แบบเสนอหัวข้อโครงงาน

CP490 วิชาเลือก (การวิเคราะห์ข้อความ)

สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์มหาวิทยาลัยศรีนครินทรวิโรฒ

ประจำปีการศึกษา 2564

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

หมายเหตุ 1. ควรแบ่งงานให้เท่าๆกันในระหว่างสมาชิก

2. คะแนนในส่วน Proposal = 20% ของวิชานี้ ( ให้คะแนนเป็นรายบุคคล ตามแต่ละส่วนที่รับผิดชอบ พิจารณาทั้งในส่วน code และ การตอบคำถาม)

3. พร้อมแนบ python notebook

4. เปลี่ยนชื่อไฟล์เป็นชื่อกลุ่ม สามารถดูชื่อกลุ่มได้ที่ <https://docs.google.com/spreadsheets/d/1YyB5HRSpJhg6qdqT1crJo_EjWRwR0txSqy0PVk4vbOc/edit#gid=1202431381>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. ชื่อโครงการ การจัดประเภทข้อความในทวิตเตอร์ที่เกี่ยวกับภัยพิบัติNLP for Text Classification ( tweet is about a real disaster)

รายชื่อนิสิต1.นางสาวต่วนอะห์ลัม ต่วนสนิ 61102010424

2.นางสาวรัตนาพร ฤกษ์พอดี 61102010426

3.นางสาวสุรีรัตน์ ยุธสุนทร 61102010429

4.นางสาวฐิติยา - 61102010528

1. ที่มาและความสำคัญ

Text classification เป็นหนึ่งในงานที่สำคัญใน Supervised machine learning (ML) เป็นกระบวนการในการกำหนดแท็ก/หมวดหมู่ให้กับเอกสาร ซึ่งช่วยให้เราสามารถจัดโครงสร้างและวิเคราะห์ข้อความโดยอัตโนมัติ ในบทความนี้จะใช้ dataset “Natural Language Processing with Disaster Tweets” ซึ่งเราจะคาดการณ์ว่าทวิตนั้นเกี่ยวกับภัยพิบัติหรือไม่ โดยใช้โมเดล Logistic Regression และ Naive Bayes โดยเปรียบเทียบ Bag-of-Words ร่วมกับ Tf-Idf และ Word Embedding ร่วมกับ Word2Vec.

1. แหล่งที่มา
   * 1. NLP for Text Classification ( tweet is about a real disaster)
     2. เขียนโดยVijaya Rani เมื่อ1 เมษายน พ.ศ.2564
     3. [https://medium.com/analytics-vidhya/nlp-tutorial-for-text-classification-in-python-8f19cd17b49e ]
2. ทฤษฎีและแนวคิดที่ใช้ในการทำงานวิจัย

**NLP : Natural Language Processing**

การพัฒนาทฤษฎีที่เกี่ยวข้องกับภาษามนุษย์โดยที่เราต้องมีความรู้ความเข้าใจเกี่ยวกับโครงสร้างของภาษา เช่น ภาษาอังกฤษ กับ ภาษาไทย จะมีโครงสร้างที่แตกต่างกัน

เราต้องนำทฤษฎีข้างต้นมาพัฒนาซอฟต์แวร์ที่สามารถติดต่อสื่อสารระหว่างมนุษย์และคอมพิวเตอร์ได้

**Bag-of-Words(BoW)**

โมเดลที่ใช้กันแพร่หลายในงานจัดแบ่งประเภทข้อความ Text Classification ในโมเดลของ BOW กลุ่มของคำจะถูกอธิบายด้วยกระเป๋าคำ Bag of words หรือกลุ่มรวมของคำ โดยไม่ได้คำนึงถึงหลักไวยากรณ์ ความถี่ที่พบ และลำดับของคำ โดยนำมาใช้เป็น Feature ในการเทรนตัวจัดแบ่งข้อความ Classifier

**Word2Vec**

Word2vec เป็นแบบจำลองที่ใช้สร้างการฝังคำหรือแปลงคำให้อยู่ในรูปแบบของเวกเตอร์ ซึ่งเวกเตอร์ของคำต่าง ๆ ถูกคำนวณจากบริบทรอบข้าง โดยใช้เทคนิคโครงข่ายประสาทเทียม( Neural Network) แบบ Encoder-Decoder มีเลเยอร์ (Layer) จำนวน 2 เลเยอร์ ซึ่งมีหลักการในการเปรียบเทียบเวกเตอร์ทางความหมายของคำ ทั้ง 2 คำ แล้วคืนค่าออกมาเป็นตัวเลขตั้งแต่ -1 ถึง 1 บ่งชี้ถึง ความใกล้เคียงทางความหมายให้ค่าน้อยไปมาก หรือพูดอีกนัยหนึ่งว่า คำที่มีบริบทการปรากฏคล้าย ๆ กัน ควรเป็นคำที่มีความหมายคล้ายกันด้วย ซึ่งวิธี Word2vec สามารถวัดค่าความละม้ายทางความหมายของเวกเตอร์ของคำ และ ใช้ร่วมกับการจำแนกข้อมูลประเภทข้อความได้ดี โดยเมื่อมีมิติของข้อมูล (Dimension) มากขึ้นจะทำให้ตัวแบบมี ความแม่นยำมากขึ้น

**Word Embedding**

การแปลงเชิงปริมาณ หรือ Word Embedding คือ การแปลงคำ (Word) เป็นตัวเลข (Numerical) ให้อยู่ใน รูปแบบของเวกเตอร์ (Vector) ที่เหมาะสมจะนำเข้าไป ในวิธีในการสร้าง feature จาก คำ วิธีหนึ่ง โดยจะทำการลดขนาด vector space ลง แต่ไม่มีความถี่เข้ามาเกี่ยวข้อง

**tf-Idf**

TF-IDF หรือ Term Frequency Inverse Document Frequency เป็นเทคนิคการคัดแยกคำตามความสำคัญ ที่ถูกใช้ในการสร้างเวกเตอร์ โดยเทคนิคนี้ใช้ในการประเมินความสำคัญของคำในข้อความทั้งหมด ความสำคัญจะมีสัดส่วนเพิ่มตามจำนวนครั้งของคำที่เกิดขึ้น ในข้อความทั้งหมด เพื่อเปรียบเทียบกับสัดส่วนผกผันของคำนั้น ๆ ในข้อความทั้งหมด หรืออีกนัยหนึ่งว่า TF-IDF จะช่วยกรองคำเด่น ๆ ออกมาจากข้อความ ซึ่งจะเพิ่ม ความสำคัญของคำที่ถูกพบโดยการให้น้ำหนักเพิ่มขึ้น ส่งผล ทำให้ตัวแบบมีความแม่นยำมากขึ้น สามารถทำนายได้ดีกว่า วิธีเวกเตอร์ของการนับคำ (Count Vectorization)

**Logistic Regression**

การวิเคราะห์การถดถอยโลจิสติก (Logistic regression analysis) เป็นการวิเคราะห์การถดถอยแบบหนึ่งที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ (Qualitative variable) การวิเคราะห์การถดถอยโลจิสติกแบ่งออกเป็น 2 ประเภท คือ การวิเคราะห์การถดถอยโลจิสติกแบบไบนารี (Binary logistic regression analysis)คือการวิเคราะห์การถดถอยที่ตัวแปรตามมี 2 ค่าคือไม่เกิดเหตุการณ์(y = 0) หรือเกิดเหตุการณ์(y = 1) และการวิเคราะ ห์การถดถอยโลจิสติกพหุกลุ่ม(Multinomial logistic regression analysis) เป็นการวิเคราะห์ที่ตัวแปร ตามที่มีค่ามากกว่า 2 กลุ่ม (Polychromous variable) การวิเคราะห์การถดถอยโลจิสติกมีวัตถุประสงค์เพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ จึงทำให้ถูกประยุกต์ใช้ในหลาย สาขาวิชา เช่น ทางด้านการแพทยใ์ช้สมการโลจิสติกทำนายโอกาสของการเกิดโรคต่างๆ ทางด้านธุรกิจใช้ในการพยากรณ์ภาวะทางด้านเศรษฐกิจ เช่น โอกาสของการเกิดหน้าเสีย

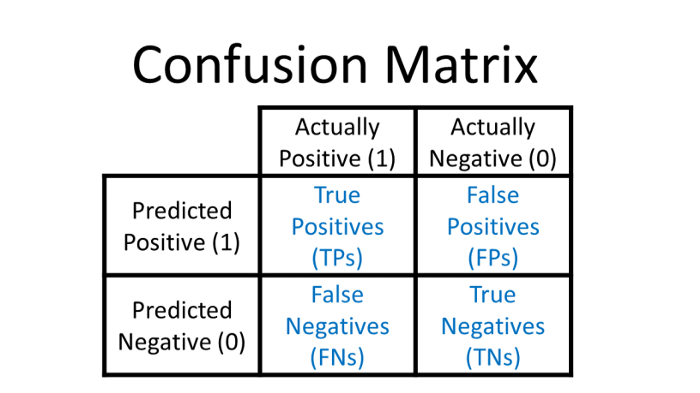
**Naive Bayes**

Naive Bayes Classifier เป็นไปตาม Bayes Theorem ทฤษฎีบทของเบย์กล่าวว่าความน่าจะเป็นตามเงื่อนไขของผลลัพธ์สามารถคำนวณได้โดยใช้ความน่าจะเป็นตามเงื่อนไขของสาเหตุของผลลัพธ์ ลักษณนาม Naive Bayes ใช้ตัวแปรอินพุตเพื่อเลือกคลาสที่มีความน่าจะเป็นหลังสูงสุด

อัลกอริทึมนี้เรียกว่าไร้เดียงสาเพราะสันนิษฐานว่าการกระจายของข้อมูลอาจเป็นแบบเกาส์เซียนหรือแบบปกติเบอร์นูลลีหรือการแจกแจงแบบพหุนาม

ข้อเสียเปรียบอีกประการหนึ่งของ Naive Bayes คือคุณสมบัติที่ต่อเนื่องจะต้องผ่านการประมวลผลล่วงหน้าและแยกออกจากกันโดย binning ซึ่งสามารถทิ้งข้อมูลที่เป็นประโยชน์ได้

**Confusion matrix**

Confusion Matrix ถือเป็นเครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนาย หรือ Prediction ที่ทำนายจากModelที่เราสร้างขึ้น ในMachine learning โดยมีไอเดียจากการวัดว่า สิ่งที่เราคิด (Model ทำนาย) กับ สิ่งที่เกิดขึ้นจริง มีสัดส่วนเป็นอย่างไร

1. **สิ่งที่แตกต่างจากบทความต้นฉบับ**

-เพิ่ม EDA สำรวจข้อมูลเพิ่มเติม ในส่วนของ คอลัมภ์ Keyword และ Location เพื่อดูความแตกต่างของ target 1 และ target 0

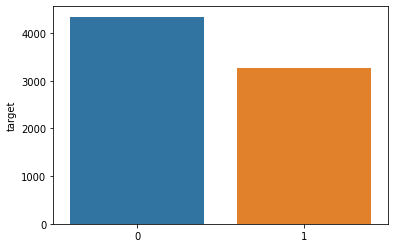
-หาความถี่ของคำใน Target = 1 (target ที่เกี่ยวข้องกับภัยพิบัติ)

-Plot Word Vectors โดยใช้ PCA

1. **แทรก link ipython notebook ของ project**

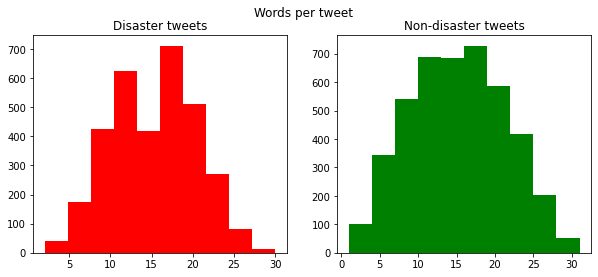
<https://colab.research.google.com/drive/1pzQ2PmQVL9OVbii5HheNOMiJEb7LdECq>

1. **การนำเสนอข้อมูล (EDA)**

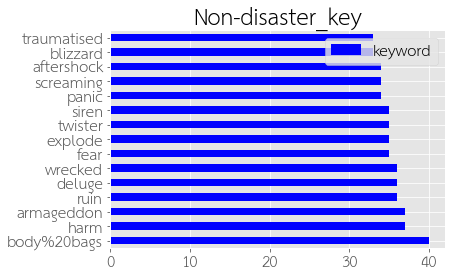
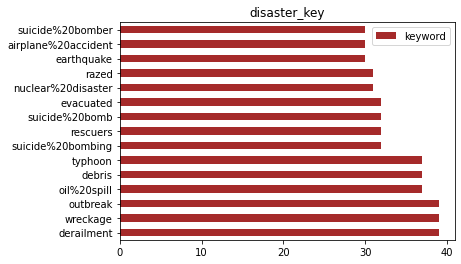
**

สำรวจข้อมูล target = 1 ,target = 0 เพื่อดูลักษณะ Class distribution แสดงทวีตคลาส 0 (ไม่มีภัยพิบัติ) มากกว่าคลาส 1 ( ทวิตภัยพิบัติ) 4342 ทวิตที่ไม่เกิดภัยพิบัติ (57%) และ 3271 ทวิตเกี่ยวกับภัยพิบัติ (43%)

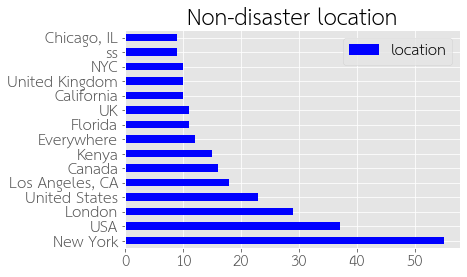
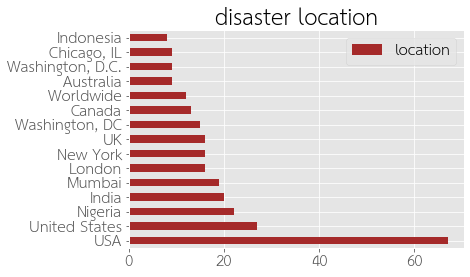
เนื่องจากข้อมูลมีความสมดุล เราจะไม่ใช้เทคนิคการปรับสมดุลข้อมูล เช่น SMOTE ขณะสร้างแบบจำลอง

**

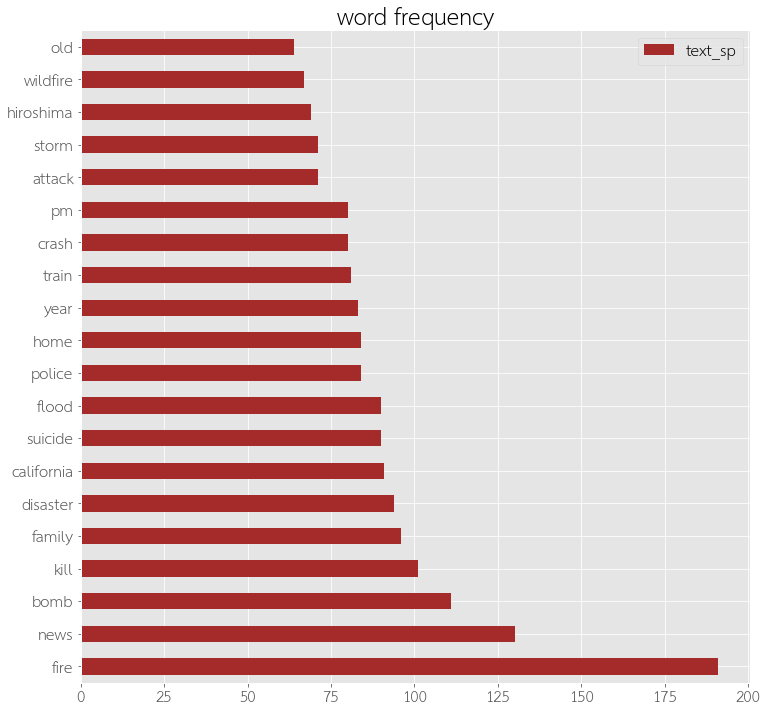
สำรวจข้อมูลจำนวนคำต่อทวิตของทั้งสอง target เพื่อแสดงจำนวนคำโดยรวมของแต่ละทวิต จำนวนคำเฉลี่ยในทวิตเกี่ยวกับภัยพิบัติ (target = 1) คือ 15.17 เมื่อเทียบกับค่าเฉลี่ย 14.7 คำในทวีตที่ไม่เกี่ยวกับภัยพิบัติ(target = 0)

**

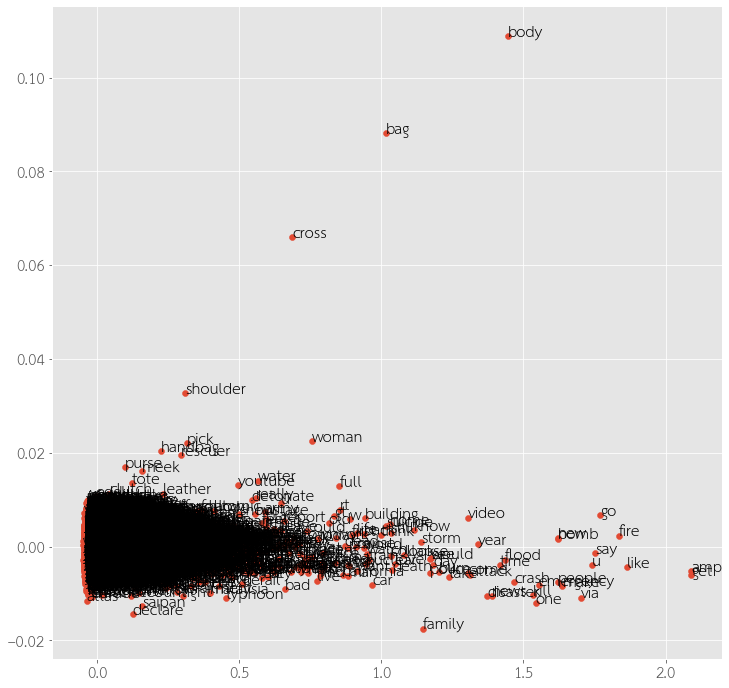
แสดง 15 คำที่พบบ่อยใน keyword ของทั้งสอง target เพื่อเปรียบเทียบว่ามีลักษณะ Keyword ซ้ำกันมากน้อยแค่ไหน เพื่อช่วยในการ select feature ในการสร้าง model ซึ่ง target 1 และ target 0 มีลักษณะ keyword ค่อนข้างแตกต่างกัน

**

แสดง 15 คำที่พบบ่อยใน Location ของทั้งสอง target เพื่อเปรียบเทียบว่ามี Location ซ้ำกันมากน้อยแค่ไหน เพื่อช่วยในการ select feature ในการสร้าง model ซึ่ง target 1 และ target 0 มีลักษณะ Location ค่อนข้างคล้ายคลึงกัน

**

แสดง 20 คำที่พบบ่อยในทวิตที่เกี่ยวข้องกับภัยพิบัติ (target = 1) ซึ่งแสดงให้เห็นลักษณะภัยพิบัติที่พบบ่อย เช่น fire ,flood,storm เป็นต้น

**กราฟแสดงการจับกลุ่มคำที่คล้ายคลึงกัน และแสดงการกระจายของคำ

1. **แผนการดำเนินงาน และการแบ่งงาน**

ตาราง 1 แสดงงาน, รายละเอียด และผู้รับผิดชอบ

หมายเหตุ โครงงานของนิสิตอาจจะมี task ที่แตกต่างไปจากนี้ สามารถปรับได้ตามความเหมาะสม

| Task | รายละเอียด | ผู้รับผิดชอบ |
| --- | --- | --- |
| overview | บอกความสำคัญของโครงงาน เข้าใจบทความต้นฉบับ และบอกสิ่งที่กลุ่มทำที่แตกต่างจากบทความต้นฉบับ ประเภทปัญหาของโครงงาน (classification, regression, sentiment analysis อื่นๆ) | รัตนาพร |
| prepossessing | ทำความสะอาดข้อมูล ระบุรายละเอียดการทำความสะอาด  ปัญหาของข้อมูล วิธีการ tokenization (ถ้ามี) | ต่วนอะห์ลัม |
| EDA | สร้างกราฟต่างๆ เพื่อแสดงถึงภาพรวมข้อมูล คำที่สำคัญ จำนวนคำทั้งหมด การกระจายตัวของคำ และลักษณะพิเศษที่ควรนำมานำเสนอ หรือมีประโยชน์ต่อโมเดล | ต่วนอะห์ลัม |
| Word representation | ทำการแปลงเป็น Word vector (feature extraction ) ระบุหลักการ และสาเหตุที่เลือกใช้ ขนาดของ vector ปัญหาที่พบ | ฐิติยา |
| Model | ระบุโมเดลที่คาดว่าจะใช้ พร้อมเหตุ ระบุข้อดี ข้อเสียของโมเดล | ฐิติยา |
| Evaluation | ระบุวิธีการวัดการประเมิน หลักการค่าตัวที่สำคัญ พร้อมเหตุผล (เช่น เลือกวัดจากค่า accuracy เพราะอะไร) | สุรีรัตน์ |
| Insight from result | จาก result คุณคาดหวังว่าจะได้เรียนรู้อะไรจาก โมเดล ปัญหาที่คาดว่าจะเกิดขึ้นกับโมเดล และ ไอเดียการแก้ไข | สุรีรัตน์ |
| อื่นๆ | งานที่อื่นๆที่เกิดขึ้นในโครงงานของคุณ | รัตนาพร |