

PROYECTO FINAL

Data mining with twitter and messages for MQTT sensors

Dulce María González Reyes A01745835

Fernando Joshua Alvarado Ortiz A01748693

Jorge Isidro Gates Zuckerberg A01745907

José Luis Madrigal Sánchez A01745419

Alejandro Bastida Cortés A01746457

José Ángel García Gómez A01745865

May, 2020

Instituto Tecnológico y de Estudios Superiores de Monterrey

Technological Innovation

Science Department

PROYECTO FINAL

¿What did we do?

First of all, we created a database with firebase in which we could supply information. Then, by means of a Python code and Twitter API, we placed various trending topics in the database for several days. Basically, all the information of the hashtags was extracted, such as the words, the screen name, tweets, etc. Below is the code used and the firebase with the data:

The image shows two screenshots. The top screenshot is a Visual Studio Code editor window titled 'twiterv5.py - Visual Studio Code'. It displays a Python script that uses the Twitter API to fetch trending topics and store them in a Firebase database. The script includes imports for 'twitter', 'json', 'collections', 'Counter', 'PrettyTable', and 'firebase'. It sets up a Twitter client with consumer key, secret, auth token, and secret. It then fetches trending topics for a specific location (Mexico) and stores them in a Firebase database. The bottom screenshot is the Firebase console, showing the 'Database' section for a project named 'tec-twitter'. The database is a Realtime Database. The structure of the database is shown as a tree view with the following nodes: 'tec-twitter' (root), 'Hashtags', 'Screen_Names', 'Status_text', 'Tweet', and 'Words'.

```
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 = 'Gs8REPBqigYbHoqmMdB1yRh7FUF6xQbMzOmrqxuCHybTRoA546'
18 OAUTH_TOKEN = '1220398224768425984-vAx1IM9YPEaNMq2tsfysMa0mPdgB16'
19 OAUTH_TOKEN_SECRET = 'cbjHHAfsv3UDz61diYAWQDa3C3gkqF1aa98n21UNC92Bz'
20
21 WORLD_WOE_ID = 1
22 US_WOE_ID = 23424977
23 MX_WOE_ID = 23424980 #trends Ciudad de Mexico
24 count = 100
25
```

Database

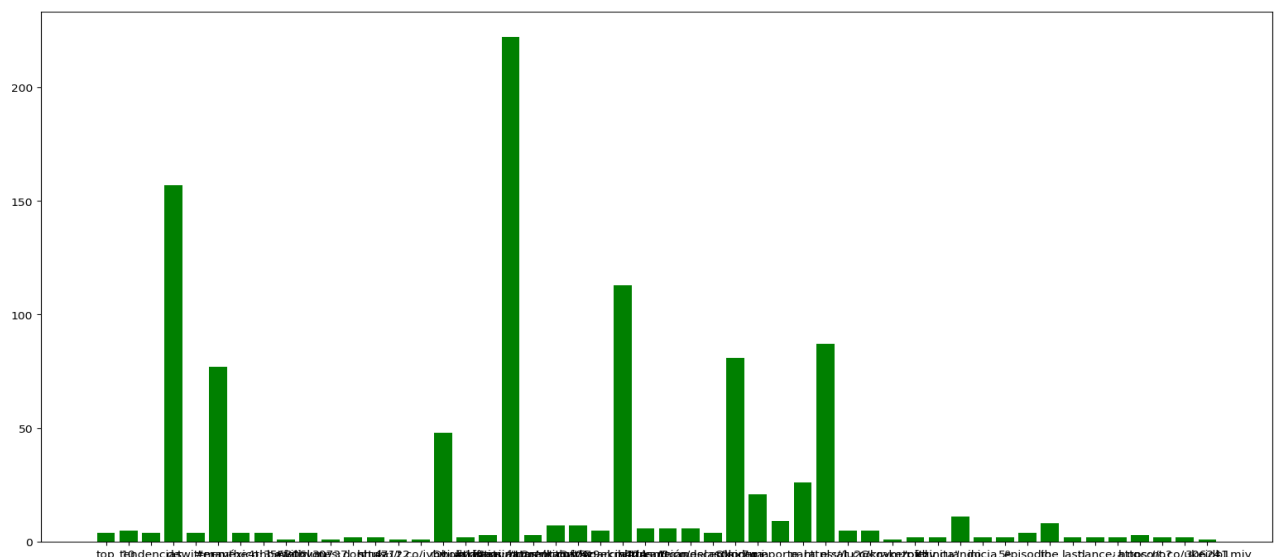
Se han activado los modos de solo lectura y estático en el visor de datos para mejorar el rendimiento del navegador. Selecciona una clave que tenga menos registros para editarlos o verlos en tiempo real

- tec-twitter
 - Hashtags
 - Screen_Names
 - Status_text
 - Tweet
 - Words

PROYECTO FINAL

Results

After having the hashtags, we made a count of the most repeated words in all of them, showing the figures by means of a histogram. Likewise, we could establish what number of words we wanted to visualize, in our case, we ended up putting 50, since we wanted to have a broader panorama to be able to find more interesting things. Therefore, the code and the histogram obtained are shown:



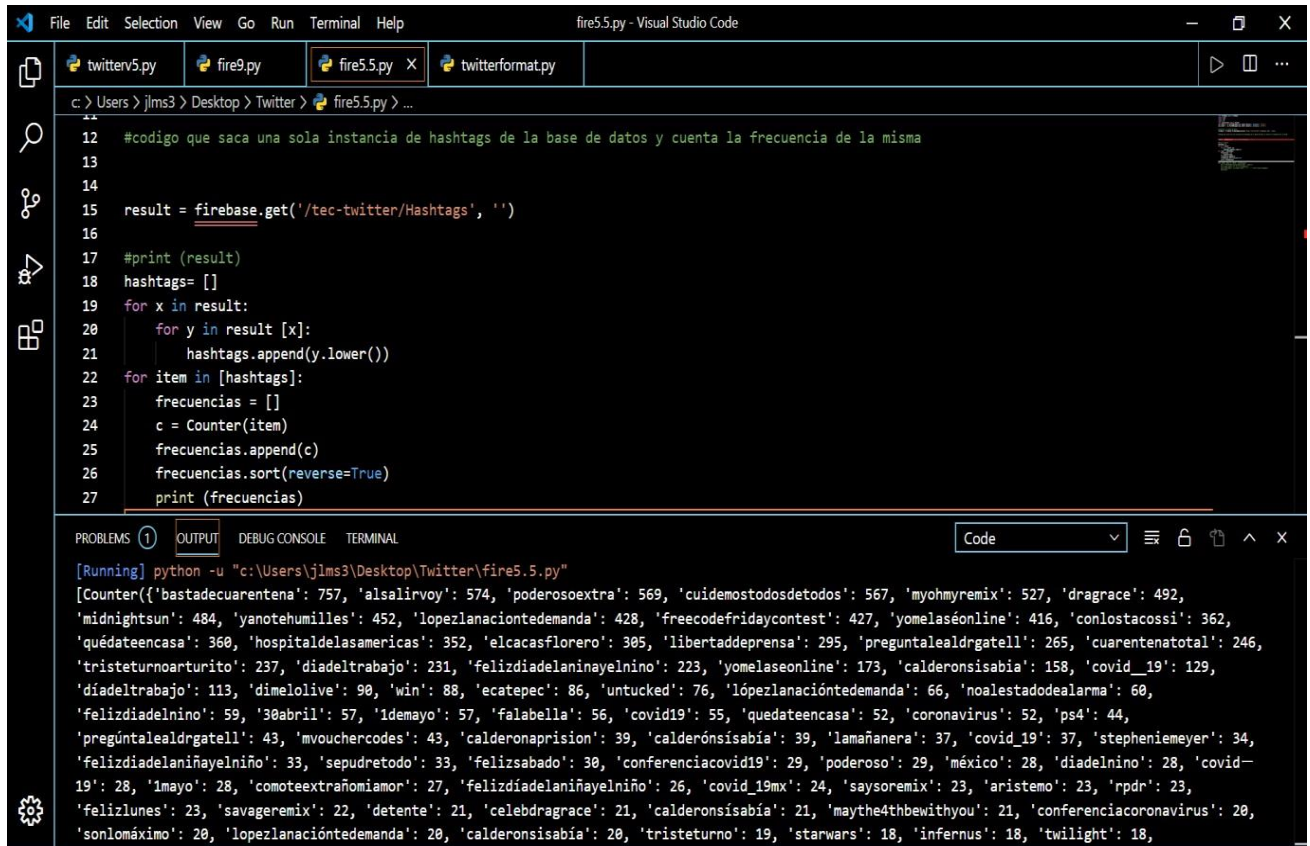
Since the words cannot be visualized very well, they will be explained in a general way. The one with the highest frequency, being more than 200, was #tristeturnoarturito, which was a trend since May 4 for Star Wars Day. On the other hand, the other words with many repetitions are basically articles and prepositions of Spanish, such as “en”, “la”, “de”. Although you can also see the word "twitter" and the "rt" for obvious reasons, and even several words form a small fragment of the same tweet, such as ("my", "contribution", "for", "the") or (“skin”, “Chinita”, “when”, “starts”, “5th”, “episode”) or (“the”, “last”, “dance”), the last example being a relationship with the series of the NBA Bulls on Netflix.

Information collected

Finally, using another python code, we generated a counter of the base hashtags and displayed the frequencies in descending order, so that we could see those that were repeated the most. And we see that

PROYECTO FINAL

many of them are related to quarantine, COVID 19 and even the government. And at the bottom of the image, it is clearly observed that Star Wars and sadturno hashtags were also a trend, that is why many related words could be seen in the highest frequencies of the histogram.



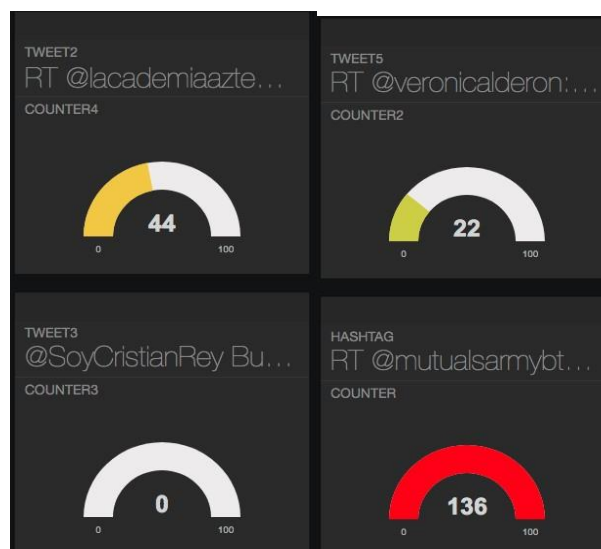
```
12 #codigo que saca una sola instancia de hashtags de la base de datos y cuenta la frecuencia de la misma
13
14
15 result = firebase.get('/tec-twitter/Hashtags', '')
16
17 #print (result)
18 hashtags= []
19 for x in result:
20     for y in result [x]:
21         hashtags.append(y.lower())
22 for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

PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL

[Running] python -u "c:\Users\jims3\Desktop\Twitter\fire5.5.py"

[Counter({'bastadecuarentena': 757, 'alsalirvoy': 574, 'poderosoextra': 569, 'cuidemostodosdetodos': 567, 'myohmyremix': 527, 'dragrace': 492, 'midnightsun': 484, 'yanotehumilles': 452, 'lopezlanacióndemanda': 428, 'freecodefridaycontest': 427, 'yomelaseonline': 416, 'conlostacossi': 362, 'quédateencasa': 360, 'hospitaldelasamericas': 352, 'elcacasflorero': 305, 'libertaddeprensa': 295, 'preguntalealdratell': 265, 'cuarentenatotal': 246, 'tristetornoarturito': 237, 'diadeltrabajo': 231, 'felizdiadelaniñayelnino': 223, 'yomelaseonline': 173, 'calderonsisabia': 158, 'covid_19': 129, 'diadeltrabajo': 113, 'dimelolive': 90, 'win': 88, 'ecatepec': 86, 'untucked': 76, 'lopezlanacióndemanda': 66, 'noalestadodealarma': 60, 'felizdiadelnino': 59, '30abril': 57, 'ldemayo': 57, 'falabella': 56, 'covid19': 55, 'quedateencasa': 52, 'coronavirus': 52, 'ps4': 44, 'preguntalealdratell': 43, 'mvouchercodes': 43, 'calderonaprision': 39, 'calderónsisabia': 39, 'lamananera': 37, 'covid_19': 37, 'stepheniemeyer': 34, 'felizdiadelaniñayelnino': 33, 'sepudretodo': 33, 'felizsabado': 30, 'conferenciacovid19': 29, 'poderoso': 29, 'mexico': 28, 'diadelnino': 28, 'covid-19': 28, '1mayo': 28, 'comoteextrañomiamor': 27, 'felizdiadelaniñayelnino': 26, 'covid_19mx': 24, 'saysoremix': 23, 'aristemo': 23, 'rpdr': 23, 'felizlunes': 23, 'savageremix': 22, 'detente': 21, 'celebrdragrace': 21, 'calderonsisabia': 21, 'maythe4thbewithyou': 21, 'conferenciacoronavirus': 20, 'sonlomaximo': 20, 'lopezlanacióndemanda': 20, 'calderonsisabia': 20, 'tristetorno': 19, 'starwars': 18, 'infennus': 18, 'twilight': 18,

After this, through freeboard.io we made a dashboard with the base information. It is worth mentioning that we had to export a json file from the firebase so that the teacher could put it on his page, from which we would take the data. So, being connected to the json, we were able to make text-type panels with some hashtags or tweets, and then put another gauge-type panel with the counter, to be able to see their presence and find out their frequency. Here are some of them:



PROYECTO FINAL

Conclusions

We can say that we liked this project a lot, since we were able to learn more about the analytics area and find more Python functions. Also, it was quite interesting to be able to use codes that made use of information from a social network, since you can really get a lot of information. But more importantly, we learned to analyze data and extract important things or in any case, find something specific.