

Group member:

Qinyu Zhao qinyuzha@andrew.cmu.edu
Kehuan Wang kehuanw@andrew.cmu.edu
Dan Li danl2@andrew.cmu.edu
Yile Jiang yilej@andrew.cmu.edu

Link to our processing video:

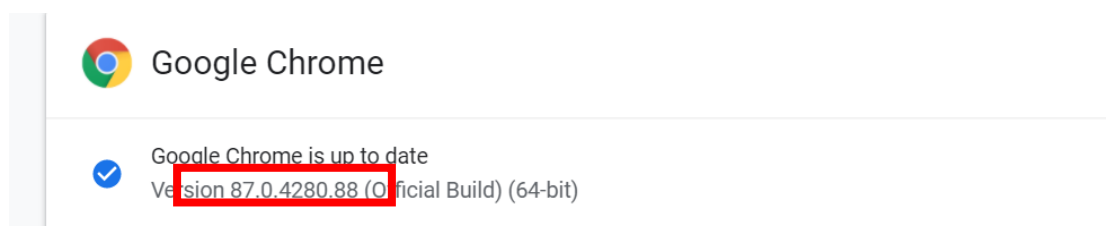
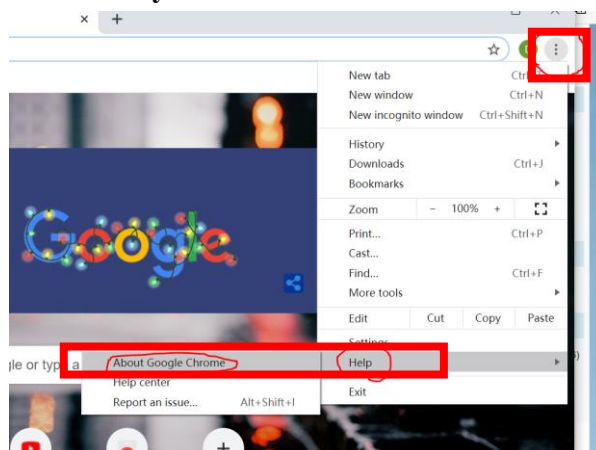
https://youtu.be/BFGWP0-kI_A

Step 1: install modules

```
pip install xlrd  
pip install xlwt  
pip install seaborn  
pip install scipy  
pip install selenium  
pip install wordcloud  
pip install jieba
```

Step 2: download WebDriver for chrome

1. Check your chrome version



2. Go to this website <https://chromedriver.chromium.org/downloads> to download the corresponding version.

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

ChromeDriver 87.0.4280.88

Supports Chrome version 87

- Resolved issue 3641: Page not getting loaded/rendered when browser window is not in focus with Chrome Beta v87 and chromedriver v(87/86)
- Resolved issue 3657: Screenshot background browser timed out

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

3. Unzip ChromeDriver.zip to get ChromeDriver.exe: move ChromeDriver.exe somewhere that Python and Selenium will be able to find it

mac: move to /usr/local/bin

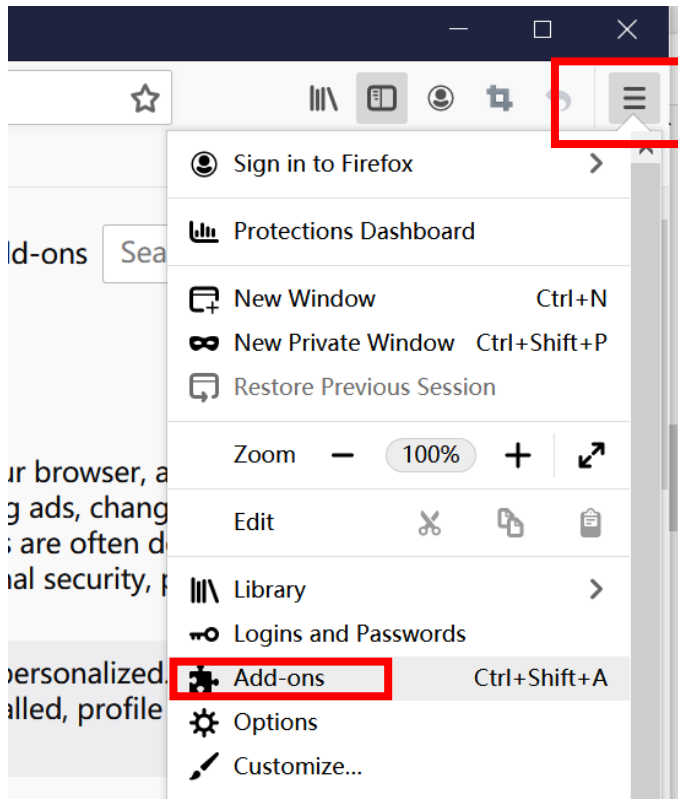
win: Put the unzipped Chromedriver.exe in its python. exe sibling. Make sure the path is in the environment variable

Enter chromedriver in cmd to test if the environment variable was added successfully

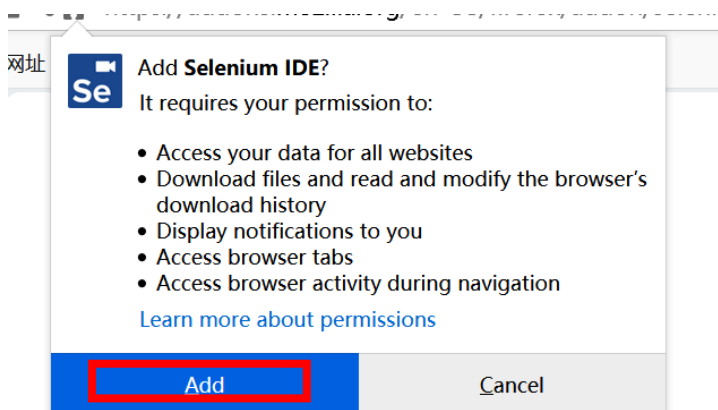
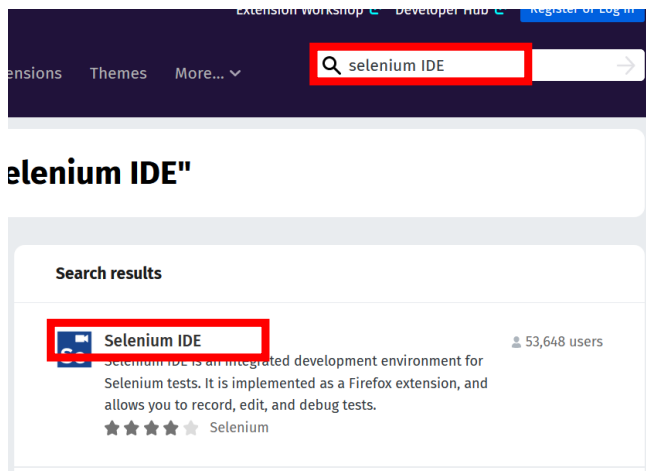
```
C:\Users\Nicole>chromedriver
Starting ChromeDriver 87.0.4280.88 (89e2380a3e36c3464b5dd1302349b135)
Only local connections are allowed.
```

Step 3: use Selenium to configure the Firefox environment

1. Download geckodriver from <https://github.com/mozilla/geckodriver/releases>
Install in the default address C:\Program Files\Mozilla Firefox
2. Open your Firefox browser and configure the Selenium IDE plug-in.
 - 2.1 open Add-ons



2.2 Search for Selenium IDE and add it to Firefox



3. Unzip geckodriver.zip to geckodriver.exe

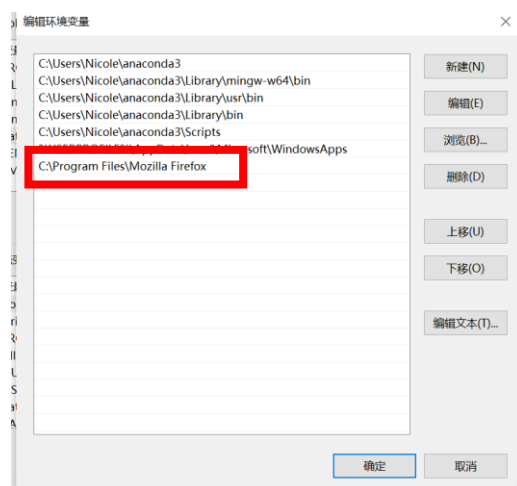
Move geckodriver.exe somewhere that Python and Selenium will be able to find it

mac: move to /usr/local/bin

win: Put the unzipped Chromedriver.exe in its python. exe sibling. Make sure the path is in the environment variable

4. Configure the environment

Move the path of Firefox (C:\Program Files\Mozilla Firefox) to environmental variables



Enter Firefox.exe in the terminal. If Firefox can start automatically, it means the installation is successful.

Step 4: Enter code below in your Jupyter notebook.

You should automatically open Baidu.com by Chrome and FireFox.

```
In [6]: from selenium import webdriver

driver = webdriver.Chrome()
driver.get("http://www.baidu.com")
driver.quit()
```

```
In [9]: from selenium import webdriver
driver = webdriver.Firefox()
driver.get('https://www.baidu.com')|
```

Step 5: Modify code parameters

Before enter command1 (crawling data and analyzing), please make sure to **modify the 'User-Agent' in BiliDanmuCrawl.py and DoubanRawDataCrawl.py.**

Please **modify the username and password in WeiboRawDataCrawl.py**

```
# auto-login
username = browser.find_element_by_css_selector('#loginName')
time.sleep(2)
username.clear()
username.send_keys('###please modify it###') # username
# password
password = browser.find_element_by_css_selector('#loginPassword')
time.sleep(2)
password.send_keys('###please modify it##') # password
browser.find_element_by_css_selector('#loginAction').click()
time.sleep(60)
```

Step 6: run the code

Using idle to run the file main.py.

Manual operation is required!

When crawling Weibo data, Weibo may ask you to verify your identity. Weibo usually sends a verification code to your mobile phone or Weibo private message. Please fill in the verification code within 60 seconds and pass the verification.

登录验证

为保障帐号安全，请使用手机号接收短信验证码以验证身份

18*****57



获取短信验证码

收不到验证码? 使用其他方式验证

请输入验证码

验证码已发送到 +86 18*****57

Enter verification code

输入验证码



54s后可重新获取

确认

收不到验证码? 使用其他方式验证