**NLP Project 1 : Open Food Facts**

Open Food Facts is a collaborative database of food products that has been created and fed by volunteers around the world. Anyone can contribute to the enrichment of this database by filling it in themselves. The goal of Open Food Facts is to share with everyone a maximum of information about food products. It contains more than 800 000 products but not all of them are perfectly written. Indeed, one of the drawbacks of this platform could be that the information is not totally accurate nor in the right format. This would lead to the information not being comparable because it is not standardized.

In order to respond to the different questions, this project will be composed of 3 major steps:

* Importation, Exploration and Cleaning of the data
* Word2Vec Model
* KMeans and Visualisation

## Importation, Exploration and Cleaning



First of all, we import the first 500k rows of the Dataset, for complexity and memory purposes.

We then explore the data in different ways, mainly looking at 3 components: Nutrition values, Ingredients and Categories:

On Nutrition, after looking at the different columns, we noticed that a lot of them were almost empty. In order to clean this data, we can only keep the columns that are atleast half full, as shown in the screenshot.

Une image contenant texte

Description générée automatiquementWe then proceeded to do a Word Cloud on Categories\_en, in order to visualize the data better:

Une image contenant texte

Description générée automatiquement

On Ingredients, we can look at the number of null values:

Une image contenant texte

Description générée automatiquement

Finally, we can choose to only keep the relevant columns in regard to the aforementioned elements:

Une image contenant texte

Description générée automatiquement

## Word2Vec

We first tokenize the different ingredients and modify them a bit to clean it up.

Une image contenant texte

Description générée automatiquement

We then looked at the frequencies and the amount of tokens that are the most repeated, also cleaned them by getting rid of Stop Words or the word “and/or”, that was not part of the StopWords dictionary.

After that, a Word Cloud is possible to visualize the data better.

Une image contenant texte

Description générée automatiquement

We can also look at bigrams, which are 2 words that work together semantically-speaking

Une image contenant table

Description générée automatiquement

## KMeans and Visualization

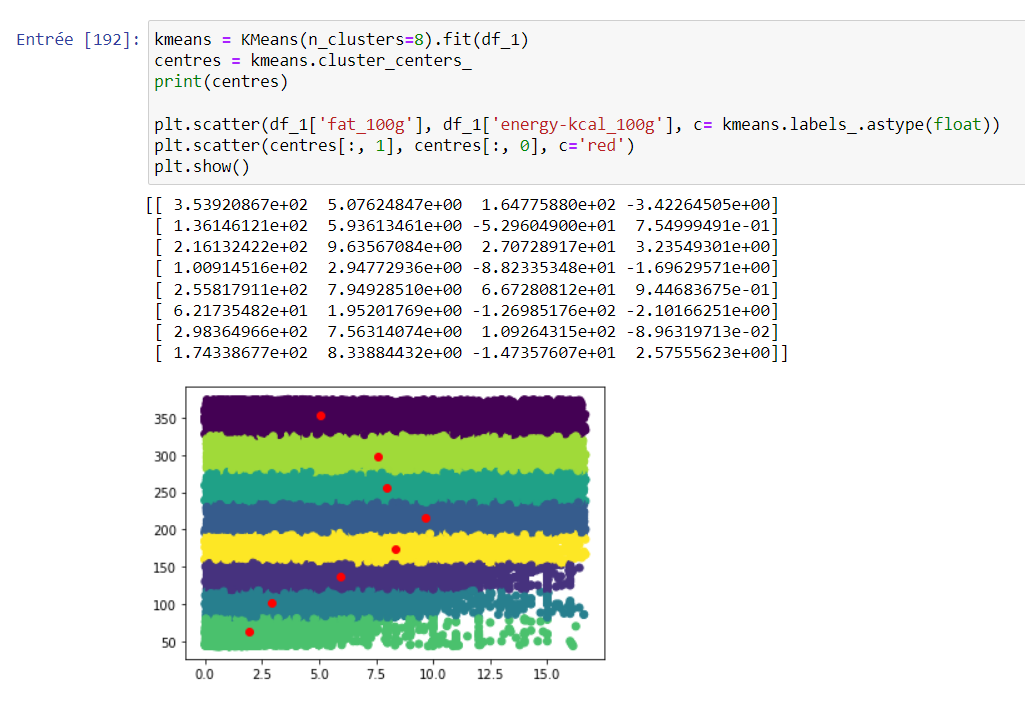
## Une image contenant texte Description générée automatiquement

The first idea of KMeans similarity that I had is to look at the similarities between categories and create clusters with it, however I ran into problems which made me unable to finalize the KMeans (compatibility between different variables types).

I then decided to look at other values that may be interesting: the energy and fat contained in products:

Une image contenant texte

Description générée automatiquement

We can choose to visualize clusters, here I chose 8 to keep a good readability.

We can see that, the more fat is contained in a food, the more it has a high energetic value. However, passed a certain energetic value, the amount of fat begins to reduce.

Conclusion:

To conclude, we have explored the open food facts data and made visualization on this data, to finally create clusters based on different values.

**Links:**

https://github.com/mehdinoureddine/NLP\_project\_1