DS4Biz - *Assignment 1* (20 Marks)

Data Collection & Preparation

Deadline: วันอาทิตย์ที่ 13 กันยายน 2563 เวลา 24.00 น.
Submission: Individual github repository – stage, commit with message, and push

Overview

จุดประสงค์ของ Assignment นี้ คือ ให้นศ.ทำการดึงข้อมูลเพื่อสร้าง dataset (staging) จากอย่างน้อย 1 หรือมากกว่า (one or more) จาก open web APIs ใดก็ได้แล้วแต่นศ. เลือกมา แต่หากนศ.ดึงข้อมูลจาก APIs มากกว่า 1 แหล่ง ข้อมูลจะต้องมีความสัมพันธ์กัน (เป็นเรื่องที่เกี่ยวข้องกัน โดยสามารถเอามาแสดงและวิเคราะห์เป็นเรื่องเดียวกันได้ เช่น ข้อมูลราคาสินค้า + ข้อมูลค่าครองชีพ ของแต่ละเมืองใหญ่ในประเทศไทย)

โดยหลังจากนศ.ทำการดึงและรวบรวมข้อมูลแล้ว ให้นศ. ใช้ภาษา Python ในการประมวลผลเบื้องต้น (ไม่ต้องมีการสร้างโมเดลในการทำนายต่าง ๆ) และวิเคราะห์ข้อมูลที่รวบรวมมา

Assignment นี้ จะต้องถูกทำลงใน Jupyter Notebook และใช้เพียง Notebook เดียวเท่านั้น (Not a script.) โค้ดที่นศ. เขียนจะ**ต้อง**มีการเขียนอธิบายที่ชัดเจนโดยใช้ Markdown cells เพื่ออธิบายแต่ละ Code cells ที่อยู่ด้านล่าง (ลำดับถัดไป) ของแต่ละ Markdown cell รวมทั้งใช้ inline (#) or block comments (""" … """) ประกอบการอธิบายโค้ดและผลลัพธ์ของการวิเคราะห์ข้อมูลของนศ.

โดยทุก ๆ Code cells จะต้องมีการเขียนอธิบายโดย Markdown cells ว่า แต่ละ Code cells นั้นทำอะไร และแปลความหมายผลลัพธ์ว่าอะไร

Tasks:

ใน Assignment นี้ นศ.จะต้องทำตามรายการต่อไปนี้เสร็จสิ้นทุกข้อ

1. เลือกอย่างน้อย 1 หรือมากกว่า (one or more) จาก open web APIs เพื่อเป็นแหล่งของการ staging ข้อมูล (Data Acquisition) หากนศ.เลือกมากกว่า 1 API APIs เหล่านั้นจะต้องสัมพันธ์กัน ตามที่อธิบายข้างต้น
2. รวบรวมข้อมูลจาก API(s) ที่นศ. เลือกโดยใช้ Python เท่านั้น และแสดงการรวบรวมข้อมูลใน Jupyter Notebook โดย Dataset ที่รวบรวมมาควรมีข้อมูลอย่างน้อย 500 records/items (ของใครน้อยกว่า 500 จะทำการหักคะแนน) โดยขึ้นอยู่กับ API ที่นศ. เลือกนศ. อาจจำเป็นต้องทำการดึงข้อมูลหลาย ๆ ครั้ง เพื่อให้ได้ข้อมูลที่เพียงพอต่อการทำ Assignment (ซึ่งการดึงหลายครั้ง ไม่มีผลต่อคะแนน)
3. ให้นศ. ทำการ Parse ข้อมูลที่รวบรวมมาและเก็บ (Save) ข้อมูลลงใน subfolder ชื่อ \data\ ซึ่งต้องอยู่ใน location เดียวกับ Jupyter Notebook โดยข้อมูลที่รวบรวมมาจะต้องอยู่ file format ที่เหมาะสมสำหรับทำการวิเคราะห์ในลำดับถัดไป เช่น (เช่น JSON, CSV, XML)
4. โหลดข้อมูลที่บันทึกไว้จากข้อ 3 แล้วทำการ represent ใน Padas Dataframe ให้นศ. แสดงการประมวลผลเบื้องต้น (Pre-processing) และทำการแสดงการตรวจสอบคุณภาพ (Quality checking) ของข้อมูลที่นศ. ได้มา ทีละขั้นตอน (Step by step) ซึ่งจำเป็นที่จะต้อง clean and filter the data ก่อนไปทำการวิเคราะห์

5. ให้นศ. ทำการวิเคราะห์และสรุปแปลผลเชิงลึกของข้อมูลที่ได้มาจากข้อมูลที่ทำความสะอาดแล้ว (cleaned dataset) โดยใช้ตารางและกราฟต่าง ๆ (Tables and Plots) ตามที่เหมาะสมของข้อมูลที่ได้มา (ยิ่งมีการแปลผลและสรุปผล ที่ลึกซึ้ง ซับซ้อน สมเหตุสมผล และการวาดกราฟที่มีความยากและซับซ้อน แต่ยังคงต้องเป็นประโยชน์ต่อการแปลผลข้อมูล ค้นพบ value ต่าง ๆ ในข้อมูล นศ.ก็จะได้คะแนนมากขึ้นตามลำดับ) โดยนศ. จะต้องอธิบายการแปลผลและ insights ที่ได้จากการวิเคราะห์ที่น่าสนใจใน Markdown cells ของ Jupyter Notebook รวมทั้งเสนอแนะการวิเคราะห์ข้อมูลนั้นเพิ่มเติมในอนาคต

Guidelines:

- ให้นศ. ทำการส่ง Assignment ซึ่ง คือ
    1. ไฟล์ .ipynb ของ Jupyter Notebook ของนศ. พร้อมข้อมูลที่รวบรวมมา ใน Github repository ของนศ. แต่ละคน โดยในแต่ละ Jupyter Notebook ของนศ. นศ.จะต้องเขียน ชื่อ นามสกุล รหัสนศ. ลงใน Markdown cell แรกของ Notebook
    2. Snapshots ของไฟล์ .ipynb ที่มี output ของทุก cells เป็น
        - ไฟล์ .html โดยการ download as ไฟล์ .ipynb ของนศ. เป็น html
        - ไฟล์ .pdf โดยการ download as ไฟล์ .ipynb ของนศ. เป็น pdf

       ทั้งนี้เพื่อป้องกันการนำไฟล์มารันใหม่แล้วได้ผลลัพธ์ไม่เหมือนเดิม จะได้ตรวจจากผลลัพธ์ที่นศ.ทำไว้
    3. ข้อมูลที่นศ.ทำการโหลดมาจาก API ใน folder \data\
- Assignment นี้เป็นงานเดี่ยว ของนศ. แต่ละคน หากมีการตรวจสอบพบการคัดลอก (Plagiarism) จะได้ 0 คะแนนในของส่วน Assignment นี้ หากมีข้อสงสัย และหากมีหลักฐานชัดเจนว่ามีการคัดลอกงานจากแหล่งใด ๆ ก็ตาม นศ.จะได้เกรด F ในวิชานี้ และส่งเรื่องต่อให้กับทางคณะฯ และสถาบันฯ ต่อไป
- Hard deadline: วันอาทิตย์ที่ 13 กันยายน 2563 เวลา 24.00 น.
    1. ส่งช้า 1-5 วัน: ลด 20% จากคะแนนตรวจที่ได้ (ขอเปิด Github ให้ส่งช้า)
    2. ส่งช้า 6-10 วัน: ลด 40% จากคะแนนตรวจที่ได้ (ขอเปิด Github ให้ส่งช้า)
    3. จะไม่มีการรับตรวจ Assignment หากส่งช้าเกิน 10 วัน โดยปราศจากหลักฐานชี้แจงเหตุผลในการส่งงานช้า ได้แก่ หลักฐานด้านการแพทย์ว่าเข้านอนโรงพยาบาลเพื่อรับการรักษา