

CS 327E Project 5, due Thursday, 10/15.

This project makes use of the [Open Food Facts](#) dataset. Before beginning the assignment, [read up](#) on the various fields that make up this dataset.

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 data contains the mongodb dump file `products.bson` along with a metadata file `products.metadata.json`.

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

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

This command should take 5-10 minutes to run. It creates a database `open_food` with a `products` collection inside it.

Create a new Python Jupyter notebook and name it `project5.ipynb`.

Translate the following SQL queries to Mongo's query language. Add each query to its own notebook cell and run them with the mongo shell.

Q1. 

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

Q2. 

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

Q3. 

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

Q4. 

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

Q5. `select code, creator, product_name, brands  
from products  
where manufacturing_places = 'Austin, TX'  
and stores = 'Whole Foods';`

Q6. `select code, creator, product_name, brands  
from products  
where brands = "Trader Joe's" and product_name is not null  
order by product_name  
limit 7;`

Q7. `select code, product_name, brands  
from products where brands in ("M&M's", "Reese's", "Mars  
Chocolate, Mars")  
  
order by product_name  
limit 8;`

- Insert a new document into the `products` collection. The document must have at least 5 fields. Read back the document you just created.
- Update the document you created and then read it back.
- Delete the document from the collection.

CS 327E Project 5 Rubric

**Due Date: 10/15/20**

Download and extract the open food facts dataset to your jupyter notebook instance. -5 no dataset or incorrect dataset found in Jupyter instance	5
Create a new Python Jupyter notebook named <code>project5.ipynb</code> . -5 incorrect file name	5
Implement queries Q1 - Q7. -10 for each missing or incorrect query -7 for each missing output	70
Run an <code>insert</code> followed by a <code>find</code> to read back the newly inserted document. -5 missing, incomplete or incorrect insert -3 missing or incorrect find	8
Run an <code>update</code> followed by a <code>find</code> to read back the newly updated document. -5 missing or incorrect update -3 missing or incorrect find	7
Run a <code>remove</code> to delete the document you inserted in the previous step. -5 missing or incorrect remove	5
<code>project5.ipynb</code> pushed to your group's private repo on GitHub. Your project <b>will not</b> be graded without this submission.	<b>Required</b>
<code>submission.json</code> submitted into Canvas. Your project <b>will not</b> 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> <p>Example:</p> <pre>{   "commit-id": "dab96492ac7d906368ac9c7a17cb0dbd670923d9",   "project-id": "some-project-id" }</pre>	<b>Required</b>
<b>Total Credit:</b>	<b>100</b>