดลองครั้งนี้ผู้วิจัยได้กำหนดระยะ local อยู่ที่ 1 หน้าวิทยานิพนธ์ เท่ากับ 7 ย่อหน้า โดยช่วงย่อหน้านี้ ได้มาจากค่าเฉลี่ยจำนวนบรรทัดต่อย่อหน้าของคลังข้อมูลทั้งหมดหารด้วยความยาว 1 หน้าวิทยานิพนธ์ คือ 30 บรรทัด

สาเหตุที่ผู้วิจัยเลือกทดลองที่ระยะ local ด้วยนั้น เนื่องจาก ผู้วิจัยพบว่า ในงานเขียนเล่มหนึ่งๆ แม้ไม่ใช่ย่อหน้าที่ลักลอกมา ก็จะพบลักษณะการใช้คำหรือรูปแบบที่แตกต่างกันออกไป ดังตัวอย่าง <w>มี<w>ความ<w>เป็น<w>**พ่อ**เลี้ยง<w>หรือ<w>**แม่**เลี้ยง<w> และ <s>การ<w>เลี้ยงดู<w>โดย<w>**บิดา**<w>/<w>**มารดา**<w>เลี้ยง<w> ที่เขียนโดยผู้เขียนคนเดียวกันแต่อยู่คนละย่อหน้า โดยจะแยกทดลองลักษณ์แต่ละชนิดข้างต้น ออกด้วยการทดลองแต่ละชุดลักษณ์อย่างละ 2 ครั้งตามช่วงข้อมูลที่ใช้ทดลอง ดังต่อไปนี้

4.3.1 การทดลองกับลักษณ์ทางภาษาประเภทค่าต่างจากชุดคำศัพท์ที่มีความถี่สูงสุดในเล่ม 100 คำแรก – จำนวน 1 ลักษณ์

ผลการทดลองประสิทธิภาพของแบบจำลองต่อลักษณ์ทางภาษาที่แสดงค่าต่างจากชุดคำศัพท์ที่มีความถี่ 100 คำแรก ทดสอบแล้วได้ผล ดังต่อไปนี้

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
แบบ global	ความถี่ 100 คำแรก	ลักลอก	0.09	0.99	0.16
		ไม่ลักลอก	0.69	0.003	0.006
แบบ Local	ความถี่ 100 คำแรก	ลักลอก	0.0	0.0	0.0
		ไม่ลักลอก	0.91	1.0	0.95

ตารางที่ 4.5 แสดงผลของลักษณ์ทางภาษาแสดงค่าต่างจากชุดคำที่มีความถี่สูงสุด 100 คำแรก

ในการทดลองขั้นนี้ ผู้วิจัยได้ทดลองเปรียบเทียบค่าเฉลี่ยของคำที่มีความถี่ที่พบสูงที่สุดในชุดข้อมูลนั้นๆ ระหว่างย่อหน้าปัจจุบันและย่อหน้าอื่นๆในเล่ม เริ่มที่ ช่วงความถี่ที่ 100 คำแรกก่อน ทั้งนี้ชุดคำที่มีความถี่สูงสุด 100 อันดับแรกมีขนาดประมาณครึ่งหนึ่งของชุดข้อมูลทั้งหมด

ผลการทดลองลักษณ์ดังกล่าว ใน 2 ระยะ คือ แบบ global และ local มีค่าความครบถ้วนที่แตกต่างกันอย่างชัดเจน ในย่อหน้าที่ลักลอกผลที่ได้จากระยะ global มีค่าถึง 0.99 ในขณะที่ระยะ local กลับทีค่าเป็น 0 แต่เมื่อพิจารณาถึงค่าความแม่นยำและค่าความครบถ้วนของย่อหน้าที่ไม่ลักลอกเพิ่มเติม พบว่า ในระยะ global ทำนายคำตอบเป็นแบบย่อหน้าที่ลักลอกเกือบทั้งหมด และ ในระยะ local ทำนายคำตอบเป็นแบบย่อหน้าที่ไม่ลักลอกทั้งหมด ดังนั้น ลักษณ์ที่มาจากการเปรียบเทียบค่าเฉลี่ยคำที่มีความถี่สูงที่สุด 100 อันดับแรก จึงไม่สามารถใช้ทำนายการตรวจเทียบภายในหาการลักลอกได้

4.3.2 การทดลองกับลักษณ์ทางภาษาประเภทค่าต่างจากชุดคำศัพท์ที่มีความถี่สูงสุดในเล่ม 50 คำแรก – จำนวน 1 ลักษณ์

ผลการทดลองประสิทธิภาพของแบบจำลองต่อลักษณ์ทางภาษาที่แสดงค่าต่างจากชุดคำศัพท์ที่มีความถี่ 50 คำแรก ทดสอบแล้วได้ผล ดังต่อไปนี้

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
แบบ global	ความถี่ 50 คำแรก	ลักลอก	0.09	0.99	0.16
		ไม่ลักลอก	0.55	0.001	0.002
แบบ Local	ความถี่ 50 คำแรก	ลักลอก	0.09	0.99	0.16
		ไม่ลักลอก	0.83	0.001	0.001

ตารางที่ 4.6 แสดงผลของลักษณ์ทางภาษาแสดงค่าต่างจากชุดคำที่มีความถี่สูงสุด 50 คำแรก

จากผลการทดลองในข้อ 4.3.4 ที่ช่วงความถี่ 100 คำแรกไม่มีประสิทธิภาพในการตรวจจับย่อหน้าที่ลักลอกได้ ผู้วิจัยจึงปรับลดช่วงความถี่ลง เหลือเพียง 50 คำแรก เพื่อทดสอบผลที่ได้จากช่วงความถี่ที่ลดลง ทั้งนี้ชุดคำที่มีความถี่สูงสุด 50 อันดับแรกมีขนาดประมาณเศษหนึ่งส่วนสี่ของชุดข้อมูลทั้งหมด

ผลการทดลองลักษณ์ดังกล่าว ใน 2 ระยะ คือ แบบ global และ local ได้ผลการทดลองที่เหมือนกันคือ ในย่อหน้าที่ลักลอกมีค่าความครบถ้วนถึง 0.99 แต่เมื่อพิจารณาถึงค่าความแม่นยำและค่าความครบถ้วนของย่อหน้าที่ไม่ลักลอกเพิ่มเติม พบว่าทั้ง 2 ระยะ ทำนายคำตอบเป็นแบบย่อหน้าที่ลักลอกเกือบทั้งหมด ดังนั้น ลักษณ์ที่มาจากการเปรียบเทียบค่าเฉลี่ยคำที่มีความถี่สูงที่สุด 50 อันดับแรก จึงไม่สามารถใช้ทำนายการตรวจเทียบภายในหาการลักลอกได้

4.3.3 การทดลองกับลักษณ์ทางภาษาประเภทค่าต่างจากชุดคำศัพท์ที่มีความถี่สูงสุดในเล่ม 15 คำแรก – จำนวน 1 ลักษณ์

ผลการทดลองประสิทธิภาพของแบบจำลองต่อลักษณ์ทางภาษาที่แสดงค่าต่างจากชุดคำศัพท์ที่มีความถี่ 15 คำแรก ทดสอบแล้วได้ผล ดังต่อไปนี้

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
แบบ global	ความถี่ 15 คำแรก	ลักลอก	0.09	0.99	0.16
		ไม่ลักลอก	0.64	0.001	0.003
แบบ Local	ความถี่ 15 คำแรก	ลักลอก	0.0	0.0	0.0
		ไม่ลักลอก	0.91	1.0	0.95

ตารางที่ 4.7 แสดงผลของลักษณ์ทางภาษาแสดงค่าต่างจากชุดคำที่มีความถี่สูงสุด 15 คำแรก

จากผลการทดลองในข้อ 4.3.5 เมื่อลดช่วงความถี่เป็น 50 คำแรกไม่มีประสิทธิภาพในการตรวจจับย่อหน้าที่ลักลอกได้ ผู้วิจัยจึงปรับลดช่วงความถี่ลง เหลือเพียง 15 คำแรก เมื่อดูในภาพรวมของคลังข้อมูลทั้งหมด ช่วงความถี่นี้ มักประกอบด้วย คำประเภทหน้าที่เป็นหมวดคำปิดและมีคำในหมวดคำเปิดปนอยู่เล็กน้อย เช่น คำกริยาที่มักพบบ่อย เช่น เป็น มี เพื่อกำจัดอิทธิพลของคำเนื้อหาประเภทคำนามที่อาจมีความหลากหลายในแต่ละผู้เขียน จึงเลือกชุดคำที่มีความถี่สูง เพียง 15 อันดับแรก

ผลการทดลองลักษณ์ดังกล่าว ใน 2 ระยะ คือ แบบ global และ local ได้ผลการทดลองที่เหมือนกันคือ ในย่อหน้าที่ลักลอกมีค่าความครบถ้วนถึง 0.99 แต่เมื่อพิจารณาถึงค่าความแม่นยำและค่าความครบถ้วนของย่อหน้าที่ไม่ลักลอกเพิ่มเติม พบว่าทั้ง 2 ระยะ ทำนายคำตอบเป็นแบบย่อหน้าที่ลักลอกเกือบทั้งหมด ดังนั้น ลักษณ์ที่มาจากการเปรียบเทียบค่าเฉลี่ยคำที่มีความถี่สูงที่สุด 50 อันดับแรก จึงไม่สามารถใช้ทำนายการตรวจเทียบภายในหาการลักลอกได้

4.3.4 การทดลองแบบ กับลักษณ์ทางภาษาประเภทคำเขียนผิด – จำนวน 6 ชุด

ผลการทดลองประสิทธิภาพของแบบจำลองเมื่อเปรียบเทียบคำเขียนผิด จำนวน 6 ชุด ความหมาย ทดสอบแล้วได้ผลดังต่อไปนี้

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
แบบ global	คำเขียนผิด 6 ชุดคำ	ลักลอก	0.0	0.0	0.0
		ไม่ลักลอก	0.91	1.0	0.95
แบบ Local	คำเขียนผิด 6 ชุดคำ	ลักลอก	0.0	0.0	0.0
		ไม่ลักลอก	0.91	1.0	0.95

ตารางที่ 4.8 แสดงผลของลักษณ์ทางภาษาชุดคำเขียนผิด 6 ลักษณ์

เมื่อใช้ลักษณ์ทางภาษาที่เปรียบเทียบคำที่เขียนผิดจำนวน 6 ชุดความหมาย รวมคำศัพท์ 12 คำ พบว่า ผลที่ได้ของทั้งช่วงการทดลองทั้งแบบ global และ local นั้นได้ผลเหมือนกัน คือ ไม่สามารถแยกประเภทของย่อหน้าที่ลักลอกออกจากย่อหน้าที่ไม่ลักลอกได้ ทั้ง 2 ช่วงการทดลองต่างทำนาย ข้อมูลทดสอบเป็นแบบย่อหน้าที่ไม่ลักลอกมาทั้งหมด ดังนั้น ลักษณ์ทางภาษาประเภทชุดคำที่เขียนผิด ดังที่พบในชุดข้อมูล จำนวน 6 ชุดความหมาย 12 คำศัพท์ ไม่ได้ช่วยในการตรวจหาย่อหน้าที่มีการลักลอกในการทดลองนี้

4.3.5 การทดลองกับลักษณ์ทางภาษาประเภทการเลือกใช้คำและรูปแบบ – จำนวน 6 ชุด

ผลการทดลองประสิทธิภาพของแบบจำลองต่อลักษณ์ทางภาษาประเภทการเลือกใช้คำและรูปแบบ จำนวน 6 ชุดคำ 13 คำศัพท์ ทดสอบแล้วได้ผล ดังต่อไปนี้

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
แบบ global	คำ,รูปแบบ 6 ชุดคำ	ลักลอก	0.0	0.0	0.0
		ไม่ลักลอก	0.91	1.0	0.95
แบบ Local	คำ,รูปแบบ 6 ชุดคำ	ลักลอก	0.0	0.0	0.0
		ไม่ลักลอก	0.91	1.0	0.95

ตารางที่ 4.9 แสดงผลของลักษณ์ทางภาษาประเภทการเลือกใช้คำและรูปแบบ 6 ลักษณ์

เมื่อใช้ลักษณ์ทางภาษากลุ่มที่เปรียบเทียบประเภทการใช้คำและรูปแบบจำนวน 6 ชุดคำ รวมคำศัพท์ 13 คำ พบว่า ผลที่ได้ของทั้งช่วงการทดลองทั้งแบบ global และ local นั้นได้ผลเหมือนกัน คือ ไม่สามารถแยกประเภทของย่อหน้าที่ลักลอกออกจากย่อหน้าที่ไม่ลักลอกได้ ทั้ง 2 ช่วงการทดลองต่างทำนาย ข้อมูลทดสอบเป็นแบบย่อหน้าที่ไม่ลักลอกมาทั้งหมด ดังนั้น ลักษณ์ทางภาษาประเภทการเลือกใช้คำและรูปแบบ ดังที่พบในชุดข้อมูล จำนวน 6 ชุดคำ 13 คำศัพท์ ไม่ได้ช่วยในการตรวจหาย่อหน้าที่มีการลักลอกในการทดลองนี้

เมื่อทดสอบประสิทธิภาพของแบบจำลองกับลักษณ์ทางภาษาที่ผู้วิจัยสนใจศึกษาทั้งหมด 5 ครั้ง แบ่งตามกลุ่มย่อยๆ ดังที่ได้เสนอผลไปแล้วในข้างต้น พบว่าลักษณ์ทางภาษาที่นำมาทดลอง นั้นยังไม่สามารถตรวจหาย่อหน้าที่มีการลักลอกได้เลย ดังนั้น นำไปเปรียบเทียบกับชุดลักษณ์ทางสถิติที่ได้ผลดีที่สุดแล้ว ปรากฏว่าลักษณ์ทางสถิติแบบที่ใช้ลักษณ์จำนวน 7 ลักษณ์ ยังคงเป็นชุดที่ให้ผลการทำนายดีที่สุด

4.3.6 การทดลองกับลักษณ์ทางภาษาประเภทคำเขียนผิด 1 ลักษณ์ - จากคำสมมติ

เนื่องจาก ในชุดข้อมูลจริงแม้จะใช้ลักษณ์ที่เป็นคำเขียนผิดที่แตกต่างกันในเล่มกับแบบจำลอง พบว่าลักษณ์ดังกล่าวไม่มีประสิทธิภาพในการหาย่อหน้าที่มีการลักลอกแม้จะพบการใช้ปนกันจริงใน

ชุดข้อมูลก็ตาม ความเป็นไปได้หนึ่ง คือลักษณ์ที่ใช้บ่งชี้ว่าย่อหน้านั้นเป็นย่อหน้าที่ลักลอกมาเช่น การเขียนคำด้วยรูปที่ต่างจากรูปที่ใช้ในย่อหน้าอื่นๆ แต่เนื่องจากในการฝึกระบบด้วยข้อมูลที่ทดลองมา ลักษณ์นั้นอาจพบในเฉพาะในย่อหน้านั้น แต่ในย่อหน้าอื่นที่ลักลอกมาเช่นกันไม่ได้มีลักษณ์นี้เป็นตัวบ่งชี้ด้วย เมื่อระบบซัพพอร์ตเว็กเตอร์แมชชีนอ่านข้อมูลฝึกฝนก็จะเห็นว่าลักษณ์ตัวนี้ใช้บ่งบอกการลักลอกได้ในย่อหน้านี้ แต่ใช้ไม่ได้ในย่อหน้าอื่นๆที่ลักลอกมา ส่งผลให้ลักษณ์ที่ตั้งใจว่าจะใช้บ่งชี้การลักลอกไม่สามารถใช้ได้ในระบบนี้ เพื่อเป็นการทดสอบข้อสันนิษฐานนี้ ผู้วิจัยจึงได้ทำการทดลองเพิ่มเติมถึงลักษณะการใช้คำผิดหรือรูปแบบที่เขียนต่างกัน ต่อแบบจำลอง โดยผู้วิจัยได้ออกแบบลักษณ์สมมติขึ้นมาตัวหนึ่ง เป็นคำสมมติที่สร้างขึ้นและไม่พบจริงในชุดข้อมูล โดยคำสมมตินั้น สามารถเขียนได้ 2 แบบ และกำหนดให้การเขียนแบบที่ 1 พบในย่อหน้าประเภทที่ไม่ลักลอก และการเขียนแบบที่ 2 พบในย่อหน้าที่ลักลอกเท่านั้น ผลการทดลองพบว่า ลักษณ์ที่มาจากคำสมมติดังกล่าว สามารถแยกประเภทของย่อหน้าที่ลักลอกและไม่ลักลอกได้ถูกต้องทั้งหมด ดังนั้น จึงสรุปได้ว่า ลักษณ์ที่เป็นการเขียนคำผิด หรือรูปแบบการเขียนที่แตกต่างกัน สามารถใช้แยกย่อหน้าที่ลักลอกและไม่ลักลอกออกจากกันได้ แต่ต้องมีความถี่ในการพบโดยทั่วไปของการลักลอก

4.4 ทดลองใช้ลักษณ์ทางภาษาร่วมกับลักษณ์ทางสถิติ

ในส่วนนี้ผู้วิจัยจะลองนำลักษณ์ชุดที่ดีที่สุดของลักษณ์ทางสถิติและลักษณ์ทางภาษาที่ได้ทดลองไปแล้วคือลักษณ์ทางสถิติจำนวน 7 ลักษณ์จากข้อมูลรับเข้าแบบคำ และลักษณ์ทางภาษาที่มีประสิทธิภาพดีที่สุด แต่สำหรับลักษณ์ทางภาษาที่เลือกนำมาทดลองกลับไม่มีผลต่อการตรวจหาย่อหน้าที่มีการลักลอก ในขั้นนี้ผู้วิจัยจำนำลักษณ์ทางภาษาที่เกี่ยวกับการเขียนคำผิด การเลือกใช้คำและรูปแบบจำนวน 12 ลักษณ์ ในระยะ global มาดูผลที่มีต่อแบบจำลองเมื่อทดลองร่วมกับลักษณ์ทางสถิติ

ชนิดของข้อมูลรับเข้า	ลักษณ์ที่ใช้	ชนิดคำตอบ	ค่าความแม่นยำ	ค่าความครบถ้วน	ค่า F-measure
แบบ global	สถิติ ภาษา 12 ชุด	ลักลอก	0.23	0.32	0.26
		ไม่ลักลอก	0.93	0.89	0.91

ตารางที่ 4.10 แสดงผลการทดลองกับลักษณ์ทางภาษาร่วมกับลักษณ์ทางสถิติ

เมื่อทดลองใช้ลักษณ์ที่ให้ผลที่ที่สุดในชุดลักษณ์ทางสถิติร่วมกับชุดลักษณ์ทางภาษาพบว่า จากเดิมที่ใช้ลักษณ์ทางสถิติจำนวน 7 ลักษณ์กับข้อมูลรับเข้าแบบคำ การตรวจจับย่อหน้าที่ลักลอกมีค่าความแม่นยำที่ 0.18 ค่าความครบถ้วนที่ 0.43 และค่า F-measure ที่ 0.26 เมื่อเพิ่มลักษณ์ทางภาษาชุดคำเขียนผิด การเลือกใช้คำและรูปแบบ จำนวน 13 คำทดสอบร่วมด้วย มีผลให้ค่าความแม่นยำเพิ่มขึ้น จาก 0.18 เป็น 0.23 แต่ค่าความครบถ้วนลดลงจากเดิม จาก 0.43 เป็น 0.32 อย่างไรก็ดี ค่า F-measure ของการใช้ลักษณ์ทางภาษามาทดสอบร่วมด้วยไม่ได้ช่วยเพิ่มค่า

F-measure จากเดิมที่ใช้ลักษณ์ทางสถิติเพียงอย่างเดียว ดังนั้น เมื่อใช้ค่าความครบถ้วนเป้นหลักในการพิจารณาประสิทธิภาพของลักษณ์ต่อแบบจำลอง ลักษณ์ที่มาจากค่าทางสถิติจากข้อมูลแบบคำจำนวน 7 ลักษณ์จึงเป็นลักษณ์ที่ดีที่สุด โดย ทำนายย่อหน้าที่ลักลอกได้ 318 ย่อหน้า ขณะที่เมื่อทดลอง ร่วมกับลักษณ์ทางภาษาทำนายย่อหน้าที่ลักลอกได้ 238 ย่อหน้า

4.5 ผลการตรวจจับย่อหน้าที่มีการลักลอกด้วยลักษณ์ที่ได้ผลดีที่สุด

จากการทดลองในข้อ 4.2, 4.3 และ 4.4 ดังผลการทดลองข้างต้น พบว่าชุดลักษณ์ที่ให้ผลดีที่สุดคือลักษณ์ทางสิถิติ จำนวน 7 ลักษณ์ ที่มาจากข้อมูลรับเข้าแบบคำ ดังรายละเอียดที่ได้แสดงไว้ในตารางที่ 4.2.4 อย่างไรก็ตาม เมื่อทดลองเปรียบเทียบลักษณ์ทางภาษากับลักษณ์ทางสถิติ รวมถึงได้เลือกใช้กลุ่มลักษณ์มาทดลองร่วมกับลักษณ์ทางสถิติทั้ง 7 ลักษณ์แล้วก็ตาม พบว่าช่วยให้ค่าความแม่นยำเพิ่มมากขึ้นแต่ทำให้ค่าความครบถ้วนลดลง เมื่อดังผลการทดลองในตารางที่ 4.10 ผู้วิจัยจึงยังคงให้แบบข้อมูลรับเข้าชนิดคำกับลักษณ์ทางสถิติพิจารณาจากค่าความครบถ้วนเป็นหลัก ลักษณ์ทางสถิติ 7 ลักษณ์จากข้อมูลรับเข้าแบบคำเป็นชุดลักษณ์ที่มีผลที่ดีที่สุด ในส่วนนี้ผู้วิจัยจะวิเคราะห์ผลการตรวจจับที่ได้จากการทดลองของลักษณ์ดังกล่าว โดยผลของชิ้นงานที่ตรวจจับผิดจะแสดงผลเป็นค่าสัดส่วนข้อความที่ทำนายผิดคำนวณด้วยสูตร ดังนี้

ผลการทำนายที่ผิดประเภท A= (จำนวนย่อหน้าประเภท A ที่ทำนายผิด/จำนวนย่อหน้าประเภท A ทั้งหมด) *100

ทั้งนี้ ผลของการทำนายผิดข้างต้นจะแสดงแยกออกเป็น 3 รูปแบบ ได้แก่

1. แสดงผลการทำนายผิดตามปริมาณการลักลอกในเล่ม ดังตารางที่ 4.11
2. แสดงผลการทำนายผิดตามความยาวของเล่มต้นฉบับ ดังตารางที่ 4.12
3. แสดงผลการทำนายผิดตามภาพรวมของความยาวของย่อหน้าทั้งแบบลักลอกและไม่ลักลอก ดังตาราง ที่ 4.13

ลำดับ	ปริมาณลักลอก	จำนวนเล่ม	ย่อหน้าที่ลักลอก				ย่อหน้าที่ไม่ลักลอก			
			สั้น	กลาง	ยาว	รวม	สั้น	กลาง	ยาว	รวม
1	น้อย	12	70 100%	11 19.64%	5 20.83%	86 57.33%	116 4.91%	308 70.64%	99 73.33%	523 17.82%
2	ปานกลาง	9	97 100%	20 21.98%	17 53.12%	134 60.91%	57 3.55%	269 83.02%	96 53.63%	422 20.00%
3	มาก	9	158 100%	17 11.89%	22 34.38%	197 53.97%	75 3.64%	274 78.74%	110 72.85%	459 17.95%

ตารางที่ 4.11 แสดงผลที่แบบจำลองทำนายข้อมูลทดสอบผิดตามประเภทของปริมาณการลักลอกในเล่ม

เมื่อพิจารณาผลของการทำนายข้อมูลทดสอบแยกตามประเภทปริมาณการลักลอกในเล่ม พบว่าในแบบย่อหน้าที่ลักลอกมาแบบมาก มีความผิดพลาดน้อยที่สุด รองลงมาคือในเล่มที่ลักลอกมาน้อย และปานกลางตามลำดับ

ลำดับ	ความยาวต้นฉบับ	จำนวนเล่ม	ย่อหน้าที่ลักลอก				ย่อหน้าที่ไม่ลักลอก			
			สั้น	กลาง	ยาว	รวม	สั้น	กลาง	ยาว	รวม
1	น้อย	10	125 100%	53 68.83%	26 81.25%	204 87.18%	24 4.96%	126 85.71%	34 73.91%	184 27.18%
2	ปานกลาง	10	99 100%	18 21.69%	22 62.86%	139 64.05%	72 5.55%	211 75.36%	66 34.55%	349 19.74%
3	มาก	10	160 100%	28 15.47%	16 21.92%	204 49.27%	205 89.91%	514 75.48%	152 3.58%	871 16.89%

ตารางที่ 4.12 แสดงผลที่แบบจำลองทำนายข้อมูลทดสอบผิดตามประเภทของความยาวต้นฉบับ

เมื่อพิจารณาผลของการทำนายข้อมูลทดสอบแยกตามขนาดความยาวของต้นฉบับ พบว่าในประเภทย่อหน้าที่ลักลอกและไม่ลักลอกมีผลการตรวจจับผิดแบบเดียวกัน คือ แบบที่มีความยาวต้นฉบับยาวน้อยมีการตรวจจับผิดมากที่สุด รองลงมาคือ แบบที่มีข้อมูลต้นฉบับยาวปานกลาง และแบบที่มีความยาวต้นฉบับแบบมากมีการทำนายผิดน้อยที่สุด จึงสรุปได้ว่าขนาดความยาวของต้นฉบับมีผลต่อความแม่นยำในการแยกประเภทย่อหน้าลักลอกและไม่ลักลอก

ย่อหน้าที่ลักลอก				ย่อหน้าที่ไม่ลักลอก			
สั้น	กลาง	ยาว	รวม	สั้น	กลาง	ยาว	รวม
325 100%	48 16.55%	44 36.67%	417 56.73%	248 4.11%	851 76.81%	305 65.59%	1404 18.47%

ตารางที่ 4.13 แสดงผลที่แบบจำลองทำนายข้อมูลทดสอบผิดตามประเภทของความยาวของย่อหน้า

เมื่อมองภาพรวมของย่อหน้าที่ลักลอกและไม่ลักลอกทั้งหมด โดยไม่อ้างอิงปริมาณการลักลอกต่อเล่มหรือความยาวของต้นฉบับ พบว่าย่อหน้าที่มีการลักลอกมีการตรวจจับผิดพลาดมากกว่าย่อหน้าที่ไม่มีการลักลอก ผลของการตรวจจับย่อหน้าที่ลักลอกมา ย่อหน้าขนาดกลางมีความแม่นยำมากที่สุด รองลงมาคือย่อหน้าขนาดยาว ทั้งนี้เป็นที่น่าสังเกตคือ แบบจำลองนี้ยังไม่สามารถตรวจจับย่อหน้าลักลอกขนาดสั้นได้เลย นอกจากนั้น ผลของแบบจำลองโดยรวมยังมีค่าที่ตรวจจับผิดเกินครึ่งหนึ่ง (56.73 เปอร์เซ็นต์) ซึ่งแสดงว่าแบบจำลองสามารถตรวจจับย่อหน้าลักลอกได้น้อยกว่าครึ่งหนึ่ง

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ในส่วนนี้ ผู้วิจัยจะสรุปผลการวิจัยที่ได้จากผลการทดลองในบทที่ 4 ตามสมมติฐานที่ตั้งไว้ก่อนการวิจัย อภิปรายผล และข้อเสนอแนะสำหรับการวิจัยครั้งต่อไป

5.1 สรุปผลการวิจัย

จากผลการทดลองในบทที่ 4 ผู้วิจัยจะสรุปผลการทดลองทั้งหมดตามสมมติฐานที่ได้ตั้งไว้ก่อนการวิจัย ทั้ง 3 ข้อ ดังนี้

สมมติฐานข้อที่ 1 แบบจำลองที่ใช้ข้อมูลแบบรับเข้าเป็นคำให้ผลดีกว่าแบบจำลองที่ใช้ข้อมูลรับเข้าเป็นตัวอักษร

จากการทดสอบ แบบจำลองที่ใช้ข้อมูลรับเข้าทั้ง 2 ประเภท คือ แบบรับเข้าเป็นคำและแบบรับเข้าเป็นตัวอักษรที่มาจากข้อมูลชุดเดียวกัน เปรียบเทียบด้วยลักษณ์ทางสถิติจำนวนเท่ากัน และใช้สูตรการคำนวณแบบเดียวกันทั้งหมด ทดลองโดยแบ่งลักษณ์ทางสถิติที่มีเป็นกลุ่มย่อยๆ ดังรายละเอียดในบทที่ 4 เมื่อทดลองเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองกับลักษณ์ทางสถิติกลุ่มต่างๆ พบว่า ผลการทดลองที่ดีที่สุดของลักษณ์ที่มาจากข้อมูลรับเข้าแบบคำ คือ แบบจำลองที่ใช้ชุดลักษณ์ทางสถิติทั้ง 7 ลักษณ์ โดยมีค่าความครบถ้วนของย่อหน้าลักลอกที่ 0.43 ตรวจจับได้ 318 ย่อหน้า จาก 735 ย่อหน้า ส่วนการทดลองที่ใช้ลักษณ์ทางสถิติจากข้อมูลรับเข้าแบบตัวอักษร ได้ผลไปในทางเดียวกันหมดคือ แบบจำลองจะทำนายข้อมูลทดสอบเป็นประเภทใดประเภทหนึ่งเพียงอย่างเดียว ทั้งนี้ แม้ในการทดลองจะมีค่าความครบถ้วน 1.0 แต่กลับมีค่าความถูกต้องต่ำ เนื่องจากแบบจำลองล้มเหลวที่จะแยกย่อหน้าทั้งสองประเภทออกจากกัน ดังนั้น สมมติฐานที่ได้ตั้งไว้ก่อนการวิจัยว่าแบบจำลองที่ใช้ข้อมูลรับเข้าแบบคำจะได้ผลดีกว่าแบบจำลองที่มีข้อมูลรับเข้าแบบตัวอักษรจึงเป็นไปตามสมมติฐาน

สมมติฐานข้อที่ 2 ลักษณ์ทางภาษาให้ผลที่ดีกว่าลักษณ์ที่ไม่ใช่ลักษณ์ทางภาษา

เมื่อเปรียบเทียบผลจากการทดสอบของลักษณ์ทางสถิติ ที่มาจากค่าทางสถิติ 3 กลุ่ม คือ ค่าทางสถิติที่มาจากคำหรือตัวอักษร ค่าทางสถิติที่มาจากคำหรือตัวอักษรและช่องว่าง และค่าทางสถิติที่แตกต่างกันระหว่างย่อหน้ากับในเล่ม ทดลองทั้งในแบบจำลองที่รับเข้าแบบคำและตัวอักษร ได้ผลการทดลองที่ดีที่สุด คือลักษณ์ทางสถิติ 7 ลักษณ์จากข้อมูลรับเข้าแบบคำ ดังที่กล่าวถึงไปแล้วในสรุปสมมติฐาน ข้อ 1 อย่างไรก็ตาม แม้จะใช้ลักษณ์ทางภาษาที่พัฒนามาจากลักษณะทางภาษาที่พบจริงในชุดข้อมูล คือ เปรียบเทียบคำเขียนผิด เปรียบเทียบการเลือกใช้คำและรูปแบบการเขียน รวมถึงเปรียบเทียบค่าเฉลี่ยของคำที่มีความถี่สูงสุดในชุดข้อมูลกับย่อหน้านั้นๆ ทั้งในระยะ global และ

local กลับไม่สามารถตรวจจับย่อหน้าที่ลักลอกได้เลย ทั้งนี้ เพื่อพิสูจน์ข้อสันนิษฐานของผู้วิจัยเกี่ยวกับลักษณะการเขียนคำผิดที่แตกต่างกันระหว่างผู้เขียน จึงออกแบบการทดลองเพิ่มเติมโดยใช้คำสมมติ 1 คำ กำหนดให้การเขียนแบบที่ 1 มาจากผู้เขียนต้นฉบับและพบเฉพาะในย่อหน้าที่ไม่ลักลอก ส่วนการเขียนแบบที่ 2 มาจากผู้เขียนคนอื่นๆและพบเฉพาะในย่อหน้าที่ไม่ลักลอก ทดสอบเฉพาะแบบเปรียบเทียบกับย่อหน้าอื่นๆในเล่ม (ระยะ global) ผลปรากฏว่า แบบจำลองสามารถทำนายผลได้ถูกต้องทั้งหมด ดังนั้น จึงสรุปได้ว่าลักษณ์ทางภาษาศาสตร์มีประสิทธิภาพที่ดีกว่าลักษณ์ทางสถิติเมื่อมีเงื่อนไขเรื่องความถี่มาเกี่ยวข้อง อย่างไรก็ตาม สำหรับผลของชุดลักษณ์ที่เลือกใช้ในการทดลองครั้งนี้ แต่ให้ผลที่แตกต่างออกไปผู้วิจัยจะอภิปรายเพิ่มเติมต่อไป

สมมติฐานข้อที่ 3 ข้อความที่ลักลอกมายิ่งยาวก็ยิ่งตรวจจับการลักลอกได้แม่นยำมากขึ้น

จากผลของการทดสอบแบบจำลองกับชุดลักษณ์ที่ให้ค่าความครบถ้วนมากที่สุด คือ ชุดลักษณ์ทางสถิติจากแบบจำลองรับเข้าแบบคำ จำนวน 7 ลักษณ์ มีค่าความแม่นยำ ที่ 0.18 ค่าความครบถ้วนที่ 0.43 และค่า F-measure ที่ 0.26 ตรวจจับย่อหน้าที่ลักลอก ได้ถูกต้อง 318 จาก 735 ย่อหน้า จำนวนย่อหน้าที่ลักลอกที่ตรวจจับผิดเมื่อแยกตามประเภทความยาวของต้นฉบับ พบว่าย่อหน้าที่ลักลอกแบบสั้นมีความผิดพลาดมากที่สุด คือ ไม่สามารถตรวจหาย่อหน้าที่ลักลอกได้เลย ประเภทที่มีความผิดพลาดรองลงมา คือ ย่อหน้าแบบยาวที่ตรวจหาผิดพลาดที่ 36.67 เปอร์เซ็นต์ ส่วนย่อหน้าที่มีการตรวจจับได้แม่นยำที่สุดคือย่อหน้าขนาดยาวปานกลาง ที่ตรวจหาผิดพลาดที่ 16.55 เปอร์เซ็นต์ ดังนั้น ข้อสมมติฐานที่ตั้งไว้ว่าข้อความที่ลักลอกมายิ่งยาวก็ยิ่งตรวจจับได้แม่นยำมากขึ้นจากผลการทดลองในที่นี้จึงไม่สามารถบอกถึงปัจจัยความยาวของย่อหน้าที่ลักลอกว่าจะมีผลต่อการตรวจจับการลักลอกแบบภายในหรือไม่

5.2 อภิปรายผล

งานวิจัยนี้ ออกแบบมาเพื่อเปรียบเทียบผลการตรวจจับย่อหน้าที่มีการลักลอก จากลักษณ์ 2 ประเภท คือ ลักษณ์ทางสถิติและลักษณ์ทางภาษา โดยลักษณ์ทางภาษาที่เลือกนำมาใช้ อ้างอิงมาจากแนวคิดหลักของการตรวจเทียบภายในหาการลักลอกงาน ที่วิเคราะห์ส่วนที่น่าสงสัยจากลักษณะทางภาษาต่างๆที่ไม่เหมือนกับส่วนอื่นๆในงานชิ้นเดียวกัน การระบุลักษณะเฉพาะในการเขียนของแต่ละผู้เขียน จากงานวิจัยที่ผ่านมา มีการใช้ลักษณ์ที่หลากหลาย ทั้งในระดับตัวอักษร ระดับคำ รวมถึงระดับโครงสร้างของประโยค ซึ่งในการทดลองนี้ การทดลองด้วยลักษณ์ทางสถิติจะทดลองกับแบบจำลองที่มีข้อมูลรับเข้าแบบตัวอักษรและแบบคำ ส่วนการทดลองด้วยลักษณ์ทางภาษาจะทดลองกับแบบจำลองที่มีข้อมูลรับเข้าแบบคำเพียงอย่างเดียว ข้อสันนิษฐานของผู้วิจัย คือ เมื่อเปรียบเทียบระหว่างแบบจำลองที่ข้อมูลรับเข้าเป็นแบบคำและแบบตัวอักษร เพื่อระบุการลักลอกงานในย่อหน้าหนึ่งๆ ข้อมูลรับเข้าแบบคำจะให้ผลที่ดีกว่า ซึ่งผลจากการทดลองก็เป็นจริง ดังสรุปสมมติฐานที่ 1

ข้างต้น ในส่วนที่เกี่ยวกับการทดลองกับลักษณ์ทางภาษา แม้จะเลือกใช้ลักษณ์ทางภาษาที่พบการใช้ไม่สม่ำเสมอกันจริงในผู้เขียนคนหนึ่งๆ และพบเป็นจำนวนหนึ่งในข้อมูล กลับไม่สามารถแยกย่อหน้าที่ลักลอกออกจากย่อหน้าที่ลักลอกได้เลย ทั้งนี้ สาเหตุที่ลักษณ์ทางสถิติเพียงอย่างเดียวมีผลต่อแบบจำลองมากกว่าที่ใช้ลักษณ์ทางภาษา ผู้วิจัยมีข้อสังเกตดังนี้

1. เมื่อเปรียบเทียบกันในย่อหน้าหนึ่งๆ ลักษณ์ทางสถิติจะมีส่วนที่แตกต่างกันเสมอ ซึ่งอาจไม่พบความแตกต่างใดๆในลักษณ์ทางภาษาเลย แม้ย่อหน้ามีคำตอบต่างกัน ดังตัวอย่าง

	ย่อหน้า	1	2	3	4	5	6	7	คำตอบ
1.1	11001_7	57.9	-83.1	13.49	-34.51	4.29	-1.35	0.55	no
1.2	11001_8	58.03	-28.97	13.56	-2.44	4.28	1.16	0.34	yes
1.3	11001_7	0	0	0	0	0	0	0	no
1.4	11001_8	0	0	0	0	0	0	0	yes

จากตัวอย่าง 2 ย่อหน้าจากข้อมูลฝึกฝนข้างต้น จะเห็นว่าลักษณ์ทางสถิติในตัวอย่างที่ 1.1 และ 1.2 มีคำตอบที่ต่างกันและมีค่าทางสถิติต่างกัน ในขณะที่ลักษณ์ทางภาษาในตัวอย่างที่ 1.3 และ 1.4 มีค่าไม่ต่างกันแม้ในย่อหน้าที่คำตอบต่างกัน เมื่อสมมติให้ตัวอย่างย่อหน้า 11001_7 และ 11001_8 เป็นงานเขียนรวม 1 ชิ้นที่ต้องการศึกษา ค่าของลักษณ์ทางสถิติที่แตกต่าง ใช้เพื่อแยกย่อหน้าที่คำตอบต่างกันได้ ในขณะที่ลักษณ์ทางภาษามีค่าของลักษณ์ไม่ต่างกัน แม้มีคำตอบต่างกัน เมื่อการทำงานของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ใช้การเรียนรู้จากระนาบข้อมูลฝึกฝนมาสร้างไฮเปอร์เพลนเพื่อใช้จัดกลุ่มชุดข้อมูลทำนาย ดังนั้น เมื่อลักษณ์ทางภาษามีค่าที่เหมือนกัน แบบจำลองจึงแยกย่อหน้าที่มีคำตอบต่างกันไม่ได้

2. ลักษณ์ทางภาษา เกี่ยวกับคำเขียนผิด การเลือกใช้คำและรูปแบบที่ต่างกัน แม้จะเลือกใช้ชุดคำที่พบเขียนผิดและใช้ต่างจริงในชุดข้อมูล คัดเลือกที่ความถี่การพบโดยรวมตามเกณฑ์ที่กำหนดไว้ ซึ่งลักษณะการใช้ปนแบบนี้ ใช้เป็นข้อสังเกตที่มีประโยชน์เพื่อระบุรูปแบบการเขียนที่ต่างออกไป กลับใช้ไม่ได้จริงกับแบบจำลอง แต่กลับให้ผลที่ต่างกันกับการทดลองด้วยคำสมมติ ที่มีค่าลักษณ์แตกต่างกันจริงพบทั้งในข้อมูลฝึกฝนและข้อมูลทดสอบ จึงกล่าวได้ว่า เมื่อการพบลักษณ์ทางภาษานั้นๆในชุดข้อมูลมีจำนวนจำกัด ลักษณ์นั้นๆ จึงใช้ไม่ได้จริงกับแบบจำลอง ทั้งนี้ ผู้วิจัยได้รวบรวมสถิติของลักษณ์ทางภาษาเกี่ยวกับคำเขียนผิด การเลือกใช้คำและรูปแบบ เปรียบเทียบกับย่อหน้าอื่นๆทั้งหมดในเล่ม (ระยะ global) ในทั้งชุดข้อมูลทั้งหมดไว้ โดยสรุปผล ออกเป็น 4 ประเภทใหญ่ๆ ตามประเภทคลังข้อมูลและคำตอบ จากนั้นแจกแจงตามจำนวนลักษณ์ที่พบต่างจากย่อหน้าอื่นๆจากทั้งหมด 12

ลักษณ์ และแสดงสัดส่วนแต่ละข้อเป็นเปอร์เซ็นต์เพื่อชี้ให้เห็นว่ามีสัดส่วนที่พบเป็นส่วนน้อยจึงมีผลต่อแบบจำลองต่างจากการทดลองกับคำสมมติ ดังนี้

2.1 ข้อมูลฝึกฝนที่เป็นย่อหน้าไม่ลักลอก

2.1.1 พบ 53,872 ย่อหน้า หรือ 98.87 เปอร์เซ็นต์ ไม่พบลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆ เลย

2.1.2 พบ 575 ย่อหน้า หรือ 1.05 เปอร์เซ็นต์ ใช้ลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆ จำนวน 1 ใน 12 ลักษณ์

2.1.3 พบ 39 ย่อหน้า หรือ 0.08 เปอร์เซ็นต์ ใช้ลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆ จำนวน 2 ใน 12 ลักษณ์

2.2 ข้อมูลฝึกฝนที่เป็นย่อหน้าลักลอก

2.2.1 พบ 5,138 ย่อหน้า หรือ 97.11 เปอร์เซ็นต์ ไม่พบลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆเลย

2.2.2 พบ 149 ย่อหน้า หรือ 2.82 เปอร์เซ็นต์ ใช้ลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆ จำนวน 1 ใน 12 ลักษณ์

2.2.3 พบ 4 ย่อหน้า หรือ 0.07 เปอร์เซ็นต์ใช้ลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆ จำนวน 2 ใน 12 ลักษณ์

2.3 ข้อมูลทดสอบที่เป็นย่อหน้าไม่ลักลอก

2.3.1 พบ 7,532 ย่อหน้า หรือ 99.09 เปอร์เซ็นต์ ไม่พบลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆเลย

2.3.2 พบ 62 ย่อหน้า หรือ 0.82 เปอร์เซ็นต์ ใช้ลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆ จำนวน 1 ใน 12 ลักษณ์ (0.82 เปอร์เซ็นต์ของย่อหน้าไม่ลักลอกในข้อมูลทดสอบ)

2.3.3 พบ 7 ย่อหน้า หรือ 0.09 เปอร์เซ็นต์ ใช้ลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆ จำนวน 2 ใน 12 ลักษณ์ (0.09 เปอร์เซ็นต์ของย่อหน้าไม่ลักลอกในข้อมูลทดสอบ)

2.4 ข้อมูลทดสอบที่เป็นย่อหน้าลักลอก

2.4.1 พบ 691 ย่อหน้า หรือ 94.01 เปอร์เซ็นต์ ไม่พบทางภาษาต่างจากย่อหน้าอื่นๆเลย

2.4.2 พบ 40 ย่อหน้า หรือ 5.44 เปอร์เซ็นต์ ใช้ลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆ จำนวน 1 ใน 12 ลักษณ์

2.4.3 พบ 4 ย่อหน้า หรือ 0.54 เปอร์เซ็นต์ ใช้ลักษณ์ทางภาษาต่างจากย่อหน้าอื่นๆ จำนวน 2 ใน 12 ลักษณ์

จากสถิติการใช้ลักษณ์ที่แตกต่างกันในข้อมูลฝึกฝนและทดสอบข้างต้น จะเห็นว่า จากจำนวนลักษณ์ทั้งหมด 12 ลักษณ์ มีการใช้ต่างกันสูงที่สุดพร้อมกันเพียง 2 ใน 12 ลักษณ์เท่านั้นและยังพบในสัดส่วนที่น้อยตามประเภทย่อหน้าและคลังข้อมูล ยิ่งไปกว่านั้น ลักษณ์ที่ใช้ต่างกันและมีความถี่มากที่สุดที่พบ คือ (พ.ศ.<+s>, พ.ศ.<-s>) รวม 348 ครั้ง รองลงมาคือ (ปรากฏ, ปรากฏ) รวม 168 ครั้ง แต่พบทั้งในข้อมูลฝึกฝนและทดสอบอยู่ในย่อหน้าทั้งลักลอกและไม่ลักลอก เหตุนี้ ลักษณ์ทางภาษาที่ใช้ทดลองในการวิจัยนี้จึงมีผลต่างจากที่ได้ในการทดลองกับคำสมมติซึ่งจะพบลักษณ์บ่งชี้ในย่อหน้าที่ลักลอกอย่างสม่ำเสมอ

3. การใช้คำเขียนผิดหรือเขียนต่างรูปแบบในผู้เขียนคนเดียวกัน ซึ่งมีผลให้ลักษณ์ประเภทนี้ใช้ไม่ได้ผลกับแบบจำลองนั้น ขัดกับแนวคิดที่ผู้เขียนจะคงลักษณะนั้นๆไว้ตลอดงานเขียนของตนเอง ในที่นี้ผู้วิจัยจะกล่าวเพิ่มเติมจากจำนวน สัดส่วน และความหลากหลายที่พบดังที่กล่าวไปแล้วในข้อ 2 โดยจะแสดงตัวอย่างลักษณะการเขียนปนกันที่พบ เพื่ออธิบายถึงลักษณะข้อมูลจริงในงานเขียนเชิงวิชาการที่ขัดแย้งกับแนวคิดข้างต้น ผู้วิจัยจะอภิปรายเพิ่มเติมจากตัวอย่างที่พบดังนี้

3.1 (อินเทอร์เน็ต, อินเตอร์เน็ต) คำชุดนี้ยังไม่พบในพจนานุกรมฉบับราชบัณฑิตยสถาน ปี พ.ศ. 2542 แต่พบในรายการศัพท์บัญญัติวิชาการที่ถ่ายถอดเสียงมาจากภาษาต่างประเทศของเวบไซค์ราชบัณฑิตยสถานให้เขียนว่า “อินเทอร์เน็ต” ซึ่งขัดแย้งกับหลักเกณฑ์การเขียนคำทับศัพท์ภาษาอังกฤษที่เผยแพร่ในเวบไซค์ราชบัณฑิตยสถาน ที่ให้ใช้ “ท” แทนเสียง “t” ในกรณีที่เป็นพยัญชนะต้น และให้ใช้ “ต” แทนตัวสะกดและการันย์ แต่มีหมายเหตุเพิ่มเติมว่า ยกเว้นในกลุ่มพยัญชนะที่ไทยนิยมใช้ “ต” ยกตัวอย่างเช่น intercom “อินเตอร์คอม” ผู้วิจัยจึงจัดคำชุดนี้อยู่ในประเภทคำเขียนต่างแทนคำเขียนผิด เมื่อสุ่มดูคำคู่นี้ในเล่มวิทยานิพนธ์ที่พบบ่อย คือ เล่มเลขที่ 11016 มีการปรากฏของคำและค่าลักษณ์ต่างไป ดังตัวอย่าง

3.1.1 ย่อหน้าที่ 11016_37 เป็นย่อหน้าแบบไม่ลักลอก มีลักษณ์ของชุดคำ (อินเทอร์เน็ต, อินเตอร์เน็ต) คือ -1 แบบเจอการใช้ปนกันทั้ง 2 คำ

*^จากการทบทวนวรรณกรรม พบว่า การใช้สิ่งเสพติดเป็นวิถีชีวิตสำคัญที่มีผลต่อสุขภาพของเด็กวัยรุ่นตอนต้น และสำหรับเด็กวัยรุ่นตอนต้นที่ติด**อินเทอร์เน็ต** สิ่งเสพติดที่สำคัญที่สุดคือ **อินเทอร์เน็ต** (Robinson and Kish , 2001)*

*วิถีชีวิตของเด็กวัยรุ่นตอนต้นที่ติด**อินเตอร์เน็ต** นอกจากจะครอบคลุมพฤติกรรมสุขภาพแล้วยังรวมถึง ระยะเวลาการติด**อินเตอร์เน็ต** และลักษณะการใช้**อินเตอร์เน็ต**ของ เด็กอีกด้วย เพราะสิ่ง*

*เหล่านี้สะท้อนถึงวิถีชีวิตของเด็กวัยรุ่น เนื่องจากการติด**อินเทอร์เน็ต**เป็นการ เปลี่ยนแปลงวิถีชีวิตของเด็กนั่นเอง (ธนิกานต์ มาฆะศิรานนท์, 2545)*

*เด็กเหล่านี้มักมีเวลาในการพักผ่อนนอนหลับน้อยกว่าปกติ จากการหมกมุ่นกับการเล่น **อินเทอร์เน็ต** (ไชยวัฒน์ บุตรพรหม,2545; Peter & Bodkin ,2007; Orzack, 2001; Wieland, 2005)^*

3.1.2 ย่อหน้าที่ 11016_38 เป็นย่อหน้าแบบไม่ลักลอก มีลักษณ์ของชุดคำ (อินเทอร์เน็ต, อินเตอร์เน็ต) คือ 1 แบบเจอการใช้คำชุดนี้ในย่อหน้าเหมือนกับย่อหน้าอื่นๆส่วนใหญ่ในเล่ม

*^จะเห็นได้ว่า การใช้**อินเทอร์เน็ต**มีความเกี่ยวข้องระยะเวลาในการใช้**อินเทอร์เน็ต**เป็นอย่างมาก ซึ่งจากการทบทวนวรรณกรรมพบว่า คนไทยใช้เวลาในการเล่น**อินเทอร์เน็ต**มากเกินไป เวลาการใช้ที่ไม่เหมาะสม เล่นกันตั้งแต่ช่วงหัวค่ำถึงดึก (ทวีศักดิ์ กออนันตกุล, 2543 อ้างถึงในไชยวัฒน์บุตรพรหม, 2545)*

*ซึ่งผลกระทบของการที่ผู้ใหญ่ใช้**อินเทอร์เน็ต**เป็นเวลานาน ๆ จะส่งผลทำให้เสียสายตา เนื่องจากจ้องจอคอมพิวเตอร์นาน ๆ การนอนหลับพักผ่อนไม่เพียงพอ เกิดอาการอ่อนเพลีย ปวดหลัง ปวดศีรษะข้างเดียว สุขภาพทรุดโทรม ส่งผลกระทบต่อหน้าที่การงาน และเกิดความเจ็บป่วยทางด้านร่างกายที่รุนแรงตามมา (Young, 1999 cited in Wang, 2001)*

*เช่นเดียวกันหากเด็กวัยรุ่นตอนต้นที่ติด**อินเทอร์เน็ต**มีระยะเวลาการใช้**อินเทอร์เน็ต**ที่นานขึ้นและมีความต่อเนื่องกัน ก็น่าจะมีผลกระทบต่อเด็กวัยรุ่นตอนต้นที่ติด**อินเทอร์เน็ต**เช่นเดียวกับผู้ใหญ่และอาจรุนแรงกว่า ไม่ว่าจะเป็นความผิดปกติทางด้านสุขภาพร่างกาย ด้านการเรียน ด้านสัมพันธ์ทางสังคม ด้านการเงิน เป็นต้น ในการวิจัยครั้งจึงได้นำการใช้**อินเทอร์เน็ต**มาศึกษา ซึ่งน่าจะมีความสัมพันธ์ทางลบกับสุขภาวะของเด็กวัยรุ่นตอนต้นที่ติด**อินเทอร์เน็ต**^*

3.1.3 ย่อหน้าที่ 11016_607 เป็นย่อหน้าแบบไม่ลักลอก มีลักษณ์ของชุดคำ (อินเทอร์เน็ต, อินเตอร์เน็ต) คือ -1 แบบเจอการใช้คำชุดนี้ในย่อหน้าต่างจากย่อหน้าอื่นๆส่วนใหญ่ในเล่ม

*^2. วางแผนร่วมกันในการควบคุมเวลาที่ใช้ โดยพยายามคิดถึงการใช้ และมีพฤติกรรมเกี่ยวกับการใช้ให้ลดลง หากมีการกลับไปใช้อีก ให้ผู้ติดระบุสาเหตุการกลับไปใช้ใหม่ว่ามีสาเหตุมาจากสิ่งใด และให้บันทึกสาเหตุและความรู้สึกโดยหลีกเลี่ยงการบันทึกด้วยคอมพิวเตอร์ ซึ่งผู้รักษาต้องคอยให้กำลังใจ และทำให้ผู้เข้ารับการรักษาเชื่อมั่นว่าเขาสามารถทำได้ที่นำเสนอมาทั้งหมดเป็นการเสนอเรื่องการติด**อินเตอร์เน็ต**ในภาพกว้าง ไม่ได้เฉพาะเจาะจงกับวัยรุ่น เหมือนที่ระบุไว้ในหัวข้อ^*

3.1.4 ย่อหน้าที่ 11016_682 เป็นย่อหน้าแบบไม่ลักลอก มีลักษณ์ของชุดคำ (อินเทอร์เน็ต, อินเตอร์เน็ต) คือ 1 แบบเจอการใช้คำชุดนี้ในย่อหน้าเหมือนกับย่อหน้าอื่นๆส่วนใหญ่ในเล่ม

*^ผลการศึกษาพบว่า การใช้**อินเทอร์เน็ต**ไม่มีความสัมพันธ์กับสุขภาวะของเด็กวัยรุ่นตอนต้นที่ติด**อินเทอร์เน็ต**อย่างมีนัยสำคัญทางสถิติ สามารถอธิบายได้ว่า เมื่อเด็กวัยรุ่นตอนต้นที่ติด**อินเทอร์เน็ต**มีการใช้อินเทอร์เน็ตในระยะเวลาที่มากหรือน้อย ก็ไม่มีความเกี่ยวข้องกับสุขภาวะของเด็กวัยรุ่นตอนต้นที่ติด**อินเทอร์เน็ต** ซึ่งไม่เป็นไปตามสมมติฐานข้อที่ 3 ที่กล่าวว่า การใช้**อินเทอร์เน็ต**มีความสัมพันธ์ทางลบกับสุขภาวะของเด็กวัยรุ่นตอนต้นที่ติด**อินเทอร์เน็ต**^*

จากตัวอย่างการใช้รูปแบบการเขียนที่ต่างกันปนของผู้เขียน จำนวน 4 ย่อหน้า อธิบายได้ว่า ผู้เขียนเลือกใช้คำว่า “อินเทอร์เน็ต” มากกว่า “อินเตอร์เน็ต” สอดคล้องกับ จากย่อหน้าทั้งหมด 692 ย่อหน้า ลักษณ์นี้มีค่า 1 ที่ 220 ย่อหน้า และ -1 ที่ 8 ย่อหน้า จากตัวอย่างที่ 3.1.1 มีการใช้ปนกันทั้ง 2 แบบในย่อหน้า เมื่อวิเคราะห์ถึงเนื้อหาภายในย่อหน้า จะพบว่าเป็นเนื้อหาที่ผู้เขียนอ้างอิงมาจากผู้อื่นจำนวน 3 ส่วน ดังที่ผู้วิจัยได้แบ่งออกเป็น 3 ย่อหน้าในตัวอย่าง จากตัวอย่างที่ 3.1.2 การใช้คำเหมือนกันทั้งย่อหน้าแม้จะอ้างอิงจากผู้อื่นมาจำนวน 2 ส่วน และจากตัวอย่างที่ 3.1.3 การใช้คำพบเพียงหนึ่งครั้งและเป็นไปตามแบบย่อหน้าอื่นๆในเล่ม เป็นการเขียนด้วยผู้เขียนเองโดยไม่มีการอ้างอิงจากที่อื่น จากตัวอย่างทั้ง 3 อาจกล่าวได้ว่า รูปแบบการเขียนที่ปนกันในงานของผู้เขียนคนนั้นๆ อาจไม่ได้แสดงลักษณะการเขียนของบุคคลนั้นๆเลย ถ้าเป็นส่วนที่อ้างอิงมาโดยตรงจากผู้อื่น ทั้งนี้ในตัวอย่างที่อ้างอิงมาทั้ง 3 ส่วน ในตัวอย่างที่ 3.1.1 มีทั้งการใช้รูปแบบต่างกันระหว่างข้อความที่อ้างมาจากผู้เขียนที่ 1 และที่อ้างมาจากผู้เขียนที่ 3 และใช้รูปแบบปนกันในข้อความที่อ้างมาจากผู้เขียนที่ 2 จากนั้น ในตัวอย่าง 3.1.3 ที่ผู้เขียนวิทยานิพนธ์เขียนขึ้นมาเองโดยไม่มีส่วนอ้างอิง ผู้เขียนเลือกใช้คำว่า “อินเตอร์เน็ต” อย่างไรก็ตามการพบเพียงหนึ่งคำอาจให้ตัดสินลักษณะการเขียนไม่ได้ผู้วิจัยจึงแสดงตัวอย่างเพิ่มเติม ในตัวอย่างที่ 3.1.4 ที่ผู้เขียนวิทยานิพนธ์คนเดิมเขียนขึ้นมาเองโดยไม่มีส่วนอ้างอิงและพบการเลือกใช้คำ “อินเทอร์เน็ต” เป็นรูปแบบเดียวกันตลอดทั้งย่อหน้า จำนวน 6 ครั้ง ซึ่งตัวอย่างที่แสดงข้างต้น ชี้ให้เห็นว่า แม้ในงานเขียนของบุคคลหนึ่งในชิ้นงานเดียวกันก็พบรูปแบบ (style) ที่ไม่สม่ำเสมอ

3.2 (ปรากฏ, ปรากฎ) ที่ผู้วิจัยเลือกคำนี้มาอธิบายเพราะเป็นคำผิดที่น่าสนใจ ตามที่ สาเหตุหนึ่งของการเขียนคำผิด คือการใช้แนบเทียบผิด เกิดจากการเทียบคำนั้นๆ กับคำอื่นที่ออกเสียงเหมือนกันจึงใช้ตัวสะกดแบบเดียวกัน [25] ทั้งนี้ ปรากฏ เป็นออกเสียงเหมือนกับ กฎ ที่มีตัวสะกดต่างกันแต่มีการใช้ที่แพร่หลายกว่า เช่น เอาไปรวมกับคำอื่นๆ เป็น กฎหมาย กฎกระทรวง เป็นต้น ในขณะที่ กฏ ที่ตัวสะกดต่างกันนั้น พบเพียงคำเดียวคือหลังคำว่า “ปรา” เท่านั้น [26] อีกทั้งรูปผิวของ “ฏ” และ “ฎ” คล้ายกันมาก ประกอบกับพบว่ามีความถี่รวมเป็นอันดับ 2 ในคลังข้อมูล (ตัวอย่างในข้อ 2) เมื่อสุ่มดูในเล่มที่พบมาก คือ เล่มที่ 29005 พบการใช้ลักษณ์ส่วนมาก คือ แบบสะกดผิด 5 ครั้ง แบบสะกดถูก 66 ครั้ง จาก 177 ย่อหน้า อย่างไรก็ตามในย่อหน้าที่ สะกด ผิด จำนวน 5 ครั้ง เป็นย่อหน้าที่ลักลอกเพียง 2 ใน 5 ครั้ง นอกจากนั้นเป็นการเขียนผิดโดยผู้เขียน

ต้นฉบับเอง ทั้งนี้ยังพบส่วนที่น่าสนใจคือ ย่อหน้าที่ลักลอกมาก็มีรูปแบบการเขียนที่ถูกเหมือนกันกับพฤติกรรมโดยมากในการเลือกใช้คำในชุดลักษณ์คำเขียนผิดและเขียนต่างผู้เขียนต้นฉบับเอง

จากตัวอย่างทั้ง 2 คำที่ยกมาข้างต้น ได้ชี้ให้เห็นว่า ลักษณ์ชุดคำเขียนผิดและรูปแบบที่ต่างกันก็มีการใช้ต่างกันได้แม้ในผู้เขียนคนเดียวกัน หรือ แม้ในย่อหน้าที่ลักลอกมาการเลือกใช้คำเขียนผิดอาจใช้เหมือนกันก็ได้แม้เป็นผู้เขียนคนละคนด้วยเหตุนี้ การใช้ลักษณ์ทางภาษากลุ่มนี้กับคลังข้อมูล จึงเป็นเหตุให้แบบจำลองไม่สามารถตรวจจับย่อหน้าที่ลักลอกได้ ตามสมมติฐาน

4. ค่าต่างของชุดคำที่มีความถี่มากที่สุดอันดับต่างๆในคลังข้อมูล โดยผู้วิจัยได้เลือกทดลองกับชุดคำที่มีความถี่มากที่สุด จำนวน 100, 50 และ 15 คำ กลับไม่มีผลกับแบบจำลอง เพื่อวิเคราะห์ว่าแนวคิดเกี่ยวกับชุดคำที่มีความถี่สูงสุด อันแสดงถึงลักษณะรูปแบบคำศัพท์ที่มักใช้ซ้ำๆ ในผู้เขียนคนหนึ่ง ซึ่งเป็นลักษณ์ที่มีประโยชน์ในการระบุตัวผู้เขียนชิ้นงาน ในการทดลองที่ใช้แล้วได้ผลดี เป็นการใช้ข้อมูลที่มีความยาวในระดับหนึ่งเปรียบเทียบกับงานเขียนชิ้นอื่นๆ ในขณะที่งานวิจัยนี้ เป็นการเปรียบเทียบระหว่างย่อหน้า ซึ่งมีขนาดหลากหลายตั้งแต่ 1 ถึงมากกว่า 1,000 คำ และในงานวิจัยบางชิ้นที่เปรียบเทียบค่าเฉลี่ยของประโยคต่อย่อหน้า หรือ ค่าเฉลี่ยคำหน้าที่ต่อประโยค จำนวนข้อมูล 1 ย่อหน้าในงานวิจัยนี้ จึงต่างจากงานอื่นๆ เช่น

4.1.1 ย่อหน้าที่ 11001_196 , 11001_197 และ 11001_198 เป็นย่อหน้าที่ไม่ลักลอก

*^2.1 ความง่วง^*

*^2.1.1 ความหมายของความง่วง^*

*^ความง่วงเป็นความรู้สึกที่เป็นนามธรรมที่บุคคลต่างรู้จักเป็นอย่างดี เพราะเป็นสิ่งที่เกิดขึ้นกับบุคคลอยู่เสมอ แต่เป็นการยากที่จะอธิบายความหมายของความง่วงให้มีลักษณะชัดเจนทางรูปธรรม จึงอธิบายความหมายของความง่วงได้ในลักษณะเชิงพฤติกรรม บุคคลที่อยู่ในสภาพง่วงนอนจะมีแนวโน้มการเคลื่อนไหวน้อยลง การพูดจะช้าและน้อยลงจนหยุดพูด หนังตาจะค่อยๆ ปิดลง การแสดงอารมณ์ทางสีหน้าจะหยุดลง จะมีลักษณะทางพฤติกรรมอื่นที่อาจพบได้ เช่น การหาว การขยี้ตา สัปหงกศีรษะ (สรยุทธ วาสิกนานนท์, 2536)^*

เมื่อใช้การระบุย่อหน้าด้วยการขึ้นบรรทัดใหม่และการแท็บ รูปแบบการเขียนวิทยานิพนธ์ภาษาไทยในชุดตัวอย่าง พบย่อหน้าที่สั้นเป็นจำนวนมาก เช่นในตัวอย่างข้อ 4.1.1 คำที่ปรากฏในย่อหน้าที่ 11001_196 และ 11001_197 เป็นเพียงตัวเลขและนามวลี มีขนาดสั้นเพียง 3 และ 6 คำเท่านั้น ส่วนย่อหน้าที่ 11001_198 มีจำนวนคำ 117 คำ ด้วยย่อหน้าที่สั้นยาวไม่เท่ากันแม้จะเปรียบเทียบกันด้วยค่าเฉลี่ยของย่อหน้า ค่าในลักษณ์นี้จึงแตกต่างกันมาก ที่ -16.8, -15.8 และ 7.25 ตามลำดับ และในเล่มเดียวกันในย่อหน้าที่ลักลอกพบมีค่าลักษณ์ที่ใกล้เคียงกับย่อหน้าที่ไม่ลักลอกด้วย

เช่น 8.29 ด้วยตัวเลขที่ใกล้เคียงกันแม้เป็นย่อหน้าที่ต่างกัน ลักษณ์นี้จึงไม่มีผลกับแบบจำลอง ปัญหานี้อาจแก้ได้หากมีการปรับค่าให้เป็นมาตรฐานโดยเอาจำนวนคำในย่อหน้ามาคำนวณด้วย ซึ่งเป็นประเด็นที่ควรมีการทดลองต่อในงานต่อไป

5.3 ปัญหาที่พบในการวิจัย

ในส่วนนี้ผู้วิจัยจะกล่าวถึงปัญหาที่พบในการทำคลังข้อมูลและข้อสังเกตเกี่ยวกับลักษณ์ทางภาษาที่เลือกใช้ที่ให้ผลที่ไม่ดีเมื่อเทียบกับลักษณ์ที่ไม่ใช้ลักษณ์ทางภาษาเลย

5.3.1 ปัญหาเกี่ยวกับการสร้างคลังข้อมูล

เนื่องจากการวิจัยนี้มุ่งหวังที่จะพัฒนาแบบจำลองที่ใช้เพื่อตรวจเทียบการลักลอกภายในในผลงานวิชาการทางภาษาไทย ผู้วิจัยจึงออกแบบคลังข้อมูลที่ใช้ในการเรียนรู้และทดสอบจากข้อมูลจริงเพื่อให้ใกล้เคียงกับสถานการณ์จริงมากที่สุด ย่อหน้าลักลอกที่นำมาปนกับข้อมูลต้นฉบับ จึงกำหนดให้มาจากข้อมูลสาขาเดียวกัน และมีคำสำคัญร่วมกันในกลุ่ม ด้วยเหตุนี้คลังข้อมูลที่ได้จึงมีความหลากหลายมาก ประกอบด้วย กลุ่มย่อยทั้งหมด 19 กลุ่ม อีกทั้งไม่สามารถควบคุมให้แต่ละกลุ่มย่อยของข้อมูลทั้งหมดมีขนาดเท่ากันได้ นอกจากนี้ การแปลงข้อมูลจาก ไฟล์ประเภท pdf มาเป็นไฟล์อักษร ทำให้สูญเสียขอบเขตของย่อหน้า การขึ้นบรรทัดใหม่ รูปแบบการเลือกใช้ตัวอักษร ตัวเอียงและตัวหนาไปทั้งหมด อีกทั้งการแปลงตัวอักษรมานั้น มีทั้งไฟล์ประเภทที่ใช้ได้ และไฟล์ประเภทที่เสียหายไม่สามารถนำมาใช้ได้ เพื่อสร้างคลังข้อมูล จำนวน 300 เล่มนั้น ผู้วิจัยต้องทดลองแปลงข้อมูลถึง 477 เล่ม ทำให้การสร้างคลังข้อมูลต้องใช้ระยะเวลามาก

ตัวอย่างข้อมูลที่เสียหาย 1- วรรณยุกต์และสระบางส่วนผิดเพี้ยนไป

จะเลื่อนไหลไปมาอยู่ภายใต้ความเป็นคนรักเพศเดียวกันเท่านั้น
และร้อยละ 35 ตามลาดับ

ตัวอย่างข้อมูลที่เสียหาย 2- ตัวอักษรผิดเพี้ยนไปทั้งหมด

ddddddddddddddddddd.

ตัวอย่างข้อมูลที่เสียหาย 3- สระบางส่วนหายไปและมีการเพิ่มช่องว่างในข้อความ

ความเป็ นมาและความสาคัญของปัญหา

ตัวอย่างข้อมูลที่เสียหาย 4- วรรณยุกต์อยู่ผิดตำแหน่ง

ขั นตอนการทานาทุกขั นตอน รวมทั งจะได้เรียนรู้ประเพณีการละเล่นของไทย

นอกจากปัญหาข้อมูลที่มีจำกัดแล้ว ในการสร้างคลังข้อมูลลักลอกเพื่อนำมาใช้ เมื่อไม่มีข้อมูลการลักลอกจริง จึงต้องสร้างข้อมูลขึ้นมาจากวิทยานิพนธ์ที่มีอยู่ อย่างไรก็ตาม ตามธรรมชาติของการเขียนงานทางวิชาการ จะต้องมีส่วนที่อ้างอิงมาจากงานของผู้อื่นอยู่เสมอ ดังที่ผู้วิจัยแสดงให้ดูถึงตัวอย่างย่อหน้าที่มีการอ้างอิงแต่มีคำตอบแบบไม่ลักลอก ซึ่งในย่อหน้านั้น ไม่สามารถตัดสินได้ว่าการเขียนปะปนกัน เป็นเพียงความผิดพลาดของผู้เขียนเอง หรือเป็นเพราะอ้างอิงเอามาโดยตรงจากผลงานของผู้อื่น จึงอาจกล่าวได้ว่า ย่อหน้าที่มีคำตอบเป็นแบบไม่ลักลอก ในคลังข้อมูลอาจเป็นข้อมูลที่ไม่ใช่ลักษณะบุคคลของผู้เขียนคนนั้นๆเลย นอกจากนี้ เรื่องความหลากหลายของหัวเรื่องวิทยานิพนธ์ที่มีทั้งหมด 19 หัวเรื่องย่อย กับชุดคำเขียนผิดที่พบมีการกระจายตัวน้อย ก็มีผลที่สำคัญต่อการพัฒนาแบบจำลอง

5.3.2 ข้อสังเกตถึงลักษณ์ทางภาษาที่เลือกใช้

จากลักษณ์ทางภาษาที่ผู้วิจัยเลือกใช้ข้างต้น เมื่อพิจารณาจากคลังข้อมูลโดยยังไม่ใช้กับแบบจำลอง พบว่าแนวคิดที่ผู้เขียนจะคงลักษณะการเขียนหรือการเลือกใช้ศัพท์แบบเดิมตลอดทั้งเล่มนั้น ในข้อมูลจริงมีความแตกต่างอยู่ เช่น เมื่อพิจารณาถึงรูปแบบลักษณ์ที่หมายถึงปีพุทธศักราช แบบที่มีช่องว่างและไม่มีช่องว่าง มีการใช้แบบปนกันทั้งที่มาจากผู้เขียนคนเดียวกัน

เช่น ตัวอย่างจากไฟล์เลขที่ 11002 บรรทัดที่ 192 และ 1217 มีการใช้ที่ปนกัน แม้จะไม่ใช่ย่อหน้าที่ลักลอก

*<s>ตั้งแต่<w>ปี<w><s><w>**พ.ศ.<w>2545**<w><s><w>ประเทศ<w>ไทย<w>*
*<w>ว่า<w>ใน<w>ปี<w><s><w>**พ.ศ.<w><s><w>2563**<w><s>*

นอกจากนั้น ยังพบการใช้ปนกันในเล่มแม้ไม่ใช่ย่อหน้าที่ลักลอก ในกรณีที่อยู่ตำแหน่งที่จำเป็นต้องมีช่องว่างได้แก่ อยู่ระหว่างตำแหน่งสุดท้ายของบรรทัด เช่น ตัวอย่างจากไฟล์เลขที่ 13015 บรรทัดที่ 241 และ 243 โดยในบรรทัดที่ 241 มีการใช้ช่องว่างหลัง พ.ศ. เมื่อเป็นตำแหน่งสุดท้ายในบรรทัด

*ใน<w>ปี<w><s><w>**พ.ศ.<w>2543<w><**s>*
*<s>ใน<w>ปี<w><s><w>**พ.ศ.<w>2538<w>***

ทั้งนี้ การใช้ปนกันในเล่ม ก็ยังพบในย่อหน้าที่มีการลักลอกจริงด้วย ดังในตัวอย่างไฟล์เลขที่ 17005 บรรทัดที่ 869 และ 1485

*<w>นักเรียน<w><s><w>**พ.ศ.<w><s><w>2543**<w><s>ที่<w>*
*<w><s>พจนานุกรม<w>ฉบับ<w>บัณ<w>ฑิตย<w>สถาน<w><s><w>**พ.ศ.***
***<w>2535**<w><s>กล่าว<w>ว่า*

นอกจากนี้ เมื่อพิจารณาถึงกลุ่มคำที่มีความหมายเดียวกันกลุ่มอื่น คือ บิดา และ พ่อ ก็พบการใช้ปนกันในย่อหน้าเดียวกันแม้ไม่ใช่ส่วนที่ลักลอก เช่น ตัวอย่างจากไฟล์เลขที่ 13006 ย่อหน้าที่ 317 บรรทัดที่ 4 และ บรรทัดที่ 5

*<w>มี<w>ความ<w>เป็น<w>**พ่อ**เลี้ยง<w>หรือ<w>**แม่**เลี้ยง<w>*
*<w>พี่น้อง<w>ต่าง<w>**บิดา**<w>หรือ<w>ต่าง<w>**มารดา**<w>*

จากตัวอย่างข้างต้น จะพบว่าแม้ไม่ใช่ส่วนที่มีการลักลอก ผู้เขียนคนเดียวกันก็เลือกใช้คำหรือรูปแบบการเขียนที่แตกต่างกันได้ ซึ่งในการวิจัยครั้งนี้ ผู้วิจัยได้เลือกใช้การเปรียบเทียบทั้งกับย่อหน้าอื่นๆทั้งเล่มในคราวเดียว(global) หรือเลือกทดลองเป็นช่วง (local) ระยะเฉลี่ย 1 หน้าวิทยานิพนธ์ แต่อย่างไรก็ตามลักษณ์ทางภาษาก็ยังไม่สามารถแยกย่อหน้าทั้ง 2 ประเภทออกจากกันได้

5.3.3 ข้อจำกัดของเครื่องมือในการวิจัย

นอกจากปัญหาที่พบในการสร้างคลังข้อมูลที่กล่าวถึงไปข้างต้น ข้อจำกัดของเครื่องมือในการวิจัยก็เป็นประเด็นใหญ่ประเด็นหนึ่ง สำหรับข้อมูลขนาดใหญ่ การกำกับข้อมูล หรือวิเคราะห์โครงสร้างโดยไม่ใช้โปรแกรมอัตโนมัติเป็นไปได้ยาก ปัจจุบันในภาษาไทย โปรแกรมอัตโนมัติในการกำกับหมวดคำยังไม่มี การทดลองในครั้งนี้แม้จะกำหนดหนึ่งหน่วยทดลองเป็นย่อหน้า แต่หน่วยที่นำมาวิเคราะห์และใช้เป็นลักษณ์นั้นอยู่ในระดับรูปผิวของคำและค่าทางสถิติเท่านั้น

6.3 ข้อเสนอแนะ

เนื่องจากงานวิจัยนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบลักษณ์ที่ใช้เพื่อพัฒนาแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน สำหรับงานเขียนวิชาการภาษาไทย ทดลองในหลายสภาพแวดล้อม เช่น ตามขนาดของต้นฉบับ ตามจำนวนที่ลักลอกมา และตามขนาดย่อหน้าที่ลักลอกมา ซึ่งผู้วิจัย ตั้งใจจะคงต้นฉบับเดิมไว้ให้มากที่สุด และครอบคลุมรูปแบบการเขียนให้มากที่สุด จึงออกแบบข้อมูลไว้ที่ 300 เล่ม อย่างไรก็ตาม กลับพบว่า คลังข้อมูลที่มีขนาดใหญ่ ทำให้การเลือกใช้ลักษณ์ทางภาษาเป็นไปได้ยาก ดังที่พบการเขียนคำต่างกันจริงแต่เมื่อพบได้น้อยเมื่อเทียบกับคลังข้อมูลขนาดใหญ่ จึงมีผลให้แบบจำลองไม่สามารถแยกย่อหน้าทั้ง 2 ประเภทออกจากกันได้ ดังนั้น การสร้างคลังข้อมูลที่มีจำนวนเล่มน้อยลงแต่มีหัวเรื่องเดียวกันทั้งหมด น่าจะส่งผลที่ดีต่อแบบจำลองมากขึ้น

จากผลการทดลองที่พบว่าแบบจำลองไม่สามารถตรวจจับย่อหน้าที่ลักลอกมาขนาดสั้นจำนวน 50-100 คำ ได้เลยนั้น น่าจะเกิดจากข้อมูลวิทยานิพนธ์จริง ที่มีย่อหน้าขนาดสั้นจำนวนมากถึง 50 เปอร์เซ็นต์ เพื่อทำให้คลังข้อมูลเหมือนของเดิมมากที่สุด ผู้วิจัยได้แทรกย่อหน้าลักลอกเข้าไปตามสัดส่วนของย่อหน้าต้นฉบับก่อนการลักลอก ซึ่งเป็นย่อหน้าขนาดสั้นถึงครึ่งหนึ่งจากทั้งหมดของย่อหน้าลักลอก ประกอบกับ ในแบบจำลองมีย่อหน้าที่ลักลอกมาน้อยเพียง 10 เปอร์เซ็นต์ เมื่อลักษณ์

ทางภาษาไม่ใช่ลักษณ์ที่ดีที่สุดการทดลอง ค่าทางสถิติที่มีพื้นฐานมาจากค่าความสั้นยาว และช่องว่างในย่อหน้า จึงเลือกคำตอบให้ย่อหน้าขนาดสั้นเป็นย่อหน้าที่ไม่ลักลอกทั้งหมด จากการทดลองใช้ลักษณ์ทางภาษายังชี้ให้เห็นว่า ย่อหน้าที่สั้นมากๆ ที่พบจริงในคลังข้อมูล เช่น ย่อหน้าที่มี แค่ 1-3 คำ ด้วยจำนวนคำที่น้อย การเลือกใช้ลักษณ์ทางภาษาจึงมีข้อจำกัดไปด้วย ยกตัวอย่างเช่น ถ้านำแนวคิดเกี่ยวกับลักษณ์ทางภาษาที่เป็นอิสระ ไม่ขึ้นอยู่กับหัวเรื่อง หรือประเภทงานเขียน คำจำพวกหน้าที่มักถูกนำมาใช้พิจารณา พอนำไปวิเคราะห์กับข้อความในย่อหน้า เช่น ^ความง่วง^ ย่อหน้าประเภทนี้ก็เป็นไปได้ที่จะไม่มีลักษณ์ทางภาษาประเภทใดที่ใช้ได้เลย คล้ายกับกรณีที่ย่อหน้าที่ใช้ในการทดลองครั้งนี้ มีค่าแทนลักษณ์เป็น 0 โดยรวมประมาณ 90 เปอร์เซ็นต์ และส่งผลต่อการทำนายผลทางเลือกหนึ่งคือ รับเข้าข้อมูลให้มีความยาวพอสมควร อาจเป็นในลักษณะหลายย่อหน้า หรือเป็นช่วงข้อมูล (window) ทั้งนี้ การทดลองด้วยข้อมูลรับเข้าตามแนวคิดนี้ประเภทของลักษณ์ที่ใช้กับแบบจำลองและค่าทางสถิติอื่นๆ ต้องศึกษาเพิ่มเติมและใช้ให้เหมาะสมกับข้อมูลด้วย

ลักษณ์ที่ใช้คำที่มีค่าความถี่สูงสุดเปรียบเทียบกับค่าเฉลี่ยของย่อหน้านั้นๆ ดังที่ผู้วิจัยได้แสดงตัวอย่างไป ในข้อ 4.1.1 ในย่อหน้าที่สั้นจะให้ค่าที่แตกต่างจากค่าเฉลี่ยมาก ผู้วิจัยเห็นว่า การเลือกเปรียบเทียบคำที่มีความถี่สูงกับคำในย่อหน้าปัจจุบัน แล้วให้ค่าลักษณ์เป็นค่าเฉลี่ยจำนวนคำในย่อหน้าที่ไม่พบในรายการคำที่มีความถี่สูงสุด กลับไม่ได้ช่วยเพิ่มประสิทธิภาพของแบบจำลอง ทั้งนี้ การปรับให้เป็นค่ามาตรฐานด้วยการนำจำนวนคำในย่อหน้ามาคำนวนด้วย ก็เป็นประเด็นที่หน้าจะศึกษาต่อไป

ประเด็นที่สำคัญอีกประเด็นหนึ่ง คือการใช้ลักษณ์ทางภาษาซึ่งโดยหลักการควรจะใช้แยกย่อหน้าที่มีการลักลอกได้ เพราะตรวจพบการใช้คำหรือรูปแปรของคำที่ต่างจากส่วนอื่นๆ ภายในเล่มนั้น แต่จากการทดลอง ลักษณ์ทางภาษาไม่สามารถใช้ตรวจจับได้เลย เหตุที่เป็นเช่นนี้ เพราะลักษณ์ที่บ่งความต่างในย่อหน้าลักลอกนั้นไม่ได้ปรากฏในย่อหน้าลักลอกอย่างสม่ำเสมอ เมื่อนำมาใช้กับแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ในที่นี้ แบบจำลองจึงไม่เห็นว่าลักษณ์นั้นใช้บ่งชี้ความแตกต่างได้อย่างไรเพราะในย่อหน้าอื่นที่ลักลอกไม่ได้ปรากฏการใช้คำที่เบี่ยงเบนไปนี้ด้วย หากจะใช้ประโยชน์จากลักษณ์ทางภาษาในลักษณะนี้ อาจจะต้องใช้แบบจำลองอื่นแทน

รายการอ้างอิง

ภาษาไทย

1. สำนักงานคณะกรรมการวิจัยแห่งชาติ, ประกาศสำนักงานคณะกรรมการวิจัยแห่งชาติ เรื่อง จรรยาบรรณนักวิจัย, ใน การประชุมครั้งที่ 3/2541. 2541.
2. มานิตย์ จุมปา, เขียนผลงานวิชาการอย่างไร ไม่ละเมิดลิขสิทธิ์และไม่ลักลอกผลงาน (*Plagiarism)* Vol. พิมพ์ครั้งที่ 2. 2556, กรุงเทพมหานคร: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.
6. สุพจน์ จันทรวิพัฒน์, การแก้ปัญหาการจัดกลุ่มข้อมูลด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีนและตัวอย่างการประยุกต์ใช้งาน. วารสารวิชาการเทคโนโลยีอุตสาหกรรม 2554. ปีที่ **7**(ฉบับที่ 1 มกราคม - มิถุนายน): p. 50-55.
7. สุรเดช บัญลือ, จักรกริช เคล้าละม่อม, and แสงนภา วันเพ็ง, การทำนายโรคพาร์กินสันโดยใช้ซัพพอร์ทเวกเตอร์แมชชีน CIT2011 & UniNOMS 2011: p. 7-12.
23. เดช ธรรมศิริ และ พยุง มีสัจ, การจำแนกข้อมูลด้วยวิธีแบบร่วมกันตัดสินใจจากพื้นฐานของเทคนิคต้นไม้ตัดสินใจ เทคนิคโครงข่ายประสาทเทียม และเทคนิคซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการเลือกตัวแทนที่เหมาะสมด้วยขั้นตอนเชิงพันธุกรรม. วารสารวิชาการพระจอมเกล้าพระนครเหนือ, 2554. **21**(2 พฤษภาคม - สิงหาคม): p. 293-303.
25. สนม ครุฑเมือง, คำที่มักเขียนผิดในภาษาไทย : การวิเคราะห์จากการตรวจผลงานการสอน. วารสารการจัดการ คณะวิทยาศาสตร์การจัดการ มหาวิทยาลัยราชภัฏลำปาง 2557. ปีที่ **7**(ฉบับที่ 1 มกราคม -มิถุนายน): p. 42-55.
26. จำนงค์ ทองประเสริฐ, ภาษาไทยไขขาน. 2528, กรุงเทพมหานคร: สำนักพิมพ์แพร่พิทยา.

ภาษาอังกฤษ

3. Stamatatos, E., *Intrinsic Plagiasirm detection using character n-gram profile*, in *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*. 2009.
4. Stamatatos, E., *A survey of modern authorship attribution methods.* J. Am. Soc. Inf. Sci. Technol., 2009. **60**(3): p. 538-556.
5. Seaward, L. and S. Matwin, *Intrinsic Plagiarism Detection Using Complexity Analysis*, in *SEPLN 2009 Workshop on Uncoverring Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, B.R. Stein, P.,Stamatatos, E., Koppel, M., Agirre, E. (eds.), Editor. 2009. p. 56-61.
8. Maurer, H., F. Kappe, and B. Zaka, *Plagiarism - A Survey.* Journal of Universal Computer Science, 2006. **12**(8 (2006)): p. 1050-1084.
9. Hexham, I., *Academic Plagiarism Defined.* 2005.

10. Park, C., *In other (people's) words : plagiarism by university students literature and lesson.* ssessment & Evaluation in Higher Education, 2003. **28**(5): p. 471-488.
11. Chong, M., L. Specia, and R. Mitkov. *Using Natural Language Processing for Automatic Dectection of Plagiarism*. in *the 4th International Plagiarism Conference (IPC-2010)*. 2010. Newcastle-upon-Tyne, UK.
12. HaCohen-Kerner, Y., A. Tayeb, and N. Ben-Dror, *Detection of Simple Plagiarism in Computer Science Papers*, in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010, Association for Computational Linguistics: Beijing, China. p. 421-429.
13. Stein, B., N. Lipka, and P. Prettenhofer, *Intrinsic Plagiarism Analysis.* Language Resources and Evaluation (LRE), 2011. **45**(1): p. 63-82.
14. Zu Eissen, S.M. and B. Stein, *Intrinsic plagiarism detection*, in *Proceedings of the 28th European conference on Advances in Information Retrieval*. 2006, Springer-Verlag: London, UK. p. 565-569.
15. Zu Eissen, M.S., B. Stein, and M. Kuling, *Plagiarism Detection Without Reference Collections* in *Advances in Data Analysis*, E. R. Decker and H. J. Lenz, Editor. 2007, Springer Berlin Heidelberg. p. 359-366.
16. Lakshmi and P. Kumar Pateriya, *A Study On Author Identification through Stylometry.* International Journal on Computer Science & Communication Networks, 2013. **2**(6): p. 653-657.
17. Halvani, O. *Register & Genre Seminar : Toward Intrinsic Plagiarism Detection*. 2010.
18. Biber, D., *Variation across speech and writing*. 1988: Cambridge University Press.
19. Li, J., R. Zheng, and H. Chen, *From fingerprint to writeprint.* Commun. ACM, 2006. **49**(4): p. 76-82.
20. Bindu, M.S. and S.M. Idicula, *Named Entity Recognizer Employing Multicalss Support Vector Machines for the Development of Question Answering Systems.* International of Computer Applications 2011. **25**(10): p. 40-46.
21. Mulloni, A., *Automatic prediction of cognate orthography using support vector machines*, in *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*. 2007, Association for Computational Linguistics: Prague, Czech Republic. p. 25-30.

22. Potthast, M., et al. *An evaluation framework for plagiarism Detection*. in *In Proceeding of the 23rd International Conference Computational Linguistics COLING 2010.* 2010. Beijing, China.

24. Argamon, S. and S. Levitan. *Measuring the Usefulness of Function Words for Authorship Attribution*. in *Proceedings of ACH/ALLC Conference 2005*. 2005. University of Victoria, Canada.

33. Zheng, R., et al., *A framework for Authorship Identification of Online Message. Writing Style Feature and Calssification technique.* Journal of America Society of Information Science Technology, 2006. **57**(3 February): p. 378-393.

Bibliography

27. Juola, P., *Authorship attribution.* Found. Trends Inf. Retr., 2006. **1**(3): p. 233-334.

28. Clough, P. and M. Stevenson *Creating a corpus of plagiarism academic texts.* 2009.

29. McCabe, D.L., *Cheating among College and University Students : A North American Perspective.* International Journal of Educational Integrity, 2005. **1**(1): p. 1-11.

30. Kock, N., *A case of academic plagiarism.* Commun. ACM, 1999. **42**(7): p. 96-104.

31. Oberreuter, G. and J.D. Vel, *Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style.* Expert Syst. Appl., 2013. **40**(9): p. 3756-3763.

32. Scuse, D. and P. Reutemann *WEKA Experimenter Turorial for Version 3-5-5*. 2007.

ภาคผนวก

ภาคผนวก ก ตัวอย่างลักษณ์ทางสถิติจำนวน 7 ลักษณ์กับแบบจำลองทั้งสองประเภท

1. ตัวอย่างลักษณ์จากแบบจำลองรับเข้าแบบคำ (ไฟล์ชนิด ARFF)

@attribute 'Average Word/P' numeric
@attribute DiffW/P numeric
@attribute S/P numeric
@attribute 'Diff S/P' numeric
@attribute S/W numeric
@attribute DiffS/W numeric
@attribute RatioW numeric
@attribute Class {no,yes}

@data

58.03,-24.97,13.55,-4.45,4.28,0.33,0.33,no
57.84,-108.16,13.51,-24.49,4.28,0.09,0.65,no
57.92,-74.08,13.51,-25.49,4.29,-0.91,0.52,no
57.91,-79.09,13.49,-31.51,4.29,-1.25,0.54,no
58.18,38.18,13.59,12.59,4.28,15.72,0.08,no
58.08,-2.92,13.56,-3.44,4.28,-0.69,0.24,no
57.84,-108.16,13.54,-11.46,4.27,2.37,0.65,no
57.9,-83.1,13.49,-34.51,4.29,-1.35,0.55,no
58.03,-28.97,13.56,-2.44,4.28,1.16,0.34,yes
57.99,-46.01,13.52,-17.48,4.29,-0.94,0.41,no
58.04,-22.96,13.57,2.57,4.28,3.08,0.32,no
57.89,-89.11,13.49,-34.51,4.29,-1.23,0.58,no
58.09,-0.91,13.56,-0.44,4.28,-0.07,0.23,no
57.82,-119.18,13.41,-65.59,4.31,-2.07,0.7,no
58.04,-21.96,13.55,-7.45,4.28,-0.47,0.31,yes
58.2,47.2,13.59,12.59,4.28,6.72,0.04,no
58.13,18.13,13.57,3.57,4.28,-0.28,0.16,no
58.2,48.2,13.59,12.59,4.28,5.72,0.04,no

รายละเอียดลักษณ์ที่ใช้ในการทดลอง

Average Word/P = ค่าเฉลี่ยจำนวนคำต่อย่อหน้าในเล่ม
DiffW/P = จำนวนคำต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม
S/P = ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่ม
Diff S/P = ค่าเฉลี่ยช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่ม
S/W = ค่าเฉลี่ยจำนวนคำต่อช่องว่างในเล่ม
DiffS/W = ค่าเฉลี่ยจำนวนคำต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม
RatioW = สัดส่วนของคำในย่อหน้าปัจจุบันต่อทั้งเล่ม
Class = คำตอบ

2. ตัวอย่างลักษณ์จากแบบจำลองรับเข้าแบบตัวอักษร (ไฟล์ชนิด ARFF)

@attribute 'Average C/P' numeric
@attribute Diffc/P numeric
@attribute S/P numeric
@attribute 'Diff S/P' numeric
@attribute S/C numeric
@attribute DiffS/C numeric
@attribute RatioC numeric
@attribute Class {no,yes}

@data

235.81,-91.19,13.55,-4.45,17.4,0.77,0.32,no
235.01,-443.99,13.51,-24.49,17.4,0.47,0.66,no
235.38,-283.62,13.51,-25.49,17.43,-4.12,0.5,no
235.29,-320.71,13.49,-31.51,17.44,-5.08,0.54,no
236.36,148.36,13.59,12.59,17.39,70.61,0.09,no
236.07,19.07,13.56,-3.44,17.41,-4.65,0.21,no
235.13,-392.87,13.54,-11.46,17.37,7.75,0.61,no
235.3,-315.7,13.49,-34.51,17.45,-5.97,0.53,no
235.78,-108.22,13.56,-2.44,17.39,4.11,0.33,yes
235.59,-191.41,13.52,-17.48,17.42,-3.65,0.41,no
235.83,-86.17,13.57,2.57,17.38,11.89,0.31,no
235.33,-301.67,13.49,-34.51,17.45,-6.26,0.52,no
236,-9,13.56,-0.44,17.4,0.1,0.24,no
235.07,-415.93,13.41,-65.59,17.52,-9.28,0.63,no
232.25,-1650.75,13.49,-31.51,17.21,24.63,1.82,no
235.79,-103.21,13.55,-7.45,17.41,-1.27,0.33,yes
236.46,193.46,13.59,12.59,17.4,25.6,0.04,no
236.19,72.19,13.57,3.57,17.4,-1,0.16,no

รายละเอียดลักษณ์ที่ใช้ในการทดลอง

Average Word/C = ค่าเฉลี่ยจำนวนตัวอักษรต่อย่อหน้าในเล่ม
DiffW/C = จำนวนตัวอักษรต่อย่อหน้าที่ต่างจากค่าเฉลี่ยในเล่ม
S/P = ค่าเฉลี่ยจำนวนช่องว่างต่อย่อหน้าในเล่ม
Diff S/P = ค่าเฉลี่ยช่องว่างต่อย่อหน้าต่างจากค่าเฉลี่ยในเล่ม
S/C = ค่าเฉลี่ยจำนวนตัวอักษรต่อช่องว่างในเล่ม
DiffS/C = ค่าเฉลี่ยจำนวนตัวอักษรต่อช่องว่างที่ต่างจากค่าเฉลี่ยในเล่ม
RatioC = สัดส่วนของตัวอักษรในย่อหน้าปัจจุบันต่อทั้งเล่ม
Class = คำตอบ

3. ตัวอย่างลักษณ์ทางภาษาชุดคำเขียนผิดและการเลือกใช้คำและรูปแบบ (ไฟล์ชนิด ARFF)

@attribute Internet numeric
@attribute Head numeric
@attribute Game numeric
@attribute Imm numeric
@attribute Era<.> numeric
@attribute Percent numeric
@attribute Appear numeric
@attribute Rule numeric
@attribute Era<s> numeric
@attribute Dad numeric
@attribute Mom numeric
@attribute 'A hundred' {no,yes}
@attribute Class {no,yes}

@data

0,0,0,0,0,0,0,0,0,0,0,no,no
0,1,0,0,0,0,0,0,0,0,0,no,no
0,0,0,0,0,0,0,0,0,0,0,yes,no
0,0,0,0,0,1,0,0,0,0,0,no,no
0,0,0,0,0,0,0,0,0,0,0,no,no
0,0,0,0,0,0,0,0,0,0,0,no,no
0,0,0,0,0,0,0,0,0,0,0,no,no
0,0,0,0,0,0,0,0,0,0,0,no,no
0,0,0,0,0,0,0,0,0,0,0,no,yes
0,0,0,0,0,0,0,0,0,0,0,no,no
0,0,0,0,0,0,0,0,0,0,0,no,no
0,0,0,0,0,0,0,0,0,0,0,yes,no

รายละเอียดลักษณ์ที่ใช้ในการทดลอง

Internet = (อินเทอร์เน็ต, อินเตอร์เน็ต)
Game = (เกม, เกมส์)
Era<.> = (พ.ศ., พศ.)
Appear = (ปรากฏ, ปรากฏ)
Era<s> = (พ.ศ.<+s>, พ.ศ.<-s>)
Mom = (แม่, มารดา)
Class = คำตอบ

Head = (ศีรษะ, ศรีษะ)
Imm = (กะทันหัน, กระทันหัน)
Percent = (เปอร์เซ็นต์, %)
Rule = (กฎ, กฏ)
Dad = (พ่อ, บิดา)
A hundred = (ร้อยละ, เปอร์เซ็นต์, %)

4. ตัวอย่างลักษณ์ทางภาษาชุดค่าเฉลี่ยคำความถี่สูงสุด 50 คำแรก (ไฟล์ชนิด ARFF)

@attribute 'Top 50' numeric
@attribute Class {no,yes}

@data
9.37,no
51.47,no
45.45,no
35.43,no
-13.68,no
3.36,no
47.46,no
59.49,no
11.38,yes
27.42,no
13.38,no
33.43,no
2.35,no
47.46,no
168.73,no
2.35,yes
-20.7,no
-5.66,no
-20.7,no
-5.66,no
-26.71,no
-14.68,no
-9.67,no
-10.68,no
-25.71,no
-18.69,no
-25.71,no
-4.66,no
4.36,no

รายละเอียดลักษณ์ที่ใช้ในการทดลอง

Top 50 = ค่าต่างจากค่าเฉลี่ยคำที่มีความถี่สูงสุด 50 คำแรกในเล่ม

ภาคผนวก ข ผลการทำนายจากแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน จากโปรแกรม WEKA

1. ผลการทำนายจากแบบจำลองที่รับเข้าแบบคำลักษณ์ทางสถิติ 7 ลักษณ์ (การทดลอง 4.2.4)

=== Run information ===

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 1 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: stat_training_char-weka.filters.unsupervised.attribute.Remove-R1
Instances: 59777
Attributes: 8
Average Word/P
DiffW/P
S/P
Diff S/P
S/W
DiffS/W
RatioW
Class
Test mode: user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)
Time taken to build model: 24875.39 seconds

=== Predictions on test set ===

inst#,actual,predicted,error,prediction
1,1:no,1:no,,1
2,2:yes,1:no,+,1
3,1:no,1:no,,1
4,2:yes,1:no,+,1
5,1:no,2:yes,+,1
6,1:no,2:yes,+,1
7,2:yes,1:no,+,1
8,1:no,2:yes,+,1
9,1:no,2:yes,+,1
10,1:no,1:no,,1
11,2:yes,1:no,+,1
12,1:no,2:yes,+,1
13,2:yes,2:yes,,1

14,1:no,2:yes,+,1
15,1:no,1:no,,1
16,1:no,1:no,,1
17,1:no,1:no,,1
18,2:yes,1:no,+,1
19,1:no,1:no,,1
20,1:no,1:no,,1
8323,1:no,2:yes,+,1
8324,1:no,2:yes,+,1
8325,1:no,2:yes,+,1
8326,1:no,2:yes,+,1
8327,1:no,2:yes,+,1
8328,1:no,1:no,,1
8329,1:no,2:yes,+,1
8330,1:no,1:no,,1
8331,1:no,2:yes,+,1
8332,1:no,1:no,,1
8334,1:no,1:no,,1
8335,1:no,1:no,,1
8336,1:no,1:no,,1

=== Evaluation on test set ===

Time taken to test model on supplied test set: 30.42 seconds

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.973	0.946	0.914	0.973	0.942	0.045	0.513	0.914	no
	0.054	0.027	0.161	0.054	0.081	0.045	0.513	0.092	yes
Weighted Avg.	0.892	0.865	0.848	0.892	0.866	0.045	0.513	0.842	

=== Confusion Matrix ===

a	b	<-- classified as
7392	209	a = no
695	40	b = yes

2. ผลการทำนายจากแบบจำลองที่รับเข้าแบบตัวอักษรลักษณ์ทางสถิติ 7 ลักษณ์ (การทดลอง 4.2.4)

=== Run information ===

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 1 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: stat_training_char-weka.filters.unsupervised.attribute.Remove-R1
Instances: 59777
Attributes: 8
Average C/P
DiffC/P
S/P
Diff S/P
S/C
DiffS/C
RatioC
Class
Test mode: user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)
Time taken to build model: 24875.39 seconds

=== Predictions on test set ===

inst#,actual,predicted,error,prediction
1,1:no,1:no,,1
2,2:yes,1:no,+,1
3,1:no,1:no,,1
4,2:yes,1:no,+,1
5,1:no,1:no,,1
6,1:no,1:no,,1
7,2:yes,1:no,+,1
8,1:no,1:no,,1
9,1:no,1:no,,1
10,1:no,1:no,,1
11,2:yes,1:no,+,1
12,1:no,1:no,,1

13,2:yes,1:no,+,1
14,1:no,1:no,,1
15,1:no,1:no,,1
16,1:no,1:no,,1
17,1:no,1:no,,1
18,2:yes,1:no,+,1
19,1:no,1:no,,1
20,1:no,1:no,,1
8323,1:no,2:yes,+,1
8324,1:no,1:no,,1
8325,1:no,1:no,,1
8326,1:no,1:no,,1
8327,1:no,2:yes,+,1
8328,1:no,1:no,,1
8329,1:no,1:no,,1
8330,1:no,1:no,,1
8331,1:no,1:no,,1
8332,1:no,1:no,,1
8333,1:no,1:no,,1
8334,1:no,1:no,,1
8335,1:no,1:no,,1
8336,1:no,1:no,,1

=== Evaluation on test set ===

Time taken to test model on supplied test set: 30.42 seconds

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.973	0.946	0.914	0.973	0.942	0.045	0.513	0.914	no
	0.054	0.027	0.161	0.054	0.081	0.045	0.513	0.092	yes
Weighted Avg.	0.892	0.865	0.848	0.892	0.866	0.045	0.513	0.842	

=== Confusion Matrix ===

a	b	<-- classified as
7392	209	a = no
695	40	b = yes

3. ผลการทำนายจากแบบจำลองที่รับเข้าแบบตัวอักษรลักษณ์ทางสถิติ 7 ลักษณ์ (การทดลอง 4.2.4)

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 1 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model "C:\\Program Files\\Weka-3-7" -seed 1
Relation: train_variation-weka.filters.unsupervised.attribute.Remove-R1
Instances: 59777
Attributes: 13
Internet
Head
Game
Imm
Era<.>
Percent
Appear
Rule
Era<s>
Dad
Mom
A hundred
Class
Test mode: user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)
Time taken to build model: 55.08 seconds

=== Predictions on test set ===

inst#,actual,predicted,error,prediction
1,1:no,1:no,,1
2,2:yes,1:no,+,1
3,1:no,1:no,,1
4,2:yes,1:no,+,1
5,1:no,1:no,,1
6,1:no,1:no,,1
7,2:yes,1:no,+,1
8,1:no,1:no,,1
9,1:no,1:no,,1

10,1:no,1:no,,1
11,2:yes,1:no,+,1
12,1:no,1:no,,1
13,2:yes,1:no,+,1
14,1:no,1:no,,1
15,1:no,1:no,,1
16,1:no,1:no,,1
17,1:no,1:no,,1
18,2:yes,1:no,+,1
19,1:no,1:no,,1
20,1:no,1:no,,1
8323,1:no,1:no,,1
8324,1:no,1:no,,1
8325,1:no,1:no,,1
8326,1:no,1:no,,1
8327,1:no,1:no,,1
8328,1:no,1:no,,1
8329,1:no,1:no,,1
8330,1:no,1:no,,1
8331,1:no,1:no,,1
8332,1:no,1:no,,1
8333,1:no,1:no,,1
8334,1:no,1:no,,1
8335,1:no,1:no,,1
8336,1:no,1:no,,1

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.912	1.000	0.954	0.000	0.500	0.912	no
	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.088	yes
Weighted Avg.	0.912	0.912	0.831	0.912	0.870	0.000	0.500	0.839	

=== Confusion Matrix ===

a	b	<-- classified as
7601	0	a = no
735	0	b = yes

ประวัติผู้เขียนวิทยานิพนธ์

นางสาว ศิวพร ทวนไธสง เกิดที่จังหวัดจันทบุรี สำเร็จการศึกษาระดับปริญญาตรี จากคณะมนุษยศาสตร์และสังคมศาสตร์ สาขาภาษาอังกฤษ เกียรตินิยม อันดับ 2 มหาวิทยาลัยนเรศวร จังหวัดพิษณุโลก ในปีการศึกษา 2545 และเข้าศึกษาต่อในหลักสูตรอักษรศาสตร์มหาบัณฑิต ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์ลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2553