**บทที่ 3**

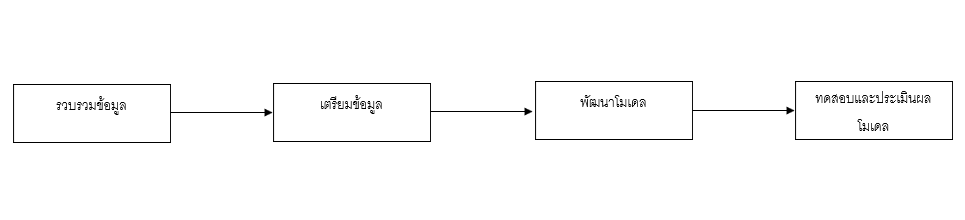
**วิธีดำเนินโครงการ**

ในการจัดทำเรื่อง การวิเคราะห์ความคิดเห็นของนักเรียน นักศึกษาวิทยาลัยอาชีวศึกษานครปฐมบนสื่อสังคมออนไลน์โดยใช้การประมวลผลภาษาธรรมชาติ ประกอบด้วยการทำงานหลัก 3 ส่วน ได้แก่ การพัฒนาโมเดลวิเคราะห์ข้อความ พัฒนา API และการพัฒนากระดานสรุปข้อมูล(Dashboard) บนเว็บแอปพลิเคชั่น ในบทนี้จะอภิปรายรายละเอียดการทำเนินงานตามผังการดำเนินงานดังนี้



**3.1** การพัฒนาโมเดลวิเคราะห์ข้อความ

ประกอบด้วยขั้นตอนการรวบรวมข้อมูล การเตรียมข้อมูล การพัฒนาโมเดล การทดสอบและประเมินโมเดล ซึ่งมีผังการดำเนินงานดังนี้



3.1.1 การรวบรวมข้อมูล

คณะผู้จัดทำได้ทำการรวบรวมข้อมูลจากนักเรียน นักศึกษาวิทยาลัยอาชีวศึกษานครปฐม จำนวน 200 คน มีจำนวนข้อมูล 800 ข้อความ สามารถแบ่งหัวข้อของข้อความได้ 4 ด้าน ได้แก่ ด้านการเรียนการสอน สิ่งอำนวยความสะดวก ทรัพยากรในการเรียนการสอน ตลอดจนสภาพแวดล้อมและความปลอดภัย จำนวนหัวข้อละ 200 ข้อความ ข้อความส่วนใหญ่ที่รวบรวมจะใช้ภาษาที่ไม่เป็นทางการ อาจมีโครงสร้างประโยคที่ไม่ถูกต้องตามไวยกรณ์ภาษาไทย ผ่านแบบฟอร์มกูเกิ้ลฟอร์ม(Google Forms) ดังรูปที่........

Graphical user interface, text, application, email

Description automatically generated

3.1.2 การเตรียมข้อมูล ในเบื้องต้นคณะผู้จัดทำได้คัดกรองข้อความที่มีเนื้อหาไม่เหมาะสม เป็นค่าว่าง ไม่สื่อความหมาย และไม่ตรงตามหัวข้อที่กำหนด นำข้อมูลที่เป็น .xls ดังรูปที่.......ให้อยู่ในรูปแบบไฟล์ .csv ดังรูปที่...........

Graphical user interface, text, application

Description automatically generated

Excel

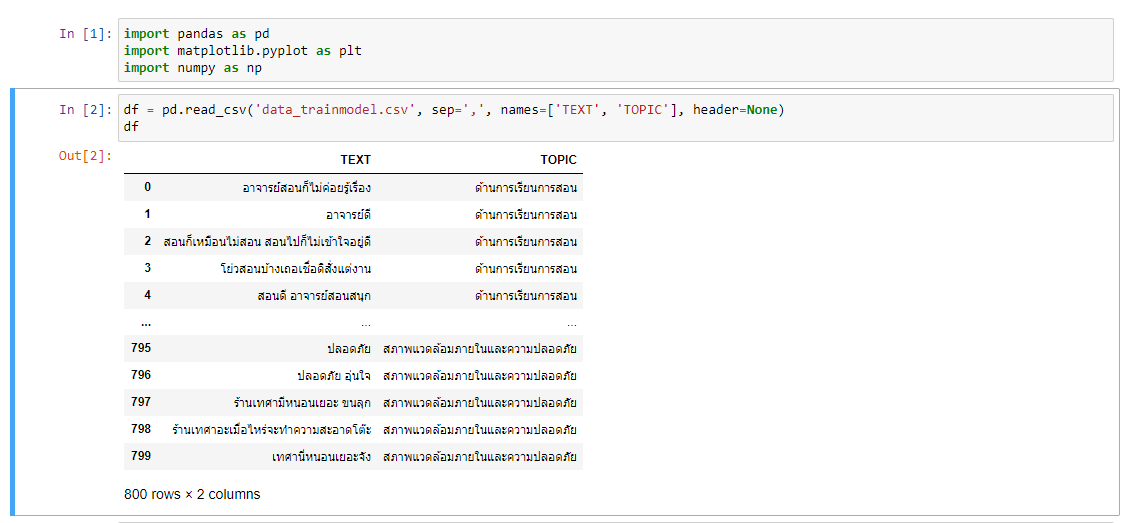
Graphical user interface, text, application

Description automatically generated

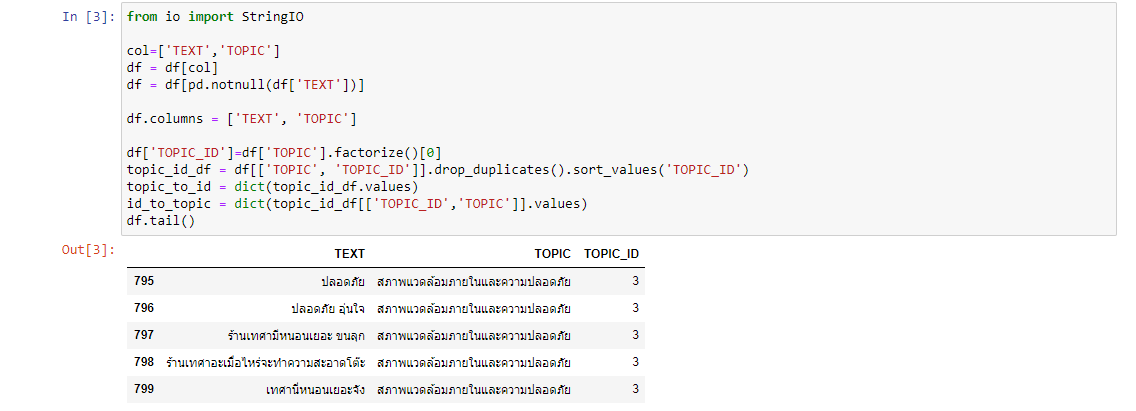
CSV

จากนั้นก็นำการประมวลภาษาธรรมชาติ (Natural Language Processing) เพื่อตัดคำที่ไม่สื่อความหมาย Stop word สัญลักษณ์ต่างๆ Word Vectorizer และBag-of-Words (BoW) ซึ่งคณะผู้จัดทำใช้ภาษา Python ในการพัฒนา โดยมีรายละเอียดดังต่อไปนี้

1. จากนั้นทำการอ่านข้อมูลจากไฟล์ดังกล่าวเก็บเข้าตัวแปร เมื่อแสดงออกมาดู จะประกอบด้วย 2 ฟิลด์ คือ TEXT และ TOPIC



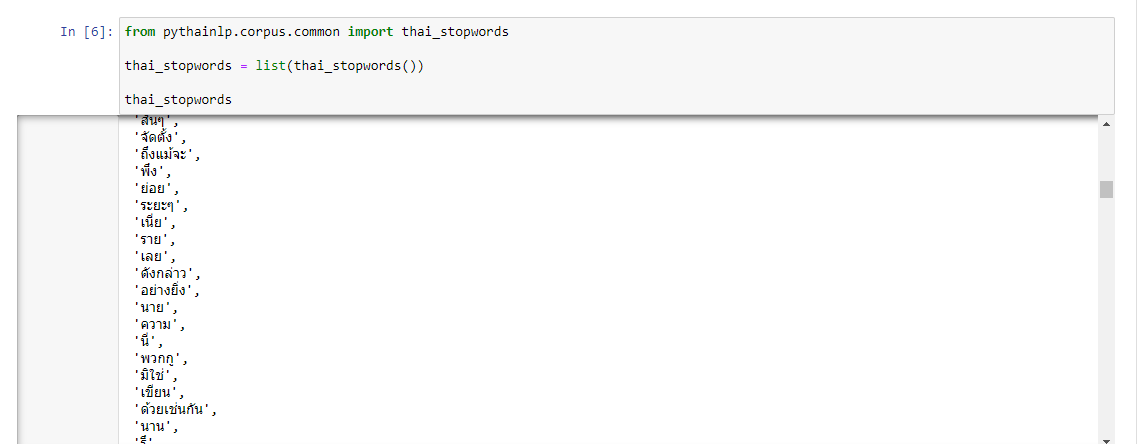
1. นำฟิลด์ TOPIC มาเข้ารหัสให้อยู่ในรูปแบบของตัวเลข ดังรูปที่.....

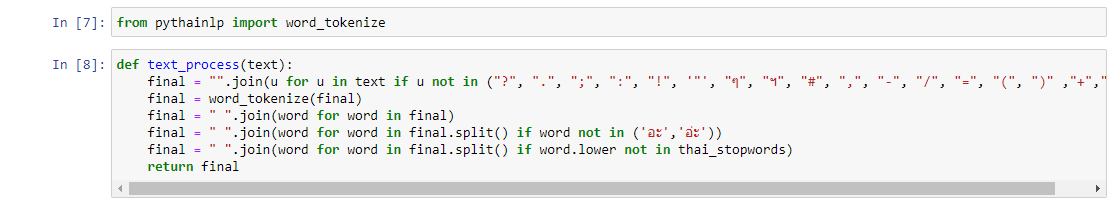


โดยฟิลด์ TOPIC\_ID 0 หมายถึง ด้านการเรียนการสอน 1 หมายถึง สิ่งอำนวยความ

สะดวก 2 หมายถึง ทรัพยากรในการเรียนการสอน 3 หมายถึง สภาพแวดล้อมและความปลอดภัย

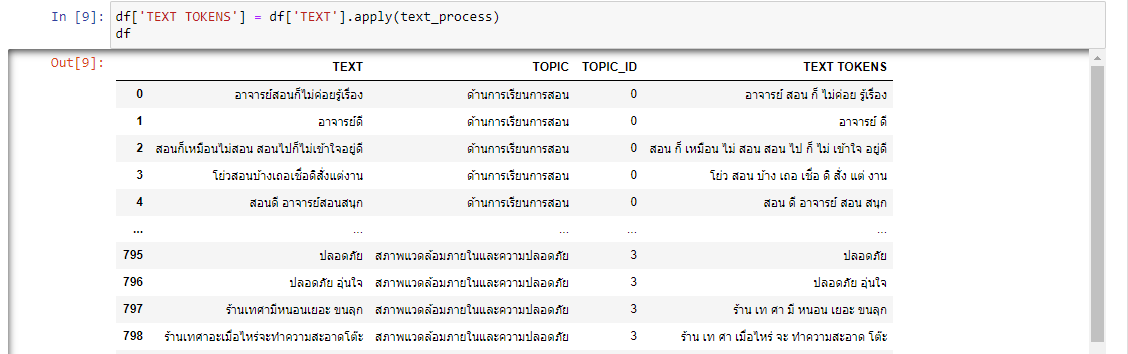
1. นำไลบารี่ thai\_stopwords มาใช้ในการตัดคำที่ไม่สื่อความหมาย เช่น ฉะนั้น เมื่อนั้น

และ จะ เป็นต้น จากนั้นสร้างฟังก์ชั่นในการตัดคำ และสัญลักษณ์ดังรูปที่....... และ.........

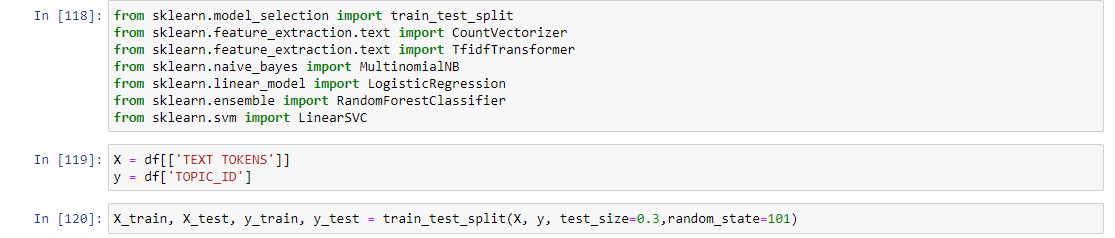
 รูปที่...........

รูปที่...........

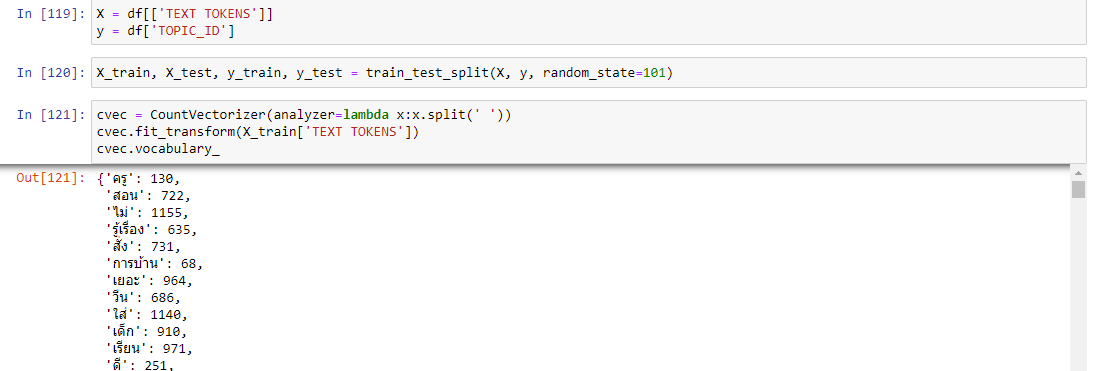
1. นำฟิลด์ TEXT มาตัดคำและสัญลักษณ์ และเก็บไว้ที่ฟิลด์ TEXT TOKENS ดังรูปที่…



1. แบ่งข้อมูลการเทรน และเทสเป็น 2 ส่วน คือ ส่วนของการเทรนโมเดลร้อยละ70 และส่วนของการทดสอบโมเดลร้อยละ 30 จากจำนวนข้อมูลจำนวน 800 ข้อความ และทำการสุ่มสลับข้อความ ดังรูปที่......

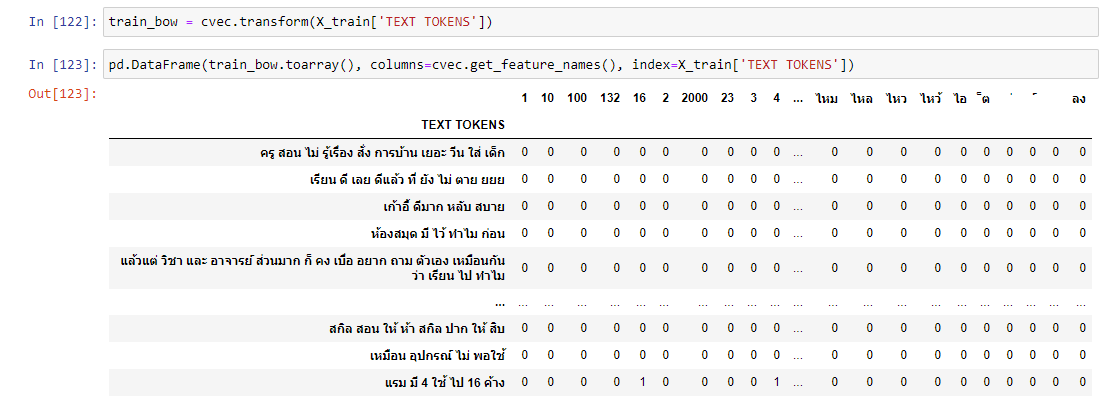


1. Word Vectorizer และ Bag-of-Words(BoW) โดยเริ่มต้นจะใช้ Count Vectorizer ของ sklearn มาช่วยในการจัดการดึงคำทั้งหมดออกมาจากข้อความ และจัดเก็บในรูปแบบ Vector ดังรูปที่.......



จากนั้นทำการสร้าง Bag-of-Words (BoW) ที่เปรียบเสมือนกับตารางที่มีแถวเป็นข้อความ

คอลัมน์เป็นคำทั้งหมด และค่าคือจำนวนคำที่ปรากฏในข้อความดังกล่าว ดังรูปที่.........



3.1.3 พัฒนาโมเดล คณะผู้จัดทำได้เลือก 4 อัลกอริธึมมาใช้ในการเทรนโมเดล ได้แก่ Random Forest, Linear SVM, Naive Bayes และ Logistic Regression มีรายละเอียดอัลกอริธึมดังนี้

1. Random Forest เป็นอัลกอรึธึมที่จะสร้างกฏขึ้นมาเองโดยการสุ่ม เช่น ถ้าเรามี y = {0,1} และ x เป็นตัวเลข 0-100 เขียนเป็นฟังก์ชั่นได้ดังนี้ y = f(x) กฎการทำนาย(Prediction) เราอาจจะเขียนได้แบบนี้

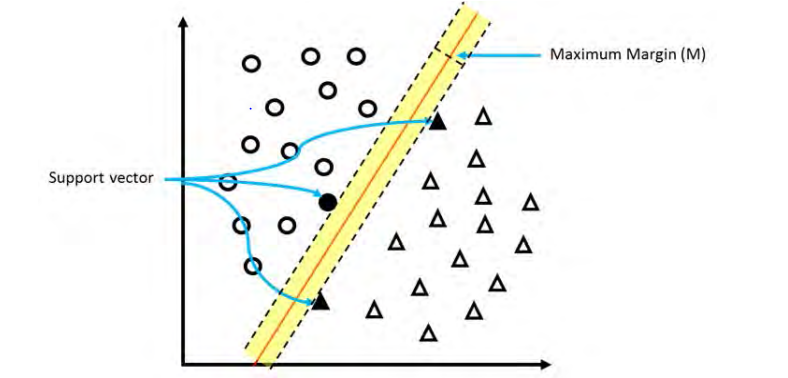
IF คะแนนสอบ x >= 80, THEN สอบผ่าน (y=1)

IF คะแนนสอบ x < 80, THEN สอบตก (y=0)

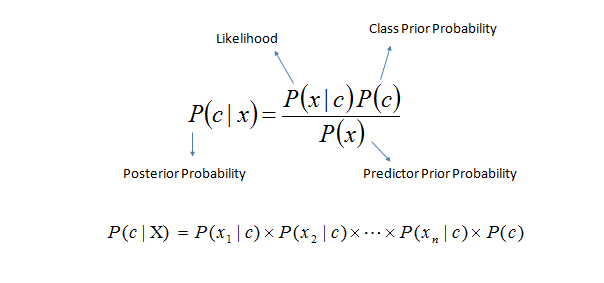
กฎดังกล่าวเกิดมาจากกระบวนการเรียนรู้ ของ Decision Tree ส่วน Random Forest คือ การสร้างโมเดลจาก Decision Tree หลายโมเดลย่อยๆ ตั้งแต่ 500 จนถึง หรือมากกว่า 1000 โมเดลเพื่อช่วยในการทำนายค่า y ตอนทำนายก็ให้แต่ละ Decision Tree ทำนายของใครของมัน และผลลัพธ์ที่ทำนายได้จะถูกเลือกโดยการโหวตของ Decision Tree ที่อยู่ภายใน Random Forest

2. Linear SVM เป็นเทคนิครูปแบบหนึ่งในการจำแนกข้อมูลบนปริภูมิคุณลักษณะ(Features Space) เป็นสองกลุ่มด้วยระนาบ โดยระนาบนี้จะมีขนาดคู่ขนาดอยู่ทั้งสองข้าง โดยที่ระนาบคู่ชนาดนี้จะสัมผัสกับข้อมูลที่อยู่ใกล้ระนาบหลักที่สุด(Support Vector) และมีระยะห่าง(Margin) จากระนาบหลักมากที่สุด(Maximum Margin)

ตัวอย่างเช่นถ้าเรามีข้อมูลที่เราทำการเทรน แล้วแบ่งเป็น 2 กลุ่มโดยเราจะเลือกใช้เส้น Linear Model ที่ไม่จำรูปแบบของข้อมูลที่ฝึกมากเกินไป(Over Fitting) โดยที่เส้นโมเดลที่เราเลือกจะมีระยะห่างระหว่าง 2 คลาสห่างกันมากที่สุด ดังรูปที่



3. Naive Bayes เป็นอัลกอริธึทในการจำแนกหหมวดหมู่โดยใช้หลักความน่าจเป็น ที่สามารถคาดการณ์ผลลัพธ์และสามารถอธิบายได้โดยทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็น สมการดังรูปที่...



อธิบายสมการ แทนตัวแปร 3 ตัว คือ

c คือ Class

x คือ Attribute

P คือ Probability (ความน่าจะเป็น)

P(c|x) Posterior probability คือ ความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น x จะมีคลาส C P(x|c) Likelihood คือ ความน่าจะเป็นที่ข้อมูลที่มีคลาส C และมีแอตทริบิวต์ x

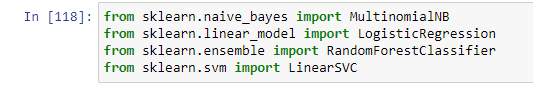
P(c) Prior probability คือ จำนวน Classที่อาจจะเกิดขึ้น/จำนวนClass ทั้งหมด หรือความน่าจะเป็นของ Class C

P(x) Predictor Prior probability คือ จำนวน Attribute ทั้งหมด

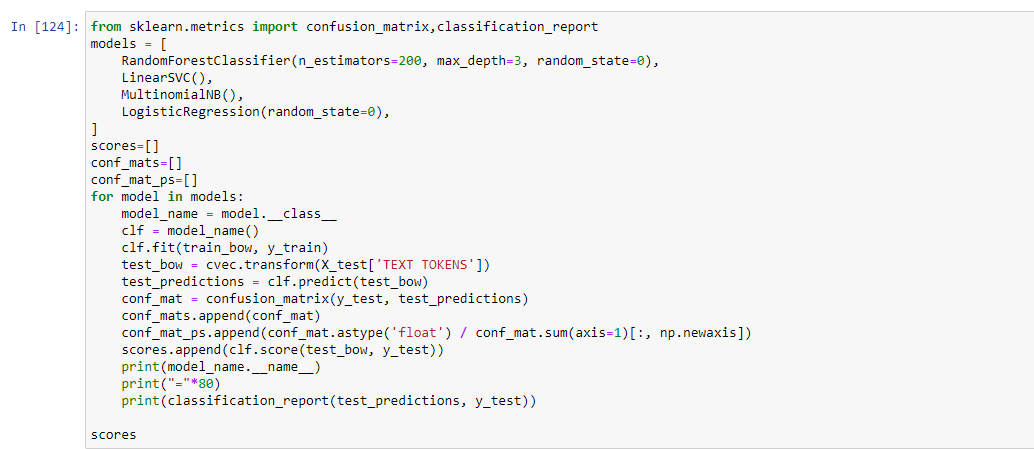
4. Logistic Regression การวิเคราะห์การถดถอยเป็นเทคนิคการสร้างแบบจําลองเชิง

คาดการณ์ชนิดหนึ่งซึ่งใช้ในการหาความสัมพันธ์ระหว่างตัวแปรที่ขึ้นต่อกัน (โดยปกติเรียกว่าตัวแปร "Y" ) และตัวแปรอิสระหนึ่งตัว (ตัวแปร "X" ) หรือชุดของตัวแปรอิสระ เมื่อใช้ตัวแปรอิสระสองตัวขึ้นไปเพื่อทํานายหรืออธิบายผลลัพธ์ของตัวแปรที่ขึ้นต่อกัน เป็นอัลกอริทึมการจําแนกประเภท มันถูกใช้เพื่อทํานายผลลัพธ์แบบไบนารีตามชุดของตัวแปรอิสระ เช่น "ใช่" หรือ "ไม่", "ผ่าน" หรือ "ล้มเหลว" เป็นต้น แต่ก็มี Multinomial logistic regression ที่สามารถใช้ได้เมื่อมีตัวแปรที่ขึ้นกับประเภทหนึ่งตัวแปรที่มีระดับที่ไม่ได้เรียงลําดับตั้งแต่สองระดับขึ้นไป เช่น รถไฟ รถบัส รถราง และจักรยาน เป็นต้น

คณะผู้จัดทำได้เลือกใช้ไลบารี่ scikit-learn ซึ่งเป็นไลบรารีในภาษา Python สำหรับการพัฒนาโปรแกรมโดยใช้การเรียนรู้ของเครื่อง จุดเด่นคือฟังก์ชันในการแบ่งประเภทข้อมูล การแบ่งกลุ่มข้อมูล การวิเคราะห์การถดถอย ลฯ และมีเครื่องมือที่ใช้ในการเทรนโมเดลด้วยอัลกอริทึมดังกล่าว คือ RandomForestClassifier LinearSVC MultinomialNB และ LogisticRegression ดังรูปที่ .....

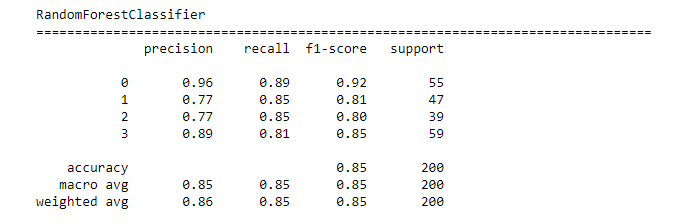


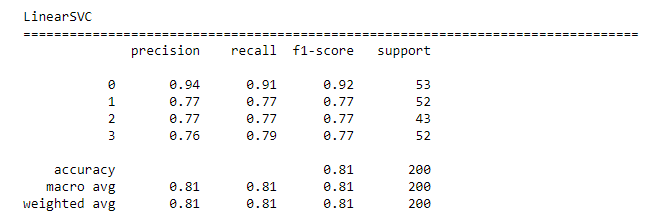
จากนั้นทำการทดสอบโมเดลโดยมีชุดข้อมูลในการฝึกเดียวกัน เพื่อหาอัลกอริทึมที่มีประสิทธิภาพ และมีความแม่นยำที่สุด ดังรูปที่ ……

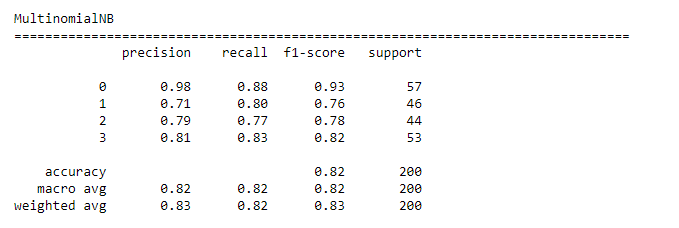


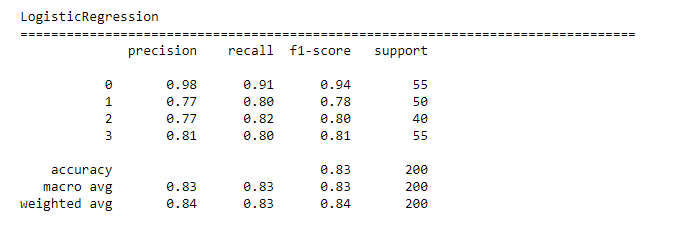
**3.1.4 ทดสอบและประเมินผลโมเดล**

จากการพัฒนาโมเดลในหัวข้อก่อนหน้านี้ คณะผู้จัดทำได้เก็บค่าผลการทดสอบโมเดลในรูปแบบ Evulate Model ของทั้ง 4 อัลกอริธึม ผลปรากฎว่า RandomForestClassifier มีความแม่นยำอยู่ที่ 85% ดังรูปที่..... LinearSVC มีความแม่นยำอยู่ที่ 81% ดังรูปที่..... MultinomialNB มีความแม่นยำอยู่ที่ 82% ดังรูปที่..... และ LogisticRegression มีความแม่นยำอยู่ที่ 85% ดังรูปที่..... ตามลำดับ

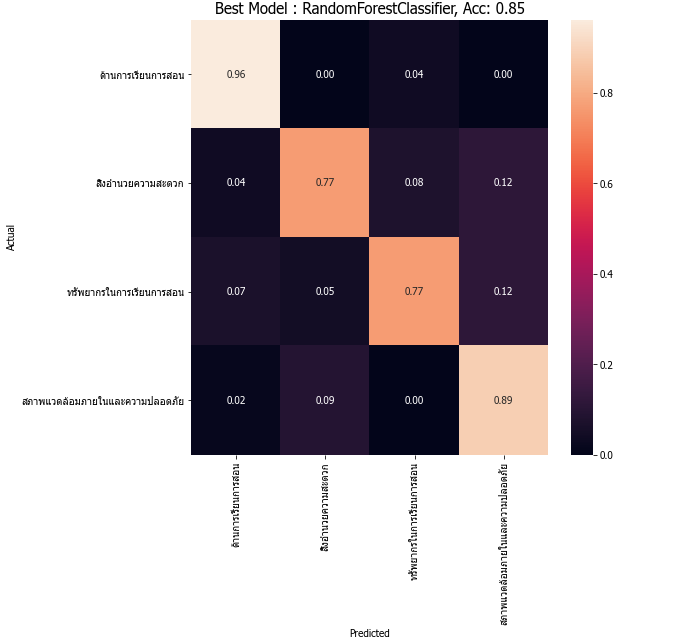








จากผลการทดสอบดังกล่าวผู้จัดทำจึงเลือกใช้อัลกอริธึม RandomForestClassifier มาใช้ในการเทรนโมเดล และแสดงในรูปแบบของ Heat Map ดังรูปทึ่.........



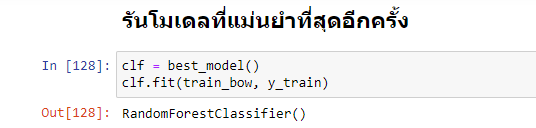
**ด้านการเรียนการสอน** โมเดลได้ทำนายผลข้อความที่ Label ว่าเป็นด้านการเรียนการสอนได้ถูกต้อง 96% โดยที่ทำนายผิดพลาดเป็นทรัพยากรในการเรียนการสอน 4%

**สิ่งอำนวยความสะดวก** โมเดลได้ทำนายผลข้อความที่ Label ว่าเป็นสิ่งอำนวยความสะดวกได้ถูกต้อง 77% โดยที่ทำนายผิดพลาดเป็นด้านการเรียนการสอน 4% ทรัพยากรในการเรียนการสอน 5% และสภาพแวดล้อและความปลอดภัย 12%

**ทรัพยากรในการเรียนการสอน** โมเดลได้ทำนายผลข้อความที่ Label ว่าเป็นทรัพยากรในการเรียนการสอนได้ถูกต้อง 77% โดยที่ทำนายผิดพลาดเป็นการเรียนการสอน 7% สิ่งอำนวยความสะดวก 8% และสภาพแวดล้อและความปลอดภัย 12%

**สภาพแวดล้อและความปลอดภัย** โมเดลได้ทำนายผลข้อความที่ Label ว่าเป็นสภาพแวดล้อมและความปลอดภัยได้ถูกต้อง 89% โดยที่ทำนายผิดพลาดเป็นการเรียนการสอน 2% สิ่งอำนวยความสะดวก 9%

จากนั้นก็พัฒนาโมเดลด้วยอัลกอริทึม RandomForestClassifier ด้วยชุดข้อมูลที่ได้เตรียมไว้ในข้างต้น



(ต่อ) API

3.1.1 ทวิตเตอร์ (Twitter)

3.1.2 การเรียนรู้ของเครื่องจักร ( Machine Learning )

3.1.3 การประมวลภาษาธรรมชาติ (Natural Language Processing )

3.1.4 การจำแนกประเภทข้อความ ( Text Classification )

3.1.5 เว็บแอพพลิเคชั่น ( Web Application )

3.1.6 ทฤษฎีแรงจูงใจ ( Motivation Theory )

3.1.7 งานวิจัยที่เกี่ยวข้อง

**3.2 ประชากรและกลุ่มตัวอย่าง**

3.2.1 ประชากรที่ใช้ในการจัดทำโครงการ ได้แก่ นักเรียนนักศึกษาวิทยาลัยอาชีวศึกษานครปฐม

3.2.2 กลุ่มตัวอย่าง ได้แก่ นักเรียนนักศึกษาสาขาเทคโนโลยีสารสนเทศ จำนวน 95 คน ข้อความที่นำมาแยกประมาณ 2,000 ข้อความ (Training Set)

**3.3 การสร้างเครื่องมือและหาคุณภาพเครื่องมือที่ใช้ในการดำเนินโครงการ**

3.3.1 เครื่องมือที่ใช้ในการดำเนินโครงการ

3.3.1.1 Visual Studio Code

3.3.1.2 Django

3.3.1.3 Bootstrap

3.3.1.4 CSS

3.3.1.5 HTML

3.3.1.6 MySQL

**3.4 การวิเคราะห์ข้อมูล**

การวิเคราะห์ข้อมูลที่รวบรวมได้จากแบบสอบถาม จำนวน 95 ชุด นำมาวิเคราะห์ด้วยวิธีทางสถิติและประมวลผลโดยแยกวิเคราะห์ดังนี้

3.4.1 แบบสอบถามข้อมูลเกี่ยวกับข้อมูลทั่วไปของกลุ่มตัวอย่าง ผู้ศึกษานำข้อมูลมาแจกแจงความถี่ และหาค่าร้อยละ

3.4.2 แบบสอบถามระบบวิเคราะห์ข้อมูลบนทวิตเตอร์ด้วยวิธีการวิเคราะห์ความรู้สึก (Sentiment Analysis) ซึ่งแบ่งออกเป็น 5 ระดับดังนี้

ระดับ 1 หมายถึง ความพึงพอใจอยู่ในระดับน้อยที่สุด

ระดับ 2 หมายถึง ความพึงพอใจอยู่ในระดับน้อย

ระดับ 3 หมายถึง ความพึงพอใจอยู่ในระดับปานกลาง

ระดับ 4 หมายถึง ความพึงพอใจอยู่ในระดับมาก

ระดับ 5 หมายถึง ความพึงพอใจอยู่ในระดับมากที่สุด

นำคะแนนของแบบสอบถามมารวมกันแล้วหาค่าเฉลี่ย เกณฑ์การแปรผลค่าเฉลี่ยเป็นดังนี้

ค่าเฉลี่ย 4.21 – 5.00 หมายถึง\*ความพึงพอใจอยู่ในเกณฑ์ระดับ มากที่สุด

ค่าเฉลี่ย 3.21 – 4.20 หมายถึง\*ความพึงพอใจอยู่ในเกณฑ์ระดับ มาก

ค่าเฉลี่ย 2.61 – 3.40 หมายถึง\*ความพึงพอใจอยู่ในเกณฑ์ระดับ ปานกลาง

ค่าเฉลี่ย 1.81 – 2.60 หมายถึง\*ความพึงพอใจอยู่ในเกณฑ์ระดับ น้อย

ค่าเฉลี่ย 1.00 – 1.80 หมายถึง\*ความพึงพอใจอยู่ในเกณฑ์ระดับ น้อยที่สุด

**3.5 สถิติที่ใช้ในการวิเคราะห์ข้อมูล**

3.5.1 ค่าสถิติร้อยละ (Percentage) เพื่อใช้แปลความหมายข้อมูลปัจจัยส่วนบุคคลของผู้ตอบแบบสอบถามดังสมการดังต่อไปนี้

กำหนดให้ p แทน ค่าร้อยละ

X แทน ข้อมูลที่ต้องการหาค่าร้อยละ

N แทน จำนวนข้อมูลทั้งหมด

3.5.2 ค่าเฉลี่ยเลขาคณิตเป็นการหาค่ากลางหรือที่เรียกว่า การวัดแนวโน้มเข้าสู่ส่วนกลางของข้อมูลคำนวณได้โดยหาผลรวมของข้อมูลทุกค่า แล้วหารด้วยจำนวนข้อมูลทั้งหมดดังในสมการ โดยใช้สมการดังต่อไปนี้

𝑥̅ =

กำหนดให้ 𝑥̅ แทน ค่าเฉลี่ยเลขาคณิต

แทน ข้อมูลหน่วยที่ ของตัวอย่างขนาด

แทน จำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

3.5.3 ส่วนเบี่ยงเบนมาตรฐาน เป็นวิธีวัดการกระจาย หากเป็นส่วนเบี่ยงเบนมาตรฐานประชากร โดยใช้สมการดังนี้

กำหนดให้ แทน ค่าส่วนเบี่ยงเบนมาตรฐาน

แทน ข้อมูลหน่วยที่ ของตัวอย่างขนาด

𝑥̅ แทน ค่าเฉลี่ยเลขคณิต

แทน จำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง