



Corporación Unificada Nacional  
de Educación Superior

**FABIAN ANDRES LOPEZ GIL**

**DIPLOMADO MACHINE LEARNING PYTHON**

**ACTIVIDAD 3**

**CUN**

**2022**

Contenido

INTRODUCCION ..... 3

OBJETIVO ..... 3

INSTALACION ..... 4

LIBRERIAS ..... 4

CODIGO ..... 4

## INTRODUCCION

Python es un lenguaje de programación que cada vez se utiliza más por las empresas y programadores que trabajan con datos (Business Intelligence, Integración de datos, Data Science, Machine Learning, Big Data...). El motivo de que cada vez cobra más importancia en su uso es en la gran cantidad de librerías existentes para realizar prácticamente todo y más aún si el objetivo es trabajar y gestionar datos, también por lo optimizado que está Python respecto a JAVA.

Para redactar este artículo he desarrollado un proceso ETL con Python desde cero en el que voy a explicar paso a paso todo lo utilizado y porqué. Para desarrollar el proceso me he inventado un objetivo y así dar sentido al desarrollo.

## OBJETIVO

La librería snsrape es un distribuidor de data para servicios de redes sociales (SNS). Nos puedes ofrecer cosas como perfiles de usuario, hashtags o búsquedas y devuelve los elementos descubiertos, por ejemplo, las publicaciones relevantes.

Desde esta base vamos a recolectar todas las personas que hayan escrito algo positivo en sus tweets, con esto, podríamos analizar la cantidad de gente que está satisfecha en la red social y que tan frecuente es para los usuarios publicar sus logros y compartirlos para la comunidad

Además de todo esto, hay que registrar en una tabla de control todo lo que va sucediendo para poder consultar qué ficheros se han procesado, si se han procesado bien, cuánto han tardado y cuando se han procesado.

## INSTALACION

### LIBRERIAS

Vamos a instalar las siguientes librerías:

- pip install numpy
- pip install pandas
- pip install iso8601
- pip install deep\_translator
- pip install mysql
- pip install pwin

pwin nos va servir para instalar:

- pwin install snsrape

```
import csv
import datetime
import uuid
import iso8601 as iso8601
import numpy
import snsrape.modules.twitter as sntwitter
import pandas as pd
from deep_translator import GoogleTranslator
from mysql import connector
```

### CODIGO

Vamos a crear una primera función que va ser nuestra configuración para la conexión a la base de datos. En este caso, configurado para MySQL:

```
def connection():
    conexion = connector.connect(host='localhost', user='root', password='')
    return conexion
```

La segunda función, la vamos a crear para la extracción de los datos:

```
def extraccion():
    maxTweets = 100
    tweets_list2 = []
    for i, tweet in enumerate(sntwitter.TwitterSearchScraper('feliz since:2020-08-01 until:2022-12-01').get_items()):
        if i > maxTweets:
            break
        tweets_list2.append([
            tweet.date,
            tweet.url,
            tweet.lang,
            tweet.retweetCount,
            tweet.id,
            tweet.content,
            tweet.user.username
        ])

    tweets_df2 = pd.DataFrame(tweets_list2,
                              columns=[
                                  'Datetime',
                                  'url of tweet',
                                  'lenguaje',
                                  're-tweets',
                                  'Tweet Id',
                                  'Text',
                                  'Username'
                              ])

    tweets_df2.head()
    tweets_df2.to_csv('text-query-tweets.csv', sep=',', index=False)
```

- maxTweets: Esta variable la definimos para poner un limite de datos
- tweets\_list2: definimos la variable en un arreglo vacio
- for i, tweet: este ciclo va recolecta la información que traemos de la librería snsrape llamando el método de Twitter y ponemos un enumerate el método agrega un contador a un objeto iterable y lo devuelve en forma de objeto enumerador.
- Este objeto enumerado va a insertar con el append a nuestra variable vacia tweets\_list2. Esto asignando unos elementos [date, url, Lang... etc]
- Una vez hecho lo anterior, vamos a utilizar pandas para el DataFrame con nuestra matriz ya construir con el append
- El head() devuelve los primeros elementos de la estructura
- to\_csv nos ayudara a leer el archivo ya con data de Excel, el cual vamos a extraer para llevarlo a la base de datos.

### Patrón:

Aquí el bucle va analizar todos los comentarios que contengan la palabra feliz entre un rango de fecha entre el 2020-08-01 y 2022-12-01.

Tercera función, transformación:

```
def transform():
    data = []
    with open('text-query-tweets.csv', encoding="utf8") as csvfile:
        spamreader = csv.reader(csvfile, delimiter=',')

        for row in spamreader:
            if (spamreader.line_num > 1):
                row[0] = datetime.datetime.fromisoformat(row[0])
                row[5] = row[5].format("utf-8")
                if (row[2] != 'es'):
                    translation = GoogleTranslator(
                        source='auto', target='es').translate(row[5])
                    row[5] = translation
                    row[5] = row[5].encode('ascii', errors='ignore')
                data.append(row)
    return data
```

Definimos una variable en vacío llamada data y posteriormente continuamos con abrir el archivo csv codificado en UTF-8

Después de leer el archivo, pasaremos al bucle for. Primero con una validación de que si es mayor a 1 continúe con el bucle formateando la hora desde el nivel 1 del arreglo y llevando una traducción a cabo con la librería deep\_translator. Con esta librería, definimos una búsqueda auto y un target que contendrá el idioma que queremos traducir la data de twitter. Seguido del método translate con el elemento del for "row[5]" que es la posición de los comentarios.

Finalizamos el bucle, insertando con append en nuestra variable data.

Cuarta función, Cargar o destino de salida

```
def load(dataclean):
    cur = dbconn.cursor()
    arraySize = len(dataclean)
    for r in range(0, arraySize):
        try:
            uuidstr = uuid.uuid4()
            cur.execute('SET NAMES utf8mb4')
            cur.execute("SET CHARACTER SET utf8mb4")
            cur.execute("SET character_set_connection=utf8mb4")
            cur.execute(
                """INSERT INTO tweets(id,fecha,url, lenguaje,`re-tweets`, `tweet ID`, contenido, usuario) VALUES( %s,%s,%s,%s,%s,%s,%s,%s)""",
                (str(uuidstr), dataclean[r][0], dataclean[r][1], dataclean[r][2], dataclean[r][3], dataclean[r][4],
                 dataclean[r][5], dataclean[r][6]))
            dbconn.commit()
        except ValueError:
            print("Error")
            print(ValueError)
    pass
```

Definimos una variable llamando nuestra conexión y el método cursor() que nos ayudara a ejecutar sentencias sql en nuestra base de datos.

Iniciamos un nuevo bucle, en caso de ser la conexión exitosa vamos a continuar con los parámetro de la base y su codificación y en seguida con insertar o cargar la data que hemos definido anteriormente.

En esta última imagen, definimos el proceso de ejecución de la aplicación.

- Proceso de extracción
- Proceso de transformación
- Con numpy creamos otra matriz con estos nuevos datos los cuales vamos a imprimir mientras carga la data a la base de datos
- Show\_db\_query: Aquí ponemos el nombre de la base de datos que vamos a utilizar
- Después dbconn para llamar la primera función que hicimos para la conexión
- Ejecutamos las sentencias
- Con Load, termina la carga total a la base de datos.

```
if __name__ == '__main__':

    print('realizando extracción')
    extraccion()

    print('realizando transformación')
    dataclean = transform()
    dataclean.pop(0)]
    an_array = numpy.array(dataclean)
    print(an_array)
    show_db_query = "use etl"
    dbconn = connection()
    cur = dbconn.cursor()
    cur.execute(show_db_query)
    for x in cur:
        print(x)

    print('realizando carga de datos')
    load(dataclean)
```

```
[datetime.datetime(2022, 11, 30, 23, 59, 24, tzinfo=datetime.timezone.utc)
'https://twitter.com/sandraalfaro/status/1598104439969415168' 'es' '11'
'1598104439969415168'
'Feliz Cumpleaños a nuestro querido Compañero @MarioBuffone70. Orgullosos estamos los Adecos, de tenerlo entre nosotros @ADemocratica Mucha salud y Bendicio
'sandraalfaro']
[datetime.datetime(2022, 11, 30, 23, 59, 23, tzinfo=datetime.timezone.utc)
'https://twitter.com/Disse_minando/status/1598104437696131073' 'pt' '0'
'1598104437696131073'
'b'@LenisAlessandra Los elementos\nEspere\ntiene tema\nqueriendo un camino\nhablar con esta mujer\n\nsteve kerchiel\n\nCuidate y no olvides ser feliz nisa'
'Disse_minando']
[datetime.datetime(2022, 11, 30, 23, 59, 23, tzinfo=datetime.timezone.utc)
'https://twitter.com/fabrozaurio/status/159810443722058241' 'es' '0'
'159810443722058241'
'Yo salto 91cm, ayer jugando me dijeron que les parecía increíble que eso pase y me felicitaron, que lindo que todo el trabajo del gym se note. Toi feliz'
'fabrozaurio']]
realizando carga de datos
```

Mostrando filas 0 - 24 (total de 100. La consulta tardó 0.002 segundos.)

SELECT \* FROM 'tweets'

Parfilando [ Editar en línea ] [ Editor ] [ Explicar SQL ] [ Crear código PHP ] [ Actualizar ]

1 > >> Mostrar todo | Número de filas: 25 | Filtrar filas: Buscar en esta tabla

Opciones extra	id	fecha	url	lenguaje	re-tweets	sevent ID	contenido	usuario
	16483571-4873-4457-b479-e41819090567	2022-11-30 23:59:59	https://twitter.com/fandemmarth/status/1598104408	pt	0	1598104408057052161	@T48UNTO hombre cuando lo vi me alegró mucho por l...	fandemmarth
	87646262-9e07-4516-bd96-052b23262d59	2022-11-30 23:59:58	https://twitter.com/fansantamaria/status/1598104584	pt	0	1598104584236720128	@Vida0881 Feliz cumpleaños	laraSantoma
	6e07334-135e-4460-895e-3c8f02c5c04	2022-11-30 23:59:58	https://twitter.com/yelinah/status/159810450417730	es	0	15981045041773017856	@yelinahonky también me salió el proter de han...	yelinah
	u1c3564-4b62-4a62-8c3f-8d05fca9616	2022-11-30 23:59:58	https://twitter.com/renatibhaca/status/159810454583	pt	0	1598104503300897024	deseada felz con todo	renatibhaca
	89e494-4744-4350-85ac-8c76a55980	2022-11-30 23:59:58	https://twitter.com/kar1n0/status/1598104525296	es	0	1598104503249646592	@krisofel Feliz	krisofel
	2b6d5d5-822d-4c17-8ac3-418d64153b6	2022-11-30 23:59:57	https://twitter.com/kar1n0/status/159810452017515	es	0	1598104502017515520	Recuerda mi donada seri, pero feliz https://t.co/...	y0_j4
	3653794-4f0c-44a4-433b-5a97956a7b4	2022-11-30 23:59:57	https://twitter.com/dmwaalk/status/15981045789550	pt	1	1598104578955038976	Nunca ser feliz hasta que hable ocho idiomas	dmwaalk
	66b344b-4b4b-4b4a-4b4a-7833b08b4a8	2022-11-30 23:59:55	https://twitter.com/aimalima/status/1598104578247	pt	0	1598104578247266304	sin miedo a ser feliz https://t.co/YVd0GdRd	Faimalima
	174e964-1008-4130-bac5-38d6b5926e	2022-11-30 23:59:57	https://twitter.com/Noemajk1/status/1598104577509	pt	0	1598104577509064704	No puedo creer que no le dije felz cumpleaños mjk	Noemajk1
	95b6301-5404-4a8f-9270-8149052a3a	2022-11-30 23:59:57	https://twitter.com/LopesAntonia/status/1598104576	und	0	1598104576569526130	@LootBorges Kkkkkk	LopesAntonia
	4c7b5b5-554-4b62-472c-a6071a3a598	2022-11-30 23:59:57	https://twitter.com/falaanderson/status/159810457	pt	0	159810457639333937	Qu le hace felz? https://t.co/Rfow0C27X0	falaanderson
	7658715-24ac-4878-90a8-9116919bde19	2022-11-30 23:59:56	https://twitter.com/Lucia537773/status/159810457	pt	0	1598104575319642112	@ParinhoDaValen espero que ningun loco y despiadad...	Lucia537773
	00a8913-6d8f-4797-8a82-4a6f7196b26	2022-11-30 23:59:55	https://twitter.com/Colpensiones/status/1598104575	es	0	1598104575130856655	@parjatosTV Hola, puedes consultar tu #HistoriaLa...	Colpensiones
	a7b6457-36c0-441e-568f-a26ba157108	2022-11-30 23:59:55	https://twitter.com/obograph/status/1598104571359	pt	0	1598104571359817728	S. estoy muy feliz de tener una nueva foto de Kati...	obograph
	1bfa4b6-3953-4c5d-8031-a12ac0b62b1	2022-11-30 23:59:55	https://twitter.com/magical_bitch/status/159810456	pt	0	1598104570336792676	ODIO A CASSANDRA CLARE! ELLA ME PAGA CON EL FINAL...	magical_bitch
	60d87db-1650-4c9e-408f-4949d818352	2022-11-30 23:59:55	https://twitter.com/Raizaguar21/status/159810456	es	0	1598104569162403540	lo feliz que me hace salir con mis amigos	Raizaguar21
	elb0d5d0-4d75-4341-a459-0ba8b6319c	2022-11-30 23:59:54	https://twitter.com/ReynaldQuinter/status/15981045	es	0	1598104565479780544	@AndyPandyCos Feliz cumpleaños para nuestra talent...	ReynaldQuinter
	x30912d3-32ca-4341-8ba7-508419a0e417	2022-11-30 23:59:54	https://twitter.com/strkuxulia/status/159810456463	pt	0	1598104564632522753	lo me casela con bellingham	strkuxulia
	732a724-63e4-4a4d-a887-99952406061	2022-11-30 23:59:54	https://twitter.com/alemc008/status/1598104564318	es	0	1598104564531879930	No me acordaba lo que me gustaba Granada hasta que...	alemc008
	ca0a2e68-29e6-4915-9856-98a0a0b0d401	2022-11-30 23:59:53	https://twitter.com/mathusafmr15/status/1598104563	pt	0	1598104563252596736	Hoy me di cuenta que perdí a una persona increíble a...	mathusafmr15
	f4a1a4d3-70d4-438a-4a02-4b3f849a6a5	2022-11-30 23:59:53	https://twitter.com/hydejiada/status/159810463310	pt	0	1598104563185516545	No entendi si es para ser feliz o mentar https://t.c...	hydejiada

1/25