



**ESCUELA POLITÉCNICA NACIONAL**  
**ESCUELA DE FORMACIÓN DE TECNÓLOGOS**

---

**BASE DE DATOS MULTIDIMENSIONAL**



ASIGNATURA:

BBMD

PROFESOR:

Ing. Juan Zaldumbide

PERÍODO ACADÉMICO:

## **Informe Recolección de datos de Twitter**

**TÍTULO:**

**INFORME: Recolección de datos de Twitter**

**ESTUDIANTE**

José Cortez

FECHA DE REALIZACIÓN: 23/ 07/ 2019

FECHA DE ENTREGA: 23/ 07/ 2019

CALIFICACIÓN OBTENIDA:

FIRMA DEL PROFESOR:



## Instalar librerías de Python para ejecución de scripts, de minería de datos en Twitter.

Librería couchdb. Para la instalación abrir una consola en modo administrador y luego ejecutar el comando “pip install couchdb”, previamente tener instalado Python y configurado variables de entorno.

```
c:\Users\Jose Cortez>pip install couchdb
Collecting couchdb
  Downloading https://files.pythonhosted.org/packages/ff/35/6660f7526c5d509b13264b27642de73754bd3d0addf56b175601c8b951e1/CouchDB-1.2-py2.py3-none-any.whl (67kB)
    100% |#####| 71kB 973kB/s
Installing collected packages: couchdb
Successfully installed couchdb-1.2
You are using pip version 19.0.3, however version 19.1.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.
```

Imagen 2 Instalación de la librería Couchd.

Librería tweepy. Comando “pip install tweepy”.

```
c:\Users\Jose Cortez>pip install tweepy
Collecting tweepy
  Downloading https://files.pythonhosted.org/packages/36/1b/2bd38043d22ade352fc3d3902cf30ce0e2f4bf285be3b304a2782a767aec/tweepy-3.8.0-py2.py3-none-any.whl
Collecting requests>=2.11.1 (from tweepy)
  Downloading https://files.pythonhosted.org/packages/51/bd/23c926cd341ea6b7dd0b2a00aba99ae0f828be89d72b2190f27c11d4b7fb/requests-2.22.0-py2.py3-none-any.whl (57kB)
    100% |#####| 61kB 786kB/s
Collecting PySocks>=1.5.7 (from tweepy)
  Downloading https://files.pythonhosted.org/packages/cd/18/102cc70347486e75235a29a6543f002cf758042189cb063ec25334993e36/PySocks-1.7.0-py3-none-any.whl
Collecting requests-oauthlib>=0.7.0 (from tweepy)
  Downloading https://files.pythonhosted.org/packages/c2/e2/9fd03d55ff70fe51f587f20bcf407a6927eb121de86928b34d162f0blac/requests_oauthlib-1.2.0-py2.py3-none-any.whl
Collecting six>=1.10.0 (from tweepy)
  Downloading https://files.pythonhosted.org/packages/bc/a9/01ffebfb562e4274b6487b4bb1ddec7ca55ec7510b22e4c51f14098443b8/charset-3.0.4-py2.py3-none-any.whl (133kB)
    100% |#####| 143kB 1.3MB/s
Collecting urllib3>=1.25.0, <1.25.1, <1.26, >=1.21.1 (from requests>=2.11.1->tweepy)
  Downloading https://files.pythonhosted.org/packages/e5/60/247f23a7121ae632d62811ba7f273d0e58972d75e58a94d329d51550a47d/urllib3-1.25.3-py2.py3-none-any.whl (150kB)
    100% |#####| 153kB 3.3MB/s
Collecting certifi>=2017.4.17 (from requests>=2.11.1->tweepy)
  Downloading https://files.pythonhosted.org/packages/69/1b/b853c7a9d4f6a6d00749e94eb6f3a041e342a885b87340b79c1ef73e3a78/certifi-2019.6.16-py2.py3-none-any.whl (157kB)
    100% |#####| 153kB 6.6MB/s
Collecting idna>=2.9, <=2.5 (from requests>=2.11.1->tweepy)
  Downloading https://files.pythonhosted.org/packages/14/2c/cd551d81dbe15200be1cf41cd03869a46fe72267450af7a6545bfc474c/idna-2.8-py2.py3-none-any.whl (58kB)
    100% |#####| 61kB ...
Collecting oauthlib>=3.0.0 (from requests-oauthlib>=0.7.0->tweepy)
  Downloading https://files.pythonhosted.org/packages/58/5e/289e98ff5ad1a321945803000c5f10f5f90eba346d13139ecdd075cfbe17/oauthlib-3.0.2-py2.py3-none-any.whl (143kB)
    100% |#####| 153kB ...
Installing collected packages: chardet, urllib3, certifi, idna, requests, PySocks, oauthlib, requests-oauthlib, six, tweepy
Successfully installed PySocks-1.7.0 urllib3-1.25.3 certifi-2019.6.16 chardet-3.0.4 idna-2.8 oauthlib-3.0.2 requests-2.22.0 requests-oauthlib-1.2.0 six-1.12.0 tweepy-3.8.0 urllib3-1.25.3
You are using pip version 19.0.3, however version 19.1.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.
```

Imagen 3 Instalación de librería Tweepy.

## Ejecución de los scripts.

El error más común al momento de ejecutar los scripts es el error 401 que es de autenticación. Imagen 4.

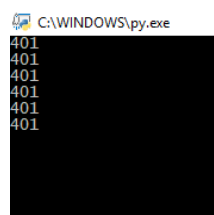


Imagen 4 Error de autenticación.

Para solucionar el error 401 se debe cambiar los tokens de validación. Imagen 5 y 6.

```
#####
###Credenciales de la cuenta de Twitter#####
#Poner aqui las credenciales de su cuenta privada, caso contrario la API bloqueara esta cuenta de ejemplo
ckey = "oPa1htMD07tLFKxCZfaHkdFa0"
csecret = "4iy1ovfrXThUEyoN10RRrHLn6bg7rNVEEw9NVEp21hDmvaFp5"
atoken = "115946548-pwHDPmtPgXybBDrlbRvpfsAzCRp0N7nKYkpaizx"
asecret = "ULn20xHUF9xj7a1gJHrYRqAvHREtjwr6QmHBUSiYB4xGD"
#####
```

Imagen 5 Token de validación fallido.

```
#####
###Credenciales de la cuenta de Twitter#####
#Poner aqui las credenciales de su cuenta privada, caso contrario la API bloqueara esta cuenta de ejemplo
ckey = "Lz3BzqqPJA7hrLE7XnY0ncdH"
csecret = "1MRn8E1Y8Ea24HBSXCvIrVTR0rx3vWA8WF0gEtaTH85FBQ4uY1"
atoken = "742503304056459264-a8gY11NQEPKAP308f73cUgUcoyIdZ00"
asecret = "L04oZ55U1XW81FPuvub78Tc0t1vTjAzINvaOecOMWxK6q"
#####
```

Imagen 6 Token valido de autenticación.

Nota: los tokens están el es script de Python.

Luego de cambiar los tokens de validación la recopilación de información se vera de la siguiente manera. Imagen 7.

```
C:\WINDOWS\spy.exe
guardado => 1153631769104658433
guardado => 1153631769242959872
guardado => 1153631769117122561
guardado => 1153631769301868544
guardado => 1153631769830264833
guardado => 1153631770136309760
guardado => 1153631770572668928
guardado => 1153631770602016768
guardado => 1153631770551697409
guardado => 1153631770893697793
guardado => 1153631770476081152
guardado => 1153631771042254848
guardado => 1153631771004551168
guardado => 1153631771153730432
guardado => 1153631771302518785
guardado => 1153631771835154432
guardado => 1153631772137185280
guardado => 1153631772162371590
guardado => 1153631772107661312
guardado => 1153631772573388800
guardado => 1153631772443389952
guardado => 1153631772522893312
guardado => 1153631773005209601
guardado => 1153631772950876165
guardado => 1153631773059715073
guardado => 1153631773311430658
guardado => 1153631773521088513
guardado => 1153631773689082411
guardado => 1153631773827448832
guardado => 1153631773739216896
guardado => 1153631774204936192
guardado => 1153631774146256896
guardado => 1153631774779629568
guardado => 1153631774783811584
guardado => 1153631775043805184
guardado => 1153631774783746048
guardado => 1153631775148728321
guardado => 1153631775291314176
guardado => 1153631775363372968
guardado => 1153631775429672960
guardado => 1153631775668572166
guardado => 1153631776142778368
guardado => 1153631776109191168
guardado => 1153631773382737920
guardado => 1153631776624873474
guardado => 1153631776767696897
guardado => 1153631776868990721
guardado => 1153631775022694401
guardado => 1153631777304371202
```

Imagen 7 Recopilación de datos de Twitter.

Vista de almacenamiento de datos en CouchDB.  
Vista de los datos opción tabla, imagen 8.

| _id                 | contributors | coordinates | created_at              | entities                  |
|---------------------|--------------|-------------|-------------------------|---------------------------|
| 1153631732186386... | null         | null        | Tue Jul 23 11:43:33 ... | {"hashtags": [], "urls... |
| 1153631733419372... | null         | null        | Tue Jul 23 11:43:33 ... | {"hashtags": [{"text...   |
| 1153631733608341... | null         | null        | Tue Jul 23 11:43:33 ... | {"hashtags": [{"text...   |
| 1153631733666983... | null         | null        | Tue Jul 23 11:43:33 ... | {"hashtags": [{"text...   |
| 1153631734442995... | null         | null        | Tue Jul 23 11:43:33 ... | {"hashtags": [{"text...   |
| 1153631734518308... | null         | null        | Tue Jul 23 11:43:33 ... | {"hashtags": [{"text...   |
| 1153631734774337... | null         | null        | Tue Jul 23 11:43:33 ... | {"hashtags": [{"text...   |
| 1153631735361335... | null         | null        | Tue Jul 23 11:43:33 ... | {"hashtags": [{"text...   |
| 1153631736032575... | null         | null        | Tue Jul 23 11:43:33 ... | {"hashtags": [{"text...   |
| 1153631736271712... | null         | null        | Tue Jul 23 11:43:34 ... | {"hashtags": [{"text...   |
| 1153631736313667... | null         | null        | Tue Jul 23 11:43:34 ... | {"hashtags": [{"text...   |

Imagen 8 Datos almacenados en CouchDB

Vista de datos opción Meta data, imagen 9.

| id                  | key                 | value   |
|---------------------|---------------------|---|
| 1153631732186386432 | 1153631732186386432 | { "rev": "1-4871907b68781a43d4996d9e88c04f58" |
| 1153631733419372546 | 1153631733419372546 | { "rev": "1-a2355e24f9eb7ee25d4606f1..."      |
| 1153631733608341504 | 1153631733608341504 | { "rev": "1-98432751fc94952e38253..."         |
| 1153631733666983936 | 1153631733666983936 | { "rev": "1-99421d5ebf8bbd084fbf17..."        |
| 1153631734442995712 | 1153631734442995712 | { "rev": "1-75559f362d7b5ae28971c..."         |
| 1153631734518308864 | 1153631734518308864 | { "rev": "1-95728566b594a027513fd5..."        |
| 1153631734774337536 | 1153631734774337536 | { "rev": "1-bbc355d36556d2b5f279da..."        |
| 1153631735361335296 | 1153631735361335296 | { "rev": "1-cdaa7c54331527f7a154273..."       |
| 1153631736032575488 | 1153631736032575488 | { "rev": "1-d1e1212ad79d91c1109294..."        |
| 1153631736271712258 | 1153631736271712258 | { "rev": "1-6652901dc280c28e9d4f8c7..."       |
| 1153631736313667592 | 1153631736313667592 | { "rev": "1-ce5fa812b801d460c76610..."        |

Imagen 9 Datos almacenados en CouchDB.

Vista de datos opción Json, imagen 10.

```

{
  "id": "1153631732186386432",
  "key": "1153631732186386432",
  "value": {
    "rev": "1-4871907b68781a43d4996d9e88c04f58"
  },
  "doc": {
    "id": "1153631732186386432",
    "rev": "1-4871907b68781a43d4996d9e88c04f58",
    "created_at": "Tue Jul 23 11:43:33 +0000 2019",
    "id_str": "1153631732186386432",
    "text": "You served. For 75 years, we've served you.\n\nOn this day in 1944, AMVETS was founded, and began assisting veterans... https://t.co/YcFpxHlxuk",
    "display_text_range": [
      0,
      140
    ],
    "source": "<a href='\"https://mobile.twitter.com/\"' rel='\"nofollow\"'>Twitter Web App</a>",
    "truncated": true,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null
  }
}

```

Imagen 10 Datos almacenados en CouchDB.

Base de datos que almacena los resultados de la recopilación, “prueba” la misma que debe ser colocada en el script de Python para que almacena la información en la tabla respectiva.

|        |         |      |
|--------|---------|------|
| prueba | 15.8 MB | 5678 |
|--------|---------|------|

Imagen 11 Base de datos que almacena la información extraída de Twitter.

Script de Python donde se realiza el cambio de nombre de la base.

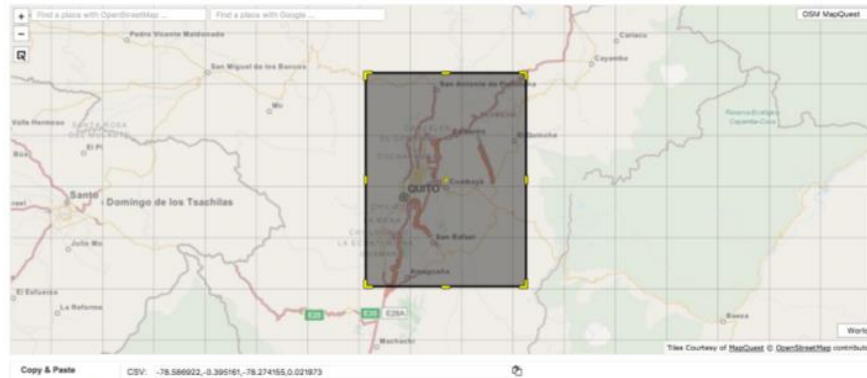
```

#Setear la URL del servidor de couchdb
server = couchdb.Server('http://localhost:5984/')
try:
    #Si no existe la Base de datos la crea
    db = server.create('prueba')
except:
    #Caso contrario solo conectarse a la base existente
    db = server['prueba']

```

Imagen 12 parte del script para asignar nombre a la base de datos.

Tipos de selección de la información.  
 Información por localización geográfica, imagen 13.



*Imagen 13 Recolección información por localización.*

Ubicación de la línea de código en el script de Python, imagen 14.

```
twitterStream.filter(locations=[-78.586922,-0.395161,-78.274155,0.021973])
```

*Imagen 14 Línea de código que filtra la información por localización.*

Resultado, imagen 15.

```
CA\WINDOWS\py.exe
Guardado => 1153638347170078720
Guardado => 1153638364765130752
Guardado => 1153638366644264960
Guardado => 1153638405215064069
Guardado => 1153638414564102144
Guardado => 1153638459992608768
Guardado => 1153638650023989249
```

*Imagen 15 Resultados de la ejecución del script.*

Información y filtrado de datos por palabras, imagen 16.

```
twitterStream.filter(track = ["ecuador", "lol", "darius"])
```

*Imagen 16 Filtrado de datos por palabras.*

Resultado, imagen 16.

```
CA\WINDOWS\py.exe
Guardado => 1153631769104658433
Guardado => 1153631769242959872
Guardado => 1153631769117122561
Guardado => 1153631769301868544
Guardado => 1153631769830264833
Guardado => 1153631770136309760
Guardado => 1153631770572668928
Guardado => 1153631770602016768
Guardado => 1153631770551697409
Guardado => 1153631770593697793
Guardado => 1153631770476081152
Guardado => 1153631771042254648
Guardado => 1153631771004551168
Guardado => 1153631771155730432
Guardado => 1153631771302516785
Guardado => 1153631771835154432
Guardado => 1153631772137185280
Guardado => 1153631772162371590
Guardado => 1153631772107661312
Guardado => 1153631772573388800
Guardado => 1153631772443389952
Guardado => 1153631772522893312
Guardado => 1153631773005096001
Guardado => 1153631772950876165
Guardado => 1153631773059715073
Guardado => 1153631773114306638
Guardado => 1153631773521088513
Guardado => 1153631773689098241
Guardado => 1153631773827448832
Guardado => 1153631773739216396
Guardado => 1153631774204936192
Guardado => 1153631774146256896
Guardado => 115363177479629568
Guardado => 1153631774783811584
Guardado => 1153631775043805184
Guardado => 1153631774783746048
Guardado => 1153631775148774573
```

*Imagen 17 resultado del filtrado por palabras.*

Nota: Al filtrar datos por ubicación se obtienen menor cantidad de resultados por el mismo tiempo de ejecución que por filtrado de palabras.

## Subir resultados al GitHub.

Crear repositorio e GitHub, imagen 18.

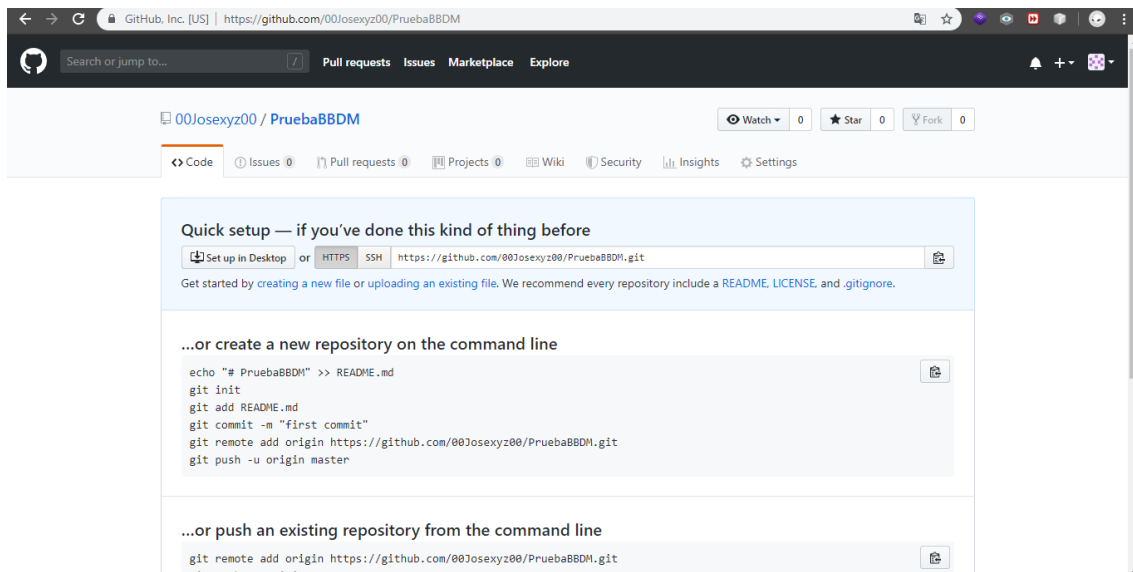


Imagen 18 Repositorio creado en GitHub.

Luego abrir una terminal de GitHub en la carpeta del proyecto e inicializarla con el comando “git init”.

```
Jose Cortez@JoseCortez MINGW64 ~/Desktop/Prueba_BBDM
$ git init
Initialized empty Git repository in C:/Users/Jose Cortez/Desktop/Prueba_BBDM/.git/
```

Imagen 19 Inicializar GitHub en carpeta del proyecto.

Después añadir lo que se va a subir con el comando “git add .” como se pretende subir todo lo que existe en la carpeta se añade un “.”, imagen 20.

```
Jose Cortez@JoseCortez MINGW64 ~/Desktop/Prueba_BBDM (master)
$ git add .
warning: LF will be replaced by CRLF in cosecha.py.
The file will have its original line endings in your working directory
warning: LF will be replaced by CRLF in tweets.html.
The file will have its original line endings in your working directory
```

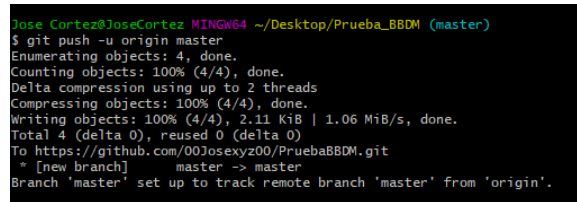
Imagen 20 Añadir los archivos que se van a subir a GitHub.

Adicionalmente, subir al repositorio local en el GitHub con comando “git commit -m ‘Prueba BBDM’”, imagen 21.

```
Jose Cortez@JoseCortez MINGW64 ~/Desktop/Prueba_BBDM (master)
$ git commit -m "Prueba BBDM"
[master (root-commit) e7ba971] Prueba BBDM
2 files changed, 128 insertions(+)
create mode 100644 cosecha.py
create mode 100644 tweets.html
```

Imagen 21 Subir archivos al repositorio local GitHub.

Por último, subir los archivos al repositorio remoto en GitHub con el comando “git push -u origin máster”, imagen 22.



```
Jose Cortez@JoseCortez MINGW64 ~/Desktop/Prueba_BBDM (master)
$ git push -u origin master
Enumerating objects: 4, done.
Counting objects: 100% (4/4), done.
Delta compression using up to 2 threads
Compressing objects: 100% (4/4), done.
Writing objects: 100% (4/4), 2.11 KiB | 1.06 MiB/s, done.
Total 4 (delta 0), reused 0 (delta 0)
To https://github.com/00Josexyz00/PruebaBBDM.git
 * [new branch]      master -> master
Branch 'master' set up to track remote branch 'master' from 'origin'.
```

*Imagen 22 Subida de archivos al repositorio remoto en GitHub.*

## 5 CONCLUSIONES

El único inconveniente que se puede presentar al momento de realizar la practica es, no tener los tokens de validación, la descarga de la información de Twitter para lo que se recomienda primero crear una cuenta y tenerla habilitada.