

Taller de Medidas de Similitud

Objetivo

Preparar una colección de datos para responder consultas y hacer búsquedas por similitud.

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.

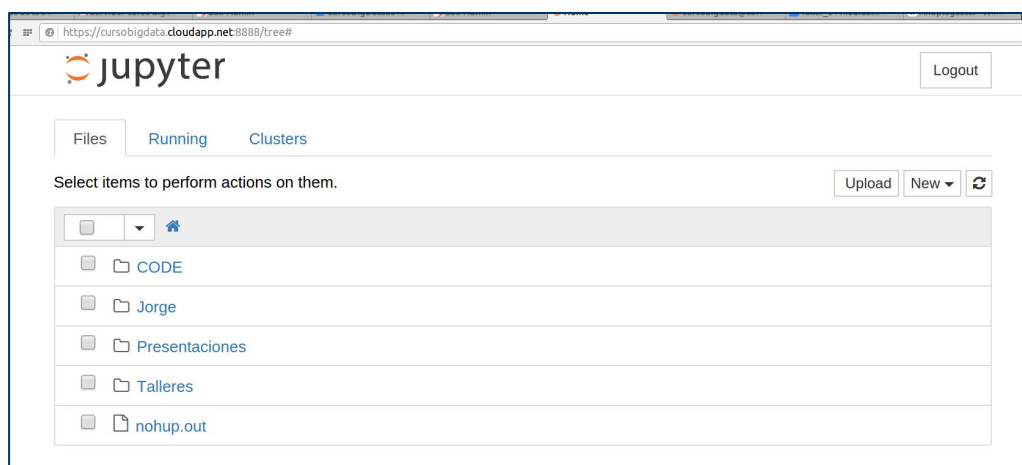


Figura 1

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

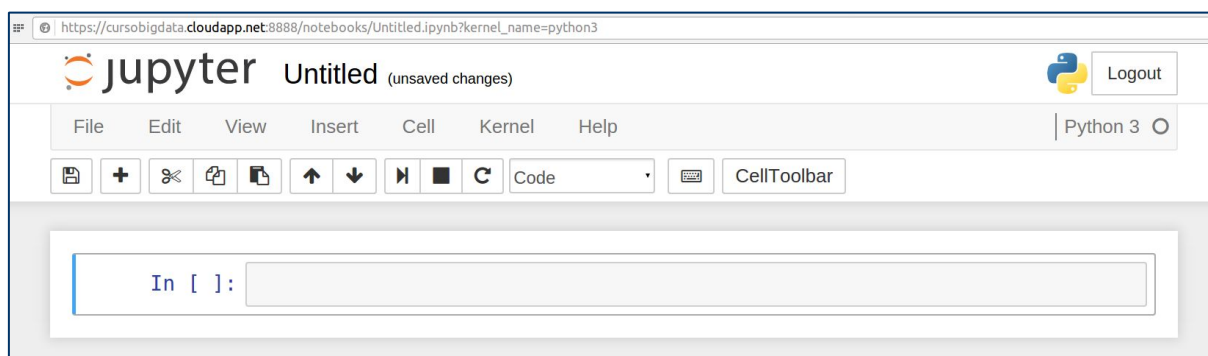


Figura 3

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**.

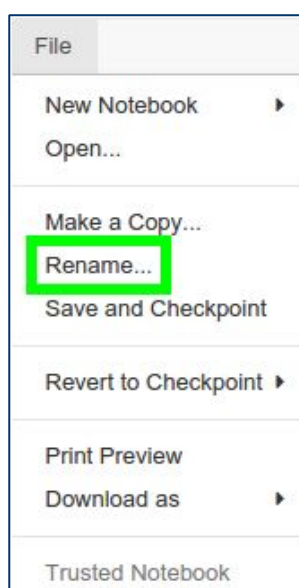


Figura 4



Rename Notebook

Enter a new notebook name:

Untitled

OK Cancel

Figura 5


Reemplace **Untitled** por **Taller_04** 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 el módulo NumPy de Python, un paquete diseñado para la manipulación de matrices, respectivamente. Escriba en la celda el comando que aparece a continuación y de clic en  (recuadro señalado de color rojo en la **Figura 6**) para ejecutar esa instrucción.

```
In [2]: import numpy as np
```

Bloque de código 1


Cargue el dataset ubicado en **/home/cursobigdata/Notebooks/Files/ExtramaritalAffairs_data.csv** a la variable **A**, que en este caso será una matriz. Escriba en la celda el comando que aparece a continuación y de clic en .

```
In [ ]: A=np.loadtxt('/home/cursobigdata/Notebooks/Files/ExtramaritalAffairs_data.csv', delimiter=',', skiprows=2)
```

```
In [ ]: otebooks/Files/ExtramaritalAffairs_data.csv', delimiter=',', skiprows= 2)
```


Bloque de código 2

```
A=np.loadtxt('/home/cursobigdata/Notebooks/Files/ExtramaritalAffairs_data.csv', delimiter=',', skiprows= 2)
```

Verifique que el número de filas y de columnas es de 6366 y 9 respectivamente. Escriba en la celda el comando que aparece a continuación y de clic en .


```
In [3]: A.shape  
Out[3]: (6366, 9)
```

Bloque de código 3

Cree una matriz de ceros con las mismas dimensiones de la matriz **A** (6366 filas, 9 columnas). Escriba en la celda el comando que aparece a continuación y de clic en .


```
In [4]: B= np.zeros(A.shape)
```

Bloque de código 4

Escoja las 8 primeras columnas de la matriz de ceros **B**. Escriba en la celda el comando que aparece a continuación y de clic en .

```
In [6]: B=B[:, 0:8]
```

Bloque de código 5

Normalice los valores de la matriz **A** y almacénelos en la matriz **B**. Escriba en la celda el comando que aparece a continuación y de clic en .

```
In [7]: for i in range(8):  
        B[:,i]= (A[:,i]-np.min(A[:,i]))/(np.max(A[:,i])-np.min(A[:,i]))
```

Bloque de código 6

Responda las siguientes preguntas:

- Nivel de satisfacción en su matrimonio (**V1**):
 - 5 → Muy bueno.
 - 4 → Bueno.
 - 3 → Regular.
 - 2 → Insatisfecho.
 - 1 → Muy insatisfecho.
- Edad (**V2**):
 - 17.5 → Por debajo de 20.
 - 22 → De 20 a 24.

27 → De 25 a 29.

32 → De 30 a 34.

37 → De 35 a 39.

42 → 40 o más.

● Número de años de casado (**V3**):

0.5 → Menos de 1 año.

2.5 → De 1 a 4 años.

6 → De 5 a 7 años.

9 → De 8 a 10 años.

13 → Más de 10 años de casado y el hijo mayor con edad inferior a 12 años.

16.5 → Más de 10 años de casado y el hijo mayor con entre 12 y 17 años de edad.

23 → Más de 10 años de casado y el hijo mayor con 18 años de edad o más.

● Número de hijos (**V4**):

0 → Ninguno.

1 → Uno.

2 → Dos.

3 → Tres.

4 → Cuatro.

5.5 → Cinco o más.

● Nivel de religiosidad (**V5**):

4 → Alto.

3 → Considerable.

2 → Leve

1 → Nulo.

● Nivel de educación (**V6**):

9 → Escuela primaria.

12 → Escuela secundaria.

14 → Educación superior (formal o no formal).

16 → Graduado de la universidad.

17 → Estudios de postgrado.


20 → Graduado de postgrado.

● Ocupación (**V7**):

- 6 → Profesional graduado de postgrado.
- 5 → Gerente, administrativo, negocios.
- 4 → Profesor, consejero, trabajador social, enfermero, artista, escritor, técnico, trabajador calificado.
- 3 → Ventas, oficinista, secretaria.
- 2 → Cultivo, agricultura, trabajador semi calificado o no calificado, otro.
- 1 → Estudiante.


● Ocupación de la pareja (**V8**):

- 6 → Profesional graduado de postgrado.
- 5 → Gerente, administrativo, negocios.
- 4 → Profesor, consejero, trabajador social, enfermero, artista, escritor, técnico, trabajador calificado.
- 3 → Ventas, oficinista, secretaria.
- 2 → Cultivo, agricultura, trabajador semi calificado o no calificado, otro.
- 1 → Estudiante.

Ahora, Cree un vector fila con sus respuestas escribiendo en la celda la instrucción que aparece a continuación y de clic en . Reemplace **V1**, **V2**, **V3**, **V4**, **V5**, **V6**, **V7** y **V8** por los números correspondientes a sus respuestas (ejemplo en el siguiente código).


```
In [9]: q= [5, 17.5, 9, 5.5, 3, 20, 5, 6]
```

Bloque de código 7

Normalice el vector **q** y asigne su valor a la variable **qn** escribiendo en la celda el comando que aparece a continuación y de clic en .


```
In [11]: qn= (q-np.min(A,0)[0:8])/(np.max(A,0)[0:8]-np.min(A,0)[0:8])
```

Bloque de código 8

Cree un vector columna para almacenar la diferencia entre su vector y cada una de las filas del dataset, escriba en la celda el comando que aparece a continuación y de clic en .


```
In [13]: r= np.zeros(B.shape[0])
```

Bloque de código 9

Cree una función que halle la diferencia entre cada uno de los registros presentes en el dataset y las respuestas que diligenció previamente. Escriba en la celda el comando que aparece a continuación y de clic en .

```
In [14]: diferencia= lambda x,y: np.sum(np.sqrt((x- y)**2))
```

Bloque de código 10

Compare el vector **qn** con el dataset. Escriba en la celda el comando que aparece a continuación y de clic en .


```
In [16]: for i in range(B.shape[0]):  
         r[i]= diferencia(qn, B[i,:])
```

Bloque de código 11

Encuentre la menor diferencia escribiendo en la celda el comando que aparece a continuación y de clic en .


```
In [17]: np.min(r)  
Out[17]: 1.3292517006802722
```

Bloque de código 12

Encuentre la fila del registro que tiene la menor diferencia escribiendo en la celda el comando que aparece a continuación y de clic en .

```
In [18]: np.argmin(r)  
Out[18]: 873
```


Bloque de código 13

Finalmente, visualice en la celda la última columna de la matriz A para la fila correspondiente a la mínima diferencia encontrada (el número de posibles infidelidades por año). Escriba en la celda el comando que aparece a continuación y de clic en .

```
In [20]: A[np.argmin(r), -1]  
Out[20]: 0.7424241999999998
```

Bloque de código 14

4. Guarde y cierre su Netbook

- 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 azul en la **Figura 7**).

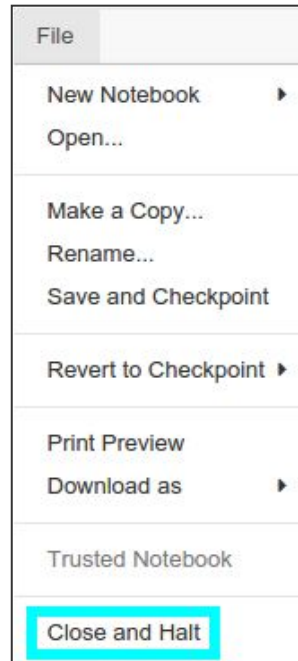


Figura 7

5. Desconectarse del servidor.

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

Fin de la Guía