CentraleSupélec

CENTRALESUPÉLEC
2025-2026

Big Data, Techniques and Platforms

Document Databases

The purpose of this assignment is to continue working with MongoDB. You will use a dataset that consists of a collection that is a sample of the *Open Food Facts* dataset. Open Food Facts is a free, online, and crowdsourced database of food products from around the world.
As stated before we will use a sample but if you want to perform your own analysis with the full dataset (that is daily updated and contains 267918 products[1])
You can find more information at this link:

`https://world.openfoodfacts.org/`

All necessary files for the assignment are:

- One data file: `openfoodfacts.bson`

In the following we provide a partial view of one of the elements of the collection. You can find a detailed but raw description of the fields of the collection at this link:

`https://static.openfoodfacts.org/data/data-fields.txt`

```
{
"_id" : "0",
"categories_hierarchy" : [
"en:beverages"
],
"traces_from_user" : "(de) ",
"packaging" : "Kunststoff",
"data_quality_warnings_tags" : [
"en:nutrition-value-very-high-for-category-proteins"
```

[1] data refers to 9th November 2021

```
],
"categories_properties_tags" : [
"all-products",
"categories-known",
"agribalyse-food-code-unknown",
"agribalyse-proxy-food-code-unknown",
"ciqual-food-code-unknown",
"agribalyse-unknown"
],
"nutrition_data_per" : "100g",
"last_editor" : "prepperapp",
"misc_tags" : [
"en:nutriscore-not-computed",
"en:nutrition-not-enough-data-to-compute-nutrition-score",
"en:nutrition-no-saturated-fat",
"en:main-countries-de-ingredients-not-in-country-language",
"en:main-countries-de-only-1-field-in-country-language"
],
"countries_lc" : "en",
"allergens_from_ingredients" : "",
"nutrition_score_beverage" : 1,
"main_countries_tags" : [ ],
"nova_group_tags" : [
"not-applicable"
],
"amino_acids_tags" : [ ],
"allergens" : "",
"nutriments" : {
"proteins_unit" : "g",
"energy-kcal_100g" : 115,
"carbohydrates_100g" : 1,
"energy" : 481,
"energy-kcal_unit" : "kcal",
"proteins_value" : 27,
"energy-kcal_value" : 115,
"energy-kcal" : 115,
"energy_unit" : "kcal",
"energy_100g" : 481,
"proteins" : 27,
"proteins_100g" : 27,
"fat_unit" : "g",
"fat_value" : 1.3,
"energy_value" : 115,
"carbohydrates" : 1,
"carbohydrates_value" : 1,
"carbohydrates_unit" : "g",
"fat_100g" : 1.3,
"fat" : 1.3
},
"rev" : 81,
"allergens_hierarchy" : [ ],
"vitamins_tags" : [ ],
"data_quality_bugs_tags" : [
"en:code-zero"
],
"packaging_tags" : [
"kunststoff"
],
"categories" : "en:beverages",
"lc" : "de",
"last_modified_by" : "prepperapp"

...
}
```

For the exercise remark the fact that you are working with real data and then you can have all the problems related with the analysis of real data (outliers, etc.).
Download the file and import it using Studio3T `data`.

1 Exercises

Now you can study data and provide the set of required queries.

1.1 Exercise: understanding data - (**1 point**).

Before starting with the queries look at the documents and provide a short description of them: the most common structure of the documents (the most common attributes, nested documents, etc.).

1.2 Exercise: querying data

Provide now the queries that answer the following questions. For this assignment you can upload on Edunao a file that includes:

- The answer to the Exercise `Understanding data`
- The query and the obtained output for the following questions. For each query you must also show how the results are returned (and give a sample – max 10 lines).

---

1. **(1 point)** The number of `products` in the collection.

2. **(1 point)** The product that has `Sharon's, sorbet, dutch chocolate` as name.

3. **(1 point)** How many times the product having `0009073102079` as `_id` has been modified. Pay attention: how do you match the value of the `_id`? Think about how to do and how you should have done this match in another context.

4. **(1 point)** The products that have `sodium` in the `nutriments` list.

5. **(1 point)** The products that have the `nutriscore_grade` equal `c`.

6. **(1 point)** How many different `creators` participated in the product creation.

7. **(2 points)** How many `creators` have created more than one product.

8. **(1 point)** The product(s) which are modified most recently.

9. **(1 point)** The products that have exactly `1` ingredient.

10. **(2 points)** The products that have 20 or more ingredients.

11. **(0,5 points)** How many products are characterized as `desserts` (`dessert` is in the `_keyword` list).

12. **(0,5 points)** How many products are characterized as `chocolate` (`chocolate` is in the `_keyword` list).

13. **(1 point)** How many products are characterized as `chocolate` and `dessert` (`chocolate` and `dessert` are in the `_keyword` list).

14. **(1 point)** How many products are characterized as `chocolate` or `dessert` (`chocolate` or `dessert` are in the (`_keyword` list).

15. **(2 points)** For the documents inside the collection provide a query that converts the type of field `categories` from a `String` to an `array` and moves data into the new attribute called `new_att_category` (each category as an element of the array.).

16. **(2 points)** How many products have `nutriscore_grade` equal to `F` and contain ingredients with `palm-oil`.