Principe

Les categories seront d'abord recuperees, puis les produits appartenant a ces categories.

I. Categories

Les categories doivent remplir certaines conditions pour limiter l'overlap:  
- le nom ("name") ne doit pas contenir de tags langage (en: , es: ...), cela permettra de selectionner uniquement les categories dont le nom est en francais.  
-l'id ne doit pas depasser 3 mots, la presence de "and" dans l'identifiant doit etre evitee.  
-seules les categories contenant moins de 6000 produits mais plus de 500 seront selectionnees.  
-les produits seront recuperes en partant de la categorie la moins peuplee, la moins suceptible d'avoir un trop grand nombre de sous categories.

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 eventuellement imaginer selectionner uniquement les categories contenant moins de 1000 produits (ou 500), et dans le cas ou la comparaison ne peut être faite efficacement on fera appel au N+1 présent dans categories\_hierarchy (dans chaque produit). Cela ralentira considerablement l'application mais les recherches seront faites avec beaucoup plus de precision.

II. Produits

Les produits suivront également des regles pour eviter les donnees inutilisables.

- On filtrera les produits de cette facon :   
 - Le code barre du produit doit etre complet (13 chiffres) et doit etre unique (indexe au prealable pour eviter les doublons)  
 - Le produit doit avoir une "categories\_hierarchy" plus longue que 2 (utilisation du n+1 voire n+2) impossible sans cela  
 - Le produit doit contenir une marque ('brands') (pour differencier deux produits ayant la meme designation par exemple)  
 - Le nutriscore du produit doit etre etabli, sans ca, pas d'outil de comparaison, pas d'utilite dans la base de donnee  
 - Au moins un point de vente doit etre etabli : c est bien sympa de savoir qu'un produit est plus sain qu'un autre, mais sans moyen de le consommer, le but de l'application est extremement limite.