09

Mãos na massa: Mesclando coleções

Chegou a hora de você executar o que foi visto na aula! Para isso, execute os passos listados abaixo.

1) Para restaurar a base de dados e a coleção do MondoDB, primeiramente disponibilize o seu servidor. Caso você ainda não tenha feito, por exemplo, faça:

```
start /b mongod --dbpath D:\MongoDB\data\db --logpath D:\MongoDB\log.txt --oplogSize 50 --smallfiles
```

2) Baixe aqui (https://s3.amazonaws.com/caelum-online-public/787-data-science-coleta-de-dados/05/ufos.bson) o arquivo **ufos.bson**, e restaure a coleção através deste arquivo, por exemplo:

```
mongorestore  --db ufos --collection ufo "D:\MongoDB\bkp_ufos\ufo\ufos.bson"
```

3) Restaurada a coleção, você pode fazer consultas à mesma, por exemplo, para ver quantos documentos há na coleção:

```
mongo ufos --eval "db.ufo.count()" --quiet
```

4) Agora, analise os dados, fazendo algumas consultas no MongoDB, através do **Robo 3T** (**Robomongo**). Primeiramente, vendo quantos documentos há por ano:

```
db.ufo.aggregate ( [
    { $group : {_id : "$Sight_Year", quantos : {$sum : 1} } },
    { $sort: {_id: 1} }
] )
```

5) Para visualizar a estrutura dos documentos, selecione o primeiro:

```
db.ufo.findOne()
```

6) Há um código responsável por mesclar as coleções **ufo.ufos** e **dbclima.clima**, você pode baixá-lo aqui (https://s3.amazonaws.com/caelum-online-public/787-data-science-coleta-de-dados/05/Mescla_Colecoes.zip) e a biblioteca de utilidades pode ser baixada aqui (https://s3.amazonaws.com/caelum-online-public/787-data-science-coleta-de-dados/04/Util.zip).

7) Em **Mescla_Colecoes.py**, você deve dizer onde o arquivo **caixa_de_areia_NUFORC.csv**, no qual serão salvas as linhas não incorporadas, será salvo:

```
if caixa_de_areia:
    print ("----------------------------Descarregando a Caixa de Areia!")
    f = open('D://Datasets//caixa_de_areia_NUFORC.csv', 'wt')
    try:
        writer = csv.writer(f)
        for i in caixa_de_areia:
```

```
                writer.writerow(i)
            print("----------------------------Gerado arquivo caixa_de_areia.csv!")
        finally:
            f.close()
    else:
        print ("----------------------------Caixa de areia vazia!")
```

Com isso feito, você pode executar a rotina.

8) Caso a rotina seja executada com sucesso, os documentos que não foram para a caixa de areia serão inseridos na coleção **clima_consolidado**. Você já pode fazer consultas, através do **Robo 3T**, por exemplo, para contar os documentos da coleção:

```
db.clima_consolidado.count()
```

9) Logo após, conte quantos documentos possuem cor, para saber quantos documentos vieram da base do Kaggle:

```
db.clima_consolidado.find (
    { cor : { $exists : 1 } },
    { _id : 1 }
).count()
```

10) Você também pode ver a estrutura de um documento que veio do arquivos **ufos.csv**:

```
db.clima_consolidado.findOne (
    { cor : { $exists : 0 } },
    { _id : 0 }
)
```

11) Agora, visualize a frequência por ano:

```
db.clima_consolidado.aggregate ( [
    { $group : {_id : "$ano", quantos : {$sum : 1} } },
    { $sort: {quantos: -1} }
] )
```

12) A frequência por estado:

```
db.clima_consolidado.aggregate ( [
    { $group : {_id : "$estado", quantos : {$sum : 1} } },
    { $sort: {quantos: -1} }
] )
```

13) E a frequência por estado e cidade:

```
db.clima_consolidado.aggregate ( [
    { $group : {_id : {estado: "$estado", cidade: "$cidade"}, quantos : {$sum : 1} } },
    { $sort: {quantos: -1} }
] )
```