Maestría Oficial en Big Data y Data Science

MongoDB+Twitter

01MBID Fundamentos de la tecnología Big Data

Alumno: Parrón Verdasco, Mario

Fecha de entrega: 15/05/2021

Índice

1.Objetivos

Crear una pequeña base de datos documental usando MongoDB.

Esta base de datos será rellenada con documentos extraídos de Twitter (Tweets), una vez esta esté rellena se realizarán consultas sobre estos datos y un visualización de los datos usando MongoDB Charts.

2.Introducción

Las cuentas de Twitter seleccionadas han sido las de los grandes Clubs de Futbol Europeo, de esta forma podemos ver qué clubs tienen un mayor impacto social.

Para esto el primer paso será seleccionar las cuentas, e ingestar sus últimos 3.200 Tweets publicados.

3.Metodología

El primer paso será preparar el soporte par estos datos, en nuestro caso se ha creado una base de datos en línea de MongoDB, para realizar las query necesarias sobre estos datos utilizaremos MongoDB Compass.

En referencia a Tweeter hemos creado un usuario Developer para poder realizar consultas en Twitter, así como incluir los tokens necesarios en el script de python proporcionado.

Una vez tenemos todo preparado seleccionamos las cuentas las cuentas, e ingestar sus últimos 3.200 Tweets publicados.

Así mismo realizaremos query de consulta para tener un mayor contexto de los datos.

Y una visualización de la información mediante Atlas.

4.Actividad

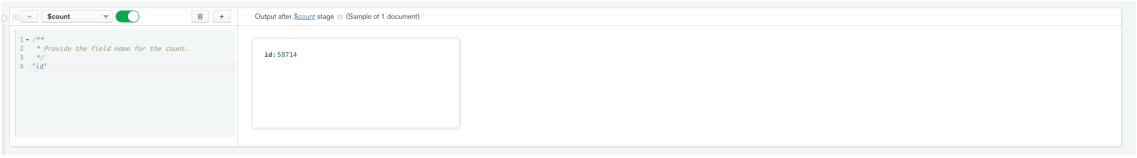
Modificación script de Python

```
# Additional information (timestamp and user)
loaded_entry['timestamp'] = datetime.now()
loaded_entry['user_alum'] = 'Mario Parrón Verdasco'
```

Consultas MongoDB

Número total de Tweets

[{$count: 'id'}]



Total de Tweets de cada cuenta

[{$group: {

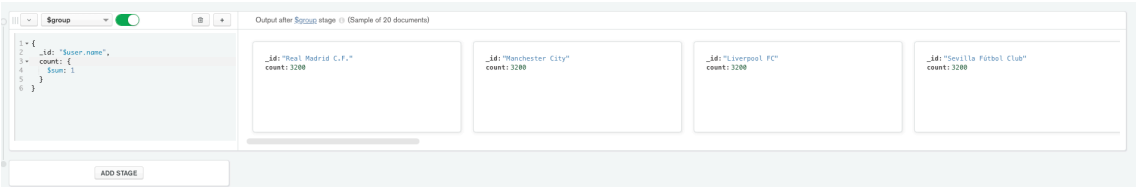
_id: "$user.name",

fieldN: {

$sum: 1

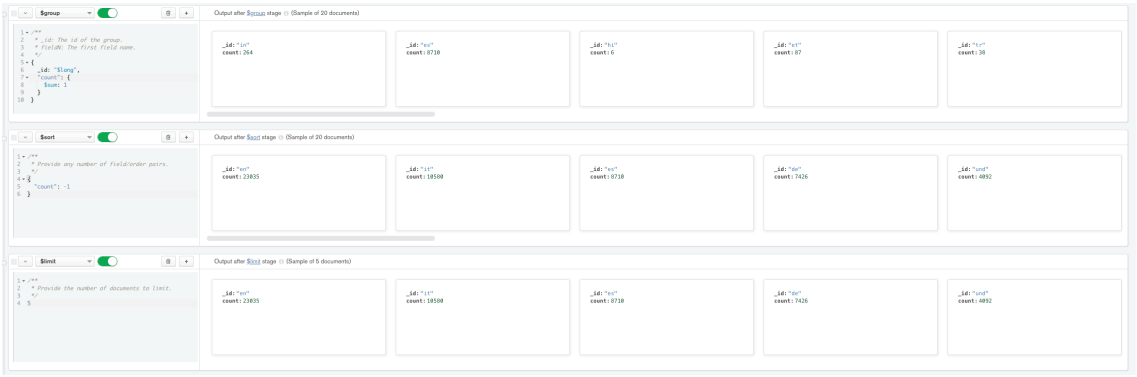
}

}}]



El ranking de los cinco primeros idiomas en los que se han escrito más Tweets.

```
[{$group: {
  _id: "$lang",
  "count": {
    $sum: 1
  }
}}, {$sort: {
  "count": -1
}}, {$limit: 5}]
```



Contar Tweets por tipo de media que lleva incrustado (e.g., foto, video, gif, etc.).

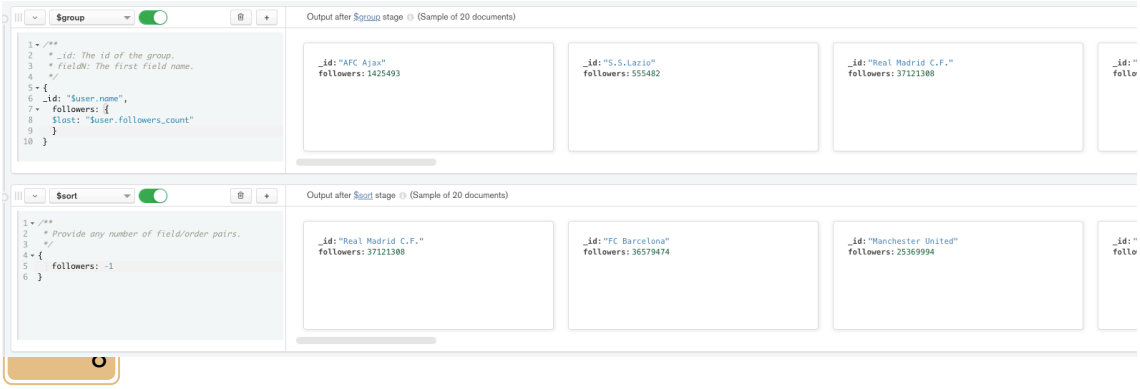
```
[{$unwind: {
  path: "$extended_entities.media",
  includeArrayIndex: 'string',
  preserveNullAndEmptyArrays: false
}}, {$group: {
  _id: "$extended_entities.media.type",
  count: {
```

```
    $sum: 1
  }
}}, {$sort: {
  count: -1
}}]
```

Ordenar las cuentas de mayor influencia a menor (de acuerdo al número de seguidores que posee cada una).

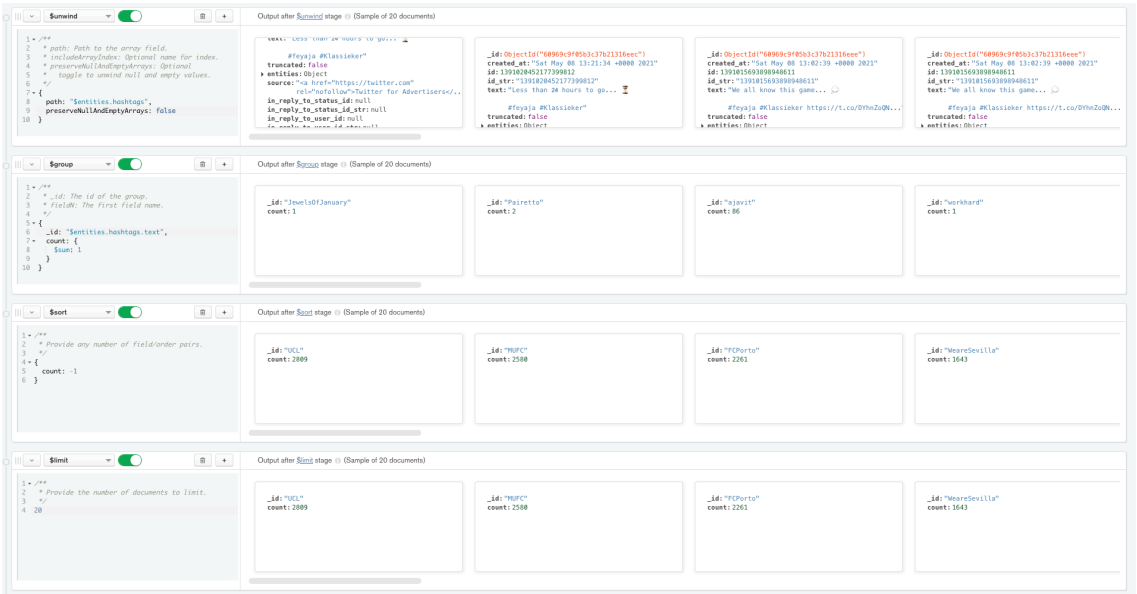
Para esta query se selecciona el último tweet de cada cuenta ($last)

```
[{$group: {
_id: "$user.name",
  followers: {
  $last: "$user.followers_count"
  }
}}, {$sort: {
   followers: -1
}}]
```



Realizar una consulta en Mongo DB para listar los 20 hashtags más utilizados.

```
[{$unwind: {
  path: "$entities.hashtags",
  preserveNullAndEmptyArrays: false
}}, {$group: {
  _id: "$entities.hashtags.text",
  count: {
    $sum: 1
  }
}}, {$sort: {
  count: -1
}}, {$limit: 20}]
```



MongoDB Charts

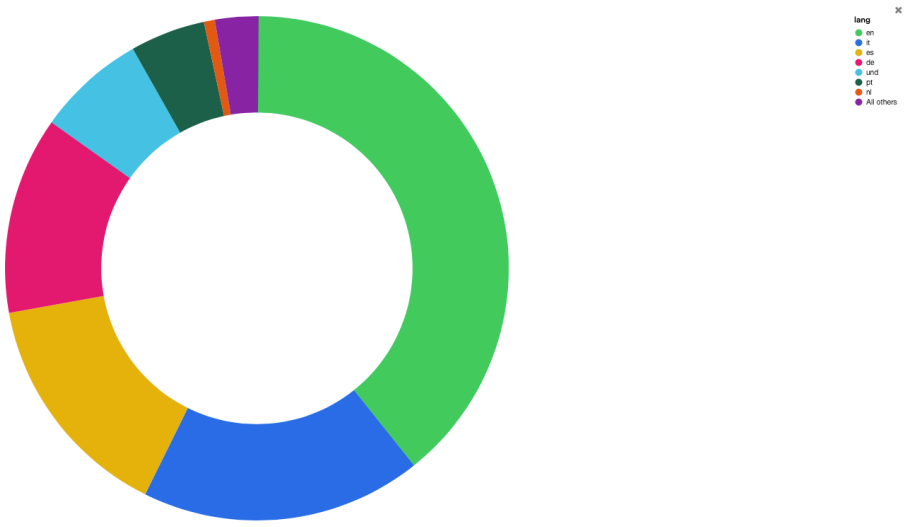
https://charts.mongodb.com/charts-actividad-tawdu/public/dashboards/609687c6-0c22-4086-8ead-5aaeafce61ae

---

Número total de Tweets

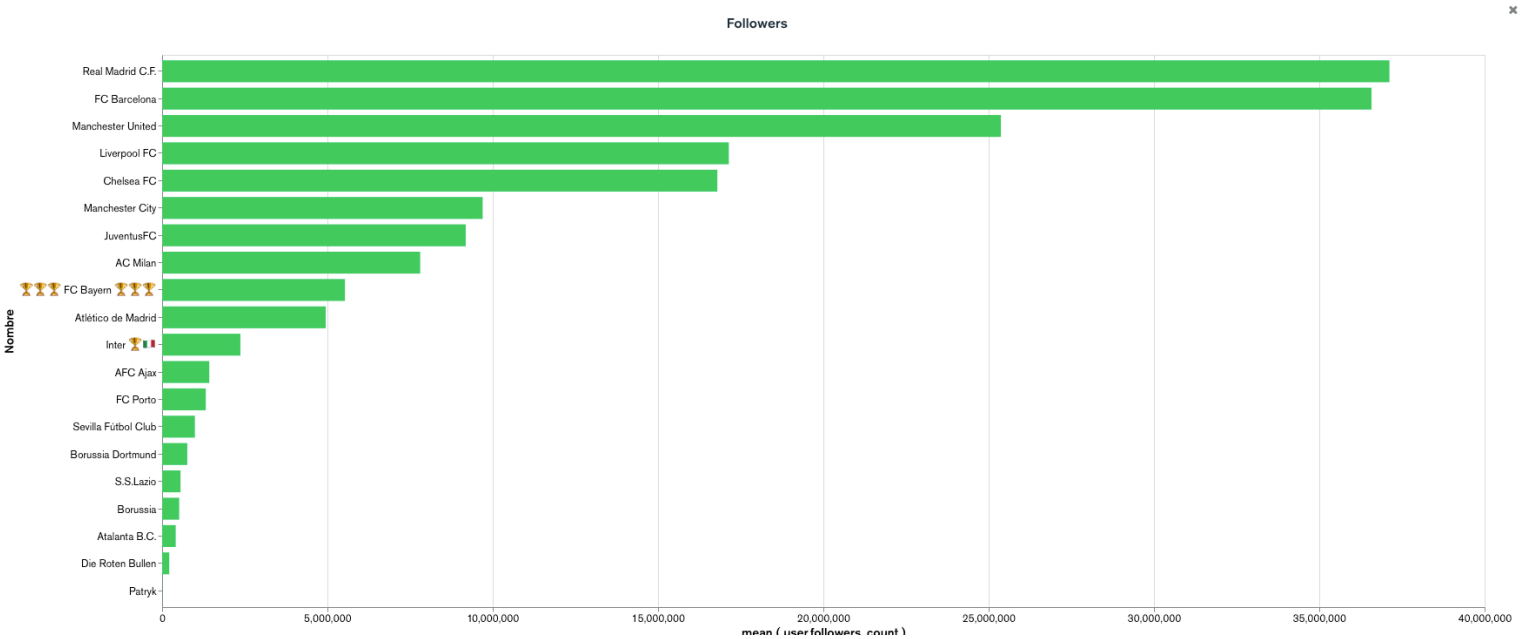
58,714

Distribución de lenguajes por tweet



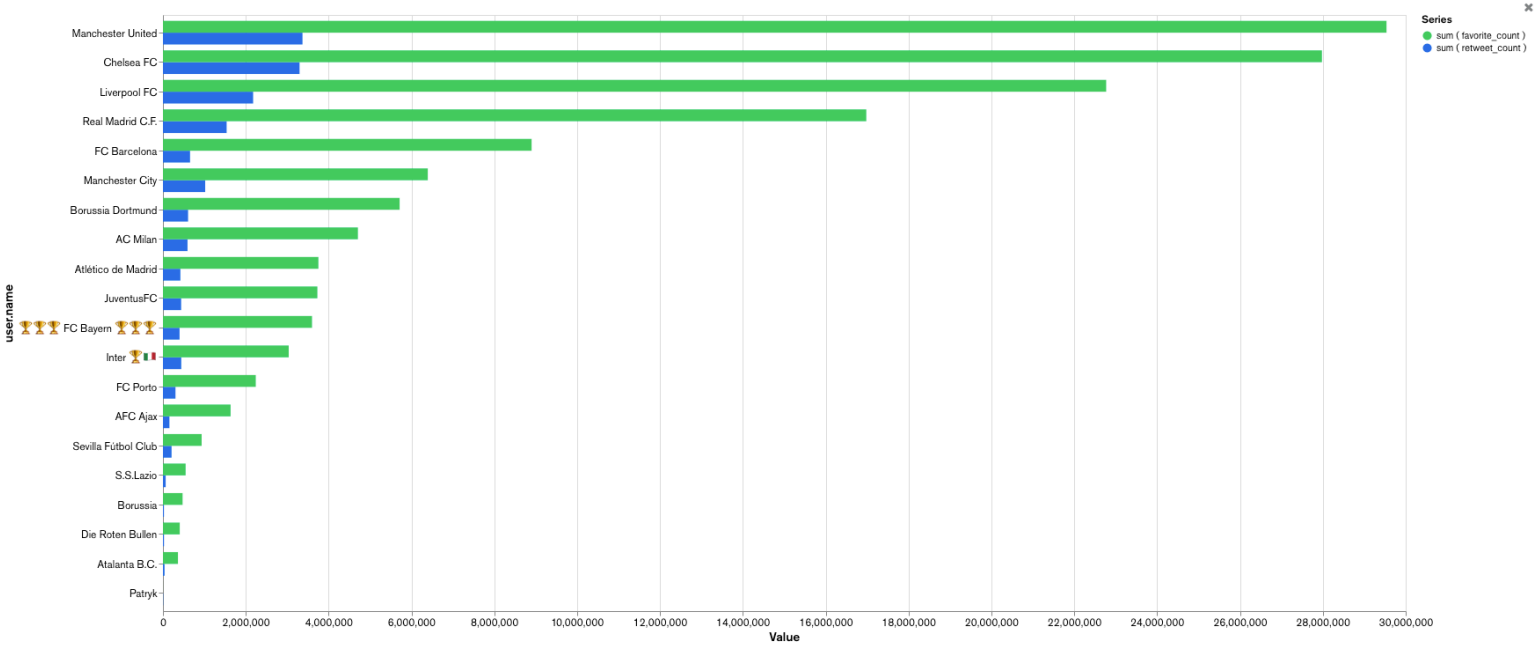
Se puede apreciar el lenguaje predominante es el inglés(verde), seguido del italiano(azul) y del español (Amarillo). En un principio esto parece debido a que los equipos Españoles también tienen una tendencia muy alta a escribir tweets en inglés.

Ranking equipos por número de seguidores



En esta gráfica podemos apreciar cómo el Real Madrid y Barcelona son los equipos con más followers, seguidos de los equipos ingleses.

Ranking equipos por número de favoritos y de retweets



En este ejemplo podemos ver cómo el Chelsea y Liverpool son más activos en redes sociales.

Hashtags retweeted

Hashtags por número de favoritos

Relación entre Favs y Retweet por cada cuenta

Se puede apreciar una relación entre número de retweets y número de favs.