

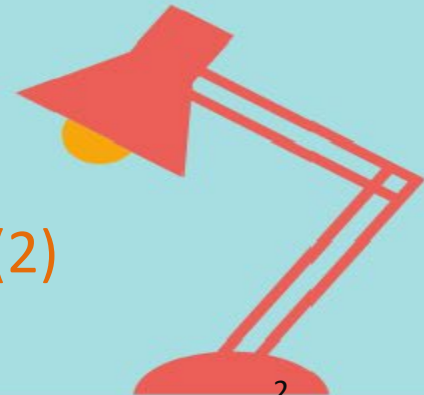
A stylized illustration of a row of books on a shelf. The books are in various colors (white, red, blue, yellow) and some have decorative patterns like stripes or a diamond. They are arranged in a slightly staggered manner.

2023-Spring Advanced Computer Programming (7)

CSIE, Asia Univ.

Course schedule

- W1-Introduction
- W2-Python libraries
- W3-BeautifulSoup(1)
- W4-BeautifulSoup(2)
- W5-
- W6-Scrapy(2)
- W7-Scrapy
- W8-Project development(1)
- W9-Midterm presentation
- W10-Web & HTTP
- W11-Flask
- W12-Flask Routes
- W13-Jinja template
- W14-Flask-form
- W15-Flask-mail
- W16-REST API
- W17-Project development(2)
- W18-Final presentation



Assignment 1

- Overview:
 - In this assignment, we will use the basic python web scraping tools urllib and BeautifulSoup to scrape data from the focustaiwan website. Please list the scraped news content and submit it to Tronclass's assignment entry.
- Objectives:
 - Learn how to obtain **the content** of a web page using web scraping.
 - To explore real html files.
 - Reflect on the possible uses of web scraping capabilities for data science.
- Instructions:
 - Go to the focustaiwan website using any browser.
<https://focustaiwan.tw/>
 - Check the tags in the html content.



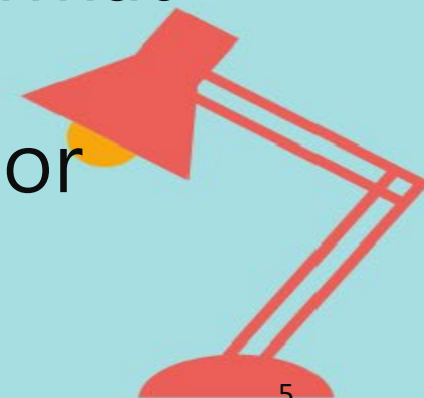
Assignment 2

- Overview:
 - In this assignment, we will use BeautifulSoup to grab all the news content from the cna website and save it to a txt file to submit to Tronclass's job entry.
- Target:
 - Learn how to **get links** to web pages using web scraping.
 - Explore crawling an entire site.
 - Reflect on possible uses of web scraping in data science.
- instruct:
 - Visit the focustaiwan website using any browser.
<https://focustaiwan.tw/>
 - Check tags in html content.



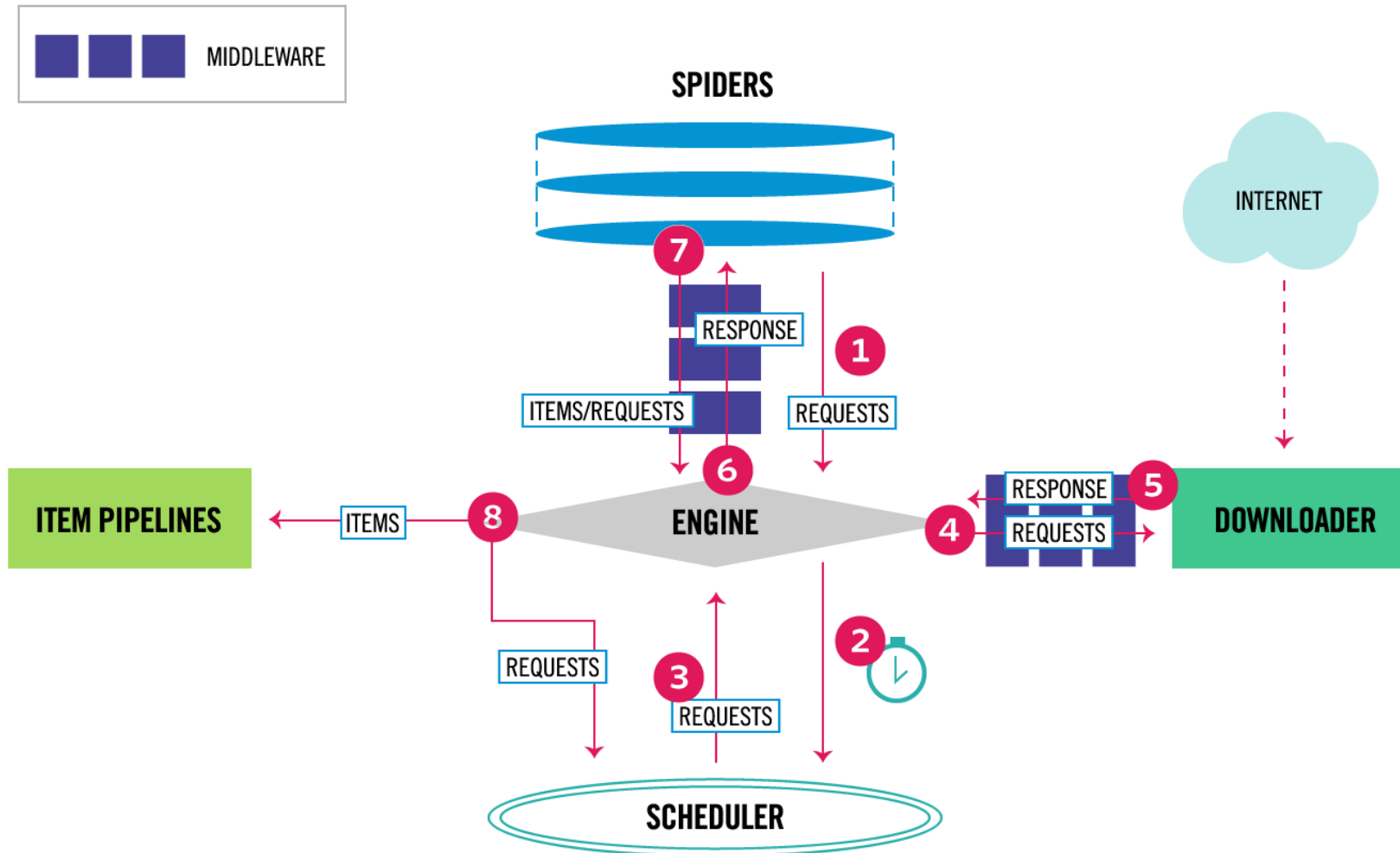
Midterm Report

- Crawling the webpage of the Department of CSIE
- https://csie.asia.edu.tw/en/associate_professors_2
- Read each teacher's specialty / Discipline expertise
- Save as a text file, which can be in txt, csv or json format
- The program part can be submitted as a notebook or py file



Scrapy crawler components

Scrapy Engine
Scheduler
Downloader
Spider
Item Pipeline
Downloader Middlewares
Spider Middlewares



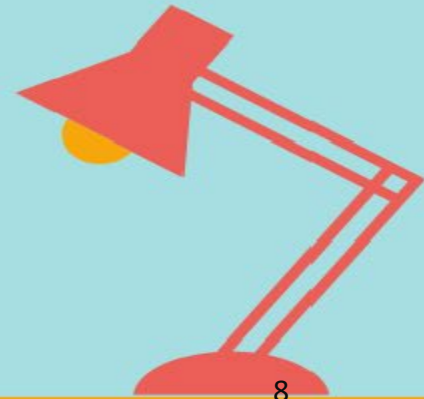
Scrapy steps

- Creating a new Scrapy project
- Writing a spider to crawl a site and extract data
- Exporting the scraped data using the command line
- Changing spider to recursively follow links
- Using spider arguments



A Minimalist End-to-End Scrapy Tutorial

- This repo contains the code for my tutorial: A Minimalist End-to-End Scrapy Tutorial (<https://medium.com/p/11e350bcdec0>).
- The website to crawl is <http://quotes.toscrape.com>.



toscrape example

Quotes to Scrape

[Login](#)

"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."

by [Albert Einstein](#) (about)

Tags: [change](#) [deep-thoughts](#) [thinking](#) [world](#)

"It is our choices, Harry, that show what we truly are, far more than our abilities."

by [J.K. Rowling](#) (about)

Tags: [abilities](#) [choices](#)

"There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle."


by [Albert Einstein](#) (about)

Tags: [inspirational](#) [life](#) [live](#) [miracle](#) [miracles](#)



Top Ten tags

[love](#)[inspirational](#)[life](#)[humor](#)[books](#)[reading](#)[friendship](#)[friends](#)[truth](#)[simile](#)

Collecting data with Scrapy



Topic-wise Practice C++ Java Python Competitive Programming Machine Learning Web Development SDE Sheet Puzzles



Collecting data with Scrapy

Last Updated : 08 Sep, 2021

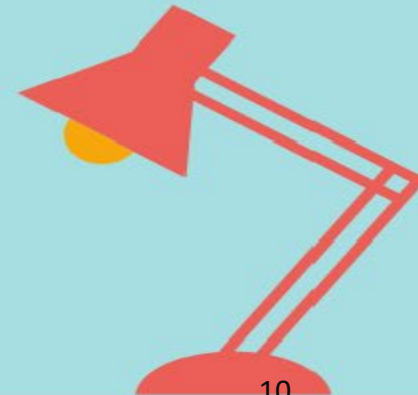
Prerequisites:

- [Scrapy](#)
- [SQLite3](#)

Scrapy is a web scraping library that is used to scrape, parse and collect web data. Now once our spider has scraped the data then it decides whether to:

- Keep the data.
- Drop the data or items.
- stop and store the processed data items.

Hence for all these functions, we are having a **pipelines.py** file which is used to handle scraped data through various components (known as a **class**) which are executed sequentially. In this article, we will be learning through the pipelines.py file, how it is used to **collect the data** scraped by scrapy using SQLite3 database language.



SQLite

- SQLite is a small database management system, and one other inclusive database is in the C++ library. The management system of the database is not the same, SQLite is not correct, and the client / server is quite good.
- SQLite Implementation SQL standard. Other used dynamic, weakly typographic SQL syntax. Other-made inset type database, application program, web page browser, hometown / client end-of-life documentary selection. It is possible to use a database engine, a working system, and an embedded system.
- SQLite Yes D. Richard Hipp erected.



Python sqlite3 -SQLite DB-API 2.0

```
import sqlite3

con = sqlite3.connect('example.db')

cur = con.cursor()

# Create table

cur.execute('CREATE TABLE stocks (date text, trans text, symbol text, qty real, price real)')

# Insert a row of data

cur.execute("INSERT INTO stocks VALUES ('2006-01-05','BUY','RHAT',100,35.14)")

# Save (commit) the changes

con.commit()

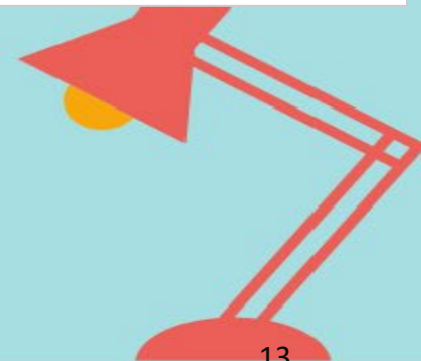
# We can also close the connection if we are done with it.
# Just be sure any changes have been committed or they will be lost.

con.close()
```

Meta Commands

```
系統管理員: Anaconda Prompt (Anaconda3)
(base) D:\Python\crawler>sqlite3
SQLite version 3.36.0 2021-06-18 18:36:39
Enter ".help" for usage hints.
Connected to a transient in-memory database.
Use ".open FILENAME" to reopen on a persistent database.
sqlite> .quit
(base) D:\Python\crawler>
```

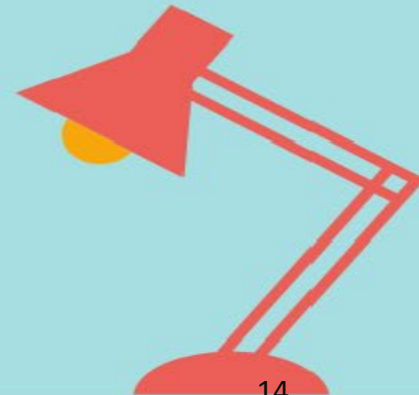
Command	Description
.show	Displays current settings
.databases	Provides database names and files
.quit	Quit sqlite3 program
.tables	Show current tables
.schema	Display schema of table
.header	Display the output table header
.mode	Select mode for the output table
.dump	Dump database in SQL text format



SQL(Structured Query Language) commands

- DDL-Data Definition Language
 - CREATE TABLE
 - DROP TABLE
 - ALTER TABLE
- DML-Data Manipulation Language
 - INSERT
 - UPDATE
 - DELETE
- DQL-Data Query Language
 - SELECT

case insensitive



SQL Examples

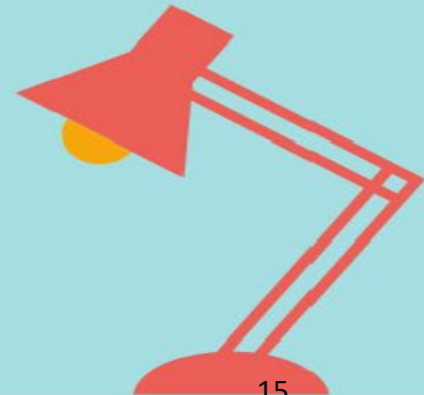
```
CREATE TABLE news (  
    news_id INTEGER NOT NULL PRIMARY KEY AUTOINCREMENT,  
    news_caption TEXT NOT NULL,  
    news_time DATETIME DEFAULT CURRENT_TIMESTAMP,  
    news_url TEXT NULL);
```

```
ALTER TABLE news ADD news_txt TEXT NULL;
```

```
INSERT INTO news (news_caption, news_url)  
VALUES ('N1', 'https://aaa.com/xxx111.aspx');
```

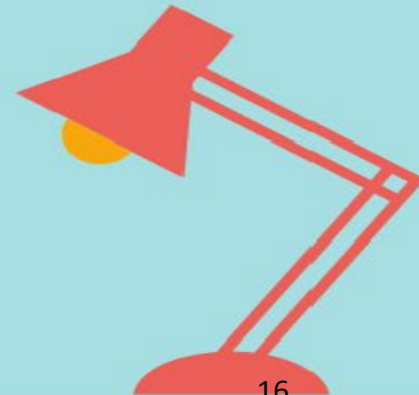
```
INSERT INTO news (news_caption, news_url)  
VALUES ('N1', 'https://aaa.com/xxx111.aspx');
```

```
SELECT * FROM news;
```




PythonAnywhere – Free online Python execution environment

- Free account is limited as follows:
 - Only one App (Application) can be created
 - Off-net access to the Internet is limited
 - CPU and storage are limited (100 seconds of CPU time a day, 512MB of storage)
 - Does not provide Jupyter (but does have IPython)
 - There can only be two Consoles (Bash and Python)



Your teacher

 pythonanywhere

Dashboard Consoles Files Web Tasks Databases

Your email address has been confirmed. You can manage your email preferences from this page. ×

[Upgrade/Downgrade Account](#) [Security](#) [Email](#) [Education](#) [API Token](#) [System Image](#)

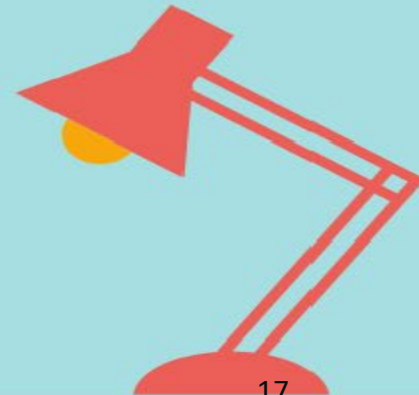
As part of our [Education beta](#), users can now nominate another user to be their "teacher". [Find out more.](#)

Your teacher

You can set who your teacher is below. **Warning!** this means they have [full access](#) to all your consoles, files and folders on PythonAnywhere, so you should make sure it's someone you trust!

You can revoke their access at any time by resetting the input field below. Blank means you don't have a teacher.

× ✓ ✗




System Image

System Image

The system image for your account determines the versions of Python that you can use and the packages that are pre-installed. [This page](#) shows the packages that are installed in each system image and in each version of Python.


Changing your system image may mean that you need to change your code. It may result in changing of your default python3 and/or editor Run button python versions. Please do not change it without reading [this help page](#) first.


Current system image: fishnchips 

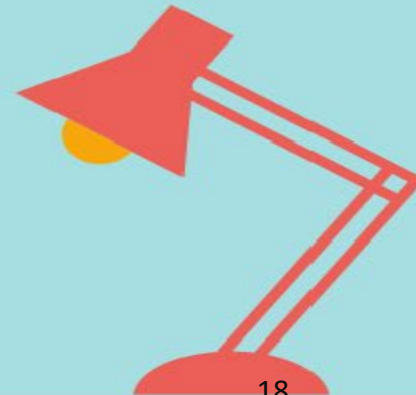
Default Python Versions

You need to upgrade your system image to be able to change your default python command.


python should run: 2.7

python3 should run: 3.8 

The editor Run button should use: 3.8 



Dashboard

 pythonanywhere

Dashboard Consoles Files Web Tasks Databases

Up here you will see instructions walking you through the key features of our Education beta.

(If after closing this helper, you want to go through it again – or try another one – go to the [Help page](#))

→

Close this tutorial

Dashboard

Welcome, [htchu](#)

CPU Usage: 0% used – 0.00s of 100s. Resets in 10 hours, 4 minutes [More Info](#)

File storage: 16% full – 80.1 MB of your 512.0 MB quota [More Info](#)

[Upgrade Account](#)

Recent Consoles

+ 5 -

You have no recent consoles.

[View all](#)

New console:

\$ Bash

>>> Python ▾

[More...](#)

Recent Files

+ 5 -

[/home/htchu/mysite/flask_app.py](#)

[/home/htchu/mysite3/urls.py](#)

[/home/htchu/mysite3/polls/urls.py](#)

[/home/htchu/mysite3/polls/views.py](#)

[/home/htchu/mysite3/manage.py](#)

[+ Open another file](#)

[Browse files](#)

Recent Notebooks

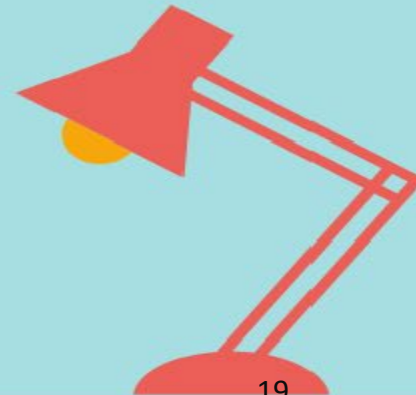
+ 5 -

Your account does not support Jupyter Notebooks. [Upgrade your account](#) to get access!


All Web apps

htchu.pythonanywhere.com

[Open Web tab](#)




Command Console

 pythonanywhere

Dashboard **Consoles** Files Web Tasks Databases

Up here you will see instructions walking you through the key features of our Education beta.


(If after closing this helper, you want to go through it again – or try another one – go to the [Help page](#))




Close this tutorial

CPU Usage: 0% used – 0.00s of 100s. Resets in 16 hours, 2 minutes [More Info](#)

Start a new console:

Python: [3.8](#) / [3.7](#) / [3.6](#) / [3.5](#) / [2.7](#) IPython: [3.8](#) / [3.7](#) / [3.6](#) / [3.5](#) / [2.7](#) PyPy: [2](#) / [3](#)
Other: [Bash](#) | [MySQL](#)
Custom: 


Your consoles:

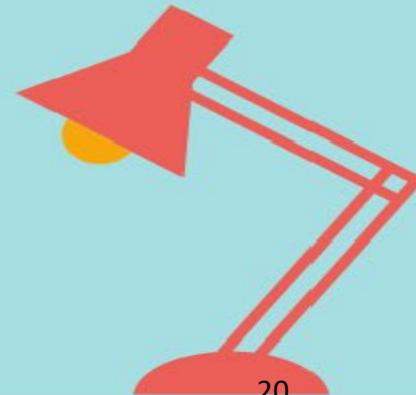
[flask_app.py \(Python3.8\)](#) 

Consoles shared with you


No-one has shared any consoles with you :-(

Running processes

 [Fetch process list](#)



Files

 pythonanywhere


Dashboard Consoles **Files** Web Tasks Databases

Up here you will see instructions walking you through the key features of our Education beta.

(If after closing this helper, you want to go through it again – or try another one – go to the [Help page](#))








→

Close this tutorial

/home/  htchu [Open Bash console here](#) **16% full** – 80.1 MB of your 512.0 MB quota [More Info](#)





























Directories

Enter new directory name [New directory](#)

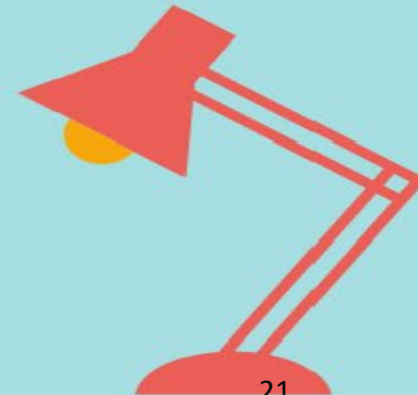
.cache/	
.config/	
.local/	
.virtualenvs/	
mysite/	
mysite2/	
mysite3/	

Files


Enter new file name, eg hello.py [New file](#)

 .bashrc	  	2021-05-04 07:38	559 bytes
 .gitconfig	  	2021-05-04 07:38	266 bytes
 .profile	  	2021-05-04 07:38	79 bytes
 .python_history	  	2021-05-04 08:18	7 bytes
 .pythonstartup.py	  	2021-05-04 07:38	77 bytes
 .vimrc	  	2021-05-04 07:38	4.6 KB
 README.txt	  	2021-05-04 07:38	232 bytes

[Upload a file](#)
100MiB maximum size




Web

 pythonanywhere

Dashboard Consoles Files **Web** Tasks Databases

Up here you will see instructions walking you through the key features of our Education beta.

(If after closing this helper, you want to go through it again – or try another one – go to the [Help page](#))




Close this tutorial

htchu.pythonanywhere.com

+ Add a new web app

Configuration for htchu.pythonanywhere.com


Reload:



Best before date:

We're happy to host your free website – and keep it free – for as long as you want to keep it running, but you'll need to log in at least once every three months and click the "Run until 3 months from today" button below. We'll send you an email a week before the site is disabled so that you don't forget to do that. [See here for more details.](#)


This site has expired, click the button below to reactivate it



Paying users' sites stay up forever without any need to log in to keep them running.



Tasks

 pythonanywhere

Dashboard Consoles Files Web **Tasks** Databases

Up here you will see instructions walking you through the key features of our Education beta.

(If after closing this helper, you want to go through it again – or try another one – go to the [Help page](#))

→

Close this tutorial

CPU Usage: 0% used – 0.00s of 100s. Resets in 15 hours, 58 minutes [More Info](#)

Scheduled tasks

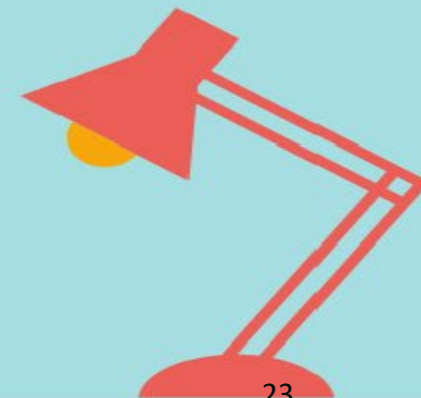
Server time: 02:15 UTC

Daily, at : UTC [Create](#)

Frequency	Time	Command	Description	Expiry	Actions
You have no tasks yet.					

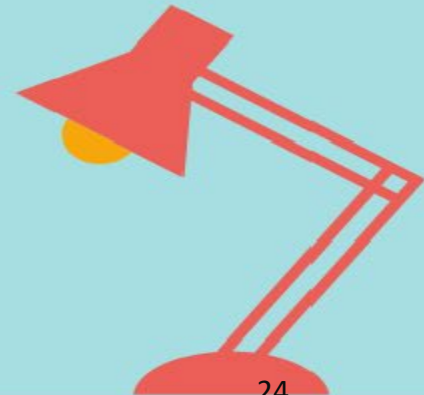
Your **Scheduled task** is a script that will run every day at a time of your choosing – you can use it to do stuff like scraping websites, or checking that your server is running.

[Paying users](#) can schedule several tasks, can run them both hourly and daily, and they never expire. Just sayin'...



Activity-1 (S-S)

- Open the Google Jamboard for the class
- Discuss what you know (K), want to know (W), and have learned (L) about this topic in the KWL chart.





Thanks!

Q&A

