Corporación Unificada Nacional
de Educación Superior

FABIAN ANDRES LOPEZ GIL

DIPLOMADO MACHINE LEARNING PYTHON

ACTIVIDAD 3

CUN

2022

Contenido

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 snscrape 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, sí 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 snscrape

```
import csv
import datetime
import uuid
import iso8601 as iso8601
import numpy
import snscrape.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-twittes',
                                  '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 snscrape 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 continue 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/1598104434722058241' 'es' '0'
 '1598104434722058241'
 '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
```