

## Ejercicio - Bases de Datos No Relacionales con Mongo

- Conexión a la Plataforma seleccionada de bases de datos no relacionales
- Generar una nueva base de datos

```
In [ ]: from pymongo import MongoClient

# Requires the PyMongo package.
# https://api.mongodb.com/python/current

client = MongoClient('mongodb+srv://hzoscarist:irxgsjdfOoSHmygn@cluster1.xaxbyn6.mongodb.net/')
```

```
In [ ]: print(client.list_database_names())

['sample_airbnb', 'sample_analytics', 'sample_geospatial', 'sample_guides', 'sample_mflix', 'sample_restaurants', 'sample_supplies', 'sample_training', 'sample_weatherdata', 'tweets', 'admin', 'local']
```

```
In [ ]: db = client.tweets
col = db.tweets
```

```
In [ ]: ## col.count_documents({'airline_sentiment': 'negative'})
```

```
Out[ ]: 9178
```

- Explorar el archivo de tweets con palabras clave definidos para el caso (incluir 5)

1. Se agrupa por la percepción del sentimiento de la aerolínea

```
In [ ]: agg_result=col.aggregate([
    { "$group": {
        "_id": "$airline_sentiment",
        "count": { "$sum": 1 }
    }}
]);
```

```
In [ ]: for i in agg_result:  
    print(i)  
  
{'_id': 'neutral', 'count': 3099}  
{'_id': 'positive', 'count': 2363}  
{'_id': 'negative', 'count': 9178}
```

### 2. Se agrupa por el motivo por el cual se tuvo una percepción negativa de la aerolínea

```
In [ ]: agg_result=col.aggregate([  
    { "$group": {  
        "_id": "$negativereason",  
        "count": { "$sum": 1 }  
    }}  
]);
```

```
In [ ]: for i in agg_result:  
    print(i)  
  
{'_id': None, 'count': 5462}  
{'_id': 'Cancelled Flight', 'count': 847}  
{'_id': 'Bad Flight', 'count': 580}  
{'_id': 'Damaged Luggage', 'count': 74}  
{'_id': 'longlines', 'count': 178}  
{'_id': 'Flight Booking Problems', 'count': 529}  
{'_id': 'Lost Luggage', 'count': 724}  
{'_id': 'Late Flight', 'count': 1665}  
{'_id': 'Customer Service Issue', 'count': 2910}  
{'_id': "Can't Tell", 'count': 1190}  
{'_id': 'Flight Attendant Complaints', 'count': 481}
```

### 3. Se agrupa por las diferentes aerolíneas

```
In [ ]: agg_result=col.aggregate([  
    { "$group": {  
        "_id": "$airline",  
        "count": { "$sum": 1 }  
    }}  
]);
```

```
In [ ]: for i in agg_result:  
    print(i)  
  
{'_id': 'Southwest', 'count': 2420}  
{'_id': 'United', 'count': 3822}  
{'_id': 'Delta', 'count': 2222}  
{'_id': 'Virgin America', 'count': 504}  
{'_id': 'US Airways', 'count': 2913}  
{'_id': 'American', 'count': 2759}
```

#### 4. Se agrupa por las diferentes timezone de los usuarios

```
In [ ]: agg_result=col.aggregate([  
    { "$group": {  
        "_id": "$user_timezone",  
        "count": { "$sum": 1 }  
    }}  
]);
```

```
In [ ]: for i in agg_result:  
    print(i)
```

```
{'_id': 'Kyiv', 'count': 2}
{'_id': 'Bucharest', 'count': 1}
{'_id': 'Mazatlan', 'count': 3}
{'_id': 'Seoul', 'count': 5}
{'_id': 'Central America', 'count': 13}
{'_id': 'Bogota', 'count': 5}
{'_id': 'Quito', 'count': 738}
{'_id': 'Central Time (US & Canada)', 'count': 1931}
{'_id': 'America/New_York', 'count': 26}
{'_id': 'Tijuana', 'count': 9}
{'_id': 'Mexico City', 'count': 3}
{'_id': 'Alaska', 'count': 108}
{'_id': 'Wellington', 'count': 1}
{'_id': 'Paris', 'count': 25}
{'_id': 'Brasilia', 'count': 23}
{'_id': 'Stockholm', 'count': 4}
{'_id': 'Kuala Lumpur', 'count': 1}
{'_id': 'Jerusalem', 'count': 3}
{'_id': 'Singapore', 'count': 2}
{'_id': 'America/Los_Angeles', 'count': 15}
{'_id': 'Tokyo', 'count': 1}
{'_id': 'Istanbul', 'count': 1}
{'_id': 'Canberra', 'count': 1}
{'_id': 'America/Atikokan', 'count': 1}
{'_id': 'Brisbane', 'count': 10}
{'_id': 'Perth', 'count': 2}
{'_id': 'Lisbon', 'count': 1}
{'_id': 'Athens', 'count': 16}
{'_id': 'Midway Island', 'count': 1}
{'_id': 'West Central Africa', 'count': 1}
{'_id': 'London', 'count': 195}
{'_id': 'Bangkok', 'count': 4}
{'_id': 'Helsinki', 'count': 9}
{'_id': 'Edinburgh', 'count': 4}
{'_id': 'Amsterdam', 'count': 74}
{'_id': 'Dublin', 'count': 17}
{'_id': 'Madrid', 'count': 7}
{'_id': 'Pacific Time (US & Canada)', 'count': 1208}
{'_id': 'Copenhagen', 'count': 2}
{'_id': 'Indiana (East)', 'count': 26}
{'_id': 'Santiago', 'count': 17}
{'_id': 'Mountain Time (US & Canada)', 'count': 369}
```

```
{'_id': 'Bern', 'count': 1}
{'_id': 'Prague', 'count': 1}
{'_id': 'Arizona', 'count': 229}
{'_id': 'Hong Kong', 'count': 2}
{'_id': 'La Paz', 'count': 3}
{'_id': 'Tehran', 'count': 17}
{'_id': 'Guam', 'count': 2}
{'_id': 'Berlin', 'count': 9}
{'_id': 'Abu Dhabi', 'count': 23}
{'_id': 'Greenland', 'count': 17}
{'_id': 'Irkutsk', 'count': 1}
{'_id': 'Caracas', 'count': 9}
{'_id': 'Taipei', 'count': 6}
{'_id': 'America/Detroit', 'count': 1}
{'_id': 'Brussels', 'count': 9}
{'_id': 'Nairobi', 'count': 2}
{'_id': 'Eastern Time (US & Canada)', 'count': 3744}
{'_id': 'America/Chicago', 'count': 37}
{'_id': 'Mid-Atlantic', 'count': 15}
{'_id': 'Monterrey', 'count': 1}
{'_id': 'None', 'count': 4820}
{'_id': 'Saskatchewan', 'count': 1}
{'_id': 'Adelaide', 'count': 7}
{'_id': 'Islamabad', 'count': 2}
{'_id': 'Buenos Aires', 'count': 14}
{'_id': 'Rome', 'count': 5}
{'_id': 'Lima', 'count': 2}
{'_id': 'Newfoundland', 'count': 1}
{'_id': 'EST', 'count': 1}
{'_id': 'Melbourne', 'count': 8}
{'_id': 'Guadalajara', 'count': 3}
{'_id': 'Hawaii', 'count': 104}
{'_id': 'Sydney', 'count': 107}
{'_id': 'America/Boise', 'count': 3}
{'_id': 'Atlantic Time (Canada)', 'count': 497}
{'_id': 'Sarajevo', 'count': 1}
{'_id': 'Beijing', 'count': 11}
{'_id': 'Casablanca', 'count': 15}
{'_id': 'Vienna', 'count': 3}
{'_id': 'Warsaw', 'count': 1}
{'_id': 'Pretoria', 'count': 1}
{'_id': 'New Caledonia', 'count': 3}
```

```
{'_id': 'New Delhi', 'count': 15}  
{'_id': 'Solomon Is.', 'count': 1}
```

### 5. Se agrupa por la ubicación desde la cual se mando el tweet

```
In [ ]: agg_result=col.aggregate([  
    { "$group": {  
        "_id": "$tweet_location",  
        "count": { "$sum": 1 }  
    }}  
]);
```

```
In [ ]: ## for i in agg_result:  
##     print(i)
```

### ○ Insights

Después de haber hecho todas estas agrupaciones se puede tener una idea del contenido de este dataset: por aerolínea fue de 2827.

- Más del 60% de los tweets tuvieron una percepción negativa de la aerolínea en la cual viajaron.
- Aunque hay una gran cantidad de casos en los que no se justifica el motivo, en aquellos casos en los cuales se menciona se ve que la gran mayoría son por el servicio al cliente. (2910 casos)
- De las 6 aerolíneas consideradas solo hay una que tuvo una cantidad considerablemente inferior con respecto a las demás. ("Virgin America" 504) Excluyendo esta última el promedio de los vuelos
- La gran cantidad de vuelos se encuentran concentrados en América del Norte.

### ○ Importar por lo menos 1000 tweets a la base de datos a una nueva estructura

- Realicé el filtro por la percepción negativa de la aerolínea a través de los tweets obteniendo más de 9000 pots.
- Exporte dichos pots en un archivo csv que será adjunto en los soportes de esta actividad.

```
In [ ]: filter={'airline_sentiment':'negative'}

result = client['tweets']['tweets'].find(
    filter=filter
)
```

```
In [ ]: ## import pprint
## pp = pprint.PrettyPrinter(indent=4)
## for doc in result:
##     pp.pprint(doc)
```

```
In [ ]: import csv

csv_columns = ['tweet_id','airline_sentiment','airline_sentiment_confidence', 'negativereason', 'negativereason_confidence']

with open('negative_sentiment.csv', 'w', encoding='utf-8') as csvfile:
    writer = csv.DictWriter(csvfile, fieldnames=csv_columns, extrasaction='ignore')
    writer.writeheader()
    for post in col.find({'airline_sentiment': 'negative'}):
        writer.writerow(post)
```

○ Construir 2 filtros de búsqueda que se considere importante dentro de la base de datos de tweets proporcionada

- El primer filtro corresponde a los tweets que muestran una percepción negativa de una aerolínea por el no buen servicio al cliente de una aerolínea en particular.
- Basta cambiar el último filtro para determinar en qué difieren, a grandes rasgos, los contendios de los tweets entre una aerolínea y otra.

```
In [ ]: filter={"$and": [
    {"airline_sentiment":"negative"},
    {"negativereason": "Customer Service Issue"},
    {"airline":"United"}]}
```

```
result = client['tweets']['tweets'].find(filter=filter)
```

```
In [ ]: ## import pprint
## pp = pprint.PrettyPrinter(indent=4)
## for doc in result:
##     pp.pprint(doc)
```

```
In [ ]:
```

- La segunda query corresponde a los tweets que muestran una percepción negativa de la aerolínea filtrada por el timezone del usuario que hizo el tweet. Para este caso utilice la ciudad de Bogotá.
- Basta cambiar el último filtro para cambiar el país del cual se quieren obtener los tweets.

```
In [ ]: filter={"$and": [
    {"airline_sentiment": "negative"},
    {"user_timezone": "Bogota"}]}
```

  

```
result = client['tweets']['tweets'].find(filter)
```

```
In [ ]: import pprint
pp = pprint.PrettyPrinter(indent=4)
for doc in result:
    pp.pprint(doc)
```

```
{  '_id': ObjectId('64a1d33068a2d70aa5c13ef0'),
  'airline': 'United',
  'airline_sentiment': 'negative',
  'airline_sentiment_confidence': 0.6726,
  'name': 'mariachan90',
  'negativereson': 'Customer Service Issue',
  'negativereson_confidence': 0.3429,
  'retweet_count': 0,
  'text': '@united understanding the situation we waited and it was opened '
          'until 10:30pm',
  'tweet_created': '2015-02-19 20:37:19 -0800',
  'tweet_id': 568630453521022976,
  'user_timezone': 'Bogota'}
{
  '_id': ObjectId('64a1d33368a2d70aa5c1513c'),
  'airline': 'Delta',
  'airline_sentiment': 'negative',
  'airline_sentiment_confidence': 1.0,
  'name': 'CamiCorreaBal',
  'negativereson': 'Lost Luggage',
  'negativereson_confidence': 0.6383,
  'retweet_count': 0,
  'text': "@JetBlue yes, We have de baggage claim, I'm so sad for the "
          'baggage and how They treat Us 😞 please we need That baggage',
  'tweet_created': '2015-02-21 10:24:24 -0800',
  'tweet_id': 569200983173156864,
  'user_timezone': 'Bogota'}
{
  '_id': ObjectId('64a1d33368a2d70aa5c1513d'),
  'airline': 'Delta',
  'airline_sentiment': 'negative',
  'airline_sentiment_confidence': 1.0,
  'name': 'CamiCorreaBal',
  'negativereson': 'Lost Luggage',
  'negativereson_confidence': 1.0,
  'retweet_count': 0,
  'text': "@JetBlue my mom's baggage is lost, in fly 1557 fort "
          'lauderdale-Bogotá today, Colombian employees treated us badly , '
          'need help please',
  'tweet_created': '2015-02-21 10:19:12 -0800',
  'tweet_id': 569199676282564609,
  'user_timezone': 'Bogota'}
{
  '_id': ObjectId('64a1d33768a2d70aa5c169d3'),
  'airline': 'American',
```

```
'airline_sentiment': 'negative',
'airline_sentiment_confidence': 1.0,
'name': 'lifeletterj',
'negativereason': "Can't Tell",
'negativereason_confidence': 1.0,
'retweet_count': 0,
'text': '@AmericanAir Close down',
'tweet_created': '2015-02-22 17:01:41 -0800',
'tweet_id': 569663351124422656,
'tweet_location': 'Indiana',
'user_timezone': 'Bogota'}
```