DATA ENGINEER ASSIGNMENT

Kwan Jun Hao

Contents

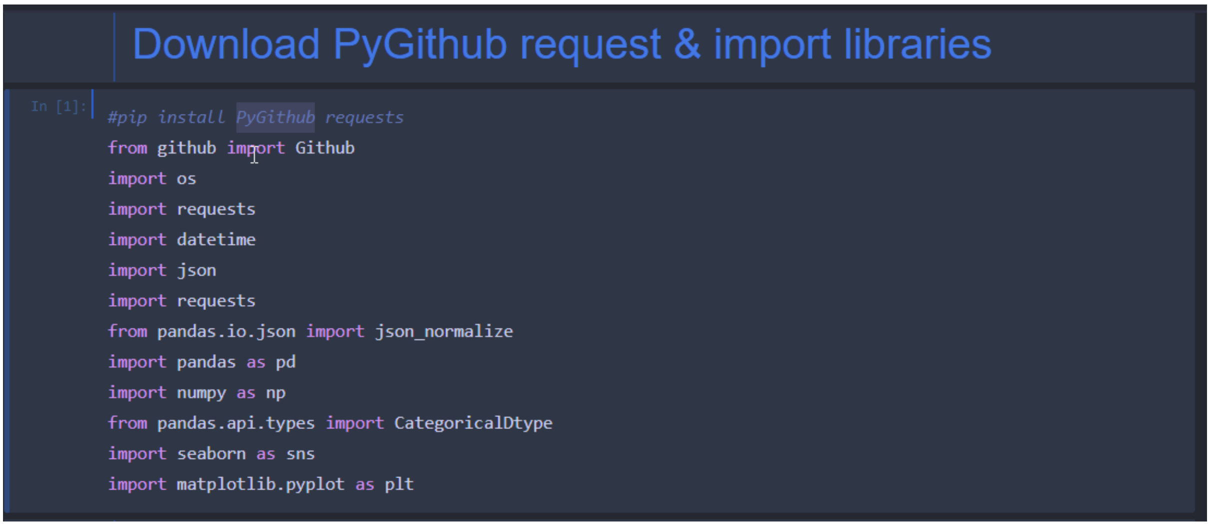
Introduction

In this guide, we will be running thru the python codes and how to operate each section. The requirements needed to run the codes are:

- Python 3
- Juypter Note Book

Open the code py_gitcommit_elt.ipynb in Juypter Note Book.

Libraries and imports

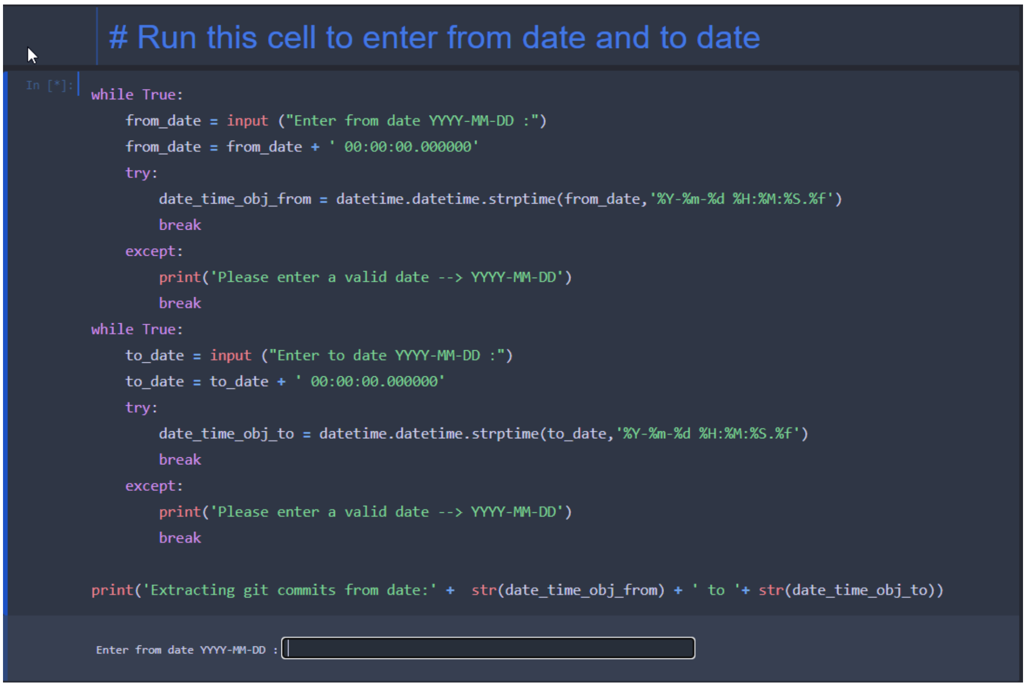


Download PyGithub request & import libraries

In [1]:

```
#pip install PyGithub requests
from github import Github
import os
import requests
import datetime
import json
import requests
from pandas.io.json import json_normalize
import pandas as pd
import numpy as np
from pandas.api.types import CategoricalDtype
import seaborn as sns
import matplotlib.pyplot as plt
```

Dynamic Date of From and To of Git commits



Run this cell to enter from date and to date

In [*]:

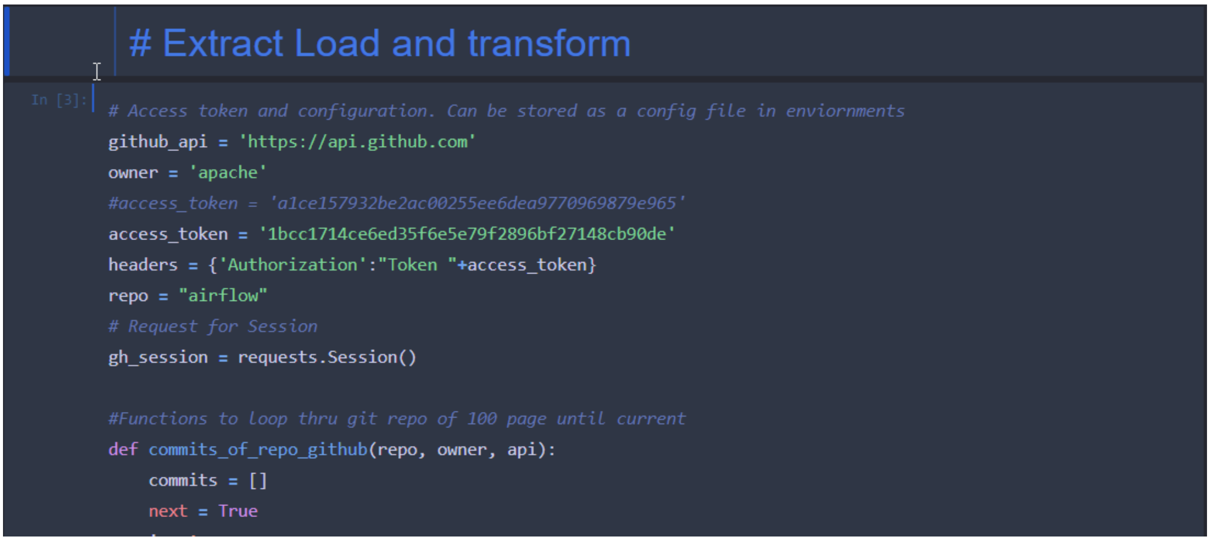
```
while True:
    from_date = input ("Enter from date YYYY-MM-DD :")
    from_date = from_date + ' 00:00:00.000000'
    try:
        date_time_obj_from = datetime.datetime.strptime(from_date,'%Y-%m-%d %H:%M:%S.%f')
        break
    except:
        print('Please enter a valid date --> YYYY-MM-DD')
        break
while True:
    to_date = input ("Enter to date YYYY-MM-DD :")
    to_date = to_date + ' 00:00:00.000000'
    try:
        date_time_obj_to = datetime.datetime.strptime(to_date,'%Y-%m-%d %H:%M:%S.%f')
        break
    except:
        print('Please enter a valid date --> YYYY-MM-DD')
        break

print('Extracting git commits from date:' +  str(date_time_obj_from) + ' to '+ str(date_time_obj_to))
```

```
Enter from date YYYY-MM-DD :
```

Enter date format of YYYY-MM-DD

Extract Load and Transformation



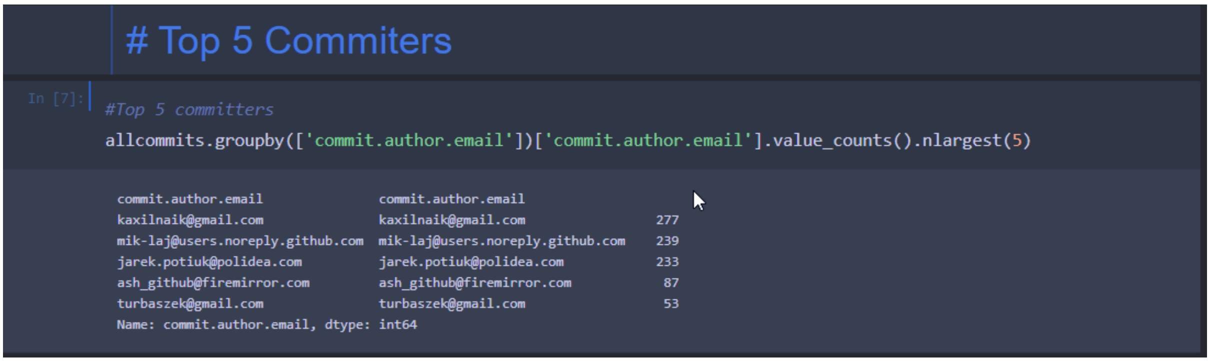
```
# Extract Load and transform
```

In [3]:

```python
# Access token and configuration. Can be stored as a config file in enviornments
github_api = 'https://api.github.com'
owner = 'apache'
#access_token = 'a1ce157932be2ac00255ee6dea9770969879e965'
access_token = '1bcc1714ce6ed35f6e5e79f2896bf27148cb90de'
headers = {'Authorization':"Token "+access_token}
repo = "airflow"
# Request for Session
gh_session = requests.Session()

#Functions to loop thru git repo of 100 page until current
def commits_of_repo_github(repo, owner, api):
    commits = []
    next = True
```

Top 5 commits



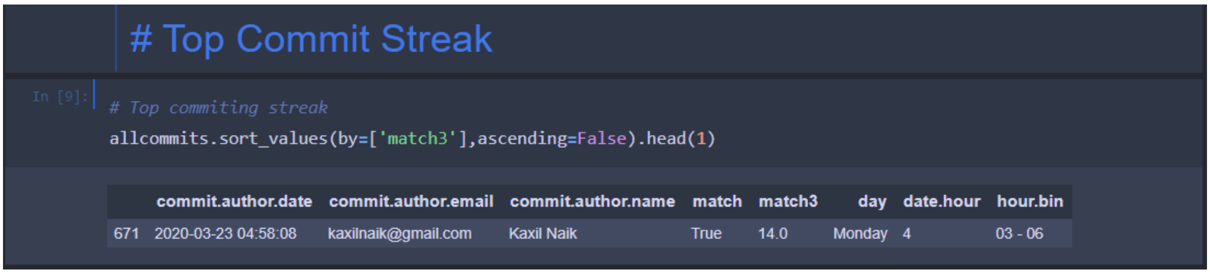
```
# Top 5 Commiters
```

In [7]:

```python
#Top 5 committers
allcommits.groupby(['commit.author.email'])['commit.author.email'].value_counts().nlargest(5)
```

```
commit.author.email               commit.author.email
kaxilnaik@gmail.com               kaxilnaik@gmail.com                  277
mik-laj@users.noreply.github.com  mik-laj@users.noreply.github.com     239
jarek.potiuk@polidea.com          jarek.potiuk@polidea.com             233
ash_github@firemirror.com         ash_github@firemirror.com             87
turbaszek@gmail.com               turbaszek@gmail.com                   53
Name: commit.author.email, dtype: int64
```

Top commit streak



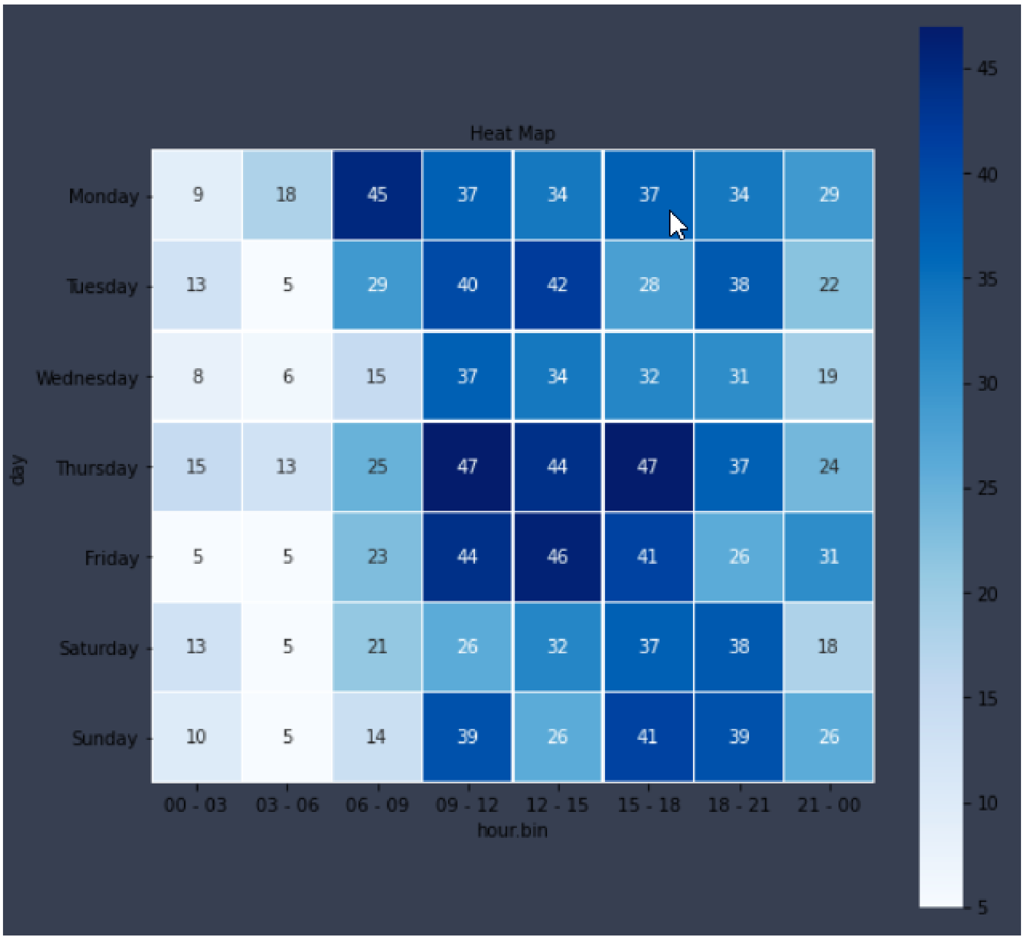
```
# Top Commit Streak
```

In [9]:

```python
# Top commiting streak
allcommits.sort_values(by=['match3'],ascending=False).head(1)
```

	commit.author.date	commit.author.email	commit.author.name	match	match3	day	date.hour	hour.bin
671	2020-03-23 04:58:08	kaxilnaik@gmail.com	Kaxil Naik	True	14.0	Monday	4	03 - 06

Heatmap



day	00 - 03	03 - 06	06 - 09	09 - 12	12 - 15	15 - 18	18 - 21	21 - 00
Monday	9	18	45	37	34	37	34	29
Tuesday	13	5	29	40	42	28	38	22
Wednesday	8	6	15	37	34	32	31	19
Thursday	15	13	25	47	44	47	37	24
Friday	5	5	23	44	46	41	26	31
Saturday	13	5	21	26	32	37	38	18
Sunday	10	5	14	39	26	41	39	26