VISTEC-depa
THAILAND
AI RESEARCH
INSTITUTE

VISTEC
VIDYASIRIMEDHI
INSTITUTE OF SCIENCE AND TECHNOLOGY

Mini Project
Sentiment Analysis

School of Information Science and Technology **(IST)**
Vidyasirimedhi Institute of Science & Technology **(VISTEC)**

Lab

Task Overview

- เนื่องด้วยในปัจจุบันการเก็บ ข้อความ จากลูกค้านั้นเป็นไปได้ด้วยความรวดเร็ว เช่นการเก็บ ข้อความจากหน้า page facebook หรือการทำ auto logging จากระบบ call center ทำให้หลายๆบริษัทนั้นมีความต้องการที่จะจำแนกหัวข้อของข้อความ เพื่อที่จะสามารถตอบสนองความต้องการของลูกค้าได้รวดเร็วขึ้น

- การนำ NLP มาใช้สร้างระบบการจำแจกข้อความนั้น จึงมีความจำเป็นอย่างมาก ดังนั้นในโจทย์นี้ต้องการให้ ผู้เข้าร่วม mini-project ทำการจำแนกข้อความจากลูกค้าตามหัวข้อดังต่อไปนี้
  - billing and payment,
  - promotions
  - other queries
  - internet
  - international dialing
  - true money
  - lost and stolen

Data Explanation

- train dataset จะประกอบด้วย ข้อความทั้งหมด 10,330 ข้อความ ซึ่งมีประเภทอยู่ใน column “destination”
- test dataset จะประกอบด้วย ข้อความทั้งหมด 1,744 ข้อความ

Train dataset

	texts	destination
0	โทรศัพท์มีปัญหาโทรออกได้ แต่เวลาคนอื่นโทรเข้าม...	billing and payment
1	สอบถามเกี่ยวกับการโอนเงินของระบบเติมเงินไปอีกห...	billing and payment
2	แจ้งเครื่องขัดข้อง พอดีดูสัญญาณไม่ได้ค่ะ	billing and payment
3	สวัสดีครับคือผมอยากทราบยอดค้างค่าบริการทั้งหมด...	billing and payment
4	พี่ครับ ผมอยากทราบว่าตอนนี้ยอดค้างของผมเท่าไหร...	billing and payment

Test dataset

	texts
0	ค่ะ พี่ค่ะ คืออยู่ๆ เน็ตเล่นไม่ได้อ่ะค่ะ
1	ครับผม อ่า ผมเปลี่ยนซิมเป็น ซิมทรูมูฟเอชอ่ะครั...
2	ผมขอเช็คข้อมูล พึ่งชำระยอดเข้าไปแต่พนักงานแจ้ง...
3	จะสอบถามเรื่องทรูแบล็คการ์ดหน่ะครับ ไม่ทราบว่า...
4	อยากจะเช็คยอดการโทรค่ะ

Submission

- ผู้เข้าเรียนต้องทำการวิเคราะห์ข้อมูล และสร้าง Machine learning model เพื่อ Classify text to destination โดยต้องจัดส่งไฟล์ดังนี้ ภายใน **31 ตุลาคม 2020** ที่ ai_academy_vistec@outlook.com

    1. Report ประกอบด้วย

        a) Presentation อธิบายการ Clean data, EDA, Model, และ Summary

        b) Python notebook กรุณาแยกเป็น 2 ไฟล์ คือ 1. Clean data and EDA และ 2. Modeling

    2. คำตอบของ Test dataset 1,744 ข้อความ
    โดยส่งคำตอบใน column “destination” ตามรูปแบบของการ submission form submission_template.csv
    ไม่เกินจำนวน 3 ไฟล์ (อาจมาจากหลาย Model) โดยตั้งชื่อไฟล์
    submission_1.csv, submission_2.csv, และ submission_3.csv

สามารถดู ได้ใน Titanic Mini-project ใน Moodle

	texts	destination
0	ค่ะ พี่ค่ะ คืออยู่ๆ เน็ตเล่นไม่ได้อ่ะค่ะ	billing and payment
1	ครับผม อ่า ผมเปลี่นนชิมเป็น ชิมทรูมูฟเอชอ่ะครั...	billing and payment
2	ผมขอเช็คข้อมูล พึ่งชำระยอดเข้าไปแต่พนักงานแจ้ง...	billing and payment
3	จะสอบถามเรื่องทรูแบล็คการ์ดหน่ะครับ ไม่ทราบว่า...	billing and payment
4	อยากจะเช็คยอดการโทรค่ะ	billing and payment

Score Benchmark

- Mini-project เป็นงานรายบุคคล คะแนนเต็ม 100 points ซึ่งคิดเป็น 20% ของคะแนนผ่าน Level 2 ทั้งหมด
  ถ้าทีมงานตรวจสอบว่าเกิดการลอกกัน หรือโกงคำตอบ จะทำการแจ้งผู้บังคับบัญชาของผู้เข้าเรียน และไม่ผ่าน Level 2
- Score ในส่วนของ Report *Python notebook ที่ส่งมาต้องสามารถรันได้ ถ้ารันไม่ได้คะแนนจะลดลง 50% (จะมีการติดต่อไปก่อนในกรณีรันไม่ได้)
  - Clean data: 0-10 points
  - EDA: 0-10 points
  - Model: 0-20 points
  - Summary: 0-10 points
- Score ในส่วนของ test dataset (ใช้คะแนนที่สูงสุดจาก submission.csv) *ต้องสามารถรันคำตอบได้จาก notebook ที่ส่งมา
  - 30 points: Micro f1 Better than 0.20
  - 35 points: Micro f1 Better than 0.30
  - 40 points: Micro f1 Better than 0.40
  - 45 points: Micro f1 Better than 0.50
  - 50 points: Micro f1 Better than 0.60