

Web Scraping (Kaggle Datasets), in general

Example: Dataset of the course 3: Natural Language Processing in TensorFlow

DeepLearning.AI TensorFlow Developer Certificat Professionnel

<https://www.coursera.org/professional-certificates/tensorflow-in-practice>

Example of the course : dataset for sarcasm detection

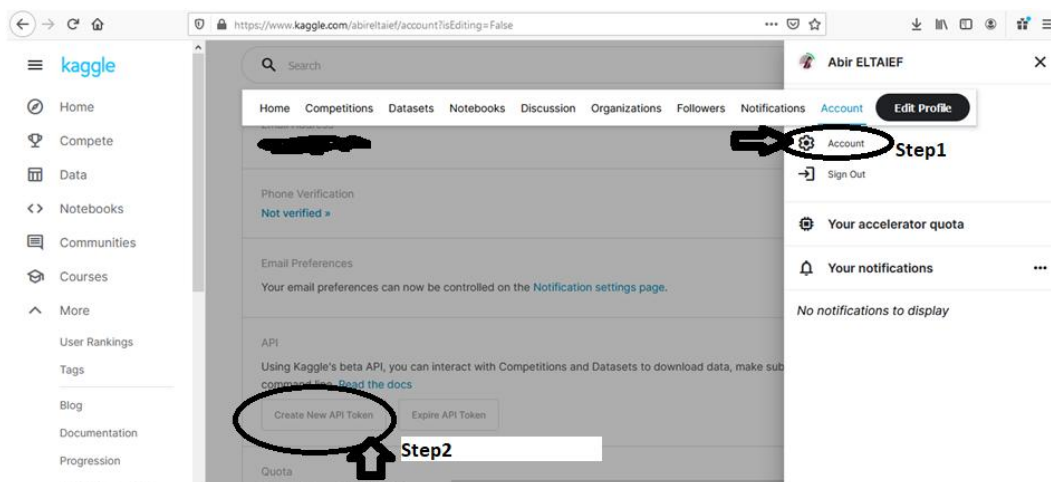
Description:

As mentioned, in the course: the web scraping (Getting the data), and cleaning it (if necessary), is a little bit beyond the scope of the course.

For all intents and purposes, and to all those wishing to get the dataset using **Kaggle API**, here, I give details of the methodology (in general using the example of the course):

Step I :

- Log in to [Kaggle](#) and access your account

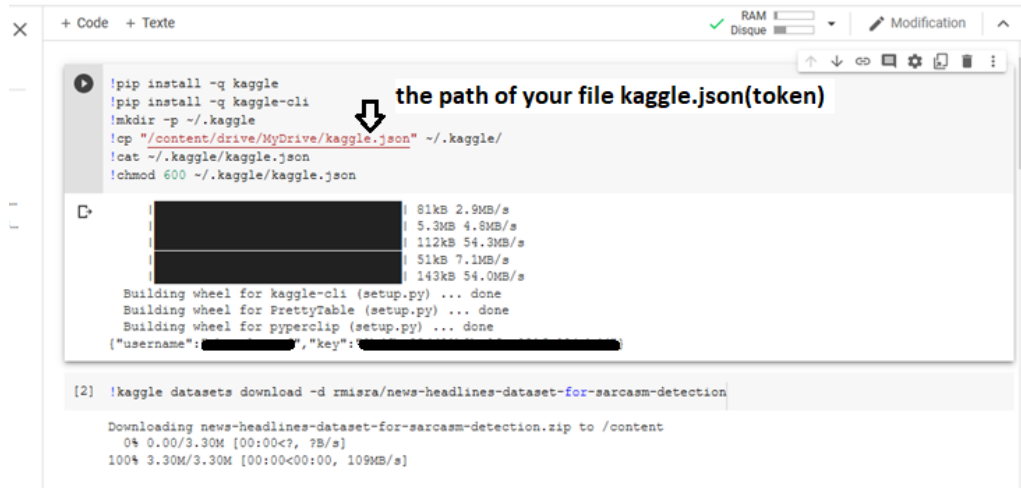


- Go to **Account**, then Click on 'create New API Token'
- Download the **kaggle.json** file which contains your **API token** : It contains something like this :

```
{"username":"aaaaaaa","key":"8aa.....ee0884ne020cfgem"}
```
- Save it **somewhere** (for example copy it to a folder mounted on google colab): **it will be the path of your token file: kaggle.json (downloaded)**

Step II :

- 1- Write this code on your notebook (on google colab) (replace my path by yours) :



```
!pip install -q kaggle
!pip install -q kaggle-cli
!mkdir -p ~/.kaggle
!cp "/content/drive/MyDrive/kaggle.json" ~/.kaggle/
!cat ~/.kaggle/kaggle.json
!chmod 600 ~/.kaggle/kaggle.json

[1] 81kB 2.9MB/s
    5.3MB 4.8MB/s
    112kB 54.3MB/s
    51kB 7.1MB/s
    143kB 54.0MB/s
Building wheel for kaggle-cli (setup.py) ... done
Building wheel for PrettyTable (setup.py) ... done
Building wheel for pyperclip (setup.py) ... done
{"username": "XXXXXXXXXX", "key": "XXXXXXXXXX"}

[2] !kaggle datasets download -d rmisra/news-headlines-dataset-for-sarcasm-detection

Downloading news-headlines-dataset-for-sarcasm-detection.zip to /content
0% 0.00/3.30M [00:00<?, ?B/s]
100% 3.30M/3.30M [00:00<00:00, 109MB/s]
```

The code :

```
!pip install -q kaggle
!pip install -q kaggle-cli
!mkdir -p ~/.kaggle
!cp "/content/drive/MyDrive/kaggle.json" ~/.kaggle/
!cat ~/.kaggle/kaggle.json
!chmod 600 ~/.kaggle/kaggle.json
```

- 2- To download the dataset « News headlines dataset for sarcasm detection from

[:https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection/home](https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection/home)

- 3- Write this line o code on your notebook :

```
!kaggle datasets download -d rmisra/news-headlines-dataset-for-sarcasm-detection
```

- 4- Then, run these classic to unzip the file and get the data :

Note : you should add the medthod `.strip()` , as I did below, to make it work(...)



The screenshot shows a Jupyter Notebook with a file explorer on the left containing 'eadlines_Dataset.json', 'eadlines_Dataset_v2.js...', and 'ines-dataset-for-sarcas...'. The main area displays the following code and output:

```
+ Code + Texte
!kaggle datasets download -d rmisra/news-headlines-dataset-for-sarcasm-detection

Downloading news-headlines-dataset-for-sarcasm-detection.zip to /content
0% 0.00/3.30M [00:00<?, ?B/s]
100% 3.30M/3.30M [00:00<00:00, 109MB/s]

[4] import zipfile
file_dezip = zipfile.ZipFile('/content/news-headlines-dataset-for-sarcasm-detection.zip')
file_dezip.extractall('/content')

[12] datastore=[]
with open('Sarcasm_Headlines_Dataset.json','r') as f:
    for line in f:
        line = json.loads(line.strip())
        datastore.append(line)

[13] datastore
{ 'article_link': 'https://100gal.theonion.com/unconditional-love-given-to-15-year-old-who-just-called-mom-a-bitch-in-middle-of-hollister',
  'headline': 'unconditional love given to 15-year-old who just called mom a bitch in middle of hollister',
  'is_sarcastic': 1},
{ 'article_link': 'https://www.theonion.com/newborn-prince-of-cambridge-begins-consolidating-power-1825477519',
  'headline': 'newborn prince of cambridge begins consolidating power by having family imprisoned in tower of london',
  'is_sarcastic': 1}
```

The code :

```
!kaggle datasets download -d rmisra/news-headlines-dataset-for-sarcasm-detection

import zipfile
file_dezip = zipfile.ZipFile('/content/news-headlines-dataset-for-sarcasm-detection.zip')
file_dezip.extractall('/content')

datastore=[]
with open('Sarcasm_Headlines_Dataset.json','r') as f:
    for line in f:
        line = json.loads(line.strip())
        datastore.append(line)
```

By Abir ELTAIEF