

Big Data, Techniques and Platforms

Document Databases & Structured Query Language

The purpose of this assignment is to continue working with MongoDB. You will use a dataset consisting of a collection sample of the *Open Food Facts* dataset. Open Food Facts is a free, online, and crowdsourced database of food products worldwide.

As stated before, we will use a sample, but if you want to perform your analysis with the full dataset (that is daily updated and contains 267918 products¹) you can find available for downloading at this link:

<https://world.openfoodfacts.org/data>

All necessary files for the assignment are:

- One data file: `openfoodfacts.json`

The file provides you a sample of the dataset. 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",
```

¹data refers to 9th November 2021

```

"data_quality_warnings_tags" : [
  "en:nutrition-value-very-high-for-category-proteins"
],
"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 store it in a directory of your computer.

1 IMPORT DATA

Import data in your MongoDB database, as you did for the lab you can take advantage of the Studio3T graphical interface.

2 EXERCISES

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

2.1 EXERCISE: UNDERSTANDING DATA - (1 POINT).

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

2.2 EXERCISE: QUERYING DATA

For this assignment you can fill in the form and each answer must include:

- The query and the obtained output for the following questions.
- A sample of the returned results (no more than 10/15 significative lines).
- Do not forget the answer to the Exercise Understanding data.

Provide the queries that answer the following questions and return:

2. **(1 point)** The number of products in the collection.
3. **(1 point)** The product that has Sharon's, sorbet, dutch chocolate as name.
4. **(1 point)** The products that have the sodium in the nutriments list.
5. **(1 point)** The products that have the nutriscore_grade equal a.
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) the most recently modified.
9. **(1 point)** The products that have exactly 1 ingredient.
10. **(1 point)** How many products are characterized by chocolate and dessert (chocolate and dessert are in the _keyword list).
11. **(1 point)** How many products are characterized by chocolate or dessert (chocolate or dessert are in the (_keyword list).

12. **(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.).
13. **(2 points)** Show the products with `nutriscore_grade` equal to `A` and contain ingredients with `palm oil`.