

CS 327E Project 6, due Thursday, 10/28.

This project makes use of the [Open Food Facts](#) dataset, which is a crowdsourced database of food products. Before beginning the assignment, read the [data dictionary](#) for this dataset so that you know what the various fields mean.

The goal of this assignment is to gain some practice writing and executing various CRUD operations against a MongoDB database. As you work through the questions, it is a good idea to consult the official [documentation](#) if you have doubts about which operators or methods to use.

Open a terminal window in JupyterLab and download the dataset from Google Cloud Storage. Run the following commands to download and extract the dataset:

```
gsutil cp gs://cs327e-open-access/open_foods.zip .
unzip open_foods.zip
```

The extracted folder contains a mongodb dump file `products.bson` which is a full database backup along with a metadata file `products.metadata.json`.

In the same terminal, restore the dump file by running the following command:

```
mongorestore -d open_food -c products dump/open_food/products.bson
```

The restore command may take up to 5 minutes to complete. It will create a MongoDB database called `open_food` with a `products` collection inside it. The collection will be populated with 309,370 documents from the Open Food Facts dataset. In addition to the data in this collection, the restore command also creates several indexes on the same collection. Creating these indexes is actually what consumes the lion's share of the restore time.

Once you have successfully run the restore command, create a new Python Jupyter notebook and name it `project5.ipynb`.

In your notebook, translate the following SQL queries into Mongo's query language. Place each query into its own notebook cell and run each cell individually.

Q1.

```
select count(*)
from products;
```

Q2.

```
select _id, product_name
from products
where categories = 'Snacks, Sweet snacks, Confectioneries,
                  Candies, Chews';
```

Q3.

```
select _id, code, product_name
from products
where last_modified_t >= 1601856000;
```

Q4.

```
select count(*)
from products
where packaging = 'Plastic';
```

Q5.

```
select _id, code, creator, product_name, brands
from products
where manufacturing_places = 'Austin, TX'
and stores = 'Whole Foods';
```

Q6.

```
select _id, creator, product_name, brands
from products
where brands = 'Trader Joes' and product_name is not null
order by product_name;
```

Q7.

```
select _id, product_name, brands
from products
where brands in ("m&m's", "mars", "Mars", "oreo", "starburst")
order by product_name
limit 5;
```

- Insert a new document into the `products` collection. The document must have a minimum of 5 attributes with non-NULL values. Read back the document you just created.
- Update the document you created in the previous step and then read it back.
- Delete the document you updated in the previous step and then query the collection to verify that it's been deleted.

CS 327E Project 6 Rubric

Due Date: 10/28/21

Download and extract the open food facts dataset to your jupyter notebook instance. -3 no dataset or incorrect dataset found in Jupyter instance	3
Run <code>mongorestore</code> to create the database and populate the collection. -3 <code>open_food</code> database or <code>products</code> collection not found	3
Implement queries Q1 - Q7. -7 for each missing, incomplete or incorrect query -3 for each missing or incorrect output	70
Run an <code>insert</code> followed by a <code>find</code> to read back the newly inserted document. -3 missing, incomplete or incorrect insert -3 missing, incomplete or incorrect find -2 for missing or incorrect output	8
Run an <code>update</code> followed by a <code>find</code> to read back the newly updated document. -3 missing, incomplete or incorrect update -3 missing, incomplete or incorrect find -2 for missing or incorrect output	8
Run a <code>remove</code> to delete the document you inserted in the previous step. -3 missing, incomplete or incorrect remove -3 missing, incomplete or incorrect find -2 for missing or incorrect output	8
<code>project6.ipynb</code> pushed to your group's private repo on GitHub. Your project will not be graded without this submission.	Required
<code>submission.json</code> submitted into Canvas. Your project will not be graded without this submission. The file should have the following schema: <pre>{ "commit-id": "your most recent commit ID from GitHub", "project-id": "your project ID from GCP" }</pre> Example: <pre>{ "commit-id": "dab96492ac7d906368ac9c7a17cb0dbd670923d9", "project-id": "some-project-id" }</pre>	Required
Total Credit:	100