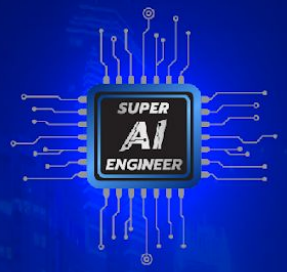


SUPER AI ENGINEER
SEASON 4

ThaiJO RESEARCHER (Data Science Hackathon)

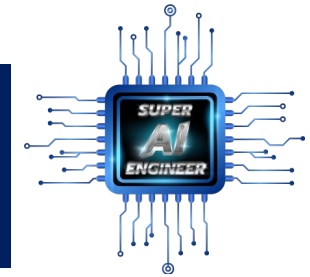
9-11 February 2024

ThaiJO RESEARCHER

Task:

- ชื่อนักวิจัยจะปรากฏอยู่ 2 columns คือ “_source.author” และ “_source.co-author”
- ลบคำนำหน้าชื่อ ตำแหนง ยศ เช่น dr. รศ. ทันตแพทย์หญิง นาย นางสาว พระครู เป็นต้น
- กำหนด ID โดยใส่ _ แล้วเรียงลำดับตามจำนวนชื่อที่มีในแต่ละ ID และต้องเป็นชื่อที่ไม่ซ้ำกัน
- 1 row อาจมีมากกว่า 1 ชื่อ ต้องแยกชื่อเหล่านั้นออกมา แล้วกำหนด ID ให้แต่ละชื่อ
- ชุดข้อมูลมีทั้งชื่อไทยและอังกฤษ ร่วมถึงภาษาต่างประเทศอื่นๆ ด้วย
- ปรับตัวอักษรภาษาอังกฤษให้ขึ้นต้นตัวใหญ่แค่ตัวแรกของชื่อและนามสกุลเท่านั้น นอกนั้นตัวเล็กทั้งหมด
- ข้อความอื่นใดที่ไม่ใช่ชื่อและนามสกุลไม่สนใจ
- ถ้าชื่อมีมากกว่า 10 ชื่อตัดมาแค่ 10 ชื่อแรกเท่านั้น

ThaiJO RESEARCHER



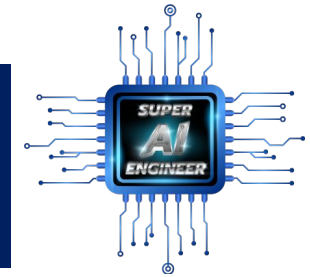
Task:

1. ถ้ามีทั้งชื่อภาษาไทย ภาษาอังกฤษ ให้แยกเป็น 2 บรรทัด (rows) เช่น

   id 1 ทีปกร คำพิทักษ์

   id 2 Teepakorn Kampitak
2. ชื่อภาษาอังกฤษ ตัวอักษรตัวแรกของชื่อและนามสกุล ใช้พิมพ์ใหญ่
3. ถ้าขีดเป็นส่วนหนึ่งของชื่อให้เก็บไว้ แต่ถ้าไม่ใช่ให้เอาออก
4. กรณีที่มีวงเล็บหลังชื่อพระ เก็บวงเล็บไว้
5. ชื่อตำแหน่ง เช่น ทันตแพทย์หญิง นาย นางสาว พระครู เอาออก
6. ถ้ามีอักขระพิเศษ เช่น * เอาออก
7. กรณีมีชื่อกลาง ซึ่งเป็นส่วนหนึ่งของชื่อ ให้เก็บเอาไว้
8. กรณีเป็นชื่อภาษาอื่นๆ ให้นำชื่อภาษานั้นๆมาตอบทั้งหมด
9. กรณีที่ไม่มีชื่อ ให้เว้นว่างช่องว่างทั้ง 10 ช่อง

ThaiJO RESEARCHER



Dataset:

- test.csv ชุดข้อมูลทดสอบ จำนวน 4,730 แถว

ต้องมีข้อมูลทั้งหมด 47,300 แถว
(พร้อมเฉลย 3 แถวแรก)

Output format:

_ID	Author	Co-author
66001	มะดารี โตะและ	มะดารี โตะและ, วัชรวุฒิ ซื่อสัตย์ และ วัชรพล พุทธรักษา

Id	name
66001_1	มะดารี โตะและ
66001_2	วัชรวุฒิ ซื่อสัตย์
66001_3	วัชรพล พุทธรักษา

sample_submission.csv ตัวอย่าง
ไฟล์ที่ใช้ส่ง

ThaiJO RESEARCHER

Evaluation Metric:

Exact Match

Kaggle link

https://www.kaggle.com/t/6485712fe6a047fc9d5f3511746023c9

ThaiJO RESEARCHER

ชื่อ (Team) สามารถเปลี่ยนได้หลังจากส่ง Submission 1 ครั้ง

ระยะเวลาการแข่ง : 9 Feb 20:00 - 11 Feb 08:00

จำนวนครั้งในการส่ง : 4 ครั้ง/วัน, ตัดรอบ 07.00 น.

- 9 Feb 20:00 - 10 Feb 07:00, 4 ครั้ง
- 10 Feb 07:01 - 11 Feb 07:00, 4 ครั้ง
- 11 Feb 07:01 - 11 Feb 08:00, 4 ครั้ง

ThaiJO RESEARCHER

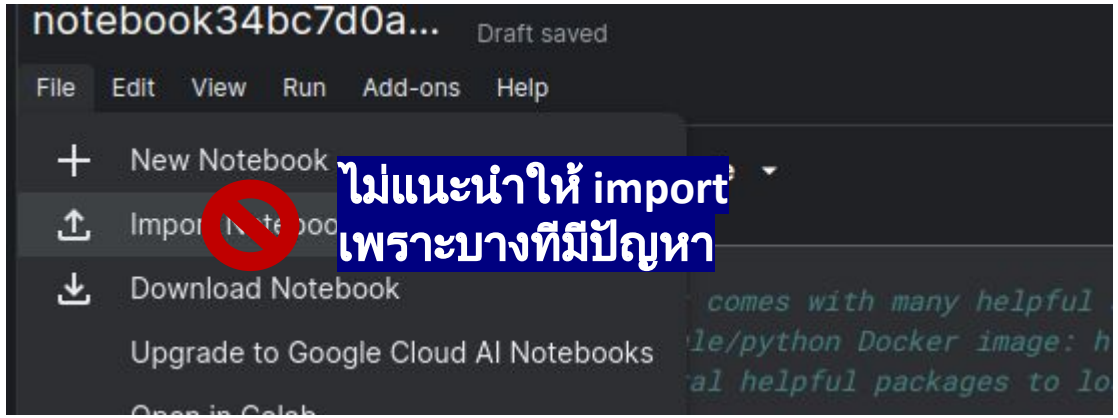
การตั้งชื่อทีมที่ถูกต้อง : 400000-ทิพย์อรุณ

ตัวอย่างการตั้งชื่อที่ผิด :

- ✗ ช่องกลางระหว่างชื่อ เช่น
  400000 - ทิพย์อรุณ
- ✗ คั่นด้วย _ เช่น
  400000_ทิพย์อรุณ
- ✗ ไม่มีคั่น เช่น
  400000 ทิพย์อรุณ
- ✗ มีนามสกุลต่อท้ายด้วย เช่น
  400000-ทิพย์อรุณ งดงามมาก
- ✗ ชื่อภาษาอังกฤษ เช่น
  400000-Tiparoon
- ✗ มีชื่ออย่างอื่นท้ายด้วยเช่น
  400000-ทิพย์อรุณ Image Search
- ✗ มีรหัสและตามด้วยคำอื่นๆ เช่น
  400000-สวัสดีครับ

ThaiJO RESEARCHER

Copy โคดมาใส่ใน cell ตรงๆ เลยจะปัญหาน้อยที่สุด



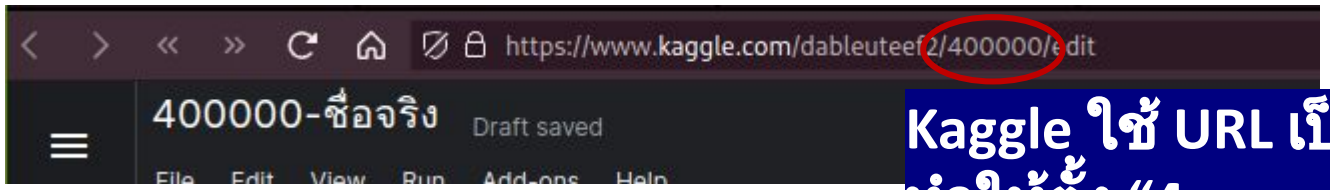
ไม่แนะนำให้ import เพราะบางทีมีปัญหา

```python
import pandas as pd
import regex as re


def remove_spaces(string):
    if string == '':
        return ''
    elif string == ' ':
        return ''
    if string[0] == ' ':
        string = string[1:]
    if string[-1] == ' ':
        string = string[:-1]
    return string

if __name__ == '__main__':
    char = {'id': [], 'name': []}
    df = pd.read_csv('thaijo/test.csv')
    df = df.fillna('')
    for idx, row in df.iterrows():
        names = []
        char['id'].extend([f"{row['_id']}_{i+1}" for i in range(10)])
        author = row['_source.author']
        if author != '':
            names.append(author)
        co_author = row['_source.co-author']
        if co_author != '':
            ca = co_author.split(',')
            for c in ca:
                if c not in names:
                    names.append(c)
        if len(names) == 0:
            char['name'].extend([None]*10)
            continue
        names = [remove_spaces(name) for name in names]
        if len(names) > 10:
            names = names[:10]
        elif len(names) < 10:
            names.extend(['']*(10-len(names)))
        char['name'].extend(names)
    char = pd.DataFrame(char)
    char.to_csv('thaijo/predict2.csv', index=False)
```

ThaiJO RESEARCHER



Kaggle ใช้ URL เป็นชื่อของ notebook
ทำให้ตั้ง “4xxxxxxx-ชื่อจริง” ซ้ำไม่ได้

เพราะฉนั้นให้ตั้งเป็น “ชื่อการแข่ง-4xxxxxxx-ชื่อจริง”


SUPER AI ENGINEER
SEASON 4
Q & A