การประยุกต์ใช้การทำเหมืองข้อความเพื่อจำแนกประเภทโรคจากอาการ

น.ส.พรรณาภรณ์ เกตุภู่พงษ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

การประยุกต์ใช้การทำเหมืองข้อความเพื่อจำแนกประเภทโรคจากอาการ

น.ส.พรรณาภรณ์ เกตุภู่พงษ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

Applying text mining for classifying disease from symptoms

Miss Pannaporn Ketpupong

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Computer Science

Department of Computer Engineering

FACULTY OF ENGINEERING

Chulalongkorn University

Academic Year 2018

หัวข้อวิทยานิพนธ์	การประยุกต์ใช้การทำเหมืองข้อความเพื่อจำแนกประเภทโรคจากอาการ
โดย	น.ส.พรรณาภรณ์ เกตุภู่พงษ์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.เกริก ภิรมย์โสภา
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	อาจารย์ นายแพทย์กฤษณ์ เจริญลาภ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ณัฐวุฒิ หนูไพโรจน์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.เกริก ภิรมย์โสภา)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(อาจารย์ นายแพทย์กฤษณ์ เจริญลาภ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.พีรพล เวทีกูล)

..... กรรมการภายนอกมหาวิทยาลัย
(นายแพทย์วิทวัส เจนบุญไทย)

พรรณาภรณ์ เกตุภู่พงษ์ : การประยุกต์ใช้การทำเหมืองข้อความเพื่อจำแนกประเภทโรคจากอาการ. (Applying text mining for classifying disease from symptoms) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ดร.เกริก ภิรมย์โสภา, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : อ. นพ. กฤษณ์ เจริญลาภ

การวินิจฉัยโรคที่คลาดเคลื่อนถือเป็นปัญหาสำคัญในวงการแพทย์ โดยปัจจุบันการวินิจฉัยโรคของแพทย์แต่ละคนจะแตกต่างกันไปตามความรู้ ความชำนาญ และประสบการณ์ที่ได้สั่งสมมา รวมทั้งการวินิจฉัยโรคในบางครั้งแพทย์อาจลืมนึกถึงโรคบางโรคไป เนื่องจากเป็นโรคที่พบเจอได้ยากหรือไม่ค่อยพบเจอในผู้ป่วย ส่งผลให้การวินิจฉัยโรคเกิดความคลาดเคลื่อน โดยหลังจากที่แพทย์ได้ทำการวินิจฉัยโรคแล้ว ขั้นตอนต่อมาคือการจำแนกรหัสไอซีดีเทนซีเอ็มให้กับคำวินิจฉัยนั้น ซึ่งถือเป็นขั้นตอนที่ยุ่งยากสำหรับแพทย์ส่วนใหญ่ ดังนั้นในงานวิจัยนี้จึงมีแนวคิดที่จะนำเสนอแบบจำลองสำหรับจำแนกประเภทโรคจากอาการ โดยการประยุกต์ใช้การทำเหมืองข้อความ เพื่อช่วยแพทย์ในการวินิจฉัยโรคและจำแนกรหัสไอซีดีเทนซีเอ็มได้ด้วยข้อมูลอาการของผู้ป่วย ซึ่งการสร้างแบบจำลองในงานวิจัยนี้จะเลือกใช้ตัวจำแนกประเภทที่นิยมใช้ในการทำเหมืองข้อความ ได้แก่ ต้นไม้ตัดสินใจ การเรียนรู้เบส์อย่างง่าย ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม มาเปรียบเทียบกันโดยใช้ระยะเวลาที่ใช้ในการสร้างแบบจำลอง ระยะเวลาที่แบบจำลองใช้ในการทำนาย กราฟเส้นโค้งอาร์โอซี อัตราผลบวกจริง อัตราผลบวกเท็จ ค่าความเที่ยง และค่าความแม่นมาเป็นตัวชี้วัด ซึ่งผลลัพธ์ที่ได้พบว่าการใช้โครงข่ายประสาทเทียมเป็นตัวจำแนกประเภทในการสร้างแบบจำลองมีความเหมาะสมที่สุดสำหรับงานวิจัยนี้ เนื่องจากให้อัตราผลบวกจริงสูงสุดที่ร้อยละ 89.03 และมีพื้นที่ใต้เส้นโค้งของกราฟเส้นโค้งอาร์โอซีมากที่สุด

ภาควิชา	ภาควิชาวิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต .....
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....
ปีการศึกษา	2561	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

5970261121 : MASTER OF SCIENCE
Classification, Data preparation, Disease, ICD-10-CM, Symptom, Text mining

Pannaporn Ketpupong : Applying text mining for classifying disease from symptoms. ADVISOR: Asst. Prof. Krerk Piromsopa, Ph.D., CHRIS CHAROENLAP, M.D.

Nowadays, misdiagnoses account for a significant portion of medical errors. This is due to the fact that each physician’s diagnosis is different depending on the physician’s knowledge, skill, and experience. In several cases, physicians may ignore uncommon diseases. Also, after the diagnosis, the physician has to provide ICD-10-CM code. This is a difficult process for most (if not all) physicians. We propose a predictive model for classifying disease from symptoms by applying text mining technique. Our research technique allows physician to diagnose and to access an ICD-10-CM code directly from symptoms. Our models are based on several classifiers such as Decision Tree, Naïve Bayes, Support Vector Machine, and Neural Network. Models from each classifier were compared using training time, predicting time, Receiver Operating Characteristic (ROC) curve, True Positive Rate (TPR), False Positive Rate (FPR), precision and accuracy. The result suggests that Neural Network gives the best TPR at 89.03%.

Department:	Department of Computer Engineering	Student's Signature .....
Field of Study:	Computer Science	Advisor's Signature .....
Academic Year:	2018	Co-advisor's Signature .....

กิตติกรรมประกาศ

ขอกราบขอบพระคุณ ผศ. ดร. เกริก ภิรมย์โสภา อาจารย์ที่ปรึกษาวิทยานิพนธ์ และ อ. นพ. กฤษณ์ เจริญลาภ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม เป็นอย่างยิ่งที่ได้เสียสละเวลาในการให้คำปรึกษา คำแนะนำ และแนวทางในการดำเนินงาน ทั้งยังให้ความช่วยเหลืออย่างเต็มกำลังเมื่อเกิดปัญหาในการดำเนินงาน ทำให้การจัดทำวิทยานิพนธ์เสร็จสมบูรณ์ด้วยดี

ขอกราบขอบพระคุณ ผศ. ดร. ณัฐวุฒิ หนูไพโรจน์ ผศ. ดร. พีรพล เวทีกุล และ นพ. วิทวัส เจนบุญไทย คณะกรรมการสอบวิทยานิพนธ์เป็นอย่างยิ่ง ที่ได้กรุณาให้คำแนะนำสำหรับนำไปปรับปรุงแนวทางในการดำเนินงาน เพื่อให้งานเป็นไปอย่างราบรื่น

ขอขอบพระคุณ คณาจารย์ทุกท่านในภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ให้ความรู้ตลอดหลักสูตร

ขอขอบคุณ โรงพยาบาลจุฬาลงกรณ์ ที่ให้ความอนุเคราะห์ข้อมูลคำวินิจฉัย เพื่อนำมาใช้ในวิทยานิพนธ์นี้

สุดท้ายขอขอบพระคุณ บิดา มารดา และขอขอบคุณเพื่อน ๆ พี่ ๆ ทุกท่านที่คอยช่วยเหลือและให้การสนับสนุนมาโดยตลอด

พรรณาภรณ์ เกตุภู่พงษ์

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง ..... ญ

สารบัญรูป ..... ฎ

บทที่ 1 1

บทนำ 1

1.1 ที่มาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของงานวิจัย 3

1.3 ขอบเขตงานวิจัย 4

1.4 ขั้นตอนและวิธีการดำเนินงานวิจัย 4

1.5 ประโยชน์ที่คาดว่าจะได้รับ 4

1.6 โครงสร้างของเนื้อหาวิทยานิพนธ์ 5

บทที่ 2 6

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 6

2.1 ทฤษฎีที่เกี่ยวข้อง 6

2.1.1 การวินิจฉัยโรค 6

2.1.2 รหัสไอซีดีเทนซีเอ็ม 7

2.1.3 การทำเหมืองข้อมูล 8

2.1.4 การจำแนกประเภท 9

2.1.5 ต้นไม้ตัดสินใจ 10
2.1.6 การเรียนรู้เบส์อย่างง่าย 10
2.1.7 ซัพพอร์ตเวกเตอร์แมชชีน 11
2.1.8 โครงข่ายประสาทเทียม 12
2.1.9 ตัววัดความแม่นยำของแบบจำลองการจำแนกประเภทข้อมูล 13
2.1.10 การเรียนรู้ของเครื่อง 14
2.2 งานวิจัยที่เกี่ยวข้อง 15
2.2.1 แนวคิดในการประยุกต์ใช้การทำเหมืองข้อความ 15
2.2.1.1 การทำเหมืองข้อความกับข้อมูลชีวการแพทย์ 15
2.2.1.2 การทำเหมืองข้อความเพื่อช่วยลดงานทางการแพทย์ 16
2.2.1.3 การทำเหมืองข้อความกับข้อมูลด้านพันธุศาสตร์ 17
2.2.1.4 การทำเหมืองข้อความกับงานด้านอื่น ๆ 18
2.2.2 การเตรียมพร้อมข้อมูลก่อนการทำเหมืองข้อความ 18
2.2.3 ตัวจำแนกประเภทที่นิยมใช้ในการทำเหมืองข้อความ 20
2.2.4 การวัดประสิทธิภาพแบบจำลองที่ได้จากการทำเหมืองข้อความ 20
2.2.5 สรุปผลของงานวิจัยที่เกี่ยวข้อง 21
2.2.6 สรุป 23
บทที่ 3 24
แนวคิดและวิธีการดำเนินงาน 24
3.1 การเก็บข้อมูล 25
3.2 การสร้างแบบจำลอง 36
3.2.1 การประมวลผลข้อมูลก่อน 36
3.2.2 การจำแนกประเภท 42
3.3 การใช้งานแบบจำลอง 52

3.4 เครื่องมือที่ใช้ในการพัฒนาแบบจำลอง 56
บทที่ 4 57
การทดสอบเครื่องมือ และการอภิปราย 57
4.1 การประเมินประสิทธิภาพของอัลกอริทึมที่ใช้ในการสร้างแบบจำลอง 57
4.2 ผลการประเมินประสิทธิภาพของอัลกอริทึมที่ใช้ในการสร้างแบบจำลอง 61
4.3 ตัวอย่างผลลัพธ์ที่ได้จากการทดสอบแบบจำลอง 63
บทที่ 5 71
บทสรุป 71
5.1 สรุปผลวิทยานิพนธ์ 71
5.2 ปัญหาและข้อจำกัดในการทำวิทยานิพนธ์ 71
5.3 แนวทางในการปรับปรุงวิทยานิพนธ์ 71
บรรณานุกรม 72
ประวัติผู้เขียน 113

สารบัญตาราง

หน้า

ตารางที่ 1 ตารางเปรียบเทียบผลของงานวิจัยที่เกี่ยวข้อง 21

ตารางที่ 2 ตารางแสดงจำนวนโรคในแต่ละหมวดหมู่ 28

ตารางที่ 3 ตารางแสดงร้อยละของจำนวนข้อมูลจากเวชระเบียนแยกตามหมวดหมู่ 29

ตารางที่ 4 ตารางแสดงร้อยละของจำนวนข้อมูลจากเว็บไซต์สาธารณะแยกตามหมวดหมู่ 30

ตารางที่ 5 ตารางแสดงผลการตรวจสอบคำสำคัญที่เกี่ยวข้องกับโรค 33

ตารางที่ 6 ตารางแสดงอัลกอริทึมของตัวจำแนกประเภทชนิดต่าง ๆ 47

ตารางที่ 7 ตารางแสดงวิธีการวิเคราะห์หาคำสำคัญของตัวจำแนกประเภทแต่ละชนิด 52

ตารางที่ 8 ตารางแสดงคำอธิบายยูสเคสของฟังก์ชันเรียกดูรายชื่อโรค 53

ตารางที่ 9 ตารางแสดงคำอธิบายยูสเคสของฟังก์ชันบันทึกชื่อโรคที่ถูกต้อง 54

ตารางที่ 10 ตารางแสดงคอนฟิวชันเมทริกซ์ขนาด 2 x 2 57

ตารางที่ 11 ตารางแสดงผลการประเมินประสิทธิภาพของอัลกอริทึมที่ใช้ในการสร้างแบบจำลอง ... 62

ตารางที่ 12 ตารางแสดงตัวอย่างบันทึกของแพทย์ที่นำมาใช้ในการทดสอบและผลลัพธ์ที่ได้จากแบบจำลอง 63

สารบัญรูป

หน้า

รูปที่ 1 ขั้นตอนการจำแนกรหัสไอซีดีเทนซีเอ็มของโรงพยาบาลจุฬาลงกรณ์ 2
รูปที่ 2 การนำระบบจำแนกประเภทโรคมาใช้งาน 3
รูปที่ 3 ส่วนประกอบของต้นไม้ตัดสินใจ 10
รูปที่ 4 เพอร์เซปตรอน 12
รูปที่ 5 หลักการทำงานของวิธี Hold Method 13
รูปที่ 6 หลักการทำงานของวิธี 5-fold Cross Validation 14
รูปที่ 7 การเรียนรู้แบบมีการสนับสนุน 15
รูปที่ 8 ขั้นตอนการดำเนินงาน 24
รูปที่ 9 จำนวนโรคทั้งหมดที่นำมาใช้ในงานวิจัย 25
รูปที่ 10 กราฟเปรียบเทียบแบบจำลองที่สร้างโดยใช้องค์ประกอบของข้อมูลที่แตกต่างกัน 27
รูปที่ 11 กราฟเปรียบเทียบแบบจำลองที่สร้างด้วยอัตราส่วนระหว่างข้อมูลจากเวชระเบียนกับข้อมูลจากเว็บไซต์ที่แตกต่างกัน 32
รูปที่ 12 กระบวนการสร้างแบบจำลอง 36
รูปที่ 13 การใช้โมดูลสต๊อปเวิร์ดตัดคำที่ไม่สำคัญ 37
รูปที่ 14 การใช้โมดูลสเตมเมอร์และเลมมาไทเซอร์เปลี่ยนรูปคำ 38
รูปที่ 15 การใช้โมดูลเลมมาไทเซอร์กับบริบทของคำเปลี่ยนรูปคำ 38
รูปที่ 16 ขั้นตอนการประมวลผลข้อมูลก่อน 39
รูปที่ 17 กราฟแสดงจำนวนของคำในแต่ละช่วงความถี่-ส่วนกลับของความถี่ของคำ 40
รูปที่ 18 กราฟเปรียบเทียบแบบจำลองที่สร้างโดยการไม่ตัดและการตัดคำที่ไม่สำคัญทางการแพทย์ 41
รูปที่ 19 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมการเรียนรู้เบส์อย่างง่ายด้วยค่าแอลฟาที่แตกต่างกัน 43

รูปที่ 20 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนด้วยค่าซีที่แตกต่างกัน 44

รูปที่ 21 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนด้วยเคอร์เนลที่แตกต่างกัน 45

รูปที่ 22 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมโครงข่ายประสาทเทียมด้วยค่าแอลฟาที่แตกต่างกัน 46

รูปที่ 23 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมต้นไม้ตัดสินใจ โดยใช้ความถี่ของคำกับความถี่-ส่วนกลับของความถี่ของคำ 48

รูปที่ 24 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมการเรียนรู้เบส์อย่างง่าย โดยใช้ความถี่ของคำกับความถี่-ส่วนกลับของความถี่ของคำ 49

รูปที่ 25 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ความถี่ของคำกับความถี่-ส่วนกลับของความถี่ของคำ 50

รูปที่ 26 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมโครงข่ายประสาทเทียม โดยใช้ความถี่ของคำกับความถี่-ส่วนกลับของความถี่ของคำ 51

รูปที่ 27 ยูสเคสของระบบจำแนกประเภทโรค 53

รูปที่ 28 ส่วนต่อประสานกับผู้ใช้ของแบบจำลอง 55

รูปที่ 29 หน้าจอสำหรับกรอกและบันทึกรหัสไอซีดีเทนซีเอ็ม 56

รูปที่ 30 การพิจารณาผลบวกจริงของรหัสไอซีดีเทนซีเอ็ม M51 58

รูปที่ 31 การพิจารณาผลบวกเท็จของรหัสไอซีดีเทนซีเอ็ม M51 58

รูปที่ 32 การพิจารณาผลลบจริงของรหัสไอซีดีเทนซีเอ็ม M51 59

รูปที่ 33 การพิจารณาผลลบเท็จของรหัสไอซีดีเทนซีเอ็ม M51 59

รูปที่ 34 กราฟเปรียบเทียบแบบจำลองที่สร้างด้วยตัวจำแนกประเภททั้ง 4 ชนิด 61

รูปที่ 35 ตัวอย่างการทำงานของแบบจำลอง กรณีที่ข้อมูลเข้าเป็นภาษาอังกฤษ 69

รูปที่ 36 ตัวอย่างการทำงานของแบบจำลอง กรณีที่ข้อมูลเข้าเป็นภาษาไทย 70

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

งานทางการแพทย์คือการรักษาผู้ป่วยให้หายจากโรคหรืออาการเจ็บป่วย โดยแพทย์จะทำการตรวจและซักประวัติผู้ป่วย เพื่อให้ทราบถึงที่มาและลักษณะอาการที่ทำให้ผู้ป่วยรู้สึกผิดปกติ จากนั้นแพทย์จะนำข้อมูลที่ได้มาทำการวินิจฉัยโรคเพื่อหาวิธีรักษาผู้ป่วย แต่ในปัจจุบันการวินิจฉัยโรคที่ผิดพลาดถือเป็นปัญหาสำคัญในวงการแพทย์ [1] โดยพบว่าการวินิจฉัยโรคเบื้องต้นของแพทย์มีโอกาสเกิดข้อผิดพลาดขึ้นได้ประมาณร้อยละ 35 [2] เนื่องจากการวินิจฉัยโรคของแพทย์แต่ละคนอาจแตกต่างกันไปตามความรู้ ความชำนาญ และประสบการณ์ที่ได้สั่งสมมา รวมถึงในบางครั้งแพทย์อาจลืมนึกถึงโรคบางโรคไป เพราะเป็นโรคที่เกิดขึ้นได้ยากหรือไม่ค่อยได้พบเจอ ทำให้การวินิจฉัยโรคของแพทย์อาจผิดพลาดไปจากที่ควรจะเป็น

หลังจากการวินิจฉัยโรค ขั้นตอนต่อมาคือการบันทึกคำวินิจฉัยของแพทย์ลงในเวชระเบียน ซึ่งในการบันทึกคำวินิจฉัยนี้จะต้องมีการบันทึกรหัสไอซีดีเทนซีเอ็ม (ICD-10-CM code) กำกับทุกครั้ง เพราะรหัสไอซีดีเทนซีเอ็มเป็นรหัสโรคสากลที่ใช้สื่อสารกันในทุก ๆ โรงพยาบาล โดยรหัสไอซีดีเทนซีเอ็มสามารถจำแนกได้จากคำที่ระบุถึงชื่อโรค ชนิดของโรค และตำแหน่งของโรคที่อยู่ในบันทึกคำวินิจฉัยบนเวชระเบียน

รูปที่ 1 คือรูปแสดงขั้นตอนในการจำแนกรหัสไอซีดีเทนซีเอ็มที่งานวิจัยนี้ได้เข้าไปศึกษา เป็นขั้นตอนของโรงพยาบาลจุฬาลงกรณ์ ซึ่งจะประกอบไปด้วยขั้นตอนต่าง ๆ [3] ดังนี้

ขั้นตอนที่ 1 แพทย์ทำการบันทึกคำวินิจฉัยลงระบบเวชระเบียน

ขั้นตอนที่ 2 แพทย์ทำการพิมพ์เอกสารคำวินิจฉัยลงใบบันทึกคำวินิจฉัย และส่งต่อให้กับเจ้าหน้าที่เวชสถิติ

ขั้นตอนที่ 3 เจ้าหน้าที่เวชสถิติทำการจำแนกรหัสไอซีดีเทนซีเอ็มของคำวินิจฉัย และบันทึกข้อมูลลงระบบเวชระเบียน

ขั้นตอนที่ 4 เจ้าหน้าที่ฝ่ายการเงินทำการตรวจสอบความครบถ้วนของข้อมูลที่เจ้าหน้าที่เวชสถิติทำการบันทึก

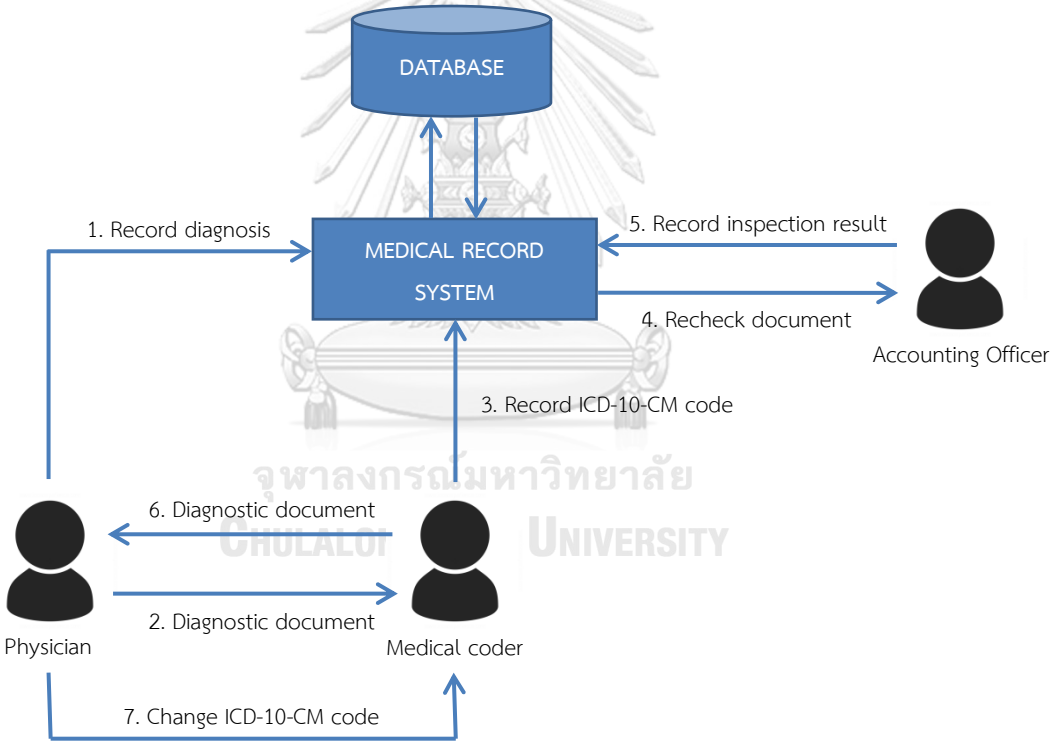
ขั้นตอนที่ 5 เจ้าหน้าที่ฝ่ายการเงินทำการบันทึกผลการตรวจสอบลงระบบเวชระเบียน

ขั้นตอนที่ 6 กรณีที่เจ้าหน้าที่ฝ่ายการเงินทำการตรวจสอบและอนุมัติเรียบร้อยแล้ว เจ้าหน้าที่เวชสถิติจะทำการพิมพ์เอกสารสรุปคำวินิจฉัยที่บันทึกรหัสไอซีดี

เทนซีเอ็มส่งให้กับแพทย์ เพื่อให้แพทย์ทำการตรวจสอบรหัสไอซีดีเทนซีเอ็มอีกครั้ง

ขั้นตอนที่ 7 กรณีที่รหัสไอซีดีเทนซีเอ็มที่ทางเจ้าหน้าที่เวชสถิติบันทึกไม่ถูกต้อง แพทย์จะแจ้งกลับมายังเจ้าหน้าที่เวชสถิติ เพื่อให้ทำการแก้ไขรหัสไอซีดีเทนซีเอ็มใหม่อีกครั้ง เมื่อเจ้าหน้าที่เวชสถิติทำการแก้ไขรหัสไอซีดีเทนซีเอ็มเรียบร้อยจะต้องเข้าสู่ขั้นตอนที่ 4 และทำขั้นตอนถัดมาทั้งหมดใหม่อีกครั้งจนกว่ารหัสไอซีดีเทนซีเอ็มที่ทำการบันทึกจะถูกต้องตรงตามคำวินิจฉัยแพทย์

จากขั้นตอนในการจำแนกรหัสไอซีดีเทนซีเอ็มข้างต้น อุปสรรคที่พบคือจำนวนบุคลากรที่มีไม่เพียงพอต่อการทำงาน ทำให้เจ้าหน้าที่เวชสถิติหนึ่งคนต้องรับหน้าที่ในการจำแนกรหัสไอซีดีเทนซีเอ็มจากแพทย์มากกว่าหนึ่งคน ส่งผลให้การทำงานล่าช้า

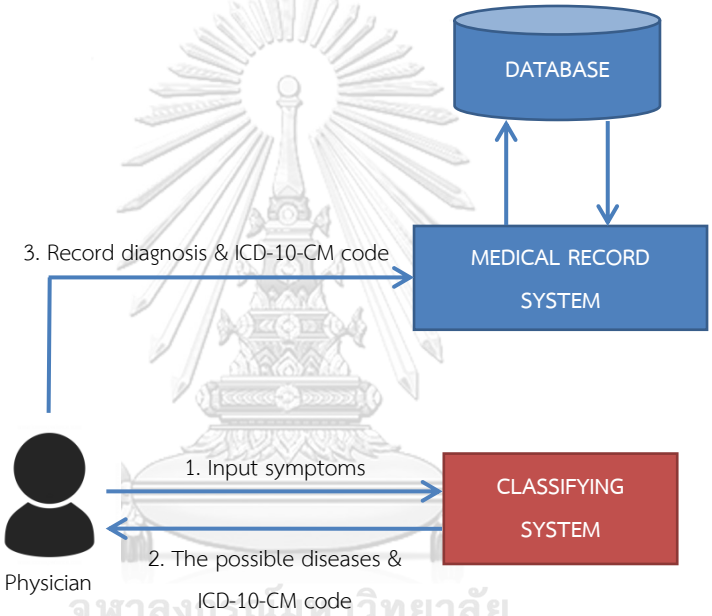


รูปที่ 1 ขั้นตอนการจำแนกรหัสไอซีดีเทนซีเอ็มของโรงพยาบาลจุฬาลงกรณ์

จากที่กล่าวมาข้างต้นจะพบว่างานทางการแพทย์นั้นมีข้อจำกัดอยู่ 2 ประการ ประการแรกคือข้อจำกัดในการวินิจฉัยโรคของแพทย์ที่อาจไม่แม่นยำเสมอไป และประการที่สองคือข้อจำกัดในการจำแนกรหัสไอซีดีเทนซีเอ็มที่ต้องใช้ระยะเวลานาน ดังนั้นงานวิจัยนี้จึงมีแนวคิดที่จะพัฒนาระบบสำหรับจำแนกประเภทโรคจากอาการ โดยการสร้างแบบจำลองเชิงทำนาย (Predictive modeling)

ขึ้นมา เพื่อใช้ในการประมวลผลหาชื่อโรคที่มีความน่าจะเป็นจากข้อมูลอาการที่ผู้ใช้กรอก และแสดงรายชื่อโรคออกมาโดยเรียงตามลำดับความน่าจะเป็นจากมากไปน้อย ซึ่งผลลัพธ์ที่ได้จะเป็นตัวช่วยแพทย์ในการตัดสินใจวินิจฉัยโรค ทำให้แพทย์สามารถวินิจฉัยโรคได้ง่ายขึ้นและช่วยเพิ่มประสิทธิภาพในการวินิจฉัยโรคของแพทย์ นอกจากนี้รายชื่อโรคที่ปรากฏในผลลัพธ์จะมีรหัสไอซีดีเทนซีเอ็มของแต่ละโรคกำกับอยู่ ทำให้แพทย์สามารถจำแนกรหัสไอซีดีเทนซีเอ็มได้สะดวกและรวดเร็วยิ่งขึ้น

รูปที่ 2 คือรูปแสดงการนำระบบจำแนกประเภทโรคมาใช้งาน โดยระบบจะเข้ามาช่วยในการวิเคราะห์ข้อมูลอาการ และแสดงรายชื่อโรคที่มีความน่าจะเป็นพร้อมทั้งจำแนกรหัสไอซีดีเทนซีเอ็มให้กับแพทย์



รูปที่ 2 การนำระบบจำแนกประเภทโรคมาใช้งาน

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อวิเคราะห์ความสัมพันธ์ระหว่างโรคและอาการ
2. เพื่อสร้างแบบจำลองสำหรับจำแนกประเภทโรคจากอาการ
3. เพื่อหาตัวจำแนกประเภทที่ดีและเหมาะสมที่สุดสำหรับการจำแนกประเภทโรค
4. เพื่อช่วยให้แพทย์สามารถทำการวินิจฉัยโรคได้ง่ายและมีประสิทธิภาพมากขึ้น
5. เพื่อช่วยให้แพทย์สามารถเข้าถึงรหัสไอซีดีเทนซีเอ็มได้สะดวกยิ่งขึ้น

1.3 ขอบเขตงานวิจัย

1. งานวิจัยนี้รองรับการจำแนกประเภทข้อความที่เป็นภาษาไทยและภาษาอังกฤษเท่านั้น
2. งานวิจัยนี้จะสร้างแบบจำลองสำหรับจำแนกประเภทโรคในหมวดกระดูกและกล้ามเนื้อ และโรคที่พบในเวชระเบียนผู้ป่วยแผนกออโธปิดิกส์ของโรงพยาบาลจุฬาลงกรณ์เท่านั้น
3. ข้อมูลที่นำมาใช้เป็นชุดข้อมูลสอน (Training dataset) สำหรับสร้างแบบจำลอง จะนำมาจากข้อมูลบนเว็บไซต์สาธารณะและข้อมูลบนเวชระเบียนผู้ป่วยแผนกออโธปิดิกส์ของโรงพยาบาลจุฬาลงกรณ์เท่านั้น
4. ข้อมูลที่นำมาใช้เป็นชุดข้อมูลทดสอบ (Test dataset) สำหรับวัดประสิทธิภาพการทำงานของแบบจำลอง จะใช้ข้อมูลจากเวชระเบียนผู้ป่วยแผนกออโธปิดิกส์ของโรงพยาบาลจุฬาลงกรณ์เท่านั้น

1.4 ขั้นตอนและวิธีการดำเนินงานวิจัย

1. ศึกษางานวิจัยที่เกี่ยวข้องกับการทำเหมืองข้อความ
2. ศึกษาความรู้และทฤษฎีที่เกี่ยวข้องกับงานวิจัย
3. ศึกษาเครื่องมือที่นำมาใช้ในการสร้างแบบจำลอง
4. ออกแบบกระบวนการสร้างแบบจำลอง
5. เก็บข้อมูลที่จะนำมาใช้สร้างแบบจำลอง
6. สร้างแบบจำลองสำหรับจำแนกประเภท
7. วัดประสิทธิภาพในการทำงานของแบบจำลอง
8. วิเคราะห์ผลการวัดประสิทธิภาพ
9. สรุปผลและจัดทำเล่มวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถใช้แบบจำลองเพื่อจำแนกประเภทโรคจากอาการได้
2. ทำให้ทราบว่าตัวจำแนกประเภทชนิดใด เหมาะสำหรับนำมาใช้ในการสร้างแบบจำลองสำหรับจำแนกประเภทโรค
3. ช่วยให้แพทย์สามารถตัดสินใจวินิจฉัยโรคผู้ป่วยได้เร็วขึ้น
4. ช่วยเพิ่มประสิทธิภาพในการวินิจฉัยโรคของแพทย์ให้มีความถูกต้องแม่นยำมากขึ้น
5. ช่วยให้แพทย์สามารถเข้าถึงรหัสไอซีดีเทนซีเอ็มได้สะดวกยิ่งขึ้น
6. สามารถนำแนวคิดของงานวิจัยนี้ไปประยุกต์ใช้กับโรคในหมวดอื่น ๆ ได้

1.6 โครงสร้างของเนื้อหาวิทยานิพนธ์

โครงสร้างของเนื้อหาวิทยานิพนธ์จะประกอบไปด้วย 5 บท ซึ่งมีรายละเอียดดังต่อไปนี้

บทที่ 1 กล่าวถึงที่มาและความสำคัญของปัญหา วัตถุประสงค์ ขอบเขตของงานวิจัย ขั้นตอนและวิธีการดำเนินงานวิจัย ประโยชน์ที่คาดว่าจะได้รับ

บทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง

บทที่ 3 กล่าวถึงแนวคิดและวิธีการดำเนินงาน

บทที่ 4 ยกล่าวถึงการทดสอบและประเมินผลงานวิจั

บทที่ 5 กล่าวถึงบทสรุป

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

งานวิจัยนี้จะสร้างแบบจำลองสำหรับจำแนกประเภทโรคจากอาการ เพื่อช่วยแพทย์ในการวินิจฉัยโรค ซึ่งขั้นตอนในการสร้างแบบจำลองจะมีการนำเอาทฤษฎีต่าง ๆ มาประยุกต์ใช้ โดยทฤษฎีที่เกี่ยวข้องมีดังต่อไปนี้

2.1.1 การวินิจฉัยโรค

การวินิจฉัยโรค (Medical diagnosis) เป็นขั้นตอนหนึ่งในการตรวจโรคของแพทย์ เพื่อให้สามารถวินิจฉัยหาสาเหตุของการเกิดโรค อาการ หรือภาวะผิดปกติต่าง ๆ ที่เกิดขึ้นกับผู้ป่วย เพื่อการรักษาและติดตามผลที่มีประสิทธิภาพ รวมถึงเพื่อประเมินสุขภาพของผู้ป่วย โดยหลักการในการวินิจฉัยโรคมี 2 วิธี [4] ดังนี้

1. **การวินิจฉัยทางคลินิก (Clinical diagnosis)**

   เป็นการวินิจฉัยโรคที่ได้จากการสอบถามอาการจากผู้ป่วย สอบถามประวัติทางการแพทย์ต่าง ๆ ของผู้ป่วยและการตรวจร่างกายเบื้องต้น ซึ่งจะช่วยในการวินิจฉัยโรคทั่วไปที่ไม่มีความรุนแรง โดยการวินิจฉัยทางคลินิกเป็นการวินิจฉัยที่พบได้บ่อยประมาณร้อยละ 80 ถึง 90 ของผู้ป่วย ตัวอย่างของโรคที่พบในการวินิจฉัยทางคลินิก เช่น โรคหวัด ท้องเสีย ท้องอืด ท้องเฟ้อ กรดไหลย้อน กระเพาะอาหารอักเสบ เป็นต้น

2. **การสืบค้นหรือการตรวจทางการแพทย์ (Medical investigation)**

   ใช้ในกรณีที่ผู้ป่วยเป็นโรคที่รุนแรง มีโอกาสพบได้น้อยประมาณร้อยละ 10 ถึง 20 ของผู้ป่วย โดยแพทย์จะทำการตรวจเพิ่มเติมจากการวินิจฉัยทางคลินิก การตรวจเพิ่มเติมจะขึ้นอยู่กับอาการของผู้ป่วย ประวัติทางการแพทย์ ผลการตรวจร่างกาย และดุลยพินิจของแพทย์ ตัวอย่างของการตรวจเพิ่มเติม เช่น การตรวจทางห้องปฏิบัติการ การตรวจภาพอวัยวะที่มีความผิดปกติทางรังสีวิทยา การตัดชิ้นเนื้อมาตรวจหรือการตรวจเซลล์ เป็นต้น

เมื่อแพทย์ได้ข้อมูลจากการสอบถามอาการ การตรวจร่างกาย และการสืบค้นแล้ว แพทย์จะสามารถวินิจฉัยได้ว่าผู้ป่วยเป็นโรคอะไรและเกิดขึ้นจากสาเหตุใด ซึ่งจะนำไปสู่การรักษาและติดตามผลผู้ป่วยอย่างถูกวิธี

ในบางครั้งหากแพทย์ไม่สามารถหาสาเหตุของโรคที่แน่ชัดได้ แพทย์อาจให้การรักษาผู้ป่วยตามดุลยพินิจว่าผู้ป่วยน่าจะเป็นโรคอะไรมากที่สุด ซึ่งเมื่อแพทย์ให้การรักษาตามนั้นแล้วผู้ป่วยหายได้ จะเรียกการวินิจฉัยแบบนี้ว่า การวินิจฉัยด้วยการรักษา (Therapeutic diagnosis)

2.1.2 รหัสไอซีดีเทนซีเอ็ม

ICD ย่อมาจาก International Classification of Diseases and Related Health Problems เป็นบัญชีจำแนกโรคระหว่างประเทศที่จัดทำขึ้นโดยองค์การอนามัยโลก (World Health Organization: WHO) เริ่มใช้ตั้งแต่ปี ค.ศ. 1893 โดยมีวัตถุประสงค์เพื่อใช้ในการจัดหมวดหมู่ของโรคและปัญหาสุขภาพต่าง ๆ ที่พบในมนุษย์ และใช้เป็นระบบรหัสโรคและรหัสปัญหาสุขภาพ ซึ่งมักจะถูกนำมาใช้ประโยชน์ในด้านระบาดวิทยา เวชสถิติ ระบบเวชสารสนเทศ การวางแผนยุทธศาสตร์ การวางแผนสุขภาพและการเบิกจ่ายค่ารักษาพยาบาล จากการบันทึกและรวบรวมข้อมูลทางสถิติ

ICD-10-CM [5] ย่อมาจาก 10th Revision of ICD Clinical Modification เป็นบัญชีจำแนกโรคระหว่างประเทศฉบับแก้ไขครั้งที่ 10 ซึ่งเป็นฉบับปรับปรุงครั้งล่าสุด ถูกปรับปรุงเมื่อปี ค.ศ. 2010 เป็นระบบที่มีองค์ประกอบสำคัญ 2 ส่วน คือ ระบบการจัดหมวดหมู่ของโรคและปัญหาสุขภาพต่าง ๆ ที่พบในมนุษย์ และระบบรหัสโรคและรหัสปัญหาสุขภาพ

ผู้ป่วยที่เข้ารับการรักษาจากโรงพยาบาลจะได้รหัสโรค 1 รหัสต่อการเข้ารักษา 1 ครั้ง โดยปกติแพทย์ที่ทำการดูแลรักษาผู้ป่วยจะเป็นผู้สรุปว่าผู้ป่วยเป็นโรคอะไร หรือมีภาวะความผิดปกติเป็นอย่างไร แม้บางครั้งอาจมีความก้ำกึ่งซึ่งยากต่อการตัดสินใจ แต่แพทย์ก็ต้องสรุปให้ได้เพื่อจะได้ทำการรักษาผู้ป่วยต่อไป

ลักษณะของรหัสไอซีดีเทนซีเอ็มเป็นรหัสที่ประกอบด้วยตัวเลขและตัวอักษร (Alphanumeric code) โดยรหัสแต่ละตัวจะขึ้นต้นด้วยอักษรภาษาอังกฤษ A-Z แล้วตามด้วยตัวเลขอารบิก 0-9 อีกประมาณ 2-4 ตัว จึงเป็นรหัสที่มีความยาว 3-5 อักขระ ตัวอย่างเช่น I10 เป็นรหัสแทนโรค Hypertension และ J18.9 เป็นรหัสแทนโรค Pneumonia

ไอซีดีเทนซีเอ็มเป็นระบบรหัสโรคและรหัสปัญหาสุขภาพพร้อมคำอธิบาย ที่ได้มีการแบ่งเนื้อหาออกเป็นบทต่าง ๆ รวมทั้งสิ้น 21 บท [6] ดังนี้

กลุ่มที่ 1	โรคติดเชื้อ	ใช้รหัส A00-B99
กลุ่มที่ 2	เนื้องอกและมะเร็ง	ใช้รหัส C00-D49
กลุ่มที่ 3	โรคเลือด	ใช้รหัส D50-D89
กลุ่มที่ 4	โรคต่อมไร้ท่อ	ใช้รหัส E00-E89
กลุ่มที่ 5	โรคจิต โรคประสาท พฤติกรรม	ใช้รหัส F01-F99

กลุ่มที่ 6	โรคสมองและระบบประสาท	ใช้รหัส G00-G99
กลุ่มที่ 7	โรคตา	ใช้รหัส H00-H59
กลุ่มที่ 8	โรคหู	ใช้รหัส H60-H95
กลุ่มที่ 9	โรคหัวใจและหลอดเลือด	ใช้รหัส I00-I99
กลุ่มที่ 10	โรคปอดและระบบหายใจ	ใช้รหัส J00-J99
กลุ่มที่ 11	โรคระบบย่อยอาหาร	ใช้รหัส K00-K95
กลุ่มที่ 12	โรคผิวหนัง	ใช้รหัส L00-L99
กลุ่มที่ 13	โรคกล้ามเนื้อและกระดูก	ใช้รหัส M00-M99
กลุ่มที่ 14	โรคไตและระบบทางเดินปัสสาวะ	ใช้รหัส N00-N99
กลุ่มที่ 15	ตั้งครรภ์ การคลอด	ใช้รหัส O00-O9A
กลุ่มที่ 16	โรคของทารกแรกเกิด	ใช้รหัส P00-P96
กลุ่มที่ 17	พิการแต่กำเนิด	ใช้รหัส Q00-Q99
กลุ่มที่ 18	อาการและอาการแสดงผิดปกติ	ใช้รหัส R00-R99
กลุ่มที่ 19	การบาดเจ็บและการได้รับพิษ	ใช้รหัส S00-T88
กลุ่มที่ 20	สาเหตุภายนอกของการบาดเจ็บ	ใช้รหัส V01-Y98
กลุ่มที่ 21	การให้บริการสุขภาพ	ใช้รหัส Z00-Z99

การให้รหัสโรค จะเริ่มจากการตรวจสอบข้อมูลโรคที่ปรากฏอยู่ในเวชระเบียน จากนั้นเปลี่ยนคำย่อทุกคำให้เป็นคำเต็มและเลือกคำหลักของโรคทั้งหมดมา เพื่อใช้ในการเปิดหารหัสไอซีดีเทนซีเอ็มจากดรรชนี และกำหนดรหัสโรคโดยเริ่มเรียงลำดับจากรหัสโรคหลัก รหัสการวินิจฉัยร่วม รหัสโรคแทรกซ้อน รหัสการวินิจฉัยอื่น และรหัสสาเหตุของการบาดเจ็บ

รหัสไอซีดีเทนซีเอ็มสามารถนำไปใช้ประโยชน์ในการวินิจฉัยสาเหตุการตาย เพื่อประเมินสาเหตุการตายที่พบบ่อย ใช้ในการวินิจฉัยโรคในฐานข้อมูลของผู้ป่วย เพื่อประเมินสถานการณ์โรคที่พบบ่อย ประเมินผลลัพธ์ของผู้ป่วยเฉพาะโรค ประเมินความต่อเนื่องของบริการเฉพาะโรค ประเมินการเข้าถึงบริการของผู้ป่วยเฉพาะโรค และประเมินประสิทธิภาพของบริการเฉพาะโรค

2.1.3 การทำเหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) [7] เป็นศาสตร์ที่จะนำไปสู่การค้นพบความรู้ในฐานข้อมูลขนาดใหญ่ (Knowledge discovery in large database) หมายถึงกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบ (patterns) และความสัมพันธ์ (associations) ที่ซ่อนอยู่ในชุดข้อมูลนั้น กระบวนการดังกล่าวมีความเป็นอัตโนมัติไม่สามารถประมวลผลได้ด้วยมือต้องใช้คอมพิวเตอร์เข้ามาช่วย เนื่องจากข้อมูลมีปริมาณมาก ผลลัพธ์จากการทำเหมืองข้อมูล คือ

ความรู้ ซึ่งเป็นรูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลหนึ่ง ๆ โดยรูปแบบนั้นจะสะท้อนถึงเหตุการณ์หรือสิ่งที่เกิดขึ้นซ้ำแล้วซ้ำอีก (repeat) จนสามารถทำนายได้ (predictable) ตัวอย่างเช่น คนที่เป็นโรคชนิดหนึ่ง มักจะมียีนรูปแบบนี้ ซึ่งความรู้ดังกล่าวสามารถนำมาใช้ในการวินิจฉัยโรคทางการแพทย์ได้ ดังนั้นรูปแบบหรือความสัมพันธ์ของลักษณะต่าง ๆ ที่พบในข้อมูลดิบนับเป็นความรู้ที่สามารถนำไปใช้ประโยชน์ในด้านต่าง ๆ ได้ เช่น ในเชิงธุรกิจ การวินิจฉัยหรือรักษาโรคทางการแพทย์ การกีฬา เป็นต้น

รูปแบบการนำเสนอผลลัพธ์ความรู้จากการทำเหมืองข้อมูล เช่น กฎความสัมพันธ์ (Association rule) กฎการจำแนกประเภท (Classification rule) เป็นต้น โดยแบบจำลองที่เป็นผลลัพธ์จากการทำเหมืองข้อมูล สามารถแบ่งได้เป็น 2 ประเภท [7] ดังนี้

1. **แบบจำลองเชิงทำนาย (Predictive/Supervised modeling)**

   เป็นผลลัพธ์ที่สร้างจากการอนุมาน (inference) ชุดข้อมูลปัจจุบัน เพื่อใช้ในการทำนายประเภทตัวอย่างในอนาคต แบบจำลองเชิงทำนายเป็นผลลัพธ์จากการทำเหมืองจำแนกประเภทข้อมูลออกเป็นกลุ่มที่ทราบล่วงหน้าตามคุณลักษณะของข้อมูลที่เรียกว่า ฉลากประเภท (class label) โดยถ้าค่าฉลากประเภทเป็นค่าไม่ต่อเนื่อง จะเรียกกระบวนการที่ใช้แยกแยะว่า การจำแนกประเภท (Classification) ถ้าค่าฉลากประเภทเป็นค่าต่อเนื่อง จะเรียกกระบวนการที่ใช้แยกแยะว่า การถดถอย (Regression)

2. **แบบจำลองเชิงพรรณนา (Descriptive/Unsupervised modeling)**

   เป็นการหาความสัมพันธ์ต่าง ๆ หรือหาวิธีการจัดกลุ่มข้อมูล (Clustering) ซึ่งไม่ได้มีจุดมุ่งหมายเพื่อการทำนาย แต่เพื่อทำให้เข้าใจถึงสาเหตุหรือปัจจัยของปัญหาหรือสิ่งที่สนใจได้ดีขึ้น เช่น เห็นการกระจายตัวของกลุ่มข้อมูล หรือเห็นความสัมพันธ์ของข้อมูลในฐานข้อมูล

2.1.4 การจำแนกประเภท

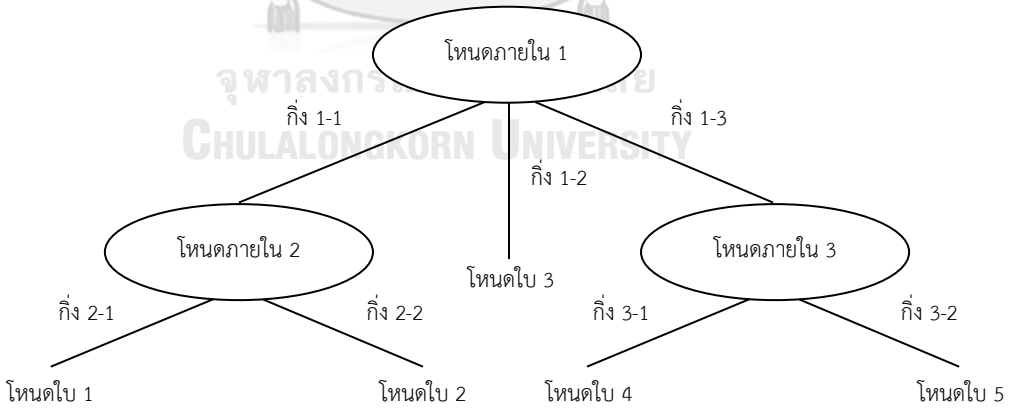
การทำเหมืองกฎการจำแนกประเภท [7] จัดเป็นการสร้างแบบจำลองเชิงทำนาย เป็นการค้นหาแบบจำลองหรือฟังก์ชันซึ่งสามารถจำแนกความแตกต่างของประเภทข้อมูล เพื่อใช้ในการทำนาย โดยการสร้างแบบจำลองจะขึ้นอยู่กับการวิเคราะห์ชุดข้อมูลสอน จึงถือเป็นการทำเหมืองประเภทหนึ่ง โดยตัวอย่างในชุดข้อมูลสอนจะมีคุณลักษณะหนึ่งซึ่งบอกประเภทของตัวอย่างเรียกว่า ฉลากประเภท ซึ่งเป็นข้อมูลแบบไม่ต่อเนื่อง (categorical) ตัวจำแนกประเภทจะถูกสอนให้เรียนรู้การจำแนกประเภทจากตัวอย่างในชุดข้อมูลสอน ผลลัพธ์คือแบบจำลองที่สร้างขึ้นเพื่อใช้จำแนกประเภทข้อมูลของตัวอย่างใหม่หรือตัวอย่างที่ไม่เคยเห็นมาก่อน

การจำแนกประเภทประกอบด้วย 2 ขั้นตอน ขั้นตอนแรกคือการสร้างแบบจำลอง โดยเซตของตัวอย่างที่เรียกว่า ชุดข้อมูลสอน ตัวอย่างแต่ละตัวจะมีคุณลักษณะหนึ่ง ซึ่งบอกค่าประเภทที่กำหนดไว้แล้วล่วงหน้า ขั้นตอนที่สองคือการนำแบบจำลองที่ได้ไปใช้ เพื่อจำแนกประเภทตัวอย่างในอนาคต ซึ่งจะต้องมีการประเมินความถูกต้องของแบบจำลองที่ได้ก่อนนำไปใช้ โดยเปรียบเทียบค่าฉลากประเภทที่ทราบล่วงหน้าของตัวอย่างในชุดข้อมูลทดสอบกับค่าผลลัพธ์การจำแนกประเภทที่ได้จากแบบจำลอง

2.1.5 ต้นไม้ตัดสินใจ

การเรียนรู้ของต้นไม้ตัดสินใจ (Decision Tree) [7] เป็นการเรียนรู้โดยการจำแนกประเภทข้อมูลออกเป็นประเภทต่าง ๆ โดยใช้คุณลักษณะของข้อมูลในการจำแนกประเภท ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ ทำให้ทราบว่าคุณลักษณะใดของข้อมูลที่เป็นตัวกำหนดการจำแนกประเภท และคุณลักษณะแต่ละตัวมีความสำคัญมากน้อยแตกต่างกันอย่างไรต่อการจำแนกประเภท ซึ่งผลลัพธ์ที่ได้จากการเรียนรู้ของต้นไม้ตัดสินใจแสดงในรูปที่ 3 ประกอบด้วย

- โหนดภายใน (internal node) คือคุณลักษณะต่าง ๆ ของข้อมูล โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้ เรียกว่า โหนดราก
- กิ่ง (branch) คือค่าของคุณลักษณะของโหนดภายในที่แตกกิ่งออกมา
- โหนดใบ (leaf node) คือกลุ่มต่าง ๆ ซึ่งเป็นผลลัพธ์ในการจำแนกประเภทข้อมูล



รูปที่ 3 ส่วนประกอบของต้นไม้ตัดสินใจ [7]

2.1.6 การเรียนรู้เบส์อย่างง่าย

การเรียนรู้เบส์อย่างง่าย (Naïve Bayesian Learning) [7] มาจากทฤษฎีของเบส์ ซึ่งใช้คำนวณหาความน่าจะเป็นที่ข้อมูลน่าจะถูกจำแนกอยู่ในประเภทใด แต่จะลดความซับซ้อนลงโดย

เพิ่มสมมติฐานที่ว่าคุณลักษณะต่าง ๆ ของข้อมูลจะไม่ขึ้นต่อกัน หรือกล่าวได้ว่าความน่าจะเป็นของข้อมูล X ที่มีคุณลักษณะ n ตัว หรือ $X = \{A_1,...,A_n\}$ จะถูกจำแนกเป็นกลุ่ม C_i มีค่าเท่ากับ

$$P(C_i \mid A_1,...,A_n) = \frac{P(A_1,...,A_n \mid C_i) \cdot P(C_i)}{P(A_1,...,A_n)} \quad [7]$$

การเรียนรู้เบส์อย่างง่ายประกอบด้วย 2 ขั้นตอน คือ ขั้นตอนการเรียนรู้ และขั้นตอนการจำแนกประเภท โดยขั้นตอนการเรียนรู้จะทำการประมาณค่าความน่าจะเป็นจากตัวอย่างในชุดข้อมูลสอน ได้แก่ ค่าประมาณความน่าจะเป็นของประเภทต่าง ๆ และค่าประมาณความน่าจะเป็นของแต่ละคุณลักษณะเมื่อรู้ประเภท สำหรับขั้นตอนการจำแนกประเภทเป็นการเลือกค่าความน่าจะเป็นสูงสุดระหว่าง C_i ของตัวอย่างใหม่ที่ทราบคุณลักษณะของข้อมูล โดยใช้สมการทฤษฎีของเบส์และค่าประมาณต่าง ๆ ที่คำนวณไว้ล่วงหน้าในขั้นตอนการเรียนรู้บนสมมติฐานแบบมีเงื่อนไข

2.1.7 ซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) [8] เป็นการใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน ซึ่งหลักการของซัพพอร์ตเวกเตอร์แมชชีนคือการหาสัมประสิทธิ์ของสมการ เพื่อสร้างเส้นจำแนกประเภทข้อมูลในขั้นตอนการเรียนรู้ และเลือกเส้นจำแนกประเภทข้อมูลที่เหมาะสมที่สุด โดยพยายามให้ระยะห่างระหว่างขอบเขตของทั้ง 2 กลุ่มมีระยะห่างมากที่สุด เพื่อลดความผิดพลาดในการจำแนกประเภท เนื่องจากถ้าระยะห่างยิ่งมากความผิดพลาดในการจำแนกประเภทก็จะมีโอกาสเกิดขึ้นน้อยลง

กำหนดให้ $(x_i, y_i), ..., (x_n, y_n)$ เป็นตัวอย่างในชุดข้อมูลสอน โดย n คือจำนวนข้อมูลตัวอย่าง m คือจำนวนมิติของข้อมูลเข้า และ y คือผลลัพธ์ที่มีค่าเป็น +1 หรือ -1 ดังสมการ

$$(x_i, y_i), ..., (x_n, y_n) \text{ เมื่อ } x \in R^m, y \in \{+1, -1\} \quad [8]$$

สำหรับปัญหาเชิงเส้น มิติข้อมูลขนาดสูงได้ถูกแบ่งเป็น 2 กลุ่มโดยระนาบตัดสินใจ ซึ่งคำนวณได้ ดังสมการ

$$(w \cdot x) + b = 0 \quad [8]$$

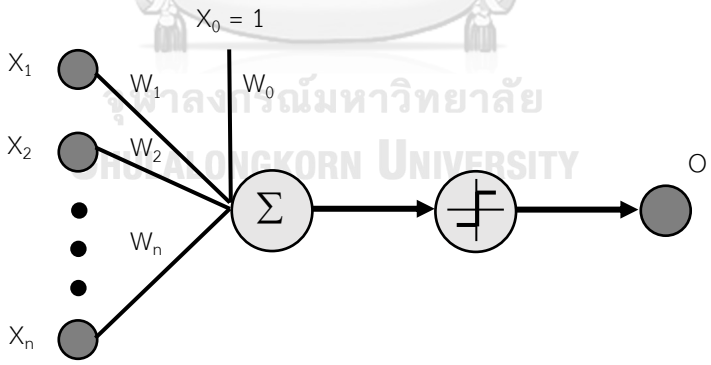
เมื่อ w คือค่าน้ำหนัก และ b คือค่าความเอนเอียง (bias) จะสามารถนำมาใช้คำนวณเพื่อทำการจำแนกประเภทข้อมูลได้ ดังสมการ

$$(w \cdot x) + b > 0 \text{ ถ้า } y_i = +1 \text{ และ } (w \cdot x) + b < 0 \text{ ถ้า } y_i = -1 \quad [8]$$

ในกรณีที่ปัญหาการจำแนกประเภทมีจำนวนประเภทมากกว่า 2 ประเภท จะใช้วิธีมัลติคลาสซัพพอร์ตเวกเตอร์แมชชีน (Multi-Class Support Vector Machine) เข้ามาช่วย โดยการนำค่าสมการจำแนกของแต่ละประเภทมาเปรียบเทียบกันด้วยวิธีต่าง ๆ ได้แก่ วิธี 1 ต่อทั้งหมด วิธี 1 ต่อ 1 และวิธีกราฟอวัฏจักรระบุทิศทาง (Directed Acyclic Graph) จากนั้นเลือกประเภทที่ให้ค่าสมการจำแนกสูงสุดมาเป็นผลลัพธ์

2.1.8 โครงข่ายประสาทเทียม

เพอร์เซปตรอน (Perceptron) [7] เป็นหน่วยย่อยที่สุดของโครงข่ายประสาทเทียม (Neural Network) ทำหน้าที่รับข้อมูลเข้าเป็นเวกเตอร์จำนวนจริงเข้ามาแล้วคำนวณหาผลรวมเชิงเส้นแบบถ่วงน้ำหนักของข้อมูลเข้า $x_1, x_2, \ldots, x_n$ โดยที่ค่า $w_1, w_2, \ldots, w_n$ ในรูปที่ 4 เป็นค่าน้ำหนักของข้อมูลเข้า และให้ข้อมูลออกเป็นค่าคงที่ที่แตกต่างกันตามค่าผลรวมที่ได้ว่ามีค่าเกินค่าขีดแบ่ง (θ) หรือไม่ ทั้งนี้ค่าข้อมูลออกจะแตกต่างกันไปตามฟังก์ชันกระตุ้น (Activation function) ที่ใช้ ข้อมูลออกที่ได้จะถูกนำไปคำนวณค่าความผิดพลาด เพื่อนำมาปรับน้ำหนักของข้อมูลเข้าต่อไป ส่วน w_0 ในรูปที่ 4 เป็นค่าลบของค่าขีดแบ่ง และ x_0 เป็นข้อมูลเข้าเทียมกำหนดให้มีค่าเป็น 1 เสมอ



รูปที่ 4 เพอร์เซปตรอน [7]

ฟังก์ชันกระตุ้นในรูปที่ 4 เป็นชนิดที่เรียกว่า ฟังก์ชันสองขั้ว (Bipolar function) จะแสดงผลของข้อมูลออกเป็น 1 ถ้าผลรวมเชิงเส้นที่ได้มีค่าเกินค่าขีดแบ่ง และเป็น -1 ถ้าไม่เกิน โดยข้อมูลออก (o) สามารถแสดงในรูปฟังก์ชันของข้อมูลเข้า ($x_1, x_2, \ldots, x_n$) ได้ดังนี้

$$o(x_1, x_2, \ldots, x_n) = \begin{cases} 1 & \text{ถ้า } w_1x_1 + w_2x_2 + \ldots + w_nx_n > \theta \text{ หรือ} \\ -1 & \text{ถ้า } w_1x_1 + w_2x_2 + \ldots + w_nx_n < \theta \end{cases} \quad [7]$$

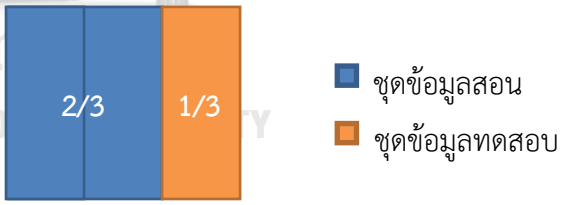
ข้อมูลออกเป็นฟังก์ชันของข้อมูลเข้าในรูปของผลรวมเชิงเส้นแบบถ่วงน้ำหนัก น้ำหนักจะเป็นตัวกำหนดว่าในจำนวนข้อมูลเข้า ข้อมูลเข้า x_i ตัวใดมีความสำคัญต่อการกำหนดค่าข้อมูลออก ตัวที่มีความสำคัญมากจะมีค่าสัมบูรณ์ของน้ำหนักมาก ในกรณีที่ผลรวมเท่ากับค่าขีดแบ่ง ค่าข้อมูลออกจะเป็น 1 หรือ -1 ก็ได้

2.1.9 ตัววัดความแม่นยำของแบบจำลองการจำแนกประเภทข้อมูล

การวัดความถูกต้องในการทำนายกลุ่มของตัวอย่างใหม่โดยตัวจำแนกประเภทชนิดต่าง ๆ สามารถวัดได้ด้วย 2 วิธี [7] ดังนี้

1. **วิธีการแบ่งชุดข้อมูลออกเป็น 2 ส่วน (Hold Method)**

   เป็นวิธีที่เหมาะกับชุดข้อมูลขนาดใหญ่ ตัวอย่างในชุดข้อมูลจะถูกแบ่งออกเป็น 2 ส่วนแบบสุ่ม ด้วยอัตราส่วนขนาดของชุดข้อมูลสอนเท่ากับ 2/3 และขนาดของชุดข้อมูลทดสอบเท่ากับ 1/3 ดังรูปที่ 5 โดยใช้ชุดข้อมูลสอนในการสร้างแบบจำลองการจำแนกประเภท และตรวจสอบความถูกต้องในการจำแนกประเภทข้อมูลใหม่ด้วยชุดข้อมูลทดสอบ ค่าความแม่นยำคำนวณได้จากอัตราส่วนระหว่างจำนวนตัวอย่างในชุดข้อมูลทดสอบที่ทำนายกลุ่มได้อย่างถูกต้องกับจำนวนตัวอย่างทั้งหมดในชุดข้อมูลทดสอบ

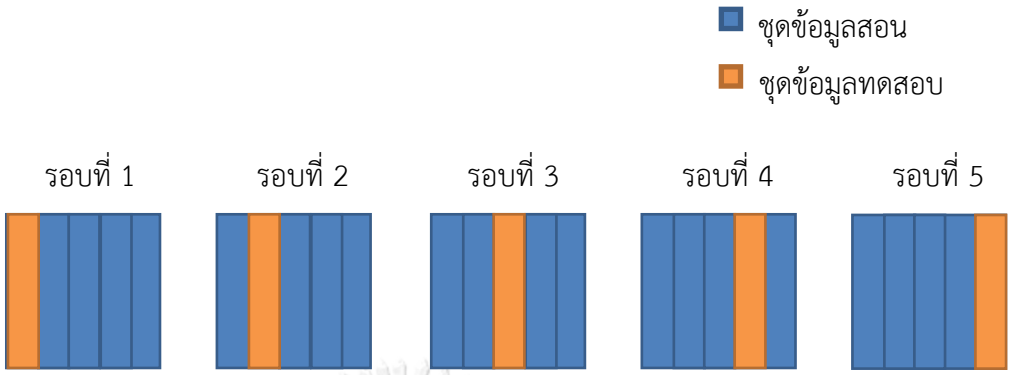


รูปที่ 5 หลักการทำงานของวิธี Hold Method

2. **วิธีการแบ่งชุดข้อมูลออกเป็น k ส่วน (K-fold Cross Validation)**

   เป็นวิธีที่เหมาะกับชุดข้อมูลจำนวนไม่มาก สมมติว่าขนาดของชุดข้อมูลเท่ากับ N ตัวอย่างในชุดข้อมูลจะถูกแบ่งออกเป็น k ส่วน โดยแต่ละชุดข้อมูลจะมีขนาด N/k วิธีนี้จะเรียนรู้ด้วยชุดข้อมูลสอนและตรวจสอบความถูกต้องในการจำแนกประเภทด้วยชุดข้อมูลทดสอบเป็นจำนวนทั้งหมด k รอบ โดยรอบที่ i จะใช้ชุดข้อมูลทดสอบชุดที่ i และใช้ชุดข้อมูลที่เหลือเป็นชุดข้อมูลสอน ดังรูปที่ 6 ค่าความแม่นยำคำนวณได้จากอัตราส่วน

ระหว่างจำนวนตัวอย่างในชุดข้อมูลทดสอบที่ทำนายกลุ่มได้อย่างถูกต้องทั้งหมด k รอบ กับจำนวนตัวอย่างทั้งหมดในชุดข้อมูล



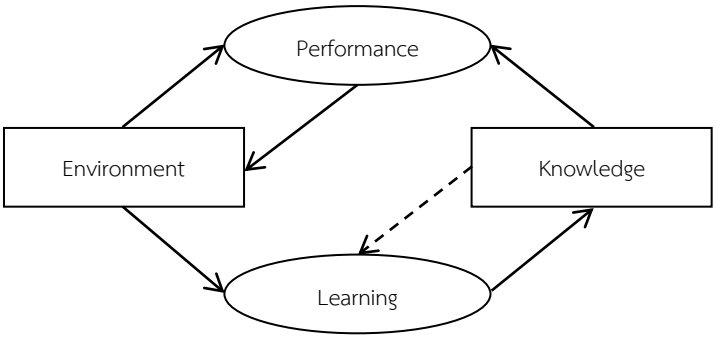
รูปที่ 6 หลักการทำงานของวิธี 5-fold Cross Validation

2.1.10 การเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง (Machine Learning) [7] เป็นสาขาหนึ่งของปัญญาประดิษฐ์ที่พยายามสอนคอมพิวเตอร์ให้สามารถเรียนรู้ได้เหมือนมนุษย์ ซึ่งวิธีการเรียนรู้ของเครื่องจะมีประสิทธิผลในการบ่งบอกคุณสมบัติและรูปแบบที่เป็นประโยชน์จากชุดข้อมูล การทำเหมืองข้อมูลได้นำวิธีการเรียนรู้ของเครื่องมาใช้ โดยเน้นที่การทำงานกับฐานข้อมูลขนาดใหญ่ เพื่อดึงความรู้จากข้อมูลที่เก็บมาใช้ให้เกิดประโยชน์ในด้านต่าง ๆ

การเรียนรู้ของเครื่องจัดเป็นการเรียนรู้เชิงอุปนัย (Induction-based learning) เป็นการให้เหตุผลโดยอาศัยข้อสังเกตหรือผลการทดลองจากหลาย ๆ ตัวอย่างมาเป็นข้อสรุป การเรียนรู้ของเครื่องจะพยายามสกัดให้ได้มาซึ่งมโนทัศน์ (concept) ซึ่งเป็นตัวแทนของชุดข้อมูลหนึ่งที่จะบ่งชี้ถึงความสัมพันธ์ของวัตถุที่อยู่ในประเภทเดียวกัน รูปแบบการเรียนรู้ของเครื่องมีทั้งหมด 3 รูปแบบ [9] ดังนี้

1. การเรียนรู้แบบมีตัวชี้นำ เป็นการเรียนรู้โดยอาศัยประเภทตัวอย่างที่ทราบล่วงหน้าจากชุดข้อมูลสอน เช่น การจำแนกประเภท เป็นต้น
2. การเรียนรู้แบบไม่มีตัวชี้นำ เป็นการเรียนรู้จากข้อมูลที่ไม่ทราบประเภทตัวอย่างล่วงหน้า แต่การเรียนรู้แบบนี้จะจัดแบ่งข้อมูลออกเป็นกลุ่ม ๆ บนพื้นฐานความเหมือนและความแตกต่างระหว่างข้อมูล เช่น การหาโครงสร้างที่ซ่อนอยู่ในข้อมูล เป็นต้น
3. การเรียนรู้แบบมีการสนับสนุน (Reinforcement learning) เป็นการเรียนรู้จากสิ่งที่สังเกตได้จากสิ่งแวดล้อมรอบตัว ดังรูปที่ 7 เช่น เกมโอเอ็กซ์ (OX) เป็นต้น



รูปที่ 7 การเรียนรู้แบบมีการสนับสนุน [9]

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 แนวคิดในการประยุกต์ใช้การทำเหมืองข้อความ

2.2.1.1 การทำเหมืองข้อความกับข้อมูลชีวการแพทย์

ข้อมูลชีวการแพทย์ถือเป็นแหล่งข้อมูลสำคัญที่สามารถนำไปใช้ประโยชน์ในการวินิจฉัย การรักษา และการป้องกันโรคได้ การทำเหมืองข้อความจึงเป็นเทคนิคที่นิยมนำมาใช้เพื่อสกัดหาความรู้จากข้อมูลชีวการแพทย์ [10] ดังจะเห็นได้จากมีงานวิจัยที่ใช้การทำเหมืองข้อความเพื่อหาปัจจัยเสี่ยงที่ทำให้เกิดโรคหัวใจ [11] และงานวิจัยอื่น ๆ ดังนี้

- งานวิจัยเรื่อง การสกัดข้อมูลจากรายงานทางพยาธิวิทยาภายในโรงพยาบาล (Information extraction from pathology reports in a hospital setting) [12] นำการทำเหมืองข้อความมาประยุกต์ใช้ในกระบวนการทำงานของโรงพยาบาลรอยัลเมลเบิร์น เพื่อสกัดหาข้อมูลที่เป็นประโยชน์จากรายงานทางพยาธิวิทยาและลดการใช้ความรู้จากผู้เชี่ยวชาญลง โดยใช้ตัวจำแนกประเภทชนิดต่าง ๆ ที่ได้รับความนิยมในการทำเหมืองข้อมูลมาเปรียบเทียบประสิทธิภาพที่ได้จากการจำแนกประเภทข้อมูลชุดเดียวกัน ซึ่งผลลัพธ์ที่ได้คือตัวจำแนกประเภทการเรียนรู้เบส์อย่างง่ายจะเหมาะสำหรับข้อมูลที่เป็นค่าไม่ต่อเนื่อง ส่วนซัพพอร์ตเวกเตอร์แมชชีนจะเหมาะสำหรับข้อมูลตัวเลข
- งานวิจัยเรื่อง การจำแนกประเภทโรคอัลไซเมอร์โดยใช้การจัดอันดับคุณลักษณะจากเอมอาร์ไอ (Feature-ranking-based Alzheimer's disease classification from structural MRI) [13] นำการทดสอบที (T-test) มาใช้ เพื่อคัดเลือกคุณลักษณะของโรคอัลไซเมอร์จากข้อมูลบนภาพถ่ายด้วยคลื่นแม่เหล็กไฟฟ้าหรือเอมอาร์ไอ และใช้ซัพพอร์ตเวกเตอร์แมชชีนมาเป็นตัวจำแนกประเภท โดยในงานวิจัยพบว่าการนำเทคโนโลยีต่าง ๆ มาใช้ร่วมกันจะช่วยเพิ่มประสิทธิภาพในการจำแนกประเภท

- งานวิจัยเรื่อง การประมาณแบบจำลองมโนทัศน์คำโดยใช้ความรู้และการแบ่งรายละเอียดสำหรับการทำเหมืองข้อความชีวการแพทย์ (Knowledge based word-concept model estimation and refinement for biomedical text mining) [14] นำเสนอแนวคิดในการสร้างแบบจำลองเชิงสถิติของมโนทัศน์คำจากการใช้ฐานความรู้เพื่อนำมาใช้กับข้อมูลทางชีวการแพทย์ โดยคำนวณค่าความน่าจะเป็นของมโนทัศน์คำจากการเลือกคำที่มีมโนทัศน์อยู่ในฐานความรู้ขึ้นมาและนำค่าเหล่านั้นมาสร้างแบบจำลอง เพื่อตีความหมายของคำตามบริบท โดยใช้การเรียนรู้เบส์อย่างง่ายมาเป็นตัวจำแนกประเภท

นอกจากการนำข้อมูลชีวการแพทย์มาใช้เพื่อการทำเหมืองข้อความแล้ว ยังมีการนำข้อมูลแนวโน้มของโรคที่ได้จากกูเกิลมาใช้ร่วมกับข้อมูลผู้ป่วย เพื่อช่วยเพิ่มประสิทธิภาพในการจำแนกประเภทอีกด้วย [15]

2.2.1.2 การทำเหมืองข้อความเพื่อช่วยลดงานทางการแพทย์

การทำงานในโรงพยาบาลมีผู้ป่วยที่ต้องรองรับในแต่ละวันเป็นจำนวนมากบวกกับจำนวนบุคลากรที่มีไม่เพียงพอ ทำให้การทำงานในโรงพยาบาลล่าช้าและเกิดความแออัดของผู้ป่วย ดังนั้นเพื่อช่วยเพิ่มประสิทธิภาพของการทำงานในโรงพยาบาล จึงมีงานวิจัยต่าง ๆ ที่นำการทำเหมืองข้อความมาประยุกต์ใช้ ดังนี้

- งานวิจัยเรื่อง การทำเหมืองข้อความเพื่อทำนายการเข้ารับการรักษาในโรงพยาบาลโดยใช้เวชระเบียนเบื้องต้นในแผนกฉุกเฉิน (Text mining approach to predict hospital admissions using early medical records from the emergency department) [16] นำการทำเหมืองข้อความมาใช้กับข้อมูลผู้ป่วยในแผนกฉุกเฉิน เพื่อใช้ในการทำนายการเข้ารับการรักษาตัวในโรงพยาบาลของผู้ป่วยและการโอนย้ายผู้ป่วย โดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นตัวจำแนกประเภท และสร้างแบบจำลองด้วยวิธีการสุ่มข้อมูลแบบความเที่ยงตรงโดยการแบ่งชุดข้อมูลเป็น 10 ส่วน (10-fold Cross Validation)
- งานวิจัยเรื่อง การทำเหมืองข้อความจากเวชระเบียนอิเล็กโทรนิกเพื่อจำแนกประเภทการเข้ารับการรักษาโรค : การวัดผลกระทบจากการเชื่อมโยงข้อมูล (Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources) [17] ใช้การทำเหมืองข้อความเพื่อช่วยตรวจหาผู้ป่วยที่มีผลการตรวจโรคเป็นบวก โดยการใช้ข้อมูลจากแหล่งข้อมูลต่าง ๆ ได้แก่ ข้อมูลทางรังสีวิทยา ข้อมูลทางพยาธิวิทยา และข้อมูลการเข้า

รักษา ซึ่งผลลัพธ์ที่ได้แสดงให้เห็นว่าการเชื่องโยงข้อมูลจากแหล่งข้อมูลต่าง ๆ เข้าด้วยกัน จะช่วยเพิ่มประสิทธิภาพในการจำแนกประเภท

งานในโรงพยาบาลนอกจากงานด้านการรักษาผู้ป่วยแล้ว ยังมีงานด้านเอกสารต่าง ๆ ที่ต้องดำเนินการต่อ หลังจากที่ผู้ป่วยได้รับการตรวจโรคจากแพทย์แล้ว เช่น การแปลงผลวินิจฉัยโรคให้เป็นรหัสโรคสากลหรือรหัสไอซีดี ซึ่งมีงานวิจัยที่เกี่ยวข้องดังนี้

- งานวิจัยเรื่อง แบบจำลองการจำแนกไอซีดี-เทน ทีเอ็ม ข้ามภาษาโดยใช้เหมืองข้อความ [3] ใช้การทำเหมืองข้อความเพื่อสร้างแบบจำลองสำหรับจำแนกไอซีดี-เทน ทีเอ็มจากการวิเคราะห์คำที่ปรากฏอยู่ในคำวินิจฉัยของแพทย์ เพื่อช่วยลดความผิดพลาดในการจำแนกไอซีดี-เทน ทีเอ็มและลดระยะเวลาในการทำงานของเจ้าหน้าที่เวชสถิติ
- งานวิจัยเรื่อง การทำเหมืองข้อความเพื่อจำแนกประเภทโรคลมบ้าหมูในเด็กด้วยรหัสไอซีดีไนน์ (ICD9-based Text Mining Approach to Children Epilepsy Classification) [18] ใช้การทำเหมืองข้อความเพื่อแปลงผลวินิจฉัยโรคลมบ้าหมูให้เป็นรหัสไอซีดีไนน์ ทำให้ลดระยะเวลาและลดผลกระทบจากการค้นหาและจัดประเภทผลวินิจฉัยของผู้ป่วยแต่ละราย

2.2.1.3 การทำเหมืองข้อความกับข้อมูลด้านพันธุศาสตร์

ในด้านพันธุศาสตร์มีการนำการทำเหมืองข้อความมาประยุกต์ใช้ เพื่อศึกษาหาความรู้และข้อมูลเกี่ยวกับยีน ดังนี้

- งานวิจัยเรื่อง โรค : การทำเหมืองข้อความและการรวบรวมข้อมูลความสัมพันธ์ระหว่างโรคกับยีน (DISEASES : Text mining and data integration of disease-gene associations) [19] นำการทำเหมืองข้อความมาประยุกต์ใช้เพื่อทำการวิเคราะห์หาความสัมพันธ์ระหว่างโรคกับยีน และสร้างแบบจำลองสำหรับทำนายชื่อโรคจากยีนที่ตรวจพบในผู้ป่วย ซึ่งข้อมูลที่ใช้ในการสร้างแบบจำลองนำมาจากการศึกษาสนิปที่กระจายตัวทั่วจีโนม (Genome wide association study)
- งานวิจัยเรื่อง การทำเหมืองข้อความแบบไร้ตัวชี้นำเพื่อประมวลผลและเพิ่มผลลัพธ์ของจีดับบลิวเอเอส (Unsupervised text mining for assessing and augmenting GWAS results) [20] นำการทำเหมืองข้อความมาจัดการกับข้อมูลแบบไร้การชี้นำ (Unsupervised text) โดยใช้เทคนิคการจัดกลุ่ม เพื่อทำการแบ่งกลุ่มยีนที่มีฟีโนไทป์เหมือนกัน ผลลัพธ์ที่ได้คือความสัมพันธ์ระหว่างยีนในรูปแบบการจัดกลุ่มตามลำดับชั้น

(Hierarchical clustering) ซึ่งจะสามารถอธิบายถึงลักษณะรูปแบบความสัมพันธ์ระหว่างโรคกับยีนได้

2.2.1.4 การทำเหมืองข้อความกับงานด้านอื่น ๆ

นอกเหนือจากงานทางด้านการแพทย์ ชีวการแพทย์ และพันธุศาสตร์แล้ว การทำเหมืองข้อความยังถูกนำมาใช้เพื่อวิเคราะห์ข้อมูลในงานด้านอื่น ๆ อีก เช่น ด้านโซเชียล และด้านวิศวกรรม ดังงานวิจัยต่อไปนี้

- งานวิจัยเรื่อง การใช้ภาษาในทวิตเตอร์ทำนายอัตราการเกิดอาชญากรรม (Language Usage on Twitter Predicts Crime Rates) [21] เป็นการวิเคราะห์ข้อมูลจากในทวิตเตอร์เพื่อทำนายอัตราการเกิดอาชญากรรม โดยวิเคราะห์การใช้ภาษาซึ่งเป็นภาษาเหมาะสมและไม่เหมาะสม (offensive) ที่ปรากฏอยู่ในทวิต เพื่อใช้เป็นแหล่งข้อมูลในการทำนายอัตราการเกิดอาชญากรรม ซึ่งตัวจำแนกประเภทที่นำมาใช้คือ ซัพพอร์ตเวกเตอร์แมชชีน โดยผลลัพธ์ที่ได้พบว่าภาษาที่ใช้บนทวิตที่แตกต่างกัน จะมีอัตราการเกิดอาชญากรรมที่แตกต่างกันด้วย
- งานวิจัยเรื่อง การอนุมานสถานที่ของผู้ใช้ทวิตเตอร์ในระยะ 10 กิโลเมตรด้วยความแม่นยำ (Inferring Twitter User Locations With 10km Accuracy) [22] เป็นการอนุมานสถานที่ของผู้ใช้จากการใช้เนื้อหาของข้อความบนทวิต จากการสร้างการกระจายตัวของคำตามภูมิศาสตร์ต่าง ๆ แล้วนำมาคำนวณหาสถานที่ของผู้ใช้ โดยใช้คำเป็นค่าน้ำหนัก ซึ่งผลลัพธ์ที่ได้พบว่าการคำนวณสถานที่จะแม่นยำในระยะ 10 กิโลเมตรจากสถานที่หลักของผู้ใช้
- งานวิจัยเรื่อง การจำแนกโครงสร้างจุลภาคขั้นสูงด้วยวิธีการทำเหมืองข้อมูล (Advanced microstructure classification by data mining methods) [23] ใช้การทำเหมืองข้อมูลเพื่อจำแนกประเภทโครงสร้างของเหล็กที่แตกต่างกันจากการประเมินค่าตัวแปรทางสัณฐานวิทยา โดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นตัวจำแนกประเภท เนื่องจากสามารถจำแนกประเภทแบบหลายประเภทได้ (Multiclass classification) และให้ผลลัพธ์ในการจำแนกประเภทดีที่สุด

2.2.2 การเตรียมพร้อมข้อมูลก่อนการทำเหมืองข้อความ

การทำเหมืองข้อความเป็นการนำข้อความจำนวนมากที่มีอยู่ในแหล่งข้อมูลมาทำการจำแนกประเภท โดยการจำแนกประเภทข้อความจะต้องแทนเอกสารเป็นคุณลักษณะ ดังนั้นจึง

ต้องสร้างเวกเตอร์ของคำขึ้นมาเพื่อใช้เป็นคุณลักษณะ โดยจะถือว่าทุก ๆ ตำแหน่งของคำที่ปรากฏในเอกสารคือหนึ่งคุณลักษณะ ซึ่งในขั้นตอนการสร้างเวกเตอร์ของคำจะประกอบไปด้วยขั้นตอนย่อย ๆ ดังนี้

1. การดึงข้อมูล (Information text retrieval) เป็นการดึงข้อมูลที่ต้องการออกมาจากแหล่งข้อมูลที่มีอยู่ เพื่อนำข้อมูลที่ได้ไปใช้ในการประมวลผลและจำแนกประเภท ซึ่งข้อมูลที่อยู่ในแหล่งข้อมูลอาจเป็นได้ทั้งข้อมูลแบบมีโครงสร้างและไม่มีโครงสร้าง
2. การคัดกรองและลบข้อมูลที่ไม่สมบูรณ์ทิ้ง เป็นการลบชุดข้อมูลที่มีข้อมูลขาดหายในบางส่วนทิ้ง ซึ่งอาจทำให้เกิดความลำเอียงในการวิเคราะห์ข้อมูล ดังนั้นจึงนำวิธีการคาดคะเนเข้ามาใช้ เพื่อช่วยลดความลำเอียงที่อาจเกิดขึ้นกับชุดข้อมูล โดยการเติมค่าข้อมูลที่ขาดหายลงไปในชุดข้อมูล ด้วยวิธีการคาดคะเน 3 รูปแบบ [24] ดังนี้
   - การคาดคะเนโดยใช้ค่าเฉลี่ย
   - การคาดคะเนโดยการเดาค่าของข้อมูลที่หายไป ด้วยค่าเดียวกันทั้งหมด
   - การคาดคะเนโดยการสร้างความแปรปรวน (variation) เล็กน้อยลงไปในกระบวนการเติมค่าขาดหาย
3. การประมวลผลข้อมูลก่อน เป็นการเตรียมพร้อมข้อมูลก่อนจะนำไปใช้สร้างแบบจำลองเพื่อให้ได้ข้อมูลที่มีคุณภาพมากขึ้น ช่วยให้การทำเหมืองข้อความมีประสิทธิภาพและมีความแม่นยำมากขึ้น [18] โดยการประมวลผลข้อมูลก่อนประกอบด้วยขั้นตอนดังนี้
   - การแปลงข้อมูลให้เป็นหน่วยย่อย (Tokenization) เป็นการแตกข้อความให้เป็นหน่วยย่อยของภาษา เช่น คำ หรือประโยค โดยจะเรียกหน่วยย่อยเหล่านี้ว่า โทเคน (Token) ซึ่งแต่ละโทเคนจะถูกคั่นด้วยช่องว่าง
   - การตัดคำที่ไม่สำคัญออก (Stop words) โดยคำที่ไม่สำคัญคือคำที่ไม่ช่วยเพิ่มประสิทธิภาพในการค้นหาข้อมูล เนื่องจากเป็นคำที่ปรากฏอยู่ในเอกสารทั่วไป
   - การเปลี่ยนรูปคำให้อยู่ในรูปแบบดั้งเดิม (Lemmatization) โดยการตัดส่วนขยายของคำออก เพื่อเพิ่มประสิทธิภาพในการทำดัชนีคำ
   - การกำหนดรูปแบบโครงสร้างของสิ่งที่สนใจตามองค์ความรู้ ซึ่งจะมีลักษณะเป็นโครงสร้างลำดับชั้นของสิ่งที่สนใจ เพื่อนำมาใช้เป็นแนวทางในการจัดการฐานข้อมูล
   - การคัดเลือกคุณลักษณะ (Feature selection) เพื่อคัดเอาเฉพาะข้อมูลที่มีคุณลักษณะที่เกี่ยวข้องกับงานเท่านั้นมาใช้ และแปลงคำให้เป็นชุดของคำ ซึ่งคำที่อยู่ในชุดเดียวกันจะเป็นคำที่อยู่ในหมวดเดียวกัน

   หลังจากทำการประมวลผลข้อมูลก่อนเสร็จแล้ว จะได้ชุดข้อมูลที่พร้อมสำหรับทำการจำแนกประเภทเพื่อสร้างแบบจำลอง โดยการจำแนกประเภทจะแบ่งออกเป็น 2

ขั้นตอน คือการเรียนรู้เพื่อจำแนกประเภทและการจำแนกประเภทตัวอย่างที่ไม่เคยเห็นมาก่อน ซึ่งจะเริ่มต้นจากการแบ่งข้อมูลที่มีอยู่ออกเป็น 2 ส่วน คือ ชุดข้อมูลสอน ใช้สำหรับการเรียนรู้เพื่อจำแนกประเภท และชุดข้อมูลทดสอบ ใช้สำหรับการจำแนกประเภทตัวอย่างที่ไม่เคยเห็นมาก่อน

2.2.3 ตัวจำแนกประเภทที่นิยมใช้ในการทำเหมืองข้อความ

ในงานวิจัยที่มีการนำการทำเหมืองข้อความมาประยุกต์ใช้กับข้อมูลชนิดต่าง ๆ เพื่อหาความสัมพันธ์ของข้อมูลและสร้างแบบจำลองสำหรับจำแนกประเภทข้อมูล การคัดเลือกตัวจำแนกประเภทที่จะนำมาใช้ในการสร้างแบบจำลองถือเป็นส่วนสำคัญของงานวิจัย เพื่อให้ได้แบบจำลองที่มีประสิทธิภาพและเหมาะสมกับงานนั้น ๆ ซึ่งจากงานวิจัยต่าง ๆ พบว่าตัวจำแนกประเภทที่นิยมนำมาใช้เพื่อสร้างแบบจำลอง หรือนำมาใช้เพื่อทดสอบหาตัวจำแนกประเภทที่ดีที่สุด ได้แก่ อัลกอริทึมของต้นไม้ตัดสินใจ การเรียนรู้เบส์อย่างง่าย และซัพพอร์ตเวกเตอร์แมชชีน โดยใช้วิธีการสุ่มข้อมูลแบบความเที่ยงตรงโดยการแบ่งชุดข้อมูลเป็น 10 ส่วน เพื่อสร้างและทดสอบประสิทธิภาพของแบบจำลอง [16] [18]

2.2.4 การวัดประสิทธิภาพแบบจำลองที่ได้จากการทำเหมืองข้อความ

การวัดประสิทธิภาพแบบจำลองในงานวิจัยที่เกี่ยวข้องกับการทำเหมืองข้อความจะใช้ชุดข้อมูลทดสอบ ซึ่งเป็นชุดข้อมูลที่ถูกแบ่งมาจากชุดข้อมูลเดียวกับชุดข้อมูลสอน มาเป็นตัววัดผลการทำงานของแบบจำลอง เพื่อนำไปใช้ในการคำนวณหาค่าความเที่ยง (precision) ค่าการระลึกได้ (recall) ค่าความแม่น (accuracy) และค่าประสิทธิภาพโดยรวม (F-measure) ของแบบจำลอง [14] โดยค่าความเที่ยงสามารถคำนวณได้จากจำนวนข้อมูลที่ทำนายประเภทถูกต้องจากจำนวนข้อมูลทั้งหมดที่ทำนายว่าเป็นประเภทที่พิจารณาอยู่ ค่าการระลึกได้คำนวณได้จากจำนวนข้อมูลที่ทำนายประเภทถูกต้อง และค่าประสิทธิภาพโดยรวมคำนวณได้จากค่าเฉลี่ยระหว่างค่าความเที่ยงและค่าการระลึกได้ [15]

2.2.5 สรุปผลของงานวิจัยที่เกี่ยวข้อง

จากงานวิจัยที่เกี่ยวข้องข้างต้นสามารถนำชนิดของชุดข้อมูล ตัวจำแนกประเภทที่ใช้ ตัวชี้วัด และผลลัพธ์ของแต่ละงานวิจัยมาเปรียบเทียบกันได้ ดังตารางที่ 1

ตารางที่ 1 ตารางเปรียบเทียบผลของงานวิจัยที่เกี่ยวข้อง

ลำดับ	งานวิจัย	ชุดข้อมูล	ตัวจำแนกประเภท	ตัวชี้วัด	ผลลัพธ์
1	แบบจำลองการจำแนกไอซีดี-เทน ทีเอ็ม ข้ามภาษาโดยใช้เหมืองข้อความ [3]	ผลคำวินิจฉัย	1. การเรียนรู้เบส์อย่างง่าย 2. ซัพพอร์ตเวกเตอร์แมชชีน 3. ต้นไม้ตัดสินใจ	1. ค่าความเที่ยง 2. ค่าการระลึกได้ 3. ค่าความแม่น	การเรียนรู้เบส์อย่างง่าย ให้ค่าความแม่นสูงสุดร้อยละ 81.41
2	การทำเหมืองข้อความเอจายล์สำหรับปัจจัยเสี่ยงของโรคหัวใจ [11]	รายงานผู้ป่วย	1. การเรียนรู้เบส์อย่างง่าย 2. ซัพพอร์ตเวกเตอร์แมชชีน 3. ต้นไม้ตัดสินใจ	1. ค่าความแม่น 2. ค่าประสิทธิภาพโดยรวม	ต้นไม้ตัดสิน ให้ค่าความแม่นสูงสุดร้อยละ 82.9 และค่าประสิทธิภาพโดยรวมสูงสุดร้อยละ 91.7
3	การสกัดข้อมูลจากรายงานทางพยาธิวิทยาภายในโรงพยาบาล [12]	รายงานทางพยาธิวิทยา	1. การเรียนรู้เบส์อย่างง่าย 2. ซัพพอร์ตเวกเตอร์แมชชีน 3. เอดาบูสต์ (Adaboost)	ค่าประสิทธิภาพโดยรวม	การเรียนรู้เบส์อย่างง่าย ให้ค่าประสิทธิภาพโดยรวมสูงสุดร้อยละ 82.3
4	การจำแนกประเภทโรคอัลไซเมอร์โดยใช้การจัดอันดับคุณลักษณะจากเอมอาร์ไอ [13]	ภาพถ่ายเอมอาร์ไอ	ซัพพอร์ตเวกเตอร์แมชชีน	1. ค่าความแม่น 2. ค่าความไว (sensitivity) 3. ค่าความจำเพาะ (specificity)	1. ค่าความแม่นร้อยละ 96.32 2. ค่าความไวร้อยละ 94.11 3. ค่าความจำเพาะร้อยละ 98.52
5	การทำเหมืองข้อความเพื่อทำนายการเข้ารับการรักษาในโรงพยาบาลโดยใช้เวชระเบียนเบื้องต้นในแผนกฉุกเฉิน [16]	เวชระเบียนผู้ป่วยในแผนกฉุกเฉิน	1. ป่าสุ่ม (Random Forest) 2. ต้นไม้สุ่มตัวอย่างมาก (Extremely Randomized Tree) 3. เอดาบูสต์ 4. การถดถอยโลจิสติก (Logistic Regression)	ค่าประสิทธิภาพโดยรวม	ซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าประสิทธิภาพโดยรวมสูงสุดร้อยละ 77.70

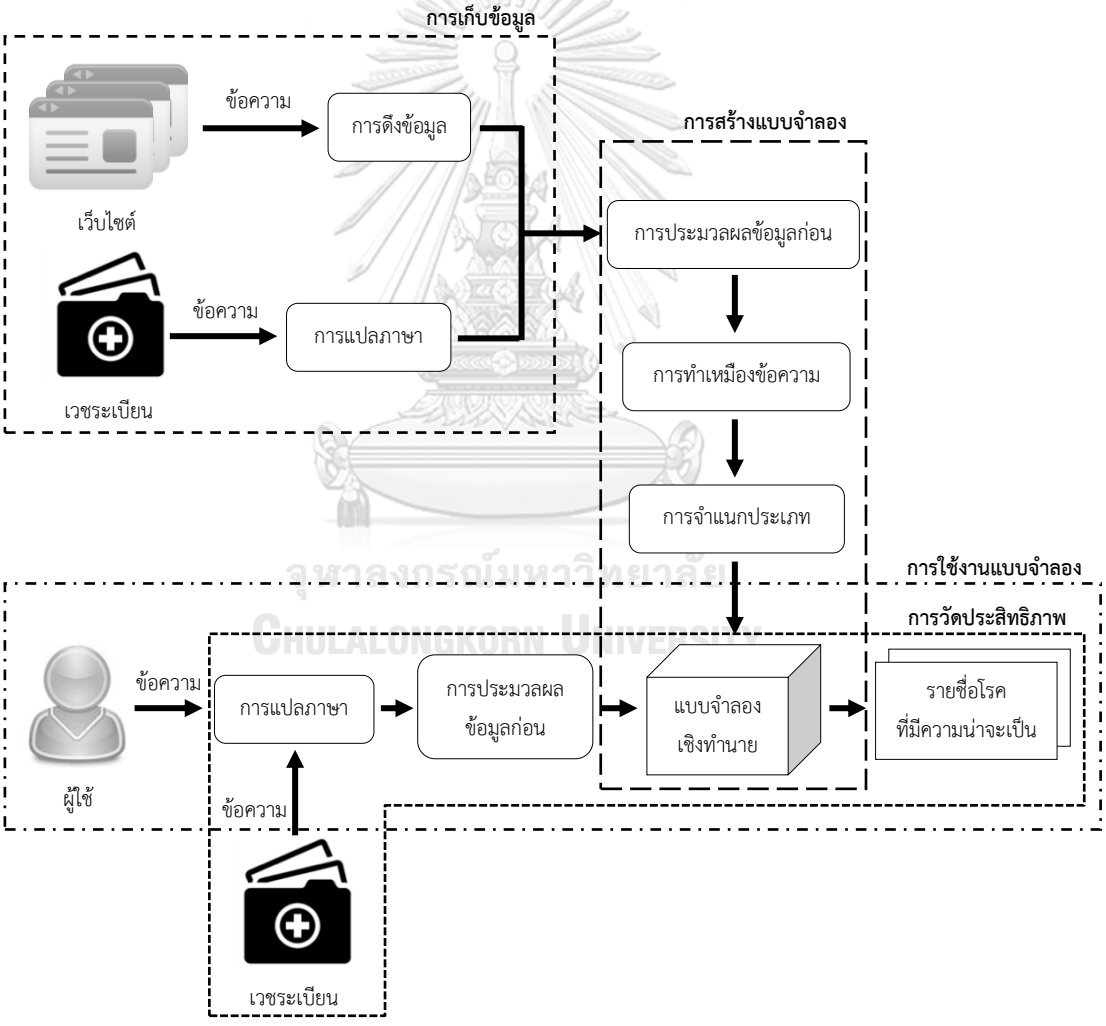
ลำดับ	งานวิจัย	ชุดข้อมูล	ตัวจำแนกประเภท	ตัวชี้วัด	ผลลัพธ์
			5. การเรียนรู้เบส์อย่างง่าย 6. ซัพพอร์ตเวกเตอร์แมชชีน		
6	การทำเหมืองข้อความจากเวชระเบียนอิเล็กโทรนิกเพื่อจำแนกประเภทการเข้ารับการรักษาโรค : การวัดผลกระทบจากการเชื่อมโยงข้อมูล [17]	เวชระเบียน 1. ข้อมูลรังสีวิทยา 2. ข้อมูลพยาธิวิทยา 3. ข้อมูลผู้ป่วย	ซัพพอร์ตเวกเตอร์แมชชีน	1. ค่าความเที่ยง 2. ค่าการระลึกได้ 3. ค่าประสิทธิภาพโดยรวม	ค่าประสิทธิภาพโดยรวมร้อยละ 92.8
7	การทำเหมืองข้อความเพื่อจำแนกประเภทโรคลมบ้าหมูในวัยเด็กด้วยรหัสไอซีดีไนน์ [18]	เวชระเบียน	วิธีการค้นหาเพื่อนบ้านใกล้สุด K ตัว (K-Nearest Neighbor)	ค่าประสิทธิภาพโดยรวม	ค่าประสิทธิภาพโดยรวมร้อยละ 71.05
8	โรค : การทำเหมืองข้อความและการรวบรวมข้อมูลความสัมพันธ์ระหว่างโรคกับยีน [19]	ข้อมูลการศึกษาสนิปที่กระจายตัวทั่วจีโนม	การรู้จำชื่อโดยใช้พจนานุกรม (Dictionary-based name entity recognition)	1. ค่าความจำเพาะ 2. อัตราผลบวกเท็จ	อัตราผลบวกเท็จร้อยละ 0.16
9	การใช้ภาษาในทวิตเตอร์ทำนายอัตราการเกิดอาชญากรรม [21]	ข้อมูลทวิตที่เปิดเป็นสาธารณะ	ซัพพอร์ตเวกเตอร์แมชชีน	ค่าความแม่น	ค่าความแม่นมากกว่าร้อยละ 95
10	การอนุมานสถานที่ของผู้ใช้ทวิตเตอร์ในระยะ 10 กิโลเมตรด้วยความแม่นยำ [22]	ข้อมูลทวิตในเกาหลี	การเรียนรู้เบส์อย่างง่าย	ค่าความแม่น	ค่าความแม่นร้อยละ 56.7

2.2.6 สรุป

จากงานวิจัยที่เกี่ยวข้อง แสดงให้เห็นว่าข้อมูลทางการแพทย์ที่มีอยู่ทั้งในโรงพยาบาลและแหล่งข้อมูลเปิดต่าง ๆ สามารถนำมาใช้ให้เกิดประโยชน์ได้ในหลากหลายด้าน โดยใช้การทำเหมืองข้อความมาวิเคราะห์หาโครงสร้างและความสัมพันธ์ต่าง ๆ ที่ถูกซ่อนอยู่ในข้อมูล ดังนั้นงานวิจัยนี้จึงมีแนวคิดที่จะนำการทำเหมืองข้อความมาประยุกต์ใช้ เพื่อวิเคราะห์หาความสัมพันธ์ระหว่างโรคและอาการของผู้ป่วยในแผนกออโธปิดิกส์ของโรงพยาบาลจุฬาลงกรณ์ และสร้างแบบจำลองขึ้นมา เพื่อใช้เป็นตัวช่วยแพทย์ในการวินิจฉัยโรคจากข้อมูลอาการของผู้ป่วย โดยการสร้างแบบจำลองจะนำอัลกอริทึมต่าง ๆ ที่ได้รับความนิยมในการทำเหมืองข้อความ ได้แก่ อัลกอริทึมของต้นไม้ตัดสินใจ การเรียนรู้เบส์อย่างง่าย ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม มาใช้เป็นตัวจำแนกประเภทในการสร้างแบบจำลอง จากนั้นนำแบบจำลองที่ได้จากตัวจำแนกประเภทแต่ละชนิดมาเปรียบเทียบกัน เพื่อหาตัวจำแนกประเภทที่เหมาะสมที่สุดสำหรับงานวิจัยนี้

บทที่ 3
แนวคิดและวิธีการดำเนินงาน

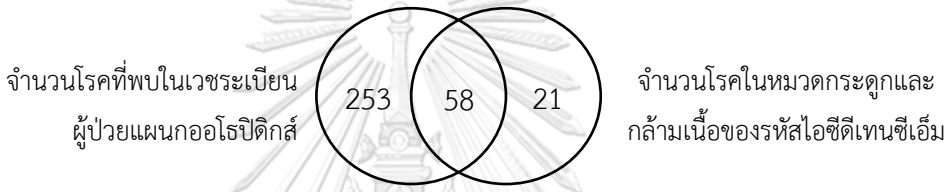
งานวิจัยนี้มีเป้าหมายเพื่อสร้างแบบจำลองสำหรับจำแนกประเภทโรค โดยใช้ข้อมูลจากเว็บไซต์สาธารณะร่วมกับข้อมูลจากเวชระเบียน เพื่อสร้างแบบจำลองด้วยตัวจำแนกประเภทชนิดต่าง ๆ และนำแบบจำลองที่ได้มาประเมินประสิทธิภาพด้วยตัวชี้วัด เพื่อหาแบบจำลองที่ดีที่สุดสำหรับนำไปให้แพทย์ใช้งาน ซึ่งองค์ประกอบของแนวคิดสามารถแบ่งได้เป็น 4 ส่วนหลัก ๆ ได้แก่ การเก็บข้อมูล การสร้างแบบจำลอง การใช้งานแบบจำลอง และการวัดประสิทธิภาพ ดังรูปที่ 8



รูปที่ 8 ขั้นตอนการดำเนินงาน

3.1 การเก็บข้อมูล

ข้อมูลที่ใช้ในการสร้างแบบจำลองจะเก็บมาจาก 2 แหล่งข้อมูล แหล่งที่ 1 คือเวชระเบียน ประกอบด้วยข้อมูลในส่วนบันทึกของแพทย์และผลการวินิจฉัยโรค ซึ่งข้อมูลในส่วนนี้จะถูกบันทึกเป็นภาษาไทยปนภาษาอังกฤษ ดังนั้นในการเก็บข้อมูลจะทำการแปลข้อมูลให้เป็นภาษาอังกฤษทั้งหมด และแหล่งที่ 2 คือเว็บไซต์สาธารณะ ประกอบด้วยข้อมูลรายละเอียดของโรคและชื่อโรค โดยการเก็บข้อมูลจากเว็บไซต์จะใช้ชื่อโรคที่พบในเวชระเบียนผู้ป่วยแผนกออโธปิดิกส์และชื่อโรคในหมวดกระดูกและกล้ามเนื้อของรหัสไอซีดีเทนซีเอ็มมาเป็นตัวค้นหา ซึ่งมีจำนวนทั้งหมด 332 โรค ดังรูปที่ 9 และใช้โมดูลพอมพ์ (Pomp) [25] เพื่อทำการดึงข้อมูลที่ปรากฏอยู่บนหน้าเว็บไซต์ โดยในงานวิจัยนี้จะดึงข้อมูลจากเว็บไซต์ที่เป็นภาษาอังกฤษเท่านั้น



รูปที่ 9 จำนวนโรคทั้งหมดที่นำมาใช้ในงานวิจัย

จากรูปที่ 9 จะพบว่าโรคที่พบในเวชระเบียนผู้ป่วยมีจำนวนทั้งหมด 311 โรค ซึ่งเป็นโรคที่อยู่ในหมวดกระดูกและกล้ามเนื้อของรหัสไอซีดีเทนซีเอ็มจำนวน 58 โรค โดยโรคทั้งหมดที่อยู่ในหมวดกระดูกและกล้ามเนื้อมีจำนวน 79 โรค จึงมีโรคอีกจำนวน 21 โรค ที่ยังไม่เคยพบในผู้ป่วยแผนกออโธปิดิกส์ ดังนั้นเพื่อให้ได้แบบจำลองที่มีประสิทธิภาพ งานวิจัยนี้จึงทำการเปรียบเทียบแบบจำลองที่สร้างจากองค์ประกอบของข้อมูลที่แตกต่างกันทั้งหมด 3 กรณี ดังนี้

กรณีที่ 1 แบบจำลองที่ใช้ข้อมูลจากเวชระเบียนทั้งหมดร่วมกับข้อมูลจากเว็บไซต์ของโรค 21 โรค ที่ยังไม่เคยพบในผู้ป่วยแผนกออโธปิดิกส์

กรณีที่ 2 แบบจำลองที่ใช้ข้อมูลจากเวชระเบียนทั้งหมดร่วมกับข้อมูลจากเว็บไซต์ของโรคทั้งหมดหรือโรคจำนวน 332 โรค

กรณีที่ 3 แบบจำลองที่ใช้ข้อมูลจากเวชระเบียนทั้งหมด

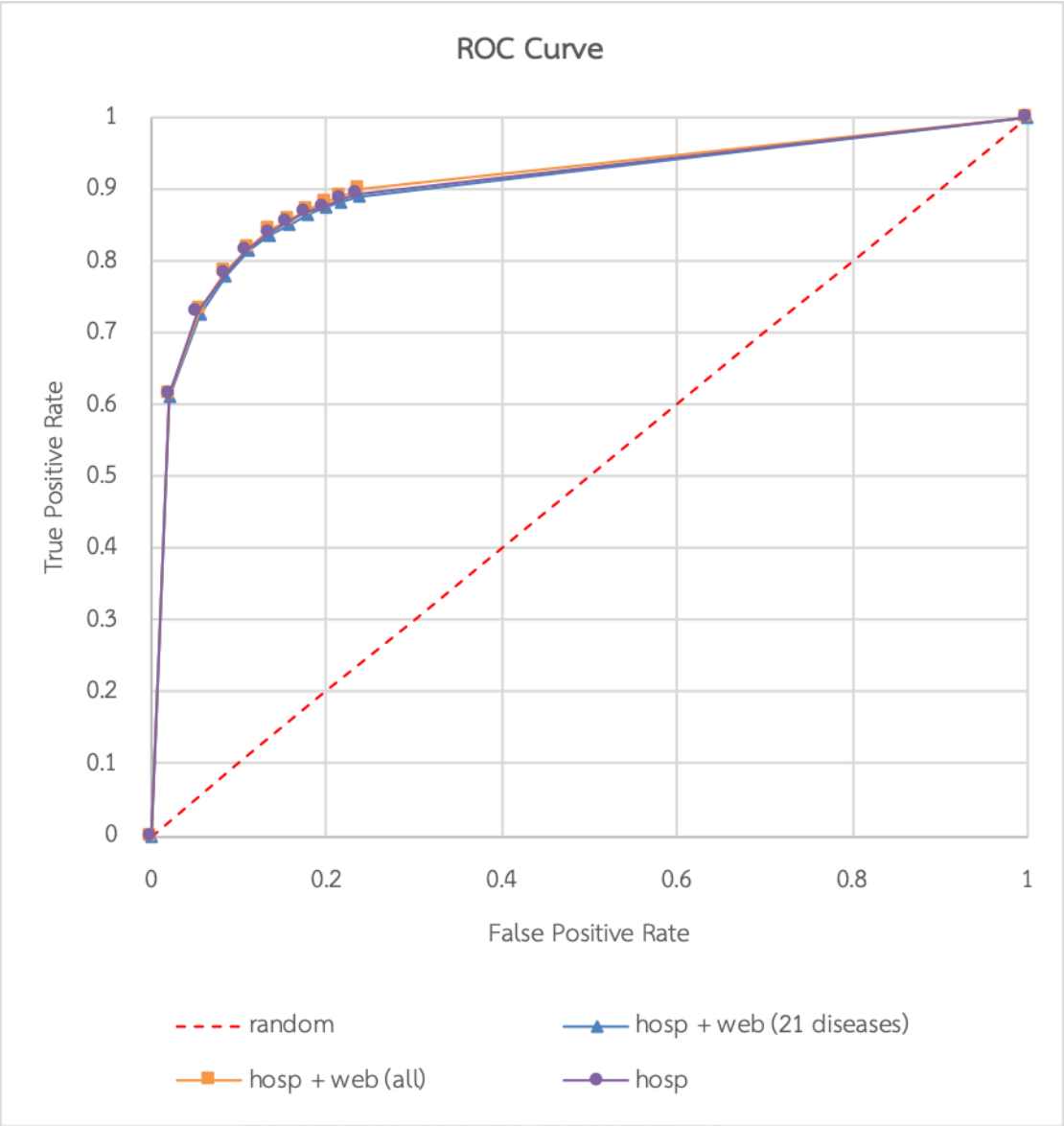
โดยชุดข้อมูลทดสอบจะใช้ข้อมูลจากเวชระเบียน ส่วนชุดข้อมูลทดสอบสำหรับ 21 โรคที่ยังไม่เคยพบในเวชระเบียน จะใช้คำสำคัญของแต่ละโรคที่ได้จากข้อมูลบนเว็บไซต์มาเป็นชุดข้อมูลทดสอบ (ตัวอย่างการทดสอบแสดงในภาคผนวก ฉ) ซึ่งในการเปรียบเทียบแบบจำลองที่ได้จากทั้ง 3 กรณี จะใช้กราฟเส้นโค้งอาร์โอซี (Receiver Operating Characteristic Curve: ROC curve) [26] เข้ามาช่วยในการพิจารณา เพื่อหาองค์ประกอบของข้อมูลที่เหมาะสมสำหรับการสร้างแบบจำลอง

กราฟเส้นโค้งอาร์โอซีเป็นตัววัดประสิทธิภาพการทดสอบ เพื่อบอกว่าการทดสอบใดมีประสิทธิภาพในการจำแนกประเภทโรคสูงที่สุดภายใต้สถานการณ์เดียวกัน โดย 1 เส้นโค้งในกราฟจะแทนการทดสอบ 1 กรณี ซึ่งการทดสอบแต่ละกรณีจะแตกต่างกันไปตามค่าตัวแปร

เส้นโค้งของการทดสอบที่มีประสิทธิภาพสูงสุดจากกราฟเส้นโค้งอาร์โอซี คือเส้นโค้งที่มีความชันสูงสุด หรือเป็นเส้นโค้งที่อยู่ใกล้มุมบนซ้ายของกราฟมากที่สุด เนื่องจากเป็นการทดสอบที่ให้อัตราผลบวกจริง (True Positive Rate) มากที่สุด และให้อัตราผลบวกเท็จ (False Positive Rate) น้อย โดยการเปรียบเทียบพื้นที่ใต้เส้นโค้งของกราฟ (Area Under Curve: AUC) ดูจากพื้นที่ใต้เส้นโค้งที่มากกว่าจะแสดงถึงประสิทธิภาพที่สูงกว่า

กราฟเส้นโค้งอาร์โอซีในงานวิจัยนี้ แต่ละจุดบนเส้นโค้งจะแทนจำนวนชื่อโรคที่แสดงอยู่ในผลลัพธ์ของแบบจำลอง กล่าวคือ จุดที่ n บนเส้นโค้ง หมายถึงจุดที่แบบจำลองแสดงชื่อโรค n โรคในผลลัพธ์ ซึ่งในกรณีที่มีชื่อโรคที่ถูกต้องอยู่ในผลลัพธ์ที่แบบจำลองแสดง จะถือว่าแบบจำลองทำนายชื่อโรคได้ถูกต้อง

รูปที่ 10 คือกราฟเส้นโค้งอาร์โอซีที่ประกอบด้วยเส้นโค้ง 3 เส้น แต่ละเส้นแทนองค์ประกอบของข้อมูล 1 กรณี โดยเส้นโค้งทุกเส้นจะถูกสร้างจากตัวจำแนกประเภทชนิดเดียวกัน เพื่อหาองค์ประกอบของข้อมูลที่เหมาะสมสำหรับนำมาใช้ในการสร้างแบบจำลอง



รูปที่ 10 กราฟเปรียบเทียบแบบจำลองที่สร้างโดยใช้องค์ประกอบของข้อมูลที่แตกต่างกัน

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 10 จะเห็นว่าเส้นโค้งที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด เป็นเส้นโค้งที่ได้จากแบบจำลองที่สร้างโดยใช้ข้อมูลจากเวชระเบียนทั้งหมดร่วมกับข้อมูลจากเว็บไซต์ของโรคทั้งหมด ดังนั้นในงานวิจัยนี้จึงเลือกใช้องค์ประกอบของข้อมูลในกรณีที่ 2 เพื่อสร้างแบบจำลอง ซึ่งชื่อโรคที่นำมาใช้ในการสร้างแบบจำลองจะมีจำนวนทั้งหมด 332 โรค โดยสามารถนำมาแจกแจงตามหมวดหมู่โรคในรหัสไอซีดีเทนซีเอ็มได้ดังตารางที่ 2

ตารางที่ 2 ตารางแสดงจำนวนโรคในแต่ละหมวดหมู่

ลำดับ	หมวดหมู่โรค	จำนวนโรค
1	Certain infectious and parasitic diseases	10
2	Neoplasms	39
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	6
4	Endocrine, nutritional and metabolic diseases	8
5	Mental, Behavioral and Neurodevelopmental disorders	2
6	Diseases of the nervous system	22
7	Diseases of the eye and adnexa	2
8	Diseases of the ear and mastoid process	2
9	Diseases of the circulatory system	20
10	Diseases of the respiratory system	3
11	Diseases of the digestive system	5
12	Diseases of the skin and subcutaneous tissue	15
13	Diseases of the musculoskeletal system and connective tissue	79
14	Diseases of the genitourinary system	4
15	Congenital malformations, deformations and chromosomal abnormalities	18
16	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	5
17	Injury, poisoning and certain other consequences of external causes	82
18	Factors influencing health status and contact with health services	10
	รวม	332

ข้อมูลที่เก็บมาจากทั้ง 2 แหล่งข้อมูล จะถูกนำมาผ่านการคัดกรองข้อมูล เพื่อตัดข้อมูลส่วนที่ไม่สมบูรณ์หรือไม่เป็นประโยชน์ต่อการจำแนกประเภททิ้ง ก่อนที่จะนำข้อมูลไปใช้ในการสร้างแบบจำลอง โดยการคัดกรองข้อมูลจะแบ่งออกเป็น 2 แบบตามชนิดของแหล่งข้อมูล ดังนี้

1. การคัดกรองข้อมูลที่มาจากเวชระเบียน จะทำการคัดเวชระเบียนที่มีข้อมูลในส่วนบันทึกของแพทย์เป็นเพียงการนัดหมายผู้ป่วย หรือไม่มีการบันทึกอาการของผู้ป่วยทิ้ง เพราะไม่สามารถนำเวชระเบียนเหล่านี้มาใช้เป็นข้อมูลเพื่อจำแนกประเภทโรคได้
2. การคัดกรองข้อมูลที่มาจากเว็บไซต์สาธารณะ จะทำการคัดข้อมูลที่มาจากเว็บไซต์ที่ลงท้ายด้วย .doc .ppt .xlsx และ .pdf ทิ้ง เนื่องจากตัวโมดูลที่นำมาใช้ในการดึงข้อมูลสามารถดึงข้อมูลได้จากบนหน้าเว็บเท่านั้น ไม่สามารถดึงข้อมูลจากไฟล์ที่ดาวน์โหลดมาได้

หลังจากทำการคัดกรองข้อมูลแล้ว จะได้ข้อมูลจากเวชระเบียนที่มีคุณสมบัติเหมาะสมจำนวน 13,521 ระเบียน จากเวชระเบียนทั้งหมด 16,261 ระเบียน และข้อมูลจากเว็บไซต์ทั้งหมด 3,320 เว็บไซต์ โดยจำนวนข้อมูลที่ได้มาจากเวชระเบียนและเว็บไซต์ สามารถคำนวณเป็นร้อยละแยกตามหมวดหมู่ของโรคได้ดังตารางที่ 3 และตารางที่ 4 ตามลำดับ

ตารางที่ 3 ตารางแสดงร้อยละของจำนวนข้อมูลจากเวชระเบียนแยกตามหมวดหมู่

ลำดับ	หมวดหมู่โรค	ร้อยละ
1	Certain infectious and parasitic diseases	0.49%
2	Neoplasms	4.27%
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	0.06%
4	Endocrine, nutritional and metabolic diseases	0.16%
5	Mental, Behavioral and Neurodevelopmental disorders	0.01%
6	Diseases of the nervous system	1.74%
7	Diseases of the eye and adnexa	0.01%
8	Diseases of the ear and mastoid process	0.01%
9	Diseases of the circulatory system	0.22%
10	Diseases of the respiratory system	0.01%
11	Diseases of the digestive system	0.03%
12	Diseases of the skin and subcutaneous tissue	0.84%
13	Diseases of the musculoskeletal system and connective tissue	55.73%

ลำดับ	หมวดหมู่โรค	ร้อยละ
14	Diseases of the genitourinary system	0.01%
15	Congenital malformations, deformations and chromosomal abnormalities	2.51%
16	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	0.47%
17	Injury, poisoning and certain other consequences of external causes	32.28%
18	Factors influencing health status and contact with health services	1.15%
	รวม	100%

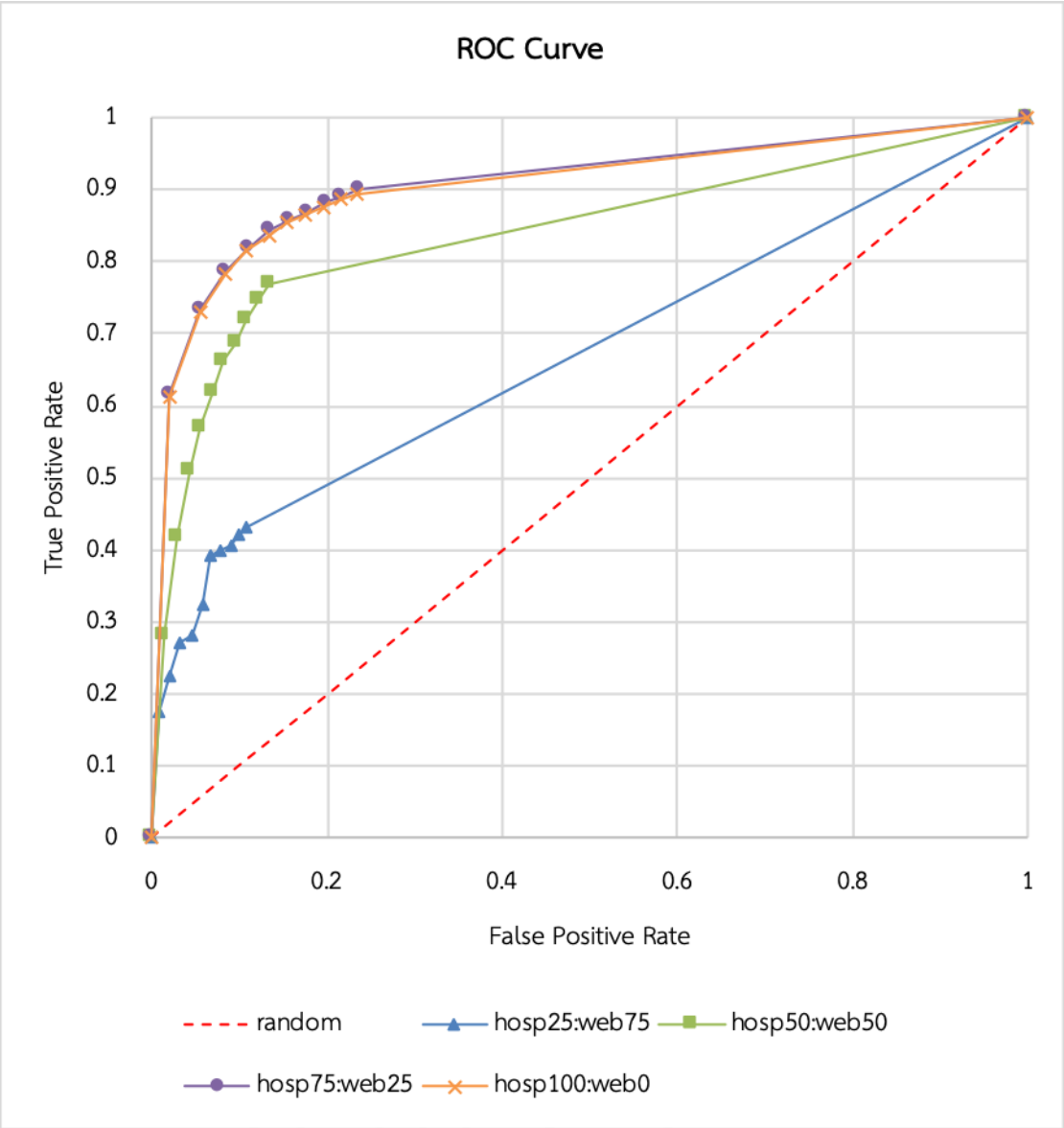
ตารางที่ 4 ตารางแสดงร้อยละของจำนวนข้อมูลจากเว็บไซต์สาธารณะแยกตามหมวดหมู่

ลำดับ	หมวดหมู่โรค	ร้อยละ
1	Certain infectious and parasitic diseases	3.01%
2	Neoplasms	11.75%
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	1.81%
4	Endocrine, nutritional and metabolic diseases	2.41%
5	Mental, Behavioral and Neurodevelopmental disorders	0.60%
6	Diseases of the nervous system	6.63%
7	Diseases of the eye and adnexa	0.60%
8	Diseases of the ear and mastoid process	0.60%
9	Diseases of the circulatory system	6.02%
10	Diseases of the respiratory system	0.90%
11	Diseases of the digestive system	1.51%
12	Diseases of the skin and subcutaneous tissue	4.52%
13	Diseases of the musculoskeletal system and connective tissue	23.80%
14	Diseases of the genitourinary system	1.20%
15	Congenital malformations, deformations and chromosomal	5.42%

ลำดับ	หมวดหมู่โรค	ร้อยละ
	abnormalities	
16	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	1.51%
17	Injury, poisoning and certain other consequences of external causes	24.70%
18	Factors influencing health status and contact with health services	3.01%
	รวม	100%

เนื่องจากชุดข้อมูลที่จะนำมาใช้เป็นชุดข้อมูลสอนได้มาจาก 2 แหล่งข้อมูล ดังนั้นจึงต้องมีการกำหนดสัดส่วนของข้อมูลจากแต่ละแหล่ง โดยในการกำหนดสัดส่วนของข้อมูลจะนำกราฟเส้นโค้งอาร์โอซีเข้ามาช่วยในการพิจารณา เพื่อหาสัดส่วนที่เหมาะสมระหว่าง 2 แหล่งข้อมูล

รูปที่ 11 คือกราฟเส้นโค้งอาร์โอซีที่ประกอบด้วยเส้นโค้ง 4 เส้น แต่ละเส้นแทนอัตราส่วนระหว่างข้อมูลจากเวชระเบียนกับข้อมูลจากเว็บไซต์ที่นำมาใช้ในการสร้างแบบจำลองที่แตกต่างกัน โดยเส้นโค้งทุกเส้นจะถูกสร้างจากตัวจำแนกประเภทชนิดเดียวกัน เพื่อหาอัตราส่วนของข้อมูลที่เหมาะสมที่สุดสำหรับนำมาใช้ในการสร้างแบบจำลอง



รูปที่ 11 กราฟเปรียบเทียบแบบจำลองที่สร้างด้วยอัตราส่วนระหว่างข้อมูลจากเวชระเบียนกับข้อมูลจากเว็บไซต์ที่แตกต่างกัน

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 11 จะเห็นว่าเส้นโค้ง hosp75:web25 เป็นเส้นที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด ดังนั้นงานวิจัยนี้จึงใช้อัตราส่วนระหว่างข้อมูลจากเวชระเบียน 75 ส่วนต่อข้อมูลจากเว็บไซต์ 25 ส่วนในการสร้างแบบจำลอง

ความน่าเชื่อถือของข้อมูลที่ได้จากเว็บไซต์จะประเมินโดยการสุ่มตัวอย่างชื่อโรคมาทั้งหมด 50 โรค เพื่อให้แพทย์ผู้เชี่ยวชาญในแผนกออโธปิดิกส์ของโรงพยาบาลจุฬาลงกรณ์ทำการตรวจสอบคำสำคัญของแต่ละโรค ซึ่งการคัดเลือกคำสำคัญจะคัดเฉพาะคำที่มีความถี่สูงสุดใน 10 อันดับแรก โดยในตารางที่ 5 คือผลที่ได้จากการตรวจสอบโดยแพทย์ ซึ่งพบว่าในแต่ละโรคจะมีคำสำคัญที่เกี่ยวข้องกับโรคนั้น ๆ เฉลี่ย 7 ต่อ 10 คำ แสดงให้เห็นว่าข้อมูลที่ได้จากเว็บไซต์มีความน่าเชื่อถือและสามารถนำมาใช้เพื่อสร้างแบบจำลองที่มีประสิทธิภาพได้

ตารางที่ 5 ตารางแสดงผลการตรวจสอบคำสำคัญที่เกี่ยวข้องกับโรค

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส	จำนวนคำสำคัญที่เกี่ยวข้อง
1	M01	Direct infections of joint in infectious and parasitic diseases classified elsewhere	7
2	M02	Postinfective and reactive arthropathies	7
3	M04	Autoinflammatory syndromes	6
4	M07	Enteropathic arthropathies	8
5	M11	Other crystal arthropathies	7
6	M12	Other and unspecified arthropathy	7
7	M13	Other arthritis	4
8	M14	Arthropathies in other diseases classified elsewhere	8
9	M15	Polyosteoarthritis	8
10	M16	Osteoarthritis of hip	3
11	M20	Acquired deformities of fingers and toes	6
12	M21	Other acquired deformities of limbs	8
13	M22	Disorder of patella	4
14	M23	Internal derangement of knee	7
15	M24	Other specific joint derangements	6
16	M25	Other joint disorder, not elsewhere classified	6
17	M27	Other diseases of jaws	7

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส	จำนวนคำสำคัญที่เกี่ยวข้อง
18	M30	Polyarteritis nodosa and related conditions	8
19	M32	Systemic lupus erythematosus	9
20	M33	Dermatopolymyositis	8
21	M34	Systemic sclerosis	8
22	M42	Spinal osteochondrosis	9
23	M45	Ankylosing spondylitis	5
24	M46	Other inflammatory spondylopathies	7
25	M48	Other spondylopathies	6
26	M51	Thoracic thoracolumbar and lumbosacral intervertebral disc disorders	7
27	M53	Other and unspecified dorsopathies, not elsewhere classified	9
28	M60	Myositis	5
29	M63	Disorders of muscle in diseases classified elsewhere	7
30	M65	Synovitis and tenosynovitis	7
31	M66	Spontaneous rupture of synovium and tendon	9
32	M70	Soft tissue disorders related to use overuse and pressure	8
33	M76	Enthesopathies lower limb excluding foot	6
34	M80	Osteoporosis with current pathological fracture	8
35	M83	Adult osteomalacia	9
36	M91	Juvenile osteochondrosis of hip and pelvis	8
37	M95	Other acquired deformities of musculoskeletal system and connective tissue	10

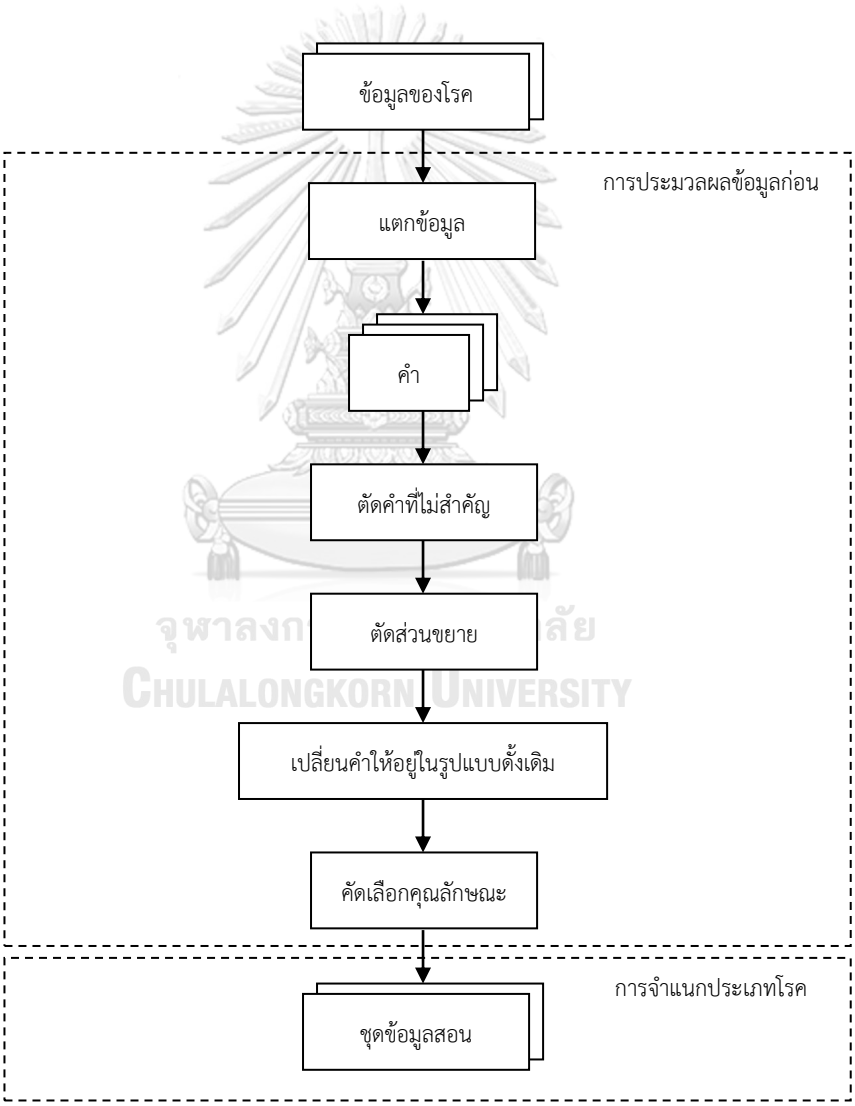
ลำดับ	รหัส ไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส	จำนวนคำสำคัญที่เกี่ยวข้อง
38	M97	Periprosthetic fracture around internal prosthetic joint	9
39	S30	Superficial injury of abdomen lower back pelvis and external genitals	9
40	S33	Dislocation and sprain of joints and ligaments of lumbar spine and pelvis	8
41	S43	Dislocation and sprain of joints and ligaments of shoulder girdle	8
42	S53	Dislocation and sprain of joints and ligaments of elbow	9
43	S54	Injury of nerves at forearm level	7
44	S55	Injury of blood vessels at forearm level	5
45	S56	Injury of muscle fascia and tendon at forearm level	8
46	S58	Traumatic amputation of elbow and forearm	7
47	S61	Open wound of wrist hand and fingers	7
48	S63	Dislocation and sprain of joints and ligaments at wrist and hand level	8
49	S66	Injury of muscle fascia and tendon at wrist and hand level	8
50	S91	Open wound of ankle foot and toes	6
	เฉลี่ย		7

3.2 การสร้างแบบจำลอง

การสร้างแบบจำลองสำหรับจำแนกประเภทโรคจากอาการ จัดเป็นการทำเหมืองข้อมูลที่ให้ผลลัพธ์เป็นแบบจำลองเชิงทำนาย โดยขั้นตอนในการสร้างแบบจำลองจะแบ่งออกเป็น 2 ส่วนหลัก ๆ คือ ส่วนการประมวลผลข้อมูลก่อน และส่วนการจำแนกประเภท

3.2.1 การประมวลผลข้อมูลก่อน

การประมวลผลข้อมูลก่อนประกอบด้วย การแตกข้อมูลออกเป็นหน่วยย่อย การตัดคำที่ไม่สำคัญ การตัดส่วนขยาย การเปลี่ยนคำให้อยู่ในรูปแบบดั้งเดิม และการคัดเลือกคุณลักษณะ ซึ่งสุดท้ายผลลัพธ์ที่ได้จะเป็นชุดข้อมูลที่พร้อมสำหรับนำไปใช้ในการสร้างแบบจำลอง ดังรูปที่ 12

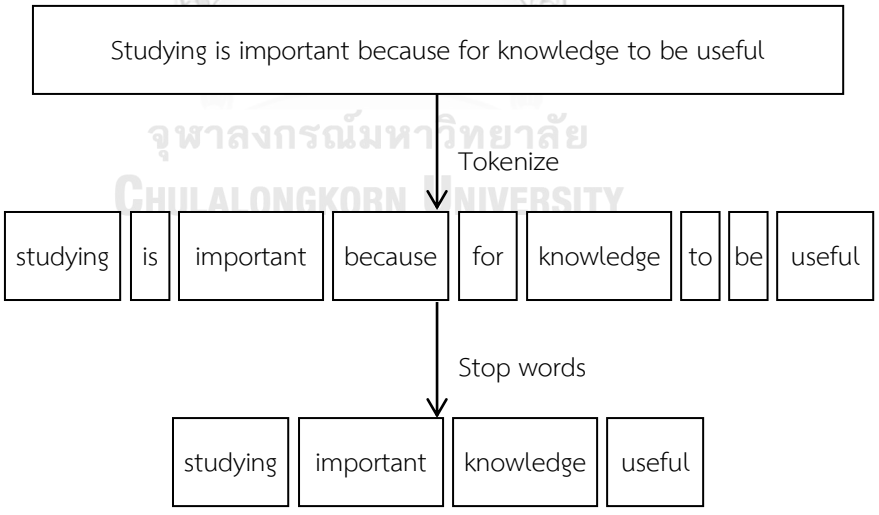


รูปที่ 12 กระบวนการสร้างแบบจำลอง

การประมวลผลข้อมูลก่อนจะใช้เอ็นแอลทีเค (Natural Language Toolkit: NLTK) [27] ซึ่งเป็นเครื่องมือที่ทำให้คอมพิวเตอร์สามารถวิเคราะห์และประมวลผลข้อมูลทางด้านภาษาศาสตร์ได้โดยใช้ไพทอน (Python) โดยมีแหล่งข้อมูลจากเวิร์ดเน็ต (WordNet) และไลบรารีที่เกี่ยวข้องกับการประมวลผลทางภาษา ทำให้สามารถตัดคำที่ไม่สำคัญ จำแนกประเภทคำ ตัดส่วนขยายคำ แปลงคำให้อยู่ในรูปแบบดั้งเดิม และวิเคราะห์ความหมายของคำได้

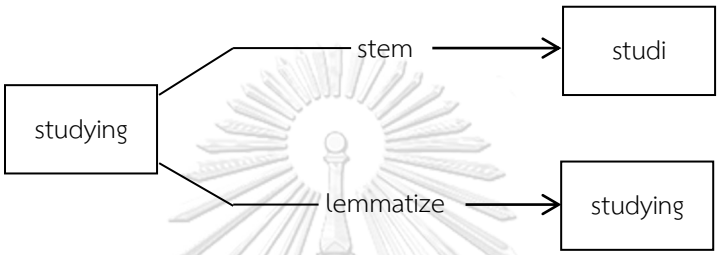
เวิร์ดเน็ตเป็นเครือข่ายของคำที่ถูกเชื่อมโยงด้วยความสัมพันธ์ต่าง ๆ โดยจะแบ่งคำออกเป็น 4 หมวดหมู่ ได้แก่ คำนาม (noun) คำกริยา (verb) คำคุณศัพท์ (adjective) และคำวิเศษณ์ (adverb) นอกจากนี้ยังมีการจัดกลุ่มคำพ้องความหมาย (synonym) หรือจัดกลุ่มคำตามความหมาย เพื่อช่วยบอกความสัมพันธ์ทางความหมายแบบต่าง ๆ เวิร์ดเน็ตจึงจัดเป็นแหล่งข้อมูลสำหรับงานทางด้านภาษาศาสตร์

ขั้นตอนในการประมวลผลข้อมูลก่อน จะเริ่มต้นจากการแตกข้อมูลออกเป็นหน่วยย่อยที่เรียกว่า คำ โดยใช้วรรคตอนเป็นตัวแบ่ง ซึ่งผลลัพธ์ที่ได้จากการแบ่งคำคือกลุ่มของคำ จากนั้นนำกลุ่มของคำที่ได้ไปผ่านการประมวลผลโดยใช้โมดูลสต๊อปเวิร์ดที่มีอยู่ในเอ็นแอลทีเค เพื่อตัดคำที่ไม่สำคัญออก ดังรูปที่ 13 โดยคำที่ไม่สำคัญคือคำที่จะปรากฏอยู่ในเอกสารทั่วไปหรือมีความถี่ของคำสูง เช่น คำที่เป็นคำบุพบท (preposition) คำกริยาช่วย (verb to be) คำสันธาน (conjunction) เป็นต้น



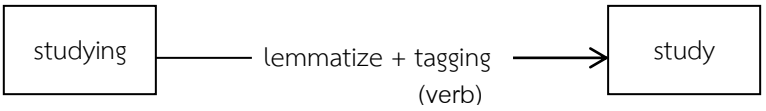
รูปที่ 13 การใช้โมดูลสต๊อปเวิร์ดตัดคำที่ไม่สำคัญ

หลังจากตัดคำที่ไม่สำคัญออก ขั้นตอนต่อมาคือการตัดส่วนขยายของคำ (Stemming) เช่น ตัด s es ing หรือ ed ออก แล้วเปลี่ยนรูปคำให้อยู่ในรูปแบบดั้งเดิม โดยในขั้นตอนนี้มีโมดูลที่สามารถนำมาใช้งานได้ 2 ตัว คือ โมดูลสเตมเมอร์ (Stemmer) และโมดูลเลมมาไทเซอร์ (Lemmatiser) ซึ่งเป็นโมดูลที่มีอยู่ในเอ็นแอลทีเค โดยการทำงานของ 2 โมดูลนี้จะให้ผลลัพธ์ที่คล้ายคลึงกัน แต่โมดูลเลมมาไทเซอร์จะให้ผลลัพธ์ของคำที่มีความถูกต้องตามหลักพจนานุกรมมากกว่า ดังรูปที่ 14 ดังนั้นในงานวิจัยนี้จึงเลือกใช้โมดูลเลมมาไทเซอร์ในการเปลี่ยนคำให้อยู่ในรูปแบบดั้งเดิม



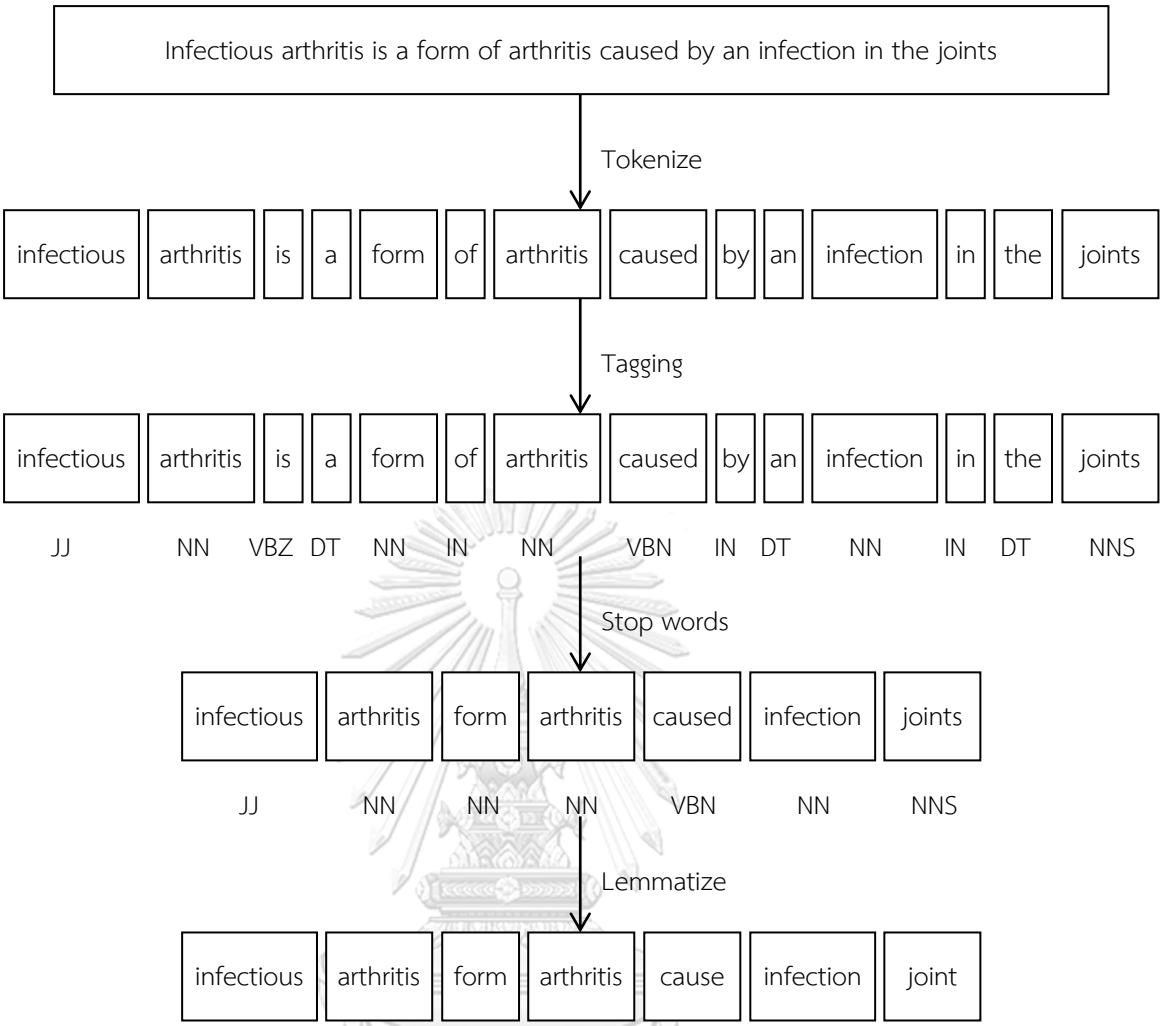
รูปที่ 14 การใช้โมดูลสเตมเมอร์และเลมมาไทเซอร์เปลี่ยนรูปคำ

การใช้โมดูลเลมมาไทเซอร์ในการเปลี่ยนคำจะให้ผลลัพธ์เป็นคำที่มีความถูกต้องตามหลักพจนานุกรมเสมอ แต่ในบางครั้งผลลัพธ์ที่ได้อาจไม่ตรงตามความต้องการเสมอไป เนื่องจากปัจจัยทางด้านบริบทของคำ เช่น ในกรณีที่มองคำว่า studying เป็นคำกริยา แต่ในตัวอย่างรูปที่ 14 จะเห็นว่าผลลัพธ์ที่ได้ยังคงเป็นคำเดิม เนื่องจากตัวโมดูลมองว่าคำนั้นอยู่ในรูปแบบดั้งเดิมของบริบทคำนาม ดังนั้นเพื่อช่วยให้การทำงานของโมดูลเลมมาไทเซอร์มีประสิทธิภาพและได้ผลลัพธ์ตรงตามความต้องการมากขึ้น งานวิจัยนี้จึงเพิ่มการระบุบริบทของคำ (Tagging) เข้าไปในขั้นตอนการเปลี่ยนคำให้อยู่ในรูปแบบดั้งเดิม ดังรูปที่ 15



รูปที่ 15 การใช้โมดูลเลมมาไทเซอร์กับบริบทของคำเปลี่ยนรูปคำ

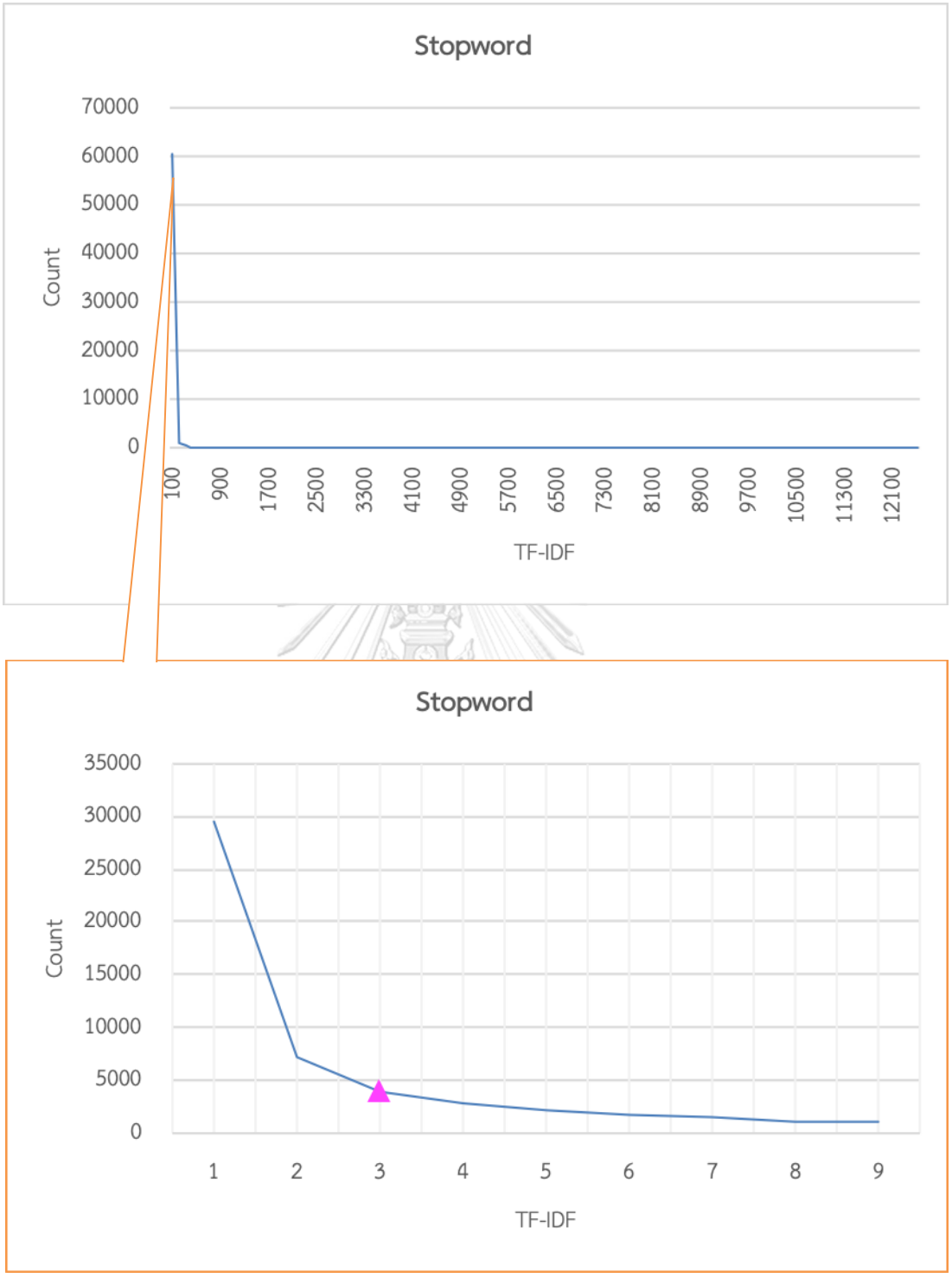
เมื่อทำการประมวลผลข้อมูลก่อนตามลำดับขั้นตอนข้างต้นแล้วจะได้ขั้นตอนการทำงานโดยรวม ดังรูปที่ 16 ซึ่งผลลัพธ์ที่ได้คือกลุ่มคำที่พร้อมนำไปทำการคัดเลือกคุณลักษณะ เพื่อใช้เป็นชุดข้อมูลสอนในการสร้างแบบจำลอง



รูปที่ 16 ขั้นตอนการประมวลผลข้อมูลก่อน

จากรูปที่ 16 หลังจากได้กลุ่มคำที่อยู่ในรูปแบบดั้งเดิมแล้ว จะทำการแปลงอักษรย่อทางการแพทย์ให้เป็นคำเต็ม จากนั้นทำการคัดเลือกข้อมูลที่ประกอบด้วยตัวอักษรเท่านั้นมาใช้เพื่อสร้างแบบจำลอง

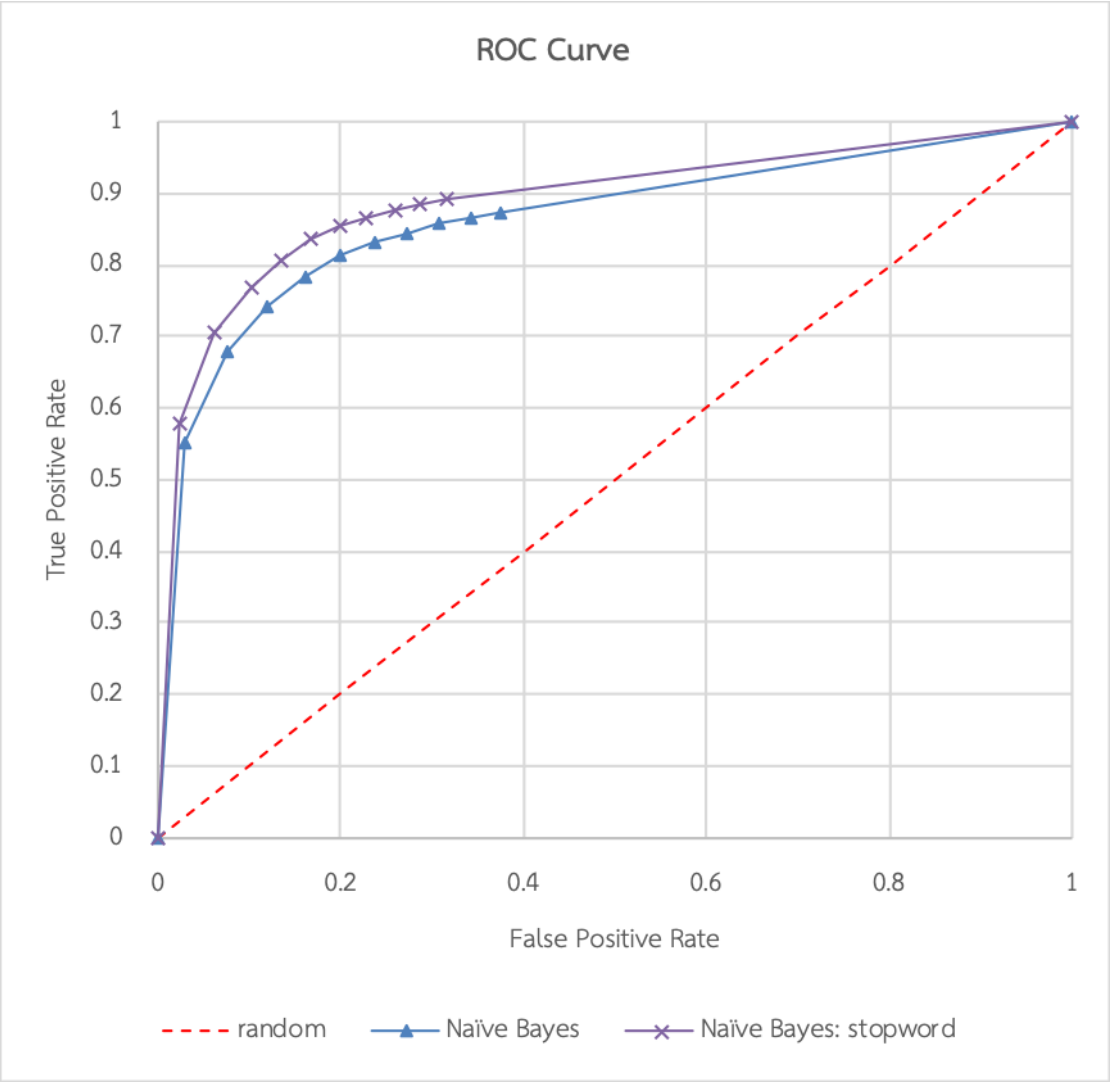
นอกจากนี้เพื่อเป็นการเพิ่มประสิทธิภาพให้กับแบบจำลอง ในงานวิจัยนี้จึงทำการตัดคำที่ไม่สำคัญทางการแพทย์ออก โดยวิเคราะห์จากค่าความถี่-ส่วนกลับของความถี่ของคำ (Term frequency–Inverse document frequency: TF-IDF) และจำนวนของคำในแต่ละช่วงความถี่-ส่วนกลับของความถี่ ซึ่งแสดงเป็นกราฟได้ดังรูปที่ 17



รูปที่ 17 กราฟแสดงจำนวนของคำในแต่ละช่วงความถี่-ส่วนกลับของความถี่ของคำ

จากผลลัพธ์ที่ได้ในกราฟรูปที่ 17 งานวิจัยนี้จะทำการตัดคำที่มีค่าความถี่-ส่วนกลับของความถี่ของคำตั้งแต่ 0 ถึง 3 ออก เพราะถือว่าเป็นคำที่ไม่สำคัญทางการแพทย์ เนื่องจากมีจำนวน

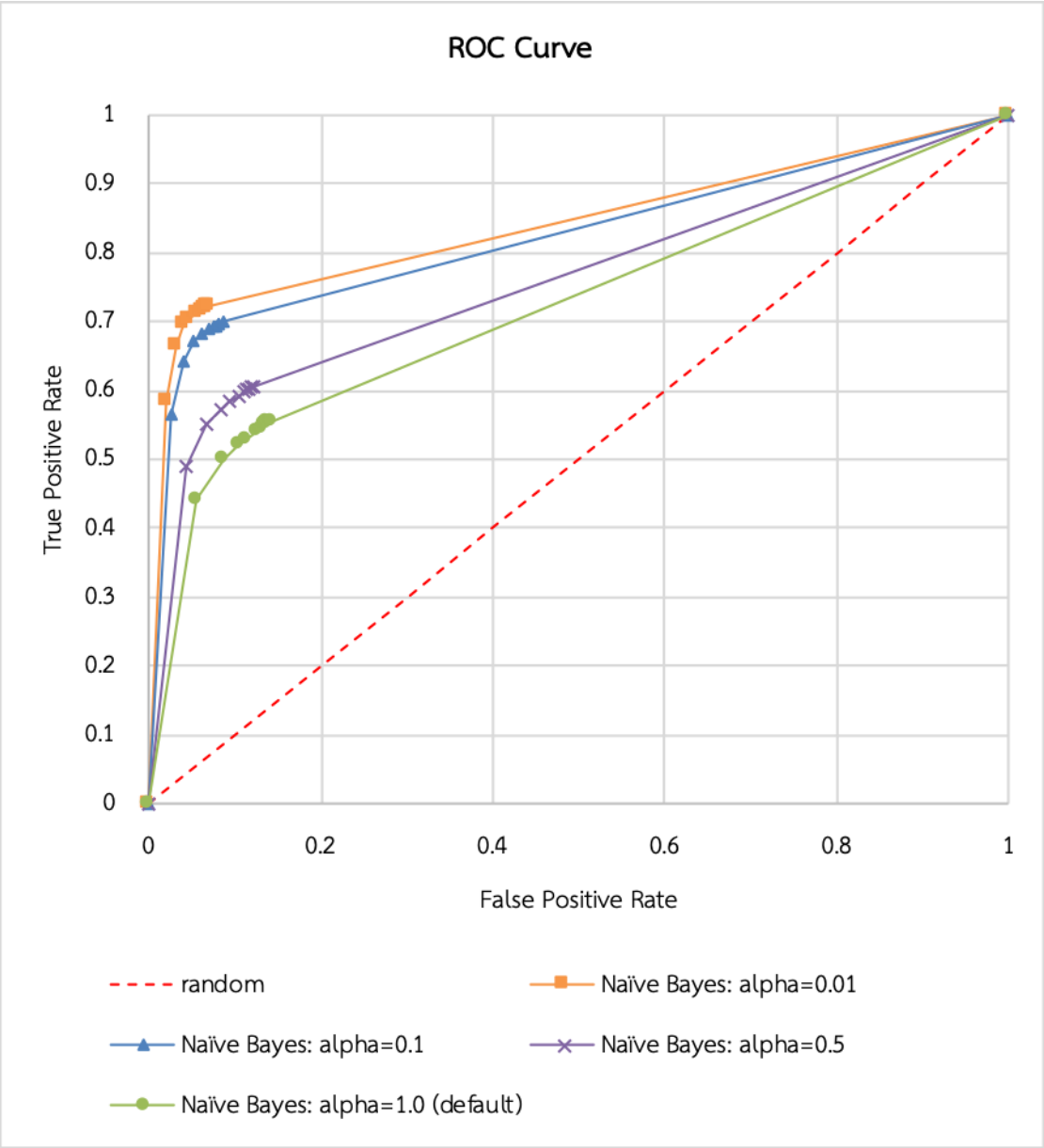
ของคำในช่วงความถี่-ส่วนกลับของความถี่นี้สูง และจากการเปรียบเทียบแบบจำลองที่ได้จากการตัดคำที่ไม่สำคัญทางการแพทย์ออกกับแบบจำลองที่ไม่ตัดคำที่ไม่สำคัญทางการแพทย์ออก ผลลัพธ์ที่ได้แสดงให้เห็นว่าการตัดคำที่ไม่สำคัญทางการแพทย์ออก จะทำให้ได้แบบจำลองที่มีประสิทธิภาพมากขึ้น ดังรูปที่ 18 จะเห็นว่าเส้นโค้ง Naïve Bayes: stopword เป็นเส้นที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด



รูปที่ 18 กราฟเปรียบเทียบแบบจำลองที่สร้างโดยการไม่ตัดและการตัดคำที่ไม่สำคัญทางการแพทย์

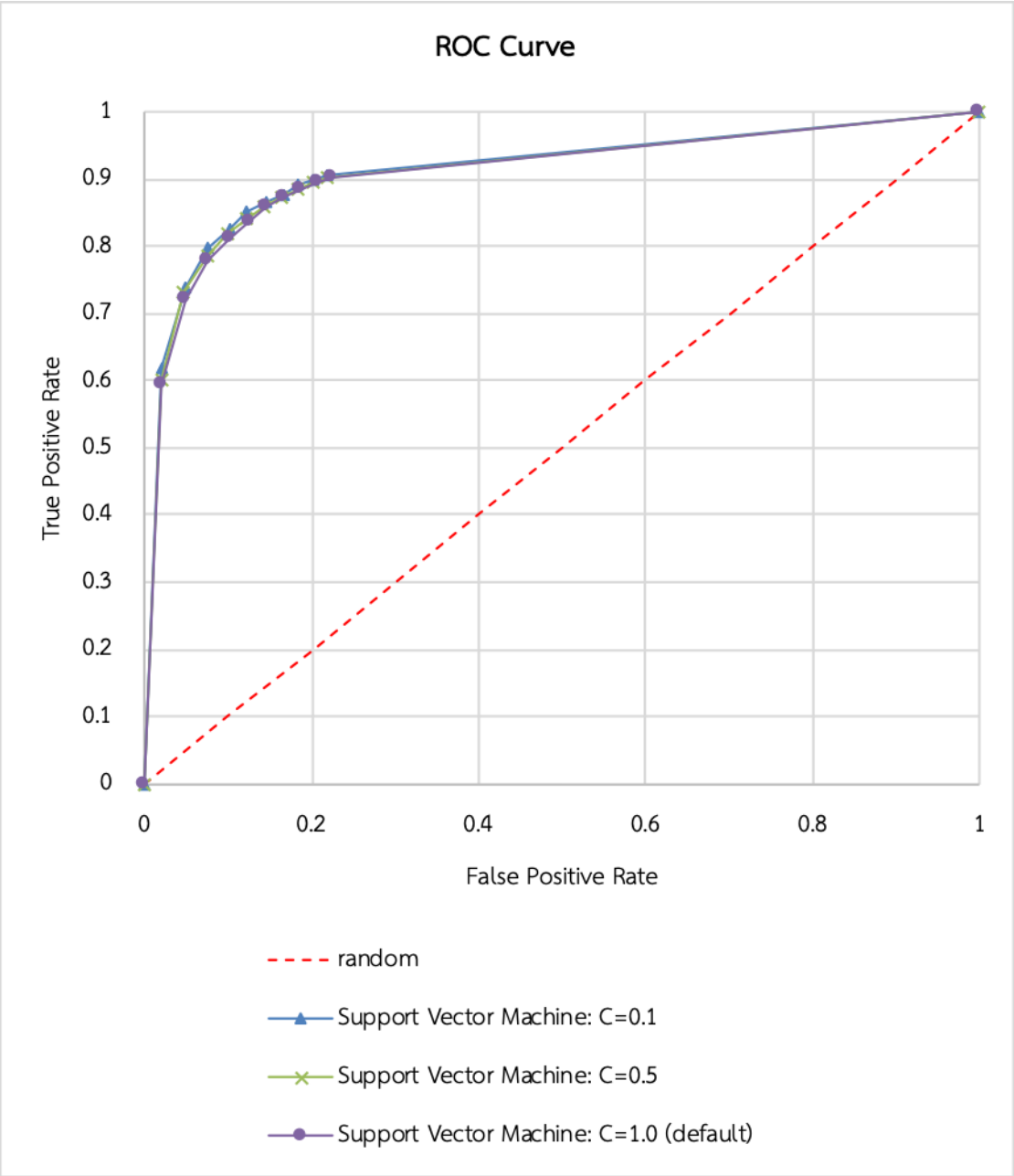
3.2.2 การจำแนกประเภท

การสร้างแบบจำลองจะใช้ตัวจำแนกประเภททั้งหมด 4 ชนิด ได้แก่ ต้นไม้ตัดสินใจ การเรียนรู้เบส์อย่างง่าย ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม มาสร้างแบบจำลองด้วยวิธีการสุ่มข้อมูลแบบความเที่ยงตรงโดยการแบ่งชุดข้อมูลเป็น 10 ส่วน โดยอัลกอริทึมของตัวจำแนกประเภทแต่ละชนิดจะนำมาจากไซคิดเลิร์น (Scikit-learn) [28] ซึ่งเป็นไลบรารีที่มีอยู่ในไพทอน โดยอัลกอริทึมของตัวจำแนกประเภทแต่ละชนิดจะประกอบด้วยตัวแปรที่แตกต่างกัน เช่น อัลกอริทึมของการเรียนรู้เบส์อย่างง่ายและโครงข่ายประสาทเทียมจะมีตัวแปรเป็นค่าแอลฟา (Alpha) ส่วนอัลกอริทึมของซัพพอร์ตเวกเตอร์แมชชีนจะมีตัวแปรเป็นค่าซี (C) และเคอร์เนล (Kernel) ดังนั้นเพื่อให้ได้แบบจำลองที่มีประสิทธิภาพ การกำหนดค่าตัวแปรในอัลกอริทึมจึงถือเป็นปัจจัยสำคัญ โดยการกำหนดค่าให้กับตัวแปรจะใช้กราฟเส้นโค้งอาร์โอซีเข้ามาช่วยในการพิจารณา ดังแสดงในรูปที่ 19 ถึง 22



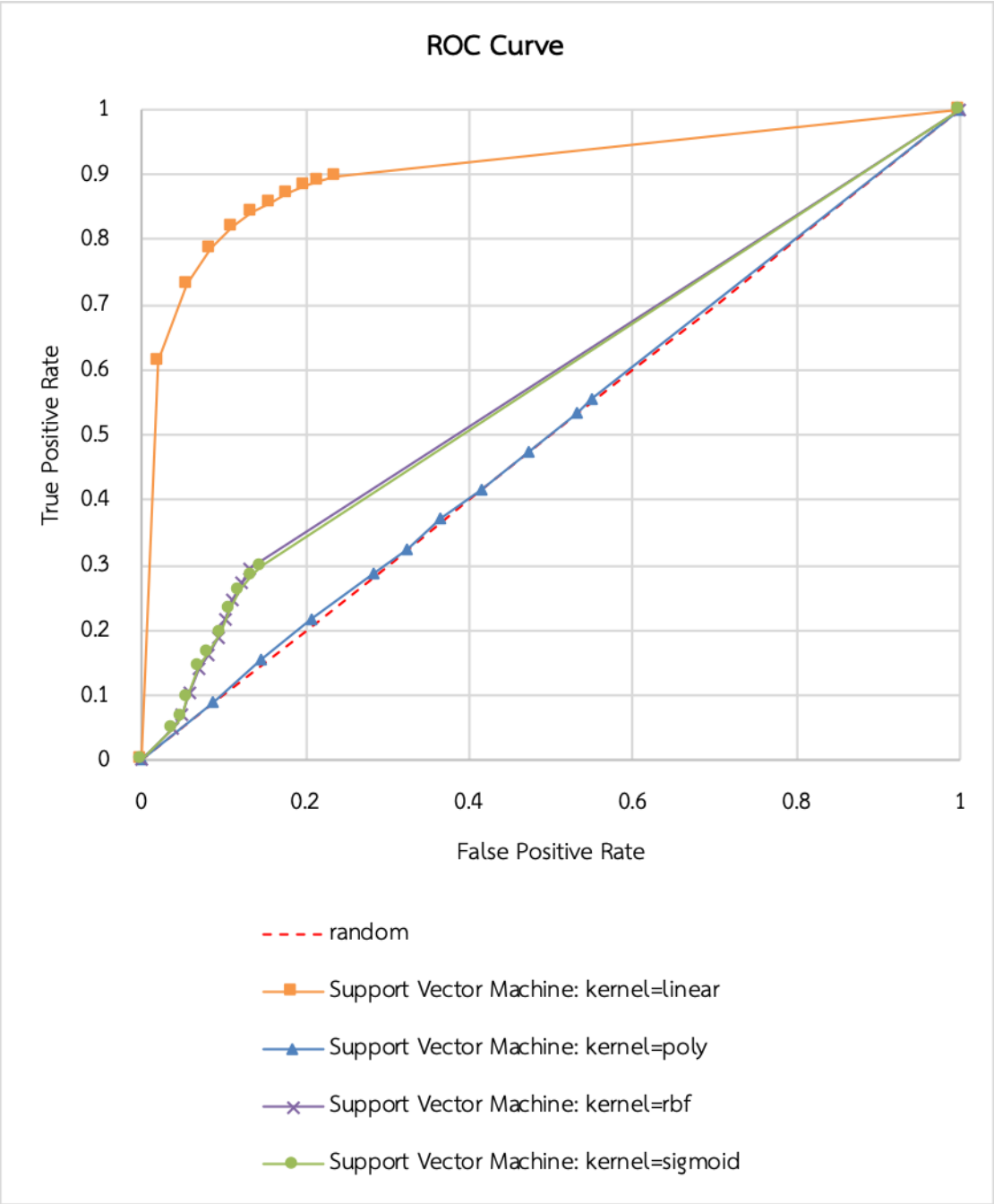
รูปที่ 19 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมการเรียนรู้เบส์อย่างง่ายด้วยค่าแอลฟาที่แตกต่างกัน

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 19 จะเห็นว่าเส้นโค้ง Naïve Bayes: alpha=0.01 เป็นเส้นที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด ดังนั้นงานวิจัยนี้จึงเลือกใช้ค่าแอลฟาเท่ากับ 0.01 ในการสร้างแบบจำลองด้วยอัลกอริทึมการเรียนรู้เบส์อย่างง่าย



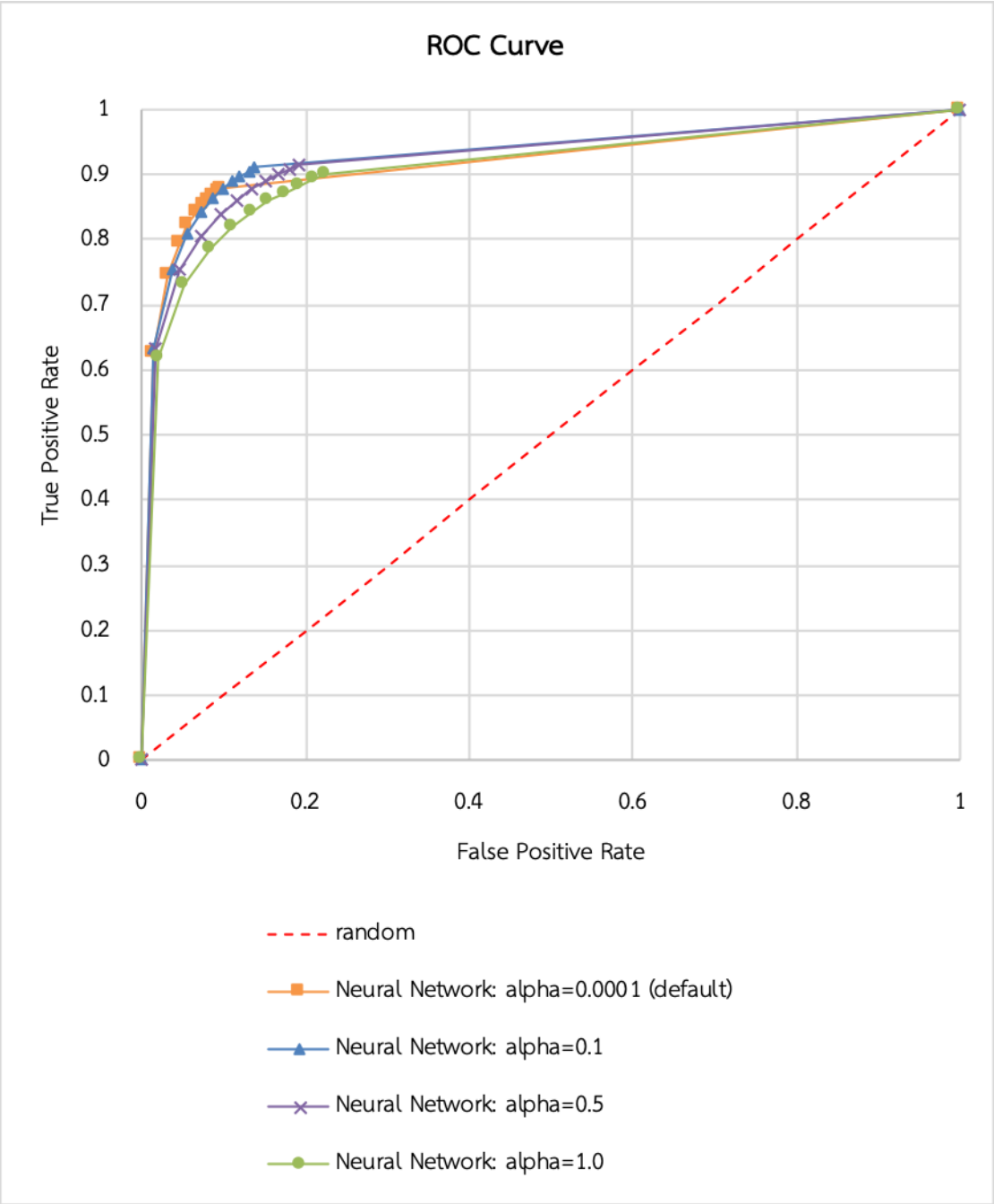
รูปที่ 20 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนด้วยค่าซีที่แตกต่างกัน

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 20 จะเห็นว่าเส้นโค้ง Support Vector Machine: C=0.1 เป็นเส้นที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด ดังนั้นงานวิจัยนี้จึงเลือกใช้ค่าซีเท่ากับ 0.1 ในการสร้างแบบจำลองด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน



รูปที่ 21 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนด้วยเคอร์เนลที่แตกต่างกัน

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 21 จะเห็นว่าเส้นโค้ง Support Vector Machine: kernel=linear เป็นเส้นที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด ดังนั้นงานวิจัยนี้จึงเลือกใช้เคอร์เนลที่เป็น linear ในการสร้างแบบจำลองด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน



รูปที่ 22 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมโครงข่ายประสาทเทียมด้วยค่าแอลฟาที่แตกต่างกัน

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 22 จะเห็นว่าเส้นโค้ง Neural Network: alpha=0.1 เป็นเส้นที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด ดังนั้นงานวิจัยนี้จึงเลือกใช้ค่าแอลฟาเท่ากับ 0.1 ในการสร้างแบบจำลองด้วยอัลกอริทึมโครงข่ายประสาทเทียม

ผลลัพธ์ที่ได้จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 19 ถึง 22 สามารถสรุปค่าตัวแปรที่เหมาะสมสำหรับอัลกอริทึมของตัวจำแนกประเภทแต่ละชนิดได้ ดังตารางที่ 6

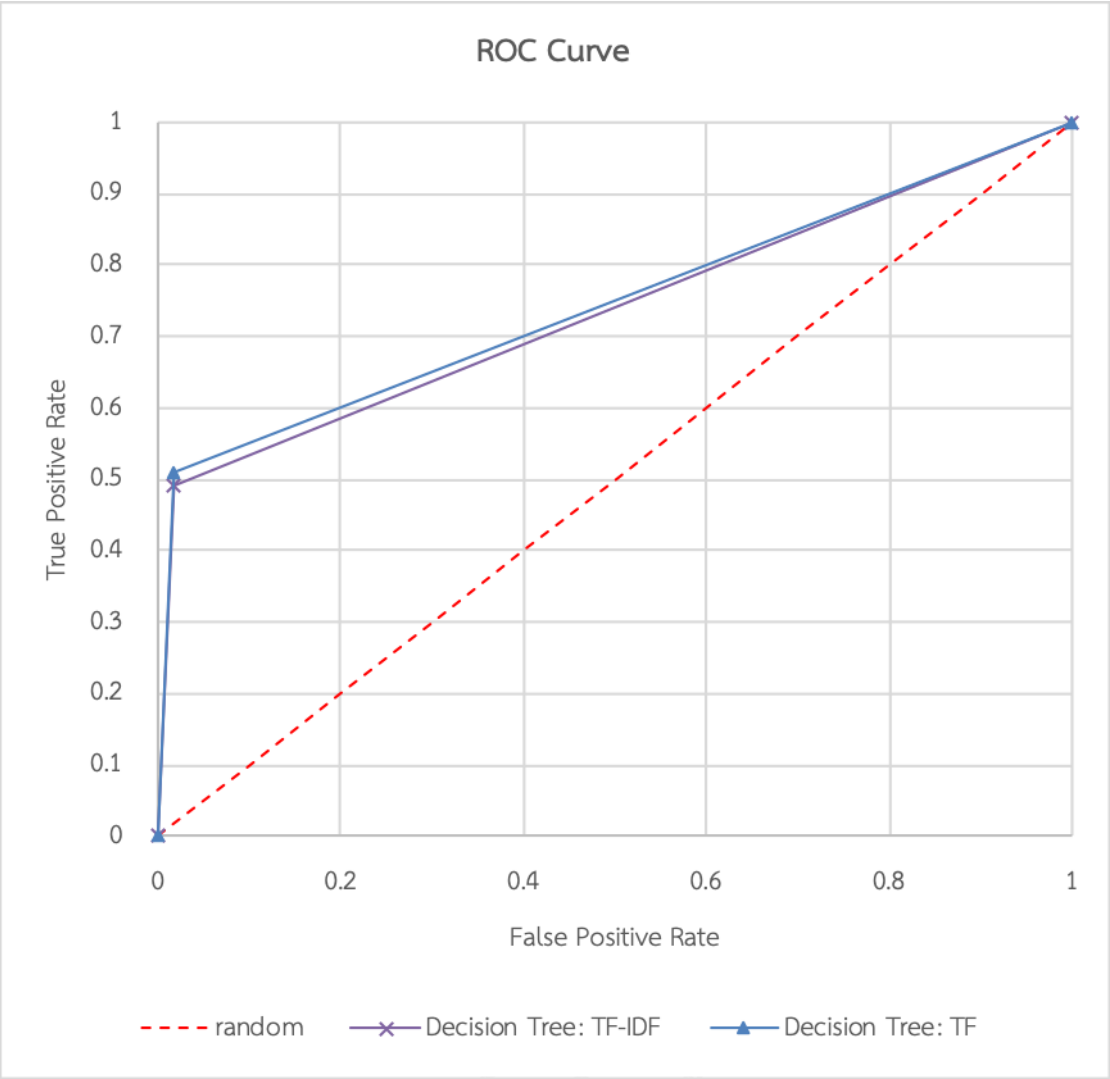
ตารางที่ 6 ตารางแสดงอัลกอริทึมของตัวจำแนกประเภทชนิดต่าง ๆ

ตัวจำแนกประเภท	ไลบรารี	การตั้งค่า
ต้นไม้ตัดสินใจ	sklearn.tree	default
การเรียนรู้เบส์อย่างง่าย	sklearn.naive_bayes	alpha=0.01
ซัพพอร์ตเวกเตอร์แมชชีน	sklearn.svm	kernel='linear', probability=True, C=0.1
โครงข่ายประสาทเทียม	sklearn.neural_network	alpha=0.1

การทำเหมืองข้อความเพื่อสร้างแบบจำลอง จะอาศัยหลักการวิเคราะห์หาคำสำคัญจากคำที่ปรากฏอยู่ในเอกสารหรือชุดข้อมูลสอน เพื่อสร้างแบบจำลองเชิงทำนาย โดยเทคนิคที่ใช้ในการวิเคราะห์หาคำสำคัญที่งานวิจัยนี้นำมาใช้ เพื่อเปรียบเทียบผลลัพธ์ที่ได้มี 2 วิธี ดังนี้

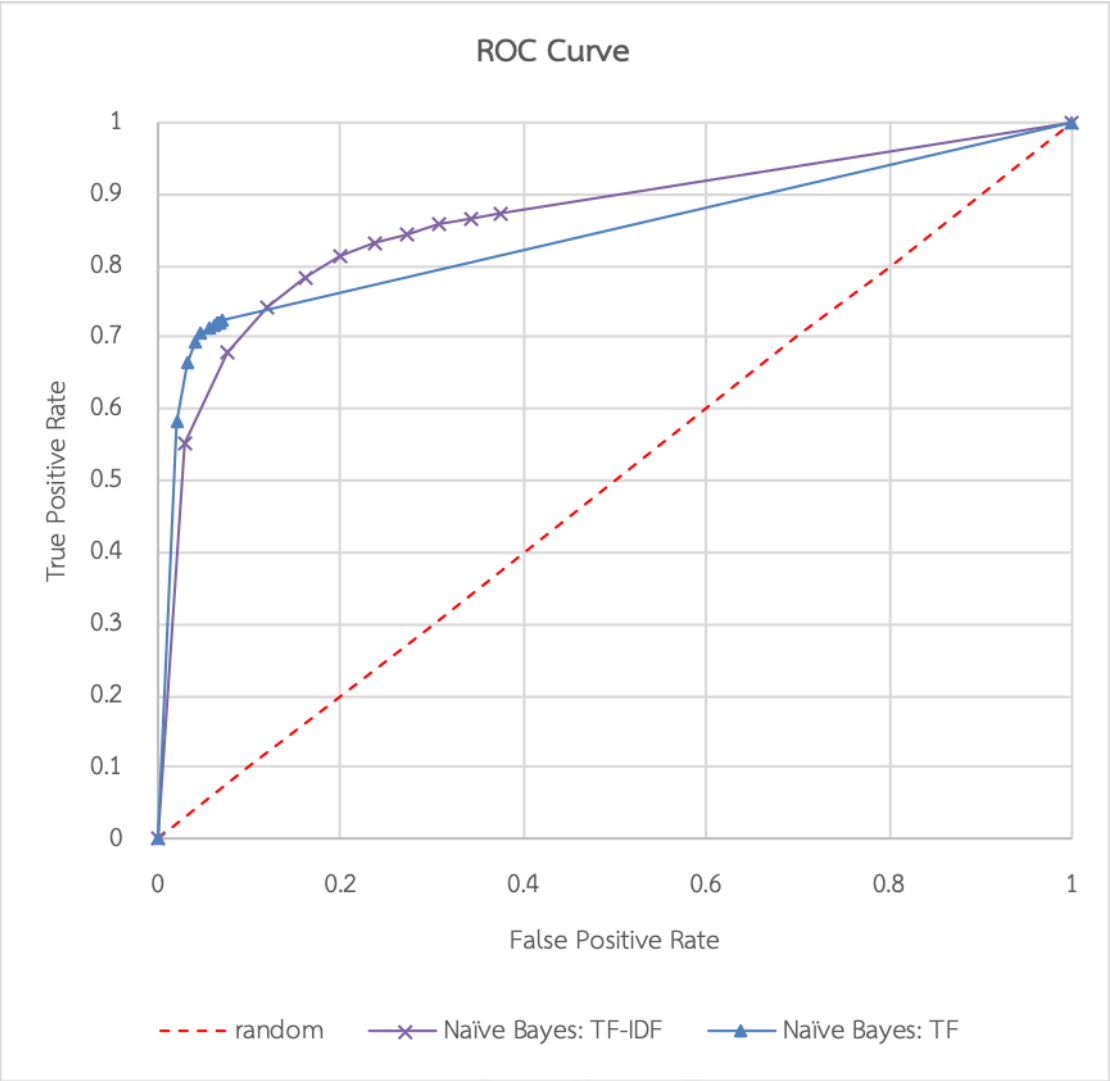
1. ความถี่ของคำ (Term Frequency: TF) เป็นวิธีการนับความถี่ของคำแต่ละคำที่ปรากฏอยู่ในเอกสาร โดยจะใช้ค่าความถี่ของคำเป็นตัวชี้วัดความสำคัญของคำ
2. ความถี่-ส่วนกลับของความถี่ของคำ เป็นวิธีการคำนวณจากผลคูณระหว่างค่าความถี่กับค่าส่วนกลับของความถี่ของคำ (Inverse Document Frequency: IDF) โดยส่วนกลับของความถี่ของคำจะแสดงให้เห็นว่าคำใดที่ปรากฏอยู่ในเอกสารหลาย ๆ ชุด คำนั้นย่อมมีความสำคัญลดลง ซึ่งถ้าผลคูณที่ได้มีค่ามากแสดงว่าคำนั้นมีความถี่สูงในเอกสารที่ใช้คำนวณ แต่มีความถี่ต่ำในเอกสารอื่น ๆ

โดยในแต่ละอัลกอริทึมของตัวจำแนกประเภทจะมีวิธีการวิเคราะห์หาคำสำคัญที่เหมาะสมที่แตกต่างกันไป โดยสามารถพิจารณาได้จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 23 ถึง 26



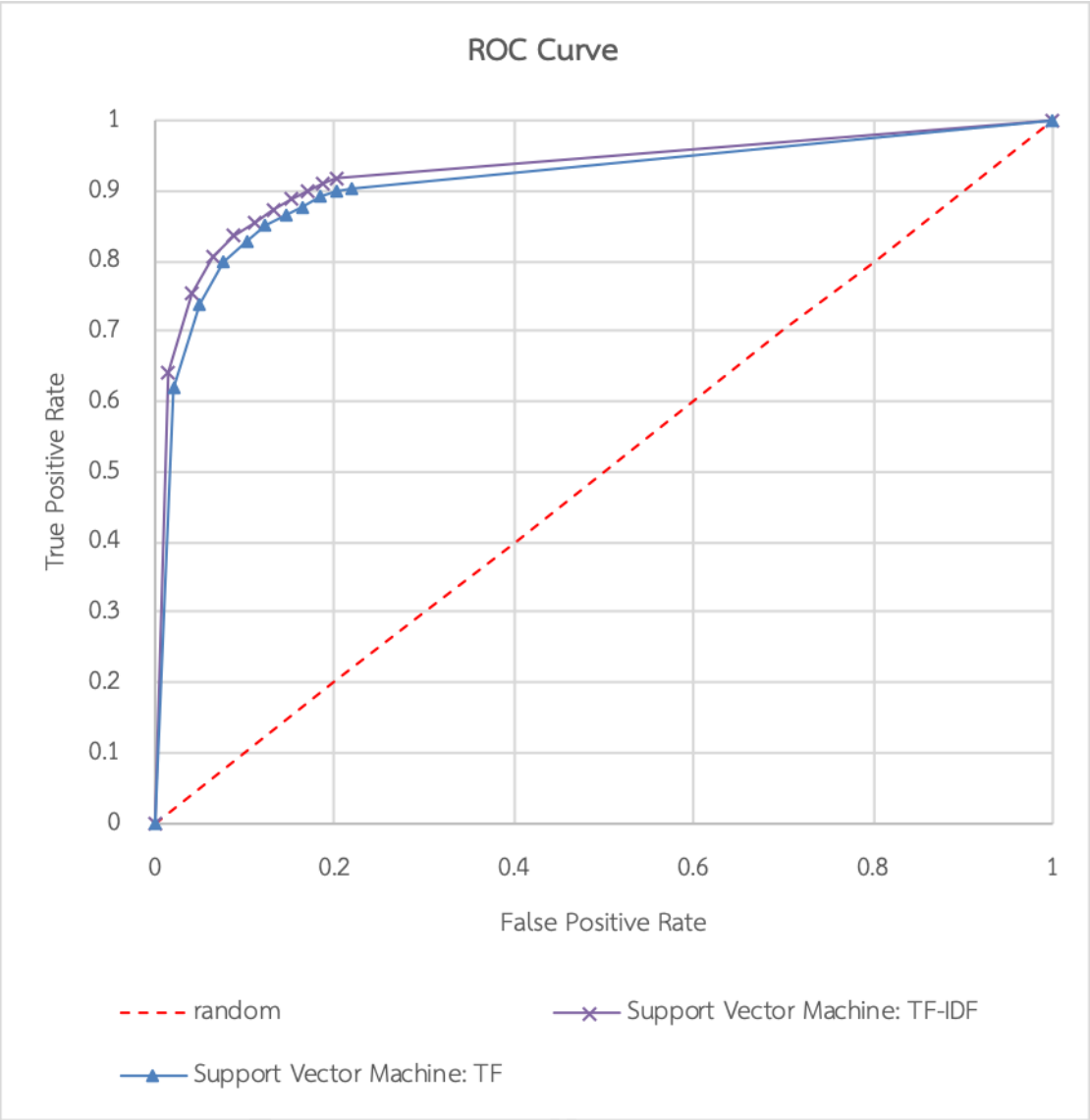
รูปที่ 23 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมต้นไม้ตัดสินใจ โดยใช้ความถี่ของคำกับความถี่-ส่วนกลับของความถี่ของคำ

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 23 จะเห็นว่าเส้นโค้ง Decision Tree: TF เป็นเส้นที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด ดังนั้นงานวิจัยนี้จึงเลือกใช้วิธีการสร้างแบบจำลองด้วยความถี่ของคำสำหรับอัลกอริทึมต้นไม้ตัดสินใจ



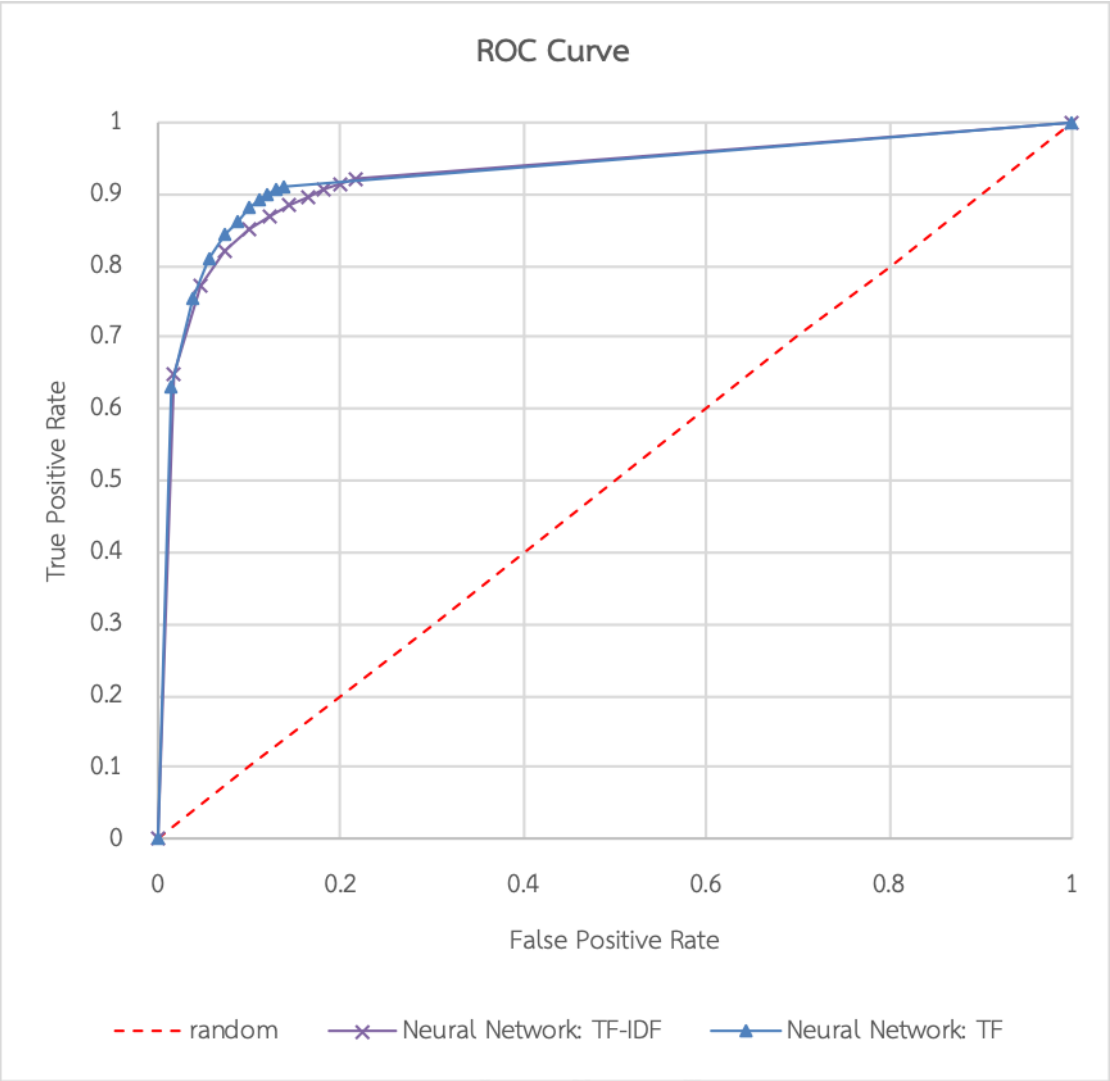
รูปที่ 24 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมการเรียนรู้เบส์อย่างง่าย โดยใช้ความถี่ของคำกับความถี่-ส่วนกลับของความถี่ของคำ

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 24 จะเห็นว่าเส้นโค้ง Naïve Bayes: TF-IDF เป็นเส้นที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด ดังนั้นงานวิจัยนี้จึงเลือกใช้วิธีการสร้างแบบจำลองด้วยความถี่-ส่วนกลับของความถี่ของคำสำหรับอัลกอริทึมการเรียนรู้เบส์อย่างง่าย



รูปที่ 25 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ความถี่ของคำกับความถี่-ส่วนกลับของความถี่ของคำ

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 25 จะเห็นว่าเส้นโค้ง Support Vector Machine: TF-IDF เป็นเส้นที่มีความชันและมีพื้นที่ใต้กราฟสูงสุด ดังนั้นงานวิจัยนี้จึงเลือกใช้วิธีการสร้างแบบจำลองด้วยความถี่-ส่วนกลับของความถี่ของคำสำหรับอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน



รูปที่ 26 กราฟเปรียบเทียบแบบจำลองที่สร้างจากอัลกอริทึมโครงข่ายประสาทเทียม โดยใช้ความถี่ของคำกับความถี่-ส่วนกลับของความถี่ของคำ

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 26 จะเห็นว่าเส้นโค้ง Neural Network: TF เป็นเส้นที่มีความชันและพื้นที่ใต้กราฟสูงสุด ดังนั้นงานวิจัยนี้จึงเลือกใช้วิธีการสร้างแบบจำลองด้วยความถี่ของคำสำหรับอัลกอริทึมโครงข่ายประสาทเทียม

ผลลัพธ์ที่ได้จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 23 ถึง 26 สามารถสรุปวิธีการวิเคราะห์หาคำสำคัญที่เหมาะสมของตัวจำแนกประเภทแต่ละชนิดได้ ดังตารางที่ 7

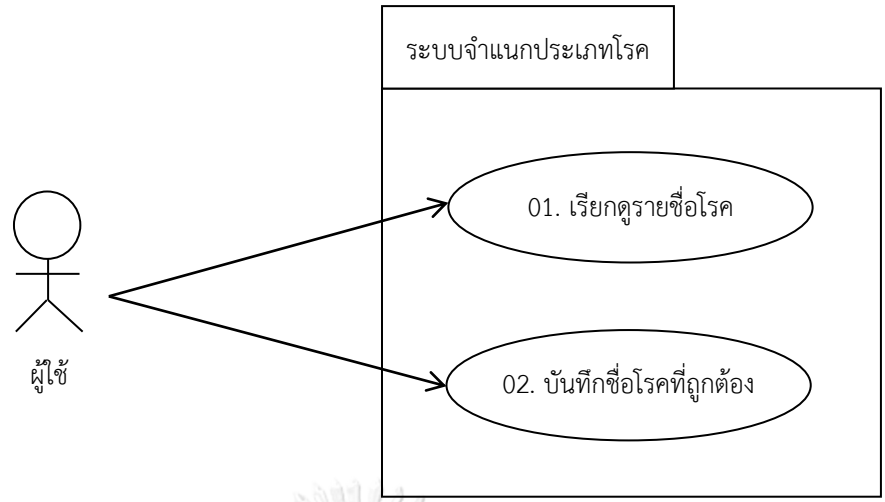
ตารางที่ 7 ตารางแสดงวิธีการวิเคราะห์หาคำสำคัญของตัวจำแนกประเภทแต่ละชนิด

ตัวจำแนกประเภท	วิธีการวิเคราะห์หาคำสำคัญ
ต้นไม้ตัดสินใจ	ความถี่ของคำ
การเรียนรู้เบส์อย่างง่าย	ความถี่-ส่วนกลับของความถี่ของคำ
ซัพพอร์ตเวกเตอร์แมชชีน	ความถี่-ส่วนกลับของความถี่ของคำ
โครงข่ายประสาทเทียม	ความถี่ของคำ

3.3 การใช้งานแบบจำลอง

การใช้งานแบบจำลอง เริ่มจากการให้ผู้ใช้กรอกและบันทึกข้อมูลอาการ ซึ่งข้อมูลที่กรอกจะต้องเป็นภาษาไทยหรือภาษาอังกฤษเท่านั้น จากนั้นระบบจะทำการแปลข้อมูลให้เป็นภาษาอังกฤษทั้งหมดโดยใช้โมดูลโกสเลท (Goslate module) [29] และทำการประมวลผลข้อมูลก่อนเช่นเดียวกับในขั้นตอนการสร้างแบบจำลอง เพื่อแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม และนำข้อมูลที่ได้ไปทำการประมวลผลผ่านแบบจำลอง เพื่อทำนายชื่อโรคที่มีความน่าจะเป็นออกมา

ฟังก์ชันในการทำงานของแบบจำลองแบ่งเป็น 2 ส่วน ส่วนที่ 1 คือฟังก์ชันการเรียกดูรายชื่อโรค เป็นส่วนที่แสดงผลลัพธ์รายชื่อโรคที่มีความน่าจะเป็นจากข้อมูลอาการที่ผู้ใช้กรอก และส่วนที่ 2 คือฟังก์ชันการบันทึกชื่อโรคที่ถูกต้อง เป็นส่วนที่ให้ผู้ใช้เลือกชื่อโรคที่คิดว่าถูกต้องจากผลลัพธ์ที่แบบจำลองแสดง โดยทั้ง 2 ฟังก์ชันนี้สามารถแสดงเป็นแผนภาพยูสเคส (Use Case Diagram) ได้ดังรูปที่ 27 และแสดงเป็นคำอธิบายของแต่ละยูสเคสได้ตามตารางที่ 8 และตารางที่ 9



รูปที่ 27 ยูสเคสของระบบจำแนกประเภทโรค

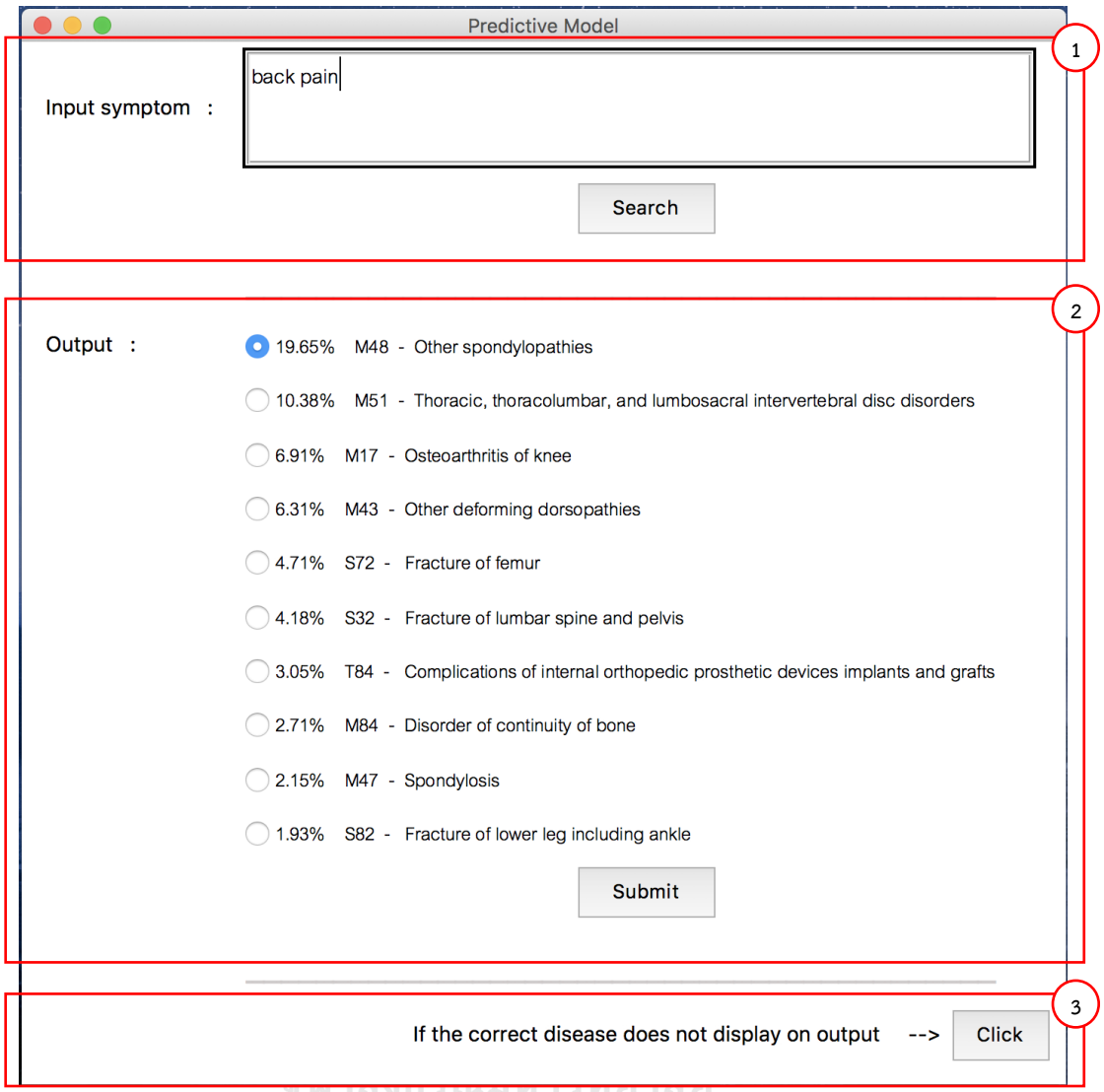
ตารางที่ 8 ตารางแสดงคำอธิบายยูสเคสของฟังก์ชันเรียกดูรายชื่อโรค

Use case ID	U01	
Use case name	เรียกดูรายชื่อโรค	
Actor	ผู้ใช้	
Pre-condition	ผู้ใช้กรอกข้อมูลอาการ	
Post-condition	ผู้ใช้สามารถเรียกดูรายชื่อโรคได้	
Flow of events	Actor	System
	2. กรอกข้อมูลอาการ 3. กดปุ่ม Search	1. แสดงหน้ากรอกข้อมูลอาการ 4. รับข้อมูลอาการ 5. ประมวลผลข้อมูลอาการผ่านแบบจำลอง 6. แสดงผลลัพธ์รายชื่อโรค

ตารางที่ 9 ตารางแสดงคำอธิบายยูสเคสของฟังก์ชันบันทึกชื่อโรคที่ถูกต้อง

Use case ID	U02	
Use case name	บันทึกชื่อโรคที่ถูกต้อง	
Actor	ผู้ใช้	
Pre-condition	ผู้ใช้ทราบชื่อโรคที่ถูกต้อง และชื่อโรคที่ถูกต้องถูกแสดง	
Post-condition	ระบบทำการเรียนรู้ใหม่ด้วยชื่อโรคที่ถูกต้อง	
Flow of events	Actor	System
	2. เลือกชื่อโรคที่ถูกต้อง 3. กดปุ่ม Submit	1. แสดงผลลัพธ์รายชื่อโรค 4. รับชื่อโรคที่ถูกต้อง 5. สอนแบบจำลองใหม่ด้วยข้อมูลอาการและชื่อโรคที่ถูกต้อง
Alternative Flows	2. กดปุ่ม Click 3. กรอกรหัสไอซีดีเทนซีเอ็มที่ถูกต้อง 4. กดปุ่ม Submit	1. แสดงผลลัพธ์รายชื่อโรค 5. รับชื่อโรคที่ถูกต้อง 6. สอนแบบจำลองใหม่ด้วยข้อมูลอาการและชื่อโรคที่ถูกต้อง

จากยูสเคสข้างต้น งานวิจัยนี้จึงมีแนวคิดในการออกแบบส่วนต่อประสานกับผู้ใช้ ดังรูปที่ 28 โดยแบ่งฟังก์ชันการทำงานออกเป็น 3 ส่วน ได้แก่ ส่วนรับข้อมูลเข้า ส่วนแสดงผลลัพธ์ และส่วนรับผลป้อนกลับ (feedback) จากผู้ใช้



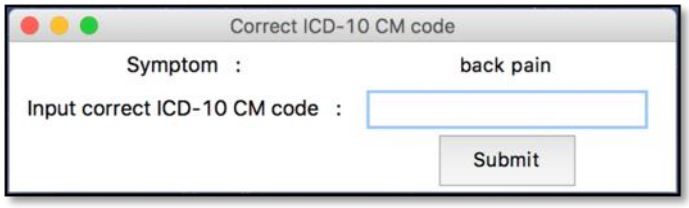
รูปที่ 28 ส่วนต่อประสานกับผู้ใช้ของแบบจำลอง

หมายเลข 1 เป็นส่วนต่อประสานสำหรับกรอกและบันทึกข้อมูลอาการ

หมายเลข 2 เป็นส่วนต่อประสานสำหรับแสดงผลลัพธ์ของแบบจำลอง ซึ่งประกอบด้วยค่าความน่าจะเป็นของโรค รหัสไอซีดีเทนซีเอ็ม และชื่อโรค โดยในส่วนนี้จะให้ผู้ใช้เลือกชื่อโรคที่คิดว่าถูกต้องมา 1 โรค เพื่อใช้เป็นชุดข้อมูลใหม่สำหรับสอนแบบจำลอง

หมายเลข 3 เป็นส่วนต่อประสานสำหรับกรณีที่ไม่มีชื่อโรคที่ผู้ใช้คิดว่าถูกต้องแสดงอยู่ใน

ผลลัพธ์ของแบบจำลอง โดยเมื่อผู้ใช้กดปุ่ม Click จะมีหน้าจอดังรูปที่ 29 ปรากฏขึ้นมา เพื่อให้ผู้ใช้ทำการกรอกและบันทึกรหัสไอซีดีเทนซีเอ็มที่คิดว่าถูกต้อง



รูปที่ 29 หน้าจอสำหรับกรอกและบันทึกรหัสไอซีดีเทนซีเอ็ม

3.4 เครื่องมือที่ใช้ในการพัฒนาแบบจำลอง

1. ฮาร์ดแวร์ที่ใช้ในการพัฒนาเครื่องมือ

   เครื่องคอมพิวเตอร์ที่ใช้ในการพัฒนา

   - หน่วยประมวลผล ความเร็ว 2.3 กิกะเฮิร์ท อินเทล คอร์ไอ 5
   - หน่วยความจำ 8 กิกะไบต์
   - ฮาร์ดดิสก์ ความจุ 256 กิกะไบต์
   - จอภาพ ขนาด 13.3 นิ้ว

2. ซอฟต์แวร์ที่ใช้ในการพัฒนาเครื่องมือ

   โปรแกรมสำหรับเครื่องคอมพิวเตอร์

   - ระบบปฏิบัติการแมคโอเอส เวอร์ชัน 10.13.3
   - เคอร์เนล ดาร์วิน เวอร์ชัน 17.4.0

   ภาษาสำหรับพัฒนาเครื่องมือในการสร้างแบบจำลอง

   - ภาษาไพทอน เวอร์ชัน 2.7

   ไลบรารีสำหรับพัฒนาเครื่องมือในการสร้างแบบจำลอง

   - ไลบรารีเอ็นแอลทีเค (NLTK)
   - ไลบรารีไซคิดเลิร์น (Scikit-learn)
     - ไลบรารีเนอีฟเบส์
     - ไลบรารีดีซีชันทรี
     - ไลบรารีเอสวีเอ็ม
     - ไลบรารีนิวรอลเน็ตเวิร์ค

บทที่ 4
การทดสอบเครื่องมือ และการอภิปราย

4.1 การประเมินประสิทธิภาพของอัลกอริทึมที่ใช้ในการสร้างแบบจำลอง

แบบจำลองที่ได้ในงานวิจัยนี้จะมีทั้งหมด 4 แบบ ซึ่งได้มาจากการใช้ตัวจำแนกประเภททั้งหมด 4 ชนิด ได้แก่ แบบจำลองจากต้นไม้ตัดสินใจ แบบจำลองจากการเรียนรู้เบส์อย่างง่าย แบบจำลองจากซัพพอร์ตเวกเตอร์แมชชีน และแบบจำลองจากโครงข่ายประสาทเทียม โดยการสร้างแบบจำลองจะใช้ชุดข้อมูลสอนและชุดข้อมูลทดสอบชุดเดียวกัน ดังนั้นการเปรียบเทียบแบบจำลองจะดูจาก

1. ระยะเวลาที่ใช้ในขั้นตอนการเรียนรู้เพื่อสร้างแบบจำลอง
2. ระยะเวลาที่แบบจำลองใช้ในการทำนายผล
3. กราฟเส้นโค้งอาร์โอซี และพื้นที่ใต้เส้นโค้ง
4. อัตราผลบวกเท็จ ค่าความเที่ยง ค่าการระลึกได้ และค่าความแม่นของแบบจำลอง

การวัดประสิทธิภาพแบบจำลองจะเริ่มจากการสร้างคอนฟิวชันเมทริกซ์ (Confusion matrix) [26] ซึ่งเป็นตารางแบบจัตุรัสที่มีจำนวนแถวเท่ากับจำนวนคอลัมน์และเท่ากับจำนวนประเภท โดยในงานวิจัยนี้จะสร้างคอนฟิวชันเมทริกซ์ที่มีขนาด 2 x 2 ดังตารางที่ 10 ซึ่งข้อมูลที่อยู่ในคอลัมน์จะเป็นประเภทที่อยู่ในชุดข้อมูลสอน ส่วนข้อมูลที่อยู่ในแถวจะเป็นประเภทที่แบบจำลองทำนายได้

ตารางที่ 10 ตารางแสดงคอนฟิวชันเมทริกซ์ขนาด 2 x 2

	ชื่อโรคที่ถูกต้อง	**ชื่อโรคที่ไม่ถูกต้อง**
แบบจำลองแสดงชื่อโรค	TP (ผลลัพธ์ที่ถูกต้อง)	FP (ผลลัพธ์ที่เกินคาด)
แบบจำลองไม่แสดงชื่อโรค	FN (ผลลัพธ์ที่หายไป)	TN (ผลลัพธ์ที่ถูกต้อง)

ชื่อโรคทั้งหมดที่นำมาใช้เพื่อคำนวณหาค่าต่าง ๆ ในตารางที่ 10 จะนำมาจากชื่อโรคที่พบในเวชระเบียนผู้ป่วยแผนกออโธปิดิกส์ และชื่อโรคในหมวดกระดูกและกล้ามเนื้อของรหัสไอซีดีเทนซีเอ็มเท่านั้น โดยในตารางที่ 10 จะประกอบด้วยค่าต่าง ๆ ดังนี้

- ผลบวกจริง (True Positive: TP) คือ จำนวนกรณีที่ผู้ป่วยเป็นโรค และแบบจำลองทำนายว่าเป็นโรค
- ผลบวกเท็จ (False Positive: FP) คือ จำนวนกรณีที่ผู้ป่วยไม่เป็นโรค แต่แบบจำลองทำนายว่าเป็นโรค
- ผลลบจริง (True Negative: TN) คือ จำนวนกรณีที่ผู้ป่วยไม่เป็นโรค และแบบจำลองทำนายว่าไม่เป็นโรค
- ผลลบเท็จ (False Negative: FN) คือ จำนวนกรณีที่ผู้ป่วยเป็นโรค แต่แบบจำลองทำนายว่าไม่เป็นโรค

การคำนวณจะแยกพิจารณาหาค่าผลบวกจริง ผลบวกเท็จ ผลลบจริง และผลลบเท็จของแต่ละโรค ดังตัวอย่างในรูปที่ 30 ถึง 33

เวชระเบียน		**แบบจำลอง**
บันทึกของแพทย์	**รหัสไอซีดีเทนซีเอ็ม**	**รหัสไอซีดีเทนซีเอ็ม**
case present with Sciatica pain right leg + weakness right leg	**M51**	S82
		M51
		M43
		. . .

กรณี ผลบวกจริง ของ M51 (TP = 1)

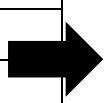
รูปที่ 30 การพิจารณาผลบวกจริงของรหัสไอซีดีเทนซีเอ็ม M51

เวชระเบียน		**แบบจำลอง**
บันทึกของแพทย์	**รหัสไอซีดีเทนซีเอ็ม**	**รหัสไอซีดีเทนซีเอ็ม**
ปวดหลังร้าวลงขาซ้าย มากกว่าขาขวา 1 เดือน อาการปวดเป็นมากขึ้นเรื่อย ๆ . . .	M48	M43
		M48
		M51
		. . .

กรณี ผลบวกเท็จ ของ M51 (FP = 1)

รูปที่ 31 การพิจารณาผลบวกเท็จของรหัสไอซีดีเทนซีเอ็ม M51

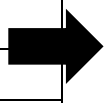
เวชระเบียน		แบบจำลอง
บันทึกของแพทย์	**รหัสไอซีดีเทนซีเอ็ม**	**รหัสไอซีดีเทนซีเอ็ม**
ล้มก้นกระแทกพื้น ปวดสะโพกขวา เดินไม่ได้	S72	S72
		T84
		M16
		. . .



กรณี ผลลบจริง ของ M51 (TN = 1)

รูปที่ 32 การพิจารณาผลลบจริงของรหัสไอซีดีเทนซีเอ็ม M51

เวชระเบียน		แบบจำลอง
บันทึกของแพทย์	**รหัสไอซีดีเทนซีเอ็ม**	**รหัสไอซีดีเทนซีเอ็ม**
แขนซ้ายอ่อนแรง 4 เดือน มีอาการชาตั้งแต่ไหล่ลงมาที่แขน . . .	**M51**	T84
		G54
		M25
		. . .



กรณี ผลลบเท็จ ของ M51 (FN = 1)

รูปที่ 33 การพิจารณาผลลบเท็จของรหัสไอซีดีเทนซีเอ็ม M51

ค่าผลบวกจริง ผลบวกเท็จ ผลลบจริง และผลลบเท็จที่ได้ของแต่ละโรคจะถูกนำมาใช้เพื่อคำนวณหาอัตราผลบวกเท็จ ค่าความเที่ยง ค่าการระลึกได้ และค่าความแม่นจากสูตรดังต่อไปนี้

1. อัตราผลบวกเท็จ สามารถคำนวณได้จาก

$$False\ Positive\ Rate = \frac{FP}{FP + TN} \quad [26]$$

อัตราผลบวกเท็จเป็นอัตราส่วนระหว่างจำนวนผลบวกเท็จกับจำนวนชื่อโรคที่ไม่ถูกต้องทั้งหมด โดยถ้าอัตราผลบวกเท็จมีค่าน้อยแสดงว่าแบบจำลองมีประสิทธิภาพมากสามารถกำจัดชื่อโรคที่ไม่ถูกต้องได้

2. ค่าความเที่ยง สามารถคำนวณได้จาก

$$Precision = \frac{TP}{TP + FP} \quad [26]$$

ค่าความเที่ยงเป็นอัตราส่วนระหว่างจำนวนผลบวกจริงกับจำนวนชื่อโรคที่ถูกแสดงทั้งหมด โดยถ้าค่าความเที่ยงมีค่าสูงแสดงว่าแบบจำลองสามารถทำนายชื่อโรคได้อย่างแม่นยำ

3. ค่าการระลึกได้ หรืออัตราผลบวกจริง สามารถคำนวณได้จาก

$$Recall = \frac{TP}{TP + FN} \quad [26]$$

ค่าการระลึกได้เป็นอัตราส่วนระหว่างจำนวนผลบวกจริงกับจำนวนชื่อโรคที่ถูกต้องทั้งหมด โดยถ้าค่าการระลึกได้มีค่าสูงแสดงว่าแบบจำลองสามารถทำนายชื่อโรคที่ถูกต้องได้มาก

4. ค่าความแม่น สามารถคำนวณได้จาก

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [26]$$

ค่าความแม่นเป็นอัตราส่วนระหว่างจำนวนชื่อโรคที่ทำนายถูกต้องกับจำนวนกรณีทั้งหมด โดยแบบจำลองที่ดีควรมีค่าความแม่นที่สูง

เมื่อได้อัตราผลบวกเท็จ ค่าความเที่ยง ค่าการระลึกได้ และค่าความแม่นของแต่ละโรคแล้ว จะนำค่าเหล่านี้ของทุกโรคมารวมกันเพื่อหาค่าเฉลี่ย โดยใช้วิธีการหาค่าเฉลี่ยขนาดเล็ก (Micro-average) กำหนดให้ μ แทนค่าเฉลี่ยขนาดเล็ก และ C แทนจำนวนประเภทหรือจำนวนโรค ดังสมการต่อไปนี้

1. ค่าเฉลี่ยขนาดเล็กของอัตราผลบวกเท็จ สามารถคำนวณได้จาก

$$False\ Positive\ Rate^{\mu} = \frac{\sum_{i=1}^{|C|} FP_i}{\sum_{i=1}^{|C|}(FP_i + TN_i)} \quad [30]$$

2. ค่าเฉลี่ยขนาดเล็กของค่าความเที่ยง สามารถคำนวณได้จาก

$$Precision^{\mu} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|}(TP_i + FP_i)} \quad [30]$$

3. ค่าเฉลี่ยขนาดเล็กของค่าการระลึกได้ หรืออัตราผลบวกจริง สามารถคำนวณได้จาก

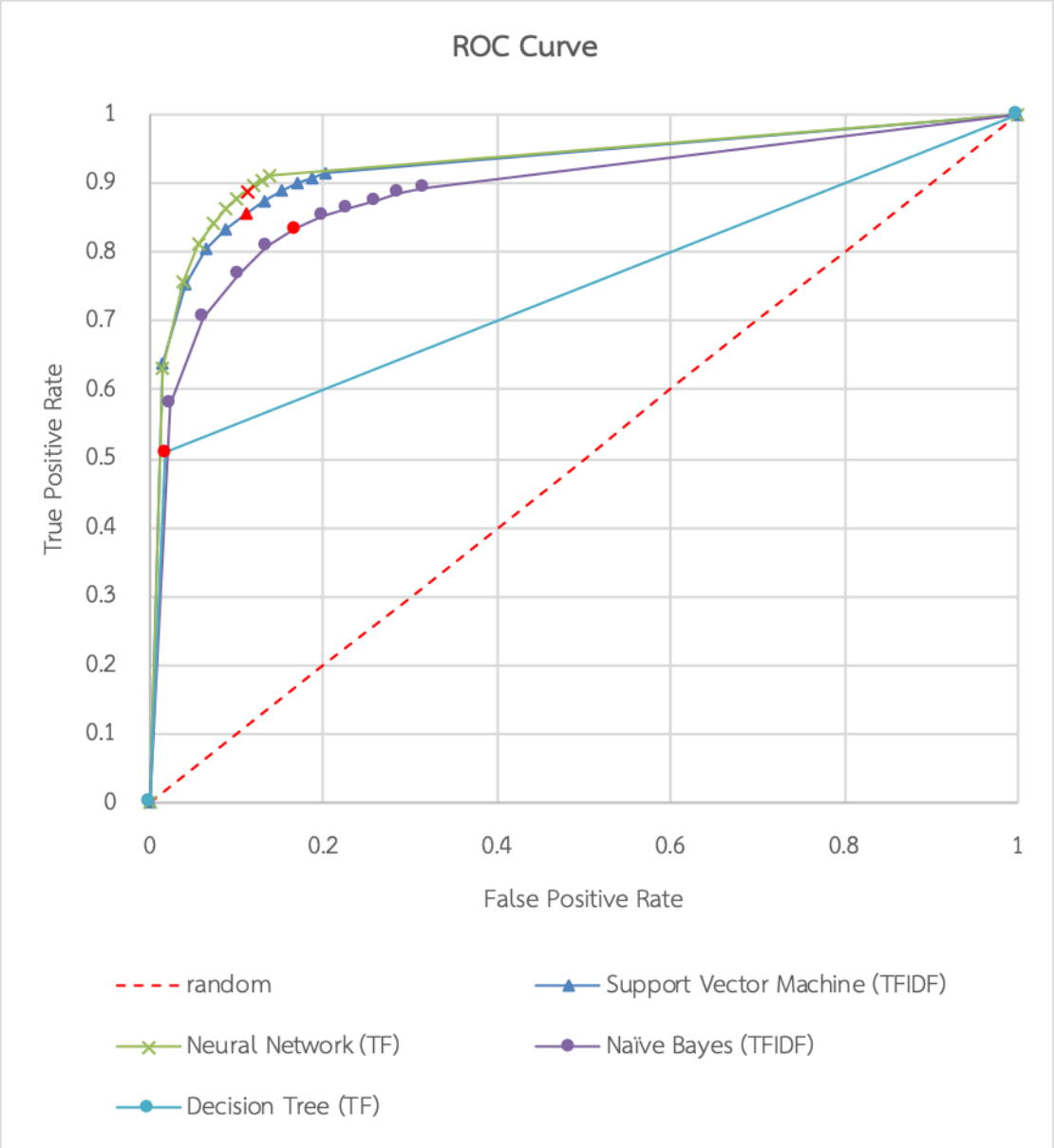
$$Recall^{\mu} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|}(TP_i + FN_i)} \quad [30]$$

4. ค่าเฉลี่ยขนาดเล็กของค่าความแม่น สามารถคำนวณได้จาก

$$Accuracy^{\mu} = \frac{\sum_{i=1}^{|C|} TP_i + TN_i}{\sum_{i=1}^{|C|}(TP_i + TN_i + FP_i + FN_i)} \quad [30]$$

4.2 ผลการประเมินประสิทธิภาพของอัลกอริทึมที่ใช้ในการสร้างแบบจำลอง

แบบจำลองในงานวิจัยนี้จะสร้างจากชุดข้อมูลสอนที่ประกอบด้วยข้อมูลจากเวชระเบียน 75 ส่วนและข้อมูลจากเว็บไซต์สาธารณะ 25 ส่วน ด้วยวิธีการสุ่มข้อมูลแบบความเที่ยงตรงโดยการแบ่งชุดข้อมูลเป็น 10 ส่วน ซึ่งอัลกอริทึมของตัวจำแนกประเภทที่นำมาใช้ในการสร้างแบบจำลองมีทั้งหมด 4 ชนิด ได้แก่ อัลกอริทึมของต้นไม้ตัดสินใจ การเรียนรู้เบส์อย่างง่าย ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม การทดสอบแบบจำลองจะใช้ข้อมูลจากเวชระเบียนมาเป็นชุดข้อมูลทดสอบ โดยผลลัพธ์จากการทดสอบแบบจำลองสามารถแสดงเป็นกราฟเส้นโค้งอาร์โอซีได้ ดังรูปที่ 34



รูปที่ 34 กราฟเปรียบเทียบแบบจำลองที่สร้างด้วยตัวจำแนกประเภททั้ง 4 ชนิด

จากกราฟเส้นโค้งอาร์โอซีในรูปที่ 34 จุดที่ n บนเส้นโค้ง หมายถึงจุดที่แบบจำลองแสดงชื่อโรคทั้งหมด n โรคในผลลัพธ์ โดยในการเปรียบเทียบอัตราผลบวกเท็จ ค่าความเที่ยง ค่าการระลึกได้ และค่าความแม่นของตัวจำแนกประเภทแต่ละชนิด จะคำนวณจากจุดตัด (cut-off point) ของแต่ละเส้นโค้ง ซึ่งเป็นจุดที่อยู่ใกล้มุมบนซ้ายมากที่สุด แล้วนำค่าเหล่านั้นมาเปรียบเทียบกัน ดังแสดงในตารางที่ 11

ตารางที่ 11 ตารางแสดงผลการประเมินประสิทธิภาพของอัลกอริทึมที่ใช้ในการสร้างแบบจำลอง

	ชนิดของตัวจำแนกประเภทที่นำมาสร้างแบบจำลอง			
	ต้นไม้ตัดสินใจ	**การเรียนรู้เบส์อย่างง่าย**	**ซัพพอร์ตเวกเตอร์แมชชีน**	**โครงข่ายประสาทเทียม**
การตั้งค่าอัลกอริทึม	default	alpha=0.01	kernel='linear', probability=True, C=0.1	alpha=0.1
วิธีการวิเคราะห์หาคำสำคัญ	ความถี่ของคำ	ความถี่-ส่วนกลับของความถี่ของคำ	ความถี่-ส่วนกลับของความถี่ของคำ	ความถี่ของคำ
เวลาที่ใช้ในการสร้างแบบจำลอง (วินาที)	21.4711	3.4228	813.3572	685.6481
เวลาที่แบบจำลองใช้ในการทำนาย (วินาที)	0.00003	0.00224	0.00194	0.00002
พื้นที่ใต้เส้นโค้ง	0.7462	0.8897	0.9257	**0.9304**
จุดตัด	1	5	5	7
อัตราผลบวกเท็จ (%)	**1.64%**	16.77%	11.12%	11.08%
ค่าความเที่ยง (%)	50.30%	23.60%	29.00%	27.11%
ค่าการระลึกได้ (%)	50.87%	83.47%	85.57%	**89.03%**
ค่าความแม่น (%)	**97.13%**	84.42%	89.44%	89.50%

จากตารางที่ 11 ระยะเวลาที่ใช้ในการสร้างและทำนายผลลัพธ์ของแบบจำลองเป็นค่าเฉลี่ยที่ได้มาจากวิธีการสุ่มข้อมูลแบบความเที่ยงตรงโดยการแบ่งชุดข้อมูลเป็น 10 ส่วน โดยระยะเวลาที่ใช้ในการทำนายของแบบจำลองแต่ละตัวจะไม่แตกต่างกัน ส่วนระยะเวลาที่ใช้ในการสร้างแบบจำลองจะแสดงให้เห็นถึงระยะเวลาที่แบบจำลองแต่ละตัวต้องใช้ในการเรียนรู้ชุดข้อมูลใหม่ที่ได้จากการใช้งานของผู้ใช้ โดยในการคัดเลือกอัลกอริทึมที่เหมาะสมสำหรับงานวิจัยนี้จะไม่นำระยะเวลาที่ใช้ในการสร้างแบบจำลองมาเป็นตัวชี้วัด เนื่องจากไม่ส่งผลต่อการใช้งานของผู้ใช้

แบบจำลองที่สร้างด้วยอัลกอริทึมของต้นไม้ตัดสินใจจะให้ค่าความแม่นสูงสุดและอัตราผลบวกเท็จต่ำสุด จากการแสดงชื่อโรคที่มีความน่าจะเป็นเพียงอันดับเดียวในผลลัพธ์ แต่เนื่องจากงานวิจัยนี้ต้องการแบบจำลองที่สามารถแสดงรายชื่อโรคที่มีความน่าจะเป็นได้หลายอันดับ เพื่อใช้เป็นตัวช่วยแพทย์ในการวินิจฉัยโรค ดังนั้นแบบจำลองที่เหมาะสมที่สุด คือ แบบจำลองที่สร้างด้วยอัลกอริทึมของโครงข่ายประสาทเทียม ซึ่งให้ค่าพื้นที่ใต้เส้นโค้งและค่าการระลึกได้สูงสุด จากการแสดงรายชื่อโรคที่มีความน่าจะเป็น 7 อันดับแรกในผลลัพธ์

สาเหตุที่ค่าความเที่ยงของอัลกอริทึมส่วนใหญ่ต่ำกว่าร้อยละ 50 เนื่องจากกรณีผลบวกเท็จที่เพิ่มขึ้น จากการแสดงชื่อโรคที่มีความน่าจะเป็นมากกว่า 1 โรคในผลลัพธ์ของแบบจำลอง แต่ผลลัพธ์ที่ถูกต้องยังคงมีเพียงโรคเดียว ทำให้มีกรณีผลบวกเท็จเพิ่มมากขึ้น ดังนั้นในงานวิจัยนี้จะไม่ใช้ค่าความเที่ยงมาเป็นตัวชี้วัด

4.3 ตัวอย่างผลลัพธ์ที่ได้จากการทดสอบแบบจำลอง

ผลลัพธ์ที่ได้จากการทดสอบแบบจำลองด้วยบันทึกของแพทย์ ถ้าชื่อโรคจากบันทึกของแพทย์บนเวชระเบียนมีปรากฏอยู่ในรายชื่อโรคที่แบบจำลองแสดงออกมา จะถือว่าแบบจำลองสามารถทำนายชื่อโรคได้ถูกต้อง ดังตารางที่ 12

ตารางที่ 12 ตารางแสดงตัวอย่างบันทึกของแพทย์ที่นำมาใช้ในการทดสอบและผลลัพธ์ที่ได้จากแบบจำลอง

ลำดับ	บันทึกของแพทย์	เวชระเบียน		แบบจำลอง	
		รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส	รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส
1	case present with Sciatica pain right leg + weakness right leg	M51	Thoracic, thoracolumbar, and lumbosacral intervertebral disc disorders	S82	Fracture of lower leg including ankle
				M51 ✓	Thoracic, thoracolumbar, and lumbosacral intervertebral disc disorders
				M43	Other deforming dorsopathies
				M48	Other spondylopathies
				M41	Scoliosis

ลำดับ	บันทึกของแพทย์	เวชระเบียน		แบบจำลอง	
		รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส	รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส
				D48	Neoplasm of uncertain behavior of other and unspecified sites
				M96	Intraoperative and postprocedural complications and disorders of musculoskeletal system, not elsewhere classified
2	ประสบอุบัติเหตุมอเตอร์ไซค์ล้ม ถูกชนทางด้านหน้า กระเด็นตกจากรถ ไม่ทราบว่าส่วนใดกระแทกพื้น ไม่มีหมดสติ พบแขนซ้ายผิดรูป ปวดขาซ้าย และมีบาดแผลบริเวณใบหน้า	S52	Fracture of forearm	S52 ✓	Fracture of forearm
				S72	Fracture of femur
				S82	Fracture of lower leg including ankle
				S22	Fracture of rib(s), sternum and thoracic spine
				S92	Fracture of foot and toe except ankle
				S42	Fracture of shoulder and upper arm
				S32	Fracture of lumbar spine and pelvis
3	ปวดหลังร้าวลงขาซ้ายมากกว่าขาขวา 1 เดือน อาการปวดเป็นมากขึ้นเรื่อย ๆ ไม่อ่อนแรง ไม่ชา ปัสสาวะ อุจจาระกลั้นได้ปกติดี	M48	Other spondylopathies	M43	Other deforming dorsopathies
				M48 ✓	Other spondylopathies
				M51	Thoracic, thoracolumbar, and lumbosacral intervertebral disc

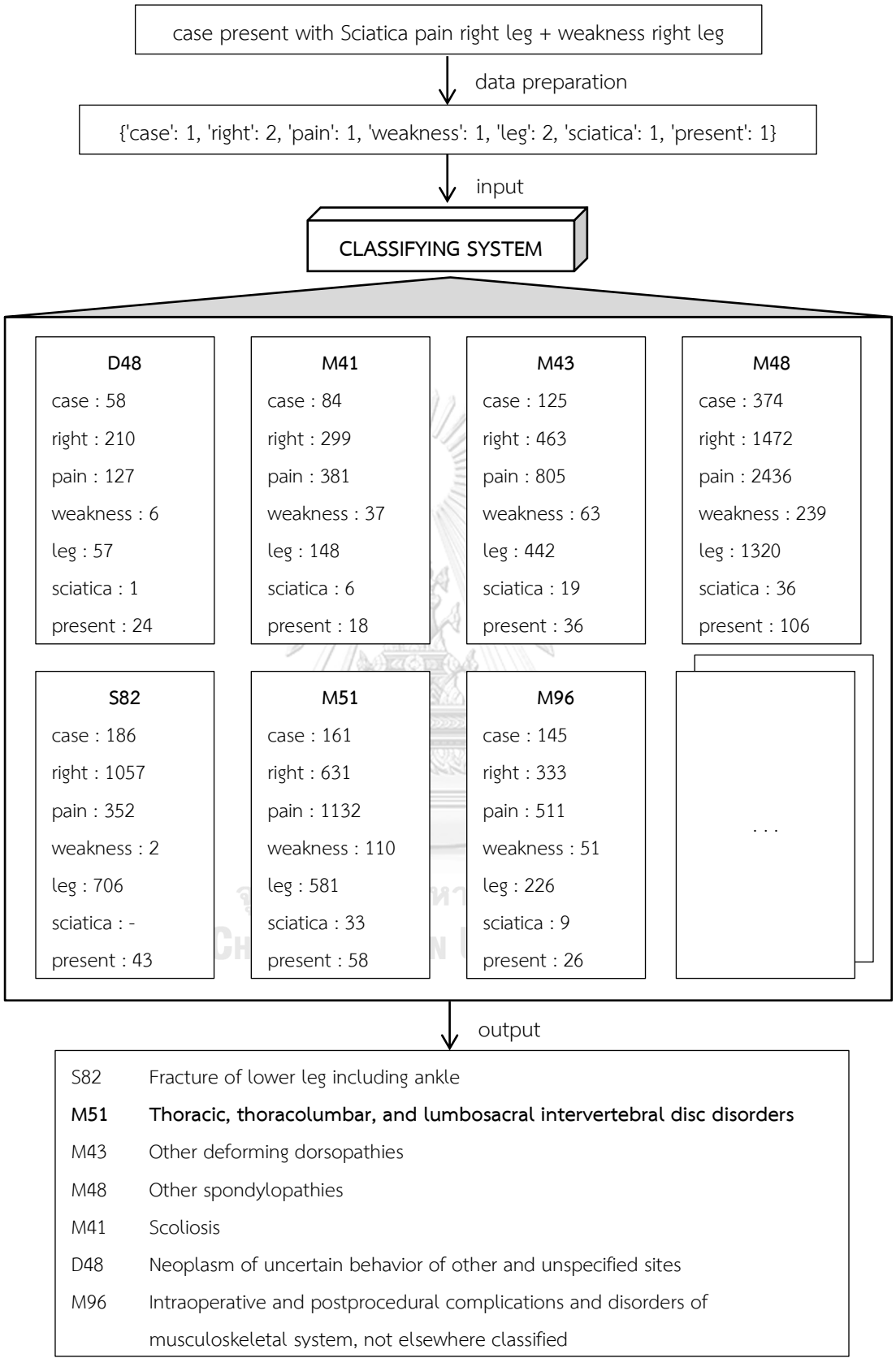
ลำดับ	บันทึกของแพทย์	เวชระเบียน		แบบจำลอง	
		รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส	รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส
					disorders
				C79	Secondary malignant neoplasm of other and unspecified sites
				M84	Disorder of continuity of bone
				M54	Dorsalgia
				S32	Fracture of lumbar spine and pelvis
4	ปวดหลังส่วนล่าง เป็นเวลานั่งหรือยืนนาน ๆ ไม่มีปวดร้าวลงขา เป็นมากขึ้น สังเกตว่าหลังโก่งมากขึ้น	M47	Spondylosis	M48	Other spondylopathies
				M46	Other inflammatory spondylopathies
				M47 ✓	Spondylosis
				T84	Complications of internal orthopedic prosthetic devices implants and grafts
				M84	Disorder of continuity of bone
				M41	Scoliosis
				D48	Neoplasm of uncertain behavior of other and unspecified sites
5	ปวดสะโพกขา เดินปกติ ผ่าตัด Hemiarthroplasty right hip หลังผ่าตัด เดินกะเผลก เจ็บขาหนีบและต้นขาเวลาขยับ ชาฝ่าเท้า	T84	Complications of internal orthopedic prosthetic devices implants and grafts	T84 ✓	Complications of internal orthopedic prosthetic devices implants and grafts
				M25	Other joint disorder, not elsewhere classified
				M24	Other specific joint

ลำดับ	บันทึกของแพทย์	เวชระเบียน		แบบจำลอง	
		รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส	รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส
					derangements
				M87	Osteonecrosis
				S72	Fracture of femur
				M16	Osteoarthritis of hip
				M96	Intraoperative and postprocedural complications and disorders of musculoskeletal system, not elsewhere classified
6	ลื่นล้มในห้องนอน ปวดสะโพกซ้าย เดินไม่ได้	S72	Fracture of femur	S72 ✓	Fracture of femur
				M17	Osteoarthritis of knee
				M51	Thoracic, thoracolumbar, and lumbosacral intervertebral disc disorders
				M48	Other spondylopathies
				S32	Fracture of lumbar spine and pelvis
				M84	Disorder of continuity of bone
				M23	Internal derangement of knee
7	อุบัติเหตุมอเตอร์ไซต์ มีขาขวาผิดรูป มีแผลที่เข้าขวา	S72	Fracture of femur	S82	Fracture of lower leg including ankle
				S72 ✓	Fracture of femur
				M17	Osteoarthritis of knee
				M43	Other deforming

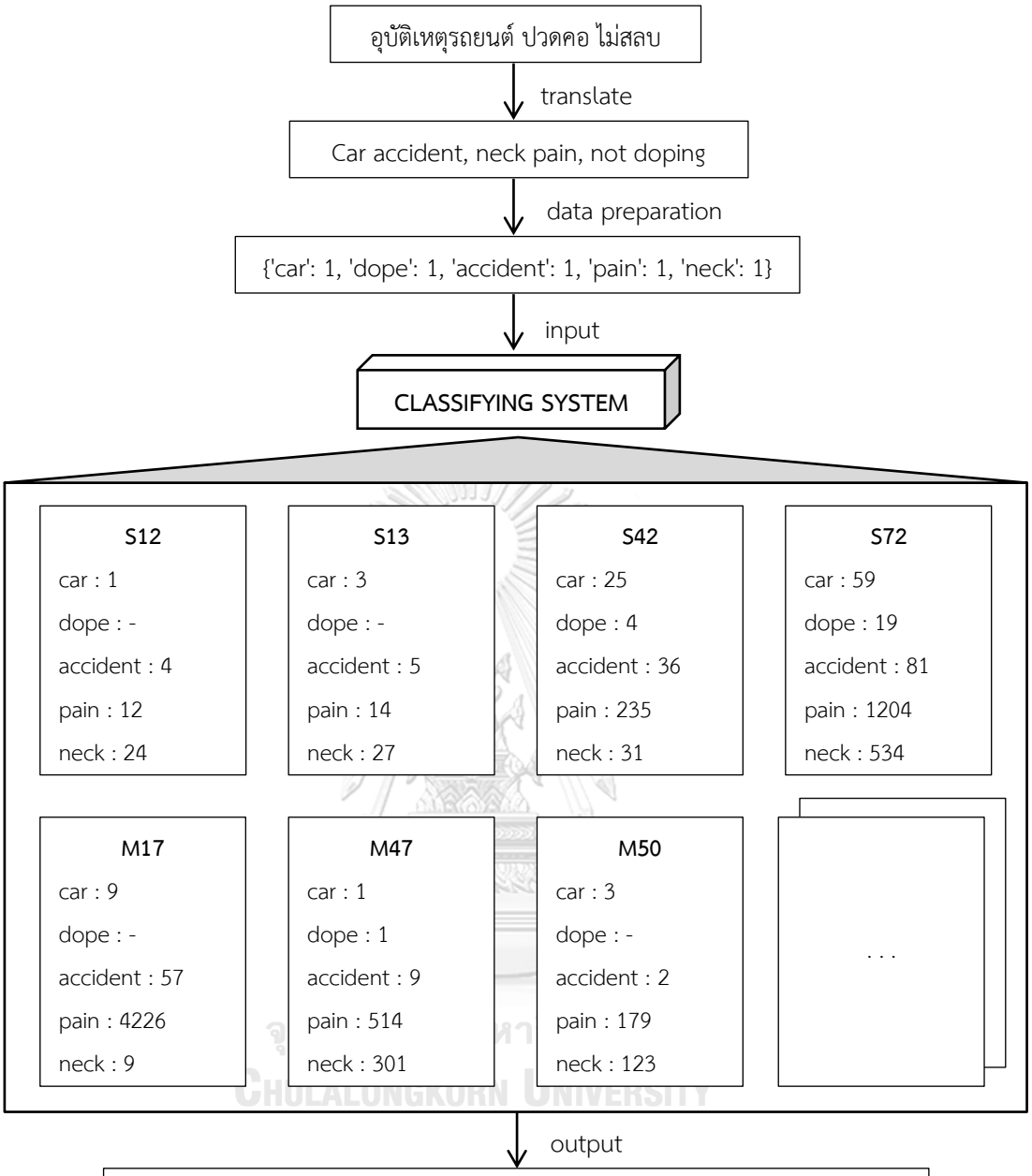
ลำดับ	บันทึกของแพทย์	เวชระเบียน		แบบจำลอง	
		รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส	รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส
					dorsopathies
				M51	Thoracic, thoracolumbar, and lumbosacral intervertebral disc disorders
				M84	Disorder of continuity of bone
				S86	Injury of muscle fascia and tendon at lower leg level
8	ล้มก้นกระแทกพื้น ปวดสะโพกขวา เดินไม่ได้	S72	Fracture of femur	S72 ✓	Fracture of femur
				T84	Complications of internal orthopedic prosthetic devices implants and grafts
				M16	Osteoarthritis of hip
				S73	Dislocation and sprain of joint and ligaments of hip
				M87	Osteonecrosis
				M25	Other joint disorder, not elsewhere classified
				M43	Other deforming dorsopathies
9	เจ็บสะโพกทั้ง 2 ข้าง ปวดมากเวลาเดิน ปวดข้างซ้ายมากกว่าขวา เดินใช้	M16	Osteoarthritis of hip	M16 ✓	Osteoarthritis of hip
				M87	Osteonecrosis
				M17	Osteoarthritis of knee
				M48	Other spondylopathies

ลำดับ	บันทึกของแพทย์	เวชระเบียน		แบบจำลอง	
		รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส	รหัสไอซีดีเทนซีเอ็ม	รายละเอียดของรหัส
	walker			M43	Other deforming dorsopathies
				M19	Other and unspecified osteoarthritis
				M25	Other joint disorder, not elsewhere classified
10	อุบัติเหตุรถยนต์ ปวดคอ ไม่สลบ	S13	Dislocation and sprain of joints and ligaments at neck level	S72	Fracture of femur
				M47	Spondylosis
				M50	Cervical disc disorders
				S42	Fracture of shoulder and upper arm
				S12	Fracture of cervical vertebra and other parts of neck
				S13 ✓	Dislocation and sprain of joints and ligaments at neck level
				M17	Osteoarthritis of knee

จากตารางที่ 12 สามารถแสดงตัวอย่างการทำงานของแบบจำลองในกรณีที่ข้อมูลเข้าเป็นภาษาอังกฤษได้ ดังรูปที่ 35 และในกรณีที่ข้อมูลเข้าเป็นภาษาไทยได้ ดังรูปที่ 36 ซึ่งผลลัพธ์ที่ได้จากแบบจำลอง นอกจากจะพิจารณาจากความถี่ของคำที่ปรากฏแล้วยังพิจารณาจากความสัมพันธ์ของคำที่ปรากฏขึ้นพร้อมกันด้วย



รูปที่ 35 ตัวอย่างการทำงานของแบบจำลอง กรณีที่ข้อมูลเข้าเป็นภาษาอังกฤษ



รูปที่ 36 ตัวอย่างการทำงานของแบบจำลอง กรณีที่ข้อมูลเข้าเป็นภาษาไทย

บทที่ 5
บทสรุป

5.1 สรุปผลวิทยานิพนธ์

งานวิจัยนี้ได้ทำการศึกษา วิเคราะห์ และพัฒนาแบบจำลองเพื่อจำแนกประเภทโรคจากอาการ โดยนำการทำเหมืองข้อความมาประยุกต์ใช้ เพื่อช่วยเพิ่มประสิทธิภาพในการวินิจฉัยโรคของแพทย์ และช่วยให้แพทย์สามารถเข้าถึงรหัสไอซีดีเทนซีเอ็มได้สะดวกยิ่งขึ้น จากการกรอกข้อมูลอาการให้กับแบบจำลอง จากนั้นแบบจำลองจะแสดงรายชื่อโรคที่มีความน่าจะเป็นเรียงตามลำดับจากมากไปน้อย พร้อมทั้งแสดงรหัสไอซีดีเทนซีเอ็มกำกับในทุก ๆ ชื่อโรค ซึ่งแบบจำลองในงานวิจัยนี้จะได้มาจากการเปรียบเทียบแบบจำลองที่ถูกสร้างขึ้นด้วยตัวจำแนกประเภทชนิดต่าง ๆ และนำผลลัพธ์ที่ได้จากแบบจำลองแต่ละตัวมาเปรียบเทียบกัน เพื่อหาแบบจำลองที่มีความเหมาะสมที่สุดสำหรับงานวิจัยนี้ โดยพบว่าแบบจำลองที่สร้างด้วยอัลกอริทึมของโครงข่ายประสาทเทียมเป็นแบบจำลองที่มีความเหมาะสมที่สุด เนื่องจากให้ค่าพื้นที่ใต้เส้นโค้งและค่าการระลึกได้สูงสุด

5.2 ปัญหาและข้อจำกัดในการทำวิทยานิพนธ์

1. ข้อมูลที่งานวิจัยนี้นำมาใช้ในการทดสอบแบบจำลอง เป็นข้อมูลจากแผนกออโธปิดิกส์ โรงพยาบาลจุฬาลงกรณ์เท่านั้น
2. ข้อมูลที่นำมาใช้ในการสร้างแบบจำลอง เป็นข้อมูลของโรคในหมวดกระดูกและกล้ามเนื้อและโรคที่พบในเวชระเบียนผู้ป่วยของแผนกออโธปิดิกส์ โรงพยาบาลจุฬาลงกรณ์เท่านั้น
3. แบบจำลองรองรับข้อมูลที่เป็นตัวอักษรภาษาไทยหรือภาษาอังกฤษเท่านั้น
4. แบบจำลองสามารถแสดงรายชื่อโรคที่มีความน่าจะเป็นไม่เกิน 10 อันดับแรก
5. กลุ่มผู้ใช้งานแบบจำลองจำกัดอยู่ในกลุ่มแพทย์เท่านั้น

5.3 แนวทางในการปรับปรุงวิทยานิพนธ์

ปรับปรุงให้แบบจำลองสามารถนำข้อมูลที่เป็นตัวเลขซึ่งอาจบ่งบอกถึงอายุ ความดันโลหิต อัตราการเต้นหัวใจ อุณหภูมิร่างกาย เป็นต้น มาใช้ประโยชน์ได้ เพื่อช่วยเพิ่มประสิทธิภาพในการทำนายของแบบจำลอง

บรรณานุกรม

1. Swaminath, G. and R. Raguram, *Medical errors - I : The problem.* Indian J Psychiatry, 2010. **52**(2): p. 110-2.
2. Sharma, H., S. Bhagat, and W. Gaine, *Reducing Diagnostic Errors in Musculoskeletal Trauma by Reviewing Non-Admission Orthopaedic Referrals in the Next-Day Trauma Meeting.* Annals of The Royal College of Surgeons of England, 2007. **89**(7): p. 692-695.
3. Jatunarapit, P., K. Piromsopa, and C. Charoeanlap, *Development of thai text-mining model for classifying ICD-10 TM*, in *2016 8th International Conference on Electronics, Computers and Artificial Intelligence*. 2016: Ploiesti, Romania. p. 1-6.
4. พวงทอง ไกรพิบูลย์. การวินิจฉัยโรค. [cited 2017 15 June]; Available from: http://haamor.com/th/การวินิจฉัยโรค/.
5. สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์. บทความรหัส *ICD-10* ที่เป็นมาตรฐานคืออะไร. [cited 2017 15 June]; Available from: https://www.etda.or.th/content/1231.html.
6. World Heatlh Organization. *Classification of Diseases (ICD).* [cited 2017 15 June]; Available from: http://apps.who.int/classifications/icd10/browse/2010/en.
7. ญาใจ ลิ่มปิยะกรณ์, ตำราวิศวฯ *100* ปี การทำเหมืองข้อมูล. 2556, กรุงเทพฯ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย. 245.
8. พรพล ธรรมรงค์รัตน์, การจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะและซัพพอร์ตเวกเตอร์แมชชีน. 2552, มหาวิทยาลัยสงขลานครินทร์.
9. วรรณพงษ์ ภัททิยไพบูลย์. *Machine Learning* คืออะไร *?* [cited 2017 10 July]; Available from: https://python3.wannaphong.com/2016/01/machine-learning-คืออะไร.html.
10. Zhu, F., et al., *Biomedical text mining and its applications in cancer research.* Journal of Biomedical Informatics, 2013. **46**(2): p. 200-211.
11. Cormack, J., et al., *Agile text mining for the 2014 i2b2/UTHealth Cardiac risk factors challenge.* Journal of Biomedical Informatics, 2015. **58**: p. S120-S127.
12. Martinez, D. and Y. Li, *Information extraction from pathology reports in a hospital setting*, in *Proceedings of the 20th ACM international conference on*

Information and knowledge management. 2011, ACM: Glasgow, Scotland, UK. p. 1877-1882.

13. Beheshti, I. and H. Demirel, *Feature-ranking-based Alzheimer's disease classification from structural MRI.* Magnetic Resonance Imaging, 2016. **34**(3): p. 252-263.
14. Yepes, A.J. and R. Berlanga, *Knowledge based word-concept model estimation and refinement for biomedical text mining.* J. of Biomedical Informatics, 2015. **53**(C): p. 300-307.
15. Wu, C., et al., *Open data mining for Taiwan's dengue epidemic.* Acta Tropica, 2018. **183**: p. 1-7.
16. Lucini, F.R., et al., *Text mining approach to predict hospital admissions using early medical records from the emergency department.* International Journal of Medical Informatics, 2017. **100**: p. 1-8.
17. Kocbek, S., et al., *Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources.* Journal of Biomedical Informatics, 2016. **64**: p. 158-167.
18. Pereira, L., et al., *ICD9-based Text Mining Approach to Children Epilepsy Classification.* Procedia Technology, 2013. **9**: p. 1351-1360.
19. Pletscher-Frankild, S., et al., *DISEASES: Text mining and data integration of disease–gene associations.* Methods, 2015. **74**: p. 83-89.
20. Ailem, M., et al., *Unsupervised text mining for assessing and augmenting GWAS results.* Journal of Biomedical Informatics, 2016. **60**: p. 252-259.
21. Almehmadi, A., Z. Joudaki, and R. Jalali, *Language usage on Twitter predicts crime rates*, in *Proceedings of the 10th International Conference on Security of Information and Networks*. 2017, ACM: Jaipur, India. p. 307-310.
22. Ryoo, K. and S. Moon, *Inferring Twitter user locations with 10 km accuracy*, in *Proceedings of the 23rd International Conference on World Wide Web*. 2014, ACM: Seoul, Korea. p. 643-648.
23. Gola, J., et al., *Advanced microstructure classification by data mining methods.* Computational Materials Science, 2018. **148**: p. 324-335.
24. Jonnagaddala, J., et al., *Coronary artery disease risk assessment from*

unstructured electronic health records using text mining. Journal of Biomedical Informatics, 2015. **58**: p. S203-S210.

25. Evgeniy Tatarkin. *Pomp*. [cited 2018 30 May]; Available from: http://pomp.readthedocs.io/en/latest/.
26. Fawcett, T., *An introduction to ROC analysis.* Pattern Recognition Letters, 2006. **27**(8): p. 861-874.
27. Bird, S., E. Klein, and E. Loper, *Natural Language Processing with Python*. 2009: O'Reilly Media Inc.
28. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.
29. Python Software Foundation. *goslate 1.5.1*. [cited 2018 1 June]; Available from: https://pypi.org/project/goslate/.
30. Sebastiani, F., *Machine learning in automated text categorization.* ACM Comput. Surv., 2002. **34**(1): p. 1-47.
31. Alkaline Software. *2018 ICD-10-CM Codes*. [cited 2018 1 November]; Available from: https://www.icd10data.com.
32. Farlex. *Acronyms and Abbreviations*. [cited 2018 1 November]; Available from: https://acronyms.thefreedictionary.com.

ภาคผนวก

ภาคผนวก ก

รหัสไอซีดีเทนซีเอ็มทั้งหมดที่นำมาใช้ในงานวิจัย [31]

ลำดับ	หมวดหมู่โรค		รหัสไอซีดีเทนซีเอ็มที่นำมาใช้
	รายละเอียด	รหัสไอซีดีเทนซีเอ็ม	
1	Certain infectious and parasitic diseases	A00-B99	A17, A18, A46, A52, A80, B18, B20, B37, B47, B91
2	Neoplasms	C00-D49	C15, C16, C18, C20, C22, C24, C34, C40, C41, C43, C44, C47, C48, C49, C50, C53, C61, C70, C73, C75, C76, C77, C78, C79, C82, C83, C84, C85, C90, C91, C96, D16, D17, D18, D21, D23, D35, D36, D48
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50-D89	D61, D64, D66, D68, D75, D76
4	Endocrine, nutritional and metabolic diseases	E00-E89	E11, E22, E46, E64, E75, E78, E83, E87
5	Mental, Behavioral and	F01-F99	F01, F80

ลำดับ	หมวดหมู่โรค		รหัสไอซีดีเทนซีเอ็ม
	รายละเอียด	รหัสไอซีดีเทนซีเอ็ม	ที่นำมาใช้
	Neurodevelopmental disorders		
6	Diseases of the nervous system	G00-G99	G03, G04, G06, G20, G31, G35, G47, G52, G54, G56, G57, G58, G62, G71, G80, G81, G82, G83, G95, G96, G97, G98
7	Diseases of the eye and adnexa	H00-H59	H16, H26
8	Diseases of the ear and mastoid process	H60-H95	H81, H91
9	Diseases of the circulatory system	I00-I99	I05, I07, I10, I20, I21, I25, I33, I48, I50, I62, I63, I69, I70, I72, I74, I77, I80, I87, I89, I97
10	Diseases of the respiratory system	J00-J99	J18, J45, J96
11	Diseases of the digestive system	K00-K95	K22, K56, K70, K80, K81
12	Diseases of the skin and subcutaneous tissue	L00-L99	L02, L03, L04, L08, L50, L59, L60, L72, L81, L89, L90, L91, L92, L97, L98
13	Diseases of the musculoskeletal system and connective tissue	M00-M99	M00, M01, M02, M04, M05, M06, M07, M08, M10,

ลำดับ	หมวดหมู่โรค		รหัสไอซีดีเทนซีเอ็มที่นำมาใช้
	รายละเอียด	รหัสไอซีดีเทนซีเอ็ม	
			M11, M12, M13, M14, M1A, M15, M16, M17, M18, M19, M20, M21, M22, M23, M24, M25, M26, M27, M30, M31, M32, M33, M34, M35, M36, M40, M41, M42, M43, M45, M46, M47, M48, M49, M50, M51, M53, M54, M60, M61, M62, M63, M65, M66, M67, M70, M71, M72, M75, M76, M77, M79, M80, M81, M83, M84, M85, M86, M87, M88, M89, M90, M91, M92, M93, M94, M95, M96, M97, M99
14	Diseases of the genitourinary system	N00-N99	N18, N19, N31, N32
15	Congenital malformations, deformations and chromosomal	Q00-Q99	Q06, Q27, Q28, Q65, Q66, Q67,

ลำดับ	หมวดหมู่โรค		รหัสไอซีดีเทนซีเอ็ม
	รายละเอียด	รหัสไอซีดีเทนซีเอ็ม	ที่นำมาใช้
	abnormalities		Q68, Q69, Q70, Q71, Q72, Q74, Q76, Q77, Q78, Q79, Q82, Q87
16	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	R00-R99	R19, R22, R26, R60, R94
17	Injury, poisoning and certain other consequences of external causes	S00-T88	S00, S02, S04, S06, S09, S12, S13, S14, S20, S22, S24, S25, S27, S29, S30, S32, S33, S34, S35, S36, S37, S38, S39, S40, S41, S42, S43, S44, S45, S46, S48, S49, S51, S52, S53, S54, S55, S56, S58, S60, S61, S62, S63, S64, S65, S66, S67, S68, S69, S70, S72, S73, S76, S79, S80, S81, S82, S83, S84, S85, S86, S87, S88,

ลำดับ	หมวดหมู่โรค		รหัสไอซีดีเทนซีเอ็ม
	รายละเอียด	รหัสไอซีดีเทนซีเอ็ม	ที่นำมาใช้
			S89, S91, S92, S93, S96, S97, S98, S99, T14, T23, T63, T79, T81, T82, T84, T85, T86, T87, T88
18	Factors influencing health status and contact with health services	Z00-Z99	Z01, Z03, Z09, Z42, Z47, Z48, Z73, Z88, Z95, Z96
	รวม		332

ภาคผนวก ข

ตัวอย่างข้อมูลบนเวชระเบียนและรหัสไอซีดีเทนซีเอ็มที่นำมาใช้ในงานวิจัย

ลำดับ	บันทึกของแพทย์	รหัสไอซีดีเทนซีเอ็ม
1	ก้อนที่เข่าขวามา 3 ปี โตขึ้นเรื่อย ๆ	L72
2	ปวดข้อมือ 8 เดือน 1 เดือน มีก้อนโตขึ้นที่ข้อมือ ขยับข้อมือไม่ได้ ปวดจึงมารพ .ill defined mass 7 cm at dorsal view of wrist Rt	M00
3	3 เดือน ขณะทำงานถูกเครื่องจักรทับมือซ้าย มีนิ้วกลางซ้ายขาด นิ้วนางผิดรูปไป ปวดเป็นแผล	M13
4	10 ปี ปวดเข่าขวามากกว่าซ้าย เดินได้ 50 มเคยฉีดยาเ .ข้าข้อเข่าข้างละ 5 เข็ม ไม่ดีขึ้น	M17
5	ปวดเข่าซ้าย 5 เดือน มีอาการปวดเข่าซ้ายเวลาเดิน แต่ยังเดินได้ประมาณ 2 กม แต่ถ้าเดินขึ้นบันไดจะปวดมากขึ้น ไม่ปวดหลัง มีเสียงกรอบแกรบ	M17
6	8 ปี ปวดเข่าขวาเป็นๆ หายๆ 3 ปี เริ่มมีอาการปวดเข่าซ้ายด้วย แต่น้อยกว่าเข่าขวา เข่าขวาได้รับการฉีดยาเข้าข้อเข่าอาการปวดดีขึ้น 3 เดือน อาการปวดเข่าขวาหายไป ปวดเข่าซ้ายมาก เดินลำบาก	M17
7	ปวดเข่าเป็น ๆ หาย ๆ 1 ปี ปวดมากขึ้น 3 เดือน เดินไกล ๆ แล้วปวดมาก ทานยาไม่ดีขึ้น	M17
8	4 เดือน ปวดเข่าซ้ายเป็น ๆ หาย ๆ เข่าบวม มีไข้บางครั้ง งอเข่าได้ มีอาการปวดเวลาเดิน เคยเจาะระบายน้ำในเข่าแล้วดีขึ้น	M17
9	case OA knee both แต่มีอาการปวดเข่าซ้ายมาก และเข่าซ้ายโก่งมาก แพทย์นัดมาผ่าตัด	M17
10	ผู้ป่วยหญิงไทย อายุ 70 ปี มาพบแพทย์ด้วยเรื่องปวดเข่าซ้ายมากขึ้นเวลาเดิน และเดินขึ้นบันได เมื่อพักอาการจะดีขึ้น	M17
11	7-8 ปีก่อน เดินได้ รองเท้าส้นสูงเท้าพลิก หลังจากนั้นปวดเข่าซ้ายมาก เดินไม่ได้ประมาณ 1 wk. 5-6 ปีก่อน ปวดมากขึ้น pain when activity	M23

ลำดับ	บันทึกของแพทย์	รหัสไอซีดีเทนซีเอ็ม
12	ผู้ป่วยปวดหลังร้าวลงขาซ้าย 5 ปี ปวดขณะเดิน ปวดที่หลังมากกว่าขา ยืนนานจะปวด เดินกระเผก ปัสสาวะ ถ่ายอุจจาระปกติ มีโรคประจำตัวเป็น ความดันโลหิตสูง	M41
13	case ผู้ป่วยเด็กหญิงไทยอายุ12ปี CC: หลังคด 1 ปี PI: แม่สังเกตว่าไหล่ไม่เท่ากัน เดินได้ปกติ ไม่ปวด FC-I มีประจำเดือนตั้งแต่อายุ 11 ปี PH: ปฏิเสธประวัติแพ้ยา	M41
14	ปวดหลัง 3 ปี ยกของแล้วมีเสียงลั่นที่หลัง แล้วมีปวดหลังร้าวถึงปลายเท้าทั้งสองข้าง เดินได้ 10 เมตร จะมีอาการปวดร้าวลงปลายเท้า ขวามากกว่าซ้าย นั่งพักแล้วดีขึ้น	M43
15	ปวดหลัง 10 ปี เตะฟุตบอลล้มแล้วมีเสียงลั่นที่หลัง 1 ปี ปวดหลังมากขึ้น ปวดร้าวลงขาขวามากกว่าซ้าย motor gr 5 all	M48
16	ปวดหลังร้าวลงต้นขาขวา 6 เดือน กินยายังไม่ดีขึ้น motor grade V all SLRT negative	M48
17	case ผู้ป่วยหญิงปวดข้อมือซ้ายมา 1 ปี เป็นๆหายๆ	M70
18	case ปวดคอร้าวลงแขน 2 ปี เวลายกแขนแล้วมีอาการปวด	M75
19	1 อาทิตย์ ขณะวิ่งออกกำลังกาย ปวดสะโพกซ้าย แต่ยังออกกำลังกายได้ ปวดมากขึ้นเรื่อยๆ เดินลงน้ำหนักได้ แต่ปวด	M84
20	1 ปี ประสบอุบัติเหตุ แขนขวาหักใส่เฝือก แล้วแขนผิดรูป แพทย์นัดมาผ่าตัด	M84
21	3 ปี ปวดสะโพกขวาเวลาเดินไกล ๆ เดินเซ แพทย์นัดมาผ่าตัด	M87
22	MCA accident 2 mo PTA มีปวดต้นคอ อ่อนแรงและชาแขน + 2 ข้างมากขึ้น อุจจาระปัสสาวะปกติ	S12
23	ผู้ป่วยหญิงไทย 60 ปี 1 วันก่อน กำลังข้ามถนนที่ปากซอยสุขุมวิท 101 มีรถ taxi ชนเข้าทีหลังด้านขวาระดับบั้นเอว หลังจากนั้นปวดหลังมาก ลุกไม่ไหว ศีรษะไม่กระแทกพื้น ไม่สลบ รู้สึกตัวดี ไป admit รพ เวชธานี.1 คืน X-ray มีกระดูกสันหลังหัก จึงขอ refer มา	S22
24	ขับ MC ชน taxi สลบตกลงในน้ำ ปวดสะโพกซ้าย เดินไม่ได้ ไม่ปวดท้อง	S32

ลำดับ	บันทึกของแพทย์	รหัสไอซีดีเทนซีเอ็ม
25	case เด็กชายไทย อายุ 7 ปี no underlying disease CC: เจ็บศอกซ้าย 3 ชั่วโมงก่อนมา รพ .PI: 3 hrs PTA โหนแป้นบาสแล้วตก เอาศอกซ้ายกระแทกพื้น ปวดศอกซ้าย บวม ขยับมือได้	S42
26	2 hr ตกจากเครื่องเล่นสูงประมาณครึ่งเมตร มือขวาเหยียดยันพื้น ไม่สลบ ข้อศอกขวาเจ็บ ผิดรูป	S42
27	2 วัน ตกจักรยาน ปวดศอกขวา	S42
28	ผู้ป่วยเด็กชายอายุ 9 ปี เล่นบนรถกระบะ พลัดตกลงมาโดยเอาศอกซ้ายยันพื้น รู้สึกเจ็บบริเวณศอกซ้ายมาก ขยับแขนซ้ายไม่ได้	S42
29	เล่นฟุตบอลล้ม แขนซ้ายปวดบวมผิดรูป มี fracture both bone Lt.forearm	S52
30	Case เด็กชายชาวพม่า อายุ 12 ปี 6 hr วิ่งเล่นหกล้ม แขนซ้ายลงกระแทกพื้น	S52
31	เตะบอล วิ่งชนกับเพื่อนแล้วหงายหลัง ล้มแขนซ้ายยันพื้น ไม่สลบ ไม่บาดเจ็บบริเวณอื่น ปวดแขนซ้ายมาก แขนซ้ายผิดรูป film: Fx both bone Lt. forearm Dx.: Closed Fx. both bone Lt. forearm Post op.: ปวดแผลเล็กน้อย นิ้วชี้กับนิ้วมือขวา ชาเล็กน้อย	S52
32	18 วัน ขี่มอเตอร์ไซด์ชนรถพ่วง มีปวดแขนซ้าย ไม่มีชาที่มือ	S52
33	พัดลมรถยนต์บาดมือ มีแผล skin loss at lt index film: no fx. --> แผลสกปรก debride + suture หลวม มี skin loss บริเวณ tip not expose bone	S61
34	โดนทำร้ายร่างกายมีแผล cut wound Rt dorsal hand, complete tear EPB, open fracture trapezium repair tendon & Rt volar slab	S66
35	อุบัติเหตุรถมอเตอร์ไซด์สลบ มีปวดขาซ้าย แผลฉีกขาดและแผลถลอกที่คอซ้าย	S72

ภาคผนวก ค
ตัวอย่างข้อมูลบนเว็บไซต์สาธารณะและรหัสไอซีดีเทนซีเอ็มที่นำมาใช้ในงานวิจัย

ลำดับ	ข้อมูลเว็บไซต์	รหัสไอซีดีเทนซีเอ็ม
1	Keratitis is the medical term for inflammation of the cornea The cornea is the dome shaped window in the front of the eye When looking at a person is eye one can see the iris and pupil through the normally clear cornea The cornea bends light rays as a result of its curved shape and accounts for approximately two thirds of the eye is total optical power with the lens of the eye contributing the remaining one third Only the very thin tear film lies between the front of the cornea and our environment The cornea is about 0 5 millimeter thick The back of the cornea is bathed in the aqueous fluid that fills the anterior chamber of the eye The cornea has a diameter of about 13 millimeters and together with the sclera the white part of the eye forms the entire outer coat of the eye	H16
2	An arthropathy is a disease of a joint Arthritis is a form of arthropathy that involves inflammation of one or more joints while the term arthropathy may be used regardless of whether there is inflammation or not Spondylarthropathy is any form of arthropathy of the vertebral column Related problems Arthropathy may also include joint conditions caused by physical trauma to joints but is traditionally used to describe the following conditions Reactive arthropathy M02 M03 is caused by an infection but not a direct infection of the synovial space See also Reactive arthritis Enteropathic arthropathy M07 is caused by colitis and related conditions Crystal arthropathy also known as crystal arthritis M10 M11 involves the deposition of crystals in the joint In gout the crystal is uric acid In pseudogout chondrocalcinosis calcium pyrophosphate deposition disease the crystal is calcium pyrophosphate Diabetic arthropathy M14 E10 E14 is caused by diabetes Neuropathic arthropathy M14 is associated with a loss of sensation Signs Bone erosions by rheumatoid arthritis 5 Arthralgia joint pain is a common but non specific sign of arthropathy Other signs may include Decreased range of motion Stiffness Effusion	M12

ลำดับ	ข้อมูลเว็บไซต์	รหัสไอซีดีเทนซีเอ็ม
	Pneumarthrosis air in a joint which is also a common normal finding Bone erosion Systemic signs of arthritis such as fatigue	
3	What are the symptoms and signs of chondromalacia patella The symptoms of chondromalacia patella are generally a vague discomfort of the inner front of the knee aggravated by activity running jumping climbing or descending stairs or by prolonged sitting with knees in a moderately bent position the so called theater sign of pain upon arising from a desk or theater seat Some patients may also have a vague sense of tightness or fullness in the knee area Occasionally if chronic symptoms are ignored the associated loss of quadriceps thigh muscle strength may cause the leg to give out Besides an obvious reduction in quadriceps muscle mass mild swelling of the knee area may occur	M22
4	What Are the Symptoms of Necrotizing Vasculitis Because this condition affects your blood vessels symptoms might occur in various parts of your body There s no single set of symptoms that can definitely indicate you have necrotizing vasculitis You might notice initial symptoms on your own without a medical test These include chills fatigue fever weight loss Other early symptoms are only detectable through a blood test These include anemia and leukocytosis which involves having a high number of white blood cells As the disease progresses symptoms can worsen and become more varied Your specific symptoms depend on what parts of your body affected You may have pain skin discoloration lesions which are usually seen on the legs ulcers In some cases the condition may be limited to your skin In other cases you might develop kidney damage or bleeding in your lungs If your brain is affected you may have difficulty swallowing speaking or moving	M31
5	These conditions nearly all present with an insidious onset of pain referred to the location of the bony damage Some notably Kienbock disease of the wrist may involve considerable swelling and Legg Calv Perthes disease of the hip causes the victim to limp The spinal form Scheuermann disease may cause bending or kyphosis of the upper spine giving a hunch back appearance	M42
6	Spondylosis is caused from years of constant abnormal pressure	M47

ลำดับ	ข้อมูลเว็บไซต์	รหัสไอซีดีเทนซีเอ็ม
	caused by joint subluxation stress induced by sports acute and repetitive trauma or poor posture being placed on the vertebrae and the discs between them The abnormal stress causes the body to form new bone in order to compensate for the new weight distribution This abnormal weight bearing from bone displacement will cause spondylosis to occur Poor postures and loss of the normal spinal curves can lead to spondylosis as well Spondylosis can affect a person at any age however older people are more susceptible	
7	The symptoms of osteomalacia include the following Bones that fracture very easily are the most common symptom Another symptom is muscle weakness This happens because of problems at the location where the muscle attaches to bone You may have a hard time walking and may develop a waddling gait Bone pain especially in your hips is also a very common symptom This dull aching pain can spread from your hips to your lower back pelvis legs and even your ribs If you also have very low levels of calcium in your blood you may have irregular heart rhythms numbness around your mouth numbness in your arms and legs spasms in your hands and feet	M83
8	How Is Osteomalacia Diagnosed Blood tests that show the following can suggest you may have osteomalacia or another bone disorder low levels of vitamin D low levels of calcium low levels of phosphorus You may also be tested for alkaline phosphatase isoenzymes High levels of these indicate osteomalacia Another blood test can check your levels of parathyroid hormone High levels of this hormone suggest insufficient vitamin D and other related problems X rays and other imaging tests can show small cracks in the bones throughout your body These cracks are called Looser transformation zones Fractures can begin there with even small injuries A bone biopsy may be required to definitively diagnose osteomalacia A needle is inserted through your skin and muscle and into your bone to obtain a small sample That sample is put on a slide and examined under a microscope Usually an X ray and blood tests are enough to make a diagnosis and a bone biopsy	M83

ลำดับ	ข้อมูลเว็บไซต์	รหัสไอซีดีเทนซีเอ็ม
	is not necessary	
9	A femoral fracture is a bone fracture that involves the femur They are typically sustained in high impact trauma such as car crashes due to the large amount of force needed to break the bone Fractures of the diaphysis or middle of the femur are managed differently from those at the head neck and trochanter see hip fractures	S72
10	A patellar dislocation is a knee injury in which the patella kneecap slips out of its normal position Often the knee is partly bent painful and swollen The patella is also often felt and seen out of place Complications may include a patella fracture or arthritis A patellar dislocation typically occurs when the knee is straight and the lower leg is bent outwards when twisting Occasionally it occurs when the knee is bent and the patella is hit Commonly associated sports include soccer gymnastics and ice hockey Dislocations nearly always occur away from the midline Diagnosis is typically based on symptoms and supported by X rays Reduction is generally done by pushing the patella towards the midline while straightening the knee After reduction the leg is generally splinted in a straight position for a few weeks This is then followed by physical therapy Surgery after a first dislocation is generally of unclear benefit Surgery may be indicated in those who have broken off a piece of bone within the joint or in which the patella has dislocated multiple times Patellar dislocations occur in about 6 per 100000 people per year They make up about 2 of knee injuries It is most common in those 10 to 17 years old Rates in males and females are similar Recurrence after an initial dislocation occurs in about 30 of people	S83

ภาคผนวก ง

ตัวอย่างอักษรย่อของแพทย์ที่นำมาใช้ในงานวิจัย [32]

ลำดับ	อักษรย่อ	คำเต็มของอักษรย่อ
1	A&E	Accident and Emergency
2	AAMRS	Automated Ambulatory Medical Record System
3	ACL	Anterior Cruciate Ligament
4	ADM	Acute Disseminated Myelitis
5	ADSC	Adenosquamous Carcinoma
6	ADSD	Adductor Spasmodic Dysphonia
7	ADT	Androgen Deprivation Therapy
8	ADTA	Anterior Deep Temporal Artery
9	ADU	Active Duodenal Ulcer
10	AE	Above Elbow
11	AEE	Allergic Eosinophilic Esophagitis
12	AEH	Acute Epidural Hematoma
13	AK	Above Knee
14	AND	Adnexa
15	CRIF	Closed Reduction and Internal Fixation
16	D	Day
17	EHL	Extensor Hallucis Longus
18	F/U	Follow up
19	FX	Fracture
20	HNP	Herniated Nucleus Pulposus
21	IMP	Impression
22	LLC	Long Leg Cast
23	Lt	Left
24	MCA	Motor Cycle Accident
25	MPS	Myofascial Pain Syndrome
26	OA	Osteoarthritis

ลำดับ	อักษรย่อ	คำเต็มของอักษรย่อ
27	PCL	Posterior Cruciate Ligament
28	PE	Physical Examination
29	PF	Patellofemoral
30	PI	Present Illness
31	PTA	Prior To Admits
32	ROM	Range Of Motion
33	Rt	Right
34	S/P	Status Post
35	THR	Total Hip Replacement
36	TIA	Transient Ischemic Attack
37	TIBC	Total Iron-Binding Capacity
38	TID	Three Times A Day
39	TIP	Tubularized Incised Plate
40	TIPS	Transjugular Intrahepatic Portosystemic Shunt
41	TKA	Total Knee Arthroplasty
42	TLIF	Transforaminal Lumbar Interbody Fusion
43	TLSO	Thoracolumbosacral Orthosis
44	TMJ	Temporomandibular Joint
45	TMP	Trimethoprim
46	TMSE	Thai Mental State Examination
47	TMT	Tarsometatarsal
48	TN	Tension
49	TNTC	Too Numerous To Count (microbiology)
50	UA	Unstable Angina
51	VA	Visual Acuity
52	VATS	Video-Assisted Thoracoscopic Surgery
53	VS	Vital Signs
54	WK	Week

ภาคผนวก จ

ผลการตรวจสอบคำสำคัญที่เกี่ยวข้องกับโรคที่ได้จากข้อมูลบนเว็บไซต์

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
1	M01	arthritis	✓	✓	✓	✓	✓
		infection	✓	✓	✓	✓	✓
		joint	✓	✓	✓	✓	✓
		disease	✓			✓	
		view					
		helminth	✓	✓			
		parasite	✓	✓	✓		✓
		inflammation	✓	✓	✓		✓
		treatment					
		symptom					
2	M02	arthritis	✓	✓	✓	✓	✓
		reactive	✓	✓	✓		✓
		disease				✓	
		infection			✓	✓	✓
		patient					
		arthropathy	✓	✓	✓	✓	✓
		joint	✓		✓		✓
		syndrome	✓				
		symptom					
		cause					
3	M04	syndrome	✓			✓	
		autoinflammatory	✓	✓	✓	✓	✓
		disease				✓	
		patient					
		fever	✓				
		disorder					
		inflammatory	✓	✓	✓		✓
		arthritis	✓	✓	✓		✓

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
		cap					
		mutation					
4	M07	arthritis	✓	✓	✓	✓	✓
		disease	✓			✓	
		symptom					
		IBD	✓	✓	✓		✓
		bowel	✓		✓		✓
		associate	✓		✓		✓
		inflammatory	✓	✓	✓		✓
		patient					
		enteropathic	✓		✓		✓
		joint	✓		✓		✓
5	M11	crystal	✓	✓	✓		✓
		arthropathy	✓	✓	✓	✓	✓
		disease				✓	
		syndrome					
		calcium	✓	✓	✓		✓
		pyrophosphate	✓	✓	✓		✓
		deposition	✓		✓		✓
		joint	✓		✓	✓	✓
		dihydrate					
		patient					
6	M12	arthropathy	✓	✓	✓	✓	✓
		disease				✓	
		unspecified	✓	✓	✓		✓
		site					
		joint	✓		✓	✓	✓
		shoulder			✓		✓
		condition					
		link					
		hand			✓		✓
		ankle			✓		✓

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
7	M13	cause	✓		✓		✓
		affect					
		disease			✓	✓	✓
		joint	✓	✓	✓	✓	✓
		hand					
		bone					
		treatment					
		arthritis	✓	✓	✓	✓	✓
		attack					
		knee					
8	M14	arthropathy	✓	✓	✓	✓	✓
		joint	✓	✓	✓	✓	✓
		arthritis	✓	✓	✓	✓	✓
		disease				✓	
		pain			✓		✓
		gout	✓				
		patient					
		foot					
		osteoarthritis	✓				
		rheumatoid	✓		✓		✓
9	M15	osteoarthritis	✓	✓	✓	✓	✓
		joint	✓	✓	✓	✓	✓
		disease				✓	
		polyosteoarthritis	✓	✓			
		pain	✓		✓		✓
		unspecified			✓		✓
		cause					
		cartilage					
		arthritis	✓	✓	✓	✓	✓
		polyarthritis	✓	✓	✓		✓
10	M16	use					
		hip	✓	✓	✓	✓	✓

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
		cause					
		exercise					
		treatment					
		joint	✓			✓	
		increase					
		bone					
		osteoarthritis	✓	✓	✓	✓	✓
		knee					
11	M20	toe	✓	✓	✓	✓	✓
		deformity	✓	✓	✓	✓	✓
		finger	✓	✓	✓	✓	✓
		foot	✓		✓		✓
		use					
		joint	✓		✓	✓	✓
		disease				✓	
		condition					
		diagnosis					
		disorder					
12	M21	deformity	✓	✓	✓	✓	✓
		limb	✓	✓	✓	✓	✓
		use					
		joint	✓				
		finger	✓		✓		✓
		disease				✓	
		diagnosis					
		toe	✓		✓		✓
		foot	✓		✓		✓
		musculoskeletal	✓	✓		✓	
13	M22	cause					
		use			✓		✓
		treat					
		knee			✓	✓	✓

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
		patella	✓	✓	✓	✓	✓
		exercise					
		treatment					
		help					
		injury	✓		✓		✓
		leg					
14	M23	cause					
		derangement	✓	✓	✓		✓
		knee	✓	✓	✓	✓	✓
		internal	✓		✓	✓	✓
		injury	✓		✓	✓	✓
		follow					
		sign					
		tear	✓		✓		✓
		ligament	✓		✓	✓	✓
		sport	✓				
15	M24	joint	✓	✓	✓	✓	✓
		derangement	✓	✓	✓	✓	✓
		specific					
		use					
		diagnosis					
		right					
		shoulder	✓				
		disorder			✓		✓
		musculoskeletal	✓	✓			
		tissue			✓	✓	✓
16	M25	disorder	✓		✓	✓	✓
		joint	✓	✓	✓	✓	✓
		right					
		use					
		shoulder	✓				
		abnormality	✓		✓	✓	✓

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
		deformity			✓		✓
		limb	✓		✓		✓
		foot					
		fistula					
17	M27	bone	✓		✓	✓	✓
		jaw	✓	✓	✓	✓	✓
		osteonecrosis	✓		✓		✓
		ONJ	✓		✓		✓
		lesion					
		cell					
		oral		✓	✓		✓
		disease				✓	
		marrow					
		infection	✓				
18	M30	polyarteritis	✓	✓	✓	✓	✓
		nodosa	✓	✓	✓	✓	✓
		blood	✓				
		artery	✓	✓	✓	✓	✓
		pan			✓		✓
		cause					
		kidney	✓				
		affect					
		disease			✓	✓	✓
		disorder				✓	
19	M32	SLE	✓	✓	✓		✓
		lupus	✓	✓	✓		✓
		disease	✓		✓	✓	✓
		cell					
		symptom				✓	
		erythematosus	✓	✓	✓		✓
		systemic	✓	✓	✓	✓	✓
		criterion			✓	✓	✓

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
		rash	✓		✓		✓
		antibody	✓				
20	M33	dermatomyositis	✓	✓	✓		✓
		muscle		✓	✓	✓	✓
		disease				✓	
		symptom					
		patient					
		myopathy	✓	✓	✓	✓	✓
		weakness			✓		✓
		polymyositis	✓		✓		✓
		myositis	✓		✓	✓	✓
		tissue		✓			
21	M34	scleroderma	✓	✓	✓		✓
		patient					
		systemic	✓	✓	✓	✓	✓
		disease				✓	
		skin	✓	✓	✓		✓
		cause					
		sclerosis	✓	✓	✓	✓	✓
		renal	✓		✓		✓
		blood	✓		✓		✓
		pulmonary			✓		✓
22	M42	spine	✓	✓	✓	✓	✓
		osteochondrosis	✓	✓	✓	✓	✓
		disease				✓	
		spinal	✓	✓	✓		✓
		disc	✓		✓		✓
		pain	✓		✓		✓
		back	✓		✓	✓	✓
		joint					
		vertebral	✓		✓		✓
		symptom				✓	

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
23	M45	cause					
		treatment					
		affect					
		joint	✓	✓	✓		✓
		disease	✓				
		low					
		exercise	✓				
		improve					
		spondylitis	✓	✓	✓	✓	✓
		bone			✓	✓	✓
24	M46	inflammatory	✓	✓	✓	✓	✓
		disease			✓	✓	✓
		spondylopathies	✓	✓	✓	✓	✓
		condition					
		tissue			✓		✓
		cause					
		diagnosis					
		musculoskeletal			✓		✓
		connective	✓		✓		✓
		spine			✓	✓	✓
25	M48	disease			✓		✓
		spondylopathies	✓	✓	✓	✓	✓
		use				✓	
		spinal	✓		✓		✓
		vertebra	✓		✓	✓	✓
		site					
		stenosis	✓				
		tissue					
		spine	✓		✓	✓	✓
		fracture					
26	M51	cause					
		condition					

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
		disc	✓	✓	✓	✓	✓
		thoracolumbar	✓	✓	✓	✓	✓
		disorder	✓			✓	
		lumbosacral	✓	✓	✓	✓	✓
		disease			✓		✓
		thoracic	✓	✓	✓		✓
		intervertebral	✓	✓	✓	✓	✓
		change					
27	M53	dorsopathies	✓	✓	✓	✓	✓
		disease			✓	✓	✓
		disorder	✓		✓		✓
		spinal	✓			✓	
		system			✓		✓
		classify					
		cervicocranial	✓				
		musculoskeletal			✓	✓	✓
		tissue			✓		✓
		cervicobrachial	✓				
28	M60	cause					
		muscle	✓	✓	✓	✓	✓
		disease				✓	
		myositis	✓	✓	✓	✓	✓
		affect					
		treatment					
		use					
		infection	✓	✓	✓	✓	✓
		myopathy	✓		✓		✓
		doctor					
29	M63	muscle	✓	✓	✓	✓	✓
		genetics	✓				
		disease				✓	
		muscular	✓		✓	✓	✓

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
		dystrophy	✓		✓		✓
		myopathy	✓		✓		✓
		tissue					
		disorder			✓	✓	✓
		soft					
		sarcoma					
30	M65	cause					
		tendon	✓	✓	✓	✓	✓
		joint	✓		✓	✓	✓
		tenosynovitis	✓	✓	✓	✓	✓
		infection	✓	✓	✓	✓	✓
		tissue			✓		✓
		synovitis	✓		✓	✓	✓
		therapy					
		injury					
		sheath			✓		✓
31	M66	rupture	✓	✓	✓	✓	✓
		tendon	✓	✓	✓	✓	✓
		spontaneous	✓	✓	✓	✓	✓
		tissue			✓		✓
		patient					
		disease				✓	
		synovium	✓	✓	✓	✓	✓
		injection	✓		✓		✓
		steroid	✓		✓		✓
		joint			✓		✓
32	M70	use	✓			✓	
		disorder	✓		✓	✓	✓
		cause			✓		✓
		pressure	✓			✓	
		soft				✓	
		overuse	✓	✓	✓	✓	✓

ลำดับ	รหัสไอซีดี เทนซีเอ็ม	คำสำคัญ	แพทย์ คนที่ 1	แพทย์ คนที่ 2	แพทย์ คนที่ 3	แพทย์ คนที่ 4	แพทย์ คนที่ 5
		tissue				✓	
		arm					
		disease				✓	
		shoulder					
33	M76	foot	✓				
		limb				✓	
		enthesopathy	✓	✓	✓	✓	✓
		exclude					
		right					
		disorder					
		diagnosis					
		leg	✓		✓	✓	✓
		hip	✓		✓		✓
		unspecified			✓	✓	✓
34	M80	fracture	✓		✓	✓	✓
		cause				✓	
		pathological	✓	✓	✓	✓	✓
		current				✓	
		osteoporosis	✓	✓	✓	✓	✓
		bone	✓		✓	✓	✓
		condition				✓	
		patient					
		disease					
		vertebra	✓		✓		✓
35	M83	osteomalacia	✓	✓	✓	✓	✓
		bone	✓	✓	✓	✓	✓
		vitamin	✓	✓	✓	✓	✓
		clinic					
		cause			✓	✓	✓
		calcium	✓	✓	✓	✓	✓
		low	✓		✓	✓	✓
		level	✓			✓	

ลำดับ	รหัสไอซีดีเทนซีเอ็ม	คำสำคัญ	แพทย์คนที่ 1	แพทย์คนที่ 2	แพทย์คนที่ 3	แพทย์คนที่ 4	แพทย์คนที่ 5
		deficiency	✓		✓	✓	✓
		symptom				✓	
36	M91	cause				✓	
		disease			✓	✓	✓
		hip	✓	✓	✓	✓	✓
		condition					
		group					
		child	✓	✓	✓	✓	✓
		joint	✓			✓	
		osteochondrosis	✓	✓	✓	✓	✓
		juvenile	✓	✓	✓	✓	✓
		occur				✓	
37	M95	tissue	✓	✓	✓	✓	✓
		disease			✓	✓	✓
		connective	✓	✓	✓	✓	✓
		disorder				✓	
		syndrome			✓		✓
		collagen	✓		✓		✓
		deformity			✓	✓	✓
		system	✓		✓		✓
		joint	✓		✓	✓	✓
		cause			✓	✓	✓
38	M97	fracture	✓	✓	✓	✓	✓
		periprosthetic	✓	✓	✓	✓	✓
		joint	✓			✓	
		prosthetic	✓	✓		✓	
		around				✓	
		internal				✓	
		encounter					
		bone	✓		✓	✓	✓
		hip	✓		✓	✓	✓
		prosthesis	✓		✓	✓	✓

ลำดับ	รหัสไอซีดี เทนซีเอ็ม	คำสำคัญ	แพทย์ คนที่ 1	แพทย์ คนที่ 2	แพทย์ คนที่ 3	แพทย์ คนที่ 4	แพทย์ คนที่ 5
39	S30	injury	✓		✓	✓	✓
		use					
		low			✓		✓
		cause				✓	
		back	✓		✓	✓	✓
		genital	✓	✓	✓	✓	✓
		superficial	✓		✓	✓	✓
		external	✓			✓	
		abdomen	✓	✓	✓		✓
		pelvis	✓	✓	✓	✓	✓
40	S33	joint	✓	✓	✓	✓	✓
		injury	✓		✓	✓	✓
		sprain	✓	✓	✓	✓	✓
		ligament	✓	✓	✓	✓	✓
		lumbar	✓		✓	✓	✓
		spine	✓		✓	✓	✓
		dislocation	✓		✓	✓	✓
		part					
		pelvis	✓		✓		✓
		rupture					
41	S43	include					
		joint	✓	✓	✓	✓	✓
		use					
		ligament	✓	✓	✓	✓	✓
		sprain	✓		✓	✓	✓
		dislocation	✓	✓	✓	✓	✓
		shoulder	✓	✓	✓	✓	✓
		injury	✓		✓	✓	✓
		cause				✓	
		girdle				✓	
42	S53	joint	✓		✓	✓	✓
		injury	✓		✓	✓	✓

ลำดับ	รหัสไอซีดี เทนซีเอ็ม	คำสำคัญ	แพทย์ คนที่ 1	แพทย์ คนที่ 2	แพทย์ คนที่ 3	แพทย์ คนที่ 4	แพทย์ คนที่ 5
		sprain	✓	✓	✓	✓	✓
		cause				✓	
		ligament	✓	✓	✓	✓	✓
		dislocation	✓	✓	✓	✓	✓
		muscle					
		tendon	✓			✓	
		elbow	✓	✓	✓	✓	✓
		strain	✓				
43	S54	injury	✓		✓	✓	✓
		nerve	✓	✓	✓	✓	✓
		cause					
		level				✓	
		forearm	✓	✓	✓	✓	✓
		muscle			✓		✓
		wrist					
		repair			✓		✓
		condition				✓	
		tendon					
44	S55	injury	✓		✓	✓	✓
		vessel	✓	✓	✓	✓	✓
		blood	✓		✓	✓	✓
		level					
		forearm	✓	✓	✓	✓	✓
		cause					
		arm					
		diagnosis				✓	
		external					
		hand					
45	S56	injury	✓		✓	✓	✓
		muscle	✓	✓	✓	✓	✓
		tendon		✓	✓	✓	✓
		forearm	✓		✓	✓	✓

ลำดับ	รหัสไอซีดี เทนซีเอ็ม	คำสำคัญ	แพทย์ คนที่ 1	แพทย์ คนที่ 2	แพทย์ คนที่ 3	แพทย์ คนที่ 4	แพทย์ คนที่ 5
		fascia			✓	✓	✓
		level				✓	
		cause					
		laceration	✓		✓		✓
		external					
		flexor	✓		✓		✓
46	S58	amputation	✓	✓	✓	✓	✓
		use					
		elbow	✓	✓	✓	✓	✓
		forearm	✓	✓	✓	✓	✓
		traumatic	✓	✓			
		injury	✓		✓	✓	✓
		hand	✓				
		limb	✓		✓	✓	✓
		patient					
		wrist					
47	S61	finger	✓	✓	✓	✓	✓
		use					
		hand	✓	✓	✓	✓	✓
		wound	✓	✓	✓	✓	✓
		open	✓	✓	✓		✓
		wrist			✓	✓	✓
		injury	✓		✓	✓	✓
		condition					
		cause				✓	
		external					
48	S63	ligament	✓	✓	✓	✓	✓
		joint	✓	✓	✓	✓	✓
		sprain	✓	✓	✓	✓	✓
		dislocation	✓	✓	✓	✓	✓
		finger	✓	✓		✓	
		injury	✓		✓	✓	✓

ลำดับ	รหัสไอซีดี เทนซีเอ็ม	คำสำคัญ	แพทย์ คนที่ 1	แพทย์ คนที่ 2	แพทย์ คนที่ 3	แพทย์ คนที่ 4	แพทย์ คนที่ 5
		cause					
		wrist	✓		✓	✓	✓
		hand	✓	✓	✓	✓	✓
		level					
49	S66	injury	✓	✓	✓	✓	✓
		tendon	✓	✓	✓	✓	✓
		muscle		✓	✓	✓	✓
		hand	✓	✓	✓	✓	✓
		fascia	✓	✓	✓	✓	✓
		level					
		wrist	✓		✓	✓	✓
		use					
		finger	✓		✓	✓	✓
		cause				✓	
50	S91	injury	✓		✓	✓	✓
		toe	✓	✓	✓	✓	✓
		use					
		wound	✓	✓	✓	✓	✓
		ankle	✓	✓	✓	✓	✓
		open		✓	✓	✓	✓
		cause					
		foot	✓	✓	✓	✓	✓
		external					
		encounter					

ภาคผนวก ฉ
ตัวอย่างผลการทำนายชื่อโรคที่ไม่พบในเวชระเบียน

ลำดับ	รหัสไอซีดี เทนซีเอ็ม	คำสำคัญ (ข้อมูลเข้า)	ผลลัพธ์ที่ได้จากแบบจำลอง (7 อันดับแรก)
1	M01	arthritis, infection, joint, parasite, inflammation	M13, M00, T84, M02, M01, M14, M06
2	M02	arthritis, reactive, arthropathy	M13, M17, M02, M00, T84, M07, M11
3	M04	autoinflammatory	M17, S72, M48, M43, S52, M04, T84
4	M07	arthritis, enteropathic, joint	M13, M00, M07, M02, M14, M06, M08
5	M14	arthropathy, joint, arthritis	M13, M00, M06, M14, M08, M02, T84
6	M15	osteoarthritis, joint, arthritis, polyarthritis	M17, M19, M13, M16, M15, M18, M06
7	M30	polyarteritis, nodosa, artery	M30, S72, M17, M48, I74, M43, M47
8	M32	erythematosus, systemic	M32, M34, M17, T84, M35, M36, S82
9	M33	dermatomyositis, muscle, myopathy, myositis	M60, M62, M33, M63, M61, M48, M54
10	M34	scleroderma, systemic, sclerosis	M34, M17, M32, G35, T84, M36, L90
11	M42	spine, osteochondrosis	M42, M92, M91, M48, M51, M41, Q76
12	M63	muscle	M17, S82, M48, S86, S56, M62, M63
13	M83	osteomalacia, bone	M17, M48, M83, S82, M85, M88, M89
14	M95	tissue, connective	C49, T84, M36, M95, M17, D17, Q69
15	M97	fracture, periprosthetic, bone, hip	S72, M97, M84, S32, S92, S82, T84

ภาคผนวก ช

ตัวอย่างโปรแกรมที่พัฒนาขึ้นมาใช้ในงานวิจัย

โปรแกรมสำหรับดึงข้อมูลจากเว็บไซต์

```
import re
from pomp.core.base import BaseCrawler
from pomp.contrib.urllibtools import UrllibHttpRequest
from pomp.core.engine import Pomp
from pomp.contrib.urllibtools import UrllibDownloader

python_sentence_re = re.compile('<p[^>]{0,}>[^}]{0,}?(?=<\/p>)|<ol[^>]{0,}>[^}]
{0,}?(?=<\/ol>)|<ul[^>]{0,}>[^}]{0,}?(?=<\/ul>)|<dl[^>]{0,}>[^}]{0,}?(?=<\/dl>)', re.I | re.M)
remove_sentence_re = re.compile('<[^>]{1,}>')

temp = []
class MyCrawler(BaseCrawler):
    def __init__(self, url):
        self.ENTRY_REQUESTS = UrllibHttpRequest(url)

    def extract_items(self, response):
        temp[:] = []
        for i in python_sentence_re.findall(response.body.decode('utf-8')):
            sentence = i.encode('utf-8').strip()
            sentence = remove_sentence_re.sub(" ", sentence)
            yield sentence

            s = "{}".format(sentence)
            if s != '' and s != ' '.join(s.split()) :
                for i in s.split('.') :
                    if i != '' :
                        i = re.sub('[^A-Za-z\s]+', " ", i)
                        i = ' '.join(i.split())
```

```
                    if i != '' and i != ' ' :
                        temp.append(i)

class CallMyCrawler() :
    def __init__(self, url) :
        pomp = Pomp(downloader=UrllibDownloader(),)
        pomp.pump(MyCrawler(url))
        self.temp = temp
```

โปรแกรมสำหรับประมวลผลข้อมูลก่อน

```
import nltk
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk import pos_tag
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords, wordnet
import warnings
from PyDictionary import PyDictionary
lemmatiser = WordNetLemmatizer()
stops = set(stopwords.words('english'))

def get_wordnet_pos(treebank_tag) :
    if treebank_tag.startswith("J") :
        return "a"
    elif treebank_tag.startswith("V") :
        return "v"
    elif treebank_tag.startswith("N") :
        return "n"
    elif treebank_tag.startswith("R") :
        return "r"
    else :
        return ""

def get_words_without_stop(sentence) :
    s = sentence.lower()
    tokens = word_tokenize(s)
    tokens_pos = dict(pos_tag(tokens))
    stopwords = set(tokens).intersection(stops)
    for sw in list(stopwords) :
        s = s.replace(' '+sw+' ', ' ')
    words = word_tokenize(s)
    words = list(words)
```

```
    return words, tokens_pos

def lemmatize_word(words, tokens_pos) :
    words_lem = []
    for i in words :
        pos_code = get_wordnet_pos(tokens_pos[i])
        if pos_code != "" :
            words_lem.append(lemmatiser.lemmatize(i, pos = pos_code))
        else :
            words_lem.append(lemmatiser.lemmatize(i))
    return words_lem

def synonym_word(words) :
    warnings.filterwarnings("ignore")
    dictionary=PyDictionary()
    for w in words.keys() :
        synonyms = dictionary.synonym(w)
        if synonyms != None :
            intersect = list(set(words)&set(synonyms))
            if len(intersect) > 0 and w in words :
                for i in intersect :
                    words[w] += words[i]
                    del words[i]
    return words
```

โปรแกรมสำหรับสร้างและทดสอบแบบจำลอง

```
import nltk
import ast
import nltk.classify
from nltk.classify.scikitlearn import SklearnClassifier
from sklearn.neural_network import MLPClassifier
import time
import openpyxl
import re
import operator

avg_time_predict = 0
avg_time_train = 0

alpha = [0, 0.1, 0.5, 1.0]
for a in alpha :
    fileXlsx = 'resultMLP.xlsx'
    wb = openpyxl.load_workbook(fileXlsx)

    for i in range(10) :
        filename = 'trainMLP'+str(i)+'.txt'

        with open(filename, 'r') as myfile:
            train = ast.literal_eval(myfile.read())

        classifier = SklearnClassifier(MLPClassifier(alpha=a))

        t0 = time.time()
        cl = classifier.train(train)
        t1 = time.time()
          avg_time_train = avg_time_train + (t1-t0)

          sheetName = 'Sheet'+str(i+1)
          ws = wb.get_sheet_by_name(sheetName)
```

```
        testName = 'testMLP'+str(i)+'.txt'
        with open(testName, 'r') as myfile1 :
            test = ast.literal_eval(myfile1.read())

        index = 2
        for t in test :
            t2 = time.time()
            predict = cl.classify(t[0])
            t3 = time.time()
            avg_time_predict = avg_time_predict + (t3-t2)

            pdist = cl.prob_classify(t[0])
            probs = {}
            for l in cl.labels() :
                probs.update({l:round(pdist.prob(l), 4)})

            sorted_prob = sorted(probs.items(), key=operator.itemgetter(1))
            result = sorted_prob[-10:]

            check = []
            for j in range(len(result)) :
                check.append(result[j][0])

            ws['A'+str(index)] = t[1]
            ws['B'+str(index)] = t[2]
            ws['C'+str(index)] = str(t[0])
            ws['D'+str(index)] = str(predict)
            ws['E'+str(index)] = str(result)

            wb.save(fileXlsx)
            index = index + 1

print('Avg Train'+str(i)+' --> '+str(avg_time_train/(len(test)*10)))
print('Avg Predict'+str(i)+' --> '+str(avg_time_predict/(len(test)*10)))
```

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวพรรณาภรณ์ เกตุภู่พงษ์
วัน เดือน ปี เกิด	2 กันยายน 2536
สถานที่เกิด	กรุงเทพฯ
วุฒิการศึกษา	สำเร็จการศึกษาระดับปริญญาตรี หลักสูตรวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
ที่อยู่ปัจจุบัน	46 ซอยโชติวัฒน์ 14 ถนนประชาชื่น บางซื่อ กรุงเทพฯ 10800