IA SCHOOL

L'ÉCOLE DE L'INTELLIGENCE ARTIFICIELLE

Mastère 2

EXAMEN Analyse de donnée avec Python : OPEN FOOD FACT

ANNEE

2020 – 2021

GEMA

GRANDES ÉCOLES DES MÉTIERS D'AVENIR

Présentation du projet

Ce projet consiste à récupérer des données sur le site **Open Food Fact.** Il s'agit d'une liste de **796 969 produits** classés par popularité (Nombre de scan de l'application Yuka). Il s'agit dans un premier temps, de constituer un dataframe à l'aide des techniques de scrapping vues en cours puis d'analyser les données obtenues. Vous trouverez ci-dessous l'url du site à scrapper.

https://fr.openfoodfacts.org/

Nous disposons d'un site dans lequel se trouve de l'information caractérisant des produits de la vie courante achetés en grande surface (nom du produit, quantité, nutri score etc…). Ce site comporte **7970 pages** avec 100 produits. La seconde partie du projet consistera à analyser les données récoltées et à répondre à l'ensemble des questions figurant dans la partie 2 du document.

Ce projet a pour but de **consolider vos acquis techniques en programmation Python** (Algorithmique, Analyse de données, Scrapping). Deux rendus sont attendus, un script .py et un fichier .ipynb avec l'ensemble des analyses demandées.

1. Scrapping de données (10 pts):

Il s'agira de créer un **script python** permettant de récupérer l'ensemble des informations pertinentes du site **Open Food Fact** en France dont le lien figure ci-dessus. Ce script devra prendre en compte le nombre de pages contenant l'information et permettra de récupérer les **796 969 produits du site** (Informations qui caractérisent le produit). On pourra utiliser pour ce faire le package **Beautifulsoup**.

Ce script devra être construit sous la forme d'un fichier .py. Il devra comprendre des commentaires, être écrit en fonction, comporter des arguments par défaut et les docstring d'usage. Il devra également comprendre un outil de mesure de temps de computation des fonctions.

Ce script permettra, entre autres, de récupérer dans un dataframe les éléments ci-dessous.

- Nom du Produit
- Code-barres (EAN/EAN-13)
- Nutri-score
- NOVA
- Eco-Score
- Caractéristiques du produit
- Quantité
- Conditionnement
- Marques
- Catégories
- Labels, certifications, récompenses
- Origine des ingrédients
- Lieux de fabrication ou de transformation
- Code de traçabilité
- Lien vers la page du produit sur le site officiel du fabricant
- Magasins
- Pays de vente
- Analyse des ingrédients :
    - Additifs
    - Ingrédients issus de l'huile de palme
- Repères nutritionnels pour 100 g
    - Matières grasses / Lipides en quantité élevée
    - Acides gras saturés en quantité élevée
    - Sucres en quantité élevée
    - Sel en faible quantité
- Comparaison avec les valeurs moyennes des produits de même catégorie (Catégories cochées)
- Informations nutritionnelles
    - Energie (kcal)
    - Nombre de calorie (Énergie (kcal))
- Impact environnemental

La récupération de variable supplémentaire à celle ci-dessus est laissée à l'appréciation de l'étudiant. Certaines variables **peuvent ne pas être présentes sur toutes pages**, il s'agira alors de **remplacer par "XXX"**.

2. Analyse de données (6 pts):

Une fois le dataset constitué, vous devrez répondre aux questions suivantes:

1. Distribution du nombre de produits par catégorie Nutri Score, Nova et Eco score.(5 *4*5 = 100 possibilités)

2. Liste des produits ayant “gluten free” optimaux (Nutri score A, Nova 1 et Ecoscore A)
3. Nombre de produit dont le ratio sucre/produit est supérieur à 0.6
4. Liste des produits dont le nombre de calories pour 100g est supérieur à 500.
5. Liste des produits contenant de l’huile de palme, vendu en France et en Belgique
6. Distribution des produits par nombre de pays de vente

Il vous est également demandé de proposer **au minimum 4 analyses** en plus de celles qui vous sont demandées ci-dessus. Plus que la complexité des analyses en elle-même, c’est leur cohérence qui sera valorisée. Vous pourrez cibler un axe de recherche et le développer. Par exemple, les produits pour les personnes ayant un régime alimentaire particulier, diabétique, végétarien, les intolérants au gluten et autres.

Une soutenance de **présentation du projet sera organisée à la durant la semaine d’examen**.

Livrables

Les livrables attendus pour chaque étudiant sont :

- Script de scrapping des données optimisé et commenté
- Notebook d'analyse de données en Python

On valorise particulièrement la capacité de l'élève à aller au-delà des analyses proposées et à expliciter son raisonnement en vue de répondre à des problématiques business.

Modalités de rendu

A. Rendu écrit

> Le rendu écrit doit se faire sur **Classroom**, dans le devoir intitulé « *Analyse de données – Mastère 2* », avec tous vos documents,
>
> **au plus tard le 25 Mai 2021**.

B. Soutenance orale

La soutenance orale se tiendra sur la semaine du 24 mai, et aura lieu en visioconférence via Google Meet.

> **Durée : 15 minutes**
>
> ⇒ 10 minutes de présentation et de question de la partie technique
> ⇒ 5 minutes de point d'amélioration sur la partie technique

Références/liens utiles :

- Beautifulsoup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- Selenium: https://selenium-python.readthedocs.io/
- Dash : https://dash.plotly.com/
- Flask:https://dev.to/admindashboards/flask-dashboard-atlantis-dark-open-source-admin-panel-with-dark-design-4l6
- Yuka: https://yuka.io/
- Nutriscore: https://fr.openfoodfacts.org/nutriscore
- Eco score: https://fr.openfoodfacts.org/eco-score-l-impact-environnemental-des-produits-alimentaires
- Nova: https://fr.openfoodfacts.org/nova