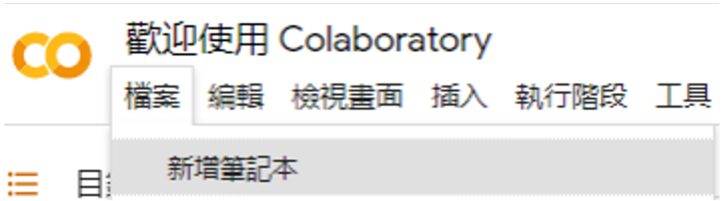
Colab使用教學

Colab

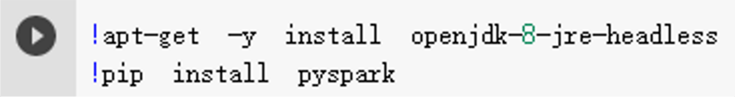
- Google Colaboratory (簡稱為 Colab) 可讓你在瀏覽器上撰寫及執行 Python，且具備下列優點：
  - 不必進行任何設定
  - 免費使用 GPU
  - 輕鬆共用
  - https://colab.research.google.com/notebooks/in

Spark安裝

1. 檔案 -> 新增筆記本



1. 安裝PySpark環境

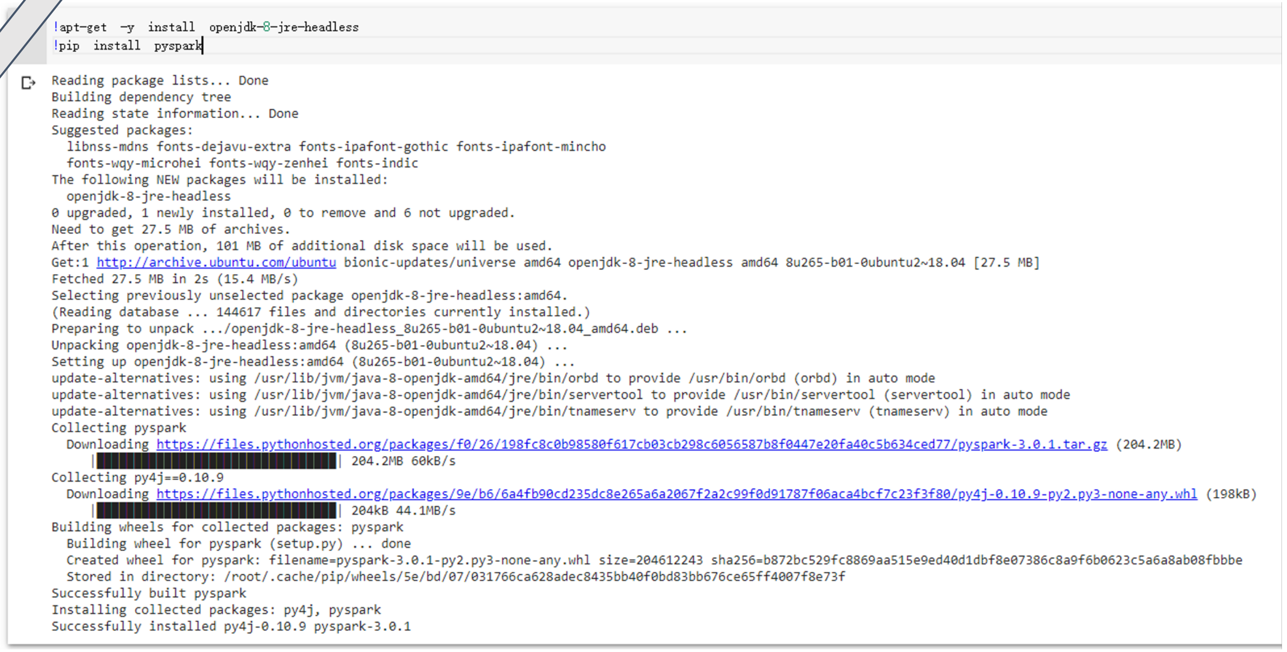


```
!apt-get -y install openjdk-8-jre-headless
!pip install pyspark
```

1. 安裝成功!

!apt-get -y install openjdk-8-jre-headless

!pip install pyspark



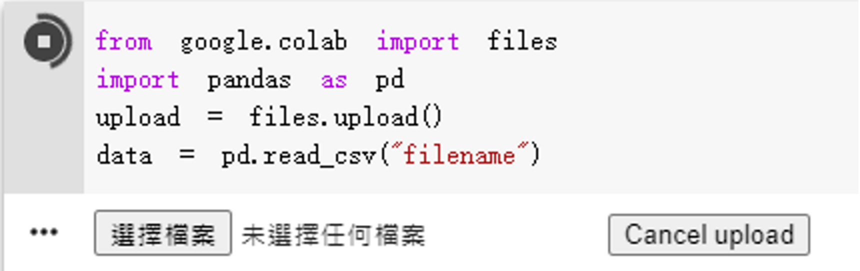
```
!apt-get -y install openjdk-8-jre-headless
!pip install pyspark
```

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
Suggested packages:
  libnss-mdns fonts-dejavu-extra fonts-ipafont-gothic fonts-ipafont-mincho
  fonts-wqy-microhei fonts-wqy-zenhei fonts-indic
The following NEW packages will be installed:
  openjdk-8-jre-headless
0 upgraded, 1 newly installed, 0 to remove and 6 not upgraded.
Need to get 27.5 MB of archives.
After this operation, 101 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 openjdk-8-jre-headless amd64 8u265-b01-0ubuntu2~18.04 [27.5 MB]
Fetched 27.5 MB in 2s (15.4 MB/s)
Selecting previously unselected package openjdk-8-jre-headless:amd64.
(Reading database ... 144617 files and directories currently installed.)
Preparing to unpack .../openjdk-8-jre-headless_8u265-b01-0ubuntu2~18.04_amd64.deb ...
Unpacking openjdk-8-jre-headless:amd64 (8u265-b01-0ubuntu2~18.04) ...
Setting up openjdk-8-jre-headless:amd64 (8u265-b01-0ubuntu2~18.04) ...
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/orbd to provide /usr/bin/orbd (orbd) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/servertool to provide /usr/bin/servertool (servertool) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/tnameserv to provide /usr/bin/tnameserv (tnameserv) in auto mode
Collecting pyspark
  Downloading https://files.pythonhosted.org/packages/f0/26/198fc8c0b98580f617cb03cb298c6056587b8f0447e20fa40c5b634ced77/pyspark-3.0.1.tar.gz (204.2MB)
     |████████████████████████████████| 204.2MB 60kB/s
Collecting py4j==0.10.9
  Downloading https://files.pythonhosted.org/packages/9e/b6/6a4fb90cd235dc8e265a6a2067f2a2c99f0d91787f06aca4bcf7c23f3f80/py4j-0.10.9-py2.py3-none-any.whl (198kB)
     |████████████████████████████████| 204kB 44.1MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.0.1-py2.py3-none-any.whl size=204612243 sha256=b872bc529fc8869aa515e9ed40d1dbf8e07386c8a9f6b0623c5a6a8ab08fbbbe
  Stored in directory: /root/.cache/pip/wheels/5e/bd/07/031766ca628adec8435bb40f0bd83bb676ce65ff4007f8e73f
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9 pyspark-3.0.1
```

PySpark相關說明: (**PySpark Documentation**: https://spark.apache.org/docs/latest/api/python/)

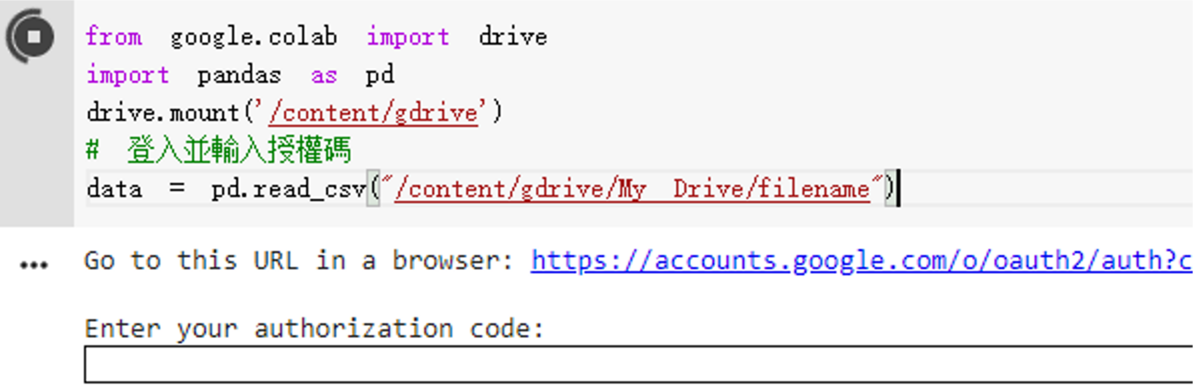
Colab匯入檔案

- 方法一、上傳檔案



```
from google.colab import files
import pandas as pd
upload = files.upload()
data = pd.read_csv("filename")
```

- 方法二、Google Drive



```
from google.colab import drive
import pandas as pd
drive.mount('/content/gdrive')
# 登入並輸入授權碼
data = pd.read_csv("/content/gdrive/My Drive/filename")
```

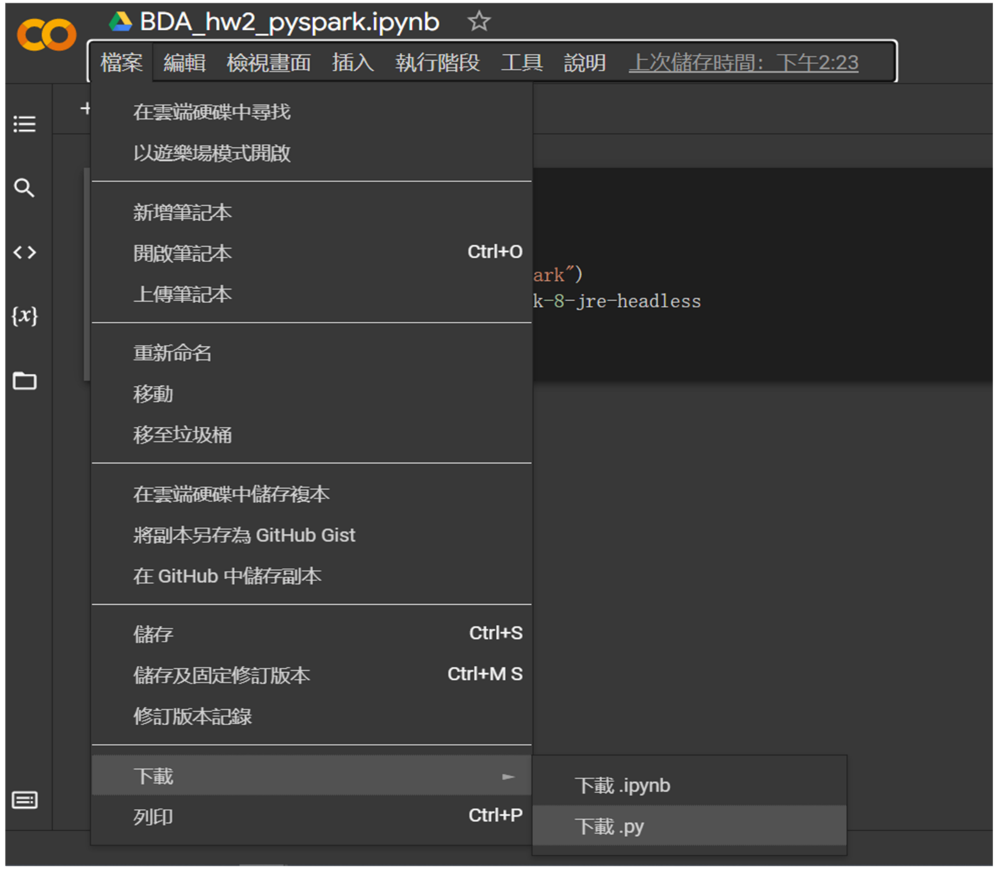
```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?c

Enter your authorization code:
```

從Colab下載成python檔

- 在作業繳交時，會需要繳交python檔，請依照下圖方法操作下載成python檔。

- 點擊 檔案>下載>下載.py



參考資料

- https://medium.com/@chiayinchen/%E4%BD%BF%E7%94%A8-google-colaboratory-%E8%B7%91-pyspark-625a07c75000

- https://ithelp.ithome.com.tw/articles/10217962