

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	3
2.Introducción.....	4
3.Metodología	5
4.Actividad.....	6

1.Objetivos

Crear una pequeña base de datos documental usando MongoDB.

Esta base de datos será rellena 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

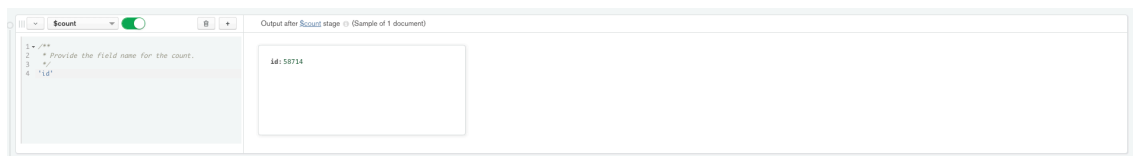
Modificación script de Python

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

Consultas MongoDB

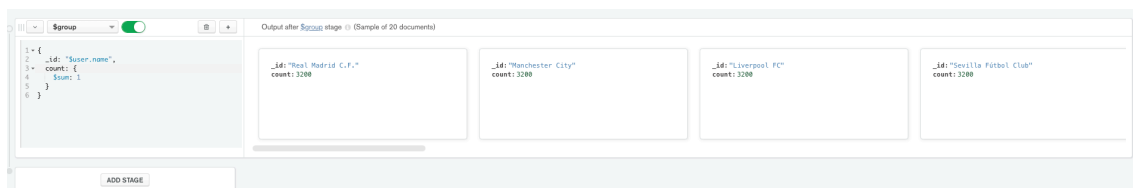
Número total de Tweets

```
[{$count: 'id'}]
```



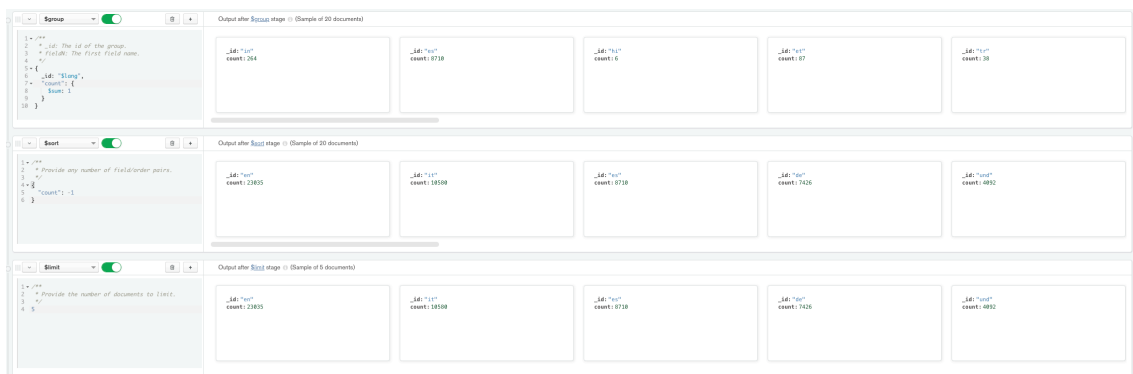
Total de Tweets de cada cuenta

```
[{$group: {
  _id: "$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

}}

The screenshot shows the MongoDB Compass interface with a query pipeline consisting of three stages: \$sum, \$group, and \$sort. The \$sum stage calculates the total count of documents. The \$group stage groups the results by user name. The \$sort stage sorts the results by the count of followers in descending order.

```
1 //**
2 // * Path to the array field.
3 // * Includes/Excludes: Optional name for index.
4 // * preserveNullAndEmptyArrays: Optional.
5 // * toggle to omit null and empty values.
6 //
7 {
8   path: "$extended.entities.media",
9   includeEmptyArrays: "string",
10  preserveNullAndEmptyArrays: false
11 }
12 }
```

Output after \$sum stage (Sample of 20 documents)

_id	count
"#Franja #Klassiker https://t.co/9p88y0M"	1126
"#Franja #Klassiker https://t.co/9p88y0M"	785
"#Franja #Klassiker https://t.co/9p88y0M"	2038

Output after \$group stage (Sample of 3 documents)

_id	count
"#Franja #Klassiker https://t.co/9p88y0M"	1126
"#Franja #Klassiker https://t.co/9p88y0M"	785
"#Franja #Klassiker https://t.co/9p88y0M"	2038

Output after \$sort stage (Sample of 3 documents)

_id	count
"#Franja #Klassiker https://t.co/9p88y0M"	1126
"#Franja #Klassiker https://t.co/9p88y0M"	785
"#Franja #Klassiker https://t.co/9p88y0M"	2038

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

}}

The screenshot shows the MongoDB Compass interface with a query pipeline consisting of two stages: \$group and \$sort. The \$group stage groups the results by user name. The \$sort stage sorts the results by the count of followers in descending order.

```
1 //**
2 // * id: The id of the group.
3 // * field: The first field name.
4 //
5 {
6   _id: "$user.name",
7   followers: {
8     $last: "$user.followers_count"
9   }
10 }
```

Output after \$group stage (Sample of 20 documents)

_id	followers
"AFC Ajax"	1425493
"S.S.Lazio"	555482
"Real Madrid C.F."	3712138

Output after \$sort stage (Sample of 20 documents)

_id	followers
"Real Madrid C.F."	3712138
"FC Barcelona"	36579474
"Manchester United"	25369994

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}]

```

The screenshot displays the MongoDB Compass interface with a query pipeline executed in four stages. The first stage, \$unwind, shows individual hashtag documents. The second stage, \$group, shows the aggregation of counts for each hashtag. The third stage, \$sort, shows the results sorted by count in descending order. The fourth stage, \$limit, shows the top 20 results.

Stage	Field	Value
\$unwind	_id	#feyja #klassiker
	created_at	"Sat May 08 13:21:34 +0000 2021"
	id	139181563090940611
	text	"We all know this game..."
\$group	_id	#feyja #klassiker
	count	1
	_id	#feyja #klassiker
	count	1
\$sort	_id	#feyja #klassiker
	count	1
	_id	#feyja #klassiker
	count	1
\$limit	_id	#feyja #klassiker
	count	1
	_id	#feyja #klassiker
	count	1

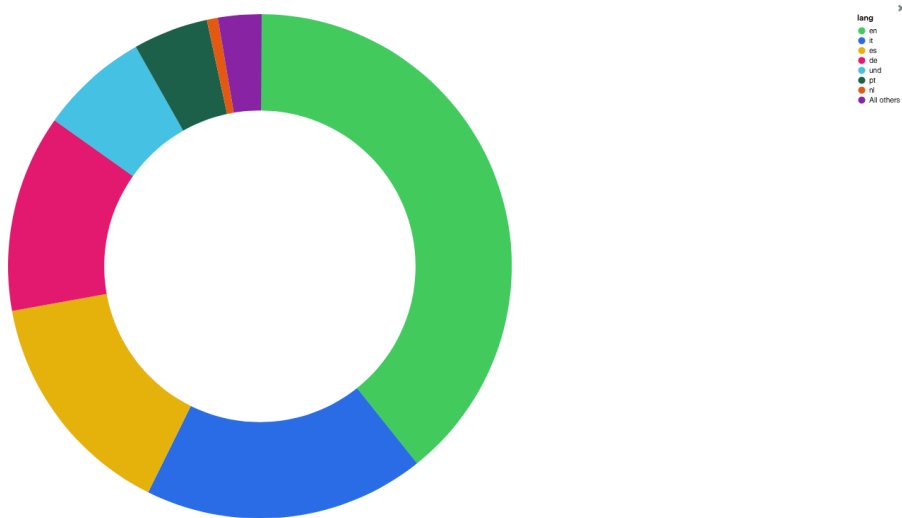
MongoDB Charts

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

Número total de Tweets

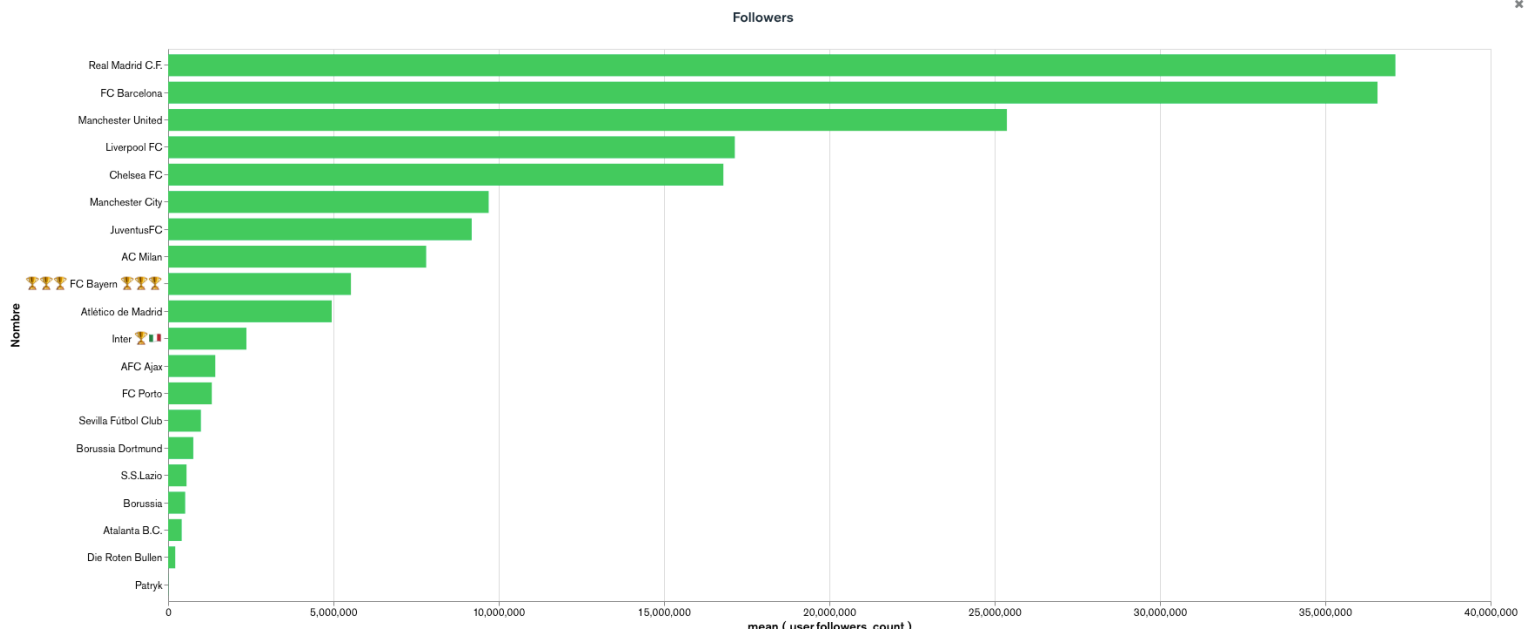
58,714

Distribución de lenguajes por tweet



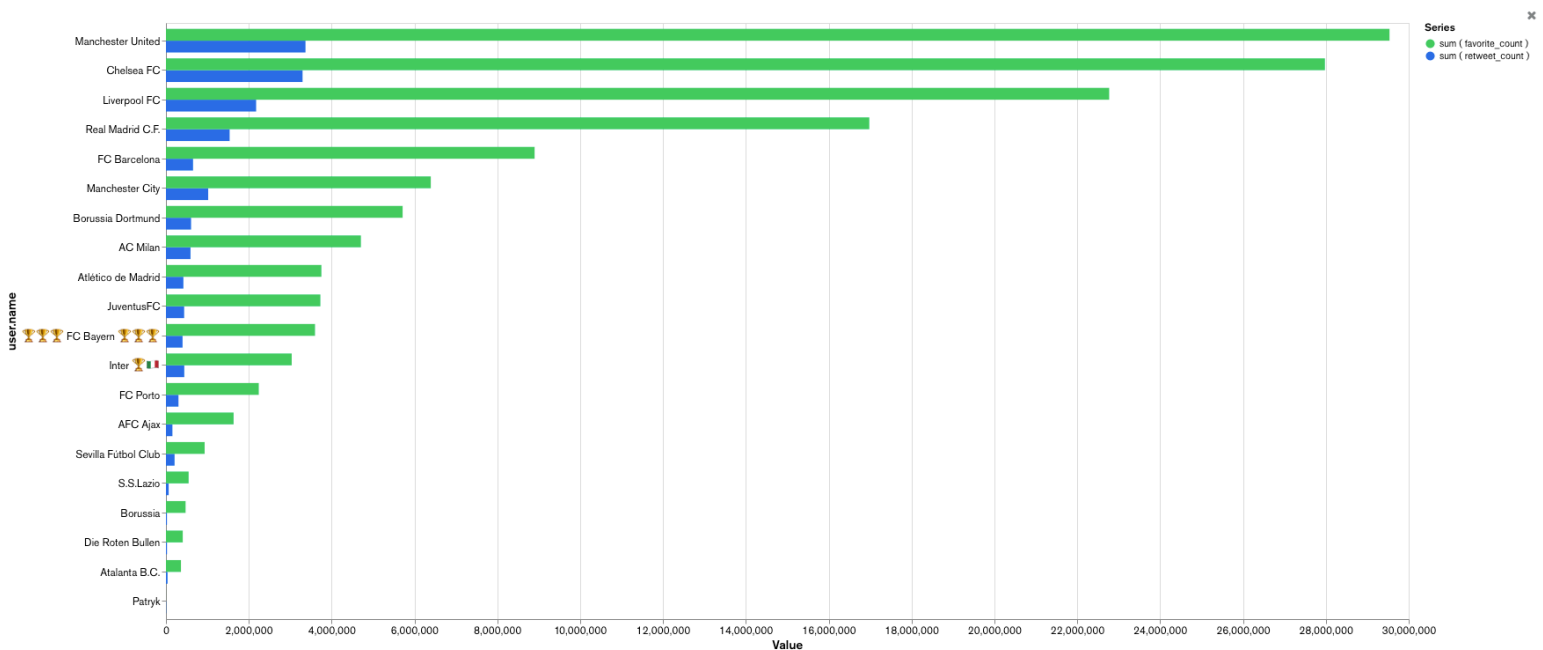
Se puede apreciar el lenguaje predominante es el inglés(verde), seguido del italiano(azul) y del español (Amarillo). En un principio esto parece debido a que los equipos Españoles también tienen una tendencia muy alta a escribir tweets en inglés.

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



Hashtags retweeted



Relación entre Favs y Retweet por cada cuenta

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

