

AI-CUP 訓練資料下載 程式範例

1. 定義

- scidm 核心使用 ckan 框架，因此所有 api 都對應 ckanapi 之開發模式，因此完整 api 使用方法可以參考：<https://docs.ckan.org/en/2.7/api/#module-ckan.logic.action.get>
- 以下列出 scidm 上的名稱與 ckan api 的英文對應
 - 群組 => group
 - 資料集 => package
 - 資料 => resource
 - 組織 => organization

1.1 python 環境套件

需要安裝 ckanapi 套件

```
$ pip install ckanapi
```

1.2 宣告

需查詢自己的 token key

資料集平台
Data Market

1. 登入後點選username

/ 使用者 / waue

資料集 動態牆 點數管理 我的最愛 管理

OWNER [全公開] 示範資料: datacon2019clipper
更新頻率 不定期 瀏覽次數 10409 下載次數 114
datacon 2019 clipper 資料示範集 [公開-需簽署] 示範資料
JPEG CSV ZIP

需申請審核 **OWNER** test578195u2934 草稿
更新頻率 不定期 瀏覽次數 0 下載次數 0
test

需申請審核 **OWNER** R1_FishLen_Dataset_CenterPoint_516f&5014p(R1新竹市魚眼交通資料集)
更新頻率 不定期 瀏覽次數 379 下載次數 5
Hsinchu city traffic image data_with point center annotation_5000p Dataset Description This data set is a fisheye lens video file of important intersections in Hsinchu City,...

需申請審核 **OWNER** R0_FishLen_Dataset_CenterPoint_(R0原新竹市魚眼交通資料集)
更新頻率 不定期 瀏覽次數 816 下載次數 34
資料集說明 此資料集為新竹市重要路口的魚眼鏡頭影片檔，由新竹市警察局提供。由於有影像+標註，可用於AI科學研發，歡迎大家申請使用。 sample file download here Dataset Description This dataset is a fisheye video file of important intersections...

需申請審核 **OWNER** R3_FishLen_Dataset_CenterPoint_18000f&160000p(R3新竹市魚眼交通資料集)
更新頻率 不定期 瀏覽次數 118 下載次數 0
2022_新竹市交通影像資料_含點中心標註_18000圖_160000點 Hsinchu city traffic image data_with point center annotation Dataset Description This data set is a fisheye lens video file of important...

追蹤者 0 資料集 14
編輯 43.5k

使用者名稱
waue

電子郵件 需申請審核
wy...@nchc.org.tw

會員註冊日
八月 28, 2017

狀態
active

API Key 需申請審核
2 69-
b.....

2. 此處可以查詢自己的api key

```
In [1]: # 全域變數
API_KEY = 'xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx' ## 此處填入 API Key
DATA_DIR = "./data" ## 此處填入 資料下載資料夾
```

```
In [2]: from ckanapi import RemoteCKAN
import requests
import hashlib
import os

ua = 'ckanapiexample/1.0 (+http://example.com/my/website)'
mysite = RemoteCKAN('https://scidm.nchc.org.tw/', apikey=API_KEY, user_agent=ua)
os.makedirs(DATA_DIR, exist_ok=True)
```

1.3 api 測試

- 參考 <https://docs.ckan.org/en/2.7/api/#module-ckan.logic.action.get>
- 其中會用到列出所有的 datasets 會用到這個 api
 - ckan.logic.action.get.package_list(context, data_dict)
- 但真正程式碼要用以下寫法。

- 執行結果如資料市集內的網址：<https://scidm.nchc.org.tw/organization/geomatics-ncku-edu-tw>

```
In [3]: organization=mysite.action.organization_show(id="geomatics-ncku-edu-tw", in
print(len(organization["packages"])))
```

14

```
In [4]: for idx in range(len(organization["packages"])):
ds_title = organization["packages"][idx]["title"]
ds_name = organization["packages"][idx]["name"]
print("{} \n dataset title = {} \n dataset id = {}".format(idx, ds_tit
```

```
[0]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_甘藍影像訓練資料集
dataset name = aicup2022_kale_training
[1]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_蔥影像訓練資料集
dataset name = aicup2022_greenonion_training
[2]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_溫網室影像訓練資料集
dataset name = aicup2022_greenhouse_training
[3]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_韭菜影像訓練資料集
dataset name = aicup2022_chinesechives_training
[4]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_結球白菜影像訓練資料集
dataset name = aicup2022_chinesecabbage_training
[5]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_花椰菜影像訓練資料集
dataset name = aicup2022_cauliflower_training
[6]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_青花菜影像訓練資料集
dataset name = aicup2022_broccoli_training
[7]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_檳榔影像訓練資料集
dataset name = aicup2022_betel_training
[8]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_蘆筍影像訓練資料集
dataset name = aicup2022_asparagus_training
[9]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_竹筍影像訓練資料集
dataset name = aicup2022_bambooshoots_training
[10]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_番荔枝影像訓練資料集
dataset name = aicup2022_custardapple_training
[11]
dataset title = AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽_葡萄影像訓練資料集
dataset name = aicup2022_grape_training
[12]
dataset title = 秋季賽作物中英文對照表
dataset name = classnameeng
[13]
dataset title = AI CUP 2022 秋季賽 - 競賽活動需知與手冊
dataset name = ai-cup-2022-manual
```

2 範例

2.1 範例一：列出指定package的resource

- 以下列出 資料集名稱 秋季賽作物中英文對照表 的所有 metadata 資料
- 範例網址 <https://scidm.nchc.org.tw/dataset/classnameeng>

```
In [11]: mydataset = mysite.action.package_show(id="aicup2022_kale_training")  
print(mydataset)
```

```
{'rating': 0.0, 'license_title': '01 政府資料開放授權', 'maintainer': '許育維',
'relationships_as_object': [], 'blockchain_record': False, 'private': False,
'maintainer_email': 'ben831013@gmail.com', 'num_tags': 2, 'update_frequenc
y': 'noscheduled', 'hot_view': 6, 'id': '95cale75-89e4-4529-a27e-5a53c7e1ea3
7', 'metadata_created': '2022-09-12T07:45:47.715792', 'pay_type': 'free', 'm
etadada_modified': '2022-09-12T08:05:21.986376', 'author': '', 'author_email':
'', 'sale_nchc_points': 0, 'acquire_url': '', 'state': 'active', 'version':
'', 'usage_announcement': '使用者得無償自本資料庫下載資料作為非營利之研發目的使
用，並同意遵守以下之授權規範。下文中，『國網資料集平台』以『本平台』為之簡稱。一、
下載之資料僅得為個人學術之研究使用，不得以營利目的為之。二、非經事前同意不得擅自將
資料內容增、刪、修改或割裂。三、不得將本資料以重置、移轉等方式交付第三人。四、本
授權非經同意不得讓與或繼承，亦不得轉授權於第三人。五、若有違反本授權規範之情事，應賠償
本中心因此遭受之損害(包括但不限於律師費用)，本平台亦得隨時終止本授權並要求刪除以下載之之
料。六、不正使用資料者本平台得隨時終止使用者使用之權利，並得視違規情節輕重禁止其於一段
時間內提出申請使用本數位資源。七、使用者應本善良管理人之注意義務妥善保管利用，若有可歸
責於使用者之故意或過失侵害本平台權益時，使用者應對本平台負擔損害賠償責任。八、本授權協
議之準據法為中華民國法。九、本授權協
```

```
-----\r\n若同意以上授權條款，請於下方說明處填寫下列
資訊：\r\n\r\n* 申請目的：', 'license_id': 'OGL-TW-1.0', 'type': 'dataset', 'r
esources': [{ 'cache_last_updated': None, 'package_id': '95cale75-89e4-4529-a
27e-5a53c7e1ea37', 'datastore_active': False, 'id': 'abb25c65-9588-403f-a649
-56d0ed940e05', 'size': 1073741824, 'state': 'active', 'sha256': 'eda38223eb
c152913317c56ccb4d32a10bc43ae888b982729ad4ac3c829f1ad3', 'hash': '', 'descri
ption': '', 'format': 'RAR', 'last_modified': '2022-09-12T07:47:40.519290',
'url_type': 'upload', 'md5': '34a9e5a86c1ea1919989d5c698751c9d', 'mimetype':
'application/rar', 'cache_url': None, 'name': 'kale.part1.rar', 'created':
'2022-09-12T07:47:41.867051', 'url': 'https://scidm.nchc.org.tw/dataset/95ca
le75-89e4-4529-a27e-5a53c7e1ea37/resource/abb25c65-9588-403f-a649-56d0ed940e
05/nchcproxy/kale.part1.rar', 'mimetype_inner': None, 'position': 0, 'revisi
on_id': 'd78b13cf-5ea6-480c-a528-4cf40c27c6ce', 'resource_type': None}, { 'ca
che_last_updated': None, 'package_id': '95cale75-89e4-4529-a27e-5a53c7e1ea3
7', 'datastore_active': False, 'id': '3bd4e155-3fb7-4b74-bf7f-281f6f235a61',
'size': 1073741824, 'state': 'active', 'sha256': '16dff7a00d632f4eec733c4416
72629d7aa3cf3709a99eb56fedb89162b10dd2', 'hash': '', 'description': '', 'for
mat': 'RAR', 'last_modified': '2022-09-12T07:48:34.424202', 'url_type': 'upl
oad', 'md5': 'd2d5cae96fd72976004981e1c90f254d', 'mimetype': 'application/ra
r', 'cache_url': None, 'name': 'kale.part2.rar', 'created': '2022-09-12T07:4
8:35.824077', 'url': 'https://scidm.nchc.org.tw/dataset/95cale75-89e4-4529-a
27e-5a53c7e1ea37/resource/3bd4e155-3fb7-4b74-bf7f-281f6f235a61/nchcproxy/kal
e.part2.rar', 'mimetype_inner': None, 'position': 1, 'revision_id': 'ac62d88
a-9e43-40c0-90e1-f87d62254a9c', 'resource_type': None}, { 'cache_last_update
d': None, 'package_id': '95cale75-89e4-4529-a27e-5a53c7e1ea37', 'datastore_a
ctive': False, 'id': '895cdd36-92a2-4309-9cff-b70f28cd45c2', 'size': 1073741
824, 'state': 'active', 'sha256': 'd52bf94d4d9ffa6adbb0d2d5661bfe7a75b8118c0
cf52bcc95c71a791fd53170', 'hash': '', 'description': '', 'format': 'RAR', 'l
ast_modified': '2022-09-12T07:55:17.273931', 'url_type': 'upload', 'md5': 'c
7e25a0aff0057c8deea5f19be5fe6a8', 'mimetype': 'application/rar', 'cache_ur
l': None, 'name': 'kale.part3.rar', 'created': '2022-09-12T07:55:18.535676',
'url': 'https://scidm.nchc.org.tw/dataset/95cale75-89e4-4529-a27e-5a53c7e1ea
37/resource/895cdd36-92a2-4309-9cff-b70f28cd45c2/nchcproxy/kale.part3.rar',
'mimetype_inner': None, 'position': 2, 'revision_id': 'd5a2e18c-2702-41f2-bb
f9-ed1c49189274', 'resource_type': None}, { 'cache_last_updated': None, 'pack
age_id': '95cale75-89e4-4529-a27e-5a53c7e1ea37', 'datastore_active': False,
'id': 'b8f00183-a15d-498d-belc-ebe749e428b5', 'size': 1073741824, 'state':
'active', 'sha256': '4261af871f7847fee294289b1caf96c51b2996c2e02554974dbc1ec
3882a8a7f', 'hash': '', 'description': '', 'format': 'RAR', 'last_modified':
'2022-09-12T08:03:32.526895', 'url_type': 'upload', 'md5': 'd506c79991c15328
ca5423766f51a565', 'mimetype': 'application/rar', 'cache_url': None, 'name':
'kale.part4.rar', 'created': '2022-09-12T08:03:33.787439', 'url': 'https://s
cidm.nchc.org.tw/dataset/95cale75-89e4-4529-a27e-5a53c7e1ea37/resource/b8f00
183-a15d-498d-belc-ebe749e428b5/nchcproxy/kale.part4.rar', 'mimetype_inner':
None, 'position': 3, 'revision_id': '0caa9947-98f3-45cc-aad5-331b3d757ad2',
'resource_type': None}, { 'cache_last_updated': None, 'package_id': '95cale75
-89e4-4529-a27e-5a53c7e1ea37', 'datastore_active': False, 'id': '5163fb44-3c
```

```
0a-4c4c-b367-1632db97970d', 'size': 130940497, 'state': 'active', 'sha256':
'558e7135bf96b64beff030281de61eb71ddce5ed0676e7941a8bb18b4c67f479', 'hash':
'', 'description': '', 'format': 'RAR', 'last_modified': '2022-09-12T08:04:0
6.718795', 'url_type': 'upload', 'md5': '0a1a974d5f1c9b6735f6317e108b2249',
'mimetype': 'application/rar', 'cache_url': None, 'name': 'kale.part5.rar',
'created': '2022-09-12T08:04:07.987132', 'url': 'https://scidm.nchc.org.tw/d
ataset/95cale75-89e4-4529-a27e-5a53c7e1ea37/resource/5163fb44-3c0a-4c4c-b367
-1632db97970d/nchcproxy/kale.part5.rar', 'mimetype_inner': None, 'position':
4, 'revision_id': 'f3173c68-80d8-4784-8833-883999a0e140', 'resource_type': N
one}}, 'usable': True, 'num_resources': 5, 'tags': [{ 'vocabulary_id': None,
'state': 'active', 'display_name': 'AI', 'id': 'f66d0014-fb32-40e7-8ec8-ff4d
57b4f5a3', 'name': 'AI'}, { 'vocabulary_id': None, 'state': 'active', 'displa
y_name': '農業', 'id': 'b419a75e-a51b-4e30-9c81-9c0ee2306d5b', 'name': '農
業'}]], 'hot_download': 0, 'access_type': 'Authorize', 'groups': [{ 'display_n
ame': '2022 AI-CUP 秋季賽', 'description': '', 'image_display_url': 'https://
scidm.nchc.org.tw/uploads/group/2022-08-17-021430.1561282022aicupfall.png',
'title': '2022 AI-CUP 秋季賽', 'id': '54d1a891-8e3b-425b-9729-e27965d8f4a2',
'name': '2022-ai-cup'}]], 'creator_user_id': '033f983c-67fe-44ef-a280-fb079b2
fc48d', 'relationships_as_subject': [], 'organization': { 'description': '支
援 2022 AI CUP 競賽資料', 'created': '2022-08-17T09:58:07.124918', 'title':
'成大測量所', 'name': 'geomatics-ncku-edu-tw', 'is_organization': True, 'stat
e': 'active', 'image_url': '2022-08-17-015807.072637geomaticsncku.png', 'rev
ision_id': 'af2f892b-89bf-41ec-be5c-d7159ffb88ab', 'type': 'organization',
'id': 'f6c7dfe4-29da-4ced-ab3b-80467df08720', 'approval_status': 'approve
d'}, 'name': 'aicup2022_kale_training', 'isopen': True, 'url': '', 'notes':
'用於AI CUP 2022農地作物現況調查影像辨識競賽-秋季賽之訓練資料集，影像資訊來源為行政院農
業委員會。', 'owner_org': 'f6c7dfe4-29da-4ced-ab3b-80467df08720', 'license_ur
l': 'https://data.gov.tw/license', 'ratings_count': 0, 'title': 'AI CUP 2022
農地作物現況調查影像辨識競賽-秋季賽_甘藍影像訓練資料集', 'revision_id': '42e228cd-19
1b-48b8-ba41-56dd71ef6aff'}
```

- 回傳結果中，有個 field 為 resources 的串列，使用for 迴圈取出內部結構
- 分析後，url 是我們要下載的連結

```
In [12]: urls = []
ids = []
for resource in mydataset["resources"]:
    print(" * {} ( {} )=> {}".format(resource["name"],resource["id"] , resou
    urls.append(resource["url"])
    ids.append(resource["id"])

    * kale.part1.rar ( abb25c65-9588-403f-a649-56d0ed940e05 )=> https://scidm.n
chc.org.tw/dataset/95cale75-89e4-4529-a27e-5a53c7e1ea37/resource/abb25c65-95
88-403f-a649-56d0ed940e05/nchcproxy/kale.part1.rar
    * kale.part2.rar ( 3bd4e155-3fb7-4b74-bf7f-281f6f235a61 )=> https://scidm.n
chc.org.tw/dataset/95cale75-89e4-4529-a27e-5a53c7e1ea37/resource/3bd4e155-3f
b7-4b74-bf7f-281f6f235a61/nchcproxy/kale.part2.rar
    * kale.part3.rar ( 895cdd36-92a2-4309-9cff-b70f28cd45c2 )=> https://scidm.n
chc.org.tw/dataset/95cale75-89e4-4529-a27e-5a53c7e1ea37/resource/895cdd36-92
a2-4309-9cff-b70f28cd45c2/nchcproxy/kale.part3.rar
    * kale.part4.rar ( b8f00183-a15d-498d-belc-ebe749e428b5 )=> https://scidm.n
chc.org.tw/dataset/95cale75-89e4-4529-a27e-5a53c7e1ea37/resource/b8f00183-a1
5d-498d-belc-ebe749e428b5/nchcproxy/kale.part4.rar
    * kale.part5.rar ( 5163fb44-3c0a-4c4c-b367-1632db97970d )=> https://scidm.n
chc.org.tw/dataset/95cale75-89e4-4529-a27e-5a53c7e1ea37/resource/5163fb44-3c
0a-4c4c-b367-1632db97970d/nchcproxy/kale.part5.rar
```

2.2 範例二：下載已知的 url 檔

2.2.1 定義下載functions, 包含下載與檢查完整度

```
In [13]: def file_as_bytes(file):
    with file:
        return file.read()

def download_file_by_resource(resource):
    local_filename = resource["name"]
    url = resource["url"]
    md5_rs = resource["md5"]

    ### download file
    local_f_path = os.path.join(DATA_DIR, local_filename)
    with requests.get(url, stream=True, headers={'Authorization': API_KEY})
        r.raise_for_status()
        with open(local_f_path, 'wb') as f:
            for chunk in r.iter_content(chunk_size=8192):
                f.write(chunk)

    ### check md5 by comparing local_file and cloud
    md5_local = hashlib.md5(file_as_bytes(open(local_f_path, 'rb'))).hexdigest()

    if not str(md5_local) == str(md5_rs):
        print("<download failed> your file's check sum == {} not equal to {}".format(md5_local, md5_rs))
        local_f_failed = local_f_path + ".failed"
        os.rename(local_f_path, local_f_failed)
        local_f_path = local_f_failed

    return local_f_path
```

2.2.2 開始下載指定資料集內resource

```
In [ ]: import time

for id in ids :
    start = time.time()
    resource = mysite.action.resource_show(id=id)
    # print(resource)
    fname = download_file_by_resource(resource)
    end = time.time()
    print("{} ({} s)".format(fname, round(end - start, 2)))

./data/kale.part1.rar (274.79 s)
./data/kale.part2.rar (275.86 s)
./data/kale.part3.rar (273.33 s)
```