Beginning your own data engineering projects

Conf42: Python 2021

Hui Xiang Chua

What is data engineering?

To ensure consistent data flow for data scientists/ users, typically in the form of a data warehouse to enable large-scale data mining, analytics and reporting purposes

Typical process

Typical process

We want to automate this process as much as possible to save time and ensure consistency and accuracy.

Use case #1

Towards Data Science listing


towards
data science
ETL
Google Sheets
Data Studio

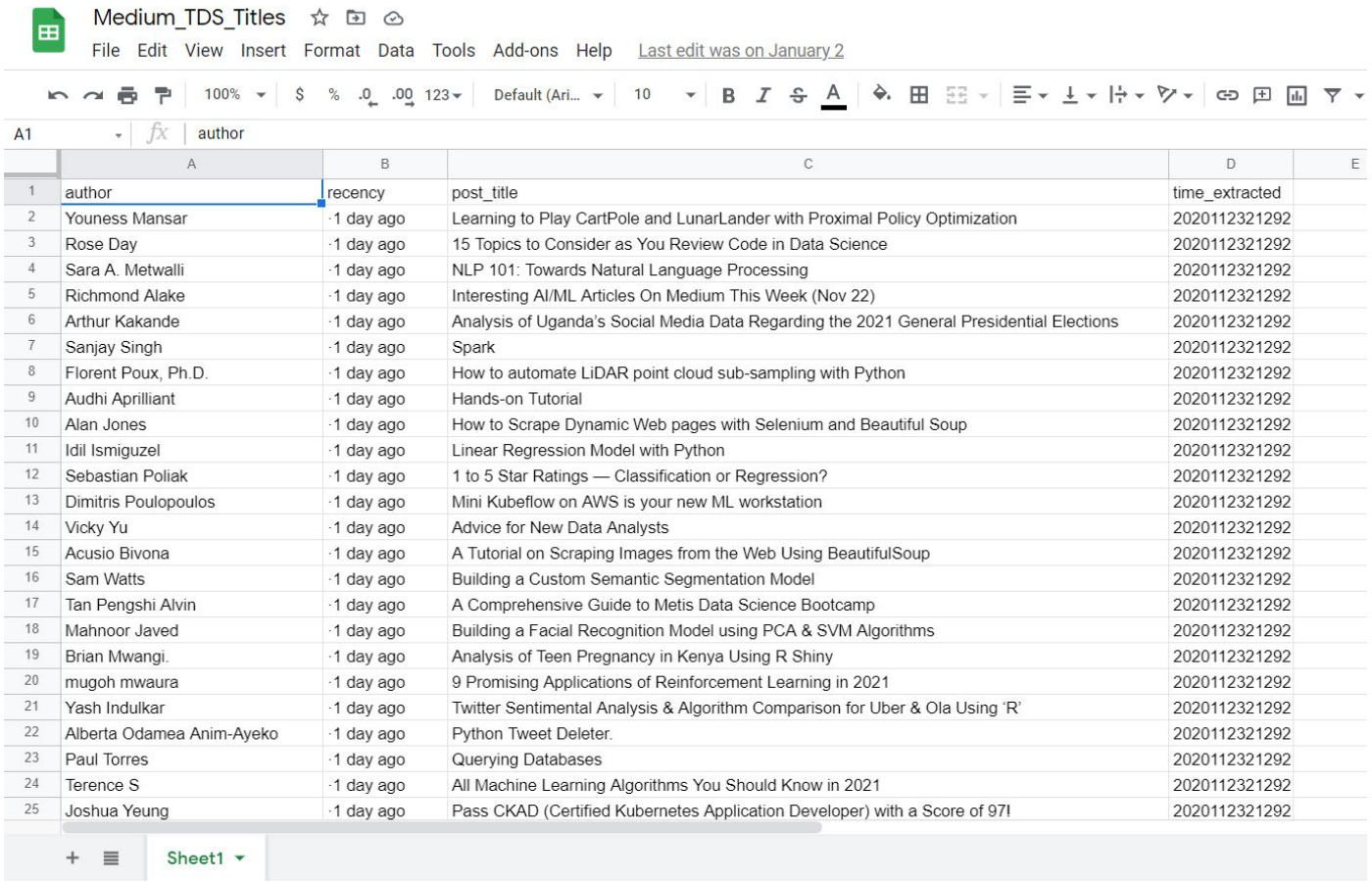
Cassie Kozyrkov · Pinned

Operationalization: the art and science of making metrics

Essential psychology for all data professionals

It's time to tackle a topic from psychology that's essential for all data professionals:

towards
data science



Medium_TDS_Titles

File Edit View Insert Format Data Tools Add-ons Help Last edit was on January 2

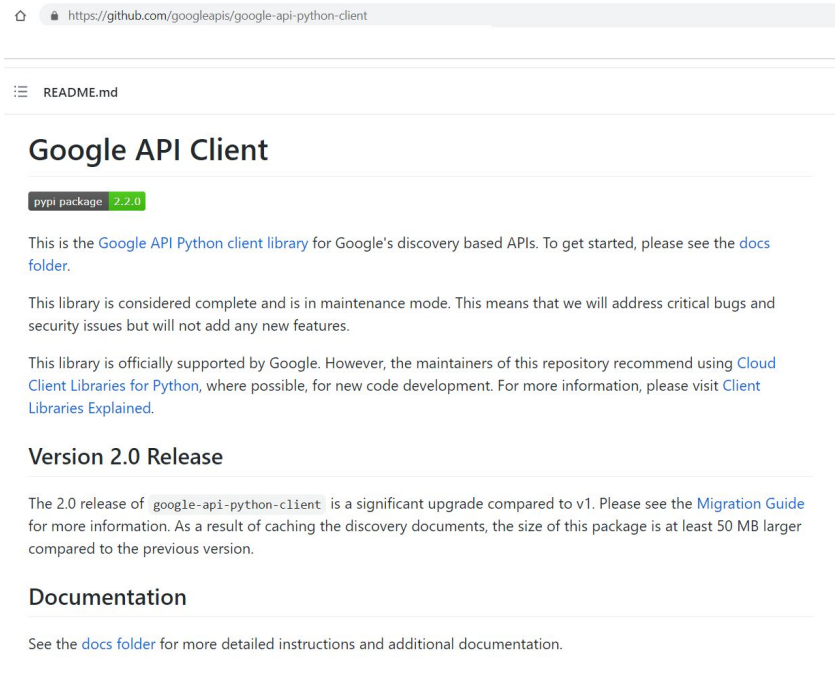
A1 author

	A	B	C	D	E
1	author	recency	post_title	time_extracted	
2	Youness Mansar	1 day ago	Learning to Play CartPole and LunarLander with Proximal Policy Optimization	20201123212920	
3	Rose Day	1 day ago	15 Topics to Consider as You Review Code in Data Science	20201123212920	
4	Sara A. Metwalli	1 day ago	NLP 101: Towards Natural Language Processing	20201123212920	
5	Richmond Alake	1 day ago	Interesting AI/ML Articles On Medium This Week (Nov 22)	20201123212920	
6	Arthur Kakande	1 day ago	Analysis of Uganda's Social Media Data Regarding the 2021 General Presidential Elections	20201123212920	
7	Sanjay Singh	1 day ago	Spark	20201123212920	
8	Florent Poux, Ph.D.	1 day ago	How to automate LiDAR point cloud sub-sampling with Python	20201123212920	
9	Audhi Aprilliant	1 day ago	Hands-on Tutorial	20201123212920	
10	Alan Jones	1 day ago	How to Scrape Dynamic Web pages with Selenium and Beautiful Soup	20201123212920	
11	Idil Ismiguzel	1 day ago	Linear Regression Model with Python	20201123212920	
12	Sebastian Poliak	1 day ago	1 to 5 Star Ratings — Classification or Regression?	20201123212920	
13	Dimitris Poulopoulos	1 day ago	Mini Kubeflow on AWS is your new ML workstation	20201123212920	
14	Vicky Yu	1 day ago	Advice for New Data Analysts	20201123212920	
15	Acusio Bivona	1 day ago	A Tutorial on Scraping Images from the Web Using BeautifulSoup	20201123212920	
16	Sam Watts	1 day ago	Building a Custom Semantic Segmentation Model	20201123212920	
17	Tan Pengshi Alvin	1 day ago	A Comprehensive Guide to Metis Data Science Bootcamp	20201123212920	
18	Mahnoor Javed	1 day ago	Building a Facial Recognition Model using PCA & SVM Algorithms	20201123212920	
19	Brian Mwangi.	1 day ago	Analysis of Teen Pregnancy in Kenya Using R Shiny	20201123212920	
20	mugoh mwaura	1 day ago	9 Promising Applications of Reinforcement Learning in 2021	20201123212920	
21	Yash Indulkar	1 day ago	Twitter Sentimental Analysis & Algorithm Comparison for Uber & Ola Using 'R'	20201123212920	
22	Alberta Odamea Anim-Ayeko	1 day ago	Python Tweet Deleter.	20201123212920	
23	Paul Torres	1 day ago	Querying Databases	20201123212920	
24	Terence S	1 day ago	All Machine Learning Algorithms You Should Know in 2021	20201123212920	
25	Joshua Yeung	1 day ago	Pass CKAD (Certified Kubernetes Application Developer) with a Score of 97!	20201123212920	

Sheet1

```
from __future__ import print_function
import numpy as np
import pandas as pd
import gspread_dataframe as gd
import gspread
from google.oauth2.service_account import Credentials
from bs4 import BeautifulSoup
from datetime import datetime
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
import time
```

To use Google API, follow the instructions here for installation and here for authentication. To use gspread and gspread_dataframe, please run pip install gspread and pip install gspread_dataframe.



https://github.com/googleapis/google-api-python-client

README.md

Google API Client

pypi package 2.2.0

This is the Google API Python client library for Google's discovery based APIs. To get started, please see the docs folder.

This library is considered complete and is in maintenance mode. This means that we will address critical bugs and security issues but will not add any new features.

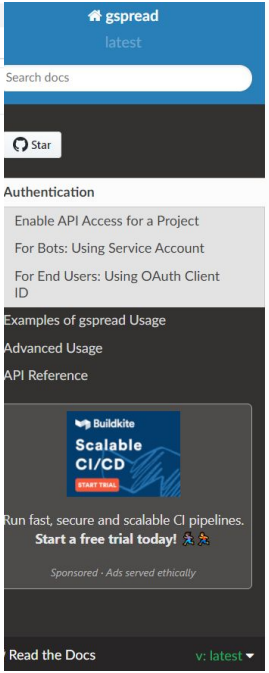
This library is officially supported by Google. However, the maintainers of this repository recommend using Cloud Client Libraries for Python, where possible, for new code development. For more information, please visit Client Libraries Explained.

Version 2.0 Release

The 2.0 release of `google-api-python-client` is a significant upgrade compared to v1. Please see the Migration Guide for more information. As a result of caching the discovery documents, the size of this package is at least 50 MB larger compared to the previous version.

Documentation

See the docs folder for more detailed instructions and additional documentation.



gspread

latest

Search docs

Star

Authentication

Enable API Access for a Project

For Bots: Using Service Account

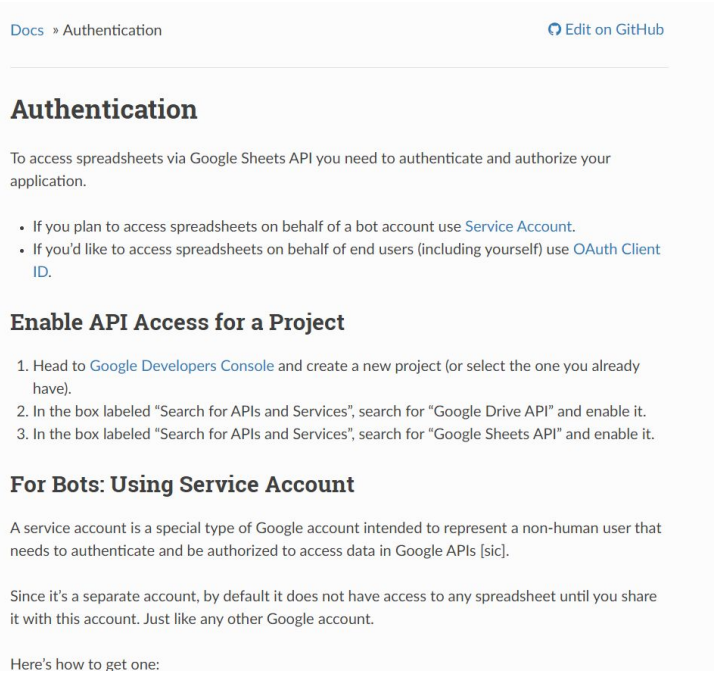
For End Users: Using OAuth Client ID

Examples of gspread Usage

Advanced Usage

API Reference

Read the Docs v: latest



Docs » Authentication

Edit on GitHub

Authentication

To access spreadsheets via Google Sheets API you need to authenticate and authorize your application.

- If you plan to access spreadsheets on behalf of a bot account use Service Account.
- If you'd like to access spreadsheets on behalf of end users (including yourself) use OAuth Client ID.

Enable API Access for a Project

1. Head to Google Developers Console and create a new project (or select the one you already have).
2. In the box labeled "Search for APIs and Services", search for "Google Drive API" and enable it.
3. In the box labeled "Search for APIs and Services", search for "Google Sheets API" and enable it.

For Bots: Using Service Account

A service account is a special type of Google account intended to represent a non-human user that needs to authenticate and be authorized to access data in Google APIs [sic].

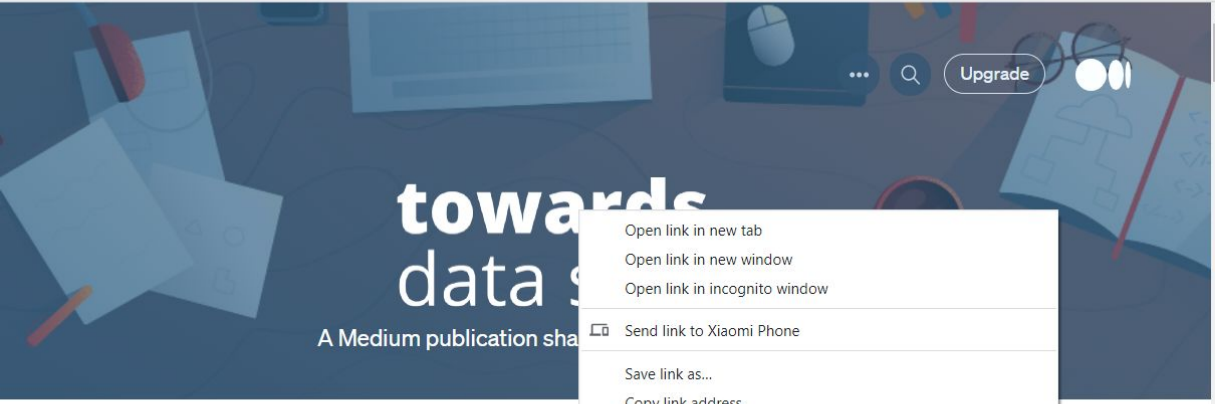
Since it's a separate account, by default it does not have access to any spreadsheet until you share it with this account. Just like any other Google account.

Here's how to get one:

```
# If modifying these scopes, delete the file token.pickle.
SCOPES = ['https://www.googleapis.com/auth/spreadsheets',
          'https://www.googleapis.com/auth/drive']

# The ID and range of a sample spreadsheet.
SAMPLE_SPREADSHEET_ID = '17e0aX_IcL6fF1GpUDaaaiGqCFijYorWtFrnKy7vUcUI' ##change this
#SAMPLE_RANGE_NAME = 'Sheet1!A:E' ##change this



credentials = Credentials.from_service_account_file('[redacted].json', scopes=SCOPES)
gc = gspread.authorize(credentials)

# Connecting with `gspread` here

ws = gc.open_by_key(SAMPLE_SPREADSHEET_ID).worksheet("Sheet1")
existing = gd.get_as_dataframe(ws)
#print(existing)

#scraping the site
options = Options()
options.add_argument('--headless')
options.add_argument('--disable-gpu')
driver = webdriver.Chrome('C:/Users/hxchu/env/Lib/site-packages/chromedriver',options=options)
site = "https://towardsdatascience.com/"
driver.get(site)
#time.sleep(3)

```

https://towardsdatascience.com

Other bookmarks | Reading list



Following | 585K Followers · Editors' Pic | About

Cassie Kozyrkov · Pinned

The AI reliability paradox

Too good to fail? The surprising way a top-performing system can hurt you

https://towardsdatascience.com/the-ai-reliability-paradox-ac8d1739206b?source=collection_home---------0----------

Open link in new tab
Open link in new window
Open link in incognito window
Send link to Xiaomi Phone
Save link as...
Copy link address
Copy Ctrl+C
Search Google for "The AI reliability paradox"
Print... Ctrl+P
AdBlock — best ad blocker
Google Translate
Inspect Ctrl+Shift+I

Elements | Console | Sources | Network | 1

```
<h1 id="96d4" class="ie if ig bj ih ii ij ik il im in io ip iq ir is it iu iv iw ah ix hw">
  <a class="cp dx" rel="noopener" href="/the-ai-reliability-paradox-ac8d1739206b?source=collection_home---------0----------">The AI reliability paradox</a> == $0
</h1>
```

Styles | Computed | Layout | Event Listeners | DOM Breakpoints | Properties | Accessibility

Filter :hov .cls

```
element.style {
}
.dx {                                   <style>
  text-decoration: none;
}
.cp {                                   <style>
  color: inherit;
}
a, button, input {                      (index):1
  -webkit-tap-highlight-color: transparent;
}
a {                                     (index):1
  color: inherit;
  text-decoration: none;
}
```

Console | What's New

Highlights from the Chrome 90 update

New CSS Flexbox debugging tools
Debug and inspect CSS Flexbox with the new CSS Flexbox debugging tools.

New Core Web Vitals overlay
Visualize page performance with the new Core Web Vitals overlay.

The image below shows how a DAG is a unidirectional, acyclic graph, where each node in the graph is a task and edges define dependencies among tasks. …

Read more · 4 min read

7

Tirthajyoti Sarkar · 10 hours ago ★

PyTest for Machine Learning — a simple

Show more

```
for i in range(20):
    try:
        loadMoreButton = driver.find_element_by_xpath("//button[text()='Show more']")
        #find_element_by_xpath("//button[contains(@aria-label,'Show more')]")
        time.sleep(2)
        loadMoreButton.click()
        time.sleep(2)
    except Exception as e:
        print(e)
        break
print("Complete")
time.sleep(10)
page = driver.page_source
driver.quit()

soup = BeautifulSoup(page, 'html.parser')

text_ls = []
for i in range(len(soup.find_all('a'))):
    text = soup.find_all('a')[i].get_text()
    text_ls.append((text))

# size of the list
size = len(text_ls)
index_ls = []
# looping till length - 2
for i in range(size - 2):

    # checking the conditions
    if text_ls[i] == text_ls[i + 1] and text_ls[i + 1] == text_ls[i + 2]:

        # printing the element as the
        # conditions are satisfied
        #print(i)
        index_ls.append((i))

```

```
text_ls
```

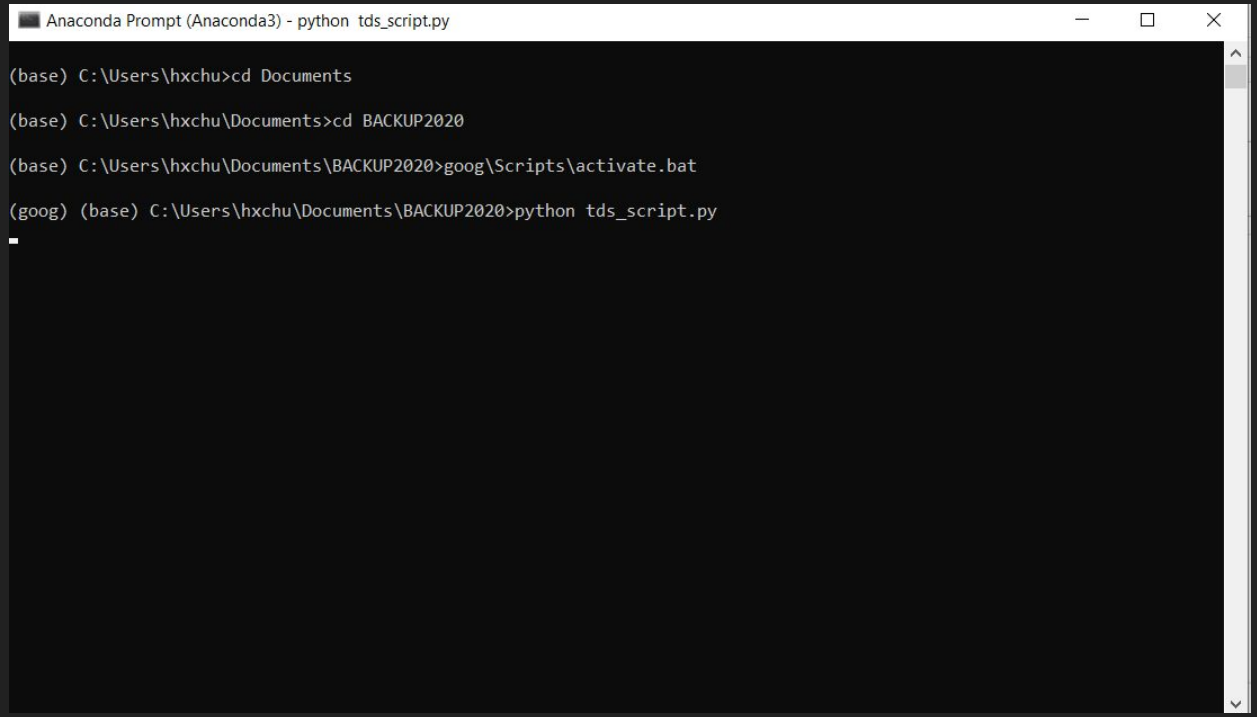
```
 '',
 'Paul May',
 '·9 hours ago',
 'Named Tuples: A Little Known Machine Learning Helper',
 'Glenn Carstens-Peters',
 'Unsplash',
 'Read more  · 8 min read',
 '',
 '',
 '',
 'Dean McGrath',
 '·11 hours ago',
 'How to Combine Python, Pandas & XlsxWriter',
 'Safar Safarov',
 'Unsplash',
 'XlsxWriter',
 'DataFrame',
 'pip',
 'Read more  · 3 min read',
 '',
```

```
data_ls = []
for i in index_ls[:-2]:
    name = text_ls[i+3]
    recency = text_ls[i+4]
    if(len(text_ls[i+5]) <20):
        title = text_ls[i+7]
    else:
        title = text_ls[i+5]
    data_ls.append((name,recency,title))

df = pd.DataFrame(np.array(data_ls))
df.columns = ['author','recency','post_title']

dt = datetime.now()
seq = int(dt.strftime("%Y%m%d%H%M%S"))
df['time_extracted']=seq


#appending with latest data
updated = existing.append(df)
print(updated)
updated = updated[['author','recency','post_title','time_extracted']]
updated = updated.drop_duplicates(['author','post_title'])
updated.fillna('', inplace=True)
updated.drop(updated.index[0], inplace=True)
#updated.dropna(axis=0,how='all',inplace=True)
ws.update([updated.columns.values.tolist()] + updated.values.tolist())



###References:
#https://stackoverflow.com/questions/23377533/python-beautifulsoup-parsing-table
#https://stackoverflow.com/questions/31328861/python-pandas-replacing-header-with-top-row
#https://www.graspingdata.tech/solved-the-caller-does-not-have-permission-using-the-api-with-a-private-goog
#https://gspread.readthedocs.io/en/latest/
```

```
data_ls = []
for i in index_ls[:-2]:
    name = text_ls[i+3]
    recency = text_ls[i+4]
    if(len(text_ls[i+5]) <20):
        title = text_ls[i+7]
    else:
        title = text_ls[i+5]
    data_ls.append((name,recency,title))
```

```
data_ls
```

```
[('Paul May',
  '·9 hours ago',
  'Named Tuples: A Little Known Machine Learning Helper'),
 ('Dean McGrath',
  '·11 hours ago',
  'How to Combine Python, Pandas & XlsxWriter'),
 ('Dan Baker',
  '·12 hours ago',
  'Deploying your Dash App to Heroku — THE MAGICAL GUIDE'),
 ('Shivangi Sareen', '·13 hours ago', 'Terminal Multiplexers to the Rescue'),
 ('Rashida Nasrin Sucky',
  '·14 hours ago',
  'A Complete Neural Network Algorithm from Scratch in Python'),
 ('Erdem Isbilen', '·14 hours ago', 'A Complete Guide to Python Lists'),
 ('Travis Tang (Voon Hao)',
  '·14 hours ago',
  'Dangers of a Pizza-Making Robot'),
 ('Brandon Smith',
  '·14 hours ago',
  'Type I and Type II Errors in COVID-19 Serology Testing')
```


Anaconda Prompt (Anaconda3) - python tds_script.py
(base) C:\Users\hxchu>cd Documents
(base) C:\Users\hxchu\Documents>cd BACKUP2020
(base) C:\Users\hxchu\Documents\BACKUP2020>goog\Scripts\activate.bat
(goog) (base) C:\Users\hxchu\Documents\BACKUP2020>python tds_script.py

https://stackoverflow.com/questions/44727232/scheduling-a-py-file-on-task-scheduler-in-windows-10

stackoverflow About Products For Teams Search...

Home

PUBLIC

Questions

Tags

Users

FIND A JOB

Jobs

Companies

TEAMS

Stack Overflow for Teams – Collaborate and share knowledge with a private group.

Free

38

Creating the exe should be the best method. But if you want to run it v
can do it in this way:

1. Launch Window's Task Scheduler
2.
3.

To e
uses
(64b
Pyth
scheduler:

Suppose the script you want to run is E:\My script.py. Instead of
instruct the task scheduler to run python.exe with the script as an argument. For example:

C:\Python27\ArcGIS10.2\python.exe "E:\My script.py"

The location of python.exe depends on your install. If you don't know where it is, you can discover its location; copy and paste the following code into a new Python script then execute the script. The script will print the location of python.exe as well as other information about your Python environment.

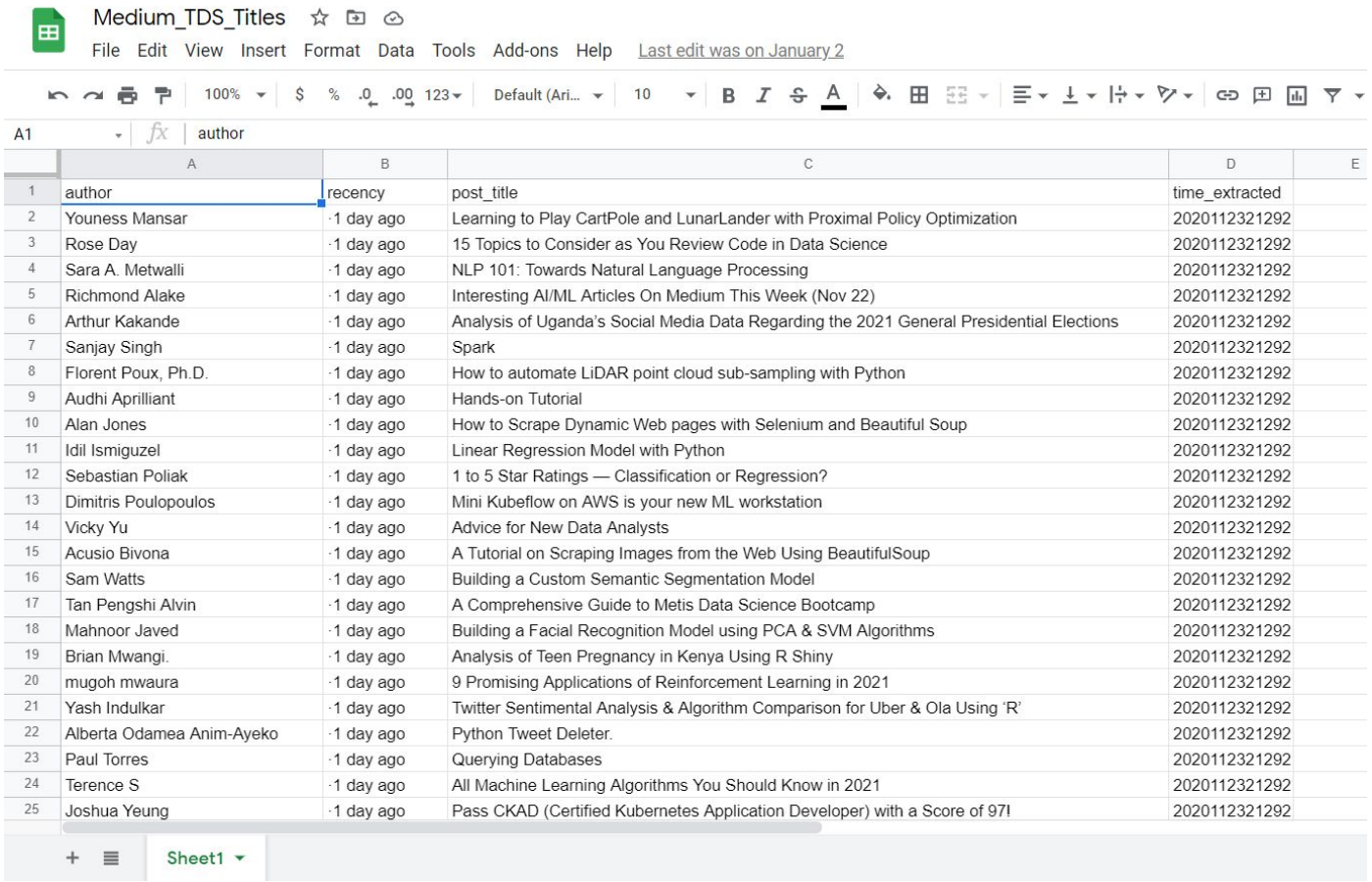
First, open the Control Panel and then click on the **Administrative Tools**:

Next, double-click on the **Task Scheduler**, and then choose the option to *'Create Basic Task…'*

After determining the location of python.exe, this is what is entered in the Action panel of the task

Type a name for your task (you can also type a description if needed), and then press Next.

towards
data science



Medium_TDS_Titles

File Edit View Insert Format Data Tools Add-ons Help Last edit was on January 2

A1 author

	A	B	C	D	E
1	author	recency	post_title	time_extracted	
2	Youness Mansar	1 day ago	Learning to Play CartPole and LunarLander with Proximal Policy Optimization	20201123212920	
3	Rose Day	1 day ago	15 Topics to Consider as You Review Code in Data Science	20201123212920	
4	Sara A. Metwalli	1 day ago	NLP 101: Towards Natural Language Processing	20201123212920	
5	Richmond Alake	1 day ago	Interesting AI/ML Articles On Medium This Week (Nov 22)	20201123212920	
6	Arthur Kakande	1 day ago	Analysis of Uganda's Social Media Data Regarding the 2021 General Presidential Elections	20201123212920	
7	Sanjay Singh	1 day ago	Spark	20201123212920	
8	Florent Poux, Ph.D.	1 day ago	How to automate LiDAR point cloud sub-sampling with Python	20201123212920	
9	Audhi Aprilliant	1 day ago	Hands-on Tutorial	20201123212920	
10	Alan Jones	1 day ago	How to Scrape Dynamic Web pages with Selenium and Beautiful Soup	20201123212920	
11	Idil Ismiguzel	1 day ago	Linear Regression Model with Python	20201123212920	
12	Sebastian Poliak	1 day ago	1 to 5 Star Ratings — Classification or Regression?	20201123212920	
13	Dimitris Poulopoulos	1 day ago	Mini Kubeflow on AWS is your new ML workstation	20201123212920	
14	Vicky Yu	1 day ago	Advice for New Data Analysts	20201123212920	
15	Acusio Bivona	1 day ago	A Tutorial on Scraping Images from the Web Using BeautifulSoup	20201123212920	
16	Sam Watts	1 day ago	Building a Custom Semantic Segmentation Model	20201123212920	
17	Tan Pengshi Alvin	1 day ago	A Comprehensive Guide to Metis Data Science Bootcamp	20201123212920	
18	Mahnoor Javed	1 day ago	Building a Facial Recognition Model using PCA & SVM Algorithms	20201123212920	
19	Brian Mwangi.	1 day ago	Analysis of Teen Pregnancy in Kenya Using R Shiny	20201123212920	
20	mugoh mwaura	1 day ago	9 Promising Applications of Reinforcement Learning in 2021	20201123212920	
21	Yash Indulkar	1 day ago	Twitter Sentimental Analysis & Algorithm Comparison for Uber & Ola Using 'R'	20201123212920	
22	Alberta Odamea Anim-Ayeko	1 day ago	Python Tweet Deleter.	20201123212920	
23	Paul Torres	1 day ago	Querying Databases	20201123212920	
24	Terence S	1 day ago	All Machine Learning Algorithms You Should Know in 2021	20201123212920	
25	Joshua Yeung	1 day ago	Pass CKAD (Certified Kubernetes Application Developer) with a Score of 97!	20201123212920	

Sheet1

This dashboard shows titles of articles published in Towards Data Science on Medium between a six-week period, 21 Nov 2020 and 2 Jan 2021. Data is scraped into a Google Sheet and then populated into tables here. This allows us to get a sense of all articles published within the day quickly. We can make use of this list to self-select articles that are of interest to us instead of solely depending on Medium's recommendations. And of course, we can also do text mining to identify popular entities/ topics/ keywords! Get to the article easily by copying the title and searching for it on Towards Data Science.

Prepared by Hui Xiang Chua. For feedback/ comments/ questions, pls write to datadoubleconfirm@gmail.com.

	author	Record Count ▾
1.	Soner Yıldırım	32
2.	Dimitris Poulopoulos	17
3.	Richmond Alake	15
4.	Terence S	14
5.	Dario Radečić	14
6.	Matt Przybyla	12
7.	Rose Day	11
8.	Emmett Boudreau	11
9.	Bharath K	10
10.	Kurtis Pykes	8

1 - 100 / 973

	author	post_title	time_extracted
1.	Salma Elshahawy, MSc.	How to Configure Github Actions the Easy Way.	20210102141454
2.	Sophia Yang	Jupyter workflow for data scientists	20210102141454
3.	Terence Shin	7 Most Recommended Skills to Learn in 2021 to be a Data Scientist	20210102141454
4.	Lauren Holzbauer	SQL Interview Prep: The Next Level	20210102141454
5.	Vinay Prabhu	A machine learning practitioner's tour of 10 under-appreciated PyPi pack...	20210102141454
6.	Yefeng Xia	Data analytics helps warehouse management	20210102141454
7.	Amanda West	Association Analysis Explained	20210102141454
8.	Renato Boemer	Trends in Data Science That Will Change Business Strategies	20210102141454
9.	Lev Maximov	Broadcasting in NumPy	20210102141454
10.	Samarth Agrawal	Understanding the Confusion Matrix from Scikit learn	20210102141454
11.	Jeroen van Zeeland	Shakespeare versus Eminem—who's the better lyricist?	20210102141454
12.	Vicky Yu	How to Present Machine Learning Results to Non-Technical People	20210102141454
13.	Yash Prakash	5 Step Guide to Setting Up a New Python Environment For Data Science	20210102141454
14.	Sun Weiran	Real-time Age, Gender and Emotion Prediction from Webcam with Keras...	20210102141454
15.	Nathan Pratt	The Baby Guessing Game	20210102141454

Use case #2
Creating a Blended face


EL
T

Battleground Singapore: Get to know your GE2020 candidates

The slate for the July 10 election is clear after Nomination Day. Find out who will be standing in your constituency and who the new faces are.

PUBLISHED: JUNE 30, 2020

Which party is standing where

People's Action Party

31 CONSTITUENCIES, **93** CANDIDATES

GRC

Chua Chu Kang | Nee Soon

See all

Progress Singapore Party

9 CONSTITUENCIES, **24** CANDIDATES

GRC

Nee Soon | West Coast

See all

Workers' Party

6 CONSTITUENCIES, **21** CANDIDATES

GRC

Sengkang | Marine Parade

See all

Singapore Democratic Party

5 CONSTITUENCIES, **11** CANDIDATES

GRC

Holland-Bukit Timah

See all

Peoples Voice

National Solidarity Party

Reform Party

Singapore People's Party

Meet the candidates

Aljunied 5-member GRC

WP


Pritam Singh


Sylvia Lim


Faisal Manap


Gerald Giam


Leon Perera

PAP


Victor Lye


Chua Eng Leong


Shamsul Kamar


NEW
Alex Yeo


NEW
Chan Hui Yuh

Ang Mo Kio 5-member GRC

PAP

```
from urllib.request import Request, urlopen
import requests
from bs4 import BeautifulSoup
import csv
import shutil
import re
import time
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
```

```
url =
"https://www.straitstimes.com/multimedia/graphics/2020/06/singapore-general-election-ge2020-
candidates/index.html"
options = Options()
options.add_argument('--headless')
options.add_argument('--disable-gpu')

driver = webdriver.Chrome(options=options)
driver.get(url)
time.sleep(3)
page = driver.page_source
driver.quit()
soup = BeautifulSoup(page, 'html.parser')
profile_pics = soup.find_all('img')
for i in range(31,369):
    if(profile_pics[i]['alt']!=''):
        pic_url = profile_pics[i]['src']
        pic_name = profile_pics[i]['alt'].replace(' ','_')
        pic_name = pic_name.replace('"','_')
        r = requests.get(pic_url, stream=True, headers={'User-agent': 'Mozilla/5.0'})
        if r.status_code == 200:
            with open(pic_name+pic_url[-4:], 'wb') as f:
                r.raw.decode_content = True
                shutil.copyfileobj(r.raw, f)
           print(i)
```

github.com/johnwmillr/Facer

Other bookmarks | Reading list

Search or jump to... / Pull requests Issues Marketplace Explore

johnwmillr / Facer

Watch 4 | Star 50 | Fork 9

Code | Issues 2 | Pull requests | Actions | Projects | Wiki | Security | Insights

master | 1 branch | 0 tags | Go to file | Add file | Code

johnwmillr Merge pull request #7 from veggiedefender/master ... 73ac53b on Sep 27, 2020 30 commits

facer	Only add image to warped/incremental if saving gif	7 months ago
tests	Add preliminary support for GIF creation	2 years ago
.gitignore	Add preliminary support for GIF creation	2 years ago
AverageFaces_RapRockCountry.png	Add example image to README	12 months ago
README.md	Add example image to README	12 months ago
requirements.txt	Update requirements.txt	12 months ago
setup.py	Add setup.py	2 years ago

README.md

About

Simple (👌) face averaging (🙂) in Python (🐍)

opencv | image-processing | face-detection | dlib | face-average | facer

Readme

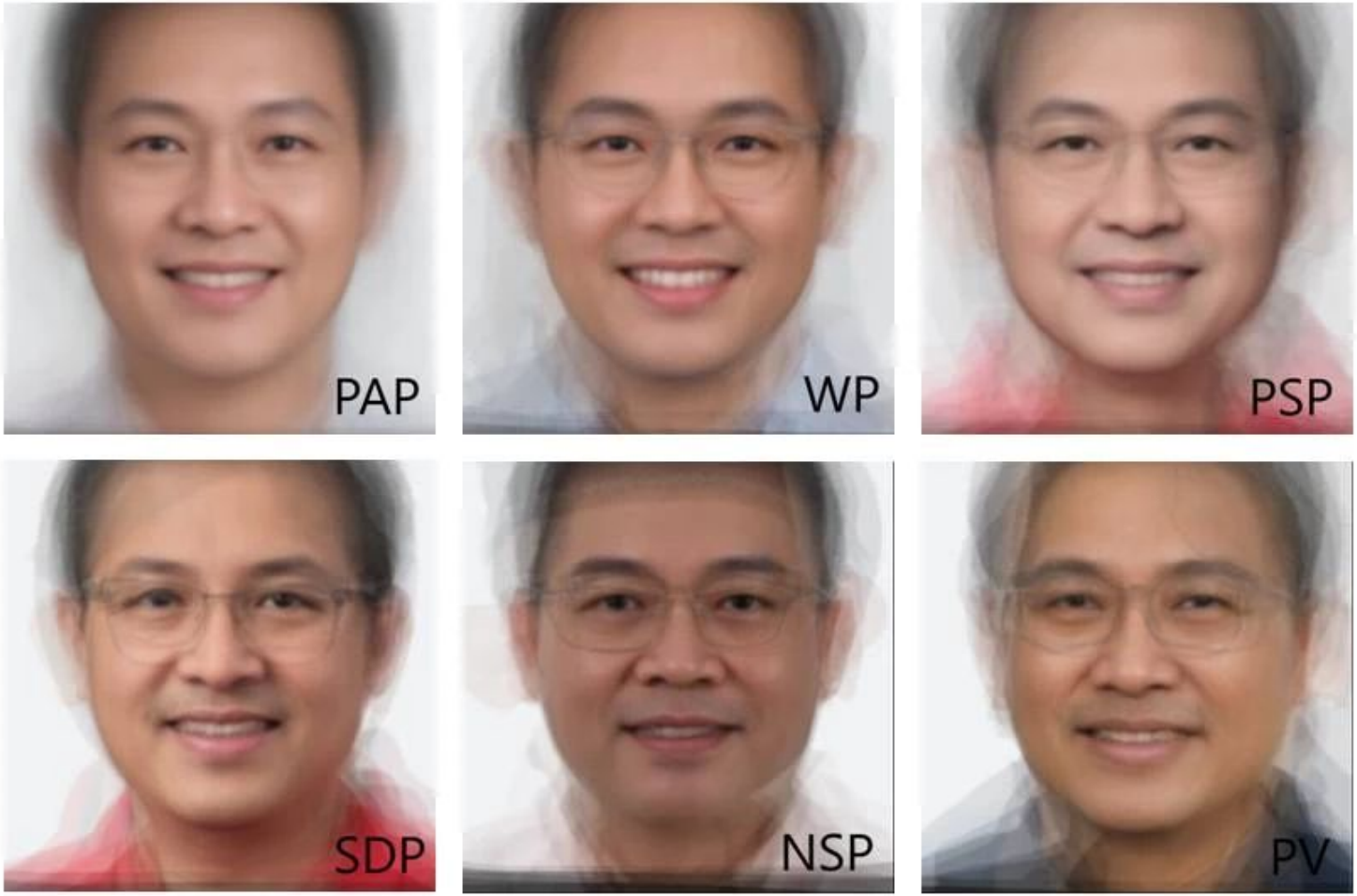
Releases

No releases published

Packages

No packages published

```
1  import cv2
2  import dlib
3  import matplotlib.pyplot as plt
4  from matplotlib import animation
5  import numpy as np
6  import math
7  import os
8  import glob
9
10 from facer.utils import similarityTransform, constrainPoint, calculateDelaunayTriangles, warpTriangle
11
```


PAP
WP
PSP
SDP
NSP
PV

Wrapping up

- Think of what analysis you/ your data science team want(s) to do
- Think about what data is necessary
- Write a script to automate the ETL (or ELT) process

DATA DOUBLE CONFIRM

DATA DIVE DAYS

Process and product of various data science tasks—from data collection, data preparation, data visualization, to basic statistical analysis and modelling. Datasets for practice available.

Selected as Top 100 Data Science Resources for 2018/2019 on MastersInDataScience.com

Connect with me

linkedin.com/in/hui-xiang-chua/

@hxchuaruns

projectosyo.wixsite.com/datadoubleconfirm