Visualización de datos geolocalizados con Plotly

Mikel Ardanaz Santamaría, Maria Bellver Carrasco,
Vicente Alberto Cifre Tomas, Alejandro Sánz Sanchez

8 de abril de 2019

En esta tarea vamos a realizar la toma de datos geolocalizados de la plataforma de Twitter, haciendo uso de la API(Application Programming Interface) integrada por Twitter, con el fín de realizar un Dashboard en el cual representemos las localizacioes de los tweets recopilados sobre un mapa mundial, así como una gráfica que muestra el número total de Tweets por horas.

1. Consideraciones Previas

Para la realización de este trabajo es necesario el registro por un lado en la API de Twitter como desarrollador para obtener las claves de acceso con las cuales son imprescindibles para poder realizar la obtención de los datos y, por otro lado, debido a que realizaremos las visualizaciones empleando la librería que proporciona la plataforma de Plotly, será necesario estar registrado en la propia página para hacer uso de dichos servicios de visualización. Mapbox es el proveedor de mapas utilizado por lo que también es necesario estar registrado en dicha plataforma.

2. Metodología de toma de datos

Para la realización de la captura de datos se ha empleado Python y se han tomado a escala mundial mediante Streaming únicamente aquellos Tweets que tenian activada la opción de la geolocalización durante los días 29 y 30 de Marzo, siendo respectivamente Viernes y Sábado, a lo largo de todo el día. Por tanto, la recopilación de datos fue contínua durante un total de aproximadamente 48 horas. Cabe destacar que el Viernes se produjo un error debido a la pérdida de conexión a internet durante unos 40 minutos.

Para realizar la captura y extracción hemos creado 2 scripts:

- ***Captura_Guardado:*** En este script se realiza la captura de los datos en Streamming y el guardado en una base de datos. Para ello, filtramos de forma que sólo se realice la captura de aquellos Tweets que contengan información sobre la geolocalización desde la cual se ha realizado el post, recopilando los siguentes datos sobre el post:

    - $Created\_at$: La fecha con día y hora a la cual ha sido creado el Tweet. Emplea como hora UTC en formato $\%d - \%m - \%Y\ \%H : \%M$.

- $User\_name$: El nombre del usuario que ha realizado el post de Twitter.
- $Text$: El texto escrito por el usuario.
- Lat : La latitud desde la cual se ha escrito el Tweet.
- Lon : La longitud desde la cual se ha escrito el Tweet.
- Rts : El número de retweets recibido para el Tweet concreto. En caso de Streaming se toma como valor defecto cero, ya que se captura en el momento que se publica.
- $Favs$: El número de veces que se ha recibido favorito, para el Tweet en concreto. En caso de Streaming se toma como valor defecto cero, ya que se captura en el momento que se publica.

Los datos se almacenan en una base de datos (MongoDB) e indicamos el orden en el cual queremos que se almacenen los campos extraídos de cada uno de los tweets.

- ***MongoExtract:*** Este script extrae de MongoDb los datos y los guarda en formato CSV.

3. Tratamiento de datos

Una vez se tienen ya todos los archivos de datos en CSV, se han unido los datos en un único archivo de texto. Tras esto se ha separado la columna de $Created\_at$ en 2 columnas, fecha y hora, para finalmente quedarnos con la última variable que varía entre la primera y la cuatragésima octava hora de captura, ya que el objetivo es representar los datos geolocalizados por horas. También, se eliminan las variables que no son útiles para la representación.

Tanto esta sección como la siguiente se puede encontrar en el notebook de Jupyter *Dashboard_Tweets*.

4. Dashboard

Como se ha comentado anteriormente, la librería plotly se ha utilizado para la implementación del dashboard. Éste consta de 3 elementos base:

- **Mapamundi animado con Tweets Geolocalizados:** Se trata de un atlas donde se ha representado con puntos la localizacion desde la cual se ha publicado cada tweet recopilado. El mapa está animado y muestra los Tweets que se han publicado cada hora.
- **Gráfico de puntos:** Representa la evolución de la cantidad total de tweets publicados por hora. Se destaca un punto (azul) que evoluciona en concordancia con el mapa.
- **Slider de hora:** Con este objeto podemos seleccionar a voluntad la hora a la cual queremos representar los tweets capturados.

Además, se dispone de dos botones, play y pause, con los que se controla el gif.

Para poder representar el mapa se han creado los distintos frames para cada hora de captura, es decir, un total de 48 frames. Se asocian a cada uno de los frames la latitud y longitud para los tweets capturados en la hora correspondiente. Éstos también contienen el conteo total de tweets para la representación del punto destacado dentro del gráfico de puntos ya que avanza acorde al mapa.

5. Conclusiones

Para poder interpretar correctamente los gráficos mostrados en el Dashboard, es necesario conocer que los datos se han tomado con hora local de España, por lo que es lógico deducir que las publicaciones varían según la zona horaria de cada región. Esto podemos verlo en el gráfico interactivo tridimensional, *Coordenadas_Horas*, en el cual se representa la longitud-latitud-hora.

Si observamos la gráfica con la disposición en el eje X (longitud) y en el eje Y (latitud), se puede apreciar que los puntos replican la forma de la distribución continental de la tierra. En cambio, si observamos la gráfica con la disposición en el eje X (longitud) y en el eje Y (horas), podemos observar 2 diagonales a las 12 h y a las 36h aproximadamente, que muestra las horas en las cuales para las longitudes representadas, no se capturan tweets debido a que es de noche en dicha ubicación.

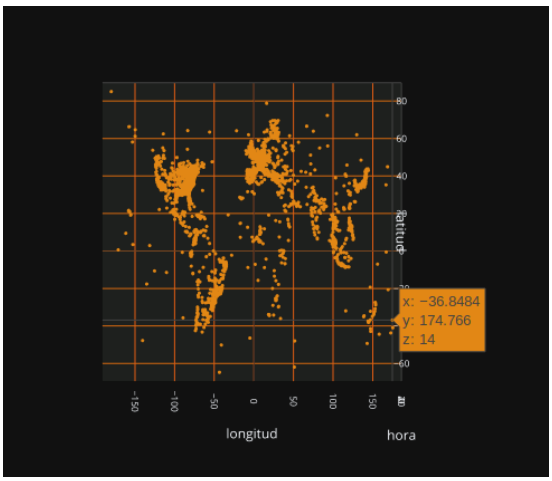

Figura 1: Longitud vs Latitud

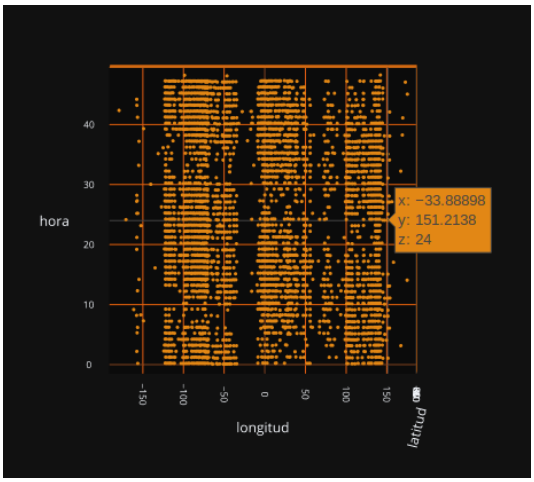

Figura 2: Longitud vs horas

Por tanto, en la región Europea en las 6 primeras horas del día 29 y 30 se observan muy poca cantidad de puntos que se incrementa en las horas restantes del día.

En cuanto a la región Asiática, concretamente China, debido al bloqueo impuesto por el gobierno Chino, se observa, en general, muy poca cantidad de publicaciones. Así mismo, se puede observar un resultado parecido en Rusia debido a la poca población en gran parte del país. En cambio, en Japón se observa una cantidad significativamente alta de puntos a lo largo de ambos días sin importar la hora.

Por otro lado, los países con más actividad en America son Estados unidos y Brasil durante las horas diurnas de cada región. Independientemente de la hora, en África y

Oceanía se aprecia una cantidad de puntos muy pequeña.

Finalmente, cabe destacar que estamos tratando sólo con tweets con geolocalización que se trata de una parte muy pequeña de la actividad real de la red social.

6. Archivos

- Código:
  - Captura de tweets y base de datos: *Captura_Guardado.py* y *MongoExtract.py*.
  - Notebook Jupyter: *Dashboard_Tweets.ipynb*
- Dashboard: *Dashboard_Tweets.html*.
- Gráfico 3D: *Coordenadas_Horas.html*.