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

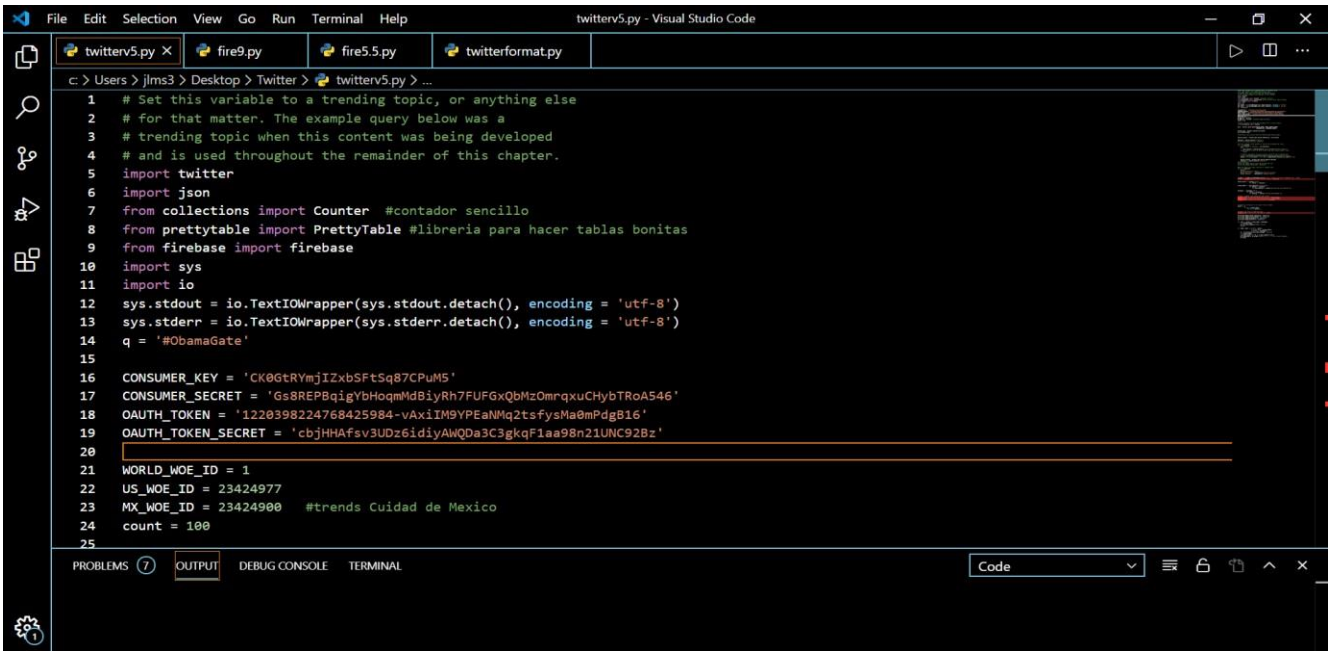
Technological Innovation

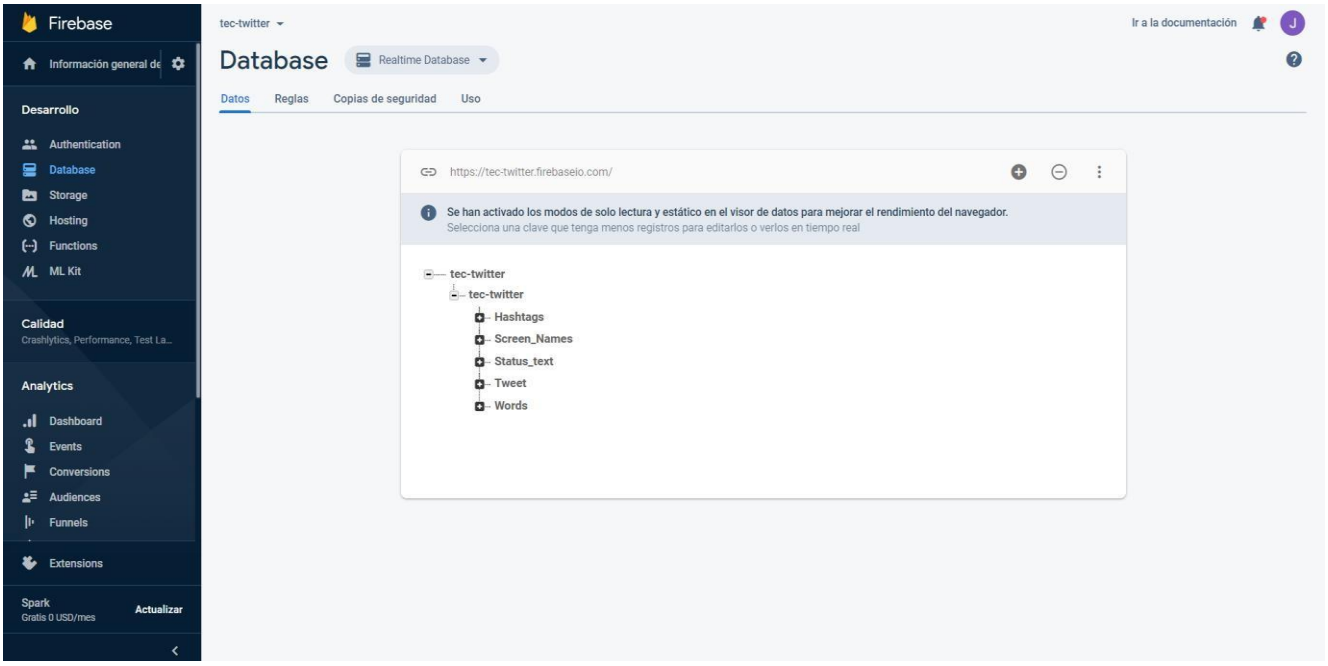
Science Department

PROYECTO FINAL

¿What did we do?

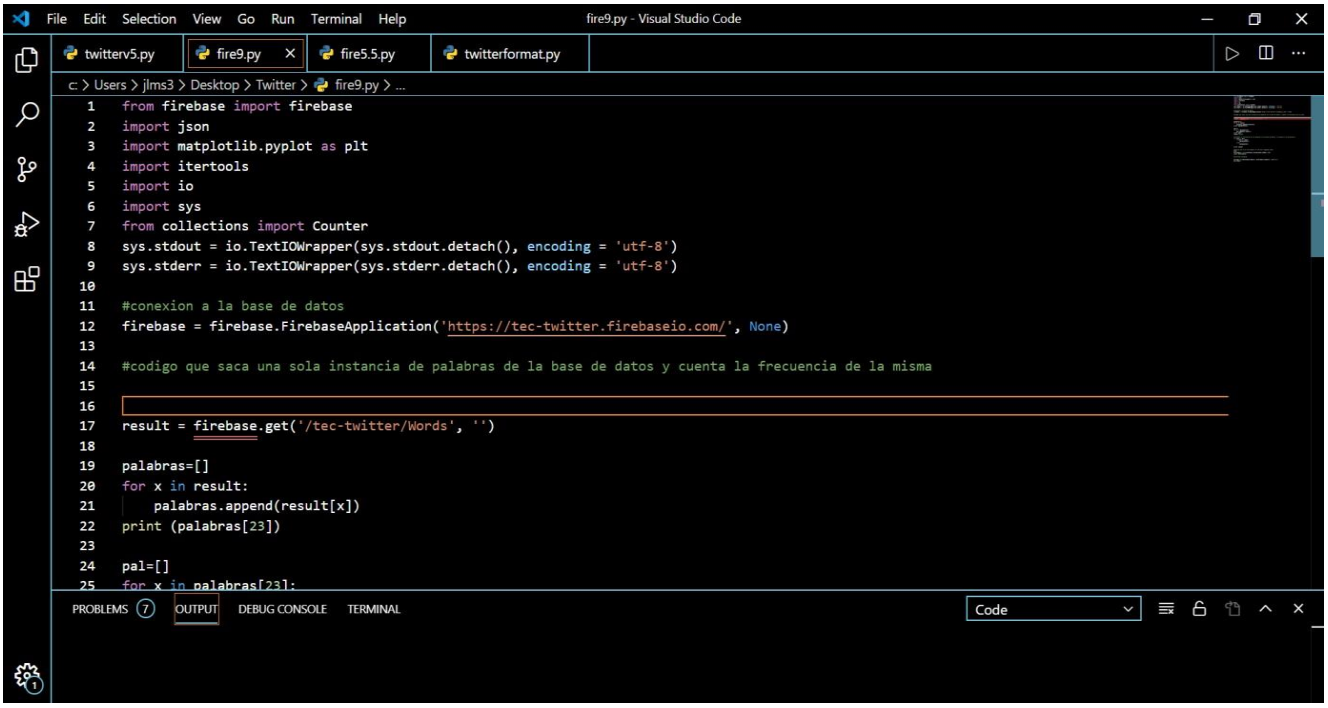
First of all, we created a database with firebase in which we could supply information. Then, by means of a Python code and Twitter API, we placed various trending topics in the database for several days. Basically, all the information of the hashtags was extracted, such as the words, the screen name,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



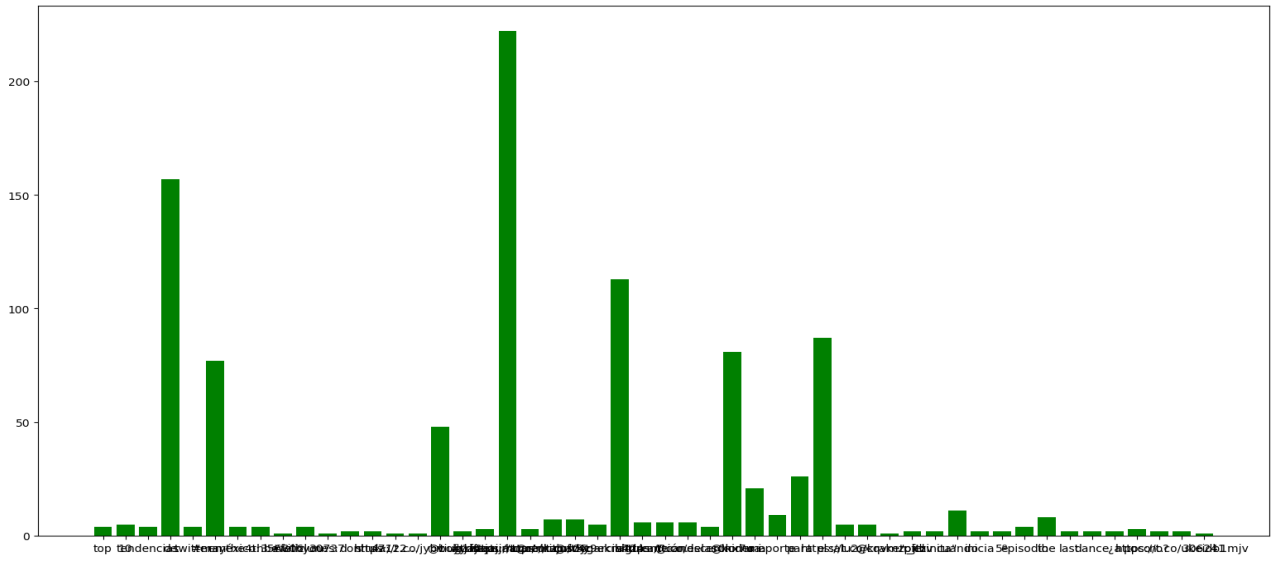


Results

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:




```
from firebase import firebase
import json
import matplotlib.pyplot as plt
import itertools
import io
import sys
from collections import Counter
sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')
sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')

#conexion a la base de datos
firebase = firebase.FirebaseApplication('https://tec-twitter.firebaseio.com/', None)

#codigo que saca una sola instancia de palabras de la base de datos y cuenta la frecuencia de la misma

result = firebase.get('/tec-twitter/Words', '')

palabras=[]
for x in result:
    palabras.append(result[x])
print (palabras[23])

pal=[]
for x in palabras[23]:
```

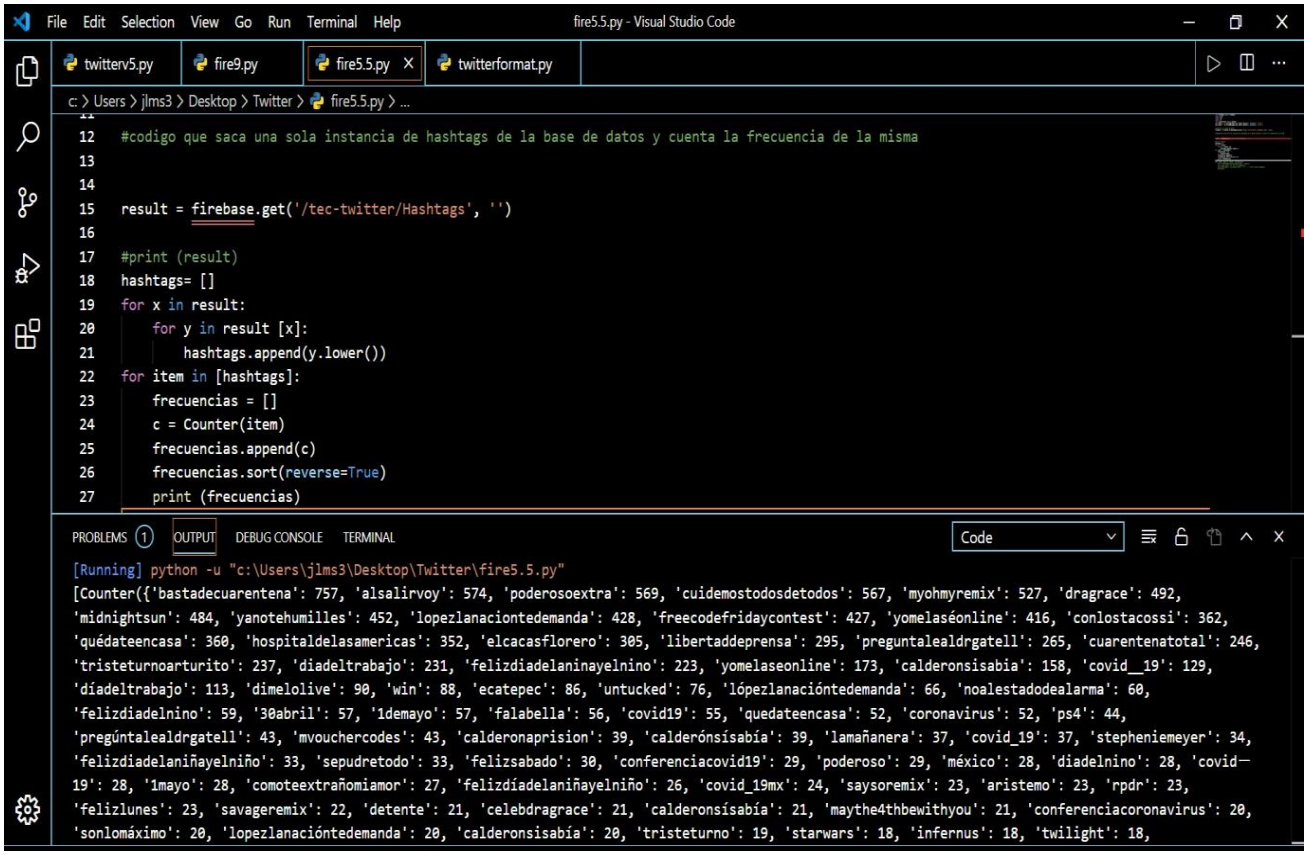


Since the words cannot be visualized very well, they will be explained in a general way. The one with the highest frequency, being more than 200, was #tristeturnoarturito, which was a trend since May 4 for Star Wars Day. On the other hand, the other words with many repetitions are basically articles 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.

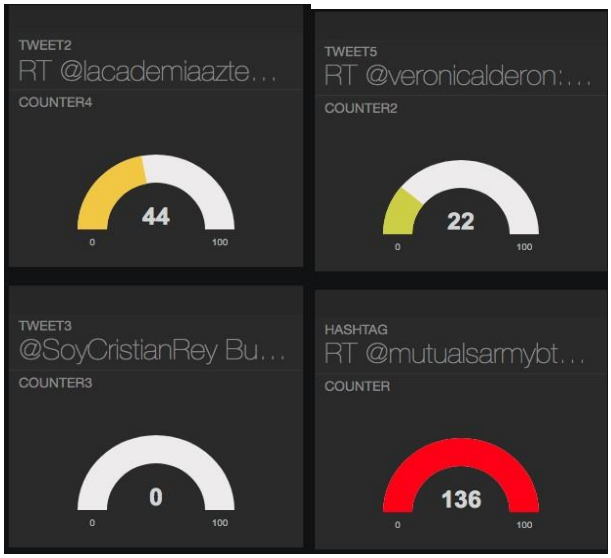
Information collected

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

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.



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:



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.