



Web Crawler Practice

Selenium



Dr. Chun-Hsiang Chan

Department of Geography,
National Taiwan Normal University



Outline

- Introduction
- Selenium
- Web Driver
- Element Indexing
- Assignment

Introduction

- Before we introduce the static web crawler, you have to know these fundamental knowledge in facilitating the understanding how the network operates.

Selenium

- Last week, we have already introduced about static web crawler, which is pretty easy and straightforward for you to conduct into data collection process.
- However, in most cases, you cannot obtain data through static web crawler because you have to give some search conditions or scroll the web page in order to see other information, which does not initially show in both screen and source code.
- To solve this dilemma, we introduce a powerful Python package, Selenium, provides several functions that automatically mocks the mouse and keyboard in your browser.

Selenium

- Basically, we call this approach as dynamic web crawler.
- The architecture and procedure are depicted as follows,
 - Open a web browser
 - Enter the account and password
 - Give the search conditions
 - Scroll to the location of targeted information
 - Extract the source code of specific information

Selenium Installation

Source: <https://www.selenium.dev/downloads/>



About Downloads Documentation Projects Support Blog English

Selenium Conf 2024 Call for Proposals is now open! Submissions close 30 April. [Learn more & submit](#)

Downloads

Below is where you can find the latest releases of all the Selenium components. You can also find a list of previous releases, source code, and additional information for Maven users.

Selenium Clients and WebDriver Language Bindings

In order to create scripts that interact with the Selenium Server (Remote WebDriver) or create local Selenium WebDriver scripts, you need to make use of language-specific client drivers.

While language bindings for [other languages exist](#), these are the core ones that are supported by the main project hosted on GitHub.



C#

Stable: [4.20.0 \(April 24, 2024\)](#)

[Changelog](#)
[API Docs](#)



Ruby

Stable: [4.20.1 \(April 25, 2024\)](#)

[Changelog](#)
[API Docs](#)



Java

Stable: [4.20.0 \(April 24, 2024\)](#)

[Changelog](#)
[API Docs](#)



Python

Stable: [4.20.0 \(April 24, 2024\)](#)

[Changelog](#)
[API Docs](#)



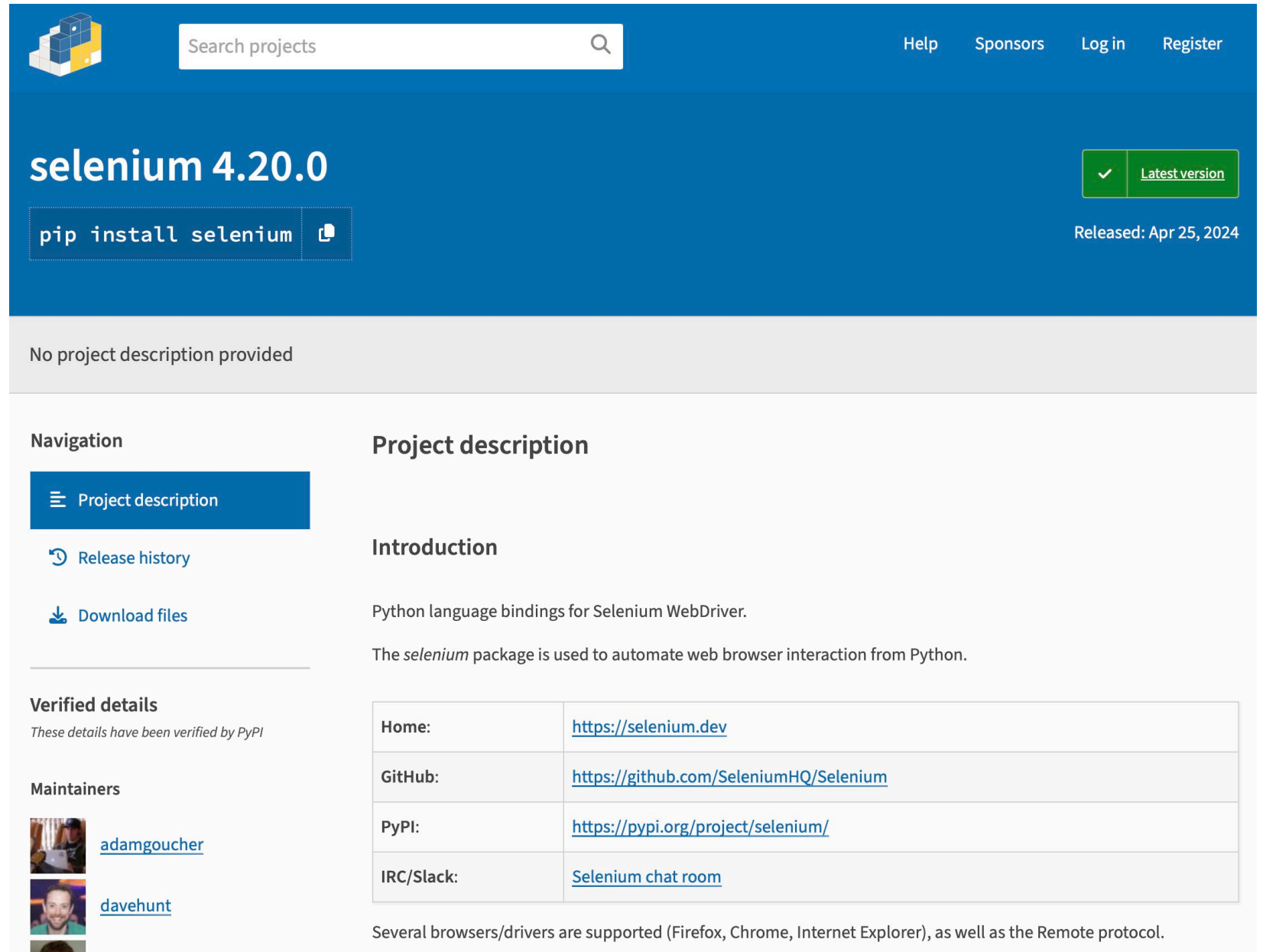
JavaScript

Stable: [4.20.0 \(April 24, 2024\)](#)

[Changelog](#)
[API Docs](#)

Display a menu

Selenium Installation



The screenshot shows the Selenium project page on PyPI. At the top, there is a search bar and navigation links for Help, Sponsors, Log in, and Register. The main header displays the project name 'selenium 4.20.0' with a 'Latest version' badge and a release date of 'Released: Apr 25, 2024'. Below this, a code block shows the installation command 'pip install selenium'. A message states 'No project description provided'. The page is divided into two columns: 'Navigation' and 'Project description'. The 'Navigation' column includes links for 'Project description', 'Release history', and 'Download files'. The 'Project description' column features an 'Introduction' section with text about Selenium WebDriver and a table of links for Home, GitHub, PyPI, and IRC/Slack. A 'Verified details' section is also present, along with a 'Maintainers' section listing 'adamgoucher' and 'davehunt'.

Search projects

Help Sponsors Log in Register

selenium 4.20.0

✓ Latest version

Released: Apr 25, 2024

```
pip install selenium
```

No project description provided

Navigation

- Project description
- Release history
- Download files

Project description

Introduction

Python language bindings for Selenium WebDriver.

The *selenium* package is used to automate web browser interaction from Python.



Home:	https://selenium.dev
GitHub:	https://github.com/SeleniumHQ/Selenium
PyPI:	https://pypi.org/project/selenium/
IRC/Slack:	Selenium chat room

Several browsers/drivers are supported (Firefox, Chrome, Internet Explorer), as well as the Remote protocol.

Verified details

These details have been verified by PyPI

Maintainers

-  [adamgoucher](#)
-  [davehunt](#)

Selenium Installation

Supported Python Versions

- Python 3.8+

Installing

If you have [pip](#) on your system, you can simply install or upgrade the Python bindings:

```
pip install -U selenium
```

Alternately, you can download the source distribution from *PyPI* <<https://pypi.org/project/selenium/#files>>, unarchive it, and run:

```
python setup.py install
```

Note: You may want to consider using [virtualenv](#) to create isolated Python environments.

現已提供深色模式
在這裡切換淺色或深色模式

關閉

為您推薦

- 唱歌與舞蹈
- 喜劇
- 運動
- 動漫卡通
- 關係
- 直播秀
- 對嘴
- 日常生活
- 美肌保養
- 遊戲
- 社會
- 穿搭
- 汽車
- 美食
- 動物
- 家庭

關注中

好友

探索

直播

個人資料

登入可關注創作者、對影片按讚以及查看評論。

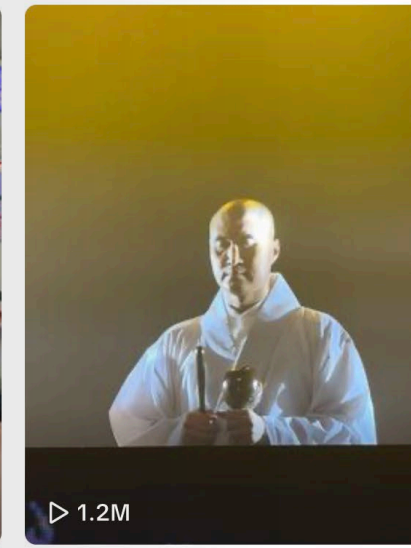
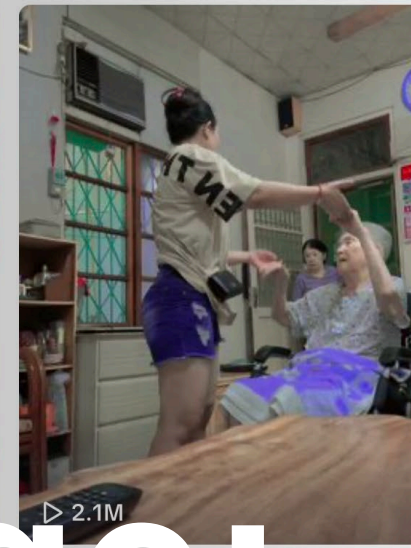
登入

創作 TikTok 特效，獲得獎勵

公司
關於 新聞編輯室 聯絡方式 工作

計劃
TikTok for Good 廣告
TikTok LIVE Creator Networks
Developers 透明度 TikTok 獎勵
TikTok Embeds

條款與政策
說明 安全 條款 隱私權政策
隱私權中心 Creator Academy
社群自律公約



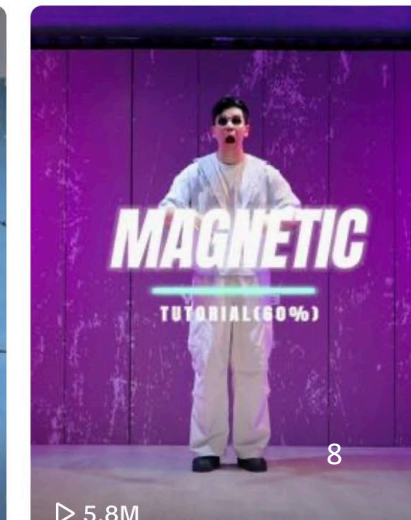
Chun-Hsiang Chan (2024) 2.8M

ggi releng teman 1.7M

metu mesti leren 2.1M

日進和尚來台super house#dj #和尚 #嘻哈 #夜店 #有趣 1.2M

#assalamualaikum #fyp #fypシ #tetepsemangat #tiktokhiburan 1.7M



Chun-Hsiang Chan (2024) 1.5M

跟我一起越過百慕達 1.2M

865.7K

784.5K

5.8M

Target Webpage!

TikTok

Web Driver

Drivers

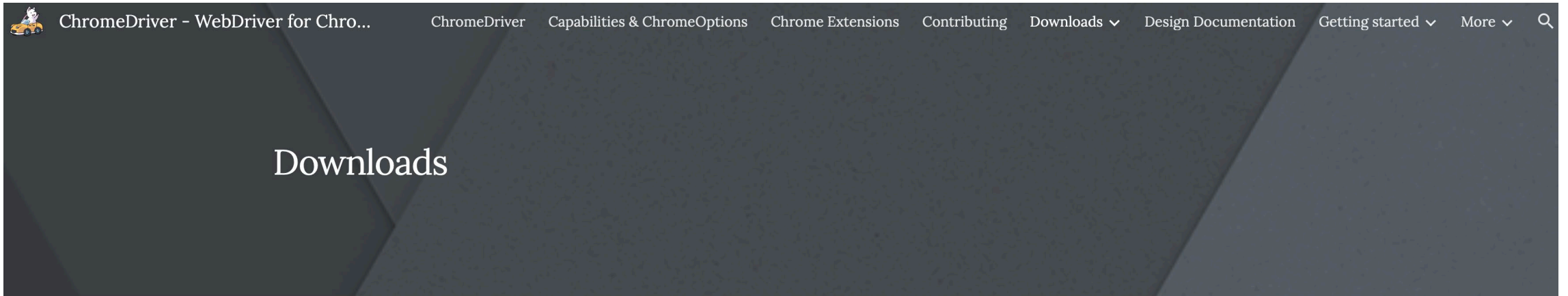
Selenium requires a driver to interface with the chosen browser. Firefox, for example, requires [geckodriver](#), which needs to be installed before the below examples can be run. Make sure it's in your *PATH*, e. g., place it in */usr/bin* or */usr/local/bin*.

Failure to observe this step will give you an error *selenium.common.exceptions.WebDriverException: Message: 'geckodriver' executable needs to be in PATH*.

Other supported browsers will have their own drivers available. Links to some of the more popular browser drivers follow.

Chrome:	https://chromedriver.chromium.org/downloads
Edge:	https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/
Firefox:	https://github.com/mozilla/geckodriver/releases
Safari:	https://webkit.org/blog/6900/webdriver-support-in-safari-10/

Web Driver



Current Releases

- **If you are using Chrome version 115 or newer, please consult [the Chrome for Testing availability dashboard](#). This page provides convenient [JSON endpoints](#) for specific ChromeDriver version downloading.**
- For older versions of Chrome, please see below for the version of ChromeDriver that supports it.

For more information on selecting the right version of ChromeDriver, please see the [Version Selection](#) page.

ChromeDriver 114.0.5735.90

Supports Chrome version 114

For more details, please see the [release notes](#).

ChromeDriver 114.0.5735.16

Supports Chrome version 114

For more details, please see the [release notes](#).

Initiation Stage

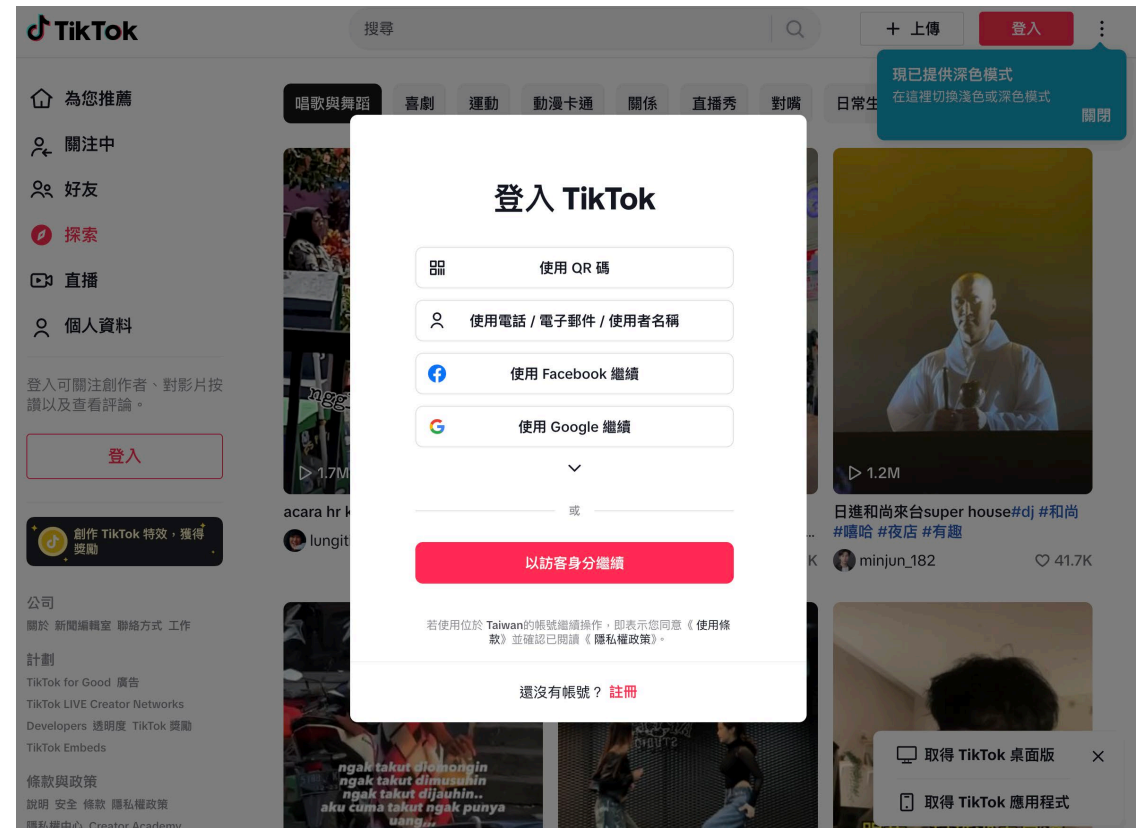
import packages

```
# import packages  
import os  
from selenium import webdriver  
from selenium.webdriver.common.by import By  
from selenium.webdriver.common.keys import Keys
```

Web Driver

- First of all, you need to start a webdriver ...

```
# start a webdriver  
browser = webdriver.Chrome()  
browser.get(url)
```



Element Indexing

- You may index the element by several methods, such as
 - Xpath
 - Name
 - ID
 - Class_name
 - Link_text
 - Tag_name
 - CSS_selector

```
# enter the webpage as a guest  
button_xpath = '/html/body/div[5]/div[3]/div/div/div/div[1]/div/div/div[3]/div/div[2]/div'  
guest_button = browser.find_element(By.XPATH, button_xpath)  
guest_button.click()
```


Element Action

- You are able to do lots of things to web elements, e.g.,
 - Send_keys
 - Click
 - Clear
 - Find_element
 - Get_attribute

```
# enter the webpage as a guest  
button_xpath = '/html/body/div[5]/div[3]/div/div/div/div[1]/div/div/div[3]/div/div[2]/div'  
guest_button = browser.find_element(By.XPATH, button_xpath)  
guest_button.click()
```



唱歌與舞蹈

喜劇

運動

動漫卡通

關係

直播秀

對嘴

日常 >



👍👍👍 #機器人舞 #機械舞

david.9786

2.8M



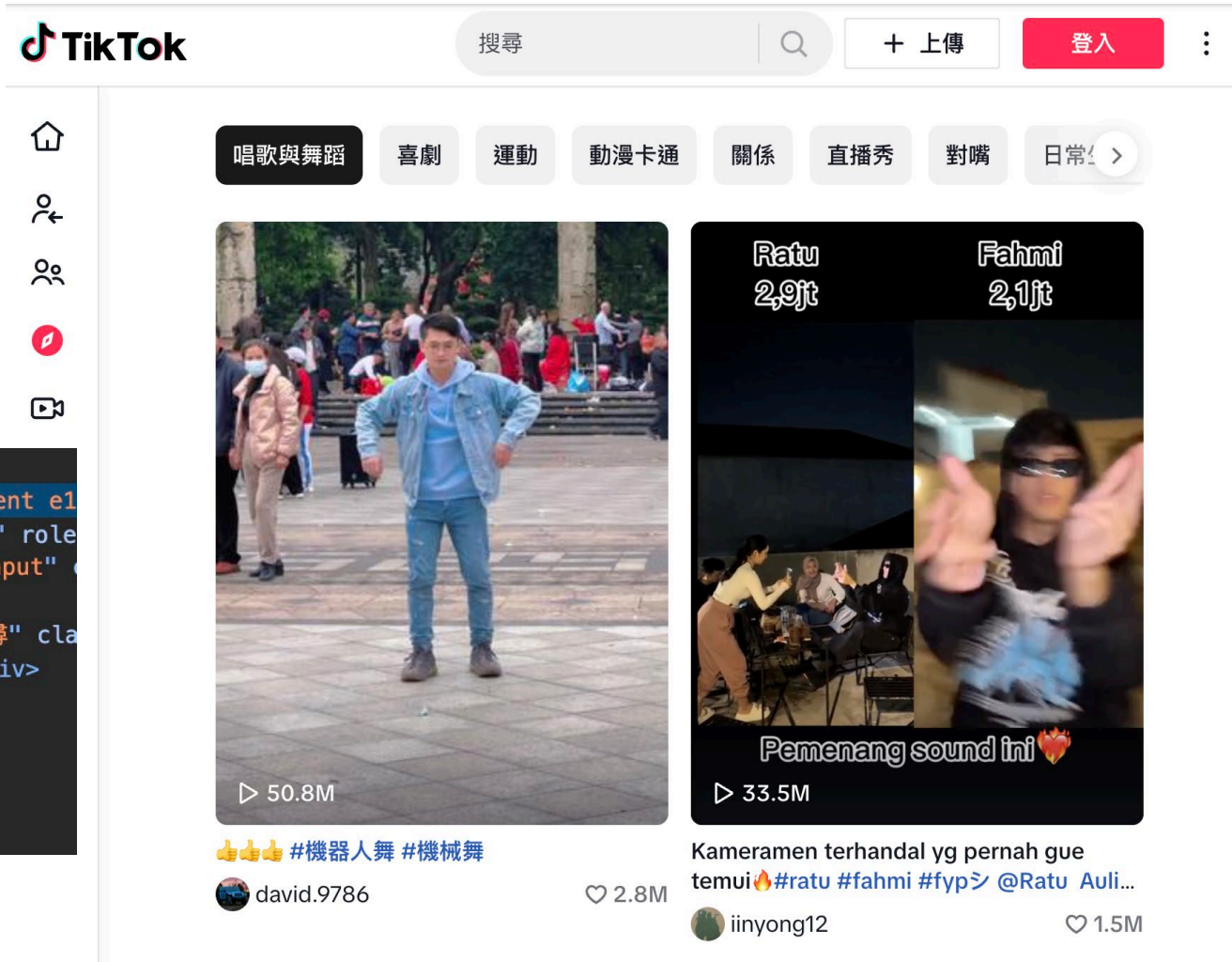
Kameramen terhandal yg pernah gue temui 🔥 #ratu #fahmi #fyp @Ratu Auli...

iinyong12

1.5M

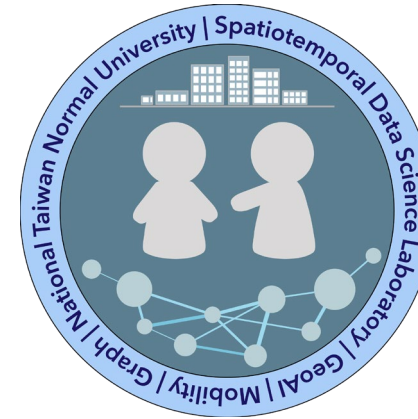
Search Textbox

```
<div class="css-1asq5wp-DivSearchFormContainer e1hi1cmj0">  
  <form data-e2e="search-box" class="search-input css-dhqzc6-FormElement e1  
    <input placeholder="搜尋" name="q" type="search" autocomplete="off" role  
      expanded="false" aria-autocomplete="list" data-e2e="search-user-input"  
    <span class="css-hck1rr-SpanSpliter e14ntknm6"></span>  
    <button data-e2e="search-box-button" type="submit" aria-label="搜尋" cla  
      ><div class="css-17iic05-DivSearchIconContainer e14ntknm8">⋮</div>  
    </button>  
    <div class="css-1mdii59-DivInputBorder e14ntknm1"></div>  
  </form>  
</div>  
</div>
```



Assignment

- Fetch top 100 post information from TikTok based on a keyword.
 - 1) Set a keyword
 - 2) Search information
 - 3) Get data (video link, view number, view date, video title, author ID, author link)
 - 4) Formulate into a CSV file



The End

Thank you for your attention!

Email: chchan@ntnu.edu.tw

Web: toodou.github.io

