

PROYECTO FINAL

Minería de datos con twitter y mensajes para sensores MQTT

Dulce María González Reyes A01745835

Fernando Joshua Alvarado Ortiz A01748693

Jorge Isidro Taymani Gates Zuckerberg A01745907

José Luis Madrigal Sánchez A01745419

Alejandro Bastida Cortés A01746457

José Ángel García Gómez A01745865

Mayo, 2020

Instituto Tecnológico y de Estudios Superiores de Monterrey

Innovación tecnológica

Departamento de Ciencias

PROYECTO FINAL

¿Qué hicimos?

Primero que nada, creamos una base de datos con firebase en la cual podríamos suministrar información. Luego por medio de un código de Python y Twitter API, fuimos colocando en la base diversos trending topics durante varios días. Básicamente se sacaba toda la información de los hashtags, como las palabras, el screen name, tweets, etc. A continuación se muestra el código utilizado y la firebase con los datos:

The image displays two screenshots related to a project. The top screenshot shows a Visual Studio Code editor window titled 'twiterv5.py - Visual Studio Code'. The editor contains a Python script with the following code:

```
1 # Set this variable to a trending topic, or anything else
2 # for that matter. The example query below was a
3 # trending topic when this content was being developed
4 # and is used throughout the remainder of this chapter.
5 import twitter
6 import json
7 from collections import Counter #contador sencillo
8 from prettytable import PrettyTable #libreria para hacer tablas bonitas
9 from firebase import firebase
10 import sys
11 import io
12 sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')
13 sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')
14 q = '#ObamaGate'
15
16 CONSUMER_KEY = 'CK0GtRYmjIZxbSft5q87CPuM5'
17 CONSUMER_SECRET = 'Gs8REP8qigYbHoqmMdB1yRh7FUF6xQbNzOmrxuCHybTRoA546'
18 OAUTH_TOKEN = '1220398224768425984-vAx1IM9YPEaNMq2tsfysMa0mPdgB16'
19 OAUTH_TOKEN_SECRET = 'cbjHHAfsv3UDz6idiyAWQDa3C3gkqF1aa98n21UNC92Bz'
20
21 WORLD_WOE_ID = 1
22 US_WOE_ID = 23424977
23 MX_WOE_ID = 23424980 #trends Ciudad de Mexico
24 count = 100
25
```

The bottom screenshot shows the Firebase Database console. The left sidebar contains the Firebase logo and a navigation menu with options: Información general de..., Desarrollo, Authentication, Database, Storage, Hosting, Functions, ML Kit, Calidad, Analytics, and Spark. The main area is titled 'Database' and shows a tree view of the database structure. The tree view shows a root node 'tec-twitter' with a subnode 'tec-twitter' which has five children: Hashtags, Screen_Names, Status_text, Tweet, and Words. The URL bar shows 'https://tec-twitter.firebaseio.com/'.

PROYECTO FINAL

Resultados

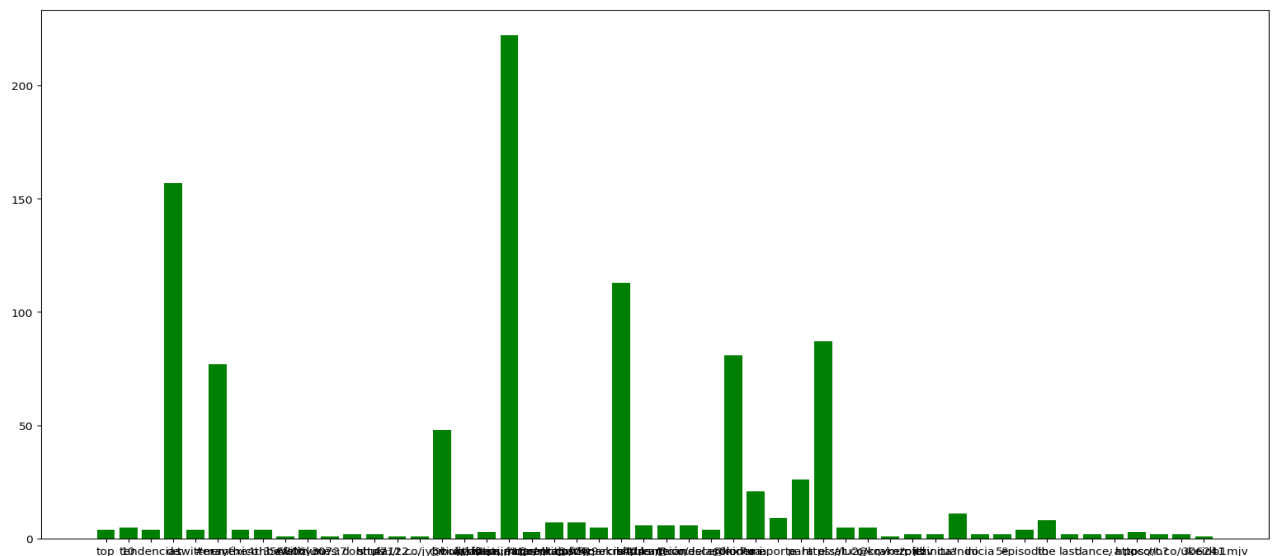
Después de tener los hashtags, hicimos un conteo de las palabras más repetidas en todos ellos, mostrando las cifras por medio de un histograma. Asimismo, nosotros podíamos establecer cuál número de palabras queríamos visualizar, en nuestro caso, terminamos poniendo 50, ya que queríamos tener un panorama más amplio para poder encontrar cosas más interesantes. Por lo tanto, se muestra el código y el histograma obtenido:

The image shows the Visual Studio Code editor interface. The top bar displays the menu (File, Edit, Selection, View, Go, Run, Terminal, Help) and the active file (fire9.py - Visual Studio Code). The file explorer on the left shows a project structure with folders like 'firebase' and 'twitter' and files like 'fire9.py', 'fire5.5.py', and 'twitterformat.py'. The main editor area displays the code for 'fire9.py', which is a Python script that uses the 'firebase-admin' SDK to connect to a Firebase database and retrieve word frequency data. The script includes imports for 'firebase', 'json', 'matplotlib.pyplot', 'itertools', 'io', 'sys', and 'collections.Counter'. It then uses 'firebase.FirebaseApplication' to connect to the database and 'firebase.get' to retrieve the data. The data is then processed to extract word frequencies. The bottom status bar shows the 'OUTPUT' tab selected, with a 'Code' dropdown and icons for Problems, Output, Debug Console, and Terminal.

```

1  from firebase import firebase
2  import json
3  import matplotlib.pyplot as plt
4  import itertools
5  import io
6  import sys
7  from collections import Counter
8
9  sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')
10 sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')
11
12 #conexion a la base de datos
13 firebase = firebase.FirebaseApplication('https://tec-twitter.firebaseio.com/', None)
14
15 #codigo que saca una sola instancia de palabras de la base de datos y cuenta la frecuencia de la misma
16
17 result = firebase.get('/tec-twitter/Words', '')
18
19 palabras=[]
20 for x in result:
21     palabras.append(result[x])
22 print (palabras[23])
23
24 pal=[]
25 for x in palabras[23]:

```



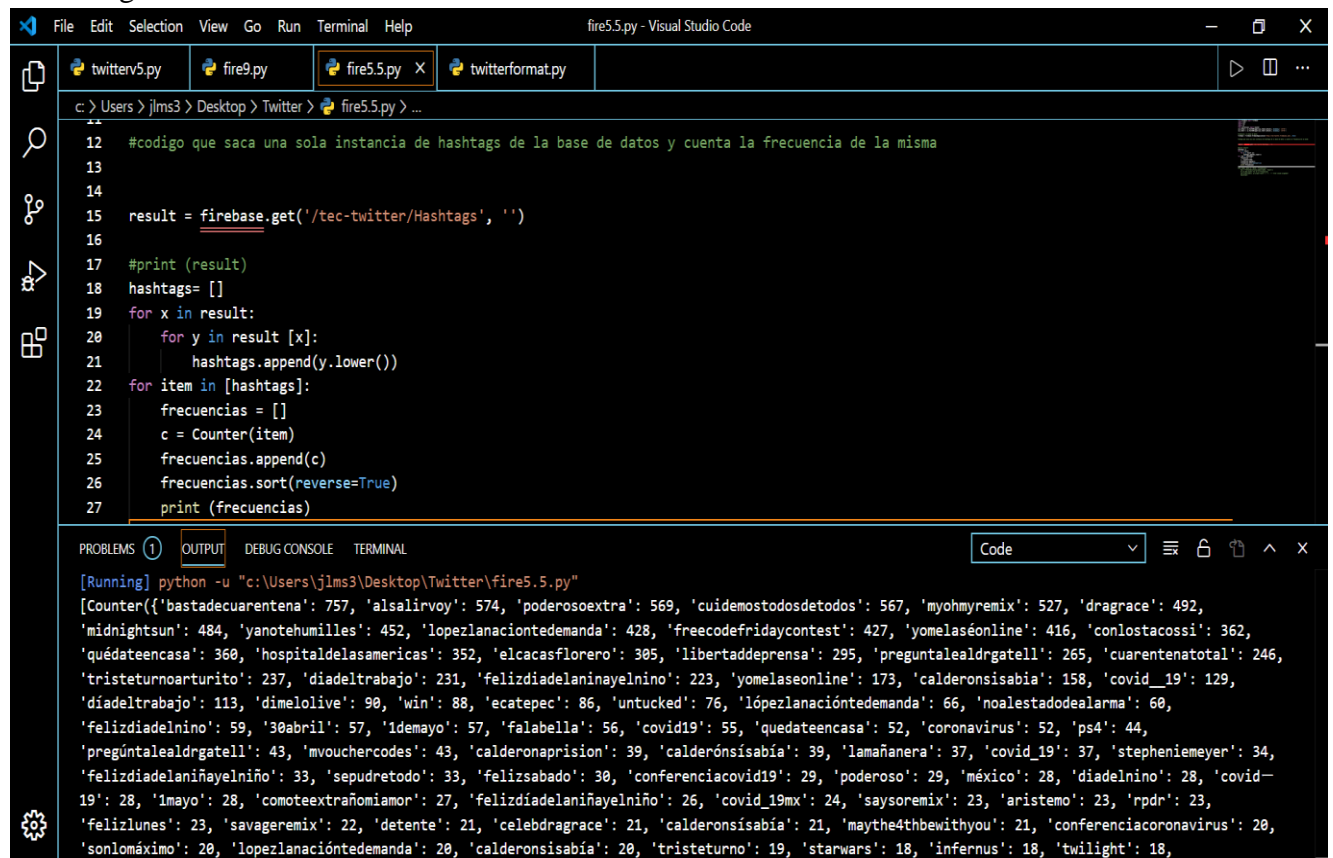
Ya que no se pueden visualizar muy bien las palabras, se explicarán de manera general. La que obtuvo una mayor frecuencia, siendo mayor a 200, fue la de #tristetornoarturito que fue tendencia desde el 4 de mayo por el Star Wars Day. En cambio, las demás palabras con muchas repeticiones son básicamente artículos y preposiciones del

PROYECTO FINAL

español, como “en”, “la”, “de”. Aunque también se puede ver la palabra “twitter” y el “rt por obvias razones, e inclusive varias palabras forman un pequeño fragmento de algún mismo tweet, como (“mi”, “aporte”, “para”, “el”) o (“piel”, “chinita”, “cuando”, “inicia”, “5º”, “episodio”) o (“the”, “last”, “dance”), siendo el último ejemplo una relación con la serie de los Bulls de la NBA en Netflix.

Información recaudada

Por último, por medio de otro código de python, generamos un contador de los hashtags de la base y mostramos las frecuencias de manera descendente, para que así pudiéramos ver aquellos que se repetían más. Y observamos que muchos de ellos están relacionados con la cuarentena, el COVID 19 e inclusive el gobierno. Y hasta abajo de la imagen se observa claramente que hashtags de Star Wars y tristeterno fueron tendencia también, por eso se pudieron ver muchas palabras relacionadas en las mayores frecuencias del histograma.



The screenshot shows a Visual Studio Code window with the file `fire5.5.py` open. The code is a Python script that queries a Firebase database for hashtags and counts their frequencies. The output window shows the results of the script execution.

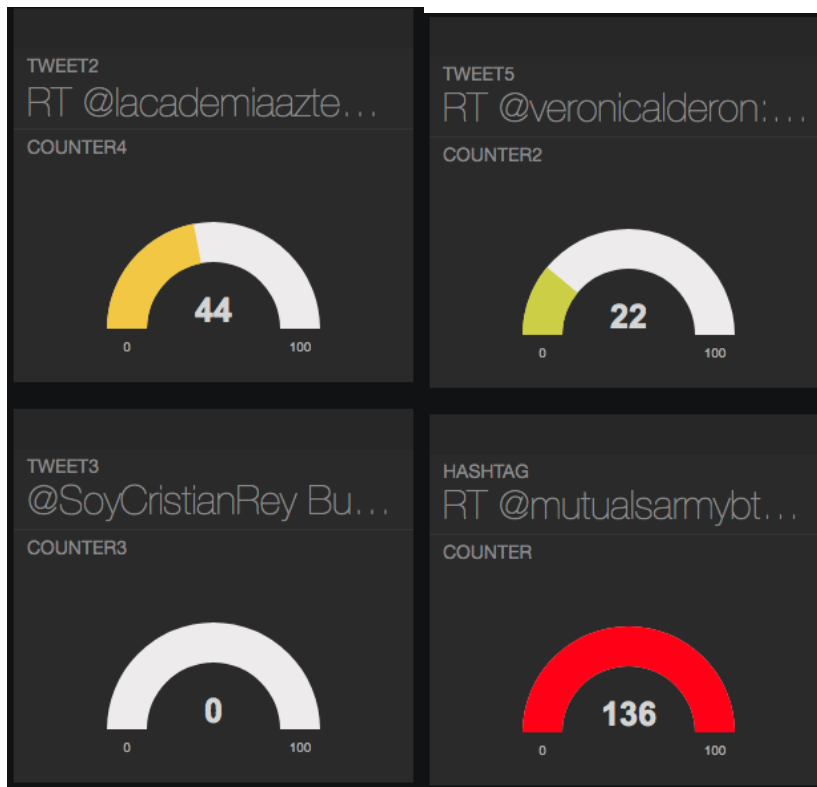
```
12 #codigo que saca una sola instancia de hashtags de la base de datos y cuenta la frecuencia de la misma
13
14
15 result = firebase.get('/tec-twitter/Hashtags', '')
16
17 #print (result)
18 hashtags= []
19 for x in result:
20     for y in result [x]:
21         hashtags.append(y.lower())
22 for item in [hashtags]:
23     frecuencias = []
24     c = Counter(item)
25     frecuencias.append(c)
26     frecuencias.sort(reverse=True)
27     print (frecuencias)
```

OUTPUT

```
[Running] python -u "c:\Users\j\ms3\Desktop\Twitter\fire5.5.py"
[Counter({'bastadecuarentena': 757, 'alsalirvoy': 574, 'poderosoextra': 569, 'cuidemostodosdetodos': 567, 'myohmyremix': 527, 'dragrace': 492, 'midnightsun': 484, 'yanotehumilles': 452, 'lopezlanaciontedemanda': 428, 'freecodefridaycontest': 427, 'yomelaseonline': 416, 'conlostacossi': 362, 'quedateencasa': 360, 'hospitaldelasamericas': 352, 'elcacasflorero': 305, 'libertaddeprensa': 295, 'preguntalealdratell': 265, 'cuarentenatotal': 246, 'tristeternoarturito': 237, 'diadeltrabajo': 231, 'felizdiadelaninayelnino': 223, 'yomelaseonline': 173, 'calderonsisabia': 158, 'covid_19': 129, 'diadeltrabajo': 113, 'dimelolive': 90, 'win': 88, 'ecatepec': 86, 'untucked': 76, 'lopezlanaciontedemanda': 66, 'noalestadodealarna': 60, 'felizdiadelnino': 59, '30abril': 57, '1demayo': 57, 'falabella': 56, 'covid19': 55, 'quedateencasa': 52, 'coronavirus': 52, 'ps4': 44, 'preguntalealdratell': 43, 'mvouchercodes': 43, 'calderonaprision': 39, 'calderonsisabia': 39, 'lamananera': 37, 'covid_19': 37, 'stepheniemeyer': 34, 'felizdiadelaninayelnino': 33, 'sepudretodo': 33, 'felizsabado': 30, 'conferenciacovid19': 29, 'poderoso': 29, 'mexico': 28, 'diadelnino': 28, 'covid-19': 28, '1mayo': 28, 'comoteextrañomiamor': 27, 'felizdiadelaninayelnino': 26, 'covid_19mx': 24, 'saysoremix': 23, 'aristemo': 23, 'rpdr': 23, 'felizlunes': 23, 'savageoremix': 22, 'detente': 21, 'celebratdragrace': 21, 'maythe4thbewithyou': 21, 'conferenciacoronavirus': 20, 'sonlomaximo': 20, 'lopezlanaciontedemanda': 20, 'calderonsisabia': 20, 'tristeterno': 19, 'starwars': 18, 'infernus': 18, 'twilight': 18,
```

Después de esto, por medio de freeboard.io hicimos un dashboard con la información de la base. Cabe mencionar que tuvimos que exportar un archivo json de la firebase para que así el profesor pudiera ponerlo en su página, de la cual tomaríamos los datos. Por lo que, al estar conectados al json, pudimos hacer paneles de tipo texto con algunos hashtags o tweets, para luego poner otro de tipo gauge con el contador, para poder ver su presencia y saber su frecuencia. A continuación se muestran algunos:

PROYECTO FINAL



Conclusión

Podemos decir que este proyecto nos gustó mucho, ya que pudimos conocer más sobre el área de analítica y encontrar más funciones de Python. Asimismo, fue bastante interesante el poder utilizar códigos que hicieran uso de información de una red social, ya que de verdad se puede sacar mucha información. Pero más importante, aprendimos a analizar datos y sacar cosas importantes o en todo caso, encontrar algo específico.