

# Taller de Distribución de Datos

---

## Objetivo

Utilizar instrucciones de programación para calcular estadísticas en una colección de datos semi-estructurados.

---

## Procedimiento

### 1. Abrir Python

Para abrir python siga las instrucciones del ítem número 1 de la **Guía de conexión** e ingrese a su carpeta personal dando doble clic sobre ella.

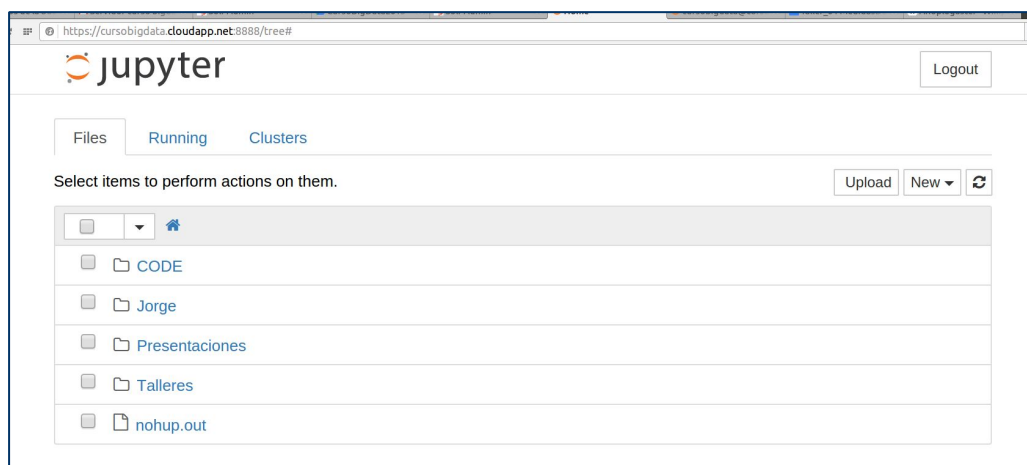


Figura 1

Una vez dentro de su carpeta de clic en **New** (recuadro de color rojo) situado al costado superior derecho de la ventana y luego de clic en **Python 3** (recuadro de color verde) como se ilustra en la **Figura 2**. Posteriormente se abrirá una pestaña en la que aparece una ventana como la mostrada en la **Figura 3**.

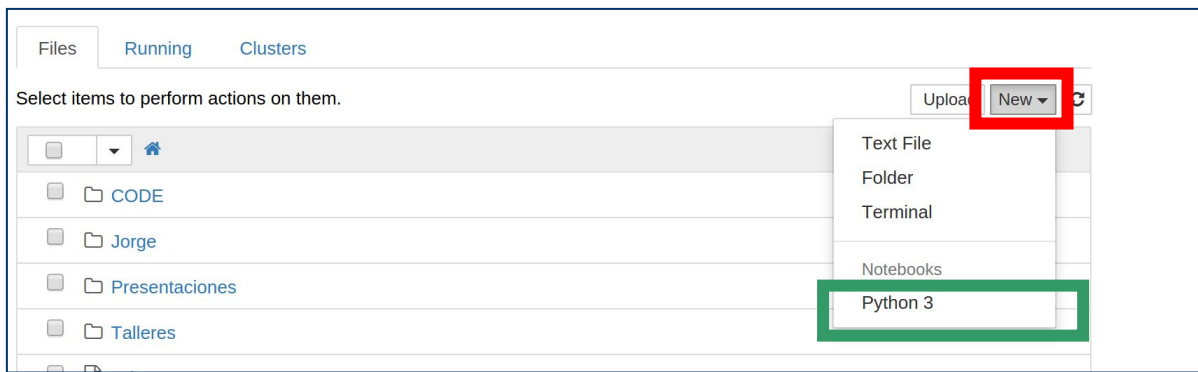


Figura 2

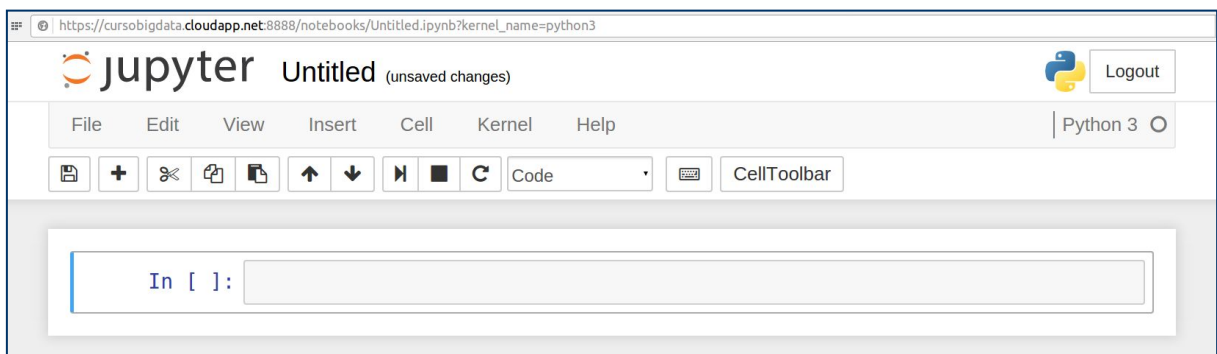


Figura 3

## 2. Renombre el Notebook

- Para renombrar el Notebook abierto de clic en **File** situado al costado superior izquierdo de la ventana y luego de clic en **Rename...** (recuadro de color verde) como se ilustra en la **Figura 4**. Posteriormente se abrirá una ventana emergente como la mostrada en la **Figura 5**.

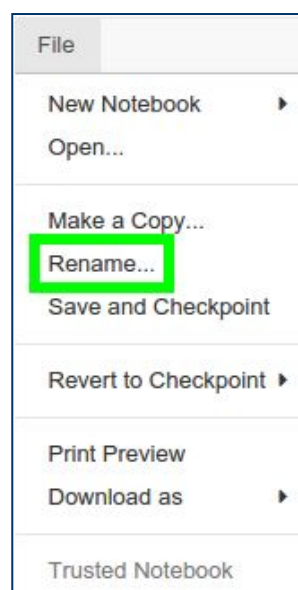


Figura 4



Figura 5


Reemplace **Untitled** por **Taller\_05** y de clic en **OK**.

### 3. Desarrollo del código.

Tenga en cuenta los botones señalados de color rojo y de color verde de la siguiente barra de herramientas.




Figura 6

Importar los módulos json y NumPy de Python, diseñados para la manipulación de archivos en formato json y matrices, respectivamente. Escriba en la celda las instrucciones que aparecen a continuación y de clic en  (recuadro señalado de color rojo en la **Figura 6**) para ejecutar esa instrucción.

```
In [1]: import json
import numpy as np
```


Bloque de código 1

Cargue el dataset ubicado en **/home/cursobigdata/Notebooks/Files/Result\_Tweets\_FV.json** a la variable **tweets**, y el número de documentos a la variable **total** para que posteriormente sea visualizado. Escriba en la celda las instrucciones que aparecen a continuación y de clic en . El número total de tweets deberá ser de 6992.

```
In [20]: tweets = json.load(open('/home/cursobigdata/Notebooks/Files/Result_Tweet
total = len(tweets)
print ("Total de tweets: ", total)

Total de tweets: 6992
```


Bloque de código 2

Para encontrar los lugares de la ciudad de Cali en donde se generan los tweets y la cantidad de publicaciones realizadas para cada lugar, escriba en la celda las instrucciones que aparecen a continuación y de clic en .

```
In [19]: lugares = {}
        for t in tweets:
            try:
                lugares[t['place']] += 1
            except:
                lugares[t['place']] = 1
        print ('Cantidad\tLugar')
        for k in lugares.keys():
            print (lugares[k], 't', k)

Cantidad      Lugar
163 t Yumbo, Valle del Cauca
1104 t Cali, Valle del Cauca
1110 t Colombia
1839 t Candelaria, Valle del Cauca
23 t Puerto Tejada, Cauca
3 t Cauca, Colombia
51 t Valle del Cauca, Colombia
374 t Pradera, Valle del Cauca
3 t Huila, Colombia
20 t Florida, Valle del Cauca
1 t Ciudad Madera, Tamaulipas
```

### Bloque de código 3

- Para hallar la hora promedio de las publicaciones y la desviación, es decir, el intervalo de tiempo en el cual se publican más tweets, se crea una función que separe las horas, los minutos y los segundos de cada tweet. Escriba en la celda las instrucciones que aparecen a continuación y de clic en .


```
In [117]: def hora(d):
        t = d.split('T')[1].replace('Z','')
        h = list(map(float, t.split(':')))
        return h

T = np.zeros((total, 3))
for i in range(total):
    T[i,:] = hora(tweets[i]['date'])

prom = list(map(float, np.mean(T,0)))
print ('Hora promedio:', prom[0], ':', prom[2])
des = list(map(float, np.std(T,0)))
print ('Desviación: (+/-)', des[0], ' horas ', des[1], ' min', des[2], ' seg')

Hora promedio: 11.403318077803204 : 29.79262013729977
Desviación: (+/-) 8.151092525330116 horas 17.54019744834843 min 17.300304305463968 seg
```

### Bloque de código 4


Para encontrar la cantidad de palabras diferentes presentes en el dataset de tweets, escriba en la celda las instrucciones que aparecen a continuación y de clic en .

```
In [119]: palabras = {}

for t in tweets:
    texto = t['text'].split()
    for w in texto:
        try:
            palabras[w] += 1
        except:
            palabras[w] = 1
print ('Total palabras diferentes:', len(palabras))

Total palabras diferentes: 19302
```

Bloque de código 5


- Para visualizar las diez (10) palabras más frecuentes y la cantidad de veces que se repite cada una de ellas (es decir, su frecuencia), se crea una función que permita la visualización de forma ordenada. Escriba en la celda las instrucciones que aparecen a continuación y de clic en .

```
In [121]: def mostrarInfo(info):
        for k in info:
            print (k[1], '\t', k[0])

frec = [(x, palabras[x]) for x in palabras.keys()]
frec.sort(key=lambda x:x[1], reverse=True)
mostrarInfo(frec[0:10])

2243    que
2175    de
1404    a
1356    la
1221    y
1167    el
1043    no
939     en
900     es
847     me
```


Bloque de código 6

- Para observar determinadas palabras (en este caso, de la fila número 50 a la 80) de la lista de las existentes en las publicaciones junto con su respectiva frecuencia, escriba en la celda la instrucción que aparece a continuación y de clic en .

```
In [122]: mostrarInfo(frec[50:80])


127     tan
123     Si
121     ni
117     qué
114     hay
113     mejor
113     o
110     muy
```

Bloque de código 7

Para observar las palabras presentes de la fila número 100 a la 130 y sus frecuencias, escriba en la celda la instrucción que aparece a continuación y de clic en .

```
In [124]: mostrarInfo(frec[100:130])
67      Se
67      mundo
66      da
64      así
64      nunca
64      algo
64      les
```


Bloque de código 8

Finalmente, para observar las palabras presentes de la fila número 2000 a la 2030 y sus frecuencias, escriba en la celda la instrucción que aparece a continuación y de clic en .

```
In [125]: mostrarInfo(frec[2000:2030])
4      volvió
4      Madre
4      mandan
4      faltan
4      conocerte
4      Padre
```

Bloque de código 9

## 4. Guarde y cierre su Notebook

- Para guardar su Notebook de clic en  (recuadro señalado de color verde en la barra de herramientas de la **Figura 6**).
- Posteriormente de clic en **File** y luego en **Close and Halt** (recuadro señalado de color verde en la **Figura 7**).

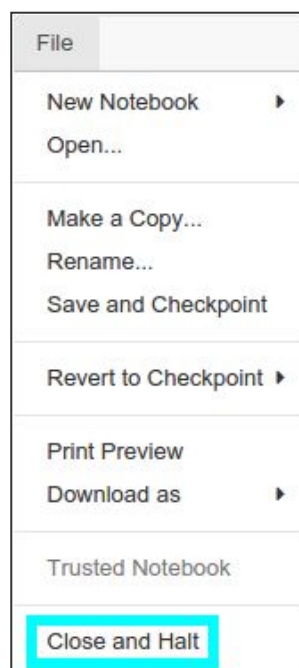


Figura 7

## **5. Desconectarse del servidor.**

- Para desconectarse del servidor, siga las instrucciones del ítem número 6 de la **Guía de conexión**.

---

**Fin de la Guia**