การวัดระดับเนื้อหาเว็บไซต์การท่องเที่ยว ตามความต้องการของนักท่องเที่ยว โดยการวิเคราะห์ข้อความในเว็บไซต์

ภูรดา จันทรัฐ

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
บัณฑิตวิทยาลัย สาขาเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
ปีการศึกษา 2561

หัวข้อสารนิพนธ์	การวัดระดับเนื้อหาเว็บไซต์การท่องเที่ยว ตามความต้องการของนักท่องเที่ยว โดยการวิเคราะห์ข้อความในเว็บไซต์
โดย	ภูรดา จันทรัฐ
สาขาวิชา	เทคโนโลยีสารสนเทศ
อาจารย์ที่ปรึกษาสารนิพนธ์	ดร. โสภณ มงคลลักษมี

บัณฑิตวิทยาลัย สถาบันเทคโนโลยีไทย-ญี่ปุ่น อนุมัติให้นับสารนิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีบัณฑิตวิทยาลัย

(รองศาสตราจารย์ ดร. พิชิต สุขเจริญพงษ์)

วันที่.....เดือน.....พ.ศ.....

คณะกรรมการสอบสารนิพนธ์

.....ประธานกรรมการ

(ดร. ภาสกร อภิรักษ์วรพินิต)

.....กรรมการ

(รองศาสตราจารย์ ดร. อรรณพ หมั่นสกุล)

.....อาจารย์ที่ปรึกษาสารนิพนธ์

(ดร. โสภณ มงคลลักษมี)

ภูรดา จันทรัฐ : การวัดระดับเนื้อหาเว็บไซต์การท่องเที่ยว ตามความต้องการของนักท่องเที่ยว โดยการวิเคราะห์ข้อความในเว็บไซต์. อาจารย์ที่ปรึกษา : ดร.โสภณ มงคลลักษมี, 76 หน้า.

การวัดระดับเนื้อหาเว็บไซต์การท่องเที่ยว ตามความต้องการของนักท่องเที่ยว โดยการวิเคราะห์ข้อความในเว็บไซต์ มีวัตถุประสงค์ในการศึกษา ได้แก่ 1) สร้างเฟรมเวิร์คเพื่อตรวจสอบคุณภาพเนื้อหาของเว็บไซต์เกี่ยวกับการท่องเที่ยว 2) เพื่อเป็นแนวทางการปรับปรุงเนื้อหาเว็บไซต์การท่องเที่ยวให้ครบถ้วนตามความต้องการของนักท่องเที่ยว โดยสามารถแบ่งขั้นตอนในการศึกษาออกเป็น 7 ขั้นตอนได้แก่ 1) ศึกษา ค้นคว้าข้อมูลและทฤษฎีการทำ machine learning และ text mining 2) วิเคราะห์ความต้องการของนักท่องเที่ยว 3) เก็บข้อมูลจากเว็บไซต์การท่องเที่ยวในประเทศไทย เพื่อนำมาใช้เป็น dataset 4) สร้างเฟรมเวิร์คของระบบการตรวจสอบเนื้อหาเว็บไซต์ โดยใช้เทคนิค machine learning และ text mining 5) ออกแบบการทดลองโดยการแบ่ง dataset เพื่อนำไปใช้ในการ training และ testing 6) นำข้อมูลมาทดลองระบบการตรวจสอบเนื้อหาเว็บไซต์ ด้วยการทดสอบแบบ single topic และ multiple topics 7) ทดสอบประสิทธิภาพความแม่นยำของระบบและ classifier สามารถสรุปผลการศึกษาได้ดังนี้

เฟรมเวิร์คของระบบที่ถูกสร้างขึ้นในครั้งนี้ สามารถตรวจสอบความครบถ้วนของเนื้อหาในเว็บไซต์ ด้วยการใช้ data mining technique โดย Latent Dirichlet Allocation (LDA) technique ถูกนำมาใช้ในการหาคำ keywords ของเนื้อหา และ TF-IDF technique ถูกนำมาใช้ในการสร้าง feature vector Random Forest Algorithm ถูกใช้เป็นแกนหลักในการตรวจสอบความครบถ้วนของเนื้อหา โดยการสร้าง classifier จำนวน 2 ตัว ประกอบด้วย binary-class classifier และ multi-class classifier สำหรับการทดลองแบบ single topic และ multiple topics ผลการทดลองแบบ single topic ได้ค่า Accuracy อยู่ที่ 0.96 สำหรับการผลการทดลองแบบ multiple topics ยังถูกแบ่งออกเป็น 2 แบบย่อย คือ มี 2 topics ในเนื้อหา และ มี 3 topics ในเนื้อหา โดยผลการทดลองในแบบ 2 topics ได้ค่า Recall อยู่ที่ 0.79 และแบบ 3 topics ได้ค่า Recall อยู่ที่ 0.78 จากผลการทดลองที่ได้แสดงให้เห็นว่าเฟรมเวิร์คนี้สามารถแยกแยะและจดจำความครบถ้วนของเนื้อหาบนเว็บไซต์ได้เป็นอย่างดี

บัณฑิตวิทยาลัย ลายมือชื่อนักศึกษา .....
สาขาวิชา เทคโนโลยีสารสนเทศ ลายมือชื่ออาจารย์ที่ปรึกษา .....
ปีการศึกษา 2561

PURADA JANTARATH : MEASURING THE CONTENT COMPLETENESS OF TOURISM WEBSITES USING TEXT MINING TECHNIQUES. ADVISOR : DR.SOPHON MONGKOLLUKSAMEE, 76 PP.

In this study, we develop a technique for measuring the completeness of content for tourism website by using text mining techniques. The two main objectives of this study. First, to create a framework for measuring the completeness of content for the website. Second, to provide a guideline for improving the website content to meet tourists' need. There are six steps in this study. The first step is to study and research on the theory of machine learning and text mining. The second step is to collect and analyze the information that tourists needs. Collecting content of tourism website in Thailand and Preparing the dataset is the third step. The fourth step is developing the framework for checking and measuring the website content. Designing the experiments, including single topic testing and multiple topics testing, and preparing the training and testing dataset is in the fifth step. Conducting experiments and measuring the performance of the proposed technique are in the sixth and seventh step, respectively.

In the framework, we use the Latent Dirichlet Allocation (LDA) technique to find the keywords of contents then TF-IDF value of those keywords to create the feature vectors. After that, we create the classifier using the Random Forest Algorithm. We train two classifiers, which are binary-class classifier and multi-class classifier, for single topic experiment and multiple topics experiment. In the single topic experiment, the accuracy score is 0.96. There two sub-experiments in the multiple topics experiment. They are two topics and three topics experiments. The recall score of the two and three topics experiment are 0.79 and 0.78. The experimental result confirms that the proposed framework works well in classifying and recognize the completeness of the website content.

Graduate School Student's Signature.....

Field of Study Information Technology Advisor's Signature.....

Academic Year 2018

กิตติกรรมประกาศ

การทำสารนิพนธ์ในครั้งนี้สำเร็จลงได้นั้น ผู้ศึกษาขอกราบขอบพระคุณในความกรุณาของ ดร. โสภณ มงคลลักษมี ที่ได้เสียสละเวลาในการให้คำปรึกษา และคำแนะนำ ตลอดระยะเวลาในการดำเนินการศึกษา อีกทั้งยังขอกราบขอบพระคุณ ดร. บุษราพร เหลืองมาลาวัฒน์ ในการชี้แนะและให้คำปรึกษาในการเริ่มต้นการทำสารนิพนธ์

ทั้งนี้ ผู้ศึกษาขอกราบ ขอบพระคุณ คณะกรรมการสอบทุกท่าน ได้แก่ ดร. ภาสกร อภิรักษ์วรพินิต และ รองศาสตราจารย์ ดร.อรรณพ หมั่นสกุล ที่กรุณาตรวจสอบแก้ไขข้อบกพร่องของสารนิพนธ์ฉบับนี้ด้วยความเอาใจใส่ รวมถึงเจ้าหน้าที่บัณฑิตวิทยาลัย และเจ้าหน้าที่คณะเทคโนโลยีสารสนเทศทุกท่านที่ให้ความสะดวกด้านอำนวยการ และประสานงานในการทำสารนิพนธ์ให้ผู้เขียนตลอดมาตลอดจนค้นคว้าหาข้อมูลในการจัดทำสารนิพนธ์ของผู้เขียนครั้งนี้สำเร็จลุล่วงไปด้วยดี

เหนือสิ่งอื่นใด ผู้ศึกษาของกราบขอบพระคุณ บิดา มารดา ของผู้ศึกษาที่เป็นกำลังใจและให้การสนับสนุนในทุก ๆ ด้านเสมอมา และเพื่อน ๆ พี่ ๆ ทุกคนในคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีไทย-ญี่ปุ่น ที่ให้การสนับสนุนและความช่วยเหลือที่ผ่านมา

ภูรดา จันทรัฐ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ภาวะเศรษฐกิจท่องเที่ยวในปี พ.ศ. 2560 ถูกคาดการณ์โดยกระทรวงการท่องเที่ยวและกีฬา ไว้ว่าสถานการณ์การท่องเที่ยวโลกในปี พ.ศ. 2560 จะเติบโตร้อยละ 3-4 ซึ่งใกล้เคียงกับการเติบโตในปี พ.ศ. 2559 สำหรับประเทศไทยเริ่มต้นด้วยจำนวนนักท่องเที่ยวต่างชาติจำนวน 9.07 ล้านคน สร้างรายได้ ได้ถึง 4.72 แสนล้านบาท แสดงให้เห็นว่าเศรษฐกิจการท่องเที่ยว ณ ปัจจุบันนี้ได้เติบโตขึ้น และมีความสำคัญมาก อีกทั้งองค์กรการท่องเที่ยวโลก (UNWTO) ยังประกาศให้ปี พ.ศ. 2560 เป็นปีแห่งการท่องเที่ยวอย่างยั่งยืนอีกด้วย [1]

เมื่อวันที่ 25 กันยายน พ.ศ. 2561 Mastercard [2] ได้เผยผลสำรวจภายใต้โครงการ Mastercard Global Destinations Cities Index เกี่ยวกับเมืองที่เป็นจุดหมายปลายทางของนักท่องเที่ยวมากที่สุดในโลก โดยมีการจัดอันดับทั้งหมด 162 ประเทศ ซึ่งตัวชี้วัดที่ใช้จัดอันดับในครั้งนี้ มาจากจำนวนนักท่องเที่ยวที่ค้างคืน และมีการใช้จ่ายผ่านบัตรเครดิต Mastercard โดยที่กรุงเทพมหานครได้อันดับ 1 ของโลก โดยมีนักท่องเที่ยวมาเยือนสูงกว่า 20 ล้านคน เป็นการครองอันดับ 1 ครั้งที่ 5 ในรอบ 6 ปี นับตั้งแต่ปีพ.ศ. 2555 และยังคงครอบอันดับ 1 ต่อเนื่องมาเป็นปีที่ 3 โดยนักท่องเที่ยวที่มาเยือนไทยมากที่สุด มาจากประเทศจีน ญี่ปุ่น และ เกาหลีใต้ อีกทั้งยังมีเมืองภูเก็ตที่อยู่ในอันดับ 12 และเมืองพัทยาที่อยู่ในอันดับ 18 ซึ่งจากผลสำรวจของ Mastercard ทำให้ทราบว่านักท่องเที่ยวเข้ามาท่องเที่ยวในประเทศไทยเป็นจำนวนมาก อย่างที่ทราบทั่วกันว่าประเทศไทยค่อนข้างมีชื่อเสียงด้านธรรมชาติและวัฒนธรรมที่เป็นเอกลักษณ์ ประกอบกับประเทศไทย เป็นประเทศที่มีค่าครองชีพที่ไม่สูงมาก เมื่อเทียบกับประเทศอื่น ๆ เช่น ฝรั่งเศส สิงคโปร์ ญี่ปุ่น หรืออีกมากมาย

ในโลกปัจจุบันต้องยอมรับว่าเทคโนโลยีหรืออิเล็กทรอนิกส์ เข้ามามีบทบาทในชีวิตมนุษย์มากขึ้นในทุก ๆ ด้าน ไม่ว่าจะเป็นด้านการศึกษา เศรษฐกิจ การแพทย์ และอีกมากมาย รวมทั้งการท่องเที่ยวด้วยเช่นกัน สำหรับการค้นหาสถานที่ท่องเที่ยวหรือแหล่งท่องเที่ยวต่าง ๆ ทั่วโลก ตัวเลือกอันดับแรกของนักท่องเที่ยวในการเข้าถึงหรือค้นหาสถานที่ท่องเที่ยวหรือแหล่งท่องเที่ยวได้ง่ายและทั่วถึง คือ การเลือกใช้เว็บไซต์หรือแอพพลิเคชั่นในการค้นหา ซึ่งในประเทศไทย การท่องเที่ยวแห่งประเทศไทย ได้ให้ความสำคัญกับการให้ข้อมูลแก่นักท่องเที่ยวเป็นอย่างมาก โดยการท่องเที่ยวแห่งประเทศไทย ได้จัดทำเว็บไซต์เพื่ออำนวยความสะดวกแก่นักท่องเที่ยวทั้งไทยและต่างชาติ หากแต่บางครั้งการให้ข้อมูลของเว็บไซต์การท่องเที่ยวแห่งประเทศไทย หรือจากเว็บไซต์การท่องเที่ยวอื่น ๆ

ยังไม่ตรงไปตามความต้องการของนักท่องเที่ยวหรือบุคคลทั่วไปที่ต้องการท่องเที่ยวมากนัก ซึ่งสามารถพบได้มากในหลายเว็บไซต์การท่องเที่ยวที่ยังมีการจัดนำเที่ยว หรือนำเสนอการท่องเที่ยวที่ยังไม่ตรงกับความต้องการของนักท่องเที่ยว

จากความเป็นมาของปัญหาดังที่กล่าวมาข้างต้นทั้งหมดนั้นผู้ศึกษาจึงมีความคิดในการสร้างเฟรมเวิร์คเพื่อตรวจสอบเนื้อหาเว็บไซต์การท่องเที่ยว ว่าเนื้อหาจากเว็บไซต์การท่องเที่ยวแต่ละเว็บไซต์การท่องเที่ยวนั้นได้แสดงข้อมูล หรือการจัดนำเที่ยว ตรงกับตามความต้องการของนักท่องเที่ยวหรือบุคคลทั่วไปครบถ้วนหรือไม่ โดยความต้องการที่นักท่องเที่ยวหรือบุคคลทั่วไปต้องการ ได้มาจากการที่ผู้ศึกษาได้ทำการสำรวจ เพื่อที่ต้องการทราบและนำความต้องการของนักท่องเที่ยวหรือบุคคลทั่วไป ไปใช้ประกอบกับการสร้าง classification ในการจดจำและแยกแยะเนื้อหาบนเว็บไซต์การท่องเที่ยว โดยที่เลือกใช้ Random forest ในการแยกแยะที่มีประสิทธิภาพที่ดีที่สุด ทั้งนี้หากผู้ที่สนใจจัดทำเว็บไซต์การท่องเที่ยว หรือผู้ที่มีเว็บไซต์การท่องเที่ยวอยู่แล้ว นำระบบหรือเฟรมเวิร์คตัวนี้ไปใช้ ก็จะสามารถช่วยให้เว็บไซต์มีการให้ข้อมูลที่ตรงกับความต้องการของนักท่องเที่ยวหรือบุคคลทั่วไป และยังช่วยให้ดึงดูดให้มีคนมาท่องเที่ยวได้มากขึ้นอีกด้วย

1.2 วัตถุประสงค์การศึกษา

1. สร้างเฟรมเวิร์คเพื่อตรวจสอบคุณภาพเนื้อหาเกี่ยวกับเว็บไซต์การท่องเที่ยว
2. เพื่อเป็นแนวทางการปรับปรุงเนื้อหาเว็บไซต์การท่องเที่ยวให้ครบถ้วนตามความต้องการของนักท่องเที่ยว

1.3 ขอบเขตของการศึกษา

1. สร้างเฟรมเวิร์คตรวจสอบเว็บไซต์การท่องเที่ยวของประเทศไทย
2. ตรวจสอบโดยใช้เทคนิค machine learning และ text mining ในการสร้างเฟรมเวิร์คที่ใช้กับภาษาไทย

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้เฟรมเวิร์คสำหรับตรวจสอบเนื้อหาเว็บไซต์การท่องเที่ยวของประเทศไทย เพื่อสื่อสารแก่นักท่องเที่ยวโดยที่เนื้อหาครบถ้วนตามความต้องการของนักท่องเที่ยว
2. ผู้จัดนำเที่ยว หรือผู้จัดทำเว็บไซต์ ได้เข้าใจและสนใจในนำเสนอข้อมูลตามความต้องการของนักท่องเที่ยว
3. คนไทยและผู้ที่สนใจให้ความสำคัญกับการนำเสนอข้อมูล และให้ความสนใจแก่นักท่องเที่ยวมากยิ่งขึ้น

บทที่ 2
ทฤษฎี เทคโนโลยี และงานวิจัยที่เกี่ยวข้อง

ในการศึกษาการวัดระดับเนื้อหาเว็บไซต์การท่องเที่ยว ตามความต้องการของนักท่องเที่ยว โดยการวิเคราะห์ข้อความในเว็บไซต์ ผู้ศึกษาจึงได้สืบค้นข้อมูล ทฤษฎี และงานวิจัย เพื่อเป็นแนวทางในการศึกษา โดยมีหัวข้อดังต่อไปนี้

2.1 ข้อมูลการท่องเที่ยว

2.2 Python Programming Language

2.3 เหมืองข้อความ

- 2.3.1 Tokenization: deepcut, pythainlp
- 2.3.2 Stop words removal
- 2.3.3 Document Topic Modeling: LDA
- 2.3.4 TF-IDF
- 2.3.5 Classification: SVM, Random Forest
- 2.3.6 Confusion Matrix

2.4 งานวิจัยที่เกี่ยวข้อง

2.1 ข้อมูลการท่องเที่ยว

องค์กรการท่องเที่ยวโลก (World Tourism Organization) ให้ความหมายคำว่าการท่องเที่ยวว่า เป็นการเดินทางใด ๆ ก็ตามที่เป็นการเดินทางตามเงื่อนไข 3 ประการ คือ

1. การเดินทางจากอยู่อาศัยไปยังสถานที่อื่นชั่วคราว แต่ไม่ใช้เป็นแหล่งที่พักถาวร
2. การเดินทางเป็นไปด้วยความเต็มใจ หรือความพึงพอใจของผู้เดินทาง
3. การเดินทางด้วยเหตุผลใด ๆ ก็ตามที่ไม่ใช่เพื่อหารายได้หรือประกอบอาชีพ

วรรณา วงษ์วานิช กล่าวว่า การท่องเที่ยว เป็นการที่ได้เดินทางไปยังสถานที่ต่าง ๆ และในการเดินทางได้มีกิจกรรมต่าง ๆ ขึ้น เช่น การเที่ยวชมสถานที่ การเดินทางเพื่อซื้อสินค้า หรือกิจกรรมทางธรรมชาติต่าง ๆ โดยแหล่งท่องเที่ยว [3] นั้นหมายถึง สิ่งหรือสถานที่ที่เกิดจากธรรมชาติและที่มนุษย์สร้างขึ้น ทั้งที่เป็นรูปธรรมและนามธรรมที่สามารถดึงดูดใจในด้านความสวยงาม ความน่าประทับใจ มีสิ่งอำนวยคงามสะดวกทั้งด้านการบริหาร ปัจจัยพื้นฐานในเรื่องระบบการสื่อสาร สาธารณโภค การขนส่ง และการต้นรับนักท่องเที่ยวอย่างมีมิตรไมตรี

จากผลการสำรวจของ Mastercard Global Destinations Cities Index แสดงให้เห็นว่าความต้องการเดินทางมายังประเทศไทยของนักท่องเที่ยวทีปริมาณมาก ซึ่งสิ่งสำคัญในการเดินทางไปยังแหล่งท่องเที่ยว คือ ข้อมูลเกี่ยวกับสถานที่ท่องเที่ยวนั้น ๆ ผู้ศึกษาจึงได้จัดทำการสำรวจเกี่ยวกับข้อมูลที่นักท่องเที่ยวต้องการมากที่สุด จำนวนทั้งหมด 126 คน ที่อยู่ในช่วงอายุ 18-60 ปี ซึ่งได้ผลสรุปจากการศึกษาครั้งนี้ว่า ข้อมูลสำคัญที่นักท่องเที่ยวต้องการเกี่ยวกับสถานที่ท่องเที่ยวประกอบด้วย

1. การเดินทาง
2. แหล่งท่องเที่ยว
3. สถานที่พัก
4. ราคา

ในงานวิจัยในนี้หัวข้อราคาจะไม่ได้นำมาใช้ในการทดสอบกับระบบ เนื่องจากข้อมูลราคามีปริมาณน้อยและไม่ชัดเจน อีกทั้งเว็บไซต์ส่วนใหญ่นำเสนอข้อมูลเกี่ยวกับราคาในแบบรูปภาพ จึงทำให้มีฐานข้อมูลที่เป็นรูปแบบข้อความน้อยมาก จึงไม่สามารถนำมาใช้เป็น dataset ในงานในการศึกษาครั้งนี้ได้

ในการศึกษาครั้งนี้ผู้ศึกษาจึงจัดเตรียม dataset เพื่อใช้ในการศึกษาเพียง 3 หัวข้อ ซึ่งแต่ละหัวข้อมีจำนวนคำและจำนวนเอกสารที่แตกต่างกันออกเป็นดังนี้

1. หัวข้อการเดินทาง
   - มีเอกสารทั้งหมด 60 เอกสาร
   - มีคำในเอกสารทั้งหมด 4,384 คำ
   - มีคำเฉลี่ยต่อเอกสาร 73.02 คำ
2. หัวข้อแหล่งท่องเที่ยว
   - มีเอกสารทั้งหมด 60 เอกสาร
   - มีคำในเอกสารทั้งหมด 2,065 คำ
   - มีคำเฉลี่ยต่อเอกสาร 34.47 คำ
3. หัวข้อสถานที่พัก
   - มีเอกสารทั้งหมด 60 เอกสาร
   - มีคำในเอกสารทั้งหมด 2,572 คำ
   - มีคำเฉลี่ยต่อเอกสาร 42.87 คำ

2.2 Python Programming Language

Python [4] เป็นภาษาในการเขียนโปรแกรมที่ได้รับความนิยามเป็นอย่างมากในปัจจุบัน เนื่องด้วยเป็นภาษาที่มีความยืดหยุ่นและเรียบง่าย ภาษา Python สนับสนุนรูปในการเขียนโปรแกรมที่หลากหลาย รวมทั้งการเขียนแบบโครงสร้างเชิงวัตถุ ภาษา Python เป็นภาษาที่ทำงานร่วมกับภาษาอื่นๆได้เป็นอย่างดี ทำให้เหมาะที่จะใช้เขียนเพื่อประสานงานโปรแกรมที่เขียนในภาษาต่างกันได้ และส่วนของโปรแกรมแบบโมดูลของภาษาอื่นๆก็สามารถนำมาใช้ร่วมกับภาษา Python ได้เป็นอย่างดี ตัวอย่างเช่น สามารถเขียนโปรแกรมในภาษา C++ และนำไปใช้กับ python ในรูปแบบโมดูล ซึ่งทำให้สามารถพัฒนาแอปพลิเคชั่นได้อย่างรวดเร็ว

โดยลักษณะเฉพาะของ python สามารถระบุได้ดังนี้:

- Python ทำงานได้รวดเร็วและมีประสิทธิภาพ

เนื่องด้วยเป็นภาษาที่ได้รับความนิยมและมีไลบรารีมากมาย ซึ่งเป็นเครื่องอำนวยความสะดวกสำหรับการเขียนโปรแกรม ตั้งแต่ขั้นพื้นฐานจนถึงขั้นฟังก์ชันสูง ทำให้สามารถทำงานที่ซับซ้อนได้ด้วภาษา Python เพียงไม่กี่บรรทัด เช่น การเขียนโค้ดเพียง 3 บรรทัด ก็สามารถสร้างเว็บเซิร์ฟเวอร์ได้

- Python สนับสนุนเทคโนโลยีอื่นๆ

ภาษา Python สามารถรองรับ COM, .Net ฯลฯ นอกจากนี้ทางเลือกบางอย่างและการเติมเต็มได้ถูกสร้างขึ้นสำหรับ Python ที่ทำให้ง่ายต่อการทำงานกับวัตถุเหล่านี้ในโหมดรวม

- Python เป็นแบบพกพา

ภาษา Python สามารถถูกเขียนไว้ในรูปแบบบสคริปต์ ซึ่งสามารถใช้ได้กับระบบปฏิบัติการอื่นเช่น Windows, Linux, UNIX, Amigo, Mac OS ฯลฯ นอกจากนี้บางรุ่นยังได้รับการปล่อยตัวออกมาเพื่อทำงานใน. Net, Java Virtual Machine (JVM) และ Nokia S60 เป็นที่น่าสนใจที่รู้ว่าผลลัพธ์ของโปรแกรมคล้ายกับนี้แพลตฟอร์มทั้งหมด

- Python ง่ายและน่ารัก

เป็นภาษาระดับสูงที่มีแหล่งเรียนรู้มากมาย นอกจากนี้ยังมีซอฟท์แวร์อื่นมากมาย ที่ช่วยให้ภาษานี้ใช้งานง่ายและสร้างแรงจูงใจให้กับผู้ใช้ต่อไป

- Python เป็น open source

แม้ว่าสิทธิทั้งหมดของโปรแกรมนี้สงวนไว้สำหรับสถาบัน Python แต่เป็น open source และไม่มีข้อจำกัดในการใช้เปลี่ยนแปลงและแจกจ่าย

รูปที่ 2.1 สัญลักษณ์โปรแกรม Python [4]

2.3 เหมืองข้อความ

เหมืองข้อความ [5] เป็นรูปแบบหนึ่งของการทำเหมืองข้อมูลและเป็นองค์ความรู้ในการประมวลผลภาษาธรรมชาติ Diane McDonald กล่าวถึงประโยชน์ในการทำเหมืองข้อความไว้ดังนี้ 1. ช่วยเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูล, 2.ปลดล้อกข้อมูลที่อยู่ในฐานข้อมูลขนาดใหญ่, 3.ช่วยปรับปรุงกระบวนการวิจัยให้ดียิ่งขึ้น, และ 4. ช่วยเพิ่มผลประโยชน์ให้กับเศรษฐกิจ

โดยการทำเหมืองข้อความจะศึกษาในด้านการค้นหาเอกสาร การเรียนรู้ระบบ สถิติ และการประมวลผลในด้านภาษา ซึ่งการทำเหมืองข้อความนี้จะทำงานด้านการจัดประเภทข้อความ (Text Categorization) การจัดกลุ่มข้อความ (Text Clustering) การสกัดข้อมูล (Extraction) การวิเคราะห์ทัศนคติ (Sentiment Analysis) การสรุปเอกสาร (Document Summarization) และการหาความสัมพันธ์ (Relation Modeling) โดยที่ใช้องค์ความรู้ด้านภาษา หลักไวยากรณ์ และหลักสถิติในการทำงาน

กระบวนการทำงานหลักๆของการทำเหมืองข้อความ มีทั้งหมด 3 ขั้นตอนหลักๆดังนี้

1. การเตรียมข้อมูล

ขั้นตอนนี้จะเป็นขั้นตอนในการเตรียมข้อมูลก่อนการใช้ระบบ ซึ่งมีหลากหลายประเภท เช่น การแปลงข้อมูล การกรองข้อมูล การหารากศัพท์ และการสร้างดัชนี

2. การใช้อัลกอริทึม

เป็นขั้นตอนการหากฎความสัมพันธ์ หรือค้นหาเอกสารที่ต้องการจากดัชนีของเอกสารด้วยเทคนิคต่าง ๆ

3. การนำเสนอผลลัพธ์

เป็นการสร้างกราฟิกหรือรูปภาพในการนำเสนอ เพื่อช่วยให้ผู้ใช้หรือผู้ศึกษามีความเข้าใจได้อย่างครบถ้วน

ในการศึกษาครั้งนี้มีเทคนิคที่เกี่ยวข้องที่ผู้ศึกษาได้เลือกใช้ดังต่อไปนี้

2.3.1 Tokenization

เป็นการตัดหรือแบ่งคำในเอกสารออกเป็นคำ ๆ สำหรับภาษาไทยการเขียนประโยคจะไม่มีช่องว่างระหว่างคำ จึงทำให้การตัดคำในภาษาไทยถือเป็นขั้นตอนที่ค่อนข้างยาก ไลบรารี่ในการทำการ tokenization มีหลากหลาย แต่ไลบรารี่ที่นิยมใช้มากในภาษาไทยมากที่สุด คือ deepcut และ pythainlp

- deepcut เป็นไลบรารี่ในการตัดคำโดยใช้เทคนิค Deep Neural Network โดยได้รับการฝึกในสถาบันที่ดีที่สุดของ NECTEC

ตัวอย่าง

ในทะเลชุมพร โดยรีสอร์ทหลายแห่งที่อยู่ริมชายหาดสามารถจัดกิจกรรมดำน้ำได้
= ใน|ทะเลชุมพร|โดย|รีสอร์ท|หลาย|แห่ง|ที่|อยู่|ริม|ชาย|หาด|สามารถ|จัด|กิจกรรม|ดำ|น้ำ|ได้

- pythainlp เป็นการประมวลผลภาษาไทยในภาษา Python เป็นโมดูลที่คล้ายกับ nltk แต่เบื้องต้นใช้ภาษาไทยแทนภาษาอังกฤษ

ตัวอย่าง

ในทะเลชุมพร โดยรีสอร์ทหลายแห่งที่อยู่ริมชายหาดสามารถจัดกิจกรรมดำน้ำได้
= ใน|ทะเล|ชุมพร|โดย|รีสอร์ท|หลาย|แห่ง|ที่อยู่|ริม|ชายหาด|สามารถ|จัด กิจกรรม|ดำน้ำ|ได้

2.3.2 Stop words removal

เป็นการทำการตัดคำที่ไม่สำคัญออก คำ stop words [6] เป็นคำที่ไม่มีความหมายต่อเนื้อหาในเอกสารนั้น ๆ อีกทั้งยังสามารถตัดทิ้งได้ และไม่ทำให้ประสิทธิภาพในการค้นหาข้อมูลแย่ลงอีกด้วย

ในภาษาไทยคำประเภทไม่สำคัญมีดังนี้

1. คำบุพบท

เป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน เช่น ของ ใน แก่ แด่ ต่อ ตั้งแต่ โดย เมื่อ กว่า กับ เป็น ดูก่อน ช้าแต่ ทาง สู่ แค่

2. คำสันธาน

เป็นคำที่ทำหน้าที่เชื่อมคำกับคำ ประโยคกับประโยคข้อความกับข้อความ ได้แก่ เพราะ เพราะว่า และ หรือ จึง ดังนั้น มิฉะนั้น ทั้ง แต่ แต่ว่า ครั้น หรือไม่ก็

3. คำสรรพนาม

เป็นคำที่ใช้แทนคำนามที่กล่าวถึงมาแล้วในประโยค เช่น ฉัน เรา เขา ดิฉัน กระผม กู คุณ ท่าน เธอ ใต้เท้า มัน สิ่ง ใคร บ้าง ต่างก็ ตัวนั้น อันโน้น อะไร ไหน ใด อย่างนี้

4. คำวิเศษณ์

เป็นคำที่ใช้ขยายคำอื่น เช่น มาก น้อย ใหญ่ เล็ก มหึมา มโหฬาร อ้วน โต สูง ไพเราะ เยอะ หลาย สวย หอม นุ่ม เผ็ดเช้า ค่ำ นี่ นั่น เอง ทั้งนี้ ค่ะ ครับ ขอรับ จ้า จ๊ะ วะ ไม่ หามิได้ บ่ ทำไม อย่างไร หมด อดีต ปัจจุบัน โบราณ หน้า

5. คำอุทาน

เป็นคำที่แสดงอารมณ์ อาการ หรือความรู้สึกของผู้พูด เช่น เอ๊ะ อ๊ะ อ๋อ เฮ้อ ว้าย โธ่ โถ อนิจจา สิ หนอย ซะ นะ น่า แหม ตายล่ะ คุณพระช่วย ชิชะ อุวะ ไม่น่าเลย โอ้โห

เนื่องจากขั้นตอนการทำ stop words removal จะเป็นขั้นตอนต่อจากการทำ tokenization จึงจะยกตัวอย่างจากตัวอย่างของ tokenization ด้านบนมาดังนี้

ในทะเลชุมพร โดยรีสอร์ทหลายแห่งที่อยู่ริมชายหาดสามารถจัดกิจกรรมดำน้ำได้

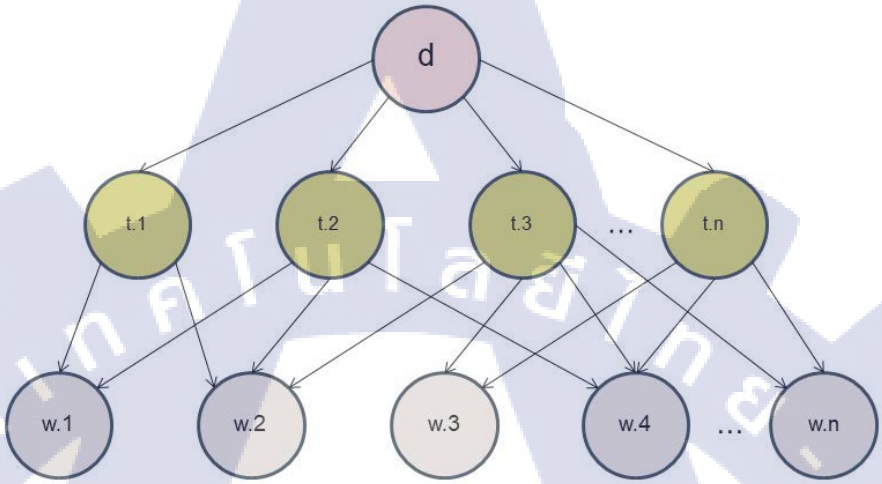
Tokenized = ใน|ทะเล|ชุมพร|โดย|รีสอร์ท|หลาย|แห่ง|ที่อยู่|ริม|ชายหาด|สามารถ|จัดกิจกรรม|ดำน้ำ|ได้

Removed stop word = ทะเล|รีสอร์ท|ที่อยู่|ริม|ชายหาด|สามารถ|จัดกิจกรรม|ดำน้ำ

2.3.3 Document Topic Modeling

เป็นการสร้างแบบจำลองการกระจายตัวของข้อมูล เพื่อนำมาจัดกลุ่มของข้อมูล โดยมีแนวคิดว่าในเอกสารเกิดจากการรวมตัวกันของหลายๆ หัวข้อ ซึ่งแต่ละหัวข้อมีคำที่น่าจะอยู่ในหมวดหมู่เดียวกัน นอกจากนี้ยังสามารถใช้เป็นรูปแบบของการทำเหมืองข้อความซึ่งเป็นวิธีที่จะทำให้ได้คำที่เป็นรูปแบบของเนื้อหาในต้นฉบับ ช่วยให้เรามีวิธีการจัดระเบียบทำความเข้าใจและสรุปรวบข้อมูลที่เป็นสาระสำคัญ เทคนิคที่ใช้สำหรับการหา document topic modeling ที่เป็นที่นิยมและใช้กันอย่างแพร่หลายก็คือ Latent Dirichlet Allocation (LDA)

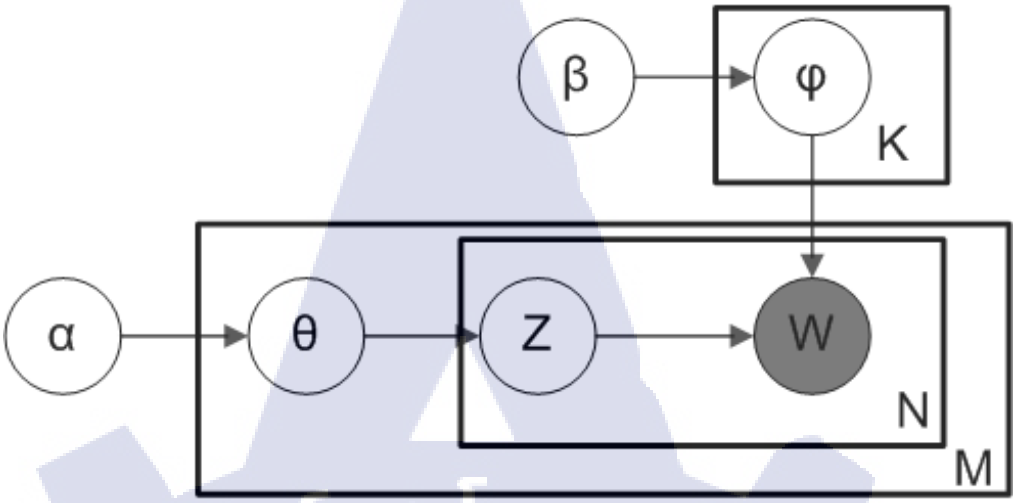
LDA เป็นแบบจำลองความน่าจะเป็น สำหรับข้อมูลที่ไม่ต่อเนื่อง เช่น คลังข้อมูล (Corpus) ซึ่งมีแนวคิดมาจากเอกสารสามารถมีการรวมกันของหัวข้อที่ซ่อนอยู่ในเอกสาร ซึ่งแต่ละหัวข้อมีการกระจายตัวของคำ โดย LDA ถูกนำไปใช้อย่างกว้างขวางในกระบวนการจัดกลุ่มของข้อมูล (Classification) และการกรองข้อมูล (Collaborative Filtering)



รูปที่ 2.2 แนวคิดของ Latent Dirichlet Allocation

จากรูปที่ 2.2 อธิบายได้ดังนี้ d หมายถึงเอกสาร โดยมีแนวคิดว่าในเอกสารสามารถมีหัวข้อได้หลายหัวข้อ t.1-t.n เป็นจำนวนหัวข้อ ซึ่งสามารถมีกี่หัวข้อก็ได้ และในแต่ละหัวข้อมีกลุ่มคำที่บ่งบอกถึงหัวข้อนั้น ๆ กระจายอยู่ โดย w.1-w.n เป็นจำนวนคำที่กระจายอยู่ในแต่ละหัวข้อ ซึ่งสามารถมีจำนวนกี่คำก็ได้เช่นกัน

LDA [7] ได้รับการนำมาใช้เป็นรูปแบบที่เป็นไปได้เชิงบวกสำหรับชุดเอกสารแนวคิดพื้นฐานที่อยู่เบื้องหลังวิธีนี้คือเอกสารจะถูกแสดงเป็นส่วนผสมแบบสุ่มตามหัวข้อแฝง แต่ละหัวข้อจะแสดงโดยการกระจายความน่าจะเป็นไปตามข้อกำหนด แต่ละบทความจะแสดงโดยการแจกแจงความน่าจะเป็นของหัวข้อ นอกจากนี้ยังมีการใช้ LDA เพื่อระบุหัวข้อในหลายๆ ด้าน เช่นการจัดหมวดหมู่ การกรองร่วมกัน และการกรองตามเนื้อหา ในกระบวนการสร้าง LDA คำที่ถูกสร้างขึ้นจากการรวมกันของหัวข้อ z ในเอกสาร d และการสุ่มตัวอย่างคำจากการกระจายคำหลัก โดยทั่วไปรูปแบบ LDA สามารถแสดงเป็นรูปแบบกราฟิกน่าจะเป็นดังแสดงในรูปที่ 2.3 มีการแสดง LDA ถึงสามระดับ ตัวแปร A และ B คือพารามิเตอร์ระดับคอร์ปัสซึ่งจะถือว่าเป็นตัวอย่างในระหว่างกระบวนการสร้างคลังข้อมูล



รูปที่ 2.3 model ของ Latent Dirichlet Allocation [7]

M หมายถึง จำนวนของเอกสาร
N หมายถึง จำนวนคำในเอกสาร
K หมายถึง จำนวนหัวข้อ
α หมายถึง hyperparameters สำหรับการกระจายหัวข้อในแต่ละเอกสาร
β หมายถึง hyperparameters สำหรับการกระจายคำต่อหัวข้อ
θ หมายถึง การกระจายหัวข้อในเอกสาร M
φ หมายถึง การกระจายคำในเอกสาร K
Z หมายถึง หัวข้อสำหรับคำในเอกสาร M
W หมายถึง คำเฉพาะ

2.3.4 TF-IDF

TF-IDF [8] ย่อมาจาก Term Frequency-Inverse Document Frequency น้ำหนัก TF-IDF เป็นน้ำหนักที่มักใช้ในการดึงข้อมูลและการทำเหมืองข้อความ น้ำหนักนี้เป็นตัวชี้วัดทางสถิติที่ใช้ในการประเมินความสำคัญของคำในเอกสารหรือคอลัมชัน ความสำคัญเพิ่มขึ้นตามสัดส่วนของจำนวนครั้งที่คำปรากฏในเอกสาร แต่ถูกชดเชยด้วยความถี่ของคำในคลัง รูปแบบของโครงการการถ่วงน้ำหนัก TF-IDF มักใช้โดยเครื่องมือค้นหาเป็นเครื่องมือหลักในการให้คะแนนและจัดอันดับความเกี่ยวข้องของเอกสารที่กำหนดให้กับข้อความค้นหาของผู้ใช้

โดยปกติน้ำหนัก TF-IDF ประกอบด้วยสองคำ: คำแรก คือ Term Frequency (TF) หรือที่เข้าใจกันว่าเป็นจำนวนครั้งของคำที่ปรากฏในเอกสาร โดยหารด้วยจำนวนคำทั้งหมดในเอกสาร คำที่สอง คือ Inverse Document Frequency (IDF) เป็นการคำนวณลอการิทึมของจำนวนเอกสารใน corpus หารด้วยจำนวนเอกสารที่คำที่เฉพาะเจาะจงปรากฏขึ้น

TF: Term Frequency ซึ่งวัดความถี่ของคำที่เกิดขึ้นในเอกสาร เนื่องจากเอกสารทุกฉบับมีความยาวแตกต่างกัน จึงเป็นไปได้ว่าคำๆหนึ่ง จะปรากฏขึ้นหลายครั้งในเอกสารที่มีความยาวมากกว่าเอกสารที่สั้นกว่า ดังนั้นระยะความถี่มักจะถูกหารด้วยความยาวของเอกสาร เป็นวิธีการธรรมดา:

TF(t) = (จำนวนครั้งของคำ t ที่ปรากฏในเอกสาร) / (จำนวนรวมของคำศัพท์ในเอกสาร)

IDF: Inverse Document Frequency ซึ่งวัดความสำคัญของคำ ในขณะที่การคำนวณข้อมูล TF ข้อกำหนดทั้งหมดถือว่ามีความสำคัญเท่าเทียมกัน อย่างไรก็ตามเป็นที่ทราบกันดีอยู่แล้วว่าคำบางคำเช่น "เป็น", "ของ" และ "ที่" อาจปรากฏขึ้นหลายครั้ง แต่ไม่มีความสำคัญ ดังนั้นเราจึงจำเป็นต้องชั่งน้ำหนักคำที่พบบ่อยลงในขณะที่ขยายขนาดที่หาได้ยากโดยการคำนวณต่อไปนี้:

IDF(t) = log_e (จำนวนของเอกสาร / จำนวนเอกสารที่มีคำว่า t อยู่ในนั้น)

<u>ตัวอย่าง</u>

ในเอกสารมีคำอยู่ทั้งหมด 100 คำ มีคำว่า “แมว” ปรากฏอยู่ 3 คำ

สูตร TF = (จำนวนครั้งของคำ t ที่ปรากฏในเอกสาร) / (จำนวนรวมของคำศัพท์ในเอกสาร)

TF = 3/100 = 0.03

ถ้าหากมีเอกสารอยู่ 10 ล้านเอกสารและมีคำว่า “แมว” ปรากฏแค่ 1,000 เอกสารสามารถ

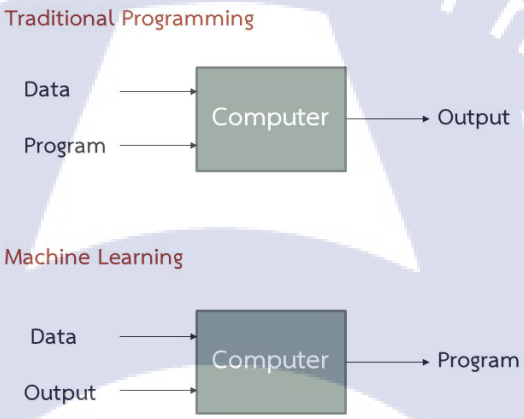
สูตร IDF = log_e (จำนวนของเอกสาร / จำนวนเอกสารที่มีคำว่า t อยู่ในนั้น)

IDF = log (10,000,000/1,000) = 4

ดังนั้นคือ TF-IDF จะคำนวณดังนี้ 0.03 * 4 = 0.12

2.3.5 Classification

Classification คือกระบวนการแยกแยะข้อมูลต่างๆ ซึ่งเป็นส่วนหนึ่งของ Machine learning ตัว Machine learning [9] นั้นคือการสอนคอมพิวเตอร์ให้ทำในสิ่งที่เป็นธรรมชาติต่อมนุษย์ โดยเรียนรู้จากประสบการณ์ อัลกอริทึมของ Machine learning ใช้วิธีคำนวณเพื่อเรียนรู้ข้อมูลโดยตรงจากข้อมูลที่ปราศจากสมการที่กำหนดไว้ในรูปแบบ อัลกอริทึมสามารถปรับตัวได้ดีขึ้นเท่ากันกับจำนวนตัวอย่างที่พร้อมสำหรับการเรียนรู้ที่เพิ่มขึ้น ซึ่งมันแตกต่างกับการเขียนโปรแกรมทั่วไป เพราะโปรแกรมจะใส่ข้อมูลและโปรแกรมเข้าไปเพื่อให้ได้ผลลัพธ์ แต่ Machine Learning จะไม่ได้โปรแกรมคำตอบ จะเป็นการใส่ข้อมูลและผลลัพธ์ เข้าไป เพื่อให้หาโปรแกรม ที่จะนำไปตอบในอนาคตได้ว่าใส่ข้อมูลไปแบบนี้ ผลลัพธ์จะเป็นอะไร

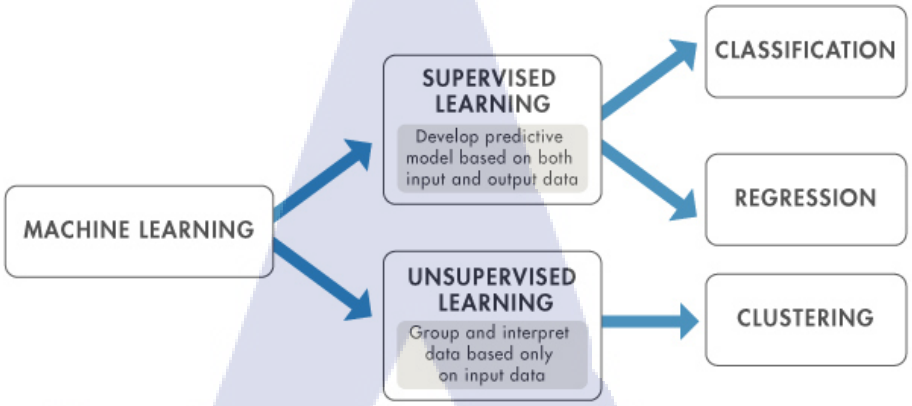


รูปที่ 2.4 กระบวนการของ Machine Learning

Machine learning ใช้เทคนิคอยู่ 2 ประเภท คือ Supervised learning และ Unsupervised learning

Supervised learning คือ การฝึกแบบจำลองฝึกแบบจำลองเกี่ยวกับข้อมูล input และ output ที่เป็นที่รู้จักเพื่อให้สามารถคาดเดาผลลัพธ์ในอนาคต

Unsupervised learning คือ การหารูปแบบที่ซ่อนอยู่ หรือโครงสร้างที่แท้จริงภายในข้อมูล input



รูปที่ 2.5 ประเภทของ Machine Learning [9]

จุดมุ่งหมายของ Supervised learning คือ การสร้างแบบจำลองที่ทำให้การคาดการณ์ขึ้นอยู่กับหลักฐานในที่ที่ไม่มีความแน่นอน อัลกอริทึมของ Supervised learning ใช้ชุดข้อมูลที่รู้จักและการตอบสนองที่รู้จักต่อข้อมูล (output) อีกทั้งยังฝึกแบบจำลอง เพื่อสร้างการคาดการณ์ที่เหมาะสมสำหรับการตอบสนองต่อข้อมูลใหม่ Supervised learning ใช้เทคนิค classification และ regression เพื่อพัฒนารูปแบบการคาดการณ์

Classification จะทำนายการตอบสนองอย่างเป็นรูปธรรม ยกตัวอย่างเช่น Email ที่ได้รับเป็นของแท้หรือสแปม หรือว่าเนื้องอกนั้นเป็นมะเร็งหรือเป็นพิษ Classification model จะจำแนกข้อมูลการป้อนข้อมูลเป็นหมวดหมู่การใช้งานทั่วไป ได้แก่ การถ่ายภาพทางการแพทย์ การรับรู้ภาพและเสียง หรือ การให้คะแนนบัตรเครดิต เป็นต้น

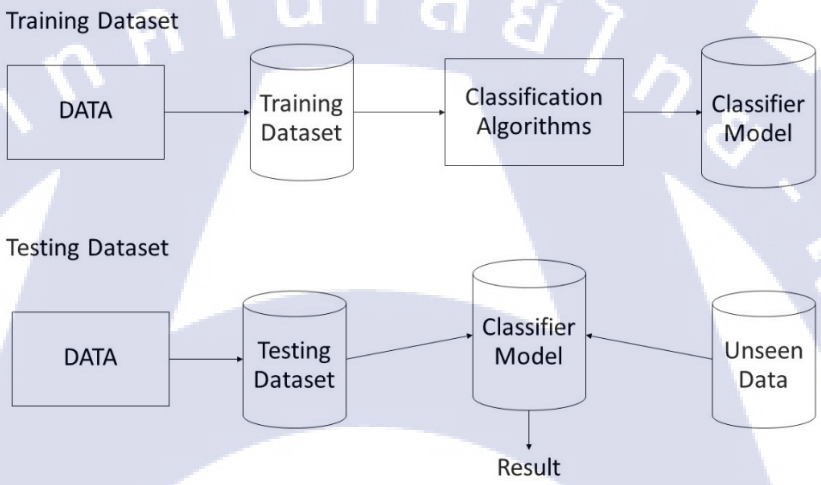
Regression จะทำนายการตอบสนองอย่างต่อเนื่อง เช่น การเปลี่ยนแปลงอุณหภูมิ หรือความผันผวนของความต้องการไฟฟ้า การใช้งานทั่วไป ได้แก่ การพยากรณ์ความต้องการไฟฟ้า และการซื้อขายอัลกอริทึม

Unsupervised learning จะค้นหารูปแบบที่ซ่อนอยู่หรือโครงสร้างที่แท้จริงภายในข้อมูล ใช้สำหรับการอนุมานจากชุดข้อมูลที่ประกอบด้วยข้อมูลที่ป้อนโดยที่ไม่มีคำตอบกำหนดไว้ โดยเทคนิคที่ใช้ใน Unsupervised learning เรียกว่า Clustering

Clustering เป็นเทคนิคที่ไม่ได้รับการเรียนรู้โดยทั่วไป ใช้สำหรับการวิเคราะห์ข้อมูลสำรวจเพื่อค้นหารูปแบบหรือการจัดกลุ่มข้อมูลที่ซ่อนอยู่ Application สำหรับ clustering รวมไปถึง การวิเคราะห์ลำดับยีน, การวิจัยการตลาด หรือการจดจำวัตถุ

การทำเหมืองข้อมูล (data mining) [10] เป็นเทคนิคที่ใช้ค้นหารูปแบบความสัมพันธ์ของข้อมูลจำนวนมากที่ซ่อนอยู่ในชุดข้อมูล โดยใช้หลักการทางสถิติและหลักคณิตศาสตร์แบบอัตโนมัติ ซึ่งจะสนใจเทคนิคของการจำแนกประเภทข้อมูลโดยเฉพาะ (Data Classification)

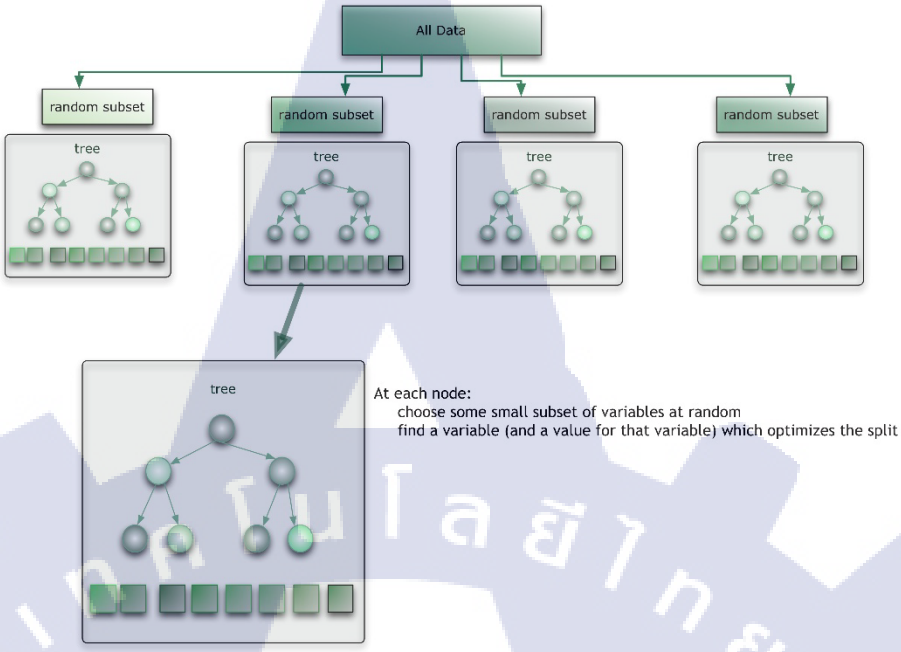
การสร้างแบบจำลองการจำแนกประเภทข้อมูล (Data Classification model) [10] จะเกิดจากการหาความสัมพันธ์ของข้อมูลในฐานข้อมูลขนาดใหญ่ โดยจะแบ่งข้อมูลทั้งหมดออกเป็น 2 กลุ่ม คือ กลุ่มข้อมูลเรียนรู้ (Training set) เป็นชุดข้อมูลที่ได้มาจากการเรียนรู้ข้อมูลส่วนใหญ่ เพื่อสร้างแบบจำลองการจำแนกประเภทข้อมูลออกมา และกลุ่มข้อมูลทดสอบ (Test set) เป็นชุดข้อมูลที่จะได้รับการตรวจสอบความถูกต้องของแบบจำลองการจำแนกประเภทข้อมูล



รูปที่ 2.6 ขั้นตอนการ training dataset และการ testing dataset

Classifier ที่ได้รับความนิยมในการใช้งานมากที่สุด และถูกเลือกใช้ในงานวิจัยในครั้งนี้ก็คือ Random Forest

Random Forest เป็นการเรียนรู้วิธีการทั้งหมดสำหรับการจำแนก, การถดถอย และงานอื่น ๆ ที่ดำเนินการโดยการสร้างความหลากหลายของต้นไม้ตัดสินใจ ณ เวลาการเทรน และการแสดงผลของคลาส ซึ่งเป็นโหมดของคลาส (การจัดหมวดหมู่) หรือการทำนายเฉลี่ย (การถดถอย) ของแต่ละต้นไม้ ในขั้นตอนการทำงานของ Random forest จะจำแนกต้นไม้หลาย ๆ ต้น โดยที่ต้นไม้แต่ละต้นจะถูกสร้างขึ้นจากกลุ่มตัวอย่างที่แตกต่างกัน จากกระบวนการของต้นไม้ตัดสินใจ (tree classification algorithm) ถูกสร้างขึ้นจนกลายเป็นป่า (forest)



รูปที่ 2.7 model of Random Forest [11]

จากรูปที่ 2.7 จะเห็นได้ว่าโมเดลของ Random forest เกิดการจากรวมตัวกันเยอะ ๆ ของ Decision tree โดยที่ Decision tree จะประกอบไปด้วย node, vector และความสัมพันธ์ของข้อมูล เมื่อนำ Decision tree หลาย ๆ ต้นมารวมกันเป็นป่า จะเกิดเป็นโมเดลของ Random forest ซึ่งเป็นโมเดลที่ทำการสุ่มเลือกชุด Training Data และสุ่มเลือก Feature ออกมาหลาย ๆ ชุด จากนั้นจะนำมาสร้าง Decision Tree หลาย ๆ ต้น แต่ละต้นก็จะให้คำตอบออกมา และในขั้นตอนสุดท้ายจะประมวลผลคำตอบของต้นไม้แต่ละต้น และแสดงเป็นผลลัพธ์สุดท้าย เช่น การทำ majority vote

นอกจากนี้ยังมีอัลกอริทึม Support Vector Machine หรือ SVM ที่เป็น classifier ที่เป็นที่นิยมอีกด้วย โดยที่รูปแบบของ อัลกอริทึม Support Vector Machine เป็นรูปแบบการเรียนรู้โมเดลพร้อมกับการเรียนรู้อัลกอริทึมที่เกี่ยวข้องซึ่งจะวิเคราะห์ข้อมูลที่ใช้สำหรับการจำแนกและการวิเคราะห์การถดถอย โดยนำมาใช้อย่างกว้างขวางในการคัดแยกในด้านการประมวลผลเป็นภาพดิจิตอล

2.3.6 Confusion Matrix

Confusion Matrix [12] เป็นตารางที่มักใช้ในการอธิบายถึงประสิทธิภาพของรูปแบบการจำแนก (Classifier) บนชุดของข้อมูลการทดสอบที่ทราบค่าที่แท้จริง confusion matrix ค่อนข้างที่จะเข้าใจง่าย แต่คำศัพท์ที่เกี่ยวข้องอาจจะทำให้เกิดความสับสน

รูปแบบ confusion matrix

n=235	Predicted: NO	Predicted: YES	
Actual: NO	TN = 60	FP = 15	75
Actual: YES	FN = 10	TP = 150	160
	70	165	

คำศัพท์พื้นฐาน:

- True positives (TP): เป็นเหตุการณ์ที่ทำนายเป็น yes แล้วผลลัพธ์เป็น yes
- True negatives (TN): เป็นเหตุการณ์ที่ทำนายเป็น no แล้วผลลัพธ์เป็น no
- False positives (FP): เป็นเหตุการณ์ที่ทำนายเป็น yes แต่ผลลัพธ์เป็น no
- False negatives (FN): เป็นเหตุการณ์ที่ทำนายเป็น no แต่ผลลัพธ์เป็น yes

สูตรอัตราที่คำนวณจาก confusion matrix สำหรับ binary classifier

- Accuracy: classifier ทำงานถูกต้องบ่อยแค่ไหน
  ตัวอย่าง (TP+TN)/total = (150+60)/235 = 0.89
- Misclassification Rate: classifier ทำงานผิดบ่อยแค่ไหน หรือที่รู้จักกันว่า “Error Rate”
  ตัวอย่าง (FP+FN)/total = (15+10)/235 = 0.11
- True Positive Rate: ผลลัพธ์เป็น yes ทำนายเป็น yes บ่อยแค่ไหน หรือที่รู้จักกันว่า “Recall”
  ตัวอย่าง TP/actual yes = 150/160 = 0.94
- False Positive Rate: ผลลัพธ์เป็น no ทำนายเป็น yes บ่อยแค่ไหน
  ตัวอย่าง FP/actual no =15/75 = 0.2
- Specificity: ผลลัพธ์เป็น no ทำนายเป็น no บ่อยแค่ไหน
  ตัวอย่าง TN/actual no = 60/75 = 0.8

- Precision: ถ้าทำนายเป็น yes classifier ทำงานถูกต้องบ่อยแค่ไหน
  ตัวอย่าง TP/predicted yes = 150/165 = 0.91
- Prevalence: มีผลลัพธ์เป็น yes เท่าไหร่
  ตัวอย่าง actual yes/total = 160/235 = 0.68

2.4 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการวัดระดับเนื้อหาเว็บไซต์การท่องเที่ยว ตามความต้องการของนักท่องเที่ยว โดยการวิเคราะห์ข้อความในเว็บไซต์ มีดังนี้

Juan Ramos [13] ได้ทำการศึกษาและตรวจสอบผลของการใช้ Term Frequency - Inverse Document Frequency (TF-IDF) เพื่อระบุว่าคำใดในคลังข้อมูลของเอกสารที่อาจจะใช้ได้ดีหรือเป็นประโยชน์มากกว่าการสืบค้นข้อมูล จากข้อมูลที่เก็บรวบรวมจะเห็นว่า TF-IDF แสดงเอกสารที่มีค่าสูงที่เกี่ยวข้องกับข้อความค้นหา หากผู้ใช้ป้อนข้อมูลในแบบสอบถามสำหรับหัวข้อใดหัวข้อหนึ่ง TF-IDF สามารถค้นหาเอกสารได้ที่มีข้อมูลที่เกี่ยวข้องในข้อความค้นหา นอกจากนี้การเข้ารหัส TF-IDF ทำได้โดยง่าย เหมาะสำหรับการสร้างพื้นฐานสำหรับความซับซ้อนที่มากขึ้นของอัลกอริทึมและระบบการสืบค้นข้อมูล

Rachsuda Jiamthapthaksin [14] ได้นำเสนอการสร้างแบบจำลองสำหรับข้อความภาษาไทย เพื่อเปลี่ยนโพสต์บน Facebook เป็นกลุ่มความสนใจของผู้ใช้ โดยเลือกใช้ Latent Dirichilet Allocation หรือ LDA หากใช้โดยตรงกับข้อความภาษาไทยจะไม่สามารถจับความสนใจของกลุ่มได้ เนื่องจากลักษณะเฉพาะของข้อมูล เช่น พิมพ์ผิดโดยเจตนา ผลงานจากงานวิจัย คือ การรวมคำแสลงภาษาไทยจากโพสต์ สำหรับการสกัดคำ การลบคำหยาบ และการประยุกต์ใช้ LDA โดยการทดลองจะทดลองบน Facebook ของนักศึกษาอาสาสมัครจากมหาลัยอัสสัมชัญ เพื่อแสดงให้เห็นถึงการลดขนาดคุณลักษณะ, การเพิ่มประสิทธิภาพของรูปแบบ และการค้นพบความสนใจกลุ่มที่มีความหมาย ซึ่งผลลัพธ์ระบบสามารถลดขนาดคุณลักษณะลงเหลือ 41.8% และระบบยังสามารถค้นหา 7 กลุ่มความสนใจจากโพสต์ของนักศึกษาอัสสัมชัญได้อีกด้วย

เกรียงกมล คำมา และจักรกฤษณ์ เสน่ห์ นมะหุต [15] ได้ศึกษาการแก้ปัญหาการวิเคราะห์เว็บไซต์อนาจารด้วยเทคนิควิเคราะห์ข้อความภาษาไทยในโครงสร้างเว็บไซต์ และ Semantic Network ซึ่งพบว่าระบบที่ได้พัฒนามานั้นไม่สามารถวิเคราะห์เว็บไซต์อนาจารที่มีเนื้อหาบางส่วนที่เป็นภาษาอังกฤษรวมอยู่ ดังนั้นผู้วิจัยจึงได้ทำการพัฒนาระบบต่อยอดเพื่อให้รองรับการวิเคราะห์เว็บอนาจารสำหรับภาษาอังกฤษ จากการทดสอบวิเคราะห์เว็บไซต์อนาจารภาษาอังกฤษ 200 เว็บไซต์ พบว่าประสิทธิภาพ ในการวิเคราะห์ถูกต้องสูงถึง 94 %

ทิชากร เนตรสุวรรณ์ [16] นำเสนอการวิเคราะห์ข่าวภาษาอังกฤษด้านอาชญากรรมด้วยเทคนิคการทำเหมืองข้อความ โดยนำเสนอการจำแนกประเภทข้อมูลข่าวอาชญากรรมออกเป็น 5 หมวดหมู่ และเลือกใช้แบบจำลองโครงข่ายประสาทเทียมในการจำแนกหมวดหมู่ โดยที่ประสิทธิภาพการจำแนกมีผลออกมาค่อนข้างดี โดยมีค่าแม่นยำเท่ากับ 84.16% ค่าความระลึก เท่ากับ 83.40% และ F-measure เท่ากับ 83.49%

คมคิด ชัชราภรณ์, ธรา อั่งสกุล และจิติมนต์ อั่งสกุล [17] ศึกษาการพัฒนาแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยว โดยแบ่งคุณลักษณะออกเป็น 2 กรณี ได้แก่ เลือกคุณลักษณะจากชื่อสถานที่ท่องเที่ยวเท่านั้น และเลือกคุณลักษณะจากชื่อและคำอธิบายของสถานที่ท่องเที่ยวนั้น ซึ่งเลือกใช้กระบวนการทั้งหมด 4 เทคนิค คือ ต้นไม้ตัดสินใจ นาอีฟเบย์ส ซัพพอร์ทแวกเตอร์แมชชีน และเคเนียเรสเนเบอร์ โดยเทคนิคที่ได้ค่าความถูกต้องที่มากและแม่นยำที่สุด คือ นาอีฟเบย์ส โดยเฉลี่ยอยู่ที่ 88.70%

กฤษฎิ์ บรรณะชัยศิริสุข [18] ได้ศึกษาวิธีการพัฒนาเครื่องมือช่วยเหลือผู้พิการทางสายตาให้เข้าถึงเนื้อหาบนหน้าเว็บไซต์ โดยเลือกใช้ Random forest ในการค้นหาเนื้อหา และนำมาเปลี่ยนรูปแบบการนำเสนอเป็นแบบใหม่ ที่สามารถให้ผู้พิการทางสายตาเข้าถึงและสามารถใช้งานได้ โดยผู้จัดทำได้สร้างเครื่องมือการเข้าเว็บไซต์ที่สามารถช่วยลดระยะเวลา รวมถึงกำจัดสิ่งแปลกปลอมที่ไม่เกี่ยวข้องกับเนื้อหาหลัก ซึ่งผู้ใช้สามารถเข้าใจถึงเนื้อหาบนเว็บไซต์ได้อย่างรวดเร็วขึ้นตามจุดประสงค์

นัทธี ศรีหาจักษ์ [19] นำเสนอการออกแบบและพัฒนาระบบในการรวบรวมสารสนเทศจากแหล่งข้อมูลที่จัดเก็บเป็นเอกสารที่มีโครงสร้าง โดยใช้เทคนิควิธีการจัดกลุ่มโดยใช้อัลกอริทึมเค-มีนส์ และตัววัดทีเอฟ-ไอดีเอฟ เพื่อบอกความเกี่ยวข้องของเอกสาร ผลลัพธ์การค้นคืนสารสนเทศในการวิจัยนี้ประเมินค่าด้วย Precision, Recall และค่า F-measure โดยได้ค่าเฉลี่ยที่ 83% 84% และ 83% ซึ่งอยู่ในระดับที่ค่อนข้างสูง

Retno Kusumaningrum, Satriyo Adhy, M. Ihsan Aji Wiedjayanto และ Suryono [20] ได้นำเสนอการใช้งาน Latent Dirichlet Allocation หรือ LDA ในการจำแนกประเภทบทความข่าวในภาษาอินโดนีเซียเป็น 5 ประเภท ได้แก่ เศรษฐกิจ, การท่องเที่ยว, อาชญากรรม, กีฬา และการเมือง การตั้งค่าการทดลองดำเนินการเพื่อประเมินผลการปฏิวัติตามวิธีการจำแนกตาม LOA โดยใช้กลยุทธ์การตรวจสอบความถูกต้อง 10-fold Cross-validation ผลการทดลองแสดงให้เห็นความถูกต้องโดยรวมประมาณ 70%

Wongkot Sriurai Phayung Meesad และ Choochart Haruechaiyasak [7] ได้เสนอแนวคิดการแทนเอกสารด้วยวิธี Topic Model ให้เอกสารสำหรับการเรียนรู้ในการจำแนกประเภทเว็บเพจ โดยใช้อัลกอริธึม Latent Dirichlet Allocation เพื่อสร้างแบบจำลองความน่าจะเป็นในการจัดกลุ่มของหัวข้อที่ซ่อนอยู่ในเอกสาร โดยจะประเมินผลอยู่ 3 วิธี ได้แก่ 1. การแทนเอกสารด้วยวิธี

Bag of Words 2. การแทนเอกสารด้วยการสร้างแบบจำลองหัวข้อ 3. การแทนเอกสารด้วยการสร้างแบบจำลองหัวข้อรวมกับหน้าเว็บข้างเคียง ซึ่งผลการทดลองพบว่าวิธีจำการแทนเอกสารด้วยการสร้างแบบจำลองหัวข้อรวมกับหน้าเว็บข้างเคียงให้ประสิทธิภาพที่สูงที่สุด โดยมีค่า F1 เท่ากับ 85.01% ซึ่งมากกกว่า Bag of Words ถึง 23.81%

ธันยพร กรานสุข และธนพล เจนสุทธิเวชกุล [21] ได้ประยุกต์ใช้เทคนิคเหมืองข้อความในการนำมาสกัดเอาคุณลักษณะเฉพาะ และสร้างแบบจำลองการจำแนกปัจจัยที่ส่งผลต่อการผิดพลาดที่ก่อให้เกิดอุบัติเหตุ โดยเปรียบเทียบประสิทธิภาพอัลกอริทึม ระหว่างอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน และอัลกอริทึมนาอีฟเบย์ ผลจากการทดสอบพบว่าอัลกอริทึมที่มีความแม่นยำ และมีประสิทธิภาพในการจำแนกได้มากกว่า โดยผลการวัดประสิทธิภาพโดยรวม เท่ากับ 97.87%

จิรเมธ ว่องพรรณงาม [22] ได้ประยุกต์โปรแกรมรับภาพความลึกจากกล้องคิเน็กต์ เพื่อตรวจสอบการสัปหงกของผู้ขับรถ เพื่อป้องกันการเกิดอุบัติเหตุ การตรวจสอบการสัปหงกจะตรวจสอบค่ามุมพิตช์ของท่าทางของศีรษะผู้ขับ ซึ่งการทดสอบได้เปรียบเทียบค่ามุมพิตช์ที่วัดจากการประมาณค่าด้วยวิธี Random forest โดยทดสอบค่าความไวกับกลุ่มภาพจำนวน 160 ภาพ ผลลัพธ์ที่ได้คือ 93.75%

ชูพันธุ์ รัตนโภคา และเมธาวี สุทธิกุล [23] ได้ออกแบบและพัฒนาระบบจัดหมวดหมู่สถานที่ท่องเที่ยวด้วยคำอธิบายภาษาไทยแบบอัตโนมัติ โดยแบ่งหมวดหมู่ออกเป็น 4 หมวดหมู่ ได้แก่ วัด ทะเล ภูเขา และที่พัก โดยใช้โปรแกรม Weka ในการเรียนรู้ข้อมูลด้วยเทคนิคซัพพอร์ทเวคเตอร์แมชชีน โดยใช้คำสำคัญของแต่ละหมวดหมู่สถานที่ละ 10 คำ และฝึกข้อมูลเพียง 120 คำเท่านั้น ให้ความแม่นยำในการจัดหมวดหมู่ถึง 95%

บทที่ 3
วิธีดำเนินการศึกษา

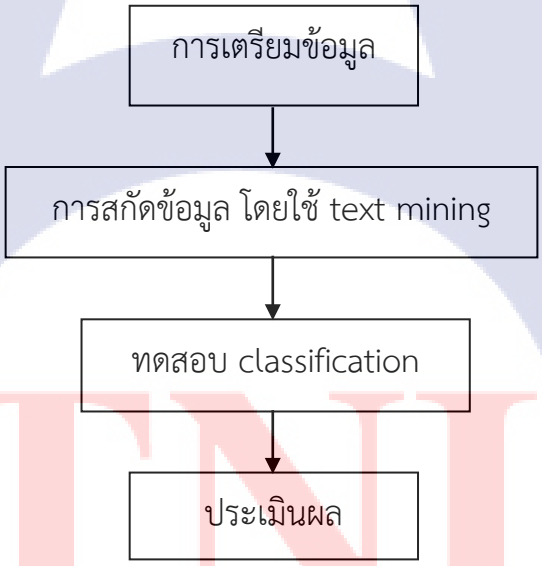
จากการศึกษาแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องเพื่อเป็นความรู้สำหรับการศึกษาในครั้งนี้ ผู้ศึกษาได้นำความรู้ที่ได้รับมาดำเนินการในการตรวจสอบเนื้อหาเว็บไซต์การท่องเที่ยวตามความต้องการของนักท่องเที่ยว โดยการวิเคราะห์ข้อความในเว็บไซต์ ซึ่งการทำงานในครั้งนี้จะประกอบด้วย 4 ส่วนหลัก คือ การหา keywords, การหา feature vector, การสร้าง classification model, และการทดสอบ classifier โดยมีขั้นตอนต่อไปนี้

3.1 ภาพรวมของเทคนิคการทำงาน

3.2 รายละเอียดการดำเนินงาน

- การสกัดข้อมูลด้วย text mining
- การสร้าง Model
- การทดสอบ classifier

3.1 ภาพรวมของเทคนิคการทำงาน

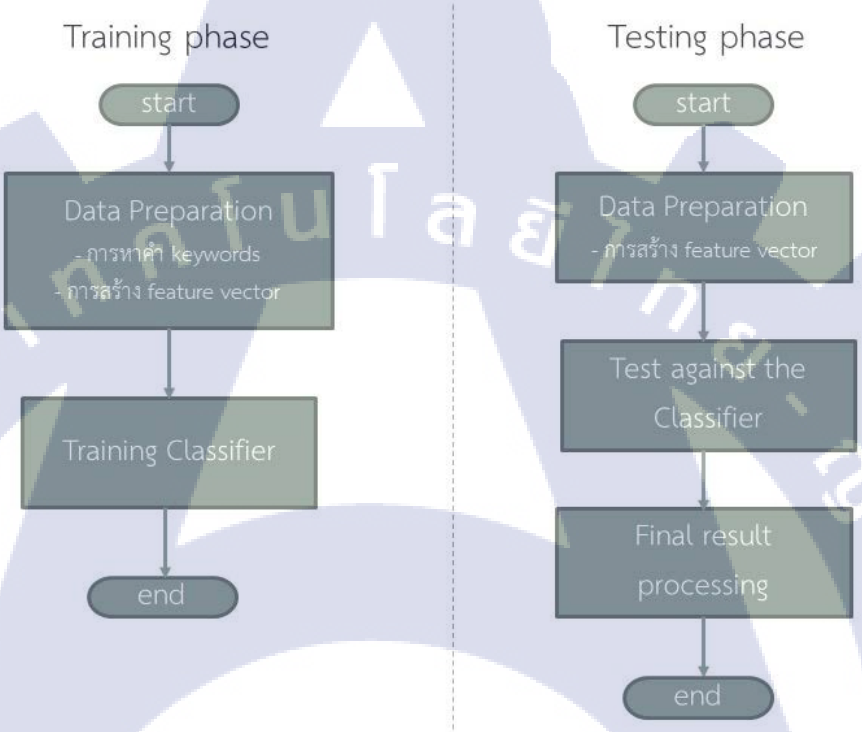


รูปที่ 3.1 Flowchart การทำงาน

ในการทำงานวิจัยในครั้งนี้ ใช้แนวทางของ Supervised Learning ดังนั้นข้อมูลในการศึกษาครั้งนี้จะถูกแบ่งออกเป็น 2 ส่วน คือ ส่วนของการทำ training และส่วนของการทำ testing

จากรูปที่ 3.2 จะเห็นได้ว่ามีการทำงานอยู่ 2 ส่วน ในส่วนของการทำ training เริ่มจากการเตรียมข้อมูล ซึ่งในการเตรียมข้อมูลของการ training จะทำงานอยู่ 2 อย่าง คือ การหาคำ keyword และการสร้าง feature vector หลังจากนั้นเมื่อทำการเตรียมข้อมูลเสร็จ ก็จะทำการ training

classifier ซึ่งเป็นขั้นตอนในการสร้าง model สำหรับงานวิจัยในครั้งนี้ ส่วนส่วนที่ 2 คือการทำ testing โดยจากทำ testing จะเริ่มจากการเตรียมข้อมูลเช่นกัน แต่ในส่วนนี้ ไม่ต้องหาคำ keyword จะทำแค่การสร้าง feature vector เท่านั้น เมื่อทำการเตรียมข้อมูลเสร็จเรียบร้อยแล้ว จะนำข้อมูลไปทดสอบผ่าน classifier หรือ model ที่ได้จากการ training เมื่อได้ทำการทดสอบเรียบร้อย ก็สามารถนำผลลัพธ์ไปวัดระดับหรือการทำสถิติ ตามขั้นตอน final result processing ตามขั้นตอนในรูปที่ 3.2

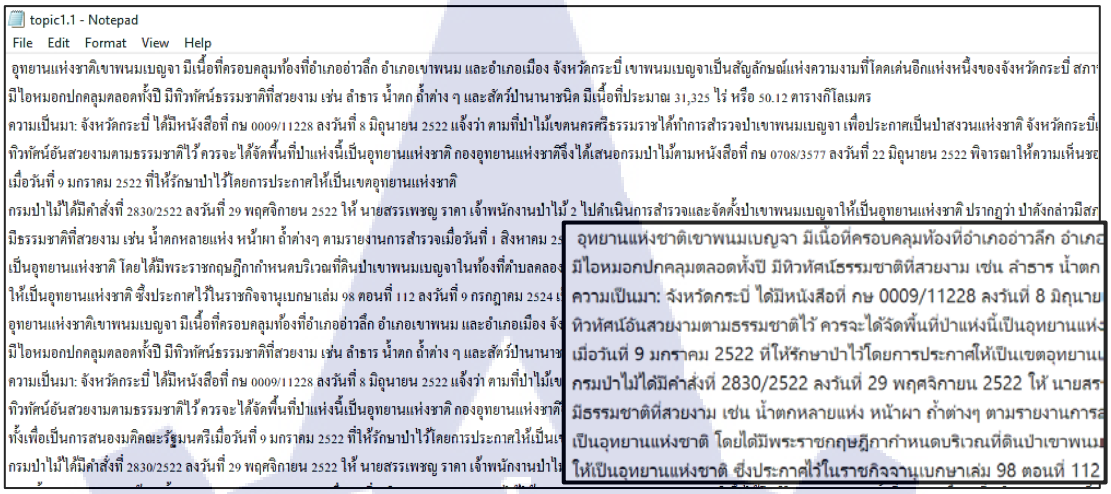


รูปที่ 3.2 ภาพรวมของเทคนิคที่นำเสนอ

3.2 รายละเอียดการดำเนินงาน

3.2.1 การเตรียมข้อมูล

ในขั้นตอนการเตรียมข้อมูล ข้อมูลที่นำมาใช้ทดสอบ เป็นข้อมูลที่ได้มาจากเว็บเพจที่มีอยู่จริง และนำข้อมูลนั้นมาแปลงในรูปแบบของ text file



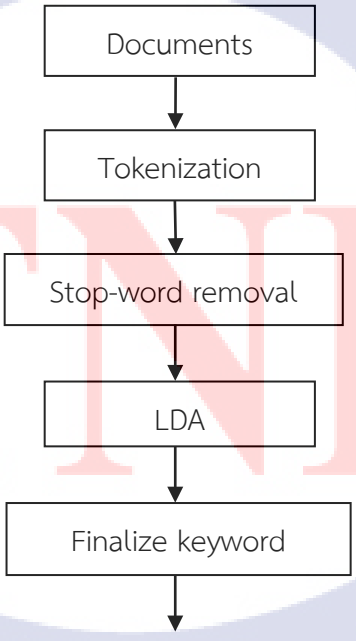
รูปที่ 3.3 ข้อมูลในรูปแบบ text file

3.2.2 การสกัดข้อมูลด้วย text mining

ในขั้นตอนการสกัดข้อมูลหรือขั้นตอนการเตรียมข้อมูล จะมีการทำงานอยู่ 2 แบบ คือ การหาคำหลักหรือคำ keywords กับการสร้าง features vector ของแต่ละเอกสาร โดยจะอธิบายขั้นตอนได้ดังนี้

3.2.2.1 การหาคำหลักหรือคำ keyword

การสกัดข้อมูลเพื่อหาคำหลักหรือคำ keywords มีกระบวนการอยู่ทั้งหมด 4 ขั้นตอน ดังนี้



รูปที่ 3.4 flowchart การหาคำ keyword

1) การแบ่งคำหรือการ tokenization เป็นการแบ่งคำออกเป็นคำ ๆ ซึ่งได้ทำการทดลองอยู่ 2 เทคนิคที่เป็นที่นิยมมากที่สุดในการแบ่งคำภาษาไทย คือ deepcut และ pythainlp จากผลการทดลองการแบ่งคำด้วยเทคนิค pythainlp ผลลัพธ์ที่ได้ออกมาเป็นไปอย่างที่ต้องการและแบ่งคำได้เป็นคำที่เกี่ยวข้องกับความต้องการของนักท่องเที่ยวมากกว่า และยังสามารถแบ่งคำออกได้เข้าใจมากกว่า จึงเลือก pythainlp ในการแบ่งคำในงานครั้งนี้

['การ', 'เดินทาง', 'รถ', 'ยนต์', 'จาก', 'ตัวเมือง', 'เชียงใหม่', 'ไป', 'ตาม', 'ทาง', 'หลวง', 'หมายเลข', 'ไป', 'ตาม', 'ถนน', 'ฝาง', 'ม่อน', 'ปิ่น', 'เลี้ยว', 'ขวา', 'ไป', 'ตาม', 'ถนน', 'รพช.', 'จะ', 'ถึง', 'บริเวณ', 'บ่อ', 'น้ำ', 'ร้อน', 'ฝาง', 'ซึ่ง', 'เป็น', 'ที่', 'ตั้ง', 'ที่', 'ทำ', 'การ', 'อุทยานแห่งชาติดอยฟ้า', 'ห่ม', 'ปก', 'รถ', 'โดยสาร', 'ประจำ', 'ทาง', 'มี', 'รถ', 'ประจำ', 'ทาง', 'ปรับ', 'อากาศ', 'ของ', 'บริษัทขนส่งจำกัด', 'และ', 'บริษัท', 'รถ', 'ร่วม', 'เอกชน', 'ระหว่าง', 'กรุงเทพ', 'ฝาง', 'เชียงใหม่', 'ฝาง', 'เมื่อ', 'ถึง', 'อำเภอ', 'ฝาง', 'จะ', 'มี', 'รถ', 'รับ', 'จ้าง', 'คอย', 'บริการ', 'รับ', 'ส่ง', 'สู่', 'ที่', 'ทำ', 'การ', 'อุทยานแห่งชาติดอยฟ้า', 'ห่ม', 'ปก', 'เดินทาง', 'ด้วย', 'รถ', 'ขับ', 'เคลื่อน', 'สี่', 'ล้อ', 'ตาม', 'ถนน', 'สายฝาง', 'บ้าน', 'ห้วยบอน', 'เมื่อ', 'ถึง', 'บ้าน', 'ห้วยบอน', 'แล้ว', 'ตรง', 'ไป', 'ตาม', 'ถนน', 'ลูกรัง', 'อีก', 'จะ', 'ถึง', 'ที่', 'ตั้ง', 'ลาน', 'กางเตินท์กิ่วลม']

ผลลัพธ์การแบ่งคำของเทคนิค deepcut

['การ', 'เดินทาง', 'รถยนต์', 'จาก', 'ตัวเมือง', 'เชียงใหม่', 'ไป', 'ตาม', 'ทางหลวง', 'หมายเลข', 'ไป', 'ตาม', 'ถนน', 'ฝาง', 'ม่อน', 'ปิ่น', 'เลี้ยวขวา', 'ไป', 'ตาม', 'ถนน', 'รพช.', 'จะ', 'ถึง', 'บริเวณ', 'บ่อน้ำร้อน', 'ฝาง', 'ซึ่ง', 'เป็นที่ตั้ง', 'ที่ทำการ', 'อุทยานแห่งชาติ', 'ดอย', 'ฟ้า', 'ห่ม', 'ปก', 'รถโดยสารประจำทาง', 'มี', 'รถประจำทาง', 'ปรับอากาศ', 'ของ', 'บริษัท', 'ขนส่ง', 'จำกัด', 'และ', 'บริษัท', 'รถร่วม', 'เอกชน', 'ระหว่าง', 'กรุงเทพ', 'ฝาง', 'เชียงใหม่', 'ฝาง', 'เมื่อ', 'ถึง', 'อำเภอ', 'ฝาง', 'จะ', 'มี', 'รถรับจ้าง', 'คอย', 'บริการ', 'รับ', 'ส่ง', 'สู่', 'ที่ทำการ', 'อุทยานแห่งชาติ', 'ดอย', 'ฟ้า', 'ห่ม', 'ปก', 'เดินทาง', 'ด้วย', 'รถ', 'ขับเคลื่อน', 'สี่', 'ล้อ', 'ตาม', 'ถนน', 'สาย', 'ฝาง', 'บ้าน', 'ห้วย', 'บอน', 'เมื่อ', 'ถึง', 'บ้าน', 'ห้วย', 'บอน', 'แล้ว', 'ตรง', 'ไป', 'ตาม', 'ถนนลูกรัง', 'อีก', 'จะ', 'ถึง', 'ที่ตั้ง', 'ลาน', 'กาง', 'เต็นท์', 'กิ่ว', 'ลม']

ผลลัพธ์การแบ่งคำของเทคนิค pythainlp

2) หลังจากทำการแบ่งคำแล้ว ก็จะตัดคำ stop-words ออก ซึ่ง stop-words หมายถึง คำที่ไม่มีความหมายต่อเนื้อหาในเอกสารนั้น ๆ อีกทั้งสามารถตัดทิ้งได้ และไม่ทำให้ประสิทธิภาพในการค้นหาต่ำลง โดยในงานวิจัยในครั้งนี้จะใช้การ stop-words ด้วยกัน 2 รูปแบบ คือ

1) stop-words ที่มากับ package ของโปรแกรม ซึ่งโดยทั่วไป stop-words ในลักษณะนี้ จะมีแค่คำบุพบท และคำสันธาน เช่น ที่ และ กับ ด้วย แต่ เพราะ ซึ่ง เป็นต้น

2) customized stop-words เป็น stop-words ที่จัดทำขึ้นเอง เนื่องจากต้องการที่จะตัดคำที่ไม่เกี่ยวข้องออก โดยที่คำที่ไม่เกี่ยวข้องนั้น จะเป็นคำที่ package ของโปรแกรมควบคุมไม่ทั่วถึง ซึ่ง stop-words ที่จัดทำขึ้นมาใหม่นั้น จะมีประเภทของคำเป็นคำบุพบท คำสันธาน คำวิเศษณ์ คำลักษณนาม คำเฉพาะ และสรรพนาม เช่น

เซนติเมตร	แผ่น
มิลลิเมตร	ภูเก็ต
ความหนาแน่น	เป็นอย่างมาก
ราชบุรี	กรุงเทพ
ยามเช้า	ประจวบคีรีขันธ์
ต่อมา	ตามลำดับ
โดยทั่วไปแล้ว	ครั้งแรก
ด้านนอก	ประกอบด้วย
จังหวัด	สังคม
เขียว	ความสูง
ขึ้นอยู่	สลับซับซ้อน

ตัวอย่างคำ stop-words ที่ทำขึ้นเอง

3) ขั้นตอนต่อมาเป็นการหา keyword โดยใช้เทคนิค Latent Dirichlet Allocation หรือ LDA ในการหา LDA คือการหาหัวข้อในเอกสาร ในขั้นตอนนี้เราต้องกำหนดจำนวนหัวข้อ และจำนวนคำในแต่ละหัวข้อ ซึ่งผู้วิจัยได้ทำการทดลองโดยทดลองจำนวนหัวข้อ 5 10 และ 15 โดยที่ในแต่ละหัวข้อกำหนดจำนวนคำเป็น 10 20 30 40 และ 50 วนครบตามลำดับ จากการที่ได้ทำการทดลอง เราเลือกใช้การหา keywords โดยใช้ LDA ที่กำหนดจำนวนหัวข้อเป็น 10 และจำนวนคำในแต่ละหัวข้อเป็น 10 คำ เนื่องจากผู้วิจัยได้ทดลองใช้จำนวนหัวข้อและจำนวนคำที่มากขึ้นหรือน้อยลงกว่านี้ ก็ไม่ได้ทำให้คำ keywords มีการเปลี่ยนแปลงมากนัก การเลือกใช้เป็น 10 topics 10 words จึงเป็นค่าที่เหมาะสมที่สุดในการหา keywords ในงานในครั้งนี้

ตัวอย่าง Output จากการหา keywords ด้วยเทคนิค Latent Dirichlet Allocation

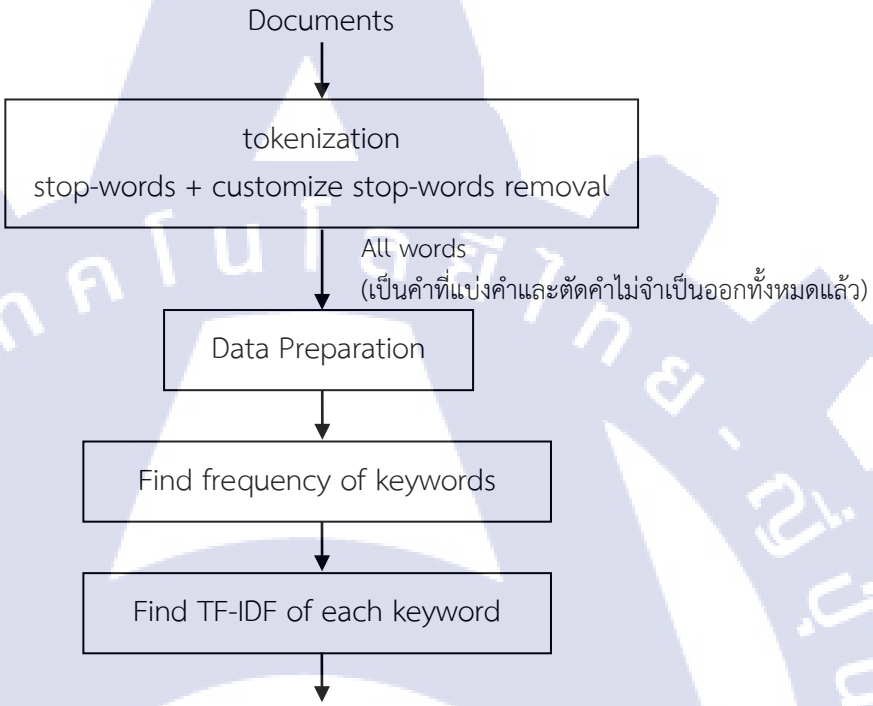
(0, '0.128*"เดินทาง" + 0.054*"เส้นทาง" + 0.052*"อุทยานแห่งชาติ" + 0.043*"รถยนต์" + 0.033*"น้ำตก" + 0.020*"ตัวเมือง" + 0.018*"สถานีรถไฟ" + 0.014*"กิโลเมตร" + 0.014*"เลี้ยวขวา" + 0.014*"เลี้ยวซ้าย"')

(1, '0.079*"อุทยานแห่งชาติ" + 0.032*"สุราษฎร์ธานี" + 0.032*"ปิง" + 0.032*"เข้าสู่" + 0.031*"เดินทาง" + 0.024*"ตะกั่ว" + 0.024*"ลี้" + 0.016*"พังงา" + 0.016*"เครื่องบิน" + 0.016*"รถด่วน"')
(2, '0.003*"จัตุรัส" + 0.003*"ชัยภูมิ" + 0.003*"ขอนแก่น" + 0.003*"ต่อรถ" + 0.003*"อุทยานแห่งชาติ" + 0.003*"เครื่องบิน" + 0.003*"สามล้อ" + 0.003*"พัน" + 0.003*"รถยนต์" + 0.003*"โป่ง"')
(3, '0.049*"เดินทาง" + 0.037*"อุทยานแห่งชาติ" + 0.037*"ทับ" + 0.037*"ร่อง" + 0.037*"เบิก" + 0.037*"เส้นทาง" + 0.037*" " + 0.037*"กล้า" + 0.025*"รถยนต์" + 0.025*"รถตู้"')
(4, '0.028*"รถบัส" + 0.028*"ออกเดินทาง" + 0.028*"เครื่องดื่ม" + 0.028*"อาหารว่าง" + 0.028*"มุ่งหน้า" + 0.028*"วีไอพี" + 0.028*"อุทยานแห่งชาติ" + 0.003*"เดินทาง" + 0.003*"รถยนต์" + 0.003*"ตัวเมือง"')
(5, '0.076*"สถานีรถไฟ" + 0.054*"รถยนต์" + 0.051*"เดินทาง" + 0.041*"เส้นทาง" + 0.039*"เรือ" + 0.028*"อุทยานแห่งชาติ" + 0.028*"กรุง" + 0.028*"ท่าเรือ" + 0.024*"น้ำตก" + 0.020*"สถานีขนส่ง"')
(6, '0.072*"เดินทาง" + 0.043*"สนามบิน" + 0.043*"สตูล" + 0.042*"เชียงราย" + 0.037*"รถยนต์" + 0.026*"แอร์" + 0.026*"หาดใหญ่" + 0.026*"ตั้งอยู่" + 0.021*"อุทยานแห่งชาติ" + 0.018*"หน้า"')
(7, '0.100*"พิษณุโลก" + 0.061*"สัก" + 0.041*"แก่ง" + 0.041*"น้ำตก" + 0.022*"ขวามือ" + 0.022*"เพชรบูรณ์" + 0.022*"มุ่งหน้า" + 0.022*"รถยนต์" + 0.022*"ป้าย" + 0.022*"ขับรถ"')
(8, '0.062*"เดินทาง" + 0.056*"รถยนต์" + 0.045*"อุทยานแห่งชาติ" + 0.034*"สถานีขนส่ง" + 0.034*"ตัวเมือง" + 0.034*"น้ำตก" + 0.023*"ส่วนตัว" + 0.023*"เลี้ยวซ้าย" + 0.023*"ริม" + 0.017*"เส้นทาง"')
(9, '0.100*"เดินทาง" + 0.068*"รถยนต์" + 0.042*"สนามบิน" + 0.034*"อุทยานแห่งชาติ" + 0.033*"เรือ" + 0.023*"ปัว" + 0.023*"บ่อ" + 0.023*"เกลือ" + 0.014*"เครื่องบิน" + 0.014*"แม่ฮ่องสอน"')

4) หลังจากได้คำมาทั้งหมด 100 คำ จากทุกเอกสารในแต่ละหัวข้อ โดยที่คำหลักที่ต้องการมีเพียงแค่ 10 คำหลัก ซึ่งเราใช้วิธีหาจากความถี่ของคำทั้ง 100 คำ คำใดที่มีความถี่มากที่สุด และเป็นคำที่เกี่ยวกับกับความต้องการในข้อนั้น ๆ จะถูกกำหนดเป็นคำหลักหรือคำที่สำคัญในหัวข้อนั้น ๆ

3.2.2.2 การสร้าง features vector ของแต่ละเอกสาร

การสร้าง features vector เริ่มจากการหาค่า TF-IDF ของคำ keywords ทุกคำ ในแต่ละหัวข้อ หลังจากนั้นก็จะนำค่า TF-IDF มาสร้างเป็น features vector เพื่อนำไปใช้ในการทำ classification ต่อไป โดยมีทั้งหมด 5 กระบวนการ



รูปที่ 3.5 Flowchart การทำงานของการหา features vector ของแต่ละเอกสาร

1) เนื่องจากได้ทำการทดลองเทคนิคการแบ่งคำมาแล้วตามด้านบน ครั้งนี้จึงนำเนื้อหาจากเว็บไซต์การท่องเที่ยวในรูปแบบของ text file มา tokenize หรือแบ่งคำด้วยเทคนิค pythainlp และทำ stop-words เอาคำไม่จำเป็นออกเช่นเดียวกับในการหาคำหลัก

2) หลังจากที่ได้คำทั้งหมดโดยที่ไม่มีคำไม่จำเป็นแล้ว ก็นำคำเหล่านั้น มาเก็บใส่ไว้ใน dataframe ซึ่งเป็นการเก็บข้อมูลอีกรูปแบบนึง เพื่อที่สามารถนำไปใช้ได้ง่ายในขั้นตอนต่อไป

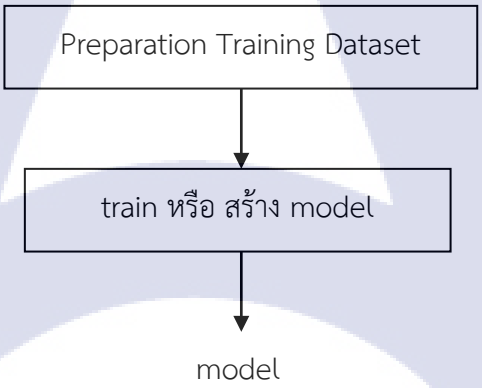
3) จากนั้นเราจะนำเอกสารที่มี keywords ของแต่ละหัวข้อมา และใช้เทคนิค vectorizer ซึ่งเป็นเทคนิคที่ใช้ในการช่วยหาคำ keywords ว่ามีคำ keywords ทั้งหมดกี่คำในเอกสารนั้น ๆ ซึ่งในเอกสารไม่จำเป็นต้องมีคำ keywords ครบทุกคำ

4) หาความถี่ของคำ keywords ที่พบในเอกสาร เมื่อได้ความถี่ของแต่ละ keywords ออกมาแล้ว จากนั้นก็นำมาหาค่า TF-IDF ซึ่งเป็นการหาค่าน้ำหนักของความถี่ของคำจากคำทั้งหมดในเอกสาร มาสร้าง feature vector ของ keywords ทั้ง 10 คำ

5) เมื่อได้ feature vector ของ keywords ในแต่ละหัวข้อมาแล้ว จะนำค่านั้นบันทึกลง file เพื่อนำข้อมูลไปใช้ในการ training data set ในขั้นตอนถัดไป

3.2.3 การสร้าง Model

ในการสร้าง Model สร้างขึ้นเพื่อใช้ตรวจสอบค่า TF-IDF ของคำ keywords กับค่า TF-IDF ที่มาจากข้อมูลใหม่จากเว็บไซต์การท่องเที่ยว ว่าข้อมูลใหม่นั้นมีค่า TF-IDF ที่ใกล้เคียงกับค่า TF-IDF ของความต้องการของนักท่องเที่ยวในข้อใดมากที่สุด เนื่องจากการทดลองนี้เป็นแบบ supervised learning คือ ต้องมีการเทรนหรือฝึกสอนข้อมูลจากข้อมูลจริง ซึ่งจะมีขั้นตอนสร้าง model ดังต่อไปนี้



รูปที่ 3.6 Flowchart การทำงานของการสร้าง model

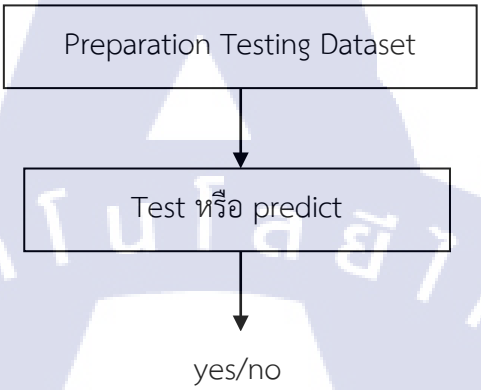
1) ในการสร้าง model อย่างแรกที่ต้องทำคือการเตรียมข้อมูลในการ train ซึ่งเป็นการเอาค่า TF-IDF ของทุก ๆ เอกสารในแต่ละหัวข้อมากำหนด label เพื่อเป็นการสอนระบบว่าค่า TF-IDF ลักษณะแบบใดใกล้เคียงหรือเป็นอยู่ในหัวข้อนั้นใด ๆ ตามความต้องการของนักท่องเที่ยว

2) จากนั้นนำข้อมูลที่เตรียมไว้มา train ด้วย classification โดยได้มีการทดลองการ train ข้อมูลกับ classification กับทั้งหมด 3 ประเภท ได้แก่ Decision Tree, Supple Vector Machine และ Random Forest โดยทั้ง 3 ประเภทได้รับความนิยมเป็นอย่างมาก ซึ่งผลลัพธ์ที่ได้ออกมา Random Forest ได้ผลลัพธ์ออกมาค่อนข้างดีที่สุด จึงเลือก classifier Random Forest มาใช้งานวิจัยในครั้งนี้

3) เมื่อ train ข้อมูลเรียบร้อยแล้ว ก็จะได้ model ออกมา ซึ่งจะเป็น model keywords ต้นแบบของในแต่ละหัวข้อในความต้องการของนักท่องเที่ยว

3.2.4 การทดสอบ classifier

การทดสอบ classifier เป็นการทดสอบเพื่อดูว่าเนื้อหาจากเว็บไซต์การท่องเที่ยวข้อมูลใหม่ที่นำมานั้น มีการให้เนื้อหาครบถ้วนตามความต้องการของนักท่องเที่ยวครบทุกข้อหรือไม่ โดยที่จะนำค่า TF-IDF ที่สร้างมาเป็น model มาเปรียบเทียบกับค่า TF-IDF จากข้อมูลใหม่ เพื่อดูความคล้ายคลึงกันของข้อมูล โดยขั้นตอนการทำงานมีดังนี้

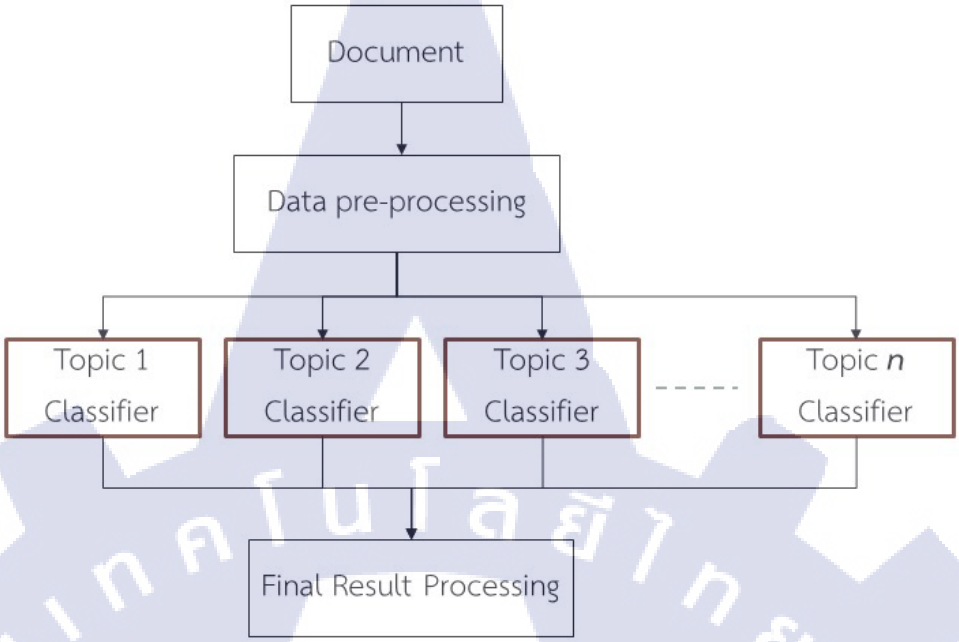


รูปที่ 3.7 Flowchart การทำงานของการทดสอบ classification

1) ในการทดสอบ model จะมีการเตรียมข้อมูลคล้ายกับการ training dataset โดยการนำค่า TF-IDF ที่ไม่เคยได้ train มาก่อน หรือระบบไม่เคยรู้จักมาก่อน แต่เนื้อหายังเป็นหัวข้อเดียวกัน มาทดสอบ และ label ให้เป็นหัวข้อนั้นๆ

2) จากนั้นนำข้อมูลมาทดสอบ โดยใช้คำสั่ง predict หรือคำสั่งทำนายข้อมูล ผลลัพธ์ในการทำการทดลองนี้จะออกมาเป็น yes กับ no ซึ่งถ้าค่า TF-IDF ตรงกับ model ในหัวข้อที่กำหนดไว้ ผลลัพธ์ก็จะออกมาเป็น yes แต่ถ้าไม่ตรงกับค่า TF-IDF ใน model ที่นำมาทดสอบ ผลลัพธ์ก็จะออกมาเป็น no

ดังนั้นจากรูปที่ 3.7 การดำเนินงานของการทำงานทั้งระบบ โดยที่เริ่มจากการนำเอกสารที่จากการเว็บเพจมา เพื่อนำไปเตรียมข้อมูลที่ผ่านการแปลงข้อมูลเป็นรูป text file การแบ่งคำ และการนำคำไม่จำเป็นออก หลังจากนั้นก็นำข้อมูลส่งไปทดสอบกับ classifier แต่ละหัวข้อ โดยการทำงานของเรา สามารถเพิ่มหัวข้อได้ตามความเหมาะสม โดยใช้การทำงานในลักษณะเดียวกันในการหาหัวข้อ หลังจากนำไปหา classifier แล้ว ก็จะได้ผลลัพธ์ออกมา สามารถนำผลลัพธ์มาสรุปผลเป็นกราฟ หรือวัดระดับในรูปแบบของดาว หรือที่เรียกว่าขั้นตอน final result processing



รูปที่ 3.8 Flowchart ภาพรวมการทำงานของระบบ

บทที่ 4
ผลการทดลอง

จากการทดลองการตรวจสอบเนื้อหาเว็บไซต์ของหัวข้อการท่องเที่ยว ทำให้ผู้ศึกษาเล็งเห็นว่าควรมีการทดสอบประสิทธิภาพของระบบ เนื่องจากเป็นการช่วยยืนยันว่าระบบมีประสิทธิภาพในการใช้งานได้จริง ในบทนี้จะแสดงให้การทดลองและผลการทดลองของระบบทั้งหมดที่ผู้ศึกษาได้ทำการทดลอง

4.1 การออกแบบการทดลอง

ในการทดลองครั้งนี้ เลือกใช้การทำ 10 fold Cross-validation ซึ่งเป็นการทดลองที่ใช้สำหรับทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือมากขึ้น ซึ่งจะใช้วิธีการแบ่งข้อมูลออกเป็น 10 ชุดเท่าๆ กัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้

ในการทดลองแต่ละครั้งจะถูกแบ่งการทดลองออกเป็น 2 รูปแบบ คือ

1. การทดลองโดยการทดสอบแบบ single topic
   ในแต่ละเอกสาร มีเพียง 1 หัวข้อ เพื่อเป็นการทดสอบประสิทธิภาพ classifier ของแต่ละหัวข้อ หากถ้ามี 3 classifier ก็จะถูกทำการทดลองแบบเดียวกันในทุก ๆ ข้อมูล
2. การทดลองโดยการทดสอบแบบ multiple topics
   ใน 1 เอกสาร มี 2 หัวข้อ และมี 3 หัวข้อ

กรณีที่เป็นไปได้แบบ 2 หัวข้อ

2 หัวข้อ	กรณีเป็นไปได้ที่ 1	กรณีเป็นไปได้ที่ 2	กรณีเป็นไปได้ที่ 3
หัวข้อที่ 1	√	√	
หัวข้อที่ 2	√		√
หัวข้อที่ 3		√	√

กรณีที่เป็นไปได้แบบ 3 หัวข้อ

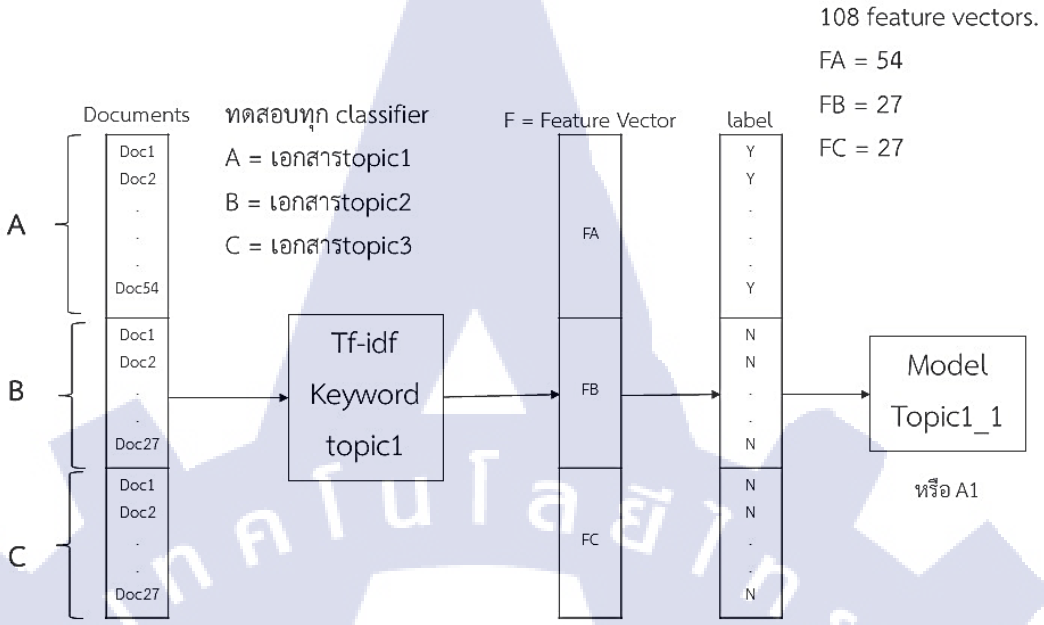
3 หัวข้อ	กรณีเป็นไปได้ที่ 1
หัวข้อที่ 1	√
หัวข้อที่ 2	√
หัวข้อที่ 3	√

4.1.1 การทดลองโดยการทดสอบแบบ single topic

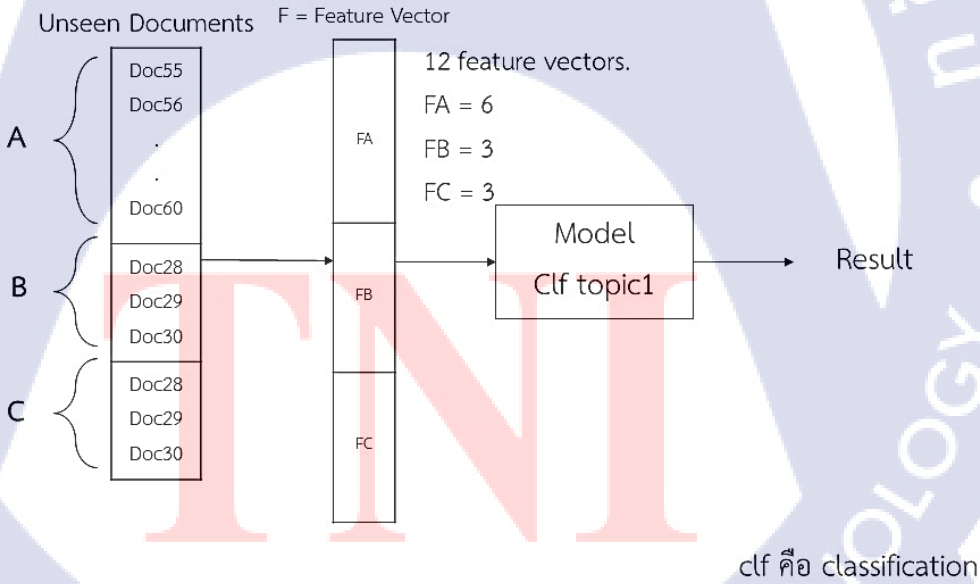
ในการทำการทดสอบแบบ single topic เป็นการทำเพื่อที่จะต้องการวัดประสิทธิภาพของ classification ของแต่ละหัวข้อ โดยที่การทำ 10 fold Cross-validation มีขั้นตอนดังนี้

1. นำเอกสารทั้งหมด โดยสมมุติให้มี 60 เอกสาร ทำการ tokenize และ stop-words removal ตามปกติ หลังจากนั้น ให้หาค่า TF-IDF จาก keyword ของแต่ละ topic
2. เมื่อได้ค่า TF-IDF ออกมาแล้ว นำค่า TF-IDF จาก keyword ของ topic เดียวกัน ทั้งหมดมาแบ่งเป็น 10 ชุดข้อมูลเท่าๆ กัน ซึ่งจะได้เป็นชุดข้อมูลละ 6 เอกสาร จากนั้นนำชุดข้อมูลมา 9 ชุด ซึ่งรวมแล้วเป็น 54 เอกสารมา label ค่าให้เป็น yes และนำค่า TF-IDF จาก keyword เดิม ใน topic อื่นมา label ให้เป็น no ทั้ง 54 เอกสารเช่นกัน
3. หลังจากนั้นเมื่อได้ label ค่าทั้งหมดแล้ว ก็นำข้อมูลไป training data set ให้ออกมาเป็นโมเดล
4. จากนั้นนำข้อมูลชุดที่เหลืออีก 1 ชุดที่ไม่ได้นำไป training ซึ่งจะมีอยู่ 6 เอกสาร นำข้อมูลนั้นมา testing กับโมเดลที่สร้างไว้ในขั้นตอนที่ 3
5. หลังจากนั้นวนนำข้อมูลแต่ละชุดมา test ให้ครบทั้ง 10 ชุดข้อมูล ก็จะได้ตัวทดสอบ classification ในแต่ละ keywords ออกมา ซึ่งนำมาช่วยทดสอบประสิทธิภาพของระบบ ในกรณีที่จะทดสอบแบบ multiple topics

ดังภาพที่ 4.1 และ ภาพที่ 4.2 โดยที่จะต้องทำแบบเดียวกันทั้งกระบวนการนี้เพิ่มในอีก 2 หัวข้อที่เหลืออยู่



รูปที่ 4.1 ขั้นตอนการ training data ของการทดสอบแบบ single topic

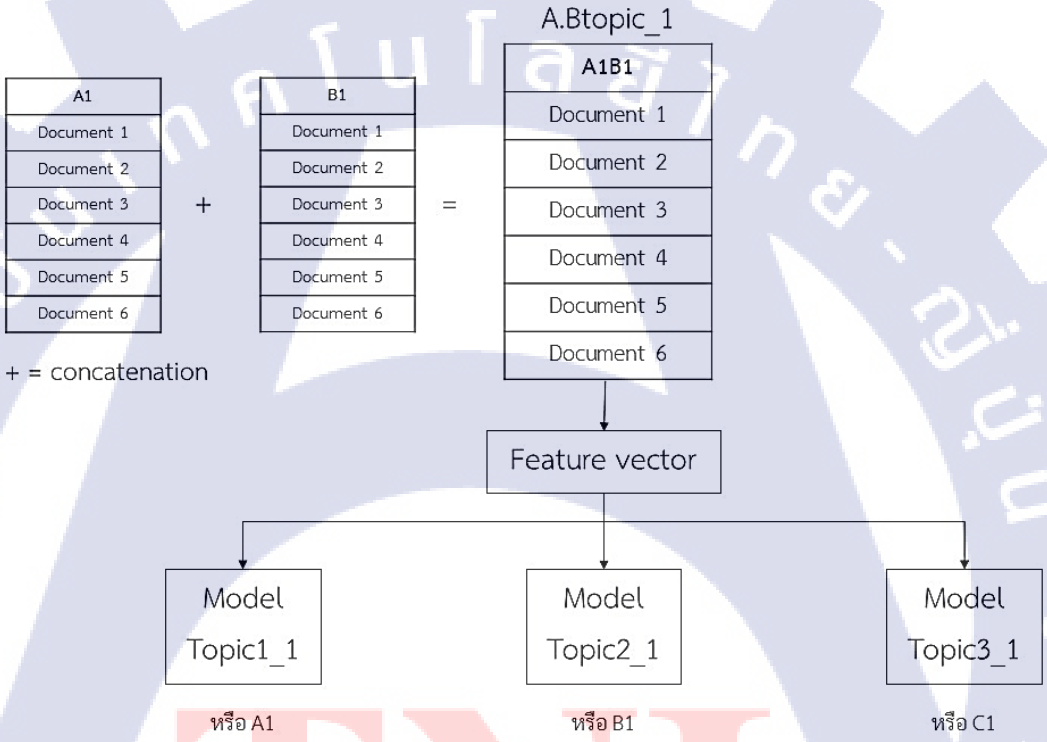


รูปที่ 4.2 ขั้นตอนการ testing data ของการทดสอบแบบ single topic

4.1.2 การทดลองโดยการทดสอบแบบ multiple topics

มี 2 แบบ ได้แก่ การรวมตัวของ 2 หัวข้อ และการรวมตัวของ 3 หัวข้อ

ขั้นตอนในการทดลองของ multiple topic นำเอาเอกสารจากแต่ละหัวข้อมารวมกันเป็นเอกสารเดียว โดยที่สามารถรวมได้ทั้ง 2 หัวข้อ และ 3 หัวข้อ หลังจากนั้นก็นำมาทำการ tokenization, stop-words removal และหาค่า TF-IDF เพื่อสร้าง feature vector และมาแบ่งข้อมูลออกเป็น 10 ชุดข้อมูลเท่าๆ กัน จะได้ชุดข้อมูลละ 6 เอกสาร นำชุดข้อมูลไป test กับ model ที่ได้จากการทดสอบในแบบ 1 หัวข้อ ดังรูปที่ 4.3



รูปที่ 4.3 ขั้นตอนการ testing data ของการทดสอบแบบmultiple topics

หลังจากนั้นก็จะได้ผลลัพธ์ออกมา ทำให้ครบทั้งทุกชุดข้อมูล หรือทำทั้งหมด 10 ครั้ง ซึ่งเมื่อได้ผลลัพธ์ออกมาแล้ว จะนำผลลัพธ์นั้นมาหาค่า confusion matrix

ในรูปแบบ multiple topics นี้ จะสามารถหาค่าได้แค่ค่าของ Recall ซึ่งเป็นค่าที่มาจากค่า False Negatives หารด้วย จำนวน total ของชุดข้อมูลทั้งหมด

4.2 การทดสอบประสิทธิภาพด้วย dataset หัวข้อการท่องเที่ยว

4.2.1 การออกแบบการทดลองหัวข้อการท่องเที่ยว

ในการทดลองนี้จะทำการทดลองกับหัวข้อการท่องเที่ยว ซึ่งในหัวข้อข่าวจะมีหัวข้ออยู่ทั้งหมด 3 หัวข้อ ได้แก่ หัวข้อการเดินทาง หัวข้อแหล่งท่องเที่ยว และหัวข้อสถานที่พัก และในแต่ละหัวข้อมีเอกสารอยู่ 60 เอกสาร จึงได้ทำการออกแบบการทดลองโดยที่แบ่งข้อมูลออกเป็น 10 ชุด ตามกฎของ 10 fold Cross-validation และในการทดลองแต่ละครั้งจะถูกแบ่งการทดลองออกเป็น 2 รูปแบบ คือ

1. การทดลองโดยการทดสอบแบบ single topic
2. การทดลองโดยการทดสอบแบบ multiple topics

4.2.2 การทดลองประสิทธิภาพของ classifier แบบ 1 หัวข้อ

<u>การทดลองที่ 1</u>

การทดลองประสิทธิภาพของ classifier แบบ 1 หัวข้อ จะแบ่งข้อมูลออกเป็นดังนี้ตารางที่ 4.1 โดยที่ช่องข้อมูลที่ highlight จะเป็นชุดข้อมูลที่นำไป test ในแต่ละการทดสอบแต่ละครั้ง

ตารางที่ 4.1 แสดงการแบ่งข้อมูลของการทำการทดลอง 10 fold Cross-validation

การทดสอบ	ชุดข้อมูล 1	ชุดข้อมูล 2	ชุดข้อมูล 3	ชุดข้อมูล 4	ชุดข้อมูล 5	ชุดข้อมูล 6	ชุดข้อมูล 7	ชุดข้อมูล 8	ชุดข้อมูล 9	ชุดข้อมูล 10
1	6	6	6	6	6	6	6	6	6	6
2	6	6	6	6	6	6	6	6	6	6
3	6	6	6	6	6	6	6	6	6	6
4	6	6	6	6	6	6	6	6	6	6
5	6	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	6	6	6
7	6	6	6	6	6	6	6	6	6	6

ตารางที่ 4.1 แสดงการแบ่งข้อมูลของการทำการทดลอง 10 fold Cross-validation (ต่อ)

การทดสอบ	ชุดข้อมูล 1	ชุดข้อมูล 2	ชุดข้อมูล 3	ชุดข้อมูล 4	ชุดข้อมูล 5	ชุดข้อมูล 6	ชุดข้อมูล 7	ชุดข้อมูล 8	ชุดข้อมูล 9	ชุดข้อมูล 10
8	6	6	6	6	6	6	6	6	6	6
9	6	6	6	6	6	6	6	6	6	6
10	6	6	6	6	6	6	6	6	6	6

ในการทำการทดสอบในรูปแบบ single topic เมื่อได้ผลลัพธ์ออกมา จะนำผลลัพธ์ที่ออกมาแสดงผ่านรูปแบบ confusion matrix ตามตารางที่ 4.2

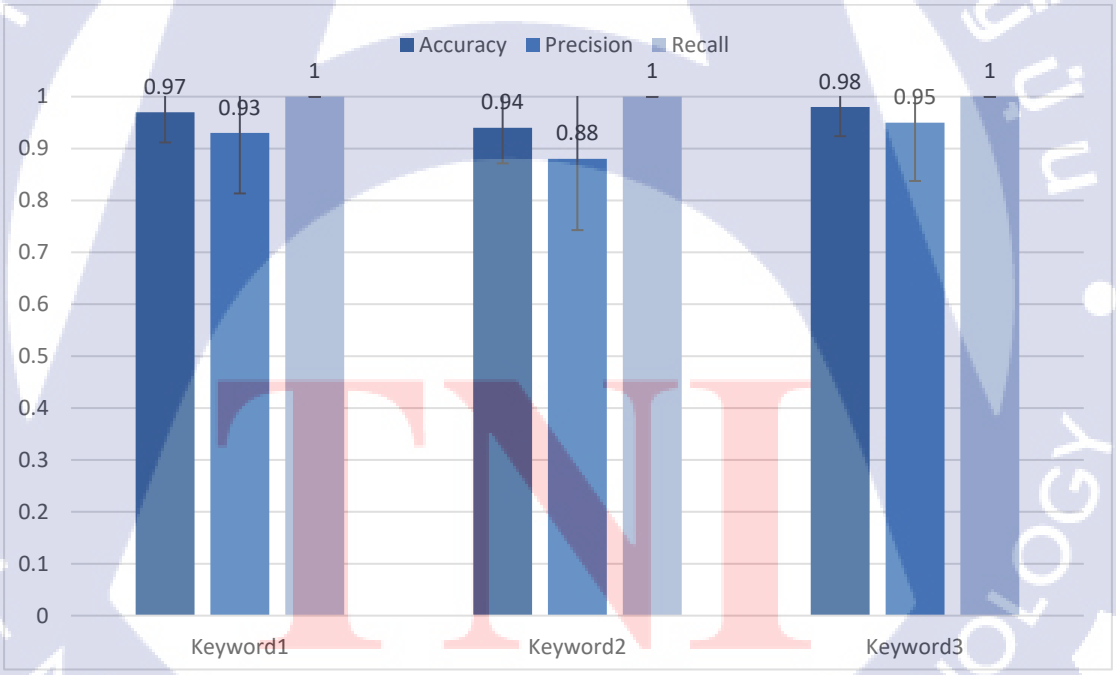
ตารางที่ 4.2 แสดงข้อมูลแบบ confusion matrix ของการทดลอง 10 fold Cross-validation แบบ single topic

	True Negatives	False Negatives	False Positives	True Positives
การทดสอบ 1	18	0	2	16
การทดสอบ 2	18	0	2	16
การทดสอบ 3	18	0	3	15
การทดสอบ 4	18	0	3	15
การทดสอบ 5	18	0	1	17
การทดสอบ 6	18	0	0	18
การทดสอบ 7	18	0	0	18

ตารางที่ 4.2 แสดงข้อมูลแบบ confusion matrix ของการทดลอง 10 fold Cross-validation แบบ single topic (ต่อ)

	True Negatives	False Negatives	False Positives	True Positives
การทดสอบ 8	18	0	0	18
การทดสอบ 9	18	0	1	17
การทดสอบ 10	18	0	1	17

จากรูปที่ 4.4 แสดงถึงประสิทธิภาพโดยรวมของ classifier แต่ละหัวข้อ โดยแสดงค่าเฉลี่ยของ Accuracy, Precision และ Recall โดยที่แทบ error bar คือค่าบวกลบของค่า SD หรือ ค่าเบี่ยงเบนมาตรฐาน



รูปที่ 4.4 แผนภูมิแสดงค่าเฉลี่ยของ Accuracy, Precision และ Recall ของทุก classifier

จากรูปที่ 4.4 จะเห็นได้ว่ากราฟ 3 แท่งแรก คือ ประสิทธิภาพ classifier ของ keywords จากหัวข้อการเดินทาง โดยที่ค่า Accuracy จะอยู่ที่ 0.97 ค่า Precision อยู่ที่ 0.93 และค่า Recall

อยู่ที่ 1 ส่วนกราฟ 3 แท่งถัดมา คือ ประสิทธิภาพ classifier ของ keywords จากหัวข้อแหล่งท่องเที่ยว โดยที่ค่า Accuracy จะอยู่ที่ 0.94 ค่า Precision อยู่ที่ 0.88 และค่า Recall อยู่ที่ 1 และกราฟ 3 แท่งสุดท้าย จะมีค่า Accuracy อยู่ที่ 0.98 ค่า Precision อยู่ที่ 0.95 และค่า Recall อยู่ที่ 1 อีกเช่นกัน

4.2.3 การทดลองประสิทธิภาพของ classifier แบบ multiple topics

การทดลองที่ 1

การทดลองประสิทธิภาพของ classifier แบบ 2 หัวข้อ

ในการการทดลองนี้จะนำเนื้อหาของ 2 หัวข้อมารวมกัน โดยมีเนื้อหาจากหัวข้อแรกทั้งหมด 6 เอกสาร และหัวข้อที่ 2 อีก 6 เอกสาร นำเอกสารมารวมกัน แล้วหาค่า TF-IDF ของทั้ง 6 เอกสาร หลังจากนั้นก็นำทั้ง 6 เอกสาร ไปสร้างค่า feature vector ในแต่ละชุดข้อมูลมา test ใน classification ซึ่งผลลัพธ์ที่เกิดขึ้นจะต้องเป็น yes เท่านั้น

เมื่อได้ผลลัพธ์ออกมาแล้ว นำผลลัพธ์มาแสดงค่าในรูปแบบ confusion matrix ซึ่งผลจะเป็นดังที่แสดงในตารางที่ 4.3 และจะมี 2 ค่า คือ True Positives และ False Negatives ที่เป็นผลลัพธ์มาจากการทดลอง

ตารางที่ 4.3 แสดงค่า True Positives (TP) และ False Negatives (FN) ของการทดลอง 10 fold Cross-validation

	หัวข้อที่ 1 และ 2		หัวข้อที่ 1 และ 3		หัวข้อที่ 2 และ 3	
	FN	TP	FN	TP	FN	TP
การทดสอบ 1	3	3	2	4	0	6
การทดสอบ 2	2	4	0	6	2	4
การทดสอบ 3	2	4	1	5	3	3
การทดสอบ 4	4	2	3	3	4	2
การทดสอบ 5	1	5	0	6	2	4

ตารางที่ 4.3 แสดงค่า True Positives (TP) และ False Negatives (FN) ของการทดลอง 10 fold Cross-validation (ต่อ)

	หัวข้อที่ 1 และ 2		หัวข้อที่ 1 และ 3		หัวข้อที่ 2 และ 3	
	FN	TP	FN	TP	FN	TP
การทดสอบ 6	0	6	1	5	0	6
การทดสอบ 7	0	6	0	6	0	6
การทดสอบ 8	0	6	0	6	0	6
การทดสอบ 9	0	6	2	4	1	5
การทดสอบ 10	2	4	2	4	0	6

เนื่องจากการทดสอบในครั้งนี้เป็นการ test ข้อมูล ไม่จำเป็นต้องมีการ train ดังนั้นค่าที่สามารถนำมาใช้ได้ จะมีเพียงแค่ค่า Recall ซึ่งจากตารางที่ 4.4 จะแสดงค่า Recall ที่เกิดจากค่า True Positives หารด้วย ค่า Actual Yes

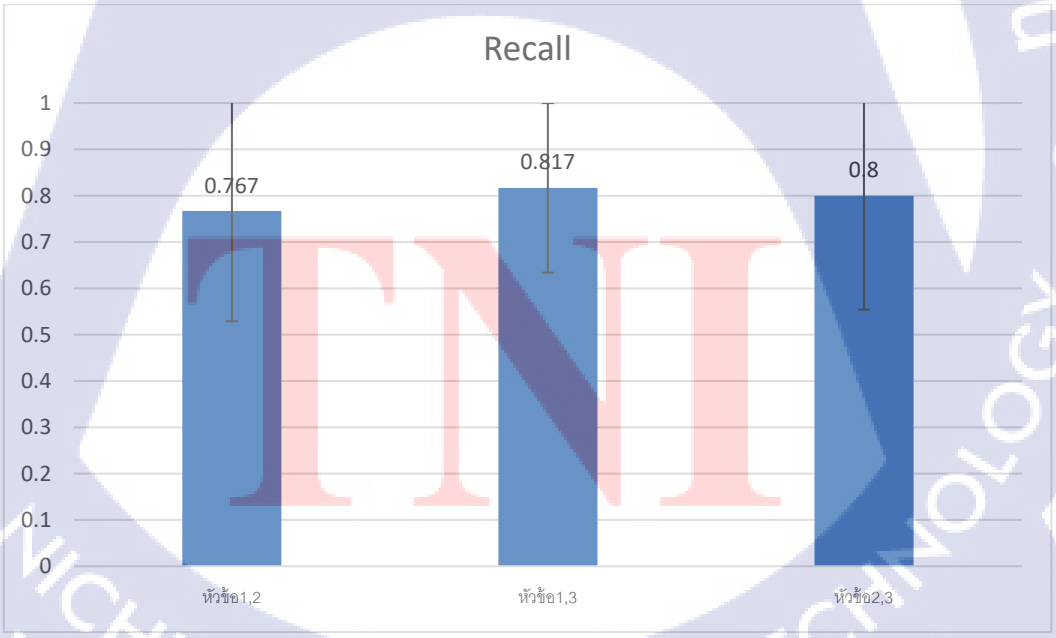
ตารางที่ 4.4 แสดงค่า Recall ของการรวมกันของ 2 หัวข้อ

	หัวข้อที่ 1 และ 2	หัวข้อที่ 1 และ 3	หัวข้อที่ 2 และ 3
การทดสอบ 1	0.5	0.67	1
การทดสอบ 2	0.67	1	0.67
การทดสอบ 3	0.67	0.86	0.5
การทดสอบ 4	0.33	0.5	0.33
การทดสอบ 5	0.83	1	0.67

ตารางที่ 4.4 แสดงค่า Recall ของการรวมกันของ 2 หัวข้อ (ต่อ)

	หัวข้อที่ 1 และ 2	หัวข้อที่ 1 และ 3	หัวข้อที่ 2 และ 3
การทดสอบ 6	1	0.83	1
การทดสอบ 7	1	1	1
การทดสอบ 8	1	1	1
การทดสอบ 9	1	0.67	0.83
การทดสอบ 10	0.67	0.67	1

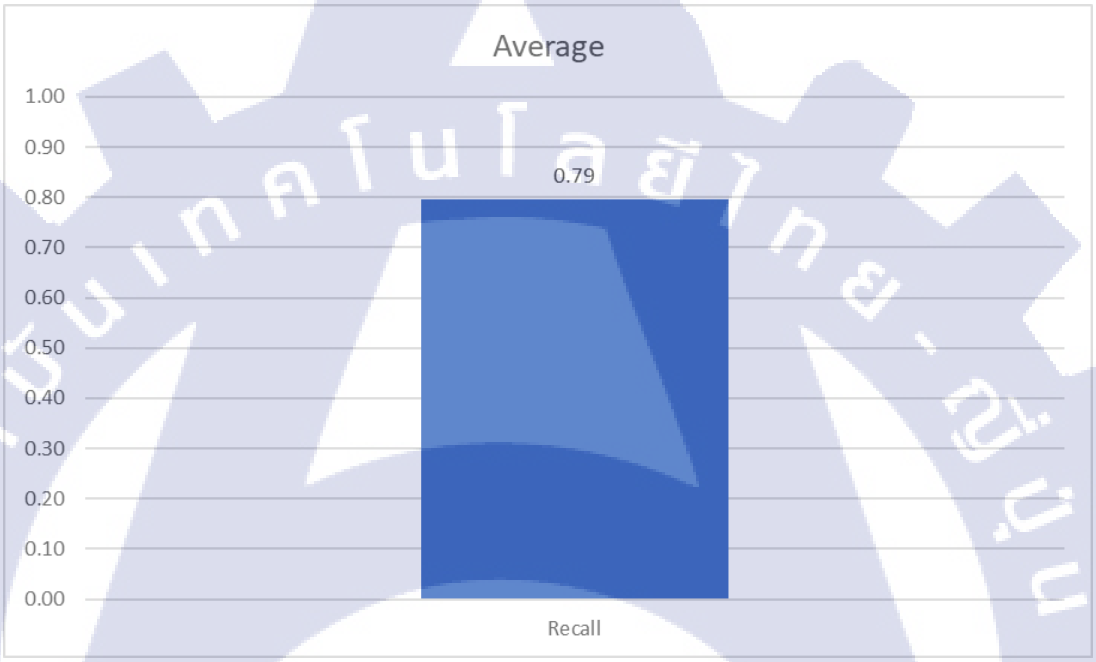
จากรูปที่ 4.5 จะแสดงค่าเฉลี่ย Recall จากข้อมูลการทดสอบทั้ง 10 การทดสอบ โดยที่แสดงค่าแยกออกเป็นทีละหัวข้อ จากตารางที่ 4.4 โดยที่ error bar เป็นค่า SD หรือค่าเบี่ยงเบนมาตรฐานจากทั้ง 10 การทดสอบ



รูปที่ 4.5 แผนภูมิค่าเฉลี่ย Recall ของการทดสอบ 10 fold cross validation ของการรวมตัวของ 2 หัวข้อ

โดยที่ค่าเฉลี่ย Recall ของการรวมเอกสารของหัวข้อที่ 1 และ 2 จะอยู่ที่ 0.77 ค่าเฉลี่ยน Recall ของการรวมเอกสารของหัวข้อที่ 1 และ 3 อยู่ที่ 0.82 และค่าเฉลี่ย Recall ของการรวมเอกสารของหัวข้อที่ 2 และ 3 อยู่ที่ 0.8

จากนั้นนำค่าเฉลี่ย Recall ของทั้ง 3 ข้อมูล มาหาค่าเฉลี่ย Recall ทั้งหมดของการรวมตัวกันของ 2 หัวข้อ ซึ่งผลออกมาอยู่ที่ 0.79 พบว่าค่อนข้างสูง ดังรูปที่ 4.6



รูปที่ 4.6 แผนภูมิค่าเฉลี่ย Recall ของการรวมตัวของ 2 หัวข้อ

<u>การทดลองที่ 2</u>

การทดลองประสิทธิภาพของ classifier แบบ 3 หัวข้อ

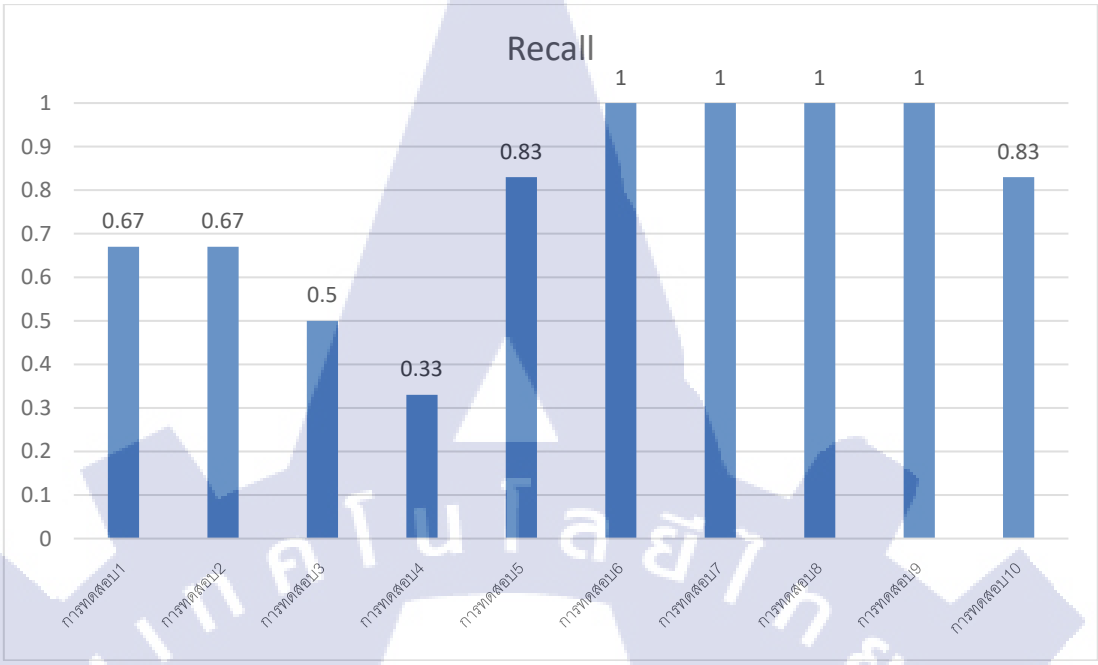
ในการการทดลองนี้จะนำเนื้อหาของ 3 หัวข้อมารวมกัน โดยมีเนื้อหาจากทั้ง 3หัวข้อมารวมกัน เป็นทั้งหมด 6 เอกสาร แล้วสร้างค่า TF-IDF ของทั้ง 6 เอกสาร หลังจากนั้นก็นำทั้ง 6 เอกสาร ไปหาค่า feature vector ในแต่ละชุดข้อมูลมา test ใน classification เช่นเดียวกันกับการทดลองที่ 1

เมื่อได้ผลลัพธ์ออกมาแล้ว นำผลลัพธ์มาแสดงค่าในรูปแบบ confusion matrix ซึ่งผลจะเป็นดังที่แสดงในตารางที่ 4.5 และจะมี 2 ค่า คือ True Positives และ False Negatives ที่เป็นผลลัพธ์มาจากการทดลอง

ตารางที่ 4.5 แสดงข้อมูลแบบ confusion matrix ของการทดลอง 10 fold Cross-validation แบบ 3 topic

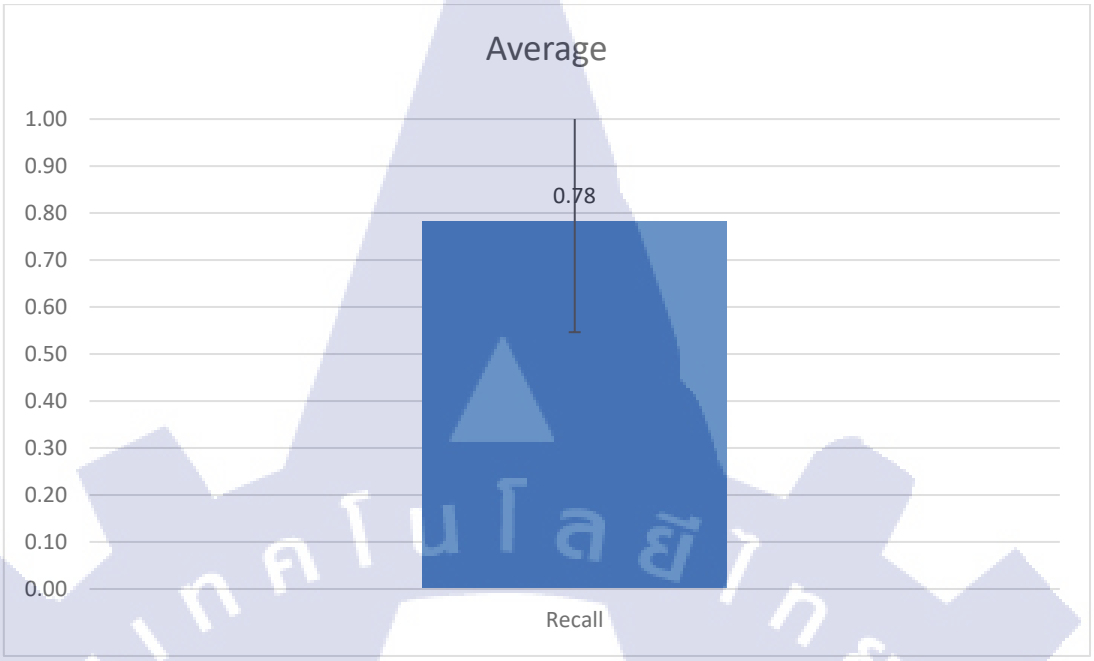
	True Negatives	False Negatives	False Positives	True Positives
การทดสอบ 1	0	0	2	16
การทดสอบ 2	0	0	2	16
การทดสอบ 3	0	0	3	15
การทดสอบ 4	0	0	4	14
การทดสอบ 5	0	0	1	17
การทดสอบ 6	0	0	0	18
การทดสอบ 7	0	0	0	18
การทดสอบ 8	0	0	0	18
การทดสอบ 9	0	0	1	17
การทดสอบ 10	0	0	1	17

เนื่องจากการทดสอบในครั้งนี้เป็นการ test ข้อมูล ไม่จำเป็นต้องมีการ train อีกเช่นกัน ดังนั้นค่าที่สามารถนำมาใช้ได้ จะเป็นค่า Recall ซึ่งจากรูปที่ 4.7 จะแสดงค่า Recall ที่เกิดจากค่า True Positives หารด้วย ค่า Actual Yes



รูปที่ 4.7 แผนภูมิ Recall ของการรวมตัวของ 3 หัวข้อ

จากนั้นนำค่า Recall จากการทั้ง 10 การทดสอบมาหาค่าเฉลี่ย เพื่อให้ทราบค่าเฉลี่ย Recall ของการตรวจสอบระบบทั้งหมด เนื่องจากเป็นการหาค่าเฉลี่ย Recall ของการรวมตัวกันทั้ง 3 หัวข้อ โดยที่ค่าเฉลี่ย Recall อยู่ที่ 0.78 ดังรูปที่ 4.8 และมี error bar เป็น ค่า SD หรือค่าเบี่ยงเบนมาตรฐาน



รูปที่ 4.8 แผนภูมิค่าเฉลี่ย Recall ของการรวมตัวของ 3 หัวข้อ

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

จากการศึกษาเรื่อง การวัดระดับเนื้อหาเว็บไซต์การท่องเที่ยว ตามความต้องการของนักท่องเที่ยว โดยการวิเคราะห์เว็บไซต์ ซึ่งการศึกษาในครั้งนี้มีวัตถุประสงค์ ข้อ ดังนี้

1. สร้างเฟรมเวิร์คเพื่อตรวจสอบคุณภาพเนื้อหาเกี่ยวกับเว็บไซต์การท่องเที่ยว
2. เพื่อเป็นแนวทางการปรับปรุงเนื้อหาเว็บไซต์การท่องเที่ยวให้ครบถ้วนตามความต้องการของนักท่องเที่ยว

สามารถสรุปผลแบ่งออกตามหัวข้อดังต่อนี้ไป

5.1 สรุปผล และอภิปรายผล

5.2 ข้อเสนอแนะ

5.3 ประโยชน์ที่ได้รับจากการทำสารนิพนธ์

5.1 สรุปผล และอภิปรายผล

ในการสรุปผลการศึกษา ผู้ศึกษาการจัดทำระบบการตรวจสอบเนื้อหาเว็บไซต์ให้ตรงกับความต้องการของนักท่องเที่ยว ซึ่งนักท่องเที่ยวมีความต้องการที่ต้องการรู้จากเว็บไซต์ ได้แก่ กา สถานที่พัก, การเดินทาง และแหล่งท่องเที่ยว จึงได้นำความต้องการของนักท่องเที่ยวมาเป็นตัวแปรในการสร้างหัวข้อในการหาข้อมูล โดยที่ผู้จัดทำได้จัดทำ framework ที่ใช้ในการตรวจสอบเนื้อหา และได้จัดทำ binary classifier ทั้งหมด 3 ตัวตามหัวข้อที่นักท่องเที่ยวสนใจ ซึ่ง classifier ทั้ง 3 ตัวนี้จะเป็นตัวช่วยในการแยกแยะหัวข้อ และยังค่อนข้างมีประสิทธิภาพในการแยกแยะได้ดีกับการทดลองในครั้งนี้อีกด้วย โดยผลการทดลองจะแบ่งออกเป็น 2 รูปแบบ คือ แบบ single topic และแบบ multiple topics ดังนี้

5.1.1 Single Topic

การทำการทดลองแบบ Single Topic ทำเพื่อตรวจสอบประสิทธิภาพของ classifier ในการแยกแยะหัวข้อของตัวเอง ซึ่งผลการทดลองโดยจะถูกแบ่งออกเป็น 3 classifier ตามหัวข้อความต้องการ โดยหัวข้อแรกการเดินทางได้ค่า Accuracy เท่ากับ 0.97 ค่า Precision เท่ากับ 0.93 และค่า Recall เท่ากับ 1 หัวข้อต่อมาหัวข้อแหล่งท่องเที่ยวได้ค่า Accuracy เท่ากับ 0.94 ค่า Precision เท่ากับ 0.88 และค่า Recall เท่ากับ 1 และหัวข้อสุดท้ายหัวข้อสถานที่พักได้ค่า Accuracy เท่ากับ 0.98 ค่า Precision เท่ากับ 0.95 และค่า Recall เท่ากับ 1 ซึ่งแสดงให้เห็นว่าความสามารถของ

classifier แต่ละตัวนี้ในการจดจำ และในการคาดเดาผลลัพธ์ของตัวเองนั้นทำงานได้ค่อนข้างเป็นอย่างดี

5.1.2 Multiple Topics

การทำการทดลองแบบ Multiple Topics ทำเพื่อตรวจสอบประสิทธิภาพของ classifier ว่าทำงานได้จริง ในกรณีที่มีเนื้อหามากกว่า 1 หัวข้อ ซึ่งในตามความเป็นจริง หน้าเว็บไซต์จะมีข้อมูลที่รวมกันอยู่หลากหลายหัวข้อเป็นปกติ โดยผลการทดลองแบบ multiple topics จะแบ่งออกเป็น 2 อย่าง ดังนี้

- การรวมตัวกันของ 2 topic จะมีค่าเฉลี่ย Recall จากการทดสอบ 10 fold Cross-validation อยู่ที่ 0.79
- การรวมตัวกันของ 3 topic จะมีค่าเฉลี่ย Recall จากการทดสอบ 10 fold Cross-validation อยู่ที่ 0.78

ซึ่งผลการทดลองนี้ได้ค่า Recall ค่อนข้างสูง แสดงให้เห็นว่าเป็นการทดลองที่มีประสิทธิภาพในการแยกแยะการตรวจสอบความครบถ้วนของข้อมูลในแต่ละเอกสาร ใน dataset ของเรานี้ มีข้อมูลแต่ละหัวข้อการท่องเที่ยวทั้งหมด 60 เอกสาร รวมทั้งหมด 180 เอกสาร ซึ่งถือว่าเป็น dataset ที่เพียงพอที่จะใช้ในการสร้าง model เพื่อจดจำความครบถ้วนของการท่องเที่ยวได้ ซึ่งส่งผลต่อทำให้เราสามารถเลือก keywords ได้ดีและค่อนข้างเจาะจง นอกจากนั้นแล้วในเอกสารทั้งหมดนี้ มีจำนวนคำที่ค่อนข้างใกล้เคียงกัน ซึ่งในหัวข้อการเดินทางมีคำเฉลี่ยอยู่ที่ 73.02 คำต่อ 1 เอกสาร หัวข้อแหล่งท่องเที่ยวมีคำเฉลี่ยอยู่ 34.47 คำ ต่อ 1 เอกสาร และหัวข้อสถานที่พักมีคำเฉลี่ยอยู่ที่ 42.87 คำ ต่อ 1 เอกสาร และถ้าเรามีเอกสารที่มากกว่านี้ classifier ของเราก็จะสามารถให้ผลการทดลองที่ดีขึ้นกว่านี้

5.2 ข้อเสนอแนะ

5.2.1 ในการทำทดลองจำนวน data หรือเอกสารที่ใช้ในการเรียนรู้ระบบ หรือการ training ต้องมีปริมาณที่มากพอสมควร เพราะหากมีข้อมูลในปริมาณที่มาก การเรียนรู้ระบบ หรือการ training dataset จะทำให้ระบบเรียนรู้ได้กว้างขวางและควบคุมเนื้อหาในหัวข้อนั้นๆ ได้เป็นอย่างดี

5.2.2 จากการทดลองพบว่าจำนวนคำในเอกสารแต่ละหัวข้อ ส่งผลกับการเรียนรู้ระบบและการทดสอบระบบเช่นกัน ถ้ามีความแตกต่างของจำนวนคำในแต่ละหัวข้อมากเกินไป จะส่งผลกระทบกับการทดสอบเป็นอย่างมาก ดังนั้นในหัวข้อแต่ละหัวข้อควรมีจำนวนคำเฉลี่ยต่างกันไม่เกิน 100-200 คำ จะทำให้มีผลการทดลองที่ดีมากยิ่งขึ้น

5.2.3 คำหลัก หรือ keyword เป็นส่วนสำคัญมากในการทำการทดสอบระบบ เนื่องจากหาก keyword ค่อนข้างเป็นคำทั่วไป ก็จะทำให้การทดสอบมีข้อผิดพลาดเยอะขึ้น เนื้อหาที่ไม่เกี่ยวข้องกับหัวข้อนั้นๆ อาจจะมี keyword ปะปนอยู่ก็เป็นได้ อีกทั้งยังสามารถแก้ไข keyword ได้โดยที่แก้ไขตัว model ซึ่งจะไม่ส่งผลกระทบกับ framework ทั้งหมด

5.2.4 ในกรณีที่จำนวนคำในแต่ละหัวข้อมีความแตกต่างกันมาก จะส่งผลต่อค่า TF-IDF ที่ได้ ต้องแก้ปัญหาด้วยการใช้วิธีการทำ sampling เพื่อหยิบจำนวนคำออกมาในแต่ละหัวข้อให้ได้ในปริมาณที่ใกล้เคียงกัน แต่อย่างไรกว่าตามจำนวนคำของแต่ละหัวข้อก็ควรมากกว่า 2000 คำ ในการทำ sampling ซึ่งในการหาปริมาณคำด้วย sampling นี้จำเป็นจะต้องศึกษาต่อในอนาคต

5.3 ประโยชน์ที่ได้จากการทำสารนิพนธ์

5.3.1 ประโยชน์ที่ผู้ศึกษาได้รับจากการศึกษา คือ

(1) ได้เรียนรู้วิธีการในการศึกษาและค้นคว้าเอกสารและงานวิจัยต่างๆ ที่เกี่ยวข้องในการศึกษาค้นคว้าหาข้อมูลในเรื่องที่ทำการศึกษา

(2) ได้เรียนรู้ความหมายและความเป็นมาของการท่องเที่ยว และได้รับรู้ความต้องการของนักท่องเที่ยวที่ต้องการจากเว็บไซต์การท่องเที่ยว

(3) ได้เรียนรู้การทำเหมืองข้อความด้วยเทคนิคต่างๆ การนำผลที่ได้จากการวิเคราะห์ทางเหมืองข้อความมาคำนวณเพื่อนำไปใช้ในการพยากรณ์ข้อมูลใหม่ที่ยังไม่เคยเกิดขึ้นมาก่อน

(4) ได้เรียนรู้การพยากรณ์ความเสี่ยง และออกแบบการทำงานของการพยากรณ์ในหัวข้อการท่องเที่ยว

5.3.2 ประโยชน์ที่ผู้อื่นได้รับจากการศึกษาครั้งนี้ คือ

(1) ผู้จัดทำเว็บไซต์ใช้งานระบบแล้ว จะเป็นประโยชน์ในการสร้างเว็บไซต์ที่ให้ข้อมูลหรือเนื้อหาที่ถูกต้องตามความต้องการของนักท่องเที่ยว

(2) ผู้ที่สนใจหรือต้องการศึกษาต่อ สามารถนำกระบวนการความคิด หรือระบบไปต่อยอดได้ เพื่อเพิ่มประสิทธิภาพให้กับระบบมากยิ่งขึ้น และยังสามารถนำไปใช้กับข้อมูลประเภทที่แตกต่างกันออกไปได้อีกด้วย

บรรณานุกรม

บรรณานุกรม

[1] สำนักงานปลัดกระทรวงการท่องเที่ยวและกีฬา, “สถานการณ์การท่องเที่ยวของไทยที่สำคัญ” *รายงานภาวะเศรษฐกิจท่องเที่ยว*, ปีที่ 6, ฉบับที่ 7, หน้า 10-17, มกราคม-มีนาคม, 2560.

[2] Julia Monti, *“Big Cities, Big Business: Bangkok, London and Paris Lead the Way in Mastercard’s 2018 Global Destination Cities Index,”* [Online]. Available: https://newsroom.mastercard.com/press-releases/big-cities-big-business-bangkok-london-and-paris-lead-the-way-in-mastercards-2018-global-destination-cities-index/ [Accessed: October 4, 2018].

[3] กรมการท่องเที่ยว กระทรวงการท่องเที่ยวและกีฬา, *“การบริหารจัดการแหล่งท่องเที่ยว,”* กรุงเทพฯ: ม.ป.พ., 2558.

[4] Masoud Nosrati, “Python: An appropriate language for real world programming,” *World Applied Programming*, vol. 1, no. 2, pp. 110-117, June 2011.

[5] วรรณวิภา วงศ์วิไลสกุล, “เหมืองข้อความและการประยุกต์ใช้,” *วารสารปัญญาภิวัฒน์*, ปีที่ 4, ฉบับพิเศษ, หน้า 157-165, พฤษภาคม, 2556.

[6] ไกรศักดิ์ เกสร, “การค้นหาข้อมูลเชิงความหมาย: แนวคิดใหม่ของโปรแกรมการค้นหา (Search Engine) และแนวทางการพัฒนาในอนาคต,” *วารสารวไลยอลงกรณ์ปริทัศน์,* ปีที่ 1, ฉบับที่ 2, หน้า 1-16, กรกฎาคม-ธันวาคม, 2554.

[7] Wongkot Sriurai Phayung Meesad and Choochart Haruechaiyasak, “Enhance web page classification by using a topic model and integrating neighboring pages,” *The journal of KMUTNB,* vol. 20, no.2, pp. 204-214, May-August 2010.

[8] *“What does TF-IDF Mean?,”* [Online]. Available: http://www.tfidf.com/ [Accessed: August 6, 2018].

[9] *“Machine Learing in MATLAB,”* [Online]. Available: https://www.mathworks.com/help/stats/machine-learning-in-matlab.html [Accessed: August 20, 2018].

[10] จุฑาทิพย์ ทิพย์พูล และ นิเวศ จิระวิชิตชัย, “การจำแนกจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมโดยใช้เทคนิคเหมืองข้อมูล,” *วารสารวิทยาศาสตร์และเทคโนโลยี มทร. ธัญบุรี,* ปีที่ 6, ฉบับที่ 1, หน้า 102-109, เมษายน 2559.

[11] Dan Benyamin, *“A Gentle Introduction to Random Forests, Ensembles, and Performance Metrics in a Commercial System,”* [Online]. Available: http://blog.citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics [Accessed: November 5, 2018].

[12] Kevin Markham, *“Simple Guide to Confusion Matrix Terminology,”* [Online]. Available: https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/ [Accessed: October 5, 2018].

[13] Ashok Koujalagi, “Determine word relevance in document queries using TF–IDF,” *International Journal of Scientific Research,* vol. 4, no.8, pp. 284-285, August 2015.

[14] Rachsuda Jiamthapthaksin, “Thai text topic modeling system for discovering group interests of facebook young adult user,” *2016 2nd International Conference on Science in Information Technology, 2nd ICSITech,* Indonesia, October 26-27, 2016, pp. 91-96.

[15] เกรียงกมล คำมา และจักรกฤษณ์ เสน่ห์ นมะหุต, “ระบบคัดกรองเว็บไซต์อนาจารด้วยเทคนิคการวิเคราะห์เว็บไซต์,” *การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 9, ICT KMUTNB 9*, กรุงเทพฯ, 9-10 พฤษภาคม 2556, หน้า 315-321.

[16] ทิชากร เนตรสุวรรณ์, *“การวิเคราะห์ข่าวภาษาอังกฤษด้านนาชญากรรมออนไลน์ด้วยเทคนิคการทำเหมืองข้อความ,”* วิทยานิพนธ์ วท.ม (เทคโนโลยีสารสนเทศ), มหาวิทยาลัยนเรศวร, พิษณุโลก, 2559.

[17] คมคิด ชัชราภรณ์, ธรา อั่งสกุล และจิติมนต์ อั่งสกุล, “แบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้เทคนิคการเรียนรู้ของเครื่อง,” *Suranaree J. Sci*, ปีที่ 6, ฉบับที่ 2, หน้า 35-58, ธันวาคม, 2555.

[18] กฤษฎิ์ บรรณะชัยศิริสุข, *“เครื่องมือท่องเว็บไซต์บนอุปกรณ์เคลื่อนที่สำหรับผู้พิการทางสายตาโดยใช้การตรวจหาเนื้อหาหลัก,”* วิทยานิพนธ์ วท.ม (วิศวกรรมคอมพิวเตอร์), จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 2558.

[19] นัทธี ศรีหาจักษ์, *"กรอบงานสารสนเทศควบรวมสำหรับการค้นคืนเอกสารมีโครงสร้างในองค์กร,"* วิทยานิพนธ์ วท.ม (วิทยาศาสตร์คอมพิวเตอร์), จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 2556.

[20] Retno Kusumaningrum et al., "Classification of indonesian news articles based on latent dirichlet allocation," *2016 International Conference on Data and Software Engineering*, ICoDSE, Indonesia, October 26-27, 2016, pp. 35-39.

[21] ธันยพร กรานสุข และธนพล เจนสุทธิเวชกุล, "การสกัดปัจจัยต่อการผิดพลาดที่ก่อให้เกิดอุบัติเหตุในระบบรถไฟด้วยการทำเหมืองข้อความ," *The Fourteenth National Conference on Computing and Information Technology*, *14th NCCIT2018,* จังหวัดเชียงใหม่, 5-6 กรกฎาคม 2561, หน้า 426-431.

[22] จิรเมธ ว่องพรรณงาม, *"ระบบเตือนภัยผู้ขับรถจากความเมื่อยล้ากรณีสัปหงกโดยใช้ภาพความลึกจากกล้องคิเน็กต์,"* วิทยานิพนธ์ วศ.ม (วิทยาศาสตร์ไฟฟ้า), จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 2558.

[23] ชูพันธุ์ รัตนโภคา และเมธาวี สุทธิกุล, "ระบบจัดหมวดหมู่สถานที่ท่องเที่ยวด้วยคำอธิบายภาษาไทยแบบอัตโนมัติผ่านเว็บเซอร์วิส," *The Tenth National Conference on Computing and Information Technology, 10th NCCIT2014,* จังหวัดภูเก็ต, 8-9 พฤษภาคม 2557, หน้า 109-114.

ภาคผนวก

ภาคผนวก ก.

ทดสอบระบบกับหัวข้อข่าว

การทดสอบระบบกับหัวข้อข่าว

Dataset

ในการทำการทดสอบกับหัวข้อข่าว เนื่องจากผู้ศึกษาต้องการหัวข้อที่แตกต่างจากเดิม และยังสามารถมีหัวข้อได้หลากหลายหัวข้อ ผู้ศึกษาจึงได้เลือกหัวข้อข่าวที่มีข้อมูลในเว็บไซต์ค่อนข้างเยอะและมีความหลากหลายของประเภทข่าว โดยการทดลองในครั้งนี้ ผู้ศึกษาได้เลือกประเภทหัวข้อข่าวออกมา 4 ประเภท ได้แก้

1. ข่าวกีฬา
2. ข่าวอาชญากรรม
3. ข่าวเศรษฐกิจ
4. ข่าวเมือง

การออกแบบการทดลอง

การออกแบบการทดลองหัวข้อข่าว มีหัวข้ออยู่ทั้งหมด 4 หัวข้อ และแต่ละหัวข้อมีเอกสารอยู่ 60 เอกสาร รวมทั้งหมด 240 เอกสาร ซึ่งในการทดลองครั้งนี้ เลือกใช้การทำ 10 fold Cross-validation ซึ่งเป็นการทดลองที่ใช้สำหรับทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือมากขึ้น ซึ่งจะใช้วิธีการแบ่งข้อมูลออกเป็น 10 ชุดเท่าๆ กัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้

ในการทดลองแต่ละครั้งจะถูกแบ่งการทดลองออกเป็น 2 รูปแบบ คือ

1. การทดลองโดยการทดสอบแบบ single topic

   ในแต่ละเอกสาร มีเพียง 1 หัวข้อ เพื่อเป็นการทดสอบประสิทธิภาพ classifier ของแต่ละหัวข้อ หากถ้ามี 4 classifier ก็จะถูกทำการทดลองแบบเดียวกันในทุก ๆ ข้อมูล

2. การทดลองโดยการทดสอบแบบ multiple topics

   ใน 1 เอกสาร มี 2 หัวข้อ, 3 หัวข้อ และ 4 หัวข้อ

   กรณีที่เป็นไปได้แบบ 2 หัวข้อ

2 หัวข้อ	กรณีเป็นไปได้ที่ 1	กรณีเป็นไปได้ที่ 2	กรณีเป็นไปได้ที่ 3	กรณีเป็นไปได้ที่ 4	กรณีเป็นไปได้ที่ 5	กรณีเป็นไปได้ที่ 6
หัวข้อที่ 1	√	√	√			
หัวข้อที่ 2	√			√	√	
หัวข้อที่ 3		√		√		√
หัวข้อที่ 4			√		√	√

กรณีที่เป็นไปได้แบบ 3 หัวข้อ

3 หัวข้อ	กรณีเป็นไปได้ที่ 1	กรณีเป็นไปได้ที่ 2	กรณีเป็นไปได้ที่ 3	กรณีเป็นไปได้ที่ 4
หัวข้อที่ 1	√	√	√	
หัวข้อที่ 2	√	√		√
หัวข้อที่ 3	√		√	√
หัวข้อที่ 4		√	√	√

กรณีที่เป็นไปได้แบบ 4 หัวข้อ

4 หัวข้อ	กรณีเป็นไปได้ที่ 1
หัวข้อที่ 1	√
หัวข้อที่ 2	√
หัวข้อที่ 3	√
หัวข้อที่ 4	√

การทำการทดลอง

1. การทดลองประสิทธิภาพของ classifier แบบ single topic

<u>การทดลองที่ 1</u>

การทดลองประสิทธิภาพของ classifier แบบ 1 หัวข้อ จะแบ่งข้อมูลออกเป็น 10 ชุดข้อมูล ตามขั้นตอนเดิมที่เคยทำการทดลองกับหัวข้อการท่องเที่ยว จากนั้นนำชุดข้อมูลทั้ง 10 ชุด ไปทดสอบกับระบบ และนำผลลัพธ์มาแสดงในรูปแบบ confusion matrix ดังตารางที่ ก.1 ดังนี้

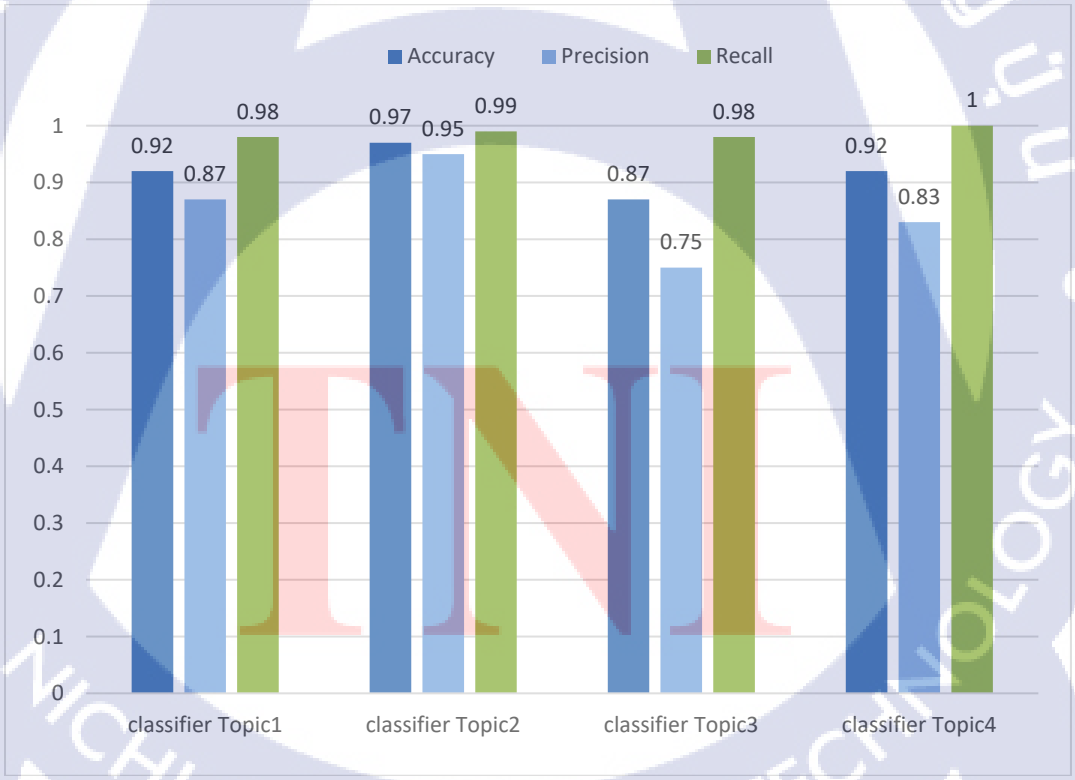
ตารางที่ ก.1 แสดงค่า confusion matrix ของการทดลอง 10 fold Cross-validation

	True Negatives	False Negatives	False Positives	True Positives
การทดสอบ 1	24	0	3	21
การทดสอบ 2	24	0	4	20
การทดสอบ 3	23	1	1	23
การทดสอบ 4	24	0	5	19
การทดสอบ 5	23	1	3	21

ตารางที่ ก.1 ตารางแสดงค่า confusion matrix ของการทดลอง 10 fold Cross-validation (ต่อ)

	True Negatives	False Negatives	False Positives	True Positives
การทดสอบ 6	24	0	3	21
การทดสอบ 7	24	0	5	19
การทดสอบ 8	22	2	3	21
การทดสอบ 9	24	0	3	21
การทดสอบ10	24	0	4	20

จากรูปที่ ก.1 แสดงถึงประสิทธิภาพโดยรวมของ classifier แต่ละหัวข้อ โดยแสดงค่าเฉลี่ยของ Accuracy, Precision และ Recall โดยที่แทบ error bar คือค่าบวกลบของค่า SD หรือ ค่าเบี่ยงเบนมาตรฐาน



รูปที่ ก.1 แผนภูมิค่าเฉลี่ย Accuracy, Precision, Recall

2. การทดลองประสิทธิภาพของ classifier แบบ multiple topics

<u>การทดลองที่ 1</u>

การทดลองประสิทธิภาพของ classifier แบบ 2 หัวข้อ

ในการการทดลองนี้จะนำเนื้อหาของ 2 หัวข้อมารวมกัน โดยมีเนื้อหาจากหัวข้อแรกทั้งหมด 6 เอกสาร และหัวข้อที่ 2 อีก 6 เอกสาร นำเอกสารมารวมกัน แล้วหาค่า TF-IDF ของทั้ง 6 เอกสาร หลังจากนั้นก็นำทั้ง 6 เอกสาร ไปสร้างค่า feature vector ในแต่ละชุดข้อมูลมา test ใน classification ซึ่งผลลัพธ์ที่เกิดขึ้นจะต้องเป็น yes เท่านั้น

เมื่อได้ผลลัพธ์ออกมาแล้ว นำผลลัพธ์มาแสดงค่าในรูปแบบ confusion matrix ซึ่งผลจะเป็นดังที่แสดงในตารางที่ ก.2 และจะมี 2 ค่า คือ True Positives และ False Negatives ที่เป็นผลลัพธ์มาจากการทดลอง

ตารางที่ ก.2 แสดงค่า True Positives และ False Negatives ของการทดลอง 10 fold Cross-validation

	หัวข้อที่ 1,2		หัวข้อ 1,3		หัวข้อ 1,4		หัวข้อ 2,3		หัวข้อ 2,4		หัวข้อ 3,4	
	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP
การทดสอบ 1	1	5	0	6	2	4	0	6	2	4	0	6
การทดสอบ 2	1	5	3	3	0	6	3	3	0	6	3	3
การทดสอบ 3	0	6	1	5	1	5	1	5	0	6	1	5
การทดสอบ 4	1	5	1	5	4	2	2	4	4	2	3	3
การทดสอบ 5	1	5	1	5	2	4	2	4	1	5	2	4
การทดสอบ 6	1	5	2	4	2	4	2	4	3	3	1	5
การทดสอบ 7	3	3	3	3	5	1	4	2	4	2	2	4
การทดสอบ 8	1	5	1	5	4	2	1	5	4	2	1	5
การทดสอบ 9	1	5	4	2	1	5	2	4	1	5	2	4
การทดสอบ10	3	3	3	3	2	4	1	5	1	5	1	5

เนื่องจากการทดสอบในครั้งนี้เป็นการ test ข้อมูล ไม่จำเป็นต้องมีการ train ดังนั้นค่าที่สามารถนำมาใช้ได้ จะมีเพียงแค่ค่า Recall ซึ่งจากตารางที่ ก.3 จะแสดงค่า Recall ที่เกิดจากค่า True Positives หารด้วย ค่า Actual Yes

ตารางที่ ก.3 แสดงค่า Recall ของการรวมกันของ 2 หัวข้อ

	หัวข้อที่ 1,2	หัวข้อที่ 1,3	หัวข้อที่ 1,4	หัวข้อที่ 2,3	หัวข้อที่ 2,4	หัวข้อที่ 3,4
การทดสอบ 1	0.83	1	0.67	1	0.67	1
การทดสอบ 2	0.83	0.5	1	0.5	1	0.5
การทดสอบ 3	1	0.83	0.83	0.83	1	0.83
การทดสอบ 4	0.83	0.83	0.33	0.67	0.33	0.5
การทดสอบ 5	0.83	0.83	0.67	0.67	0.83	0.67
การทดสอบ 6	0.83	0.67	0.67	0.67	0.5	0.83
การทดสอบ 7	0.5	0.5	0.33	0.33	0.33	0.67
การทดสอบ 8	0.83	0.83	0.33	0.83	0.33	0.83
การทดสอบ 9	0.83	0.33	0.83	0.67	0.83	0.67
การทดสอบ10	0.5	0.5	0.67	0.83	0.83	0.83

จากรูปที่ ก.2 จะแสดงค่าเฉลี่ย Recall จากข้อมูลการทดสอบทั้ง 10 การทดสอบ โดยที่แสดงค่าแยกออกเป็นทีละการรวมกันของเอกสารแต่ละหัวข้อ

รูปที่ ก.2 แผนภูมิ Recall ของการรวมตัวของ 2 หัวข้อ

<u>การทดลองที่ 2</u>

การทดลองประสิทธิภาพของ classifier แบบ 3 หัวข้อ

ในการการทดลองนี้จะนำเนื้อหาของ 3 หัวข้อมารวมกัน โดยมีเนื้อหาจากหัวข้อแรกทั้งหมด 6 เอกสาร และอีก 2 หัวข้อ อีก 6 เอกสาร นำเอกสารมารวมกัน แล้วหาค่า tf-idf ของทั้ง 6 เอกสาร หลังจากนั้นก็นำทั้ง 6 เอกสาร ไปสร้างค่า feature vector ในแต่ละชุดข้อมูลมา test ใน classification ซึ่งผลลัพธ์ที่เกิดขึ้นจะต้องเป็น yes เท่านั้น

เมื่อได้ผลลัพธ์ออกมาแล้ว นำผลลัพธ์มาแสดงค่าในรูปแบบ confusion matrix ซึ่งผลจะเป็นดังที่แสดงในตารางที่ 4 และจะมี 2 ค่า คือ True Positives และ False Negatives ที่เป็นผลลัพธ์มาจากการทดลอง

ตารางที่ ก.4 แสดงค่า True Positives และ False Negatives ของการทดลอง 10 fold Cross-validation

	หัวข้อ 1,2,3		หัวข้อ 1,2,4		หัวข้อ 1,3,4		หัวข้อ 2,3,4	
	FN	TP	FN	TP	FN	TP	FN	TP
การทดสอบ 1	0	6	3	3	0	6	0	6
การทดสอบ 2	3	3	0	6	3	3	3	3

ตารางที่ ก.4 แสดงค่า True Positives และ False Negatives ของการทดลอง 10 fold Cross-validation (ต่อ)

	หัวข้อ 1,2,3		หัวข้อ 1,2,4		หัวข้อ 1,3,4		หัวข้อ 2,3,4	
	FN	TP	FN	TP	FN	TP	FN	TP
การทดสอบ 3	1	5	1	5	0	6	1	5
การทดสอบ 4	2	4	4	2	4	2	3	3
การทดสอบ 5	1	5	2	4	2	4	2	4
การทดสอบ 6	2	4	2	4	1	5	3	3
การทดสอบ 7	4	2	6	0	4	2	2	4
การทดสอบ 8	1	5	4	2	1	5	1	5
การทดสอบ 9	3	3	1	5	3	3	1	5
การทดสอบ10	3	3	3	3	3	3	1	5

เนื่องจากการทดสอบในครั้งนี้เป็นการ test ข้อมูล ไม่จำเป็นต้องมีการ train ดังนั้นค่าที่สามารถนำมาใช้ได้ จะมีเพียงแค่ค่า Recall ซึ่งจากตารางที่ ก.5 จะแสดงค่า Recall ที่เกิดจากค่า True Positives หารด้วย ค่า Actual Yes

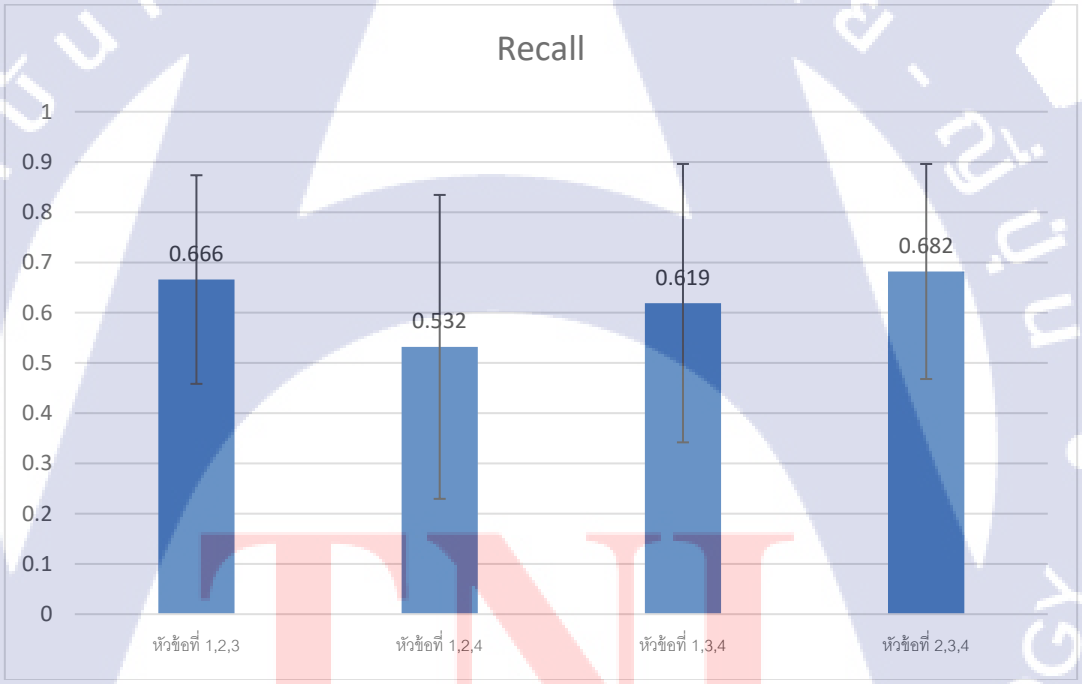
ตารางที่ ก.5 แสดงค่า Recall ของการรวมกันของ 3 หัวข้อ

	หัวข้อที่ 1,2,3	หัวข้อที่ 1,2,4	หัวข้อที่ 1,3,4	หัวข้อที่ 2,3,4
การทดสอบ 1	1	0.5	1	1
การทดสอบ 2	0.5	1	0.5	0.5
การทดสอบ 3	0.83	0.83	1	0.83
การทดสอบ 4	0.67	0.33	0.33	0.5
การทดสอบ 5	0.83	0.33	0.33	0.33
การทดสอบ 6	0.67	0.67	0.87	0.5

ตารางที่ ก.5 แสดงค่า Recall ของการรวมกันของ 3 หัวข้อ (ต่อ)

	หัวข้อที่ 1,2,3	หัวข้อที่ 1,2,4	หัวข้อที่ 1,3,4	หัวข้อที่ 2,3,4
การทดสอบ 7	0.33	0	0.33	0.67
การทดสอบ 8	0.83	0.33	0.83	0.83
การทดสอบ 9	0.5	0.83	0.5	0.83
การทดสอบ10	0.5	0.5	0.5	0.83

จากรูปที่ ก.3 จะแสดงค่าเฉลี่ย Recall จากข้อมูลการทดสอบทั้ง 10 การทดสอบ โดยที่แสดงค่าแยกออกเป็นทีละการรวมกันของเอกสารแต่ละหัวข้อ



รูปที่ ก.3 แผนภูมิ Recall ของการรวมตัวของ 3 หัวข้อ

<u>การทดลองที่ 3</u>

การทดลองประสิทธิภาพของ classifier แบบ 4 หัวข้อ

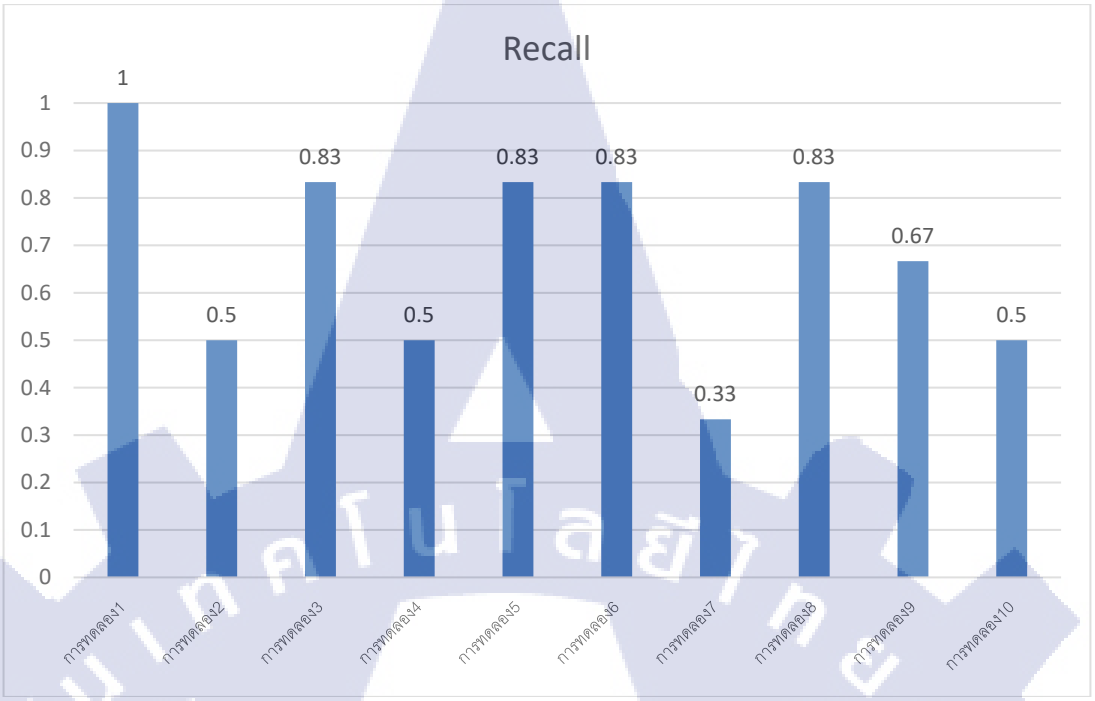
การทดลองประสิทธิภาพของ classifier แบบ 4 หัวข้อ หลังจากทำการทดลองข้อมูลกับระบบแล้ว ผลลัพธ์ที่ได้ออกมา จะนำมาแสดงค่าในรูปแบบ confusion matrix ซึ่งผลจะเป็นดังที่

แสดงในตารางที่ ก.6 และจะมี 2 ค่า คือ True Positives และ False Negatives ที่เป็นผลลัพธ์มาจากการทดลอง

ตารางที่ ก.6 แสดงค่า confusion matrix ของการทดลอง 10 fold Cross-validation

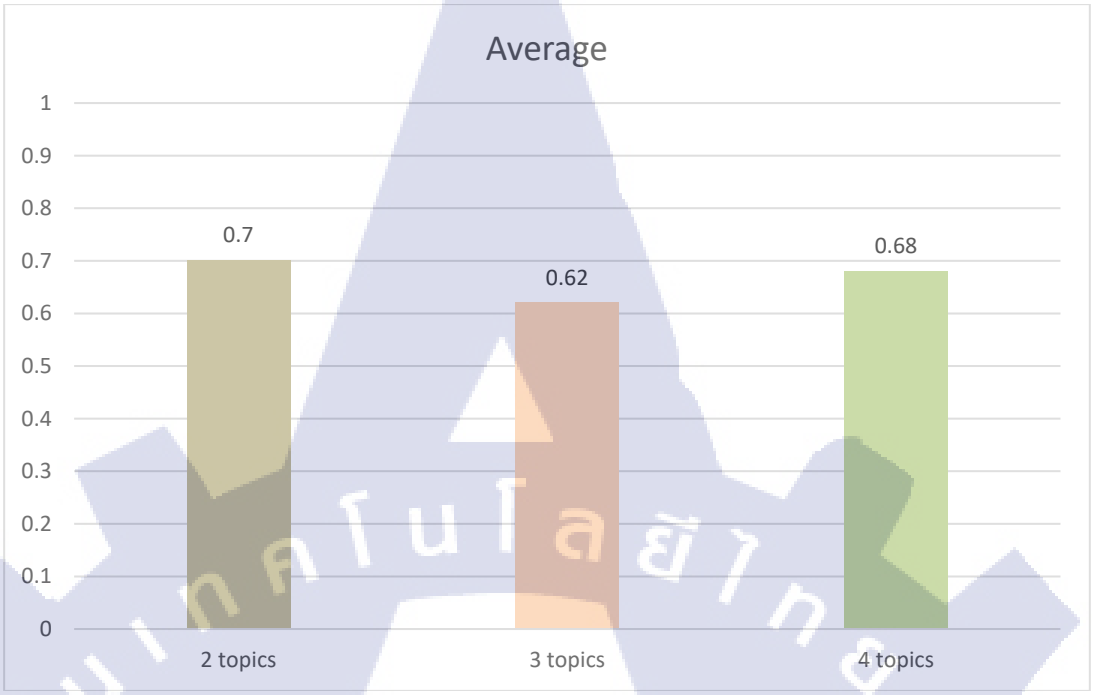
	True Negatives	False Negatives	False Positives	True Positives
การทดสอบ 1	0	0	0	24
การทดสอบ 2	0	0	3	21
การทดสอบ 3	0	0	1	23
การทดสอบ 4	0	0	4	20
การทดสอบ 5	0	0	2	22
การทดสอบ 6	0	0	1	23
การทดสอบ 7	0	0	5	19
การทดสอบ 8	0	0	1	23
การทดสอบ 9	0	0	2	22
การทดสอบ10	0	0	5	19

จากรูปที่ ก.4 จะแสดงค่าเฉลี่ย Recall จากการรวมกันของเอกสารทั้ง 4 หัวข้อ ทั้ง 10 การทดสอบ โดยที่แสดงค่าแยกออกเป็นทีละการรวมกันของเอกสารแต่ละหัวข้อ ดังนี้



รูปที่ ก.4 แผนภูมิ Recall ของการรวมตัวของ 4 หัวข้อ

จากการทดสอบประสิทธิภาพ classifier แบบ multiple topics จะเห็นได้ว่าการทดสอบประสิทธิภาพจะสามารถแสดงได้เพียงแค่ค่า Recall เนื่องจากการทดสอบประสิทธิภาพ classifier แบบ multiple topics จะมีการ train ข้อมูลเฉพาะแค่ yes เพียงเท่านั้น จากรูปที่ ก.5 จะแสดงค่าเฉลี่ยของค่า Recall จากการรวมตัวของเอกสารทั้ง 2, 3 และ 4 หัวข้อ



รูปที่ ก.5 แผนภูมิค่าเฉลี่ย Recall ของการรวมกันของเอกสารทุกหัวข้อ

สรุปผลการทดลอง

ในการสรุปผลการศึกษาที่ผู้ศึกษาได้ศึกษาในหัวข้อข่าว โดยมีหัวข้อหรือประเภทของข่าวทั้งหมด 4 หัวข้อ ในงานครั้งนี้ ได้แก่ หัวข้อข่าวกีฬา, หัวข้อข่าวอาชญากรรม, หัวข้อข่าวเศรษฐกิจ และหัวข้อข่าวการเมือง จึงได้จัดทำ binary classifier ทั้งหมด 4 ตัวตามหัวข้อที่กล่าวมาข้างต้น ซึ่ง classifier ทั้ง 4 ตัวนี้จะเป็นตัวช่วยในการแยกแยะหัวข้อ และยังค่อนข้างมีประสิทธิภาพในการแยกแยะได้ดีกับการทดลองในครั้งนี้อีกด้วย โดยผลการทดลองจะแบ่งออกเป็น 2 รูปแบบ คือ แบบ single topic และแบบ multiple topics ดังนี้

5.1.1 Single Topic

การทำการทดลองแบบ Single Topic ทำเพื่อตรวจสอบประสิทธิภาพของ classifier ในการแยกแยะหัวข้อของตัวเอง ซึ่งผลการทดลองโดยจะถูกแบ่งออกเป็น 4 โดยหัวข้อแรกข่าวกีฬาได้ค่า Accuracy เท่ากับ 0.92 ค่า Precision เท่ากับ 0.87 และค่า Recall เท่ากับ 0.98 หัวข้อต่อมาหัวข้อข่าวอาชญากรรมได้ค่า Accuracy เท่ากับ 0.97 ค่า Precision เท่ากับ 0.95 และค่า Recall เท่ากับ 0.99 หัวข้อต่อมาหัวข้อข่าวเศรษฐกิจได้ค่า Accuracy เท่ากับ 0.87 ค่า Precision เท่ากับ 0.75 และค่า Recall เท่ากับ 0.98 และหัวข้อสุดท้ายหัวข้อข่าวการเมืองได้ค่า Accuracy เท่ากับ 0.92 ค่า Precision เท่ากับ 0.83 และค่า Recall เท่ากับ 1 ซึ่งแสดงให้เห็นว่าความสามารถของ classifier แต่ละตัวนี้ในการจดจำ และในการคาดเดาผลลัพธ์ของตัวเองนั้นทำงานได้ค่อนข้างเป็นอย่างดี

5.1.2 Multiple Topics

การทำการทดลองแบบ Multiple Topics ทำเพื่อตรวจสอบประสิทธิภาพของ classifier ว่าทำงานได้จริง ในกรณีที่มีเนื้อหามากกว่า 1 หัวข้อ ซึ่งในตามความเป็นจริง หน้าเว็บไซต์จะมีข้อมูลที่รวมกันอยู่หลากหลายหัวข้อเป็นปกติ โดยผลการทดลองแบบ multiple topics จะแบ่งออกเป็น 3 อย่าง ดังนี้

- การรวมตัวกันของ 2 topic จะมีค่าเฉลี่ย Recall จากการทดสอบ 10 fold Cross-validation อยู่ที่ 0.70
- การรวมตัวกันของ 3 topic จะมีค่าเฉลี่ย Recall จากการทดสอบ 10 fold Cross-validation อยู่ที่ 0.62
- การรวมตัวกันของ 4 topic จะมีค่าเฉลี่ย Recall จากการทดสอบ 10 fold Cross-validation อยู่ที่ 0.68

ซึ่งผลการทดลองนี้ได้ค่า Recall ปานกลาง แสดงให้เห็นว่าเป็นการทดลองที่มีประสิทธิภาพในการแยกแยะการตรวจสอบความครบถ้วนของข้อมูลในแต่ละเอกสารในระดับปานกลาง หากแต่ใน dataset ครั้งนี้เว็บเพจในแต่ละเว็บไซต์ค่อนข้างใช้คำในการเขียนข่าวที่หลากหลาย จึงทำให้บางเอกสารไม่พบ keywords ในหัวข้อนั้น ๆ ส่งผลให้ผลการทดลองไม่สูงมากนัก ในอนาคตสามารถปรับเปลี่ยนโดยการหาเอกสารเพิ่มมากขึ้น หรือปรับเปลี่ยนการหาคำ keywords เพื่อให้ประสิทธิภาพในการแยกแยะมีประสิทธิภาพที่ดียิ่งขึ้นไปอีก

ภาคผนวก ข.

ทดสอบระบบกับหัวข้อการท่องเที่ยวเชิงนิเวศ

การทดสอบระบบกับหัวข้อการท่องเที่ยวเชิงนิเวศ

Dataset

ในการทำการทดสอบกับหัวข้อการท่องเที่ยวเชิงนิเวศ ผู้ศึกษาได้ศึกษาแหล่งท่องเที่ยวเชิงนิเวศในประเทศไทย จึงนำหัวข้อการท่องเที่ยวเชิงนิเวศมามาทดสอบกับระบบ ซึ่งได้ทำการเลือกหัวข้อจากมาตรฐานคุณภาพแหล่งท่องเที่ยวเชิงนิเวศ โดยพัฒนาจากสำนักพัฒนาแหล่งท่องเที่ยว กรมการท่องเที่ยว กระทรวงการท่องเที่ยวและกีฬา ประกอบด้วย 4 หัวข้อ

1. ทรัพยากรธรรมชาติ และสิ่งแวดล้อม ของสถานที่ท่องเที่ยว
2. กิจกรรมของการท่องเที่ยวเชิงนิเวศ เช่น การเดินป่า ดำน้ำ ปีนเขา ส่องสัตว์
3. การให้ข้อมูลเกี่ยวกับแหล่งท่องเที่ยว เช่น ช่วงเวลา สภาพอากาศ วิธีการท่องเที่ยวที่ถูกต้อง หรือการปลูกจิตสำนึกให้แก่นักท่องเที่ยว
4. การมีส่วนร่วมของชุมชน

การออกแบบการทดลอง

การออกแบบการทดลองหัวข้อการท่องเที่ยวเชิงนิเวศ มีหัวข้ออยู่ทั้งหมด 4 หัวข้อ โดยที่มีหัวข้อที่ 1 และ 2 อยู่หัวข้อละ 30 เอกสาร และหัวข้อที่ 3 และ 4 หัวข้อละ 15 เอกสาร รวมทั้งหมด 90 เอกสาร ซึ่งในการทดลองครั้งนี้ เลือกใช้การทำ 5 fold Cross-validation ซึ่งเป็นการทดลองที่ใช้สำหรับทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือมากขึ้น ซึ่งจะใช้วิธีการแบ่งข้อมูลออกเป็น 5 ชุดเท่าๆ กัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ ในการทดลองแต่ละครั้งจะถูกแบ่งการทดลองออกเป็น 2 รูปแบบ คือ

1. การทดลองโดยการทดสอบแบบ single topic
   ในแต่ละเอกสาร มีเพียง 1 หัวข้อ เพื่อเป็นการทดสอบประสิทธิภาพ classifier ของแต่ละหัวข้อ หากถ้ามี 4 classifier ก็จะถูกทำการทดลองแบบเดียวกันในทุก ๆ ข้อมูล
2. การทดลองโดยการทดสอบแบบ multiple topics
   ใน 1 เอกสาร มี 2 หัวข้อ, 3 หัวข้อ และ 4 หัวข้อ

กรณีที่เป็นไปได้แบบ 2 หัวข้อ

2 หัวข้อ	กรณีเป็นไปได้ที่ 1	กรณีเป็นไปได้ที่ 2	กรณีเป็นไปได้ที่ 3	กรณีเป็นไปได้ที่ 4	กรณีเป็นไปได้ที่ 5	กรณีเป็นไปได้ที่ 6
หัวข้อที่ 1	√	√	√			

หัวข้อที่ 2	√			√	√	
หัวข้อที่ 3		√		√		√
หัวข้อที่ 4			√		√	√

กรณีที่เป็นไปได้แบบ 3 หัวข้อ

3 หัวข้อ	กรณีเป็นไปได้ที่ 1	กรณีเป็นไปได้ที่ 2	กรณีเป็นไปได้ที่ 3	กรณีเป็นไปได้ที่ 4
หัวข้อที่ 1	√	√	√	
หัวข้อที่ 2	√	√		√
หัวข้อที่ 3	√		√	√
หัวข้อที่ 4		√	√	√

กรณีที่เป็นไปได้แบบ 4 หัวข้อ

4 หัวข้อ	กรณีเป็นไปได้ที่ 1
หัวข้อที่ 1	√
หัวข้อที่ 2	√
หัวข้อที่ 3	√
หัวข้อที่ 4	√

การทำการทดลอง

1. การทดลองประสิทธิภาพของ classifier แบบ single topic

<u>การทดลองที่ 1</u>

การทดลองประสิทธิภาพของ classifier แบบ 1 หัวข้อ จะแบ่งข้อมูลออกเป็น 10 ชุดข้อมูล ตามขั้นตอนเดิมที่เคยทำการทดลองกับหัวข้อการท่องเที่ยว จากนั้นนำชุดข้อมูลทั้ง 10 ชุด ไปทดสอบกับระบบ และนำผลลัพธ์มาแสดงในรูปแบบ confusion matrix ดังตารางที่ ข.1 ดังนี้

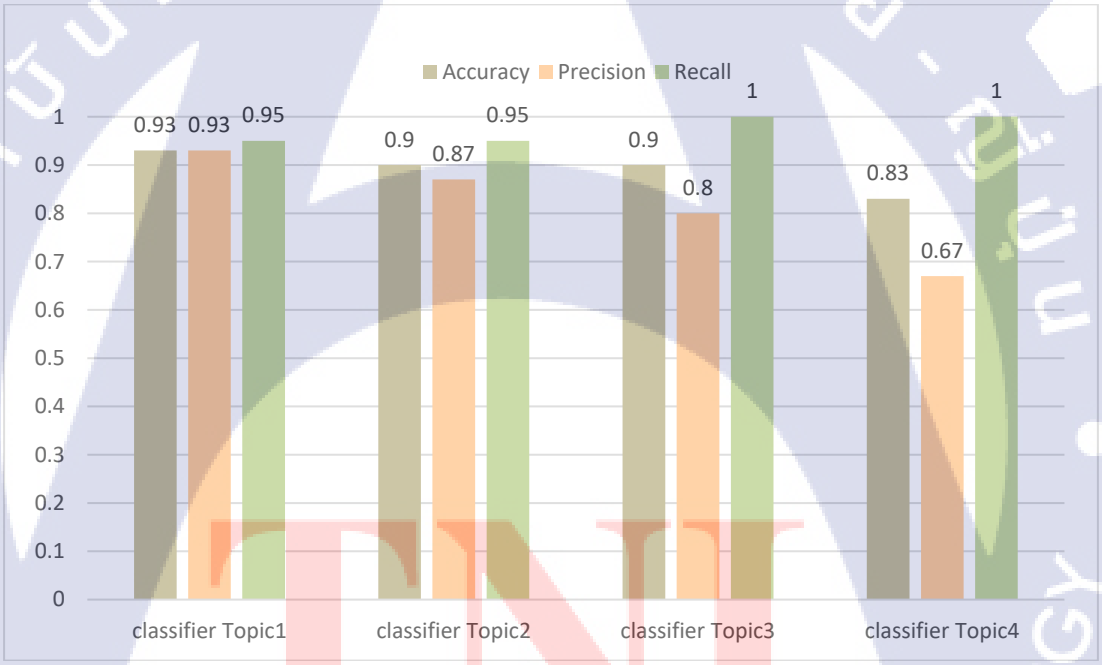
ตารางที่ ข.1 แสดงค่าconfusion matrix ของการทดลอง 5 fold Cross-validation

	True Negatives	False Negatives	False Positives	True Positives
การทดสอบ 1	11	1	0	12
การทดสอบ 2	12	0	3	9

ตารางที่ ข.1 confusion matrix ของการทดลอง 5 fold Cross-validation (ต่อ)

	True Negatives	False Negatives	False Positives	True Positives
การทดสอบ 3	12	0	1	11
การทดสอบ 4	11	1	3	9
การทดสอบ 5	12	0	4	8

จากรูปที่ ข.1 แสดงถึงประสิทธิภาพโดยรวมของ classifier แต่ละหัวข้อ โดยแสดงค่าเฉลี่ยของ Accuracy, Precision และ Recall โดยที่แทบ error bar คือค่าบวกลบของค่า SD หรือ ค่าเบี่ยงเบนมาตรฐาน



รูปที่ ข.1 แผนภูมิค่าเฉลี่ย Accuracy, Precision, Recall

2. การทดลองประสิทธิภาพของ classifier แบบ multiple topics

<u>การทดลองที่ 1</u>

การทดลองประสิทธิภาพของ classifier แบบ 2 หัวข้อ

ในการการทดลองนี้จะนำเนื้อหาของ 2 หัวข้อมารวมกัน โดยมีเนื้อหาจากหัวข้อแรกทั้งหมด 3 เอกสาร และหัวข้อที่ 2 อีก 3 เอกสาร นำเอกสารมารวมกัน แล้วหาค่า TF-IDF ของทั้ง 3 เอกสาร

หลังจากนั้นก็นำทั้ง 3 เอกสาร ไปสร้างค่า feature vector ในแต่ละชุดข้อมูลมา test ใน classification ซึ่งผลลัพธ์ที่เกิดขึ้นจะต้องเป็น yes เท่านั้น

เมื่อได้ผลลัพธ์ออกมาแล้ว นำผลลัพธ์มาแสดงค่าในรูปแบบ confusion matrix ซึ่งผลจะเป็นดังที่แสดงในตารางที่ ข.2 และจะมี 2 ค่า คือ True Positives และ False Negatives ที่เป็นผลลัพธ์มาจากการทดลอง

ตารางที่ ข.2 แสดงค่า True Positives และ False Negatives ของการทดลอง 5 fold Cross-validation

	หัวข้อที่ 1,2		หัวข้อ 1,3		หัวข้อ 1,4		หัวข้อ 2,3		หัวข้อ 2,4		หัวข้อ 3,4	
	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP
การทดสอบ 1	0	3	2	1	1	2	1	2	1	2	2	1
การทดสอบ 2	1	2	2	1	2	1	1	2	0	3	1	2
การทดสอบ 3	1	2	2	1	3	0	1	2	2	1	1	2
การทดสอบ 4	1	2	1	2	2	1	1	2	1	2	2	1
การทดสอบ 5	2	1	2	1	2	1	2	1	1	2	2	1

เนื่องจากการทดสอบในครั้งนี้เป็นการ test ข้อมูล ไม่จำเป็นต้องมีการ train ดังนั้นค่าที่สามารถนำมาใช้ได้ จะมีเพียงแค่ค่า Recall ซึ่งจากตารางที่ ข.3 จะแสดงค่า Recall ที่เกิดจากค่า True Positives หารด้วย ค่า Actual Yes

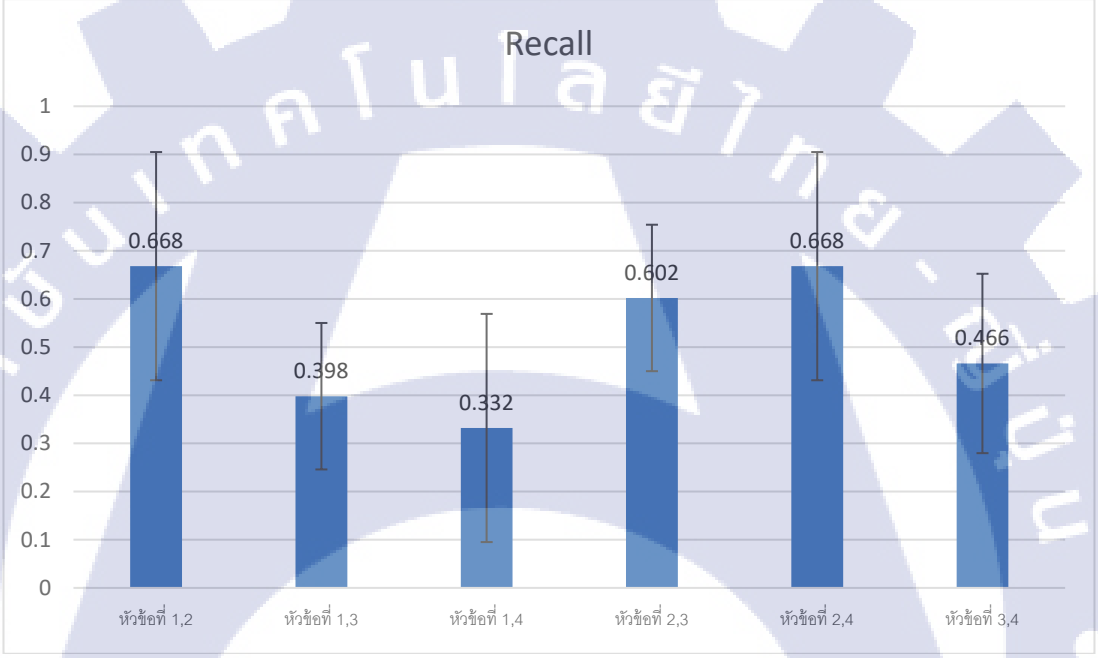
ตารางที่ ข.3 แสดงค่า Recall ของการรวมกันของ 2 หัวข้อ

	หัวข้อที่ 1,2	หัวข้อที่ 1,3	หัวข้อที่ 1,4	หัวข้อที่ 2,3	หัวข้อที่ 2,4	หัวข้อที่ 3,4
การทดสอบ 1	1	0.33	0.67	0.67	0.67	0.33
การทดสอบ 2	0.67	0.33	0.33	0.67	1	0.67
การทดสอบ 3	0.67	0.33	0	0.67	0.33	0.67
การทดสอบ 4	0.67	0.67	0.33	0.67	0.67	0.33

ตารางที่ ข.3 แสดงค่า Recall ของการรวมกันของ 2 หัวข้อ (ต่อ)

	หัวข้อที่ 1,2	หัวข้อที่ 1,3	หัวข้อที่ 1,4	หัวข้อที่ 2,3	หัวข้อที่ 2,4	หัวข้อที่ 3,4
การทดสอบ 5	0.33	0.33	0.33	0.33	0.67	0.33

จากรูปที่ ข.2 จะแสดงค่าเฉลี่ย Recall จากข้อมูลการทดสอบทั้ง 5 การทดสอบ โดยที่แสดงค่าแยกออกเป็นทีละการรวมกันของเอกสารแต่ละหัวข้อ



รูปที่ ข.2 แผนภูมิ Recall ของการรวมตัวของ 2 หัวข้อ

<u>การทดลองที่ 2</u>

การทดลองประสิทธิภาพของ classifier แบบ 3 หัวข้อ

ในการการทดลองนี้จะนำเนื้อหาของ 3 หัวข้อมารวมกัน โดยมีเนื้อหาจากหัวข้อแรกทั้งหมด 6 เอกสาร และอีก 2 หัวข้อ อีก 6 เอกสาร นำเอกสารมารวมกัน แล้วหาค่า TF-IDF ของทั้ง 6 เอกสาร หลังจากนั้นก็นำทั้ง 6 เอกสาร ไปสร้างค่า feature vector ในแต่ละชุดข้อมูลมา test ใน classification ซึ่งผลลัพธ์ที่เกิดขึ้นจะต้องเป็น yes เท่านั้น

เมื่อได้ผลลัพธ์ออกมาแล้ว นำผลลัพธ์มาแสดงค่าในรูปแบบ confusion matrix ซึ่งผลจะเป็นดังที่แสดงในตารางที่ ข.4 และจะมี 2 ค่า คือ True Positives และ False Negatives ที่เป็นผลลัพธ์มาจากการทดลอง

ตารางที่ ข.4 แสดงค่า True Positives และ False Negatives ของการทดลอง 5 fold Cross-validation

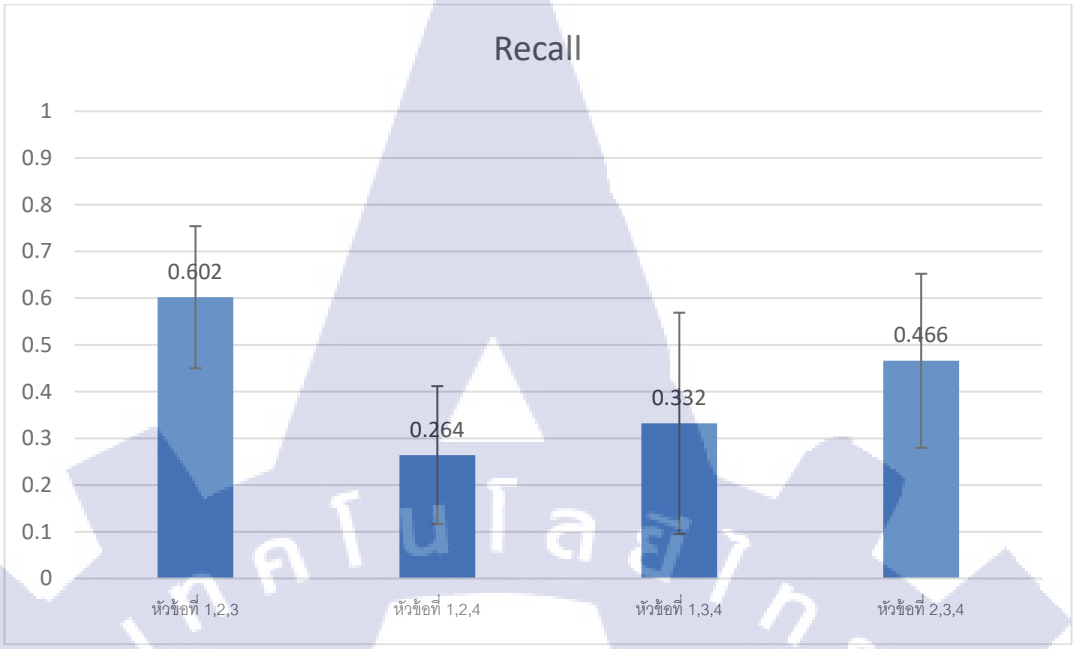
	หัวข้อ 1,2,3		หัวข้อ 1,2,4		หัวข้อ 1,3,4		หัวข้อ 2,3,4	
	FN	TP	FN	TP	FN	TP	FN	TP
การทดสอบ 1	2	1	2	1	2	1	1	2
การทดสอบ 2	1	2	2	1	1	2	1	2
การทดสอบ 3	1	2	3	0	3	0	2	1
การทดสอบ 4	1	2	2	1	2	1	2	1
การทดสอบ 5	1	2	2	1	2	1	2	1

เนื่องจากการทดสอบในครั้งนี้เป็นการ test ข้อมูล ไม่จำเป็นต้องมีการ train ดังนั้นค่าที่สามารถนำมาใช้ได้ จะมีเพียงแค่ค่า Recall ซึ่งจากตารางที่ ข.5 จะแสดงค่า Recall ที่เกิดจากค่า True Positives หารด้วย ค่า Actual Yes

ตารางที่ ข.5 แสดงค่า Recall ของการรวมกันของ 3 หัวข้อ

	หัวข้อที่ 1,2,3	หัวข้อที่ 1,2,4	หัวข้อที่ 1,3,4	หัวข้อที่ 2,3,4
การทดสอบ 1	0.33	0.33	0.33	0.67
การทดสอบ 2	0.67	0.33	0.67	0.67
การทดสอบ 3	0.67	0	0	0.33
การทดสอบ 4	0.67	0.33	0.33	0.33
การทดสอล 5	0.67	0.33	0.33	0.33

จากรูปที่ ข.3 จะแสดงค่าเฉลี่ย Recall จากข้อมูลการทดสอบทั้ง 5 การทดสอบ โดยที่แสดงค่าแยกออกเป็นทีละการรวมกันของเอกสารแต่ละหัวข้อ



รูปที่ ข.3 แผนภูมิ Recall ของการรวมตัวของ 3 หัวข้อ

การทดลองที่ 3

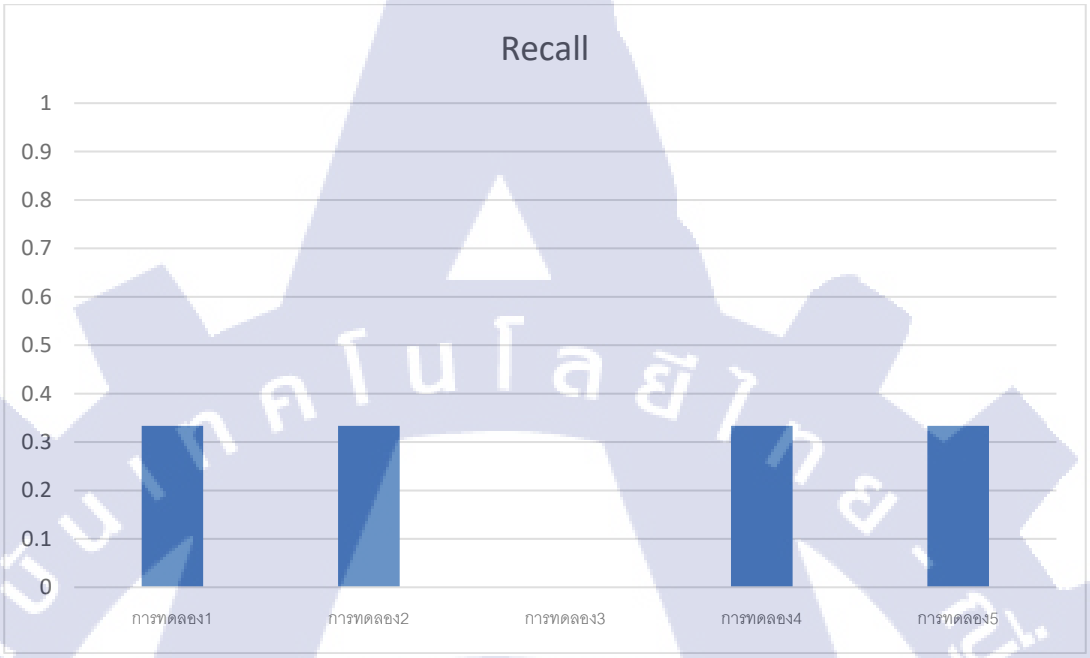
การทดลองประสิทธิภาพของ classifier แบบ 4 หัวข้อ

การทดลองประสิทธิภาพของ classifier แบบ 4 หัวข้อ หลังจากทำการทดลองข้อมูลกับระบบแล้ว ผลลัพธ์ที่ได้ออกมา จะนำมาแสดงค่าในรูปแบบ confusion matrix ซึ่งผลจะเป็นดังที่แสดงในตารางที่ ข.6 และจะมี 2 ค่า คือ True Positives และ False Negatives ที่เป็นผลลัพธ์มาจากการทดลอง

ตารางที่ ข.6 แสดงค่า confusion matrix ของการทดลอง 5 fold Cross-validation

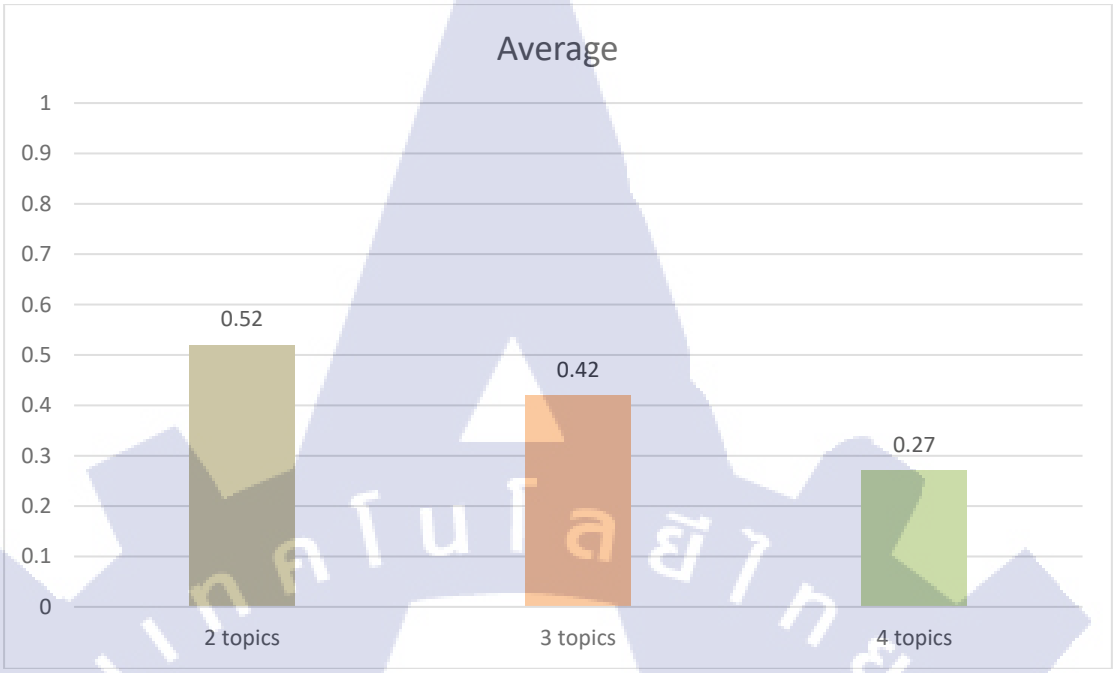
	True Negatives	False Negatives	False Positives	True Positives
การทดสอบ 1	0	0	2	10
การทดสอบ 2	0	0	4	8
การทดสอบ 3	0	0	5	7
การทดสอบ 4	0	0	2	10
การทดสอบ 5	0	0	4	8

จากรูปที่ ข.4 จะแสดงค่าเฉลี่ย Recall จากการรวมกันของเอกสารทั้ง 4 หัวข้อ ทั้ง 5 การทดสอบ โดยที่แสดงค่าแยกออกเป็นทีละการรวมกันของเอกสารแต่ละหัวข้อ ดังนี้



รูปที่ ข.4 แผนภูมิ Recall ของการรวมตัวของ 4 หัวข้อ

จากการทดสอบประสิทธิภาพ classifier แบบ multiple topics จะเห็นได้ว่าการทดสอบประสิทธิภาพจะสามารถแสดงได้เพียงแค่ค่า Recall เนื่องจากการทดสอบประสิทธิภาพ classifier แบบ multiple topics จะมีการ train ข้อมูลเฉพาะแค่ yes เพียงเท่านั้น จากรูปที่ ข.5 จะแสดงค่าเฉลี่ยของค่า Recall จากการรวมตัวของเอกสารทั้ง 2, 3 และ 4 หัวข้อ



รูปที่ ข.5 แผนภูมิค่าเฉลี่ย Recall ของการรวมกันของเอกสารทุกหัวข้อ

สรุปผลการทดลอง

ในการสรุปผลการศึกษาที่ผู้ศึกษาได้ศึกษาในหัวข้อการท่องเที่ยวเชิงนิเวศ โดยมีหัวข้อหรือมาตรฐานทั้งหมด 4 หัวข้อ ในงานครั้งนี้ ได้แก่ หัวข้อทรัพยากรธรรมชาติ, หัวข้อกิจกรรม, หัวข้อการให้ข้อมูล และหัวข้อการมีส่วนร่วมของชุมชน จึงได้จัดทำ binary classifier ทั้งหมด 4 ตัวตามหัวข้อที่กล่าวมาข้างต้น ซึ่ง classifier ทั้ง 4 ตัวนี้จะเป็นตัวช่วยในการแยกแยะหัวข้อ และยังค่อนข้างมีประสิทธิภาพในการแยกแยะได้ดีกับการทดลองในครั้งนี้อีกด้วย โดยผลการทดลองจะแบ่งออกเป็น 2 รูปแบบ คือ แบบ single topic และแบบ multiple topics ดังนี้

5.1.1 Single Topic

การทำการทดลองแบบ Single Topic ทำเพื่อตรวจสอบประสิทธิภาพของ classifier ในการแยกแยะหัวข้อของตัวเอง ซึ่งผลการทดลองโดยจะถูกแบ่งออกเป็น 4 โดยหัวข้อแรกทรัพยากรธรรมชาติได้ค่า Accuracy เท่ากับ 0.93 ค่า Precision เท่ากับ 0.93 และค่า Recall เท่ากับ 0.95 หัวข้อต่อมาหัวข้อกิจกรรมได้ค่า Accuracy เท่ากับ 0.90 ค่า Precision เท่ากับ 0.87 และค่า Recall เท่ากับ 0.95 หัวข้อต่อมาหัวข้อการให้ข้อมูลได้ค่า Accuracy เท่ากับ 0.90 ค่า Precision เท่ากับ 0.80 และค่า Recall เท่ากับ 1 และหัวข้อสุดท้ายหัวข้อการมีส่วนร่วมของชุมชนได้ค่า Accuracy เท่ากับ 0.83 ค่า Precision เท่ากับ 0.67 และค่า Recall เท่ากับ 1 ซึ่งแสดงให้เห็น

ว่าความสามารถของ classifier แต่ละตัวนี้ในการจดจำ และในการคาดเดาผลลัพธ์ของตัวเองนั้นทำงานได้อยู่ในเกณฑ์พอใช้ได้

5.1.2 Multiple Topics

การทำการทดลองแบบ Multiple Topics ทำเพื่อตรวจสอบประสิทธิภาพของ classifier ว่าทำงานได้จริง ในกรณีที่มีเนื้อหามากกว่า 1 หัวข้อ ซึ่งในตามความเป็นจริง หน้าเว็บไซต์จะมีข้อมูลที่รวมกันอยู่หลากหลายหัวข้อเป็นปกติ โดยผลการทดลองแบบ multiple topics จะแบ่งออกเป็น 3 อย่าง ดังนี้

- การรวมตัวกันของ 2 topic จะมีค่าเฉลี่ย Recall จากการทดสอบ 5 fold Cross-validation อยู่ที่ 0.52
- การรวมตัวกันของ 3 topic จะมีค่าเฉลี่ย Recall จากการทดสอบ 5 fold Cross-validation อยู่ที่ 0.42
- การรวมตัวกันของ 4 topic จะมีค่าเฉลี่ย Recall จากการทดสอบ 5 fold Cross-validation อยู่ที่ 0.27

ซึ่งผลการทดลองนี้ได้ค่า Recall ค่อนข้างต่ำ จะสังเกตว่าค่าประสิทธิภาพของ classifier ในแต่ละตัวจะอยู่ในเกณฑ์ค่อนข้างสูง แต่เมื่อทดสอบประสิทธิภาพของ classifier ที่เกิดจากการรวมกันของเอกสารที่มีหลายหัวข้อ พบว่าค่าประสิทธิภาพของระบบค่อนข้างต่ำ เนื่องจากจำนวนคำในเอกสารในแต่ละหัวข้อ มีจำนวนคำที่แตกต่างกันมาก จึงส่งผลให้เห็นว่าเมื่อนำเอกสารของแต่ละหัวข้อมารวมกัน จำนวนคำของแต่ละหัวข้อมีความแตกต่างมากเกินไป ทำให้การค้นหาและแยกแยะผิดพลาด และทำให้ค่าของ Recall แย่ลง จากการที่ผู้ศึกษาได้ทำการทดลองพบว่าจำนวนคำในเอกสารแต่ละหัวข้อควรมีจำนวนคำเฉลี่ยต่างกันไม่เกิน 100-200 คำ ซึ่งหากในเอกสารมีคำเฉลี่ยของแต่ละหัวข้อต่างกันไม่มาก จะส่งผลให้ประสิทธิภาพของระบบและผลลัพธ์ดียิ่งขึ้น ในงานอนาคตสามารถปรับเปลี่ยนจำนวน dataset หรือ ทำ sampling ในการแก้ไขความแตกต่างของจำนวนคำ