Principe – Utilisation de l'API

Les catégories seront d'abord récupérées, puis les produits appartenant à ces catégories.

I. Catégories

Les catégories doivent remplir certaines conditions pour limiter l'overlap:
- le nom ("name") ne doit pas contenir de tags langage (en: , es: ...), cela permettra de sélectionner uniquement les catégories dont le nom est en français.
-l'id ne doit pas dépasser 3 mots, la présence de "and" dans l'identifiant doit être évitée.
-seules les catégories contenant moins de 6000 produits mais plus de 500 seront sélectionnées.
-les produits seront récupères en partant de la catégorie la moins peuplée, la moins susceptible d'avoir un trop grand nombre de sous catégories.

Ces mesures ne seront bien évidemment pas suffisantes pour éviter totalement l'overlap mais l'exclusion des catégories "principales", celles contenant 90% des autres catégories (boissons, snacks, aliment d'origine végétale) limitera fortement les catégories inutiles ou toute comparaison serait impertinente (oui de l'eau est plus saine que du coca cola, les deux appartiennent a la catégorie boisson mais ne partagent comme trait que le mode d'ingestion, pas pertinent donc).

On peut éventuellement imaginer sélectionner uniquement les catégories contenant moins de 1000 produits (ou 500). Cela ralentira considérablement l'application mais les recherches seront faites avec beaucoup plus de précision.

II. Produits

Les produits suivront également des règles pour éviter les données inutilisables.

- On filtrera les produits de cette façon :
    - Le code barre du produit doit être complet (13 chiffres) et doit être unique (indexe au préalable pour éviter les doublons).
    - Le produit doit contenir une marque ('brands') (pour différencier deux produits ayant la même désignation par exemple).
    - Le nutriscore du produit doit être établi, sans ca, pas d'outil de comparaison, pas d'utilité dans la base de donnée.