

Extraction des ingrédients depuis les fiches techniques de produits alimentaires

Pierre MASSÉ

Juin 2020

Résumé

La gestion de l'information produit est devenu un enjeu de société majeur ces dernières années. Les scandales sanitaires récents ont déclenché une prise de conscience collective des consommateurs, en parallèle de la mise en place de réglementations de plus en plus contraignantes pour l'ensemble des acteurs de la filière [3][4].

La mise en place d'outils mettant en oeuvre les principes du Machine Learning appliqués au traitement du langage permettrait d'aider les opérationnels de la gestion de l'information à interpréter plus vite et mieux les documents mis à disposition par les fournisseurs du groupe. Le présent rapport détaille la mise en place d'un outil permettant d'extraire les listes d'ingrédients des fiches techniques transmises par les fabricants des produits.

Dans une première partie, on détaillera le contexte métier autour de la gestion de l'information produit. Nous nous analyserons ensuite les données disponibles, et les possibilités de les exploiter. Nous expliciterons alors les raisons qui ont poussé à choisir le cas d'usage, et la manière dont il pourrait être mis en oeuvre au sein du Groupe Pomona. La quatrième partie sera consacrée à la construction du modèle à proprement parler, et la dernière partie ouvrira la réflexion sur les travaux à venir.

TABLE DES MATIÈRES

I Contexte métier	6
1 Description du groupe	6
1.1 Le métier du Groupe Pomona	6
1.2 La décentralisation	7
1.2.1 Les Directions fonctionnelles	7
1.2.2 Les clients du groupe	8
1.2.3 Premier niveau de décentralisation : les branches	8
1.2.4 Le second niveau de décentralisation : les succursales	11
2 La gestion de l'information produit	11
2.1 L'information produit	13
2.1.1 Utilisations de l'information produit	13
2.1.2 Des produits bruts aux produits transformés	14
2.1.3 Les grands types d'information	14
2.2 Le processus	18
2.2.1 Le fournisseur est propriétaire des informations produit	18
2.2.2 La notion de produit et d'article	18
2.2.3 Les acteurs	20
2.2.4 Les contrôles	20
2.3 Les outils informatiques associés	21
2.3.1 Les branches faiblement outillées	21
2.3.2 Le GIP	23
2.3.3 Le PIM	23
II Les données	28
3 Le périmètre produit	28
3.1 Les produits non-alimentaires	28
3.2 Accessibilité de la donnée en fonction des branches	29
3.3 Analyses quantitatives	29
3.3.1 Comparatifs entre les branches	29
3.3.2 Les grands types de produits	30

4 Les données utilisables, issues du PIM	31
4.1 Données structurées	34
4.1.1 Description des données structurées	34
4.1.2 Analyse de ces données structurées	35
4.2 Données non structurées	42
4.2.1 Les libellés	42
4.2.2 Les listes d'ingrédients	42
4.3 Pièces jointes	43
4.3.1 Fiches techniques fournisseur	44
4.3.2 Étiquettes produit	45
4.4 Récapitulatif de la complétude des données	46
4.5 Analyse qualitative des données	47
4.5.1 Évaluation de la qualité des données	47
4.5.2 Types de pdf possédés	47
4.6 Les données « manuellement étiquetées »	47
4.6.1 Pour répondre à quel besoin ?	47
4.6.2 Mode de constitution de l'échantillon	47
4.6.3 Méthodologie de l'étiquetage manuel	48
4.6.4 Règles de gestion pour l'étiquetage manuel	48
4.6.5 Confrontation avec le contenu du PIM	48
III Les objectifs de ce projet	50
5 Les cas d'usage	50
5.1 Objectifs : Qualité et productivité	50
5.2 La préalimentation d'information	50
5.3 Le contrôle des informations transmises	50
5.3.1 Le contrôle à la saisie fournisseur	51
5.3.2 L'aide aux vérifications Pomona	51
5.3.3 Les contrôles en masse asynchrones	51
6 Le choix du cas d'usage	51
6.1 La représentation dominante des fiches techniques	51
6.2 Les multiples formats et le besoin de « spatialisation »	51
6.3 Les informations « spatialisées »	52
6.4 La complexité dans la représentation des données logistiques	52
6.5 L'identification d'une liste d'ingrédient par son contenu	52

6.6 Conclusion quant au choix du cas d'usage	52
IV Construction du modèle	53
7 Les principes généraux	53
7.1 Contenu du texte d'une liste d'ingrédients	53
7.2 Limitation à l'identification des listes d'ingrédients	53
7.3 Conversion de documents en texte	53
8 Construction d'un modèle simple « ouvert »	54
8.1 Entraînement	55
8.1.1 Périmètre	55
8.2 Conversion en blocs de texte	55
8.3 Train/Test split	55
8.4 Entrainement du modèle	55
8.5 Calcul de la similarité	55
8.6 Illustration des résultats obtenus	55
8.7 Pistes d'améliorations identifiées	56
9 Utilisation des données manuellement étiquetées	56
9.1 Chargement des données manuellement étiquetées	56
9.2 Découpage des textes en blocs	57
9.3 Train/Test split	58
9.4 Entraînement du modèle	59
9.5 Illustration des prédictions obtenues	59
10 Mesure de la performance	62
10.1 Accuracy	62
10.1.1 Approche naïve	63
10.1.2 Avec du « text-postprocessing »	64
10.2 Fonctions de « loss » spécifiques	64
10.2.1 Distance de Levenshtein	66
10.2.2 Distance de Dameray-Levenshtein	66
10.2.3 Distance de Jaro	66
10.2.4 Distance de Jaro-Wrinkler	66
10.2.5 Métriques non retenues	66
10.2.6 Conclusion sur la métrique à utiliser	66

11 Transfer learning	66
11.1 Principe du pré-entraînement	66
11.2 Illustration de l'impact sur la performance	66
12 Hyperparameter tuning	67
12.1 Les paramètres ajustables	67
12.1.1 La prise en compte des « n-grams » dans la tokenization	67
12.1.2 L'application de « n-grams » de blocs	67
12.1.3 L'utilisation d'expressions régulières dans le split des blocs	67
12.1.4 Applications d'autres fonctions de similarité	67
12.2 Application d'une grid search	67
V Travaux subséquents	68
13 Opérationnalisation de cette maquette	68
13.1 Client et sponsor métier	68
13.2 Sélection du use case	68
13.3 Mise en place d'une organisation projet	68
13.3.1 Identification des compétences nécessaires	68
13.3.2 Choix d'un cadre méthodologique projet	68
13.3.3 Développement côté PIM	68
13.4 Industrialisation du code du modèle	69
13.5 Monitoring de la performance du modèle	69
14 Extension des fonctionnalités offertes	69
14.1 Prise en compte de nouveaux types de pièces jointes	69
14.2 Utilisation d'outil d'OCR pour les pdf non structurés	69
14.3 Mise en place d'outil de spatialisation des textes	69
14.4 Construction d'outils d'extraction de données connexes à la composition	69
14.5 Élargissement aux données nutritionnelles	70
14.6 Évaluation de la performances sur d'autres familles de produits	70
VI Annexes	71
A Figures, tableaux et bibliographie	71

B Exemples de pièces jointes et ground truth	73
B.1 Fiches techniques	73
B.1.1 Fiche technique sel Cerebos	74
B.1.2 Fiche technique olives Valtonia	75
B.1.3 Fiche technique Panna Cotta Nestlé	76
B.1.4 Fiche technique confiture Andros	78
B.1.5 Fiche technique ciboulette La case aux épices	80
B.1.6 Fiche technique poivron El Arenal	82
B.1.7 Fiche technique mélange trappeur Terre Exotique	85
B.2 Étiquettes produit	86
B.2.1 Étiquette curry Grain d'ailleurs	86
B.2.2 Étiquette madeleines Saint Michel	87
B.2.3 Étiquette lentilles Soufflet	88
B.2.4 Étiquette pannacotta Nestlé	89
B.2.5 Étiquette sauce soja Kikkoman	90
B.2.6 Étiquette mélange trappeur Terre Exotique	91
B.3 Étiquetage manuel des données	92
B.3.1 Règles de gestion pour l'étiquetage	92
C Les notebooks de ce projet	96
C.1 Analyse quantitative	97
Analyse des données du PIM	107
C.2 Génération de l'échantillon de données manuellement étiquetées	119
C.3 Modèle « ouvert »	124
C.4 Modèle basé sur les données manuellement étiquetées	130
C.5 Mesure de la performance	140
D Le code des différents modules	148
D.1 Gestion du fichier de configuration - Module conf	148
D.2 Extraction des données du PIM - Module pimapi	148
D.3 Conversion des pièces jointes en textes - Module pimpdf	151
D.4 Transformateurs et estimateurs spécifiques - Module pimest	153

Première partie

CONTEXTE MÉTIER

Chapitre 1 DESCRIPTION DU GROUPE

L'objet de l'ensemble de cette première partie est de donner sur le Groupe Pomona des éclairages nécessaires à la compréhension du cas d'usage développé. Bien d'autres aspects sur la société, pourraient être mentionnés (ex : des indicateurs sur l'activité, l'histoire du groupe...) mais ils seront omis car non indispensables à la compréhension du sujet. Plus de détails sur le groupe sont accessibles sur le site web de la société[10].

1.1 Le métier du Groupe Pomona

Le Groupe Pomona est une société de distribution livrée de produits alimentaires à destination des professionnels des métiers de bouche. L'activité du groupe consiste uniquement à acheter et revendre de la marchandise, à l'exclusion de toute activité de fabrication ou de transformation¹. Le groupe Pomona est une société de *distribution*. Elle ne possède d'ailleurs pas d'actif industriel (autre que des entrepôts logistiques) ni d'agrément pour transformer les marchandises.

Cette activité d'achat/vente se fait dans la majorité des cas sous le régime du *négoce*, à savoir que le groupe acquiert la propriété des marchandises qu'il commercialise avant de la céder à ses clients. L'autre régime est celui dit de la *prestation (logistique)*. Dans ce cas, par le jeux d'écritures comptables, la valorisation du stock disparaît des comptes du groupe. Néanmoins, indépendamment de cet aspect purement comptable, l'ensemble :

des flux de documents : commandes d'achat, factures fournisseur, commandes de vente, factures clients

des flux financiers : paiements fournisseur, paiements client

des flux physiques : réception et stockage, préparation et expédition

¹. de très rares cas de transformation existent (ex : mûrissement de fruits, filetage de poisson) mais sont extrêmement exceptionnels

restent largement inchangés.

Pour résumer, l'activité de l'ensemble des entités du groupe pourraient se résumer via le schéma présenté à la FIGURE 1 page 7

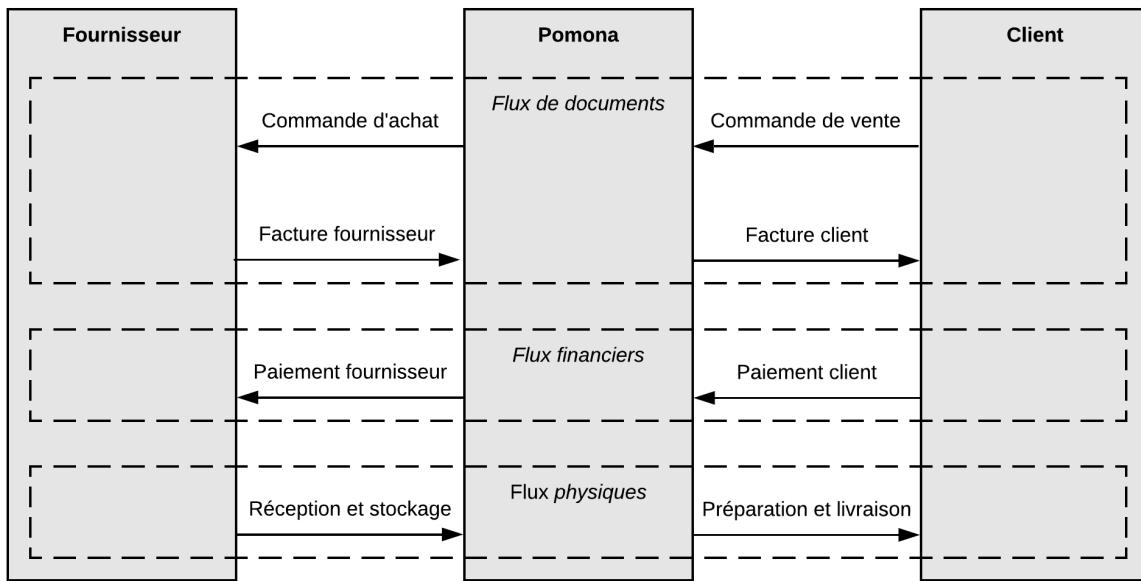


FIGURE 1 – Les flux métier avec les partenaires commerciaux

Le métier du groupe est d'être un grossiste, qui achète et revend des produits alimentaires² sans produire ou transformer quoi que ce soit.

1.2 La décentralisation

Le Groupe Pomona est un groupe fortement décentralisé, avec des organisations largement indépendantes les unes des autres.

1.2.1 Les Directions fonctionnelles

Pour des raisons évidentes de recherche de synergies ou de conformité réglementaires, certaines activités restent toutefois mutualisées à la maille du groupe. Il s'agit des organisations suivantes :

La Direction Administrative et Financière (DAF) : regroupe les équipes comptables groupe, l'autorité interne et la consolidation financière

2. dans la grande majorité des cas, cf. *Les produits non-alimentaires* 3.1 page 28

La Direction Qualité Sécurité et Environnement (DQSE) : est en charge de définir et contrôler l'application des standard de qualité

La Direction des Systèmes d'Information (DSI) : développe et maintient en condition opérationnelles les systèmes d'information du groupe

La Direction Technique et Logistique (DTL) : est en charge des projets immobiliers (entrepôts), des négociations avec les transporteurs et joue un rôle de conseil interne sur les sujets logistiques

La Direction des Ressources Humaines : se charge de l'ensemble des aspects en lien avec le recrutement, la paye et les sujets sociaux

La Direction Commerciale groupe (DCG) : définit une stratégie et des bonnes pratiques commerciales et marketing

1.2.2 Les clients du groupe

Afin de comprendre l'organisation du groupe, il est nécessaire de connaître la typologie de ses clients. Comme mentionné précédemment, le groupe s'adresse exclusivement aux professionnels des métiers de bouche. Aucune marchandise n'est vendue à des particuliers. Les principales typologies de clients sont les suivantes :

Les Sociétés de Restauration : elles exploitent les restaurants d'entreprise, certaines cantines d'établissements d'enseignement, les maisons de retraite, ...

Les Marchés Publics : regroupent les clients qui dépendent des collectivités (écoles, hôpitaux, prisons, ...)

La restauration commerciale : est l'ensemble des restaurants à vocation commerciale, qu'ils soient chaînés (hippopotamus, O'Tacos, ...) ou indépendants (« le restaurant du coin »)

Les spécialistes : il s'agit des détaillants spécialisés qui s'adressent aux particuliers. Boulanger, pâtissier, bouchers, traiteurs, vente à emporter, ...

Les Grandes et Moyennes surfaces (la GMS) : sont les enseignes de la grande distribution. En général, l'accès à ces clients est compliqué par les règles mises en place par leurs centrales d'achat. Il représentent en général qu'un canal de vente d'opportunité.

Les trois premières de ces catégories représentent ce que l'on appelle la *Restauration Hors Domicile (RHD)* (ou parfois également la Restauration Hors Foyer, RHF).

1.2.3 Premier niveau de décentralisation : les branches

Le Groupe Pomona est divisé en branches, qui sont des unités opérationnelles indépendantes, et qui ont toute latitude pour gérer leurs stratégie et politique commerciales, la gestion de leurs achats, leur stratégie marketing, ... En particulier, les systèmes d'information ne sont pas identiques entre les différentes branches. Afin d'éviter de se concurrencer entre elles, leurs domaines d'activité respectifs ont été partitionnés par familles de produit commercialisés, segments client cibles et géographie.

Les branches RHD

Les branches RHD s'adressent aux clients de la Restauration Hors Domicile (cf. section 1.2.2 page 8) en France. Elles se répartissent ce marché en travaillant des gammes de produits distinctes. Il s'agit des branches historiques du groupe, qui représentent l'essentiel de son chiffre d'affaire. La répartition par produit est la suivante :

PassionFroid : spécialiste des *produits surgelés, de la viande fraîche et des produits laitiers*

ÉpiSaveurs : spécialiste des produits qui se conservent à température ambiante : *produits d'épicerie, conserves, boissons et consommables de cuisine non-alimentaires*

TerreAzur : spécialistes des *Fruits et Légumes frais, et Produits De la Mer frais*

La non-concurrence entre les branches est assurée par le fait qu'elles ne commercialisent pas les mêmes produits. Bien que nommées RHD, elles peuvent également vendre leurs produits à la grande distribution (GMS : Grandes et Moyennes Surfaces), mais généralement ces marchés sont verrouillés par les centrales d'achat des grandes enseignes. La branche TerreAzur arrive toutefois à prendre des parts de marché significatives sur ce segment. Les branches RHD utilisent le progiciel SAP comme système de gestion. La branche TerreAzur est en cours de déploiement, en 2020 environ les 2 tiers des succursales travaillent avec ce progiciel.

Les branches spécialistes

Les branches spécialistes s'adressent aux clients dits spécialistes (cf. section 1.2.2 page 8) en France. Elles sont en mesure de commercialiser tout type de produit pour répondre aux besoins de leurs clients. En particulier, elles peuvent tout à fait commercialiser certains produits qui sont également vendus par les branches RHD. Elles se répartissent la clientèle spécialiste de la manière suivante :

Délice et Création : s'adresse aux *Boulanger et Pâtissiers*

Saveurs d'Antoine : s'adresse aux *Bouchers, Charcutiers et Traiteurs*

Relais d'Or : s'adresse à la *restauration commerciale indépendante*

Comme pour les branches RHD, ces branches peuvent lorsqu'elles en ont l'opportunité vendre leurs produits à la GMS.

L'étranger

Bien que le Groupe Pomona soit une société dont l'essentiel de l'activité est faite sur le marché français, deux réseaux sont en cours de constitution sur des pays limitrophes. Ces branches sont susceptibles de travailler tout type de produit, à destination de tout type de client. Elles sont positionnées sur les marchés suivants :

Pomona Suisse : présente sur le marché Suisse

Pomona Iberia : présente sur le marché Espagnol

On peut synthétiser la répartition de l'activité par branche de la manière présentée à la FIGURE 2 page 10.

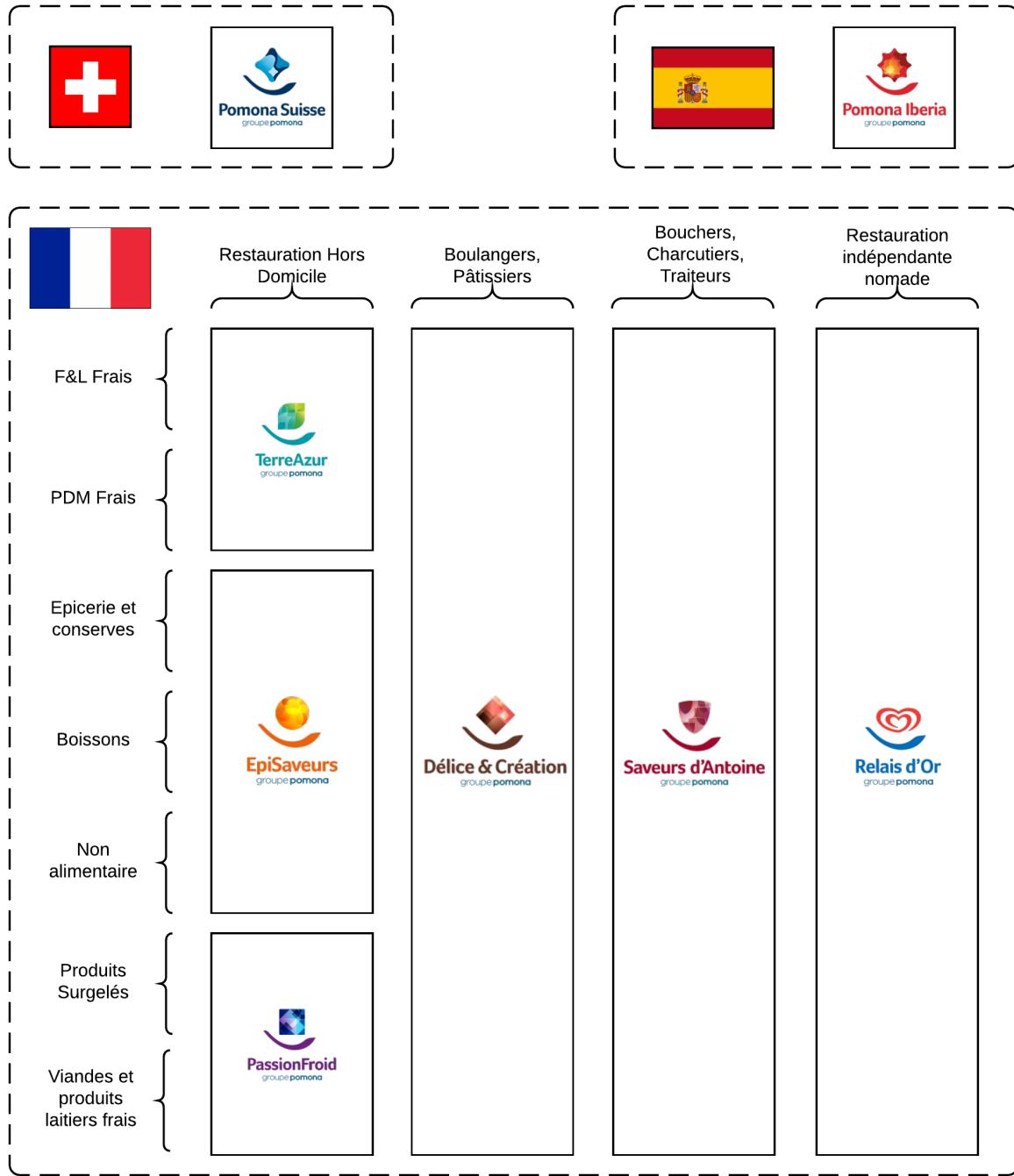


FIGURE 2 – La répartition de l'activité des branches

1.2.4 Le second niveau de décentralisation : les succursales

Chacune des branches est elle-même à son tour décentralisée en un réseau d'entrepôts régionaux : les succursales (parfois également appelées simplement « régions »). Ces succursales sont gérées comme des PME indépendantes, avec un directeur et un compte de résultat qui leur est propre. Si certaines négociations avec des fournisseurs ou des clients nationaux sont parfois menées par les branches, les succursales sont autonomes dans :

- la définition de leur assortiment, même si des contraintes s'appliquent
- la stratégie de développement commercial
- la négociation des prix d'achat
- la négociation des prix de vente
- la politique de rémunération de leurs employés

À ce titre, elles ont leurs propres équipes d'achat, leurs équipes commerciales (télévente et vente route), leurs équipes administratives et évidemment leurs équipes logistiques (essentiellement en entrepôt et les chauffeurs livreurs en charge des livraisons client).

Certaines activités restent de la responsabilité des équipes centrales des branches, comme : la négociation avec les clients ou les fournisseurs nationaux, la constitution de l'assortiment commun (les produits que toutes les succursales doivent détenir), la gestion des référentiels de données de base métier, ...

Un exemple de maillage régional est présenté en FIGURE 3 page 12, sachant que ce maillage régional est différent pour chacune des branches.

Chapitre 2

LA GESTION DE L'INFORMATION PRODUIT

Ce chapitre a pour vocation à éclairer les aspects métier en lien avec la gestion de l'information produit. C'est le seul processus métier qui sera détaillé dans la mesure où c'est uniquement celui qu'il est nécessaire de connaître pour comprendre les cas d'usage développés ultérieurement.



FIGURE 3 – Le maillage régional de la branche ÉpiSaveurs

2.1 L'information produit

2.1.1 Utilisations de l'information produit

Conformité réglementaire

La gestion de l'information produit est essentiellement une contrainte réglementaire à statisfaire. Comme mentionné au préambule, la réglementation autour de l'information des consommateur s'est sans cesse complétée au cours des dernières années. Un des textes centraux est le règlement n°1169/2011 dit INCO (INformation COnsommateur)[3][4]. C'est ce règlement qui définit l'ensemble des informations qui doivent être étiquetées sur le produit (liste d'ingrédients, tableau de données nutritionnelles, ...), mais également affichée au client lors de commande en ligne sur les sites de e-commerce. Il s'agit principalement d'informations relatives à la sécurité alimentaire (ex : les allergènes) ou la santé (ex : informations nutritionnelles).

Attentes client

Les consommateurs finaux (les « convives ») étant de plus en plus sensibles au contenu de leur assiette, les clients du groupe sont de plus en plus demandeurs d'informations relatives aux produits qu'ils commandent. Ils demandent donc régulièrement des informations qui vont au-delà de ce qui est normalement prévu par la réglementation.

De plus, sur certains marchés pour lesquels des contrats courant sur de longues périodes - jusqu'à un an - sont établis (les marchés publics sont très concernés), il n'y a pas d'échantillonnage des produits. La seule manière pour ces clients d'évaluer la qualité des produits est de se référer aux documents contenant les informations produit, fournis par les distributeurs.

Gestion

Certaines informations relatives aux produits sont nécessaires pour des raisons administratives. Par exemple, la gestion des taxes sur les produits alimentaires est complexe :

- les taux de TVA sont variables en fonction du type de produit
- des taxes spécifiques s'appliquaient aux produits contenant de l'huile ou de la farine
- des règlements particuliers s'appliquent aux alcools
- ...

D'autres informations, comme la nomenclature douanière, sont nécessaires pour effectuer les déclarations auprès des douanes européennes.

Un autre type d'information capital pour la gestion des flux d'achat et de vente sont les informations logistiques, qui définissent par exemple le nombre d'unités consommateur dans le colis, le nombre de colis sur une palette, ... Une gestion rigoureuse de ces informations est indispensable pour que les flux d'achat

ou de vente soient correctement exécutés (que les quantités commandées soient les bonnes, que les montants facturés soient corrects, ...).

2.1.2 Des produits bruts aux produits transformés

Le niveau d'exigence en termes d'information produit est variable en fonction du niveau de transformation de ce produit. Par exemple, sur des fruits et légumes frais, à peu de choses près seul le pays d'origine doit être affiché au client. Sur une barre chocolatée, ou un plat cuisiné, il sera nécessaire d'afficher :

- une liste d'ingrédients (mettant en évidence les allergènes)
- un tableau de données nutritionnelles (protéines, glucides, ...)
- une dénomination réglementaire

2.1.3 Les grands types d'information

On se focalisera dans ce paragraphe sur les informations relatives aux *produits alimentaires*.

La composition

La première grande famille de données réglementaires sont les données de composition. Elles détaillent quels sont les ingrédients qui sont mis en oeuvre dans la fabrication des produits. Évidemment, la composition a en général plus de sens pour les produits transformés que pour les produits bruts. Elle peut prendre la forme d'un texte listant la liste des ingrédients (l'étiquetage de ce texte est en général obligatoire sur les emballages des produits), ou d'un tableau.

Les ingrédients incluent également les additifs. Il s'agit de substances ajoutées à la recette pour répondre à des fonctions particulières (colorant, exhausteur de goût, émulsifiant, ...). Elles ne représentent en général qu'un pourcentage en masse négligeable dans la composition totale du produit.

Le pourcentage en masse est parfois inclus sur certains ingrédients. La réglementation l'oblige dans certains cas, par exemple quand l'ingrédient en question est mentionné dans la dénomination du produit (pour une *tarte aux framboises*, la proportion de framboise doit être mentionnée dans la composition).

Enfin, un aspect à la fois réglementaire et particulièrement important est la présence d'allergènes dans la composition. Le règlement INCO[3][4] impose de mettre en évidence les allergènes relevant d'une des 14 catégories suivantes :

1. Céréales contenant du gluten, à savoir blé, seigle, orge, avoine, épeautre, kamut ou leurs souches hybrides, et produits à base de ces céréales
2. Crustacés et produits à base de crustacés
3. Œufs et produits à base d'œufs
4. Poissons et produits à base de poissons
5. Soja et produits à base de soja

6. Lait et produits à base de lait (y compris le lactose)
7. Fruits à coque, à savoir : amandes, noisettes, noix, noix de cajou, noix de pécan, noix du Brésil, pistaches, noix de Macadamia ou du Queensland, et produits à base de ces fruits
8. Céleri et produits à base de céleri
9. Moutarde et produits à base de moutarde
10. Graines de sésame et produits à base de graines de sésame
11. Anhydride sulfureux et sulfites
12. Lupin et produits à base de lupin
13. Mollusques et produits à base de mollusques

Il peut y avoir deux niveaux de présence d'un allergène dans un produit (au-delà de la simple absence) :

intentionnellement mis en oeuvre : dans le cas où un ingrédient allergène fait volontairement partie de la recette. Ex : présence de moutarde dans un plat cuisiné.

contamination croisée : par exemple lorsque le produit fini est issu d'une chaîne de transformation qui traite un ingrédient allergène, mais que cet ingrédient ne fait pas partie de la recette. Ce cas de figure est en général mis en évidence par des mentions telles que "*Peut contenir des traces de soja*" ou bien "*Transformé dans un atelier processant également des fruits à coques et du sésame*".

Les informations nutritionnelles

Une autre grande famille d'information produit sont les informations nutritionnelles. Elles détaillent la quantité des principaux nutriments contenus dans les produits. Certains d'entre eux sont rendus obligatoires par le règlement INCO [3][4] cf. l'exemple de tableau à la TABLE 1 page 15, et d'autres sont optionnels, comme par exemple la quantité de fer, de calcium, ...

Informations nutritionnelles	Pour 100g	Pour un biscuit	% des AJR pour un biscuit
Énergie	1674 kJ 398 kcal	209 kJ 50 kcal	3 %
Protéines	3.0 g	1.0 g	3 %
Matières grasses	13.0 g	1.6 g	2 %
dont acides gras saturés	5.8 g	0.7 g	4 %
Glucides	66 g	8.2 g	3 %
dont sucres	48 g	6.1 g	7 %
Fibres alimentaires	2.5 g	0.3 g	
Protéines	3.3 g	0.4 g	1 %
Sel	0.41 g	0.05 g	1 %

TABLE 1 – Exemple de tableau de données nutritionnelles

La réglementation rend obligatoire de mentionner les informations nutritionnelles de cette table pour 100g, ou 100mL de produit (pour les boissons).

Les informations nutritionnelles peuvent également se présenter sous forme d'allégations, qui ont des définitions précises dans la réglementation. Ces allégations peuvent être : *sans sel, faible en sucres, riche en fibres, ...*

Les origines

Du fait de la complexification des opérations de transformation et de la complexification des flux d'échanges de marchandises, l'origine des produits alimentaires est une notion qui n'est pas définie avec précision dans l'absolu. Il n'y a donc pas non plus de réglementation précise sur le sujet, si ce n'est que l'information produit doit toujours être présentée de manière loyale au consommateur. On peut se donner une règle simple pour définir l'origine d'un produit alimentaire : plus il est brut, plus va compter l'origine de ses ingrédients ; plus il est transformé, plus va compter le lieu de dernière transformation.

Par exemple, sur des morceaux piécés de viande fraîche, on aura des origines multiples en fonction du pays de naissance, d'élevage ou d'abattage de la bête. Et à l'inverse, sur un steak haché cette information n'aura aucun sens dans la mesure où il est produit d'un assemblage de « minerais » pouvant provenir de multiples pays. L'industriel pourra choisir de communiquer sur le fait que la viande a été transformée en steak dans telle usine par exemple.

Les données logistiques

On appelle données logistiques essentiellement le plan de palettisation et de conditionnement du produit. Il s'agit de la définition de la « hiérarchie logistique » du produit. Cette hiérarchie se base d'abord sur la définition d'une « unité de base » qui est la plus petite unité légalement détaillable (i.e. qui porte l'ensemble des informations réglementaire pour sa commercialisation). Ces notions ont été standardisées par l'organisme international de standardisation GS1[6]. Deux exemples pour illustrer :

- pour un boîte de sachets de thé, l'unité de base est la boîte car les sachets de thé ne portent pas les informations nécessaires à leur commercialisation
- pour un paquet de barres chocolatées (comme celles qu'on peut trouver au détail en boulangerie), l'unité de base est la barre car elle porte l'ensemble des mentions réglementaires sur son emballage

La hiérarchie logistique est à la fois :

- la définition des niveaux successifs d'emballage des produits : combien d'unités de base dans un paquet, combien de paquets dans un carton, combien de cartons sur une palette, ...
- la définition du contenu de l'unité de base (ex : le nombre de sachets de thé, le nombre de doses dans une boîte d'aides culinaires, ...)

Les données logistiques concernent également les durées de vie du produit (qu'il s'agisse de Date Limite de Consommation ou Date de Durabilité Minimale) :

- la durée de vie totale à fin de production
- la durée de vie garantie à la livraison

Parfois, certaines contraintes d'approvisionnement peuvent être mentionnées :

- unités commandables (ex : on ne peut commander que des cartons complets)
- multiples de commande (ex : on ne peut commander les cartons que 10 par 10 pour des raisons de montage des palettes)
- minimum de commande (ex : il faut commander au minimum 30 cartons)

mais elles sont dépendantes d'un accord entre l'industriel et son client distributeur et ne sont donc pas à proprement parler des informations produit.

Les données administratives et financières

Les données dites administratives et financières regroupent le reste des informations de gestion pour lesquelles il existe des contraintes réglementaires. Il s'agit :

- du taux de TVA du produit
- de sa nomenclature douanière et du pays d'origine au sens de la déclaration d'échange de biens^[5]
- de toute autre taxe applicable au produit

Les labels

Afin de garantir des qualités spécifiques à certains produits, des organismes de certification ont mis en place des labels pouvant s'appliquer aux produits. En général, ils se basent sur des cahiers des charges et peuvent être assortis d'audits de certification ou de contrôle. Ils peuvent garantir des méthodes de production ou transformation, des lieux de production, des caractéristiques de leurs ingrédients, des pratiques commerciales équitables, ... Les types de labels les plus connus sont :

- les produits Biologiques
- les origines protégées (Appellation d'Origine Protégée, Indication Géographique Protégée, *viandes de France, Bleu Blanc Cor, Régions UltraPériphériques d'Europe...*)
- les pratiques commerciales équitables (Max Havelaar, ...)
- les modes de production respectueux de l'environnement (Aquaculture Stewardship Council, Marine Stewardship Council, Roundtable on Sustainable Palm Oil, Nordic Swan, ...)
- la qualité « générale » des produits (Label Rouge, ...)

Les données marketing

Certaines données marketing font également partie de l'information produit. La plus évidente est la marque commerciale du produit, qui parfois définit totalement le produit. Par exemple, on sait ce qu'est un Snickers, de la Mousline, du Nutella, ... Les produits peuvent également porter d'autres allégations marketing, non réglementaires ou labelisantes : Élu produit de l'année, Vu à la télé, Issu de notre savoir-faire centenaire, ...

2.2 Le processus

2.2.1 Le fournisseur est propriétaire des informations produit

Comme présenté à la section 1.1 page 6, le Groupe Pomona n'a pas d'activité de fabrication ou de transformation de marchandises. À ce titre, l'ensemble des données produits ne peuvent être déterminées que par les fournisseurs de ces produits. L'ensemble des entités du groupe s'appuient donc sur les données transmises par les industriels ou producteurs de marchandises.

Il peut arriver que certains produits soient achetés par Pomona à d'autres négociants non-producteurs. Dans ce cas, de la même manière que le groupe Pomona a la responsabilité de collecter puis transmettre les informations produit à ses clients, ces fournisseurs négociants doivent eux-même aller chercher l'information produit et la transmettre à Pomona.

Dans tous les cas, c'est le fournisseur qui est responsable de produire et de transmettre l'information produit aux entités du groupe Pomona.

2.2.2 La notion de produit et d'article

Un mot sur la modélisation des données adoptée est nécessaire pour comprendre les grandes lignes du processus. Comme il a été vu à la section 1.2.3 page 8, certains produits sont susceptibles d'être commercialisés par plusieurs branches du groupe.

De plus, du fait que les systèmes d'information ne sont pas identiques entre les branches, certaines contraintes imposent parfois des différences de modélisation, des duplications volontaires de codes pour répondre à ces contraintes. Une illustration de ce point pour clarifier : la facturation client pour une canette de soda peut se faire au litre (permet de comparer les prix entre les différents conditionnements et les différentes marques) ou à la cannette (permet de se faire une idée du coût portion d'un produit). Or, la possibilité de pouvoir facturer un même article dans plusieurs unités différentes n'est pas une fonctionnalité offerte par tous les systèmes d'information. En particulier, ÉpiSaveurs peut gérer dans ce cas un unique article et le facturer dans l'unité de son choix en fonction des demandes des clients. Mais Délice et Création (qui possède un système d'information différent) doit dupliquer cet article car une seule unité de facturation est possible pour un article donné.

Enfin, au-delà des contraintes liées au SI, certaines pratiques imposent de laisser aux branches une indépendance forte dans la gestion de leurs référentiels articles. Il faut savoir que commercialiser sous un même code article des produits qui sont similaires permet d'obtenir des gains de productivité à plusieurs niveaux :

- on économise des emplacements en entrepôt (un emplacement ne peut contenir qu'un article)
- on gagne du temps administratif dans la gestion des prix : le foisonnement d'articles impose de gérer plus de prix client

— ...

La contrepartie à adopter cette pratique est qu'il n'est alors plus possible de différencier ces produits similaires, par exemple pour leur appliquer des prix de vente distincts, ou bien offrir la possibilité à un vendeur de garantir au client la livraison d'un produit plutôt que l'autre. Néanmoins, en fonction de la clientèle adressée, certains produits pourront être considérés comme similaires, alors que pour d'autres ils ne seront pas interchangeables. Cet exemple est détaillé dans la FIGURE 4 page 19.

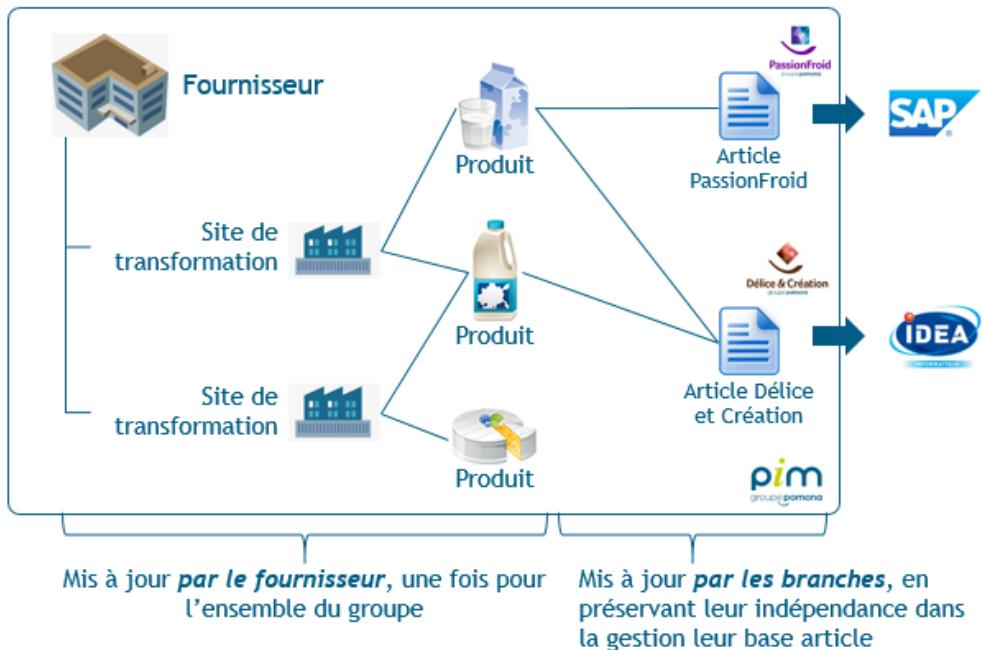
On différencie donc les deux notions suivantes :

les produits : ils représentent une marchandise physique produite par un fournisseur. Ce sont les produits qui portent les *informations produit* décrites à la section 2.1.3 page 14. Un produit ne peut appartenir qu'à un seul fournisseur. Le référentiel produit est unique pour l'ensemble du groupe.

les articles : ce sont les objets qui sont gérés par les branches dans leurs systèmes de gestion respectifs.

Leurs attributs sont très liés au système d'information qui les porte. Chaque branche gère de manière autonome son référentiel article, incluant les liens qui sont faits entre produits et articles.

Cette modélisation permet de répondre à l'ensemble des contraintes présentées dans ce paragraphe.



Dans cet exemple fictif, le type de conditionnement du lait n'a aucun impact sur les clients de la branche Délice et Crédit. Elle commercialise donc sous un même code article deux produits distincts.

Pour des raisons de contrainte de conservation, la branche PassionFroid a quant à elle choisi de ne commercialiser qu'un seul produit - en brique. Elle pourrait choisir d'ouvrir un nouveau code article pour le lait en bouteille (non représenté sur le schéma ci-dessus).

FIGURE 4 – La distinction entre produit et article

2.2.3 Les acteurs

L'acheteur

L'acheteur peut être en succursale, ou à la centrale d'achat de la branche (cf. 1.2.3 page 8 et 1.2.4 page 11). Il est responsable de l'assortiment, i.e. des articles qui sont commercialisé par sa branche, ou sa succursale. Il est à l'origine de la création d'un nouveau produit, dans la mesure où c'est lui qui va décider de le référencer ou pas. La gestion de l'information produit ne représente qu'une petite partie de l'ensemble de ses responsabilités (qui peuvent inclure négociations fournisseurs, gestion du sourcing, gestion des approvisionnements, prévisions d'évolution des prix, ...).

Le fournisseur

Il a été vu que le Groupe Pomona - dans sa qualité de distributeur - n'est pas en mesure de déterminer seul les informations produit sur les marchandises qu'il commercialise. À ce titre, les équipes des fournisseurs sont en charge de transmettre les information produit à Pomona. En général, ce sont deux types de profil qui effectuent cette tâche :

- les commerciaux (aux sens large, incluant les assistants commerciaux)
- les ingénieurs qualité

Le gestionnaire de référentiel - SEGER

Le gestionnaire de référentiel travail au SService de GEstion des Référentiels (SEGER), qui sont des équipes positionnées au niveau des branches (cf. 1.2.3 page 8). Ils sont responsables de la qualité des données dans les référentiels métier (articles, fournisseurs, clients, ...). La gestion des données de base dans les référentiels est leur mision principale.

L'ingénieur qualité

L'ingénieur qualité travaille à la DQSE ou en succursale (cf. 1.2.1 page 7 et 1.2.4 page 11). Dans le processus de gestion de l'information produit, il est en charge de contrôler la qualité des données, mais également leur conformité réglementaire (ex : on ne peut qualifier un produit de « faible en sel » que s'il comporte moins de n grammes de sel pour 100 grammes de produit.) La gestion de l'information produit ne représente qu'une partie des responsabilités de l'ingénieur qualité.

2.2.4 Les contrôles

Comme cela a été vu dans la section 2.1.1 page 13, il est nécessaire que les différentes entités du groupe soient en possession d'une information produit fiable. Or, avoir des données de qualité nécessite des efforts de la part des métiers, en particulier lorsque le processus n'est pas entièrement porté en interne dans la société.

À ce titre, plusieurs étapes de contrôle ont été définies dans le processus de gestion du référentiel de données produit et article :

lorsque le fournisseur a saisi les données produit : la personne à l'origine de la demande de référencement (en général, un acheteur) doit contrôler la cohérence des données produit

lorsque le demandeur a demandé la création d'un article : le gestionnaire de référentiel valide à nouveau la cohérence des informations produit

après la création article, de manière asynchrone : le service qualité contrôle par échantillonnage les données d'une partie des produits et articles qui ont été modifiés pendant une période.

Le retour d'expérience montre que ces contrôles, loin d'être redondants, sont nécessaires pour avoir une qualité de données acceptable. Ces processus de contrôle sont décrits à la FIGURE 5 page 22.

Les contrôles effectués à chacune des étapes sont les suivants :

contrôle de la complétude des données : vérification que les données transmises comportent l'ensemble des données attendues

contrôle de cohérence entre les données : vérification que les informations transmises sont cohérentes entre elles (ex : un allergène présent dans la liste d'ingrédients du produit a bien été signalé comme allergène par ailleurs)

contrôle de la cohérence avec les pièces jointes : en plus de données structurées, les fournisseurs transmettent également des fichiers portant des informations produit (ex : l'étiquette produit ou le visuel de l'emballage). Ces pièces jointes sont décrites à la section 4.3 page 43. La personne en charge du contrôle vérifie que les données transmises sont cohérentes avec ces documents.

2.3 Les outils informatiques associés

Comme vu dans la description des branches du groupe (voir section 1.2.3 page 8), les outils informatiques ne sont pas tous les mêmes sur l'ensemble des branches. Ainsi, les outils utilisés pour la gestion de l'information produit ne sont pas les mêmes.

2.3.1 Les branches faiblement outillées

Les branches étrangères (Pomona Suisse, Pomona Iberia), spécialistes (Délice et Création, Saveurs d'Antoine) et la branche TerreAzur sont aujourd'hui faiblement outillées. Cela signifie que l'information produit est en général stockée uniquement sous la forme de fichiers (essentiellement les pièces jointes, décrites à la section 4.3 page 43). L'ensemble des échanges avec les fournisseurs se font par mail, et les articles sont créés directement dans les systèmes de gestion par les gestionnaires de référentiel. Les liens entre les articles et les informations produit ne sont pas matérialisés dans les systèmes informatiques.

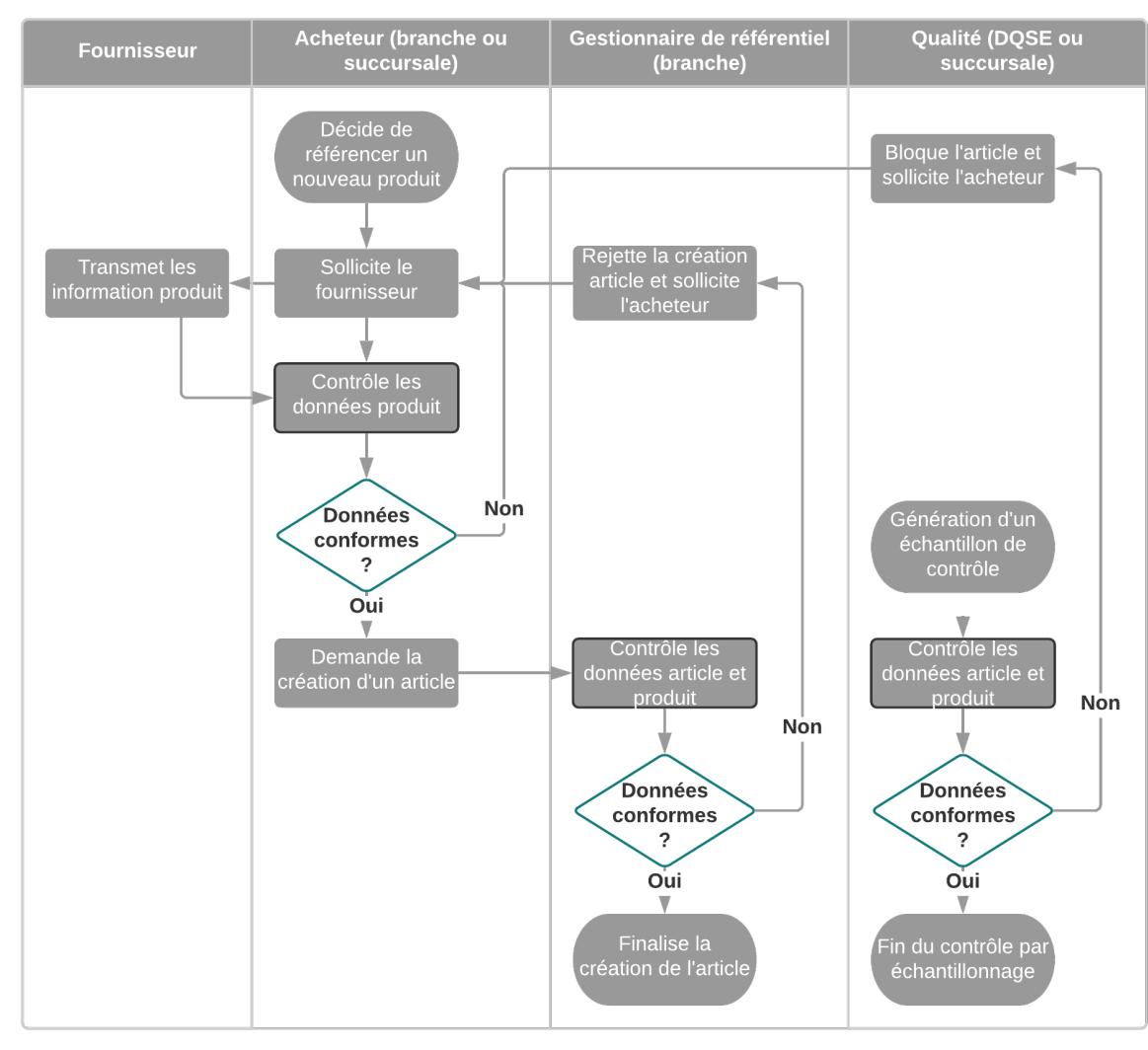


FIGURE 5 – Le processus de création article

2.3.2 Le GIP

Le GIP (Gestion de l'Information Produit) est utilisé sur la branche PassionFroid. C'est un système de gestion de l'information produit qui est maintenant obsolescent et en cours de remplacement. Il a toutefois le mérite de permettre le stockage dans une application des données et des pièces jointes relatives aux produits, avec la possibilité d'accéder aux informations produit à partir des identifiants des articles. C'est ce système qui a permis de pouvoir alimenter les sites de e-commerce PassionFroid et ÉpiSaveurs avec les informations produit. Il est toutefois ancien, et ne propose pas de fonctionnalité d'export en masse fiable. Il s'agit d'une application qui n'est pas ouverte aux utilisateurs externes au groupe, et les échanges avec les fournisseurs passent donc par des échanges de mails.

2.3.3 Le PIM

Le PIM (Product Information Management) est un système de gestion de l'information produit qui a été mis en production en mai 2019, pour la branche ÉpiSaveurs. Il a vocation à être déployé sur l'ensemble des branches du groupe. En terme de responsabilités, il vise à gérer la totalité des informations produit, mais également d'être maître sur les référentiels articles. Ce sera le système de référence pour tous les autres systèmes consommant de l'information produit, ou des données article.

Description générale de l'outil

C'est un système qui porte l'ensemble du processus de gestion de l'information produit, tel que décrit à la FIGURE 5 page 22. Il est accessible aux fournisseurs du groupe, qui viennent directement mettre à disposition les données et les pièces jointes. Les fonctionnalités caractéristiques de ce système par rapport à d'autres systèmes informatiques sont :

- la gestion de workflows (processus), qui permet d'orchestrer l'activité de l'ensemble des acteurs du système
- la possibilité de paramétriser un modèle de données relativement complet, avec des centaines de métadonnées sur les différents objets
- la présentation de l'interface utilisateur via le navigateur, qui permet de s'adresser simplement à des acteurs hors du groupe (pas de client lourd à installer)
- la gestion performante de pièces jointes (documents informatiques) en grandes quantité
- Une gestion de versions des objets robuste, permettant d'auditer l'historique ou de restaurer des données dans un état précédent

Cet outil porte entre autres les fonctionnalités de contrôle des informations, comme illustré à la FIGURE 6 page 24. L'intégration du PIM au sein des systèmes informatiques du Groupe Pomona est décrite dans la FIGURE 7 page 24.

The screenshot shows the PIM software interface for product control. On the left, there's a sidebar with various icons. The main area has a header with the PIM logo and navigation links. Below the header, there's a section titled "Section en cours de modification" (Modification in progress) with a blue background. It contains a table with nutritional values for 100g and ingredients. To the right, there's a green panel titled "NUTRITION : OK" and another panel titled "DÉCLARATIF DES DONNÉES NUTRITIONNELLES". At the bottom, there are tabs for "Fiche technique fournisseur" and "Données logistiques".

Cette capture d'écran montre l'outil de contrôle des données produit. Sur la partie gauche, le contenu des pièces jointes est affiché (visuel de l'emballage en haut, fiche technique en bas), sur la droite les données qui ont été transmises par le fournisseur.

FIGURE 6 – Une capture d'écran du PIM

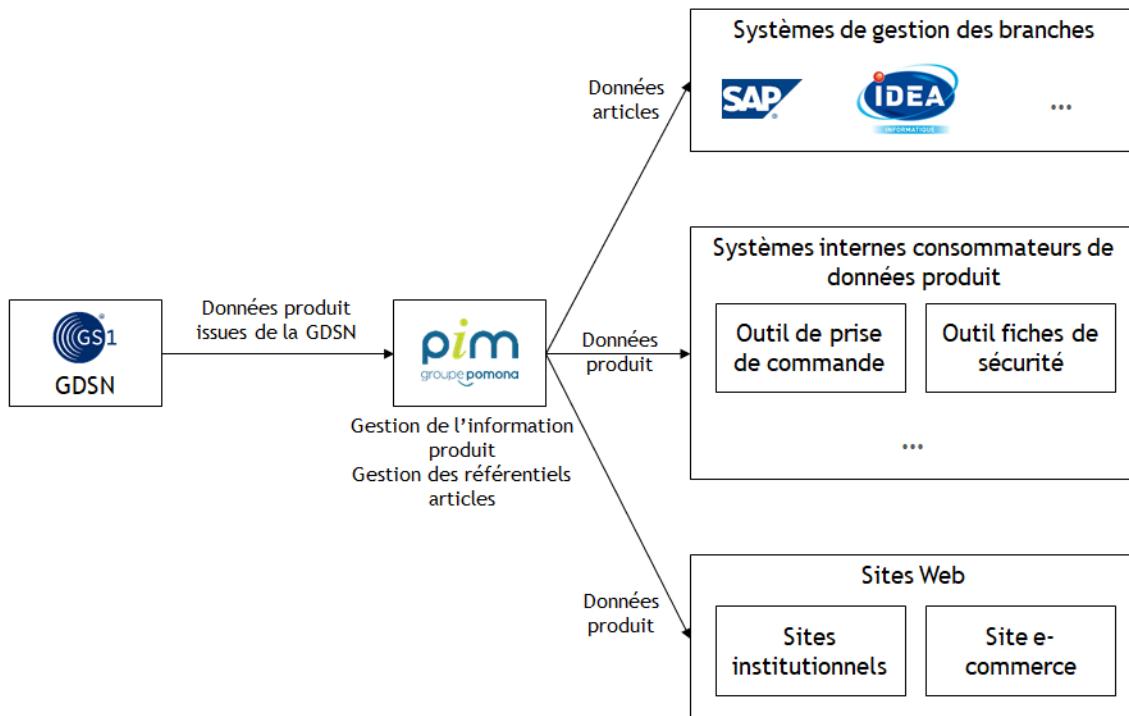


FIGURE 7 – L'intégration du PIM au sein des systèmes du Groupe Pomona

Le socle technologique (et un mot de vocabulaire)

Ce logiciel a été construit sur la base de l'outil de gestion de contenu généraliste Nuxeo. Il s'agit d'un outil de GED (Gestion Électronique de Documents), et tous les objets (produits, fournisseurs, articles, ...) qu'il stocke sont nommés « documents ». Dans le cadre du PIM, ce qui est habituellement appelé « document » au sens d'un fichier informatique (document pdf, image png, vidéo, ...) est appelé « pièce jointe ». On utilisera ce vocabulaire dans le présent rapport.

Le logiciel Nuxeo est adapté par une société - Keendoo - qui pré-paramètre la solution généraliste Nuxeo pour en faire un outil spécialisé dans la gestion d'information pour les produits alimentaires.

La dernière « couche » de développement a été réalisé de manière spécifique par les équipes de développement du Groupe Pomona.

La base de données sous-jacente à l'application est une base NoSQL MongoDB.

La GDSN

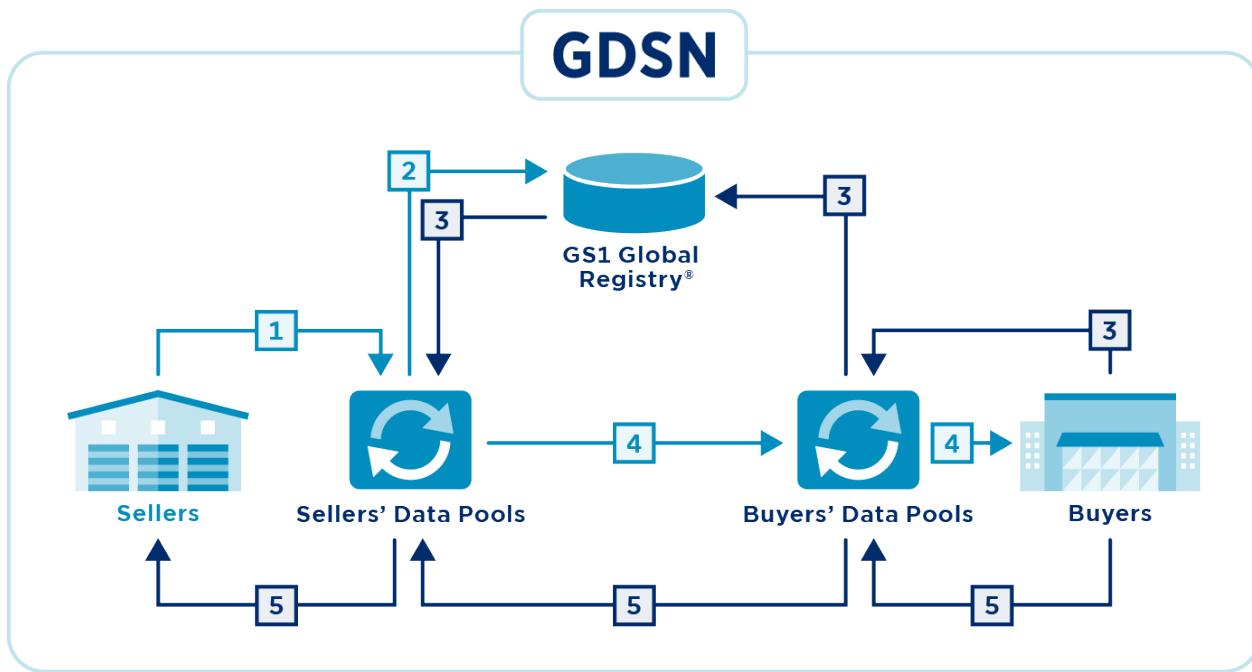
La GDSN (Global Data Synchronization Network) est un réseau d'échange de données produit entre industriels, distributeurs, restaurateurs, ... Ce réseau est exploité par des opérateurs privés, mais le format et la chorégraphie des échanges ont été standardisés par l'organisme de standardisation GS1. Son schéma de principe est décrit à la FIGURE 8 page 26. Sans rentrer dans le détail, au sein du Groupe Pomona l'utilisation qui en est faite est de récupérer les informations depuis ce réseau d'échange, afin de préalimenter les données produit pour les fournisseurs. Cette fonctionnalité permet de faire gagner du temps aux fournisseurs pour leur éviter une partie de ressaisie, mais également de limiter les erreurs. Toutefois, cette fonctionnalité ne permet pas à elle seule de garantir une parfaite qualité de données. Il s'agit uniquement d'un « tuyau », si les données en entrée ne sont pas correctes, elles ne seront pas correctes en sortie. Pour aller plus loin dans la compréhension de ce réseau, il est possible de consulter les ressources mises en ligne par GS1 [7][8].

L'identification des objets dans le PIM

Un dernier point à connaître à propos du PIM, est la manière d'identifier l'ensemble des objets en son sein. Chaque objet géré (ainsi que toute version archivée) porte un identifiant unique, nommé *uid* qui est totalement univoque. Pour la suite, on se basera sur ces uid pour faire référence à des produits stockés dans le PIM. Une illustration est présentée à la FIGURE 9 page 26.

Les API

Un des aspects intéressants du PIM pour l'exploitation en masse des données produit, est qu'il expose des API permettant d'aller requérir l'ensemble de son contenu. Cela concerne à la fois les données dites structurées, mais également les pièces jointes. Cela rend les données produit de la branche ÉpiSaveurs bien plus simplement accessibles que celles des autres branches.



1. Loading of company data
 2. Registering of company data
 3. Subscription to seller's data pool
 4. Publishing of company data
 5. Confirmation receipt of company data

FIGURE 8 – Schéma de principe de la GDSN

La captura de pantalla muestra la interfaz de administración de productos PIM. En la barra superior, se observa la URL <https://produits.groupe-pomona.fr/nuxeo/ui/#!/doc/a2b33d4b-5c39-404c-aa65-a41e87bac7e9>. La parte central del sitio muestra un producto de PiM's Orange en paquet de 150 g. Los datos visibles incluyen el nombre del producto (**PIM'S ORANGE EN PAQUET 150 G LU**), su UID (**PIMP-0000005795**), una imagen del producto y un botón para la versión 1.1. A la izquierda, hay un menú vertical con iconos para navegar entre diferentes secciones como Espace Fournisseurs, Mondelez France, etc. Abajo, se detallan las secciones de administración: **ADMINISTRATION POMONA**, **FICHES DE CONTRÔLES** y **IDENTIFICATION PRODUIT**.

L'uid du produit affiché est mis en évidence dans l'url

FIGURE 9 – L'uid d'un produit

La reprise de données initiale et la migration

Un point à avoir également en tête est le mode d'alimentation initial du PIM pour ÉpiSaveurs et le processus dit d'enrichissement qui ont été décidés. Les données ont été transférées du système de gestion historique (le GIP, cf. section 2.3.2 page 23) vers le PIM, sans transformation métier. Cela signifie qu'il n'y a eu ni correction, ni enrichissement de données lors du chargement initial. Il a été décidé de lancer ces travaux après le démarrage, directement dans le système PIM, en utilisant les fonctionnalités de sollicitation des fournisseurs qu'il propose. La manière de procéder est la suivante :

- les produits qui ont été créés lors de la reprise de données initiale sont envoyés aux fournisseurs pour qu'ils corrigent et complètent les informations produit
- le processus est ensuite le même que pour un nouveau référencement (en particulier, les contrôles et les renvois au fournisseur, tels que présenté à la FIGURE 5 page 22)
- lorsque la validation finale des gestionnaires de référentiels est effectuée, le produit est considéré comme « migré » et ses données sont réputées correctes

Les données du PIM étaient donc incomplètes et vraisemblablement incorrectes au démarrage, et en juin 2020, ces travaux sont toujours en cours. Il existe donc des produits dans le système, pour lesquels la revue et la correction des données n'a pas encore eu lieu.

Deuxième partie

LES DONNÉES

Chapitre 3 LE PÉRIMÈTRE PRODUIT

3.1 Les produits non-alimentaires

Un petit aparté est nécessaire concernant les produits non-alimentaires. Si l'essentiel des produits commercialisés par les branches du groupe sont des produits alimentaires, comme évoqué précédemment une partie de l'activité commerciale se fait tout de même autour de produits non-alimentaires. Ces produits restent malgré tout destinés exclusivement aux professionnels des métiers de bouche, et il s'agit de consommables (par opposition à des articles d'électroménager, de la vaisselle non-jetable, ...).

On distingue en général deux catégories de produits non-alimentaires :

- les produits dits « d'hygiène »
- les produits dits « de chimie »

Les produits de chimie regroupent les produits qui doivent faire l'objet d'une fiche de données de sécurité au sens du règlement Européen No 1907/2006 dit « REACH » (Registration, Evaluation, Authorisation and Restriction of Chemicals) [2].

On appelle produits d'hygiène tous les autres produits non-alimentaires. L'appellation « d'hygiène » est donc réductrice, dans la mesure où cette large famille regroupe les consommables de nettoyage (éponges, papiers absorbants, ...) mais également tout type d'autres consommables (serviettes de tables, gobelets en plastiques, pics à brochettes, boîtes de produits à emporter, ...).

La commercialisation de produits non-alimentaires existe au sein du groupe, mais on se focalisera pour la suite sur les produits alimentaires qui reste le cœur de métier.

3.2 Accessibilité de la donnée en fonction des branches

Comme vu à la section 2.3 page 21, les systèmes d'information associés à la gestion de l'information produit offrent des niveaux d'accès hétérogènes à la donnée produit. Le récapitulatif par branche est le suivant :

ÉpiSaveurs : on peut simplement accéder à l'ensemble des données produit, structurées, non structurées (i.e. textes longs) et pièces jointes

PassionFroid : on a uniquement la possibilité d'exporter manuellement les données structurées articles depuis le système de gestion SAP. Elles permettent de produire quelques analyses quantitatives. Il est difficile de faire des exports en masse de l'outil de gestion de l'information produit GIP (cf. section 2.3.2 page 23).

TerreAzur : idem PassionFroid, si ce n'est qu'en plus le système GIP n'est pas utilisé au sein de cette branche.

Délice et Création : le système d'information ne permet pas d'exporter les données et donc de produire des indicateurs détaillés. On peut toutefois avoir des informations quantitatives de la part des opérationnels.

Saveurs d'Antoine : idem Délice et Création

Pomona Suisse : la branche est en cours de structuration, et les référentiels articles ne sont pas partagés entre les succursales. Il n'est pas possible d'obtenir d'information quantitative sur ces données.

Pomona Iberia : idem Pomona Suisse

Pour les analyses quantitatives, on pourra se baser sur des extractions uniquement pour les branches RHD (ÉpiSaveurs, PassionFroid, TerreAzur). L'ensemble des analyses portant sur les branches RHD sont produites sur la base d'extractions de leur système de gestion SAP.

3.3 Analyses quantitatives

Les graphes de cette section ont été produits via le code présenté en annexe, au chapitre C.1 page 97. Les données pour les branches spécialistes (Délice et Création et Saveurs d'Antoine) sont issues d'informations fournies par le métier, hors système. Dans l'ensemble de cette section, on raisonnera à la maille *article* (cf. la définition article vs. produits, présentée à la section 2.2.2 page 18).

3.3.1 Comparatifs entre les branches

En termes de volumétrie article (cf. FIGURE 10 page 30, c'est TerreAzur qui possède le référentiel le plus étendu (environ 62 000 articles de marchandises actifs). Cela s'explique par le fait que cette branche commercialise essentiellement des produits bruts, non-préemballés (ex : des cagettes de fruits ou de légumes). Or, ces produits ne sont pas clairement identifiés, par exemple par un GTIN. Au démarrage de SAP pour

cette branche, afin de limiter la charge sur les gestionnaires de référentiels, le parti a été pris de créer en avance de phase l'ensemble des articles susceptibles d'être commercialisés. Cela s'est traduit par la création d'un grand nombre d'articles, du fait de l'application « brutale » de la combinatoire des différents critères pouvant définir un produit. Un exemple (fictif) serait, sur les pommes :

- 8 variétés possibles (Gala, Golden, ...)
- 4 calibres possibles
- 6 conditionnements possibles (plateau 6kg, plateau 4,5kg, ...)
- 2 catégories (I, II)
- 8 origines (France, Espagne, ...)

ce qui donne un total de 3072 articles uniquement sur cette gamme de produits.

Viennent ensuite PassionFroid, et ÉpiSaveurs, qui sont les autres « grosses » branches historiques du Groupe.

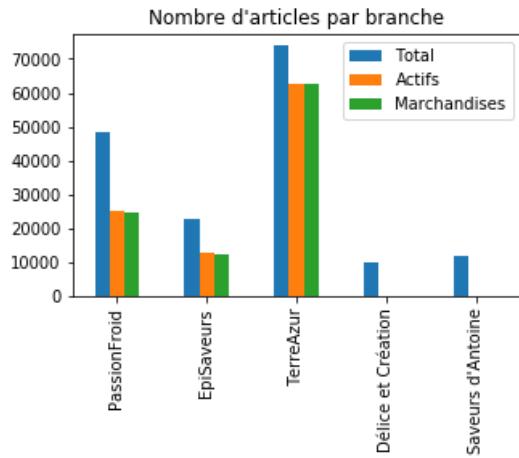


FIGURE 10 – Volumétrie article par branche

Une analyse du recouvrement des référentiels montre que dans l'ensemble, les branches ne travaillent pas les mêmes articles (cf. FIGURE 11 page 31). PassionFroid commercialise certains produits des branches ÉpiSaveurs et TerreAzur, mais cela s'explique par une petite entité luxembourgeoise qui travaille des produits de tout type de stockage. Une réserve toutefois par rapport à cette analyse de recouvrement produit : elle sous-estime vraisemblablement lesdits recouvrements, dans la mesure où la présence de doublons n'est pas prise en compte.

3.3.2 Les grands types de produits

Comme montré à la FIGURE 12 page 32, on voit bien (en plus du fait que les articles étaient peu partagés entre les branches) que :

- les deux types d'articles (négocie et presté) sont utilisés par les 3 branches
- c'est tout de même PassionFroid qui fait l'utilisation majoritaire d'articles prestés

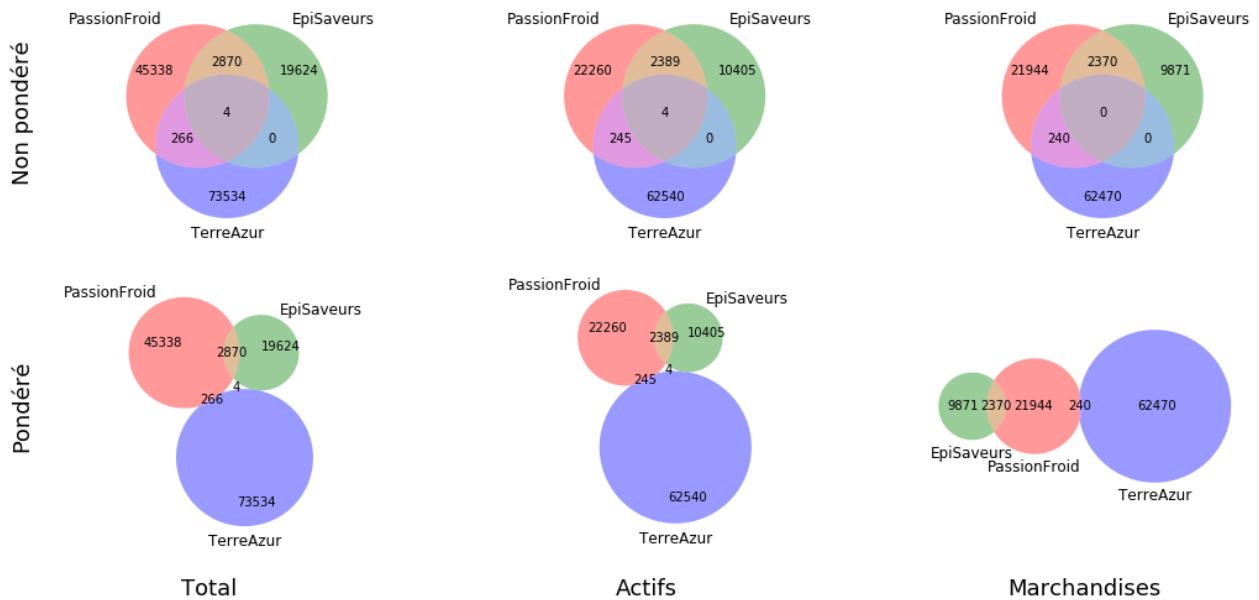


FIGURE 11 – Recouvrements entre branches RHD

- les branches *ne partagent pas* l'utilisation des autres « catégories » (groupe de marchandises, conditions de stockage, hiérarchie produit, ...)

Les indicateurs sont également récapitulés dans la TABLE 3 page 33.

Possibilité d'extension : - ajouter la vision produit (via les FIA) - faire une sorte d'heatmap qui montre la forte correlation entre les différentes variables catégorielles (et les définitions associées. Ex : ZELAB + SA = Saurisserie, etc...)

Chapitre 4

LES DONNÉES UTILISABLES, ISSUES DU PIM

Comme vu au chapitre 3 page 28, les données produit ne sont simplement accessibles que pour la branche ÉpiSaveurs. On se focalisera donc sur cette branche pour la suite de cette étude, ainsi que sur les produits alimentaires (en excluant donc les produits d'hygiène et de chimie, cf. section 3.1 page 28). Contrairement au chapitre précédent, ici on travaillera à la maille *produit* (cf. la distinction produit vs. article section 2.2.2 page 18). L'ensemble des tableaux et graphes de ce chapitre ont été produit via le notebook "Analyse des données du

Répartition des articles selon les features catégorielles

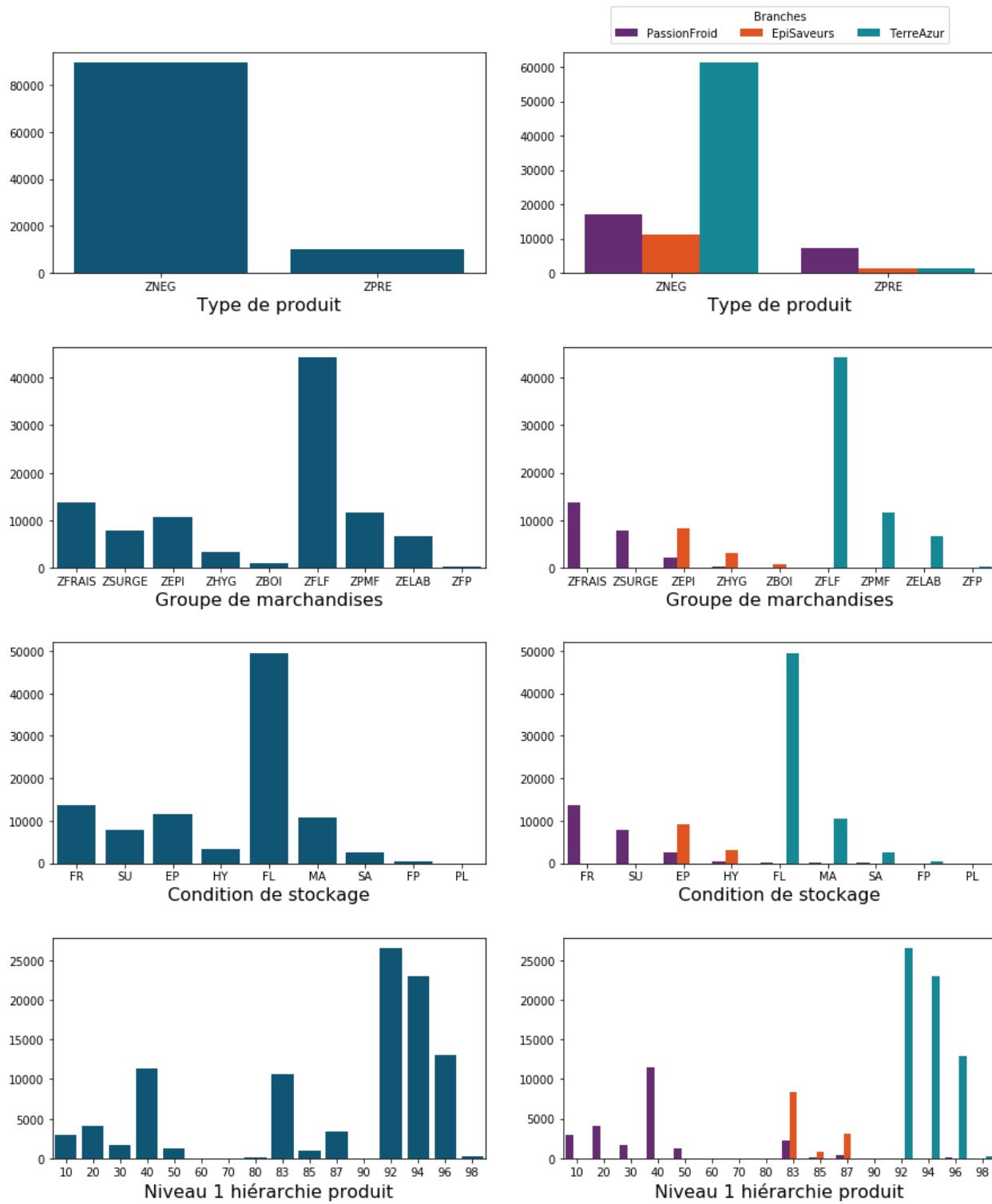


FIGURE 12 – Répartition des articles en fonction des variable catégorielles

Branche Type de produit	PassionFroid	EpiSaveurs	TerreAzur	Total
ZNEG - Article de négoce	17166	11048	61273	89487
ZPRE - Article de prestation	7388	1193	1437	10018
Branche Groupe de marchandises	PassionFroid	EpiSaveurs	TerreAzur	Total
ZSURGE - Surgelés	7756	-	-	7756
ZFRAIS - Frais	13785	6	4	13795
ZEPI - Epicerie	2298	8305	-	10603
ZBOI - Boissons	126	826	-	952
ZHYG - Hygiène	350	3078	-	3428
ZFLF - Fruits et Légumes	4	-	44133	44137
ZPMF - Produits de la mer	142	-	11594	11736
ZELAB - Produits élaborés	91	-	6644	6735
ZFP - Fleurs et plantes	-	-	297	297
ZAUTRE - Autres	2	26	38	66
Branche Condition de stockage	PassionFroid	EpiSaveurs	TerreAzur	Total
SU - Surgelés	7758	-	-	7758
FR - Frais	13781	6	3	13790
EP - Epicerie	2430	9155	-	11585
HY - Hygiène	344	3080	-	3424
FL - Fruits et légumes	78	-	49508	49586
MA - Marée	126	-	10501	10627
FP - Fleurs et plantes	-	-	286	286
SA - Saurisserie	34	-	2408	2442
PL - Publicié	2	-	1	3
Branche Niveau 1 hiérarchie produit	PassionFroid	EpiSaveurs	TerreAzur	Total
10 - Beurre, oeufs, fromage	3010	6	1	3017
20 - Elaborés	4150	2	6	4158
30 - Garnitures et fruits	1701	-	-	1701
40 - Produits carnés	11413	-	-	11413
50 - Produits de la mer	1214	-	2	1216
60 - Consommables	1	-	-	1
70 - Emballage	-	1	-	1
80 - Publicité sur le lieu de vente	34	25	37	96
83 - Epicerie	2306	8296	-	10602
85 - Liquides	135	836	-	971
87 - Hygiène et entretien	348	3075	-	3423
90 - Services	10	-	-	10
92 - Fruits	35	-	26543	26578
94 - Légumes	37	-	22929	22966
96 - Produits de la mer Frais	160	-	12891	13051
98 - Fleurs - plantes	-	-	301	301

TABLE 3 – Utilisation des variables catégorielles article au sein des branches RHD

PIM", présenté en annexe C.1 page 107.

4.1 Données structurées

4.1.1 Description des données structurées

Les données dites structurées sont l'ensemble des données qui peuvent prendre leurs valeurs dans un domaine restreint. Par exemple, ce sont les données booléennes, les choix issus de listes déroulantes, les valeurs numériques... Les principales données structurées pour les produits alimentaires dans le PIM sont :

le code du produit : calculé par le système

le fournisseur : référence croisée vers le code du fournisseur

le type de produit : épicerie, boisson alcoolisée, hygiène, chimie, boisson non-alcoolisée

le GTIN du produit : identifiant numérique unique, utilisé entre autres pour l'étiquetage sous forme de code à barres [9]

le type d'unité de base : paquet, boîte, sachet, rouleau, bouteille, pot, ...

les poids : brut, net, net égoutté (pour les conserves)

le volume : pour les produits liquides

les durées de vie : le type (Date Limite de Consommation ou Date de Durabilité Minimale) et la durée (totale à fin de production, garantie à livraison)

les modes de conservation avant/après ouverture : à température ambiante, au réfrigérateur puis à consommer sous 2 jours, ...

les labels : le(s) label(s) s'appliquant au produit (cf. section 2.1.3 page 17)

les régimes particuliers : Halal, Casher, Sans porc, Végétarien, Végétalien, ...

les caractéristiques spéciales : sans OGM, non traité par ionisation

la présence d'allergènes : le niveau de présence de chacun des 14 allergènes réglementés (cf. section 2.1.3 page 14) : absence, présence ou traces

les matières grasses utilisées : palme, beurre, coco, tournesol, palmiste, ...

les additifs présents : les codes Exxx et les fonctions des additifs mis en oeuvre [1][11]

les données nutritionnelles obligatoires : pour 100g ou 100mL, valeur énergétique (en kJ et kcal), matières grasses, dont acides gras saturés, Glucides, dont sucres simples, Fibres, Protéines, simplement

les données nutritionnelles facultatives : vitamines, minéraux, omégas, ...

les allégations nutritionnelles : riche en, faible en, sans,... associé à un nutriment défini dans les 2 points précédents

le nutriscore : note allant de A à E, définie dans la loi Santé de janvier 2016

le taux de TVA : un des quatre taux définis dans la réglementation française

le code nomenclature douanière : code identifiant les marchandises défini par les douanes pour la Déclaration d'Échange de Biens [5]

le pays d'origine pour la DEB : le pays d'origine à déclarer dans la Déclaration d'Échange de Biens [5]

les informations logistiques : il s'agit du plan de conditionnement et de palettisation du produit. Elles regroupent les différents niveaux et les quantités pour passer de l'un à l'autre (ex : 3 boites dans un cartons, 64 cartons dans une palette), les poids et dimensions de ces niveaux logistiques, leurs GTIN,

...

4.1.2 Analyse de ces données structurées

Le statut des produits

Comme cela a été présenté à la FIGURE 5 page 22, le processus de gestion de l'information produit fait intervenir de multiples acteurs et peut donc être assez long. Un produit qui n'est pas arrivé au bout du processus et fait l'objet de contrôles et validations successives n'est pas réputé comme portant des informations correctes. Dans le PIM, la « localisation » du produit dans le processus est portée par son statut (c'est une simple donnée catégorielle). Seuls les produits au statut « Validé » sont considérés comme portant des données valides. La répartition des produits pour chacun des statuts est présenté à la FIGURE 13 page 35.

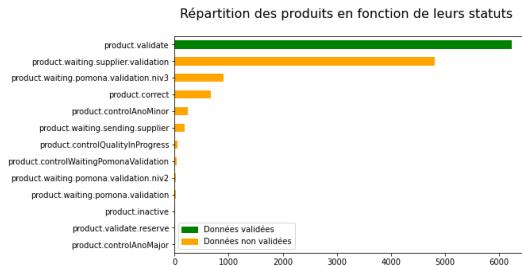


FIGURE 13 – Répartition des produits par statut

state	
product.validate	6240
product.waiting.supplier.validation	4812
product.waiting.pomona.validation.niv3	906
product.correct	671
product.controlAnoMinor	245
product.waiting.sending.supplier	192
product.controlQualityInProgress	48
product.controlWaitingPomonaValidation	32
product.waiting.pomona.validation.niv2	29
product.waiting.pomona.validation	21
product.inactive	12
product.validate.reserve	3
product.controlAnoMajor	1

TABLE 4 – Répartition des produits par statut

De plus, comme cela a été présenté dans la section migration 2.3.3 page 27, certains produits ont fait l'objet d'un contrôle et d'une correction après le démarrage, mais d'autres non. Il est nécessaire de prendre également cet aspect en compte si l'on souhaite avoir la vision des produits qui sont censés avoir des données correctes. La détermination du statut des produits vis-à-vis du processus de migration se fait au travers de facettes :

- les produits qui ont été créés lors de la reprise de données initiale portent une facette "beginningMigration"
- les produits créés dans le PIM après le démarrage (cas des nouveaux référencements) ne portent pas cette facette

- les produits issus de la migration (portant la facette "beginningMigration") sont envoyés aux fournisseurs pour qu'ils complètent les données
- une fois que le processus de validation de l'information produit est terminé (validation des gestionnaires de référentiels), une facette "endMigration" est apposée sur le produit

La répartition des produits fonction de leurs statuts de migration est présenté à la TABLE 5 page 36. Les produits portant des données correctes, du point de vue de la migration, sont :

- ceux qui ont été créés après le démarrage (donc qui ont été créés avec le processus et les règles de gestion cible)
- ceux qui ont été créés à la reprise, mais qui sont identifiés comme ayant terminé le processus de migration (portant la facette de fin)

	Facette fin de migration : Non	Facette fin de migration : Oui
Créé après le démarrage	1657	0
Créé au démarrage	7212	4343

TABLE 5 – Répartition des produits par statut de migration

Si on combine le statut des produits avec leurs statut de migration, on peut calculer un statut « Données valides » qui doit permettre d'identifier des produits avec des données de qualité. C'est ce statut qui sera utilisé dans les paragraphes suivants, lorsqu'on fera des analyses relatives à la qualité des données. On nommera ces statuts « En qualité » et « Hors qualité » pour la suite de ce document. La volumétrie des produits en fonction de chacun de ces statuts est présentée à la TABLE 6 page 36.

	Répartition produit par qualité
Hors qualité	8622
En qualité	4590

TABLE 6 – Répartition des produits par « qualité des données »

Données d'identification

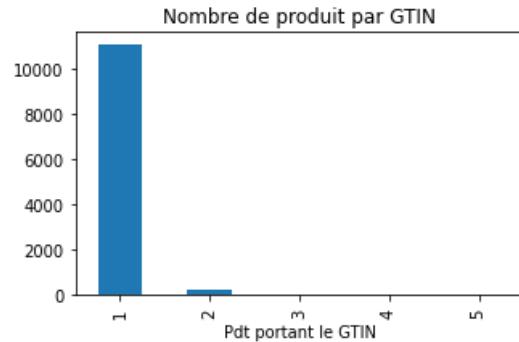
Un exemple et une description statistique de ces données est présentées aux TABLE 8 page 39 et TABLE 9 page 39. On peut en tirer les conclusions suivantes :

- Les codes produits sont bien des identifiants uniques des produits : on a autant de valeurs uniques que de valeurs renseignées
- Une proportion importante de produits (91%) porte un GTIN (cf. [9]), ce qui reflète une « maturité » des filières produits avec lesquelles travaille ÉpiSaveurs.
- Ces GTIN sont également censés être des identifiants uniques des produits. Néanmoins, comme présenté à la FIGURE 14 page 37, il peut arriver que le même GTIN soit présent sur plusieurs produits. Quelques

illustrations et explications sont présentées en annexe, dans le notebook d'analyse des données.

- La distribution numérique des produits par fournisseur est importante. Il existe plus de 600 fournisseurs pour les 13000 produits, et l'essentiel des fournisseurs n'ont qu'un nombre très limité de produits. On peut avoir une vision à la FIGURE 15 page 37 ou à la FIGURE 16 page 38.

TODO : illustrer ici les graphes suivants.



Pdt portant le GTIN	Nb de GTIN
1	11079
2	228
3	21
4	20
5	1

TABLE 7 – Nombre de produits par GTIN

FIGURE 14 – Nombre de produits par GTIN

Distribution des fournisseurs fonction du nombre de produit par fournisseur

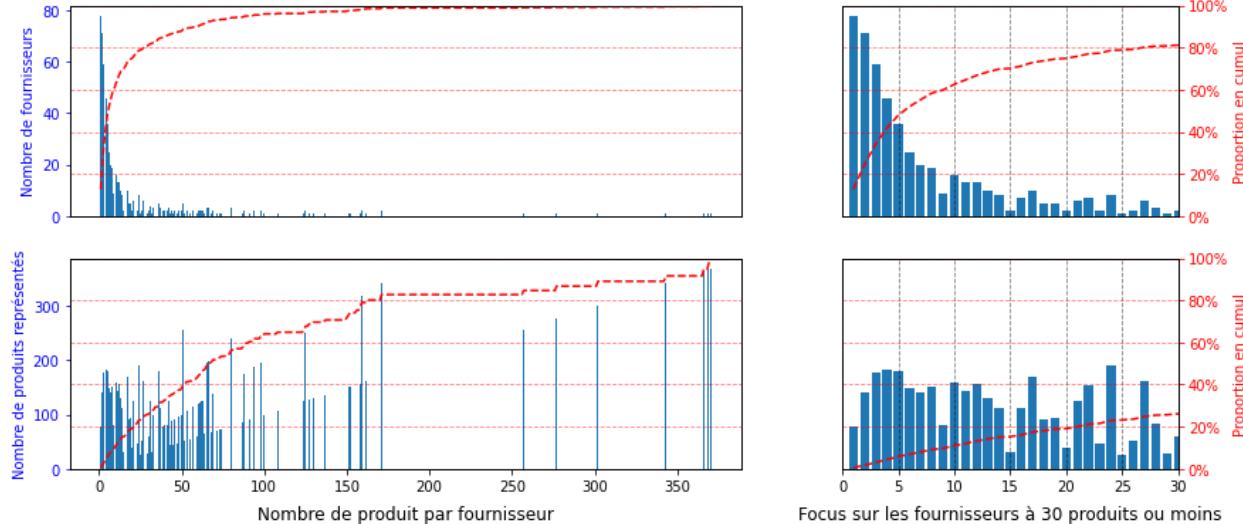


FIGURE 15 – Distribution des fournisseurs par nombre de produits

TODO : continuer l'analyse par domaine de données.

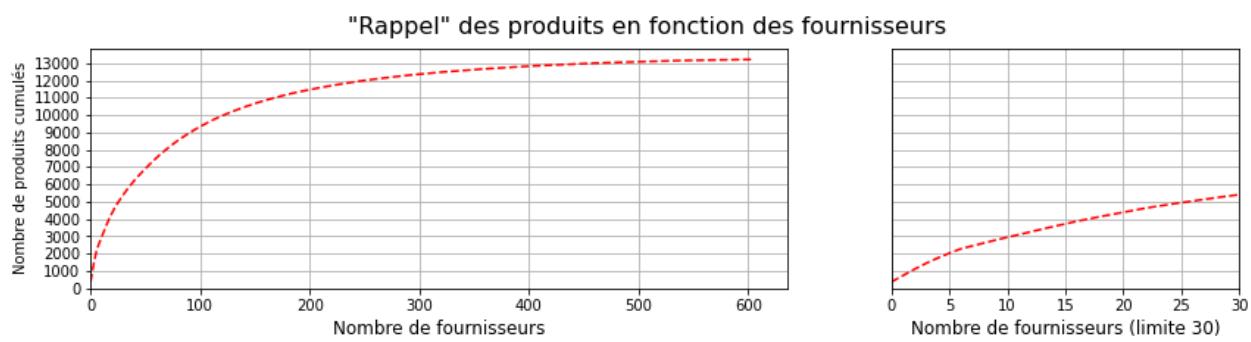


FIGURE 16 – Le Pareto des produits en fonction des fournisseur

uid	Code produit	Code fournisseur	Type de produit	GTIN	data_ok
976ba68b-b8c8-4be1-885d-8176bc591bf3	PIMP-00000004202	PIMF-0000000350	grocery	3039825400044	True
24c6c44a-636c-494c-8ed4-55938e876a42	PIMP-0000003605	PIMF-0000000357	grocery	9002100033675	False
cfb9793-4e5f-48e7-8dc0-d7653fe47d70	PIMP-0000012864	PIMF-0000000147	grocery	3256540003125	False
8bccddc40-d795-4c7e-8c68-8aabd7b2690fc	PIMP-0000008144	PIMF-0000000313	grocery	3275923050778	True
68318fb3-b8cd-483f-e348-bf856acb344c	PIMP-0000003253	PIMF-0000000440	grocery	8001250008503	True

data_ok	Code produit	Code fournisseur	Type de produit	GTIN
Hors qualité	count	8622	8622	7572
	unique	8622	550	7359
	top	PIMP-0000007080	PIMF-0000000250	
En qualité	freq	1	315	64
	count	4590	4590	4463
	unique	4590	339	4139
En qualité	top	PIMP-0000007612	PIMF-0000000179	288
	freq	1	306	3079

TABLE 8 – Exemples de codes d'identification

uid	Unité de base	Poids net	Poids brut	Poids net égoutté	Volume	data_ok
976ba68b-b8c8-4be1-885d-8176bc591bf3	SAC	5,000	5,050	-	-	True
24c6c4a-63ac-494c-8ed4-5593ee876a42	BTE	1,050	1,140	-	-	False
cfb9793-4e5f-4857-8d0-d7053fe47d70	PAQ	320,000	330,000	-	-	False
8bceddc40-d795-4c7e-8cb8-aacd7b2690fc	BTE	0,055	0,182	-	-	True
68318fb3-b8cd-483t-8348-bf8564cb344c	PAQ	1,000	1,066	-	-	True

TABLE 10 – Exemples de dimensions

data_ok	Unité de base	Poids net	Poids brut	Poids net égoutté	Volume
count	8622	8250,000	8250,000	679,000	855,000
unique	29	-	-	-	-
top	BTE	-	-	-	-
freq	2209	-	-	-	-
mean	-	3,628	-	-	-
std	-	74,276	-	-	-
min	-	0,000	3,310	1,541	5,426
25%	-	0,500	52,524	1,072	44,001
50%	-	1,000	0,000	0,000	0,000
75%	-	3,000	0,548	0,500	0,500
max	-	4900,000	1,145	1,560	4,125
count	4590	4590,000	4730,500	2,415	1000,000
unique	26	-	3,332	6,050	1000,000
top	SAC	-	4730,500	347,000	1058,000
freq	1080	-	-	-	-
mean	-	2,184	-	-	-
std	-	3,116	2,405	1,345	9,645
min	-	0,000	3,313	1,235	98,485
25%	-	0,450	0,000	0,000	0,000
50%	-	1,000	0,515	0,425	0,500
75%	-	3,000	1,075	1,000	0,960
max	-	33,474	3,160	2,335	1,832
			40,589	10,000	3100,000

TABLE 11 – Description des dimensions sur le dataframe

uid	Durée de vie totale	Durée minimale restante	Type de conservation	Conservation avant ouv.	Conservation après ouv.	Température	data_ok
976ba68b-b8c8-4be1-885d-8176bc591bf3	1080.0	730.0	AM	ambientTemperatureoilAndDryPlace	-	True	
24c6c4a-63ac-494c-8ed4-5593ee8f6a42	548.0	180.0	AM	coolAndDryPlace	-	False	
cfb9793-4e5f-48e7-8d0-d7053fe47d70	24.0	15.0	AM	coolAndDryPlace	-	False	
8bccddc40-d795-4c7e-8c6a-8aacd7b2690fc	730.0	365.0	AM	protectedFromHeatAndBurglaryPlace	-	True	
68318fb3-b8cd-483t-8348-bf8564cb344c	540.0	360.0	AM	protectedFromHeatAndBurglaryPlace	-	True	

TABLE 12 – Exemples de conservation

data_ok	Durée de vie totale	Durée minimale restante	Type de conservation	Conservation avant ouv.	Conservation après ouv.	Température
count	5526.000	5983.000	8229	6288	6267	8
unique	-	-	2	7	18	8
top	-	-	AM	coolAndDryPlace	coolAndDryPlace	10
freq	645.580	347.550	8208	5676	2762	1
mean	489.808	373.920	-	-	-	-
std	0.000	0.000	-	-	-	-
min	360.000	180.000	-	-	-	-
25%	540.000	280.000	-	-	-	-
50%	900.000	480.000	-	-	-	-
75%	9999.000	9999.000	-	-	-	-
max	3416.000	3482.000	4590	3688	3679	13
count	-	-	2	7	18	5
unique	-	-	AM	coolAndDryPlace	coolAndDryPlace	15
top	-	-	4577	2595	1750	5
freq	661.385	358.287	-	-	-	-
mean	483.434	390.790	-	-	-	-
std	0.000	0.000	-	-	-	-
min	360.000	180.000	-	-	-	-
25%	540.000	360.000	-	-	-	-
50%	730.000	480.000	-	-	-	-
75%	9999.000	9999.000	-	-	-	-

TABLE 13 – Description des conservations sur le dataframe

4.2 Données non structurées

4.2.1 Les libellés

Plusieurs libellés permettent de faire référence à chaque produit, avec des usages distincts :

Libellé temporaire unité de besoin : il s'agit simplement d'un libellé qui est choisi par la personne à l'origine de la demande de référencement produit (souvent l'acheteur, cf. FIGURE 5 page 22), afin que le fournisseur comprenne sur quel produit porte le référencement.

Désignation du produit fournisseur : c'est le libellé qui est donné par le fournisseur afin d'identifier simplement son produit.

Code article interne fournisseur : c'est l'identifiant du produit, dans le système de gestion du fournisseur. Il s'agit généralement de codes, avec autant de formats distincts que de fournisseur venant contribuer.

Marque commerciale du produit : il s'agit du nom de la marque commerciale du produit.

Dénomination réglementaire : c'est le libellé qui doit décrire de manière neutre le produit, sans notion liée au marketing (telle que la marque, entre autres), et qui doit obligatoirement figurer sur l'emballage du produit.

Des exemples de libellés sont présentés à la TABLE 14 page 42.

Libellé temporaire	Désignation produit fournisseur	Code interne fournisseur	Marque commerciale	com-	Dénomination réglementaire
VIN BEAUJO NOUVEAU(75CLX6) D MOREL	BEAUJOLAIS NOUVEAU	BN 2019	DOMINIQUE MOREL		VIN AOP
COCKTAIL FRT NATUREL BTE 5/1X3 DUNE	COCKTAIL DE FRUITS A L'EAU - FORMAT 5/1	COCEODUR	DUNE Restau- ration	COCKTAIL DE FRUITS A L'EAU	DE
Sauce barbecue en coupelle 20 g GYMA	Sauce barbecue en coupelle 20 g GYMA	400937	GYMA		Sauce Barbecue
Confiture abricot en bocal 450 g VALADE	Confiture abricot en bocal 450 g VALADE	PF100006	VALADE EN CORREZE		Confiture d'abricots
Confiture extra de figue en bocal verre	BONNE MAMAN CONF FIGUES VIOLETTTE 370	20000929	BONNE MAN		Confiture Extra de Figues violettes.

TABLE 14 – Exemples de libellés produit

4.2.2 Les listes d'ingrédients

L'autre grand type de donnée non structurées sont les listes d'ingrédients. La construction des listes d'ingrédients doit suivre les règles suivantes, même si l'application n'est pas toujours parfaitement respectée :

- elle doit détailler l'ensemble des ingrédients, y compris les additifs et les arômes

- elle doit être triée par ordre d'importance pondérale décroissante (i.e. les ingrédients les plus représentatifs en poids doivent être cités en premier)
- la quantité de certains ingrédients (en pourcentage de la masse) par exemple ceux mis en valeur sur l'étiquetage ou dans la dénomination de vente (ex. gâteau aux fraises, pizza au jambon)

Même s'il ne s'agit pas d'une exigence réglementaire, le Groupe Pomona demande à ses fournisseurs de ne pas distinguer les ingrédients par phase comme cela se fait parfois. Cela signifie, par exemple, séparer une partie de la composition du produit (la pâte de la garniture pour une tarte, la sauce et les raviolis, ...). De telles pratiques peuvent parfois induire le consommateur en erreur, comme par exemple dans la liste d'ingrédients suivante (s'applique à des chips de légumes) :

Légumes 64% (betterave, panais, carottes, patates douces), huile de tournesol, sel marin.

Sans l'artifice d'avoir regroupé les légumes en une seule phase, le premier ingrédient de la liste aurait pu être l'huile de tournesol, qui est un ingrédient moins attractif pour un consommateur de chips de légumes.

Quelques exemples de listes d'ingrédients sont présentées à la TABLE 15 page 43.

Désignation produit fournisseur	Liste d'ingrédients
HARICOTS BLANCS À LA TOMATE	Haricots blancs (UE et non UE), eau, sel, tomate concentré, antioxydant : acide citrique.
POIVRE VERT DÉSHYDRATÉ	100% poivre vert
Crème de marron de l'Ardèche en boîte 500 g CLEMENT FAUGIER	Châtaignes 50%, sucre, marrons glacés, sirop de glucose, eau, extrait naturel de vanille
Velouté de poireaux pommes de terre hyposodé en sachet 800 g NEFF MADA	Pomme de terre 32 %, amidon modifié de pomme de terre, féculle de pomme de terre, maltodextrine de blé, poireau 8 %, arômes naturels, sucre, oignon, antiagglomérant : E551 "nano", plantes aromatiques, curcuma
Poivre Kampot rouge en pot 50 g TERRE EXOTIQUE	Poivre de Kampot rouge 100%

TABLE 15 – Exemples de listes d'ingrédients

Les contraintes ci-dessus s'appliquant aux listes d'ingrédients font qu'en général, il s'agit d'une énumération d'ingrédients, sans doublon.

4.3 Pièces jointes

Dans le PIM, les pièces jointes - intéressantes au titre de l'information produit - gérées au niveau du produit sont les suivantes :

- les fiches techniques fournisseur
- les étiquettes produits

D'autres types de pièces jointes sont gérées, mais elles ne seront pas décrites dans le détail (car non pertinente pour le cas d'usage présenté) :

les certificats des labels : ce sont des documents pdf produits par les organismes de certification attestant qu'un produit peut porter un label. Ils sont mis à disposition par les fournisseurs

les images produit : ce sont des visuels des produits (photographies ou images construites) qui visent à être utilisées dans des catalogues ou sur les sites de vente en ligne

les fiches logistiques : ce sont des fichiers qui viennent compléter les informations de la fiche technique lorsque cette dernière ne porte pas les informations relatives à la hiérarchie logistique.

les fiches techniques et argumentaires Pomona : ce sont des documents pdf produits par le PIM, stockées sur les articles, qui reprennent des informations produit et article (cf. section 2.2.2 page 18). Elles sont transmises aux clients ou utilisés par les commerciaux Pomona. Elles permettent d'avoir une présentation uniforme de l'ensemble de l'assortiment (les fiches techniques fournisseur ont des formats très variables)

4.3.1 Fiches techniques fournisseur

Généralités sur les fiches techniques

Une fiche technique fournisseur est un document, d'une à une dizaine de pages, qui reprend l'essentiel des informations techniques à propos du produit. Elles portent globalement l'ensemble des informations produit telles que présentées à la section 2.1.3 page 14. C'est le document le plus complet vis-à-vis des informations produit, il porte en général des informations complémentaires à toutes les informations présentes sur l'emballage du produit (les étiquettes). En général, une fiche technique ne porte d'information que pour un unique produit, mais dans le cas d'assortiments, les informations peuvent être relatives à plusieurs d'entre eux. Cf. la fiche technique pour l'assortiment de confitures, présentée en annexe B.1.4 page 78. Les fiches techniques fournisseur sont des pièces jointes qui sont collectées par la branche ÉpiSaveurs depuis la mise en place du logiciel de gestion historique GIP (cf. section 2.3.2 page 23), et il s'agit d'une données obligatoire dans le PIM également.

Le format des fiches techniques

Dans le PIM, ces pièces jointes sont collectées et stockées sous forme de pdf (les autres formats de fichier ne sont pas autorisés). Sauf de rares exceptions, il s'agit de fichiers issus de logiciels de traitement de texte, au format A4 portrait. Ce sont des documents techniques, et sont donc en général très structurés, avec des paragraphes, sous-paragraphes, tableaux, ... Comme cela se constate aisément sur les exemples présentés en annexe, même si les informations portées sont sensiblement toujours les mêmes, les formats de ces documents sont extrêmement variables. Il n'y a aucun standard réglementaire ou normatif relatif à la construction de ces fiches, chaque industriel constitue donc son propre format. On a donc un foisonnement de formats différents.

Focus sur le mode de présentation de quelques informations

Les informations produits sont en général présentées de la manière suivantes dans les fiches techniques :

Liste d'ingrédients : elle est quasiment toujours présentée sous forme de texte (identique à la liste d'ingrédients telle qu'affichée sur l'emballage du produit), mais elle peut également être présentée sous forme de tableau. Cf. la fiche technique présentée en annexe B.1.6 page 82 (celle-ci ne porte pas de liste d'ingrédients sous forme de texte)

Allergènes : il peuvent être :

- soit mis en évidence dans la liste d'ingrédients via la police (gras, souligné, majuscules)
- soit listés hors de la liste d'ingrédient, sous forme de texte. Cf. la fiche technique présentée en annexe B.1.3 page 76
- soit listés sous forme d'un tableau reprenant l'ensemble des allergènes réglementaires, et le niveau de présence associé (présence, contamination croisée ou absence). Cf. la fiche technique présentée en annexe B.1.5 page 80

Données nutritionnelles : les données nutritionnelles sont en général présentées sous forme de tableau dans les fiches techniques. De plus, les informations nutritionnelles sont souvent données pour 100 grammes (ou 100 millilitres), ce qui est réglementaire, mais également parfois pour une portion. La taille de la portion est définie arbitrairement par le fournisseur. Cf. la fiche technique de la préparation pour panna cotta B.1.3 page 76.

Données logistiques : les informations relatives à la hiérarchie logistiques se présentent souvent sous forme de tableaux. L'interprétation de ces tableaux est généralement complexe et nécessite un peu de réflexion.

4.3.2 Étiquettes produit

Généralités sur les étiquettes

L'étiquette produit est la partie de l'emballage du produit qui porte les informations produit. Selon la technologie de l'emballage - très liée au fait que le produit est plutôt brut ou plutôt transformé - il peut s'agir :

- d'une photo de l'étiquette collante apposée sur l'extérieur de l'emballage du produit (cf. exemple des étiquettes de lentilles B.2.3 page 88 ou de sauce soja B.2.5 page 90)
- d'un applet de l'emballage, ou son bon-à-tirer, qui est le document qui est ensuite envoyé aux chaînes de production pour impression (cf. exemple du bon à tirer pour la préparation pour panna cotta B.2.4 page 89)
- ou de tout autre visuel montrant une partie indissociable de l'emballage physique du produit (ex : un photo de la face de l'emballage qui porte les informations produit, cf. l'étiquette des madeleines B.2.2 page 87)

Pour résumer, l'étiquette est une pièce jointe qui représente l'information produit qui « voyage avec le produit ». La cohérence entre les données de l'étiquette et l'information produit transmises aux client est un enjeu majeur, dans la mesure où ce sont en général les informations portées par le produit physique qui sont correctes. La collecte systématique des étiquettes produit est une nouveauté arrivée avec la mise en place du PIM pour ÉpiSaveurs (mai 2019). En théorie, les étiquettes mises à disposition par les fournisseurs devraient systématiquement porter les informations réglementaires. Or, du fait de la relative nouveauté de cette collecte, les pièces jointes transmises ne sont pas toujours conformes (cf. l'étiquette des lentilles B.2.3 page 88, qui ne portent aucune information nutritionnelle ou de composition).

Le format des étiquettes

Comme les fiches techniques, ces pièces jointes sont collectées et stockées sous forme de pdf (les autres formats de fichier ne sont pas autorisés). Du fait que les natures mêmes de ces pièces jointes sont diverses (cf. paragraphe précédent), il n'existe pas de format prédominant.

Focus sur le mode de présentation de quelques informations

Les étiquettes portent moins d'information que les fiches techniques. Par exemple, elles ne portent pas d'information sur la hiérarchie logistique, les données administratives (tel que le taux de TVA ou la nomenclature douanière), les codes d'identification (hormis le GTIN qui est présent sur le code à barre), ... Les durées de vie ne sont pas mentionnées : sur l'emballage d'un produit seule la date limite apparaît, et elle dépend du lot de production. En règle générale, hormis quelques allégations volontairement affichées par l'industriel, l'étiquette ne porte que les informations réglementaires. Les information produit positionnées sur l'étiquette se présentes de la manière suivante :

Liste d'ingrédients : elle est toujours présentée sous forme d'un texte énumérant les ingrédients (cf. la section 4.2.2 page 42 qui détaille le contenu d'une liste d'ingrédients)

Allergènes : ils sont uniquement mis en évidence dans la liste d'ingrédients, par l'utilisation d'une police spécifique (gras, souligné, majuscule, ...). Cela peut se constater par exemple sur l'étiquette de madeleines B.2.2 page 87 ou celle de la sauce soja B.2.5 page 90

Données nutritionnelles : les données nutritionnelles se présentent souvent sous forme d'un tableau, toujours avec les valeurs pour 100 grammes ou 100 millilitres, et parfois pour une portion. Il peut arriver, plutôt pour les produits peu transformés, que ces données soient simplement écrite sous forme de texte tel que

Énergie : 1101 kJ/260 kcal, Glucides : 57 g, Sucres : 52 g, Protéines : 4,3 g, Sel : 7,1 g

4.4 Récapitulatif de la complétude des données

Mettre ici un ou plusieurs tableaux récapitulatifs illustrant les données possédées quantitativement.

Montrer également la complétude en fonction du statut, de la date de création, et des facettes de migration.

Mettre l'exemple du champ "acide gras trans >1%"

4.5 Analyse qualitative des données

4.5.1 Évaluation de la qualité des données

Montrer qu'un sondage basique fait que la qualité actuelle est perfectible

4.5.2 Types de pdf possédés

Dire ici qu'il y a finalement beaucoup de pdf qui possèdent des textes extractibles vs. uniquement des images.

4.6 Les données « manuellement étiquetées »

4.6.1 Pour répondre à quel besoin ?

La qualité des données actuellement présentes dans le système fait que la cohérence entre les listes d'ingrédients du PIM et celles présentes dans les fiches techniques n'est pas assurée. Comme dans tout modèle de machine learning se basant sur des données, il est indispensable ici d'avoir des données dont on est sûrs de la qualité. Il a donc été décidé d'étiqueter manuellement un échantillon représentatif de fiches techniques, en leur faisant correspondre leurs listes d'ingrédients.

4.6.2 Mode de constitution de l'échantillon

Le code pour la constitution de l'échantillon est présenté dans le notebook en annexe C.2 page 119.

Les règles de gestion adoptées pour la constitution de cet échantillon sont les suivantes :

- On constitue un échantillon de 500 fiches techniques étiquetées
- On se limite aux produits d'Épicerie et de Boisson non alcoolisée, car ce sont les produits pour lesquels la réglementation impose d'afficher une liste d'ingrédients aux consommateurs
- On stratifie cet échantillon par type de produit (Épicerie vs. Boisson non alcoolisée)
- On se limite évidemment à des produits qui possèdent une fiche technique

Il aurait pu être intéressant de se limiter aux produits « en qualité » (cf. les définitions données à la section 4.1.2 page 36 sur les statuts des produits), mais l'étiquetage manuel avait été fait avant l'analyse de ces statuts. La constitution de cet échantillon s'est traduite par la génération d'un fichier csv de 500 lignes, avec 3 colonnes :

- l'uid du produit de l'échantillon
- la désignation produit fournisseur

- une colonne vide, visant à accueillir les listes d'ingrédients lors de l'étiquetage manuel

4.6.3 Méthodologie de l'étiquetage manuel

Cette activité s'est faite simplement, dans un environnement Windows :

- les pièces jointes ont été téléchargées localement, dans des dossiers portant l'uid du produit associé
- le csv contenant les uid et les désignation produit fournisseur a été ouvert dans le tableur Microsoft Excel
- en prenant chaque uid séquentiellement, il était relativement rapide de rechercher localement la fiche technique correspondante dans l'explorateur de fichier et de l'ouvrir
- le plus souvent, il était possible de copier/coller la liste d'ingrédient dans le tableur Excel
- le fichier a ensuite été à nouveau sauvegardé au format csv, afin de pouvoir être chargé dans pandas

Une passe de nettoyage des caractères spéciaux issus des copier/coller a ensuite été faite dans excel.

4.6.4 Règles de gestion pour l'étiquetage manuel

Malgré le fait que l'exercice semble à priori peu complexe, en pratique un nombre assez élevé de cas particuliers ont nécessité de prendre des décisions, parfois arbitraires. En général, il s'agit de décisions à prendre lorsque des mentions qui ne sont pas des ingrédients au sens strict sont incluses dans le texte des ingrédients. Le principe de base qui a été retenu est d'essayer de coller au plus à ce qui est attendu dans le PIM. Par exemple, dans le PIM on demande de retirer les préfixes du type « Ingrédients : » car ces mentions sont reprises telles sur les sites internet, ce qui peut se traduire par l'affichage de « Ingrédients : Ingrédients : ... ».

L'ensemble des règles sont documentées à l'annexe B.3.1 page 92.

4.6.5 Confrontation avec le contenu du PIM

Il est possible de comparer les liste d'ingrédients du PIM, et ceux des données étiquetées manuellement. Si on prend uniquement en compte les égalités strictes, on a un niveau de cohérence exactement à 10% (50 produits sur les 500 du périmètre). Des exemples de liste d'ingrédients en écart sont présentées à la TABLE 16 page 49. Il apparaît clairement que l'essentiel des écarts sur cet échantillon sont dus à des ajustements de forme (mise en majuscule des allergènes, retrait de parenthèses, retours à la ligne, ...).

Ingrédients du PIM	Ingrédients de la ground truth
Céréales (farine de FROMENT (27,5%), céréales complètes (15,1%) (farine complète de FROMENT (15%), farine complète de SEIGLE, farine complète d'ORGE) sucre graisses végétales (palme, palmiste, colza) poudre de cacao maigre (5,1%) amidon de BLE LAIT écrémé en poudre (équivalent lait 16%) amidon de pomme de terre poudres à lever (carbonates de sodium et d'ammonium, diphosphates) LACTOSE et protéines de LAIT sel arôme émulsifiant : lécithine de tournesol agent de traitement de la farine : E223 (SULFITES).	Céréales [farine de froment (27.5%), céréales complètes (15.1%) (farine complète de froment (15 %), farine complète de seigle , farine complète d'orge)] sucre graisses végétales (palme, palmiste, colza) poudre de cacao maigre (5.1%) amidon de blé lait écrémé en poudre (équivalent lait 16%) amidon de pomme de terre poudres à lever (carbonates de sodium et d'ammonium, diphosphates) lactose et protéines de lait sel arôme émulsifiant : lécithine de tournesol agent de traitement de la farine : E223 (sulfites) Peut contenir des traces de fruits à coque, de soja, d'oeuf et de sésame
cheddar fondu(44%) (eau, LACTOSERUM, cheddar(5%) (LAIT, ferment, acidifiant (sulfates de sodium (E514)), enzymes), LAIT écrémé, maltodextrine, antioxydant (phosphate de sodium (E339)), CREME, épaisseurs (octényle succinate d'amidon sodique (E1450), gomme xanthane (E415)), acidifiant (sulfates de sodium (E514))), eau, huile de palme, épaisseurs (amidon modifié de tapioca, amidon modifié de maïs), maltodextrine, acidifiants (sulfates de sodium (E514), citrate de sodium (E331), acide phosphorique (E338), acide lactique (E270)), arôme, antioxydant (phosphate de sodium (E339)), colorants (bêta carotène (E160a), annatto (E160b), extrait de paprika (E160c)), émulsifiant (stéaroyl-2-lactylate de sodium (E481), mono et diglycérides d'acides gras alimentaires (E471))	cheddar fondu (44%) (eau, lactosérum, cheddar (5%) (lait, ferment, acidifiant (sulfates de sodium (E514)), enzymes), lait écrémé, maltodextrine, antioxydant (phosphate de sodium (E339)), crème, épaisseurs (octényle succinate d'amidon sodique (E1450), gomme xanthane (E415)), acidifiant (sulfates de sodium (E514)), eau, huile de palme, épaisseurs (amidon modifié de tapioca, amidon modifié de maïs), maltodextrine, acidifiants (sulfates de sodium (E514), citrate de sodium (E331), acide phosphorique (E338), acide lactique (E270)), arôme, antioxydant (phosphate de sodium (E339)), colorants (bêta carotène (E160a), annatto (E160b), extrait de paprika (E160c)), émulsifiants (stéaroyl-2-lactylate de sodium (E481), mono et diglycérides d'acides gras alimentaires (E471))
Pommes* (80%), framboises* (20%)	pommes* (80%), framboises* (20%) *Ingrédient issu de l'agriculture biologique
Eau, graines de MOUTARDE, vinaigre d'alcool, sel, vin blanc (contient SULFITES), sucre, épices, acidifiant (acide citrique), conservateur (DISULFITE de potassium)	Eau, graines de moutarde, vinaigre d'alcool, sel, vin blanc (contient sulfites), sucre, épices, acidifiant (acide citrique), conservateur (disulfite de potassium)
Pâte à tartiner aux NOISETTES et au cacao 81,5 % (sucre, huile de palme, NOISETTES 13%, LAIT écrémé en poudre 8,7%, cacao maigre 7,4%, émulsifiants : lécithines [SOJA] , vanilline), farine de FROMENT 16%, levure de bière, extrait de malt d'ORGE, sel, LAIT écrémé en poudre, émulsifiants : lécithines [SOJA] , protéines de FROMENT, protéines de LAIT, eau.	pâte à tartiner aux noisettes et au cacao 81,5 % (sucre, huile de palme, noisettes 13%, lait écrémé en poudre 8,7%, cacao maigre 7,4%, émulsifiants : lécithines [soja] ; vanilline), farine de froment 16%, levure de bière, extrait de malt d'orge, sel, lait écrémé en poudre, émulsifiants : lécithines [soja] ; protéines de froment, protéines de lait, eau. Le chocolat utilisé est un chocolat pur beurre de cacao.
Sirop de glucose, sucre, amidons transformés, acidifiants : acide citrique, acide malique, correcteurs d'acidité : citrate tricalcique, malate acide de sodium, agent d'enrobage : cire de carnauba, arôme, concentrés de fruits et de plantes : carthame, citron, colorants : carmins, bleu patenté V, lutéine, sirop de sucre inverti.	sirop de glucose ; sucre ; amidons transformés ; acidifiants : acide citrique, acide malique ; correcteurs d'acidité : citrate tricalcique, malate acide de sodium ; agent d'enrobage : cire de carnauba ; arôme ; concentrés de fruits et de plantes : carthame, citron ; colorants : carmins, bleu patenté V, lutéine ; sirop de sucre inverti.
Huile de colza, jaunes d'OEUVFS, purée de tomates, eau, vinaigre, câpres (2.8 %) (câpres, eau, sel), cornichons (2.8 %) (cornichons, vinaigre, sel), sucre, sel, amidon modifié, acidifiant : glucono-delta-lactone, conservateur : sorbate de potassium, antioxydant : E385, épice, arôme naturel aneth.	Huile de colza (70.45 %), jaunes d'oeufs (6.2 %), purée de tomates, eau, vinaigre, câpres (2.8 %) (câpres, eau, sel), cornichons (2.8 %) (cornichons, vinaigre, sel), sucre, sel, amidon modifié, acidifiant : glucono-delta-lactone, conservateur : sorbate de potassium, antioxydant : E385 (74 mg/kg), épice, arôme naturel aneth.
Lentilles, eau, sel, arôme naturel (CELERI)	Lentilles, eau, sel, arôme naturel (céleri)
Ingrédients : Sucre, pâte de cacao, NOISETTES (23%) , beurre de cacao, BEURRE concentré, émulsifiant (lécithine de SOJA), arôme, LAIT écrémé en poudre. Cacao : 46% minimum dans le chocolat noir. PEUT CONTENIR AUTRES FRUITS À COQUE.\t\t\t	Sucre, pâte de cacao, NOISETTES (23%) , beurre de cacao, BEURRE concentré, émulsifiant (lécithine de SOJA), arôme, LAIT écrémé en poudre. Cacao : 46% minimum dans le chocolat noir. PEUT CONTENIR AUTRES FRUITS À COQUE.
Huile de colza, eau, vinaigre d'alcool, jaunes d'OEUVFS, sucre, estragon (2,5%), échalote (2,4%), sel, cerfeuil, jus de citron à base de concentré, piment de Cayenne, arômes, amidon modifié, colorant : bêta-carotène, conservateur : sorbate de potassium.	Huile de colza, eau, vinaigre d'alcool, jaunes d'oeufs, sucre, estragon (2,5%), échalote (2,4%), sel, cerfeuil, jus de citron à base de concentré, piment de Cayenne, arômes, amidon modifié, colorant : bêta-carotène, conservateur : sorbate de potassium.Ce produit peut contenir des traces de moutarde.

TABLE 16 – Exemples d'écart entre les données étiquetées et celles du PIM

Troisième partie

LES OBJECTIFS DE CE PROJET

Chapitre 5 LES CAS D'USAGE

5.1 Objectifs : Qualité et productivité

Comme présenté précédemment, 1 on dépense de l'énergie, et 2 on a une qualité de données qui est perfectible. Un traitement automatique des documents mis à disposition permettrait de décharger les personnes qui interviennent dans le processus (fournisseurs, acheteurs, gestionnaires de référentiels, ingénieurs qualité) et de garantir une meilleure pertinence de l'information produit.

5.2 La préalimentation d'information

Préalimenter les informations, sous réserve d'avoir un outil suffisamment fiable, permettrait de faire gagner du temps aux fournisseurs. Trois obstacles :

- cela n'a d'intérêt que si le système est capable de produire de l'information structurée avec une fiabilité élevée (par exemple, 80% de données correctes serait un minimum)
- cela entre en concurrence directe avec le système GDSN présenté au paragraphe portant le même nom à la section 2.3.3 page 25. Or, ce système est justement spécialement conçu pour faire transiter les informations produit des fournisseurs aux distributeurs, avec une standardisation des échanges
- cela apporte l'essentiel de la valeur aux fournisseurs, mais pas au Groupe

5.3 Le contrôle des informations transmises

Cf. le schéma présenté à la FIGURE 5 page 22. Plus le taux de détection des erreurs est élevé, et plus les erreurs sont détectées tôt, mieux c'est :

- la qualité des données s'en trouve évidemment améliorée

- le processus est plus court en temps, en évitant les aller-retours
- on décharge l'ensemble des acteurs, en limitant la

5.3.1 Le contrôle à la saisie fournisseur

Si on alerte le fournisseur au moment où il saisit, on peut dès le début du processus éviter une erreur. Cela pourrait avoir lieu quand il soumet ses données, on fait tourner un traitement et on remonte des avertissements dans l'IHM du PIM.

5.3.2 L'aide aux vérifications Pomona

Lors des contrôles, on pourrait également remonter les incohérences détectées entre pièces jointes et données à contrôler.

5.3.3 Les contrôles en masse asynchrones

Enfin, il pourrait être pertinent de faire tourner de manière asynchrone des contrôles de qualité de données sur l'ensemble de la base. TODO : détailler un peu le pourquoi c'est nécessaire (exemple du champ acides gras trans), et un des points difficiles : acquittement des avertissements non pertinents.

Chapitre 6

LE CHOIX DU CAS D'USAGE

6.1 La représentation dominante des fiches techniques

On a beaucoup de fiches techniques, pas beaucoup d'étiquettes ou de fiches logistiques (cf. la table qui va bien.) On va donc plutôt dans un premier temps tenter de travailler avec les fiches techniques.

6.2 Les multiples formats et le besoin de « spatialisation »

Chaque fournisseur décide évidemment du format de document qu'il souhaite produire. On a donc beaucoup de formats différents, et un Pareto finalement trop « mou » pour envisager de construire des templates pour récupérer les informations automatiquement (cf. FIGURE 16 page 38).

6.3 Les informations « spatialisées »

Les données que l'on souhaite récupérer sont globalement de 3 types :

- données de composition
- données nutritionnelles
- données logistiques

Comme vu dans la section 4.3.1 page 44, un grand nombre d'informations sont spatialisées. Par exemple, la représentation des données nutritionnelles se fait régulièrement sous forme de tableau. Or, c'est un peu compliqué à interpréter, car il fut réussir à interpréter un tableau, et à en sortir des couples de clé / valeur.

6.4 La complexité dans la représentation des données logistiques

Les données logistiques sont souvent difficile à interpréter pour un humain, donc cette activité peut paraître difficile à déléguer à une machine. Mettre 2-3 exemples.

6.5 L'identification d'une liste d'ingrédient par son contenu

Il est possible de dire si un texte est une liste d'ingrédients, juste en lisant ce texte. Par exemple, « 14g » peut être une quantité de glucides, de lipides, le poids d'une pièce unitaire, ... Mais un texte tel quel <mettre ici un exemple> a de forte chance d'être le contenu d'une liste d'ingrédients.

6.6 Conclusion quant au choix du cas d'usage

Tant la préalimentation, que l'aide au contrôle des données seraient faisable d'un point de vue technique. Il suffirait pour cela de publier un service, qui fonctionnerait de la manière suivante :

- le PIM appelle le service, avec un message contenant l'uid du produit à contrôler ou préalimenter
- le serveur récupère du PIM les données nécessaires au contrôle ou à la préalimentation
- en retour, il renvoie au PIM soit l'état du contrôle (OK, erreur, avertissement, avec les précisions nécessaires), soit les données telles qu'elles doivent être préalimentées
- le PIM, sur la base de ce retour, affiche le résultat du contrôle ou bien alimente les données et les présente à l'utilisateur

Au vu des différentes contraintes listées dans ce document, on s'attachera à extraire les listes d'ingrédients des produits alimentaires de la branche EpiSaveurs depuis les fiches techniques fournisseur, en se basant sur le contenu textuel de ces documents.

Quatrième partie

CONSTRUCTION DU MODÈLE

Chapitre 7 LES PRINCIPES GÉNÉRAUX

7.1 Contenu du texte d'une liste d'ingrédients

En général, chaque ingrédient sera présent une seule fois dans la liste (cf. section 4.2.2 page 42)

Le calcul d'embeddings via des modèles tels que SVD ou Word2Vec fait peu de sens.

l'extraction des textes se fait au format Bag Of Words, sans utiliser de notion d'IDF. L'utilisation de TF semble églement ne pas amener de valeur à priori.

7.2 Limitation à l'identification des listes d'ingrédients

On est sur une taxonomie d'informations limitée dans les fiches techniques.

On pourrait envisager de classifier l'ensemble des textes présents dans les fiches techniques.

Mais l'absence de données étiquetées rend cette tâche impossible. La charge d'étiquetage d'un nombre représentatif de blocs de texte de fiches techniques est trop importante pour être mise en oeuvre dans le cadre de ce projet.

7.3 Conversion de documents en texte

dire ici qu'on utilise principalement pdfminer vs. d'autres outils d'OCR.

De plus, on partira dans un premier temps sur une transformation basique d'un document en texte, sans passer par une analyse de la localisation des textes sur le document (cf. les difficultés présentées dans la section 6.2 page 51).

Chapitre 8

CONSTRUCTION D'UN MODÈLE SIMPLE

« OUVERT »

Le fonctionnement global de ce premier modèle (présenté à la FIGURE 17 page 54) ne respecte pas les principes du Machine Learning. Il permet juste d'éprouver la méthode pressentie, ainsi que de se faire une idée de l'efficacité d'un modèle de ce type. En effet, même si on utilise des briques d'extraction de features depuis des textes, il manque une partie de mesure de la performance, indispensable pour pouvoir évaluer et améliorer la pertinence du modèle. Les illustrations de ce chapitre sont issues du notebook présenté en annexe C.3 page 124, et le code des classes utilisées (IngredientExtractor et PIMIngredientExtractor) est inclus dans le module pimest, en annexe D.4 page 153. Ce modèle n'utilise pas les données étiquetées manuellement (présentées à la section 4.6 page 47), mais se base simplement sur les listes d'ingrédients du PIM.

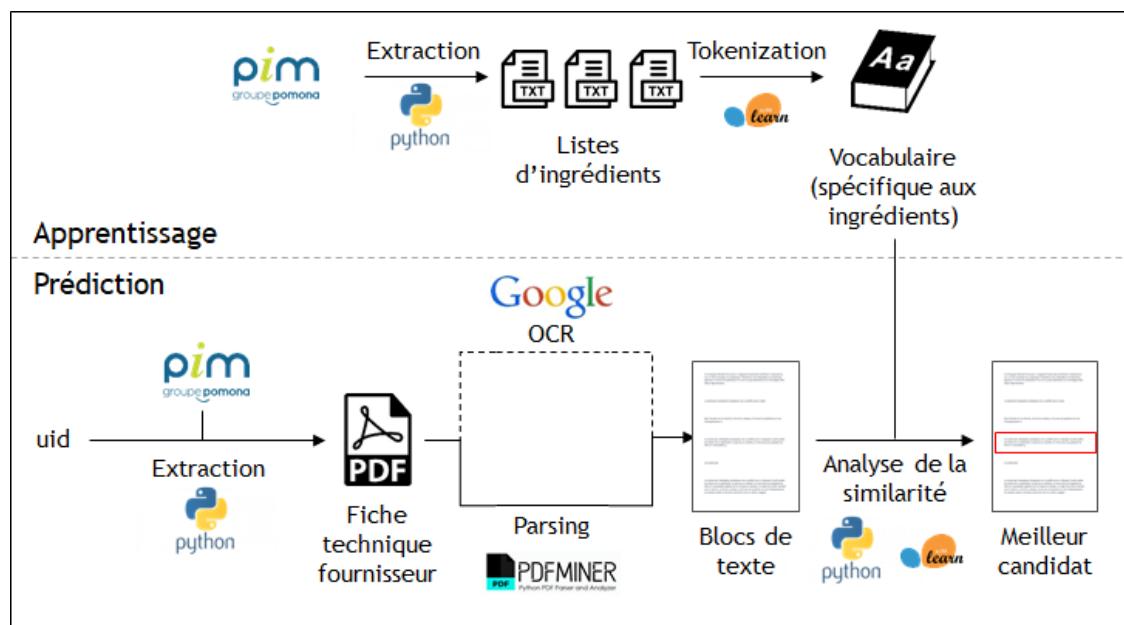


FIGURE 17 – Schéma de principe du « modèle ouvert »

8.1 Entraînement

8.1.1 Périmètre

Pour l’entraînement de ce modèle, on va uniquement se limiter aux produits d’épicerie ou de boissons non-alcoolisées. En effet, ce sont pour ces produits que la réglementation impose d’afficher en clair la composition aux consommateurs. On se limitera aussi aux produits qui portent une liste d’ingrédients, et qui sont « En qualité » (cf. les définitions données à la section 4.1.2 page 35 sur les statuts des produits). TODO : je me suis arrêté ici.

8.2 Conversion en blocs de texte

On utilise la bibliothèque PDFMiner.six. Elle nous sort un long string qui contient le texte entier du document. On applique une « bête » fonction : on splitte ce string quand on observe 2 retours à la lignes consécutifs. Le code est présenté en annexe. Pour le moment, vu la proportion importante de PDF dont le contenu est extractible

8.3 Train/Test split

On fait un split 50/50, on se base sur les uid pour identifier les produits. Sur le train set, on récupère les listes d’ingrédients du PIM.

8.4 Entrainement du modèle

L’entraînement est basique : on constitue seulement un vocabulaire en utilisant la fonctionnalité mise à disposition dans scikit-learn.

8.5 Calcul de la similarité

On calcule la similarité cosinus entre chacun des blocs, et le vocabulaire. On prend, systématiquement l’argmax de la similarité qu’on propose comme liste d’ingrédients.

8.6 Illustration des résultats obtenus

Mettre ici les résultats sur quelques fiches techniques présentées en annexe. Spoiler : rien que comme ça, les résultats sont encourageants.

8.7 Pistes d'améliorations identifiées

En plus de la mesure de la performance, qui est indispensable avant de pouvoir procéder à des ajustements.

Pistes identifiées :

- Faire un découpage du gros texte en blocs plus malin, potentiellement avec des expressions régulières
- Faire des « ngrams de blocs », ce qui permettrait de parfois fusionner des blocs qui ont été séparés (car contenaient des retours à la ligne successifs)
- Essayer une autre manière de calculer la similarité ?

Chapitre 9

UTILISATION DES DONNÉES

MANUELLEMENT ÉTIQUETÉES

Comme présenté à la section 4.6.5 page 48 relative à la comparaison entre les données du PIM et celles récupérées lors de l'étiquetage, il y a un grand nombre d'écart. Or, si on entraîne le modèle et qu'on mesure sa performance sur des données de mauvaise qualité, on aura de mauvais résultats. On va donc construire un modèle se basant sur les données manuellement étiquetées. Le fonctionnement de ce modèle est présentés à la FIGURE 18 page 57. La méthodologie utilisée à cette partie est présentée dans le notebook « Modèle basé sur les données manuellement étiquetées » en annexe C.4 page 130. Les différents transformateurs et estimateurs spécifiques sont définis dans le module pimest, inclu en annexe D.4 page 153.

9.1 Chargement des données manuellement étiquetées

La toute première étape est la constitution d'un dataframe contenant :

- les uid pour indexer les produits
- les listes d'ingrédients manuellement étiquetées depuis les fiches techniques
- le contenu de chacune des fiches techniques au format texte

On commence par charger les données du fichier csv contenant les uid et les listes d'ingrédients. Ensuite, un pipeline scikit learn d'acquisition des données est lancé. Il s'agit de 3 transformateurs en série, qui effectuent les travaux suivants :

- construction du chemin pointant vers les fiches techniques (sur la base des uid)
- construction d'une feature contenant les données des fichiers, en binaire
- construction du texte complet de la fiche technique (en se basant sur la library pdfminer.six)

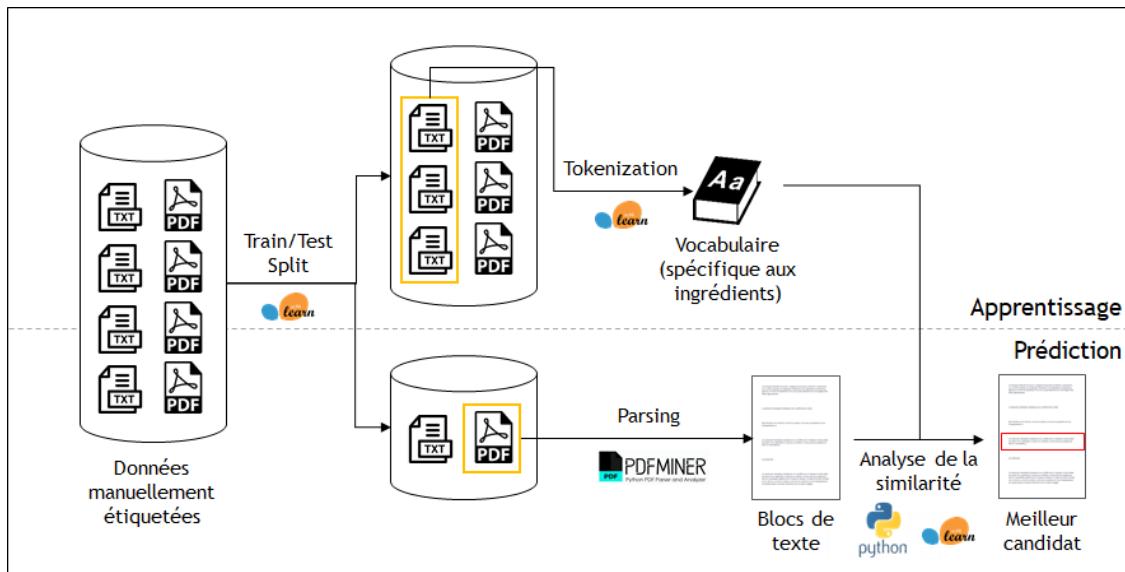


FIGURE 18 – Schéma de principe du modèle basé sur les données étiquetées

Le résultat du lancement de ce pipeline est présenté à la TABLE 17 page 58.

9.2 Découpage des textes en blocs

Le second travail est le découpage des textes en blocs. Dans un premier temps, on va simplement effectuer ce découpage en splittant le texte lorsque deux retours à la ligne successifs sont détectés. Un exemple de découpage est présenté ci-dessous.

30/12/19	Poids net	Contaminants / Contaminating
Date d'impression :	Poids brut	Ionisation / Irradition
Remarque :	Origine	
Les informations contenues dans cette fiche technique sont données de bonne foi, en l'état actuel de nos connaissances, et selon les indications communiquées par le producteur ou le fournisseur. Il appartient au client de vérifier la conformité de la marchandise par rapport à l'usage qu'il en fait.	/ net weight	OGM / GMO
Création :	/ gross weight	Pesticides/ Pesticides
12/06/12	/ Origin	Métaux Lourds
12 rue René Cassin	0,07 Kilogramme	/ Heavy Metals
37390 NOTRE DAME	0,125 Kilogramme	Allergènes et leurs dérivés (si présents)
Télé :	CANADA	/ Allergens (if existing)
02 47 85 55 00	/ General information	Conformité à la directive 1999/2/CE (22/02/99)
Fax :02 47 41 33 32	Informations générales	Produit non ionisé et ne contenant pas d'ingrédients ionisés.
FICHE TECHNIQUE	DLUO conseillée / "Best before date" recommandé	Not irradiated accordingly with the Reg 1999/2/CE (22/02/99).
Mélange du trappeur, 70 g	Nomenclature douanière / Customs code	Free from GMO
Trapper blend, 70g	Conditions idéales de stockage	Ne contient pas d'OGM, est non soumis à l'étiquetage sur les OGM
Code article KEREX	/ Conditions of storage	Conforme à la directive 396/2005 /CE
Nom latin (si disponible)	Ingrédients :	In accordance with Reg 396/2005 /CE
/ EAN Code	Conserver dans un endroit frais et sec	Conforme au règlement 1881/2006 /CE
Code barre	Store in a cool dry place	In accordance with Reg 1881/2006 /CE...
/ KEREX Code	5 ans / 5 years	
/ (Latin name)	0910999900	
TEEPTRAPPEUR	Sucre, poivre noir, coriandre, légumes déshydratés (ail, oignon, poivron rouge), sel de mer, sucre d'érable, arôme d'étable naturel, huile végétale (canola)	Gluten
X	Sugar, black pepper, coriander, dehydrated vegetables (garlic, onion, red bell pepper), sea salt, maple sugar, natural maple aroma, vegetable oil (canola)	Crustacés
3760063322262	/ Ingredients	Oeufs
		Poisson
		Soja
		Lait
		Fruits à coque - Arachides
		Céleri
		Moutarde
		Sésame
		Sulfites
		Lupin
		Mollusques

text

NESCAFÉ® SPÉCIAL FILTRE
 \n\nDose individuelle de 2 g\nTechnologie micro-grains\nCODE EAN (UC)\n\nn3033710076017\n\nDENOMINATION LEGALE DU PRODUIT\nDESCRIPTION DU PRODUIT\nnCafé instantané et café torréfié moulu.\n\nUne dominante Arabica pour l'arôme et une pointe de Robusta pour le \ncorsé, associés à une torréfaction légère pour un café équilibré et peu \namer.\nSachet dose pour une tasse.\n\nnDOSAGE PRECONISÉ\nn\nnMODE OPERATOIRE\nnPour obtenir\nn1 café Court (DA)\n\nn1 café Long (DA)\n\nnEau\nn\nn7\nn\nn12\nn\nncl\nn\nnNESCAFÉ®\n\nnSPÉCIAL FILTRE\nn\nn2\nn2\nn\nng\nn\nng\nn\nnA reconstituer avec de l'eau.\n\nnTempérature de l'eau : 75°C\nnPour une qualité optimale, utilisez de l'eau filtrée.\n\nnIngrédients : Café instantané, café torréfié moulu (3%).\n\nnINGRÉDIENTS\nn\nnPROFIL GUSTATIF\nn\nnIntensité\nn\nnConditionné sous atmosphère protectrice.\n\nnENGAGEMENT QUALITÉ\nn\nn NESTLÉ a un système de management de la qualité, le NMS (NESTLÉ \nManagement System), en cohérence avec les systèmes ISO 9001 ...

LENTEILLES BLONDES 4mm\n\nRéférence PQG007-3.22.1\nnVersion\nnDate d'application :\\nPage 1/2\\n\\nG\\n\\n15/10/2019\\n\\nPrésentation\\n\\nCaractéristi -\\n\\nques \\n\\nphysico-\\nchimiques \\n\\nDéfinition\\n\\nOrigine\\nDénomination \\n\\négale\\n\\nLentilles de couleur brun clair. Elles sont de forme biconvexe et \\npossèdent une peau assez épaisse. Leur diamètre est compris \\nentre 4mm et 5mm\\n\\nChine, Canada, France, Italie, USA, Turquie\\n\\nLentilles blondes\\n\\nProcess\\n\\nNettoyage, épirrage, triages\\n\\nConservation\\n\\n36 mois à l'abri de la chaleur et de l'humidité\\n\\nCritères d'analyses\\n\\nMoyenne/Tolérance\\n\\nMéthodes\\n\\nHumidité\\n\\nMatières minérales étrangères\\n\\nMatières végétales étrangères\\n\\nGraines\\n\\nImpropres\\n\\nBrises\\n\\nGermées\\n\\nCalibre 4-5 mm\\n\\n11,5% / 16%\\n\\n0,05% / 1%\\n\\n0,15% / 0,5%\\n\\n0,5% / 1%\\n\\n0,4% / 1%\\n\\n0,05% / 1%\\n\\n95% / 90%\\n\\nNF V03707\\n\\nMicrobiologie\\n\\nIl n'existe pas de réglementation concernant les exigences microbiologiques \\npour ce produit.\\n\\nPesticides\\nMét...

FICHE TECHNIQUE

PRODUIT FINI

n°000100

Purée de Poire Sans Sucres Ajoutés

Date d'application : 05/05/2014

Page : 1/2

Coupelettes Aluminium 120 x 95 g

Définition : Ce produit est une purée de fruits obtenue à partir des parties comestibles des fruits (après broyage et sans concentration notable). Ce produit est sans sucres ajoutés : il contient uniquement les sucres naturellement présents dans les fruits.

La purée présente une texture homogène et légèrement granuleuse.

La stabilité du produit est obtenue par pasteurisation et dosage à chaud.

Aspects nutritionnels

Désignation et liste des ingrédients

Valeurs nutritionnelles (pour 100 g)

Désignation légale : Purée de Poires sans sucres ajoutés *

Contient les sucres naturellement présents dans les fruits

Liste des ingrédients : Poire 99,9%, antioxydant : acide ascorbique.

Matières grasses

Energie : 65 kcal

Acides gras saturés : n

Glycides : n

Fibres alimentaires : n

Pro...
n

TABLE 17 – Exemples du contenu de fiches techniques au format texte (tronqués)

/ Gluten	Absence	/ Total plat count (APC)
/ Crustaceans	Absence	E. Coli
/ Eggs	Absence	/
/ Fish	Absence	/ Salmonella
/ Soy	Absence	/ Yeasts
/ Milk	Absence	/ Moulds
/ Peanuts and Treenuts	Absence	/ Total aflatoxin
/ Celery and celeriac		B1 aflatoxin
/ Mustarde		/
/ Sésame		
/ Sulphites		NF V05-051 < 6 000 000 / g
/ Lupin		NF V08-053 < 10 / g
/ Shellfish		NF V08-052 Absence dans 25g
Absence		NF V08-059 < 10 000 / g
Absence		NF V08-059 < 10 000 / g
Absence		Kit Enzymatique < 10 ppb
Absence		Kit Enzymatique < 5 ppb
Absence		
Absence		
<hr/>		
Caractères microbiologiques		
<hr/>		
/ Microbiological characteristics		
<hr/>		
Microorganismes aérobies 30 °C		
Escherichia coli		
Salmonelles		
Levures		
Moisiures		
Aflatoxine Total		
Aflatoxine B1		
<hr/>		

On constate que le découpage n'est pas idéal, cf. la fiche technique de ce produit, présentée en annexe B.1.7 page 85. Les séparations des cellules des tableaux de cette fiche ne sont pas prises en compte, et on a des blocs trop étendus.

9.3 Train/Test split

Dans la mesure où l'on possède assez peu de données, on va conserver un échantillon assez important dans le jeu d'entraînement : 400 produits (soit 80% des données disponibles).

9.4 Entraînement du modèle

On fait tourner de la même manière que sur le modèle dit « ouvert », à savoir qu'on ne préprocesse pas les données avant d'appliquer le CountVectorizer.

9.5 Illustration des prédictions obtenues

Un échantillon des prédictions obtenues est présenté dans la TABLE 18 page 61. Pour éviter d'avoir des listes d'ingrédients prédites prenant trop de place dans cette table, celles dont la longueur dépasse 500 caractères ont été filtrées avant génération de cet échantillon. Les résultats présentés à cette table sont donc vraisemblablement biaisés, dans la mesure où les très longues listes prédites doivent avoir plus de chance d'être erronées.

Les grandes tendances qui se dégagent à l'analyse de cette liste sont les suivantes :

- globalement, les résultats sont bons. On retrouve régulièrement des morceaux de texte qui sont similaires à la liste cible
- une erreur qui revient régulièrement est le fait que le découpage en blocs est parfois imparfait, on sélectionne « trop large »
- à l'inverse, le modèle n'a pas retiré des listes d'ingrédients prédites des mentions qui ont été rétirées lors de l'étiquetage manuel (cf. les règles d'annotation présentées en annexe B.3.1 page 92) : les préfixes de type « Liste d'ingrédients : », les allégations telles que « Teneur totale en sucres : 60g pour 100g »
...
- le modèle semble plus performant lorsque la liste d'ingrédients réelle est longue. On le vérifiera dans le chapitre relatif à la mesure de la performance du modèle

Un mot sur les cas où la liste d'ingrédients cible ou prédite sont vides :

- Les listes d'ingrédients cible sont vides lorsque la pièce jointe ne mentionnait pas de liste d'ingrédients. Cela peut arriver, et les produits concernés n'ont pas été sortis de l'échantillon. Il est important de pouvoir aussi mesurer les faux positifs, qui sont nombreux avec cette technique de choix systématique du meilleur candidat
- Les listes d'ingrédient prédites sont vides lorsque l'outil de parsing des pdf (pdfminer.six) n'a extrait aucun texte. C'est le cas quand la pièce jointe était un document imprimé qui a été scanné. Le texte n'est présent que sous forme d'image (cf. la fiche technique du sel en annexe B.1.1 page 74)

TABLE 18: Extrait des résultats de la prédiction

Liste d'ingrédients cible	Liste d'ingrédients prédicté
<p>sucré*, LAIT en poudre*, beurre de cacao*, pâte de cacao*, émulsifiant : lécithine de tournesol (E322), extrait de vanille*</p> <p>* matière première issue de l'agriculture biologique</p> <p>cacao : 27% minimum</p>	<p>Liste des Ingrédients :</p> <p>sucré*, LAIT en poudre*, beurre de cacao*, pâte de cacao*, émulsifiant : lécithine de tournesol (E322), extrait de vanille*</p> <p>* matière première issue de l'agriculture biologique</p>
<rien>	<rien>
<p>Amidon de maïs* - Lait écrémé* - Sel - Fécule de pomme de terre* - Tomate* - Oignon* - Arômes naturels - Poivres* 3 % (poivre vert*, poivre blanc*, poivre noir*) - Huile de tournesol* - Extrait de levure* - Sucre caramélisé* - Ail* - Maltodextrine de maïs*.</p> <p>* issus de l'agriculture biologique</p>	<p>Amidon de maïs* - Lait écrémé* - Sel - Fécule de pomme de terre* - Tomate* - Oignon* - Arômes naturels - Poivres* 3 % (poivre vert*, poivre blanc*, poivre noir*) - Huile de tournesol* - Extrait de levure* - Sucre caramélisé* - Ail* - Maltodextrine de maïs*.</p> <p>* issus de l'agriculture biologique</p>
semoule de blé dur supérieure et de l'eau	Ingédients : semoule de blé dur supérieure et de l'eau
<rien>	<p>Boisson gazeuse aromatisée au jus de fruit à base de concentré</p> <p>S.Pellegrino Orange 33 cl (Aranciata)</p> <p>S.Pellegrino Citron 33 cl (Limonata)</p> <p>Le plaisir des fruits à l'italienne</p>
<p>Eau, maltodextrine, sel, arômes, sucre, arôme naturel de citronnelle, amidon modifié, ail en poudre, épices (com-bava, curcuma), extraits d'épices (gingembre, poivre), stabilisant (gomme xanthane).</p>	<p>Eau, maltodextrine, sel, arômes, sucre, arôme naturel de citronnelle, amidon modifié, ail en poudre, épices (com-bava, curcuma), extraits d'épices (gingembre, poivre), stabilisant (gomme xanthane).</p>
<p>Sucre, cacao maigre en poudre (beurre de cacao : 11% minimum), arôme vanille.</p> <p>Cacao : 32% minimum</p>	<p>Sucre, cacao maigre en poudre (beurre de cacao : 11% minimum), arôme vanille.</p>
<p>Sucre ; sirop de glucose ; graisse de palme ; humectant : sirop de sorbitol ; gélatine ; acidifiant : acide citrique ; arôme.</p>	<rien>
Gésier de dinde émincé 50%, graisse de canard 47%, sel, arômes naturels de poivre.	Gésier de dinde émincé 50%, graisse de canard 47%, sel, arômes naturels de poivre.
Purée de tomates mi réduite (64%), sucre, vinaigre, amidon modifié, sel, acidifiant : acide citrique	<p>Liste des ingrédients : Purée de tomates mi réduite (64%), sucre, vinaigre, amidon modifié, sel, acidifiant : acide citrique</p>
<rien>	<p>Boisson gazeuse aromatisée au jus de fruit à base de concentré</p> <p>S.Pellegrino Orange 33 cl (Aranciata)</p> <p>S.Pellegrino Citron 33 cl (Limonata)</p> <p>Le plaisir des fruits à l'italienne</p>
<p>Sirop de glucose-fructose, framboises 35%, sucre, gélifiant : pectines, acidifiant : acide citrique.</p>	<p>Liste ingrédients : Sirop de glucose-fructose, framboises 35%, sucre, gélifiant : pectines, acidifiant : acide citrique.</p> <p>Préparée avec 35g de fruits pour 100g de produit fini.</p> <p>Teneur totale en sucres : 60g pour 100g.</p>
Café instantané, café torréfié moulu (3%).	<ul style="list-style-type: none"> - NESTLÉ a un système de management de la qualité, le NMS (NESTLÉ Management System), en cohérence avec les systèmes ISO 9001 et ISO 22000. - Etiquetage conforme à la réglementation en vigueur sur les OGM - Ce produit ne contient pas d'ingrédients ionisés. - Certifications usines : ISO 9001, FSSC 22000, ISO 14001 et OHSAS 18001. - Agrément sanitaire : site non soumis à agrément sanitaire
<p>sucré, pâte de cacao, beurre de cacao, cacao maigre en poudre, émulsifiant : lécithine de tournesol (E322), arôme vanille</p> <p>cacao : 50% minimum</p>	<p>Liste des Ingrédients :</p> <p>sucré, pâte de cacao, beurre de cacao, cacao maigre en poudre, émulsifiant : lécithine de tournesol (E322), arôme vanille</p>

Continued on next page

TABLE 18: Extrait des résultats de la prédiction

Liste d'ingrédients cible	Liste d'ingrédients prédicté
Eau, huile de tournesol, beurre 9,5 %, jaune d'oeuf 6 %, amidon modifié, sel, jus de citron concentré, amidon de maïs, épaississants (gomme guar, gomme xanthane), protéines de pois, sucre, arôme naturel, curcuma, extrait de paprika.	Eau, huile de tournesol, beurre 9,5 %, jaune d'oeuf 6 %, amidon modifié, sel, jus de citron concentré, amidon de maïs, épaississants (gomme guar, gomme xanthane), protéines de pois, sucre, arôme naturel, curcuma, extrait de paprika.
Piment rouge fort équeuté* (85%), cumin, ail moulu (3%), eau, arôme naturel d'ail, sel, sorbate de potassium. (* présence de sulfites)	<p>A) Ingrédients :</p> <p>Piment rouge fort équeuté* (85%), cumin, ail moulu (3%), eau, arôme naturel d'ail, sel, sorbate de potassium. (* présence de sulfites)</p> <p>B) Origines des ingrédients :</p> <ul style="list-style-type: none"> - Piments : Maroc - Ail : chine - Sel, arôme, épices, sorbate de potassium : Europe <p>I)</p> <p>A) Critères Physico-chimiques :</p> <p>B) Critères microbiologiques :</p>
Sucre, amidon de maïs, arôme vanille	ajouter le produit à la préparation avec les autres ingrédients de la recette et mélanger pour homogénéiser
Salicornes de culture, eau, sel, acide citrique	Se consomment en légumes d'accompagnement avec toutes préparations de poissons ou crustacés ou aussi avec des viandes blanches
cèpes 70% (Boletus edulis et respective famille), huile de tournesol, blanquette 5% (Tuber borchii Vitt.), oignon, beurre, sel, protéines du lait, farine de riz, amidon de maïs, dextrose, extrait de levure, épices, arôme truffée, arômes naturels, antioxydant : acide l-ascorbique (E 300).	<p>CODE DU PRODUIT : NOME DU PRODUIT : FORMAT : BARCODE (EAN13) : NOMENCLATURE COMBINÉE :</p> <p>Ingrédients : cèpes 70% (Boletus edulis et respective famille), huile de tournesol, blanquette 5% (Tuber borchii Vitt.), oignon, beurre, sel, protéines du lait, farine de riz, amidon de maïs, dextrose, extrait de levure, épices, arôme truffée, arômes naturels, antioxydant : acide l-ascorbique (E 300).</p>
Eau, haricots verts, sel.	Allergène Égoutter, ne pas rincer. Faire sauter 3 minutes avec de la matière grasse.

Chapitre 10

MESURE DE LA PERFORMANCE

Comme vu aux chapitres précédents, il est indispensable de mesurer la performance de nos modèles. On le fera sur le modèle se basant sur les données manuellement étiquetées. Le principe est présenté à la FIGURE 19 page 62.

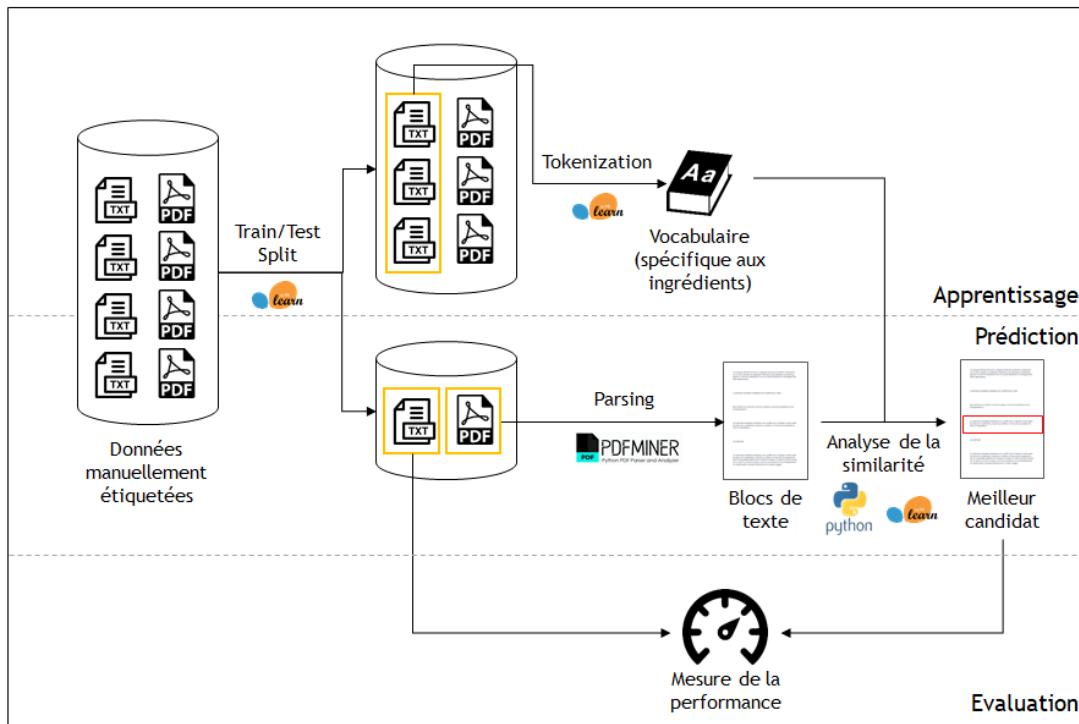


FIGURE 19 – Illustration de la méthodologie de mesure de la performance

10.1 Accuracy

La métrique qui tombe le plus sous le sens est l'accuracy (on utilisera le terme anglais pour éviter les confusions avec la notion de « precision » telle qu'elle est utilisée par exemple dans le f1-score). On mesure simplement la proportion de prédictions qui sont égales à la ground truth. La méthodologie utilisée est détaillée dans le notebook « Mesure de la performance » présenté en annexe C.5 page 140.

10.1.1 Approche naïve

Description de cette approche

L'approche « naïve » consiste simplement à mesurer la proportion de textes prédits strictement égaux à la ground truth. Or, comme cela a été vu précédemment :

- les règles d'étiquetage manuel, détaillées à l'annexe B.3.1 page 92, montrent que des transformations sont parfois appliquées au contenu des listes d'ingrédients des pièces jointes avant d'établir la ground truth
- les textes à comparer sont longs (jusqu'à quelques centaines de caractères), cf. TABLE 18 page 61
- la mise en forme, en particulier les retours à la ligne, ne sont pas positionnés aux mêmes endroits. Dans le parsing des documents pdf, lorsque le texte revient à la ligne après avoir atteint le bord de la page, on a un retour à la ligne. Ce comportement n'a pas été reproduit dans l'établissement de la ground truth
- le découpage en blocs de texte, de manière simple, produit des textes qui ne sont pas toujours le reflet du contenu spatialisé de ce document (cf. l'exemple donné à la section 9.2 page 57, et la fiche technique associée en annexe B.1.7 page 85)

On s'attend donc à avoir une « accuracy naïve » faible.

Les résultats obtenus

Comme présenté dans notebook « Analyse de la performance » (cf. annexe C.5 page 140), les résultats obtenus sont conformes à l'attendu : l'accuracy est très faible. Elle vaut 1% (1 échantillon sur 100 produits) lorsqu'on la mesure sur l'échantillon de test après entraînement sur l'échantillon d'entraînement. Le seul produit pour lequel la liste d'ingrédients a été correctement identifiée porte les ingrédients suivants.

Sirope de glucose, sucre, eau, stabilisants (E440i, E440ii, E415), acidifiants (E330, E450i), conservateur (E202).

Si on effectue une cross-validation sur l'ensemble des données étiquetées, en appliquant un découpage en 10 folds, on obtient une accuracy moyenne de $1.80\% \pm 1.89\%$.

Nécessité d'améliorer cette métrique

On a vu que l'accuracy calculée de manière naïve porte un jugement sévère sur la performance du modèle. Par exemple, à la troisième ligne de la TABLE 18 page 61, on voit bien que la liste d'ingrédients prédite est identique à la ground truth, si ce n'est qu'les retours à la ligne ne sont pas positionnés exactement au même endroit. Comme on souhaite pouvoir ajuster le modèle, il est nécessaire d'avoir une métrique de mesure de la performance qui soit plus précise.

10.1.2 Avec du « text-postprocessing »

Le principe

Afin de pallier ces problèmes de mise en forme de texte, on assouplit un peu les contraintes par rapport à l'égalité stricte. En effectuant un traitement de text processing, à la fois sur la ground truth et les résultats du modèle, on va comparer des textes un peu plus « standardisés ». Les traitements effectués sont les suivants :

- On passe le texte en minuscules
- On retire la ponctuation
- On remplace tous les « whitespaces » (retours à la ligne, espaces multiples, tabulations, ...) par des espaces simples
- On retire les accents

L'ensemble de ces transformations sont faites en utilisant les fonctionnalités proposées par le CountVectorizer de la bibliothèque scikit-learn (cf. le notebook en annexe C.5 page 140 et le module pimest inclus en annexe D.4 page 153).

Les résultats

Si on évalue cette nouvelle accuracy avec le text processing, sur l'échantillon d'entraînement après entraînement sur l'échantillon d'entraînement, on obtient une accuracy de 14% (14 listes d'ingrédients correctement prédites sur 100). Ces 14 listes d'ingrédients sont présentées à la TABLE 19 page 65.

De la même manière que précédemment, si on fait une cross-validation sur l'ensemble des données manuellement étiquetées, on obtient une accuracy de $16.60\% \pm 3.35\%$.

Les limites de cette métrique

Cette métrique est déjà plus intéressante que l'approche naïve, mais elle a quand même un défaut majeur : elle a une vision encore trop binaire des résultats. En effet, que le texte soit identique à un préfixe près (cf. la quatrième ligne de la TABLE 18 page 61), ou qu'il n'ait rien à voir (d'autres exemples sont présents dans cette même table), elle considérera la prédiction comme erronée. Or, il est important d'identifier les cas où le modèle s'est complètement trompé par rapport à ceux où il a quand même identifié le bon bloc contenant les ingrédients.

10.2 Fonctions de « loss » spécifiques

On peut aussi définir des fonctions de loss, qui permettent d'être plus fin qu'une simple évaluation OK / KO du résultat du modèle. On calculera une distance entre le résultat du modèle, et de la ground truth.

Liste d'ingrédients cible	Liste d'ingrédients prédicté
Gésier de dinde émincé 50%, graisse de canard 47%, sel, arômes naturels de poivre.	Gésier de dinde émincé 50%, graisse de canard 47%, sel, arômes naturels de poivre.
Edulcorants sorbitol, isomalt, sirop de maltitol, aspartame, mannitol, sel d'aspartame-acesulfame, acesulfame-k, sucralose ; gomme base (contient de la lecithine de SOJA), aromes, épaisseur gomme arabique, humectant glycerol, colorant E171, agent d'enrobage cire de carnauba, colorant E163, antioxydant BHA. Contient une source de PHENYLALANINE.	Edulcorants sorbitol, isomalt, sirop de maltitol, aspartame, mannitol, sel d'aspartame-acesulfame, acesulfame-k, sucralose ; gomme base (contient de la lecithine de SOJA), aromes, épaisseur gomme arabique, humectant glycerol, colorant E171, agent d'enrobage cire de carnauba, colorant E163, antioxydant BHA. Contient une source de PHENYLALANINE.
mini poivrons jaunes, eau, sucre, sel, affermissant chlorure de calcium : E509, acidifiant : acide citrique, vinaigre, antioxydant : vitamine C E330	mini poivrons jaunes, eau, sucre, sel, affermissant chlorure de calcium : E509, acidifiant : acide citrique, vinaigre, antioxydant : vitamine C E330
Farine de BLE, huile de colza non hydrogénée, OEUFS de poules élevées en plein air (21%), sucre, stabilisant : glycérin, sirop de glucose-fructose, émulsifiant : mono- et diglycérides d'acides gras, poudres à lever : diphosphates et carbonates de sodium (BLE), féculé, sel, arôme.	Farine de BLE, huile de colza non hydrogénée, OEUFS de poules élevées en plein air (21%), sucre, stabilisant : glycérin, sirop de glucose-fructose, émulsifiant : mono- et diglycérides d'acides gras, poudres à lever : diphosphates et carbonates de sodium (BLE), féculé, sel, arôme.
Pommes de terre 59,5 % - Céleris 40 % - Amidon de maïs - Sirop de glucose de maïs - Huile de colza - Emulsifiants : E322, E471 - Stabilisant : E450i - Curcuma - Conservateur : E223 - Antioxydant : E304 - Acidifiant : E330.	Pommes de terre 59,5 % - Céleris 40 % - Amidon de maïs - Sirop de glucose de maïs - Huile de colza - Emulsifiants : E322, E471 - Stabilisant : E450i - Curcuma - Conservateur : E223 - Antioxydant : E304 - Acidifiant : E330.
<rien>	<rien>
Amidon de maïs* - Lait écrémé* - Sel - Fécule de pomme de terre* - Tomate* - Oignon* - Arômes naturels - Poivres* 3 % (poivre vert*, poivre blanc*, poivre noir*) - Huile de tournesol* - Extrait de levure* - Sucre caramélisé* - Ail* - Maltodextrine de maïs*. * issus de l'agriculture biologique	Amidon de maïs* - Lait écrémé* - Sel - Fécule de pomme de terre* - Tomate* - Oignon* - Arômes naturels - Poivres* 3 % (poivre vert*, poivre blanc*, poivre noir*) - Huile de tournesol* - Extrait de levure* - Sucre caramélisé* - Ail* - Maltodextrine de maïs*. * issus de l'agriculture biologique
Farine de FROMENT, poudre de LACTOSERUM, sucre, poudre d'OEUF entier, poudres à lever : (E450, E500), matière grasse LAITIERE, sel. Peut contenir des traces de : soja, fruits à coques, lupin.	Farine de FROMENT, poudre de LACTOSERUM, sucre, poudre d'OEUF entier, poudres à lever : (E450, E500), matière grasse LAITIERE, sel. Peut contenir des traces de : soja, fruits à coques, lupin.
Eau, maltodextrine, sel, arômes, sucre, arôme naturel de citronnelle, amidon modifié, ail en poudre, épices (com-bava, curcuma), extraits d'épices (gingembre, poivre), stabilisant (gomme xanthane).	Eau, maltodextrine, sel, arômes, sucre, arôme naturel de citronnelle, amidon modifié, ail en poudre, épices (com-bava, curcuma), extraits d'épices (gingembre, poivre), stabilisant (gomme xanthane).
OEUFS, farine de BLE, sucre, amidon de BLE, stabilisants : sorbitols- glycérin, cacao maigre en poudre (3,5%), émulsifiants : E472b - E477, poudres à lever : E450 - E500, sirop de glucose, LAIT écrémé en poudre, sel, conservateur : E202, épaisseur : E410, arôme.	OEUFS, farine de BLE, sucre, amidon de BLE, stabilisants : sorbitols- glycérin, cacao maigre en poudre (3,5%), émulsifiants : E472b - E477, poudres à lever : E450 - E500, sirop de glucose, LAIT écrémé en poudre, sel, conservateur : E202, épaisseur : E410, arôme.
Sirop de glucose, sucre, eau, stabilisants (E440i, E440ii, E415), acidifiants (E330, E450i), conservateur (E202).	Sirop de glucose, sucre, eau, stabilisants (E440i, E440ii, E415), acidifiants (E330, E450i), conservateur (E202).
Flageolets verts. Jus : eau, sel, affermissant : chlorure de calcium (E509)	Flageolets verts. Jus : eau, sel, affermissant : chlorure de calcium (E509)
Carottes, eau, sucre, sel, vinaigre d'alcool, acidifiant : acide lactique. Présence fortuite de CELERI	Carottes, eau, sucre, sel, vinaigre d'alcool, acidifiant : acide lactique. Présence fortuite de CELERI
Légumes 43,2 % (pomme de terre, oignon, carotte, tomate, poireau) - Amidon modifié de pomme de terre - Extrait de levure - Sirop de glucose de maïs - Huile de colza - Sucre - Arôme naturel - Ail - Curcuma.	Légumes 43,2 % (pomme de terre, oignon, carotte, tomate, poireau) - Amidon modifié de pomme de terre - Extrait de levure - Sirop de glucose de maïs - Huile de colza - Sucre - Arôme naturel - Ail - Curcuma.

TABLE 19 – Prédictions identifiées comme correctes après postprocessing

10.2.1 Distance de Levenshtein

Brève description de chacune de ces distances.

10.2.2 Distance de Dameray-Levenshtein

Brève description de chacune de ces distances.

10.2.3 Distance de Jaro

Brève description de chacune de ces distances.

10.2.4 Distance de Jaro-Wrinkler

Brève description de chacune de ces distances.

10.2.5 Métriques non retenues

Distance de Hamming

10.2.6 Conclusion sur la métrique à utiliser

Une fois que cela aura été fait.

Chapitre 11

TRANSFER LEARNING

11.1 Principe du pré-entraînement

Expliquer qu'il s'agit d'une approche hybride des 2 modèles précédents On effectue une entraînement à la fois sur une partie des listes d'ingrédients du PIM, et sur une partie des données étiquetées. On vérifie ensuite, uniquement sur

11.2 Illustration de l'impact sur la performance

Ici, on montre l'impact sur la performance, du fait d'intégrer des listes d'ingrédients. On met en abscisse le nombre de listes d'ingrédients qu'on ajoute, et en ordonnée la performance du modèle (avec barre d'erreurs, via cross validation). On regarde si l'effet est positif : cela montrera s'il est intéressant d'avoir plus de données.

On regarde si on observe une saturation : cela montrera si on a déjà suffisamment de données sous forme de listes d'ingrédients dans le PIM, ou bien si ce serait intéressant d'en acquérir plus.

Chapitre 12

HYPERPARAMETER TUNING

On peut, dans l'optique d'améliorer la performance du modèle, ajuster certains paramètres et d'évaluer l'impact via une grid search. On fera tourner sur le modèle avec transfer learning.

12.1 Les paramètres ajustables

12.1.1 La prise en compte des « n-grams » dans la tokenization

On peut utiliser les n-grams lors de la tokenisation.

12.1.2 L'application de « n-grams » de blocs

Voir si dans la recherche du meilleur candidat, on s'autorise la constitution de « n-grams » de blocs.

12.1.3 L'utilisation d'expressions régulières dans le split des blocs

Voir si certaines expressions régulières pour splitter les blocs procurent de meilleurs résultats.

12.1.4 Applications d'autres fonctions de similarité

Voir l'impact d'utiliser d'autres manières de calculer la similarité.

1 - autre chose que la similarité cosinus (fonction du nombre de mots du bloc et de la proportion de mots issus du vocabulaire des ingrédients)

2 - en appliquant du TF et du TF-IDF

12.2 Application d'une grid search

Illustrer ici les résultats d'une grid search ou d'une random search si trop gourmand.

Cinquième partie

TRAVAUX SUBSÉQUENTS

Chapitre 13

OPÉRATIONNALISATION DE CETTE MAQUETTE

13.1 Client et sponsor métier

Estimation du ROI et identification d'un sponsor et d'un client.

13.2 Sélection du use case

Préalimentation ou appui au contrôle de données ?

13.3 Mise en place d'une organisation projet

13.3.1 Identification des compétences nécessaires

Nécessite des compétences diverses : développement côté PIM, compétences infra, définition du niveau de criticité de cette fonctionnalité (pour définition du monitoring et des plans de reprise d'activité)

13.3.2 Choix d'un cadre méthodologique projet

Scrum, c'est ce qu'on connaît le mieux.

13.3.3 Développement côté PIM

Développement des fonctionnalités telles qu'elles ont été présentées dans la section sur le choix du use case.

13.4 Industrialisation du code du modèle

Refactoring de certaines classes (e.g. : le requester, qui porte trop de responsabilités) Poursuite de l'écriture de tests unitaire pour avoir une couverture > 80%. Mise en place d'un processus de déploiement continu. Rédaction de la documentation : revue des docstring et mise en place d'un build Sphinx

13.5 Monitoring de la performance du modèle

voir la manière dont on peut superviser le niveau de performance du modèle. Capture-t-on en direct le retour des utilisateurs dans le PIM ?

Chapitre 14

EXTENSION DES FONCTIONNALITÉS

OFFERTES

14.1 Prise en compte de nouveaux types de pièces jointes

Aller chercher également les étiquettes.

14.2 Utilisation d'outil d'OCR pour les pdf non structurés

Intégrer ce qui a déjà été fait autour des solutions cloud, Google ou Azure.

14.3 Mise en place d'outil de spatialisation des textes

Charge importante, mais pourrait être utilisé pour d'autres sujets

14.4 Construction d'outils d'extraction de données connexes à la composition

Détermination des allergènes sur la base du contenu des listes d'ingrédients.

14.5 Élargissement aux données nutritionnelles

Si spatialisation faisable, tenter de récupérer des données nutritionnelles.

14.6 Évaluation de la performances sur d'autres familles de produits

Construction d'une nouvelle ground truth sur des fiches techniques de PassionFroid et TerreAzur, et évaluation de la performance du modèle.

Sixième partie

ANNEXES

Annexe A FIGURES, TABLEAUX ET BIBLIOGRAPHIE

LISTE DES TABLEAUX

1	Exemple de tableau de données nutritionnelles	15
2	Volumétrie article par branche	30
3	Utilisation des variables catégorielles article au sein des branches RHD	33
4	Répartition des produits par statut	35
5	Répartition des produits par statut de migration	36
6	Répartition des produits par « qualité des données »	36
7	Nombre de produits par GTIN	37
8	Exemples de codes d'identification	39
9	Description des codes d'identification sur le dataframe	39
10	Exemples de dimensions	40
11	Description des dimensions sur le dataframe	40
12	Exemples de conservation	41
13	Description des conservations sur le dataframe	41
14	Exemples de libellés produit	42
15	Exemples de listes d'ingrédients	43
16	Exemples d'écart entre les données étiquetées et celles du PIM	49

17	Exemples du contenu de fiches techniques au format texte (tronqués)	58
18	Extrait des résultats de la prédition	60
18	Extrait des résultats de la prédition	61
19	Prédictions identifiées comme correctes après postprocessing	65

TABLE DES FIGURES

1	Les flux métier avec les partenaires commerciaux	7
2	La répartition de l'activité des branches	10
3	Le maillage régional de la branche ÉpiSaveurs	12
4	La distinction entre produit et article	19
5	Le processus de création article	22
6	Une capture d'écran du PIM	24
7	L'intégration du PIM au sein des systèmes du Groupe Pomona	24
8	Schéma de principe de la GDSN	26
9	L'uid d'un produit	26
10	Volumétrie article par branche	30
11	Recouvrements entre branches RHD	31
12	Répartition des articles en fonction des variable catégorielles	32
13	Répartition des produits par statut	35
14	Nombre de produits par GTIN	37
15	Distribution des fournisseurs par nombre de produits	37
16	Le Pareto des produits en fonction des fournisseur	38
17	Schéma de principe du « modèle ouvert »	54
18	Schéma de principe du modèle basé sur les données étiquetées	57
19	Illustration de la méthodologie de mesure de la performance	62

BIBLIOGRAPHIE

- [1] Conseil de l'Union Européenne. Règlement n°1333/2008 sur les additifs alimentaires, dec 2008. <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:354:0016:0033:FR:PDF>.

- [2] Conseil de l'Union Européenne. Règlement n°1907/2006 dit REACH, dec 2006. <https://bit.ly/2Jm05v9>.
- [3] Conseil de l'Union Européenne. Règlement n°1169/2011 dit INCO, nov 2011. https://www.senat.fr/europe/textes_europeens/ue0120.pdf.
- [4] Direction Générale de la Concurrence, de la Consommation et de la Répression des Fraudes. Étiquetage des denrées alimentaires : nouvelles règles européennes, jan 2015. <https://www.economie.gouv.fr/dgccrf/etiquetage-des-denrees-alimentaires-nouvelles-regles-europeennes>.
- [5] Direction Générales des Douanes et Droits Indirects. Notions essentielles sur la Déclaration d'Échanges de Biens. <https://www.douane.gouv.fr/notions-essentielles-sur-la-declaration-dechangers-de-biens>.
- [6] GS1. GDSN Trade Item Implementation Guide, nov 2019. https://www.gs1.org/docs/gdsn/tiig/3_1/GDSN_Trade_Item_Implementation_Guide.pdf.
- [7] GS1 France. Le réseau GDSN, le canal pour l'échange d'informations produits. <https://www.gs1.fr/Notre-offre/Le-reseau-GDSN-le-canal-pour-l-echange-d-informations-produits>.
- [8] GS1 Global. Global Data Synchronisation Network. <https://www.gs1.org/services/gdsn>.
- [9] GS1 Global. GS1 General Specifications. https://www.gs1.org/docs/barcodes/GS1_General_Specifications.pdf.
- [10] Groupe Pomona. Site institutionnel du groupe pomona. <https://www.groupe-pomona.fr/>.
- [11] Wikipedia. Liste des additifs alimentaires. https://fr.wikipedia.org/wiki/Liste_des_additifs_alimentaires.

Annexe B

EXEMPLES DE PIÈCES JOINTES ET GROUND TRUTH

B.1 Fiches techniques

B.1.1 Fiche technique sel Cerebos



Fiche technique

CEREBOS® Sel Gros Alimentaire



Version 1.9

Page 1 / 1

date d'impression: 26.10.2010

No.-CAS:	7647-14-5	No.-EINECS:	231-598-3
Apparence:			produit blanc, cristallin
Analyses chimiques	Spécification	Typique	Méthodes
• Chlorure de sodium	> 99,8 %	99,9 %	ASTM 534-98
• Teneur en eau	< 0,1 %	0,02 %	ISO 2483
• Insolubles dans l'eau	< 0,01 %	0,005 %	ISO 2479
• Anti-agglomérant E 535	< 20 mg/kg		EuSalt AS 004
Granulométrie	Typique	Méthodes	
• > 3,15 mm	1 %	EN 1235	
• 1,00 - 3,15 mm	89 %		
• < 1,00 mm	10 %		
Propriétés physiques:		Méthodes	
• Masse volumique apparente	1.100 - 1.300 kg/m³	EN 1236	
Sur demande:			
• Iode (sous la forme de NaI) exprimé en I	15 - 20 mg/kg	EuSalt AS 002	
• Fluor (sous la forme KF)	250 mg/kg	EuSalt AS 017	
• Législation nationale: 212,5 - 287,5 mg F/kg			
Réglementation sur les denrées alimentaires, Impuretés et contaminants:			
Conforme au CODEX ALIMENTARIUS.			
Domaine d'application			
Sel gros de table ou de cuisine de qualité alimentaire pouvant être supplémenté en iode ou en iodure et fluor.			
Site de conditionnement			
Salines Cérébos et de Bayonne à Dombasle (France 54) et à Mouguerre (France 64) -			
Groupe esco			
Stockage			
Il est conseillé de ne pas gerber les palettes. Le CHLORURE DE SODIUM étant un produit hygroscopique, tout conditionnement ouvert devra être stocké à l'abri de l'humidité.			
Moyens de livraison			
• sur demande			

Les données précédentes résultent de nos contrôles qualité. Ces données ne dispensent pas l'utilisateur d'un contrôle à réception et ne sont pas forcément des garanties de vente. L'utilisateur est seul responsable du choix du produit en fonction de l'application souhaitée

esco - european salt company GmbH & Co.KG
 Headquarters • Landschaftstraße 1 • 30159 • Hannover • Allemagne • ☎ +49-(0)511-85030-0 ☎ ...-131
 esco benelux nv • Park Lane, Culliganlaan 2G bus 1 • B-1831 • Diegem • Belgique • ☎ +32-2711-0160 ☎ ...-0161
 esco france s.a.s • 49 Avenue Georges Pompidou • F-92693 • Levallois-Perret Cedex • France • ☎ +33(0)1.49.64.59.00 ☎ ...-1.49.64.59.10
 Vatel S.A. • Apartado 211-Sobralinho • P-2616-956 • Álverca • Portugal • ☎ +35-1219-5184-20 ☎ ...-39
 esco Espagne S.L. • Joan d'Austria, 39-47 • 08005 • Barcelona • Espagne • ☎ +34 (93) 2247238 ☎ +34 (93) 2214193
 esco Nordic AB • Drakegatan 10 • 401 23 • Göteborg • La Suède • ☎ +46-31 773 70-01 ☎ ...-02
 www.esco-salt.com Certification EN ISO 9001:2008

B.1.2 Fiche technique olives Valtonia

copram	FICHE TECHNIQUE	Date : 10/12/2019	V 1.2
	OLIVES NOIRES CONFITES DENOYAUTEES 4/4 VALTONIA		

INGREDIENTS :	Olives, eau, sel, stabilisateur de couleur : E579
---------------	---

MARQUE :	VALTONIA
BOITAGE :	4/4
ORIGINE :	Espagne
EAN 13 :	3061435001137
EAN colis :	3061435101134

CARACTERISTIQUES PHYSICO-CHIMIQUES ET ORGANOLEPTIQUES	
PH :	5.5 à 8.0
TAUX DE SEL :	2.0 à 3.5 %
Calibre :	30/33 unités / 100 g d'olives entières
ODEUR ET SAVEUR :	Franche et caractéristique
COULEUR :	Noire à robe de moine
PRÉSENCE DE NOYAUX :	1 noyau pour 100 fruits 2 fragments pour 100 fruits
DEFAUTS AUTRES CRITERES	Se référer au code des olives

CARACTERISTIQUES MICROBIOLOGIQUES	
Produit conforme à l'arrêté d'octobre 1997 relatif au contrôle de la stabilité des produits appétisés et assimilés (Norme AFNOR NF V08-401).	

VALEURS NUTRITIVES MOYENNES POUR 100 G DE PRODUIT EGOUTTE								
Kj	Kcal	Protéines	Matières grasses	Dont AGS	Glucides	Dont sucres	Sel	Fibres
515	125	0.5 g	13.0 g	2.2 g	0 g	0 g	2.0 g	3.0 g
INFORMATIONS COMPLEMENTAIRES								
Présence d'allergènes :	Non			Pesticides :	Conforme à la législation en vigueur.			
OGM et ionisation :	Absence			Métaux lourds :	Conforme à la législation en vigueur.			

CONDITIONNEMENT		
Contenance :	Poids net :	Poids net égoutté :
850 ml	800 g	360 g
Nombre de boîtes / carton :	Nombre de cartons /couche :	Nombre de couches / palettes :
6	12	12

CONSERVATION	
DLUO avant ouverture	3 ans. Stocker à température ambiante à l'abri de l'humidité
Après ouverture	au réfrigérateur dans un récipient alimentaire avec son liquide de couverture une semaine

Les informations contenues dans cette fiche sont celles dont nous disposons à la date de validation. Elles sont donc susceptibles d'être modifiées ultérieurement. Pour toute réclamation ou demande d'information merci de préciser impérativement le lot et/ou date de fabrication portés sur le couvercle.

Copram - Société de commercialisation des produits alimentaires du Maroc
 17, Quatrième Rue-Zone industrielle B.P. 168 – 13745 VITROLLES cedex – France
 RC Salon B 315 712/299 - Siret : 315 712 299 00017
 NAF : 511N - N° T.V.A. FR 03 315 712 299

B.1.3 Fiche technique Panna Cotta Nestlé

**Panna Cotta
NESTLÉ Docello®
Etui de 600 g (2 x 300g)
pour 50 portions**

Nestlé docello®

CODE EAN
9002100034771

DESCRIPTION DU PRODUIT	BÉNÉFICES CLÉS DU PRODUIT			
Préparation en poudre pour Panna Cotta.	Avec arôme naturel. Facile à mettre en œuvre et à personnaliser.			
INGRÉDIENTS				
Sucre, dextrose, maltodextrine, stabilisants (E460, E450, E516, E401, E404), épaississant (E407), arôme naturel (lait). Peut contenir : fruits à coque, œuf, soja et gluten.				
ALLERGÈNES MAJEURS	ENGAGEMENT QUALITÉ			
Conformément aux réglementations vigueur :	NESTLÉ a un système de management de la Qualité certifié par les normes ISO 9001 et FSSC 22000. Etiquetage conforme à la réglementation en vigueur sur les OGM. Ce produit ne contient pas d'ingrédients ionisés.			
MODE D'EMPLOI				
1. Porter à ébullition le mélange de lait et de crème liquide. 2. Hors du feu, verser en pluie tout en fouettant la préparation pour Panna Cotta puis mélanger jusqu'à parfaite homogénéisation. Porter à nouveau le mélange à ébullition. 3. Verser la préparation dans des ramequins, faire refroidir puis stocker en chambre froide (entre 0°C et +3°C). Servir frais. Suggestion: pour une Panna Cotta encore plus onctueuse vous pouvez réduire le dosage à 130g par litre de liquide.				
DOSAGES				
Produit déshydraté	Dosages			
	Base	Lait	Crème liquide	Nombre de portions (90g)
	150 g	0,5 L	0,5 L	12
	300 g	1 L	1 L	25
600 g	2 L	2 L	50	
UTILISATION				
Pour une texture plus souple et un rendement amélioré : un sachet (300 g) + 2 L de crème liquide + 1 L de lait = 36 portions (90 g). Idéal dans la composition d'un café gourmand. Facilement personnalisable, ce dessert se consomme nature ou parfumé (thé vert, cardamome, badiane, pulpe de fruits...), mais également accompagné de fruits frais, de Sauces Desserts NESTLE Docello®, de coulis de fruit...				

Version du 05/07/2018

Les dernières mises à jour sont disponibles sur notre site internet www.nestleprofessional.fr

Page 1 / 2



**Panna Cotta
NESTLÉ Docelio®
Etui de 600 g (2 x 300g)
pour 50 portions**

Nestlé
docelio®

CODE EAN
9002100034771

DÉCLARATION NUTRITIONNELLE

	Pour 100 g	Par portion (90 g)*
Énergie	1563 kJ 368 kcal	811 kJ 195 kcal
Matières grasses	0 g	14 g
- dont acides gras saturés	0 g	9,0 g
Glucides	91 g	14 g
- dont sucres	86 g	14 g
Fibres alimentaires	2,1 g	0,3 g
Protéines	0 g	2,2 g
Sel	0,87 g	0,19 g

Ce produit entre dans la catégorie GEMRCN** des desserts à limiter à hauteur de 3/20 repas maximum.

3/20 repas maximum

**Groupe d'Étude des Marchés Restauration Collective et Nutrition

*préparée avec du lait demi-écrémé et de la crème liquide (34% de matières grasses)

AVANTAGES ET BENEFICES DU PRODUIT

Crémeux et fondant, ce dessert est facile et rapide à mettre en œuvre.

CONSERVATION - STOCKAGE

Durabilité minimale : 24 mois

À conserver au sec et à l'abri de la chaleur dans son emballage d'origine. Bien refermer après chaque utilisation.

DONNÉES LOGISTIQUES

	Type UC / UD	Code EAN	Poids Net	Poids Brut	Dimensions (L x l x H) en mm
Unité consommateur (UC)	Boîte	9002100034771	600 g	691 g	80 x 185 x 200
Unité de distribution (UD)	Carton	9002100034788	3,6 kg	4,49 kg	388 x 253 x 216
Palette - Gerbabilité : OUI	Palette	7613033074578	129,6 kg	187 kg	1200 x 800 x 1014

Codes internes Nestlé			Code Douanier	Pays de production	Nbre UC par UD	Nbre UD par Couche	Couches par Palette	Nbre UD par Palette	Nbre UC par Palette
M74C/09	12202814	43825851	2106909855	Serbie	6	9	4	36	216

Nestlé France S.A.S. Noisiel 542 014 428 RCS MEAUX, NOISIEL © Reg. Trademark of Société des Produits Nestlé S.A.



Nestlé
PROFESSIONAL

Créateur de Solutions Culinaires & Boissons

Version du 05/07/2018

Les dernières mises à jour sont disponibles sur notre site internet www.nestleprofessional.fr

Page 2 / 2

B.1.4 Fiche technique confiture Andros



Route de Oinville
28 700 AUNEAU
Tél : + 33 2 37 33 16 33 - Fax : + 33 2 37 33 16 91



	05/01/2017
	-

DENOMINATION COMMERCIALE		
Colis assorti de 60 bocaux de confitures et marmelade 30g Bonne Maman		
15 Confiture Myrtilles et Cassis + 15 Confiture Fraises et Groseilles + 15 Confiture Abricots et Pêches + 15 Marmelade Oranges et Mandarines		
DENOMINATION LEGALE DE VENTE		
Confiture de Myrtilles et de Cassis - Confiture de Fraises et de Groseilles - Confiture d'Abricots et de Pêches - Marmelade d'Oranges et de Mandarines		
CARACTERISTIQUES		
Variété(s) :	myrtille cassis, fraises groseilles, abricots pêches, oranges mandarines	
Conservation :	A conserver au frais après ouverture.	Date de durabilité minimale : Voir sur unité de vente
Gencod :	3 60858 082942 3	Code article : 50082942
Codes douaniers :	Confiture Myrtilles et Cassis: 20079939 Confiture Fraises et Groseilles: 20079933 Confiture Abricots et Pêches: 20079939 Marmelade Oranges et Mandarines: 2007993110	
Confiture de myrtilles et de cassis <u>Ingrediénts</u> : fruits (myrtilles 41%, cassis 9%), sucre, sucre roux de canne, jus de citrons concentré, gélifiant : pectine de fruits.		
Confiture de fraises et de groseilles <u>Ingrediénts</u> : fruits (fraises 27 %, groseilles 23 %), sucre, sucre roux de canne, jus de citrons concentré, gélifiant : pectine de fruits.		
Confiture d'abricots et de pêches fruits (abricots 34%, pêches 16%), sucre, sucre roux de canne, jus de citrons concentré, gélifiant : pectine de fruits. Malgré tous nos soins, cette confiture peut contenir des noyaux.		
Marmelade d'oranges douces et de mandarines <u>Ingrediénts</u> : fruits (oranges douces 37%, mandarines 3%), sucre, sucre roux de canne, jus de citrons concentré, gélifiant : pectine de fruits.		
ARGUMENTAIRE PRODUIT Les confitures Bonne Maman duo, des alliances subtiles où le deuxième fruit vient rhabasser la saveur et la gourmandise du premier fruit		



VALEURS NUTRITIONNELLES MOYENNES													
		Myrtilles Cassis			Fraises Groseilles			Abricots Pêches			Oranges Mandarines		
		pour 100 g	pour 30 g	% AR* par ration	pour 100 g	pour 30 g	% AR* par ration	pour 100 g	pour 30 g	% AR* par ration	pour 100 g	pour 30 g	% AR* par ration
ENERGIE	En kJoules (kJ) :	1016	305	4%	1018	305	4%	1031	309	4%	1022	307	4%
	En kcalories (kcal) :	239	72	4%	240	72	4%	243	73	4%	241	72	4%
Matières grasses (g) :		0,2	0	0%	0,3	0	0%	0,2	0	0%	0,1	0	0%
Dont Acides Gras saturés		0	0	0%	0	0	0%	0	0	0%	0	0	0%
Glucides (g) :		58	17	7%	58	17	7%	59	18	7%	59	18	7%
Dont sucres:		58	17	19%	58	17	19%	59	18	20%	59	18	20%
Fibres (g):		1,8	1	—	1,5	0	—	1,3	0	—	1	0	—
Protéines (g):		0,5	0	0,3%	0,5	0	1,0%	0,6	0	0,4%	0,4	0	0,2%
Sel (g):		0	0	0%	0	0	0%	0	0	0%	0	0	0

* Apports de Référence (Apports de référence pour un adulte type (8400kJ/2000kcal). Ces valeurs et les portions peuvent varier selon l'âge, le sexe et l'activité physique

CARACTERISTIQUES PHYSICO - CHIMIQUES et ORGANOLEPTIQUES				
Parfums	Myrtilles Cassis	Fraises Groseilles	Abricots Pêches	Oranges Mandarines
Texture	onctueuse			gélifiée
Brix	60 ° brix			
Goût	Caractéristique des fruits			
Couleur	noire	rouge vif	orange	orange
pH	3,00 à 3,4	3,00 à 3,15	3,20 à 3,40	3,00 à 3,15

MENTIONS COMPLEMENTAIRES				
Allergène(s) obligatoire(s) présent(s) :	NON			
Allergène(s) obligatoire(s) pouvant être introduit(s) de manière non intentionnelle :				
Présence OGM, ionisation :				
Les produits ANDROS sont exempts :				
- d'Organisme Génétiquement Modifié et / ou d'ingrédients provenant d'Organisme Génétiquement Modifié				
- d'ingrédient ionisé y compris dans les matières premières et les ingrédients utilisés.				

RECOMMANDATIONS				
GEMRCN	—	Programme 1 fruit pour la récré		

CONDITIONNEMENT et PALETTISATION				
Bocal verre 44 ml / Capsule métallique TO 48 BM				
		Unité	COLIS	PALETTE
Format :	1 colis assorti de 60 pots de 30 g			
GENCOD :	3 60858 082942 3			03 60858 814419 1
Poids net (Kg) :	1,800			280,800
Poids brut (Kg) :	4,843			778,000
Largeur (mm) :	236			800
Longueur (mm) :	146			1200
Hauteur (mm) :	219			1459
Nombre de pièces :				156
Nombre de colis :				156
Nombre de couches :				6
Nombre de colis par couche :				26

COMMENTAIRES :				
En cas d'incohérence entre la fiche technique et l'emballage, veuillez noter que seul l'emballage fait foi.				

B.1.5 Fiche technique ciboulette La case aux épices

 LA CASE AUX ÉPICES	FICHE TECHNIQUE CIBOULETTE TUBULAIRE FLAPPERS	117801 Indice : f Date : 20/11/2014 Page : 1 / 2
DESCRIPTION DU PRODUIT		
 CARTONS de 8 Flapper's	Liste des ingrédients Ciboulette	
Caractéristiques organoleptiques La Ciboulette tubulaire est constituée de feuilles séchées de l' <i>Allium schoenoprasum</i> .		
Dénomination légale Ciboulette tubulaire		
Conservation : DDM (Date de Durabilité Minimale) DDM : 36 mois DDM minimum à réception : 18 mois		Commentaires Origine : Chine
Conditions de stockage A conserver dans un endroit frais et sec		
CARACTÉRISTIQUES PHYSICO-CHIMIQUES		VALEURS NUTRITIONNELLES MOYENNES pour 100g
Humidité :	max 8 %	Valeur énergétique : soit
aW :	-	- kcal - kJ
Cendres totales :	max 13 %	Matières grasses : dont acides gras saturés :
Cendres ins. dans l'acide :	max 2 %	Glucides : dont sucre : Protéines : Sel : Fibres alimentaires :
Huiles essentielles :	- mL/100g	- g - g - g - g - g
Commentaires Physico-chimie		
CARACTÉRISTIQUES MICROBIOLOGIQUES		Critères
Flore totale / g		-
Entérobactéries / g		1 000
Coliformes totaux / g		-
Escherichia coli / g		100
Salmonella		absence dans 25g
Staphylocoques à coagulase positive / g		100
Listeria monocytogenes		absence dans 1g
Anaérobies sulfito-réducteurs / g		-
Clostridium perfringens / g		1 000
Bacillus cereus / g		1 000
Levures /g		-
Moisiures /g		-
Commentaires Microbiologie		
Epicez, saucez, dosez !		

Epicez, sauez, dosez !





FICHE TECHNIQUE
CIBOULETTE TUBULAI RE FLAPPERS

117801
Indice : f
Date : 20/11/2014
Page : 2 / 2

CONDITIONNEMENT	Emballage primaire	Emballage secondaire	Emballage tertiaire	Palette
Poids net	60 g	0.48 kg	- kg	47.04 kg
Poids brut	123 g	1.184 kg	- kg	139.03 kg
Longueur	- mm	315 mm	- cm	120 cm
Largeur	- mm	160 mm	- cm	80 cm
Hauteur	215 mm	220 mm	- cm	175 cm
Materiel	-	Carton	-	-
Marquage	-	-	-	-
GENCOD	3344540027736	3344540026111	-	-
Nb. emballages primaires	-	8	-	-
Plan de palettisation	Nb colis par rangée : 14	Nb rangées par palette : 7	Nb total colis par palette :	98

ALLERGENES (selon la directive 2007/68/CE)

Présence

All1	Céréales contenant du gluten (à savoir blé, seigle, orge, avoine, épeautre, Kamut ou leurs souches hybrides) et Produits à base de ces Céréales.	Non
All2	Poissons et Produits à base de Poissons.	Non
All3	Crustacés et Produits à base de Crustacés.	Non
All4	Oeufs et Produits à base d'Oeufs.	Non
All5	Arachides et Produits à base d'Arachides.	Non
All6	Soja et Produits à base de Soja.	Non
All7	Lait et Produits à base de lait (y compris lactose)	Non
All8	Fruits à coque, à savoir amandes, noisettes, noix, noix de cajou, noix de pécan, noix du Brésil, pistaches, noix de macadamia et noix du Queensland et Produits à base de ces fruits.	Non
All9	Graines de sésame et Produit à base de graines de sésame.	Non
All10	Céleri et Produits à base de céleri.	Non
All11	Moutarde et Produits à base de moutarde.	Non
All12	Anhydride sulfureux et sulfites en concentrations de plus de 10 mg/kg ou 10 mg/l exprimées en SO2.	Non
All13	Lupin et Produits à base de lupin.	Non
All14	Mollusques et produits à base de mollusques.	Non

Commentaires Allergènes

ATTESTATIONS

Conformément à la réglementation européenne sur les OGM (règlements CE n°1829/2003 et 1830/2003), ce produit est un ingrédient conventionnel et ne nécessite donc pas d'étiquetage spécifique.

Conformément à la directive 1999/2/CE, ce produit n'est pas ionisé et ne contient aucun ingrédient soumis à ionisation.

Cette fiche technique est conforme au règlement UE 1169/2011.

CERTIFICATIONS



Une gamme complète de sauces et d'épices pour les professionnels de la restauration...



B.1.6 Fiche technique poivron El Arenal

CONSERVAS EL ARENAL S.L.	FICHES TECHNIQUES POIVRONS ROUGES ENTIERS	<i>Date création: 10/02/2015 Date révision: 10/02/2015 Version: 1</i>								
Nom du produit	POIVRONS ROUGES ENTIERS	FORMAT								
DESCRIPTION		<i>PRODUIT CONSISTANT DE POIVRONS ROUGES ENTIERS DE LA VARIETE CAPSICUM ANNUM L, PELÉS DEPOURVUS DU CALICE ET TIGE ET PRACTIQUEMENT DE GRAINES CONSERVÉS DANS UNE SAUMURE COMPOSÉE D'EAU, DE SEL, ET D'ACIDE CITRIQUE DANS DES BOÎTES HERMETIQUES ET PASTEURISÉES JUSQU'A OBTENIR LEUR STERILITÉ COMMERCIALE</i>								
INGRÉDIENTS		<table border="1"> <tr> <td>POIVRONS ROUGES</td><td>62,50%</td></tr> <tr> <td>EAU</td><td>36,96%</td></tr> <tr> <td>SEL</td><td>0,34%</td></tr> <tr> <td>ACIDE CITRIQUE</td><td>0,20%</td></tr> </table>	POIVRONS ROUGES	62,50%	EAU	36,96%	SEL	0,34%	ACIDE CITRIQUE	0,20%
POIVRONS ROUGES	62,50%									
EAU	36,96%									
SEL	0,34%									
ACIDE CITRIQUE	0,20%									
DURÉE DE CONSERVATION	<i>LA DATE LIMITE DE CONSOMMATION EST DETERMINÉE PAR LA MANIPULATION ET LES CONDITIONS DE TRANSPORT ET DE STOCKAGE. DANS DES CONDITIONS NORMALES (TEMPÉRATURE ENTRE 10-25 °C ET HUMIDITÉ INFÉRIEURE À 75 %) ELLE EST DE 36 MOIS</i>									
CONDITIONS DE STOCKAGE ET RECOMMANDATIONS D'UTILISATION	<i>Protéger de la lumière, et garder à température ambiante, une fois ouvert, garder au frais dans un récipient non métallique à une température entre 1 - 3° C, pendant une période maximum de 3 jour.</i>									
INFORMATION DIÉTÉTIQUE	<i>Produit sans conservant ni colorants. Produit apte pour les végétariens</i>									
SPECIFICATIONS DE FERMETURES	<table border="1"> <tr> <td>Pourcentage minimum d'embrochement</td><td>45%</td></tr> <tr> <td>Pourcentage minimum de compacité</td><td>80%</td></tr> </table>		Pourcentage minimum d'embrochement	45%	Pourcentage minimum de compacité	80%				
Pourcentage minimum d'embrochement	45%									
Pourcentage minimum de compacité	80%									

CONSERVAS EL ARENAL S.L.	FICHES TECHNIQUES POIVRONS ROUGES ENTIERS	<i>Date création: 10/02/2015</i> <i>Date révision: 10/02/2015</i> <i>Version: 1</i>
---------------------------------	--	---

DESTINATION	Consommation humaine en général
--------------------	---------------------------------

NORMES DE QUALITÉ	CONTENU	1/2 KG
	POIDS NET (g)	400
	POIDS ÉGOUTTÉ (g)	250
	ESPACE DE TÊTE (MM)	< 15
	VACUUM (CM Hg)	> 10
	PH	4.0 ± 0.3
	ODEUR	TYPIQUE
	COULEUR	BRILLANT ROUGE FONCÉ
	TEXTURE	TYPIQUE
	SAVEUR	TYPIQUE

VALEURS NUTRITIONNELLES <i>En accord avec le règlement CE 1169/2011</i>	Energie	109 kJ
	Energie	26 Kcal
	Graisses	0,2 G
	Dont saturées	0 G
	Carbohydrates	4,5 G
	Dont sucre	2 G
	Fibre	1,5 G
	Protéines	0,8 G
	Sel	0,4 G

MINERAUX/VITAMINES <i>(valeur moyenne/100g poids égoutté)</i>	SODIUM (mg)	160
	CALCIUM mg)	20
	FER (mg)	0,4
	VIT. A (IU)	520
	VIT.C (mg)	45

CONSERVAS EL ARENAL S.L.	FICHES TECHNIQUES POIVRONS ROUGES ENTIERS	<i>Date création: 10/02/2015 Date révision: 10/02/2015 Version: 1</i>
---------------------------------	--	---

PARAMÉTRES MICROBIOLOGIQUES	INCUBATION A 37°C PENDANT 7 JOURS (AFNOR V08-408)	SANS ALTERATION
------------------------------------	---	-----------------

INFORMATION ALLERGÉNIQUE	Liste des composants qui causent des allergies ou une hypersensibilité (CE Régulation 1169/2011. Annexe II)	<i>Présent dans le produit</i>	
		<i>oui / Non</i>	<i>Ingrédients</i>
	Céréales contenant du gluten et produits à base de céréales contenant du gluten (blé, maïs, seigle, orge, avoine, épeautre, Canut ou de leurs hybrides)	NON	
	Crustacés et fruits de mer	NON	
	Les oeuf et produits dérivés	NON	
	Les poissons et produits dérivés	NON	
	Cacahuète et produits dérivés	NON	
	Soja et produits à base de soja, à l'exception de l'huile et de graisse de soja entièrement raffinées	NON	
	Lait, produits laitiers et dérivés (inclus la lactose et les protéines de lait)	NON	
	Noix: amandes, noisettes, noix, noix de cajou, noix de pécan, noix du Brésil, noix de pistaches, noix de macadamia, et produits dérivés	NON	
	Céleri et produit dérivés	NON	
	Moutarde et produits dérivés	NON	
	Graines de sésame et produit dérivés	NON	
	Le dioxyde de soufre et sulfites en concentrations de plus de 10 mg / kg ou 10 mg / litre en termes de total SO	NON	
	Lupin et produits dérivés	NON	
	Molusques et produits dérivés	NON	

CONFORMITÉ	Ne contient pas et n'a pas été fabriqué avec des produits GMO
	Conforme à la législation pour les pesticides
	Conforme avec la législation de résidus de métaux
	Conforme avec la législation de traçabilité Règlement CE 178/2002) et HACCP's (Règlement CE 852/2004).

B.1.7 Fiche technique mélange trappeur Terre Exotique



Date d'impression : 30/12/19

Création : 12/06/12

Remarque :

Les informations contenues dans cette fiche technique sont données de bonne foi, en l'état actuel de nos connaissances, et selon les indications communiquées par le producteur ou le fournisseur. Il appartient au client de vérifier la conformité de la marchandise par rapport à l'usage qu'il en fait.

FICHE TECHNIQUE Mélange du trappeur, 70 g

Trapper blend, 70g

Code article KEREX / KEREX Code	TEEPTRAPPEUR	Poids net / net weight	0,07 Kilogramme
Nom latin (si disponible) / (Latin name)	X	Poids brut / gross weight	0,125 Kilogramme
Code barre / EAN Code	3760063322262	Origine / Origin	CANADA

Informations générales / General information

DLUO conseillé / "Best before date" recommended	5 ans / 5 years
Nomenclature douanière / Customs code	0910999900
Conditions idéales de stockage / Conditions of storage	Conserver dans un endroit frais et sec Store in a cool dry place
Ingrédients : / Ingredients	Sucre, poivre noir, coriandre, légumes déshydratés (ail, oignon, poivron rouge), sel de mer, sucre d'érythritol, arôme d'érythritol naturel, huile végétale (canola) Sugar, black pepper, coriander, dehydrated vegetables (garlic, onion, red bell pepper), sea salt, maple sugar, natural maple aroma, vegetable oil (canola)

Contaminants / Contaminating

Ionisation / Irradiation	Conformité à la directive 1999/2/CE (22/02/99) Produit non ionisé et ne contenant pas d'ingrédients ionisés. Not irradiated accordingly with the Reg 1999/2/CE (22/02/99).		
OGM / GMO	Free from GMO Ne contient pas d'OGM, est non soumis à l'étiquetage sur les OGM		
Pesticides/ Pesticides	Conforme à la directive 396/2005 /CE In accordance with Reg 396/2005 /CE.		
Métaux Lourds / Heavy Metals	Conforme au règlement 1881/2006 /CE In accordance with Reg 1881/2006 /CE..		
Allergènes et leurs dérivés (si présents) / Allergens (if existing)	Gluten	/ Gluten	Absence
	Crustacés	/ Crustaceans	Absence
	Oeufs	/ Eggs	Absence
	Poisson	/ Fish	Absence
	Soja	/ Soy	Absence
	Lait	/ Milk	Absence
	Fruits à coque - Arachides	/ Peanuts and Treenuts	Absence
	Céleri	/ Celery and celeriac	Absence
	Moutarde	/ Mustarde	Absence
	Sésame	/ Sésame	Absence
	Sulfites	/ Sulphites	Absence
	Lupin	/ Lupin	Absence
	Mollusques	/ Shellfish	Absence

Caractères microbiologiques / Microbiological characteristics

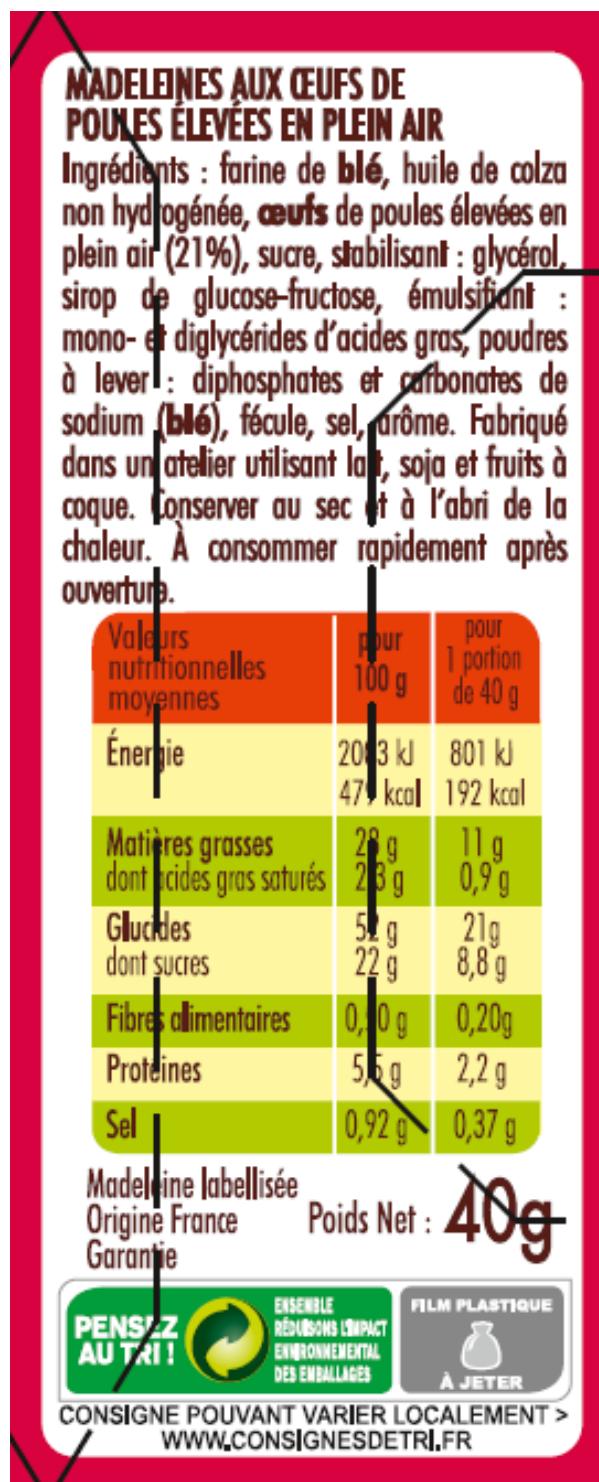
Microorganismes aérobies 30 °C	/ Total plat count (APC)	NF V05-051 < 6 000 000 / g
Escherichia coli	/ E. Coli	NF V08-053 < 10 / g
Salmonelles	/ Salmonella	NF V08-052 Absence dans 25g
Levures	/ Yeasts	NF V08-059 < 10 000 / g
Moisiures	/ Moulds	NF V08-059 < 10 000 / g
Aflatoxine Total	/ Total aflatoxin	Kit Enzymatique < 10 ppb
Aflatoxine B1	/ B1 aflatoxin	Kit Enzymatique < 5 ppb

B.2 Étiquettes produit

B.2.1 Étiquette curry Grain d'ailleurs



B.2.2 Étiquette madeleines Saint Michel

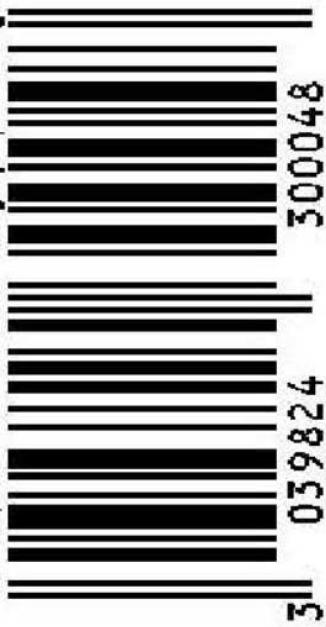


B.2.3 Étiquette lentilles Soufflet

**LENTELLER LINSEN 4mm
GREEN LENTILS 4mm
LINZEN 4mm**

Trace possible de céréales contenant du gluten - Possible traces of gluten containing cereals

Mögliche Spuren von Getreide, die Gluten enthalten - Mogelijk spoor van graangewassen met gluten



5 KG

VP

B.2.4 Étiquette pannacotta Nestlé



B.2.5 Étiquette sauce soja Kikkoman



B.2.6 Étiquette mélange trappeur Terre Exotique



B.3 Étiquetage manuel des données

B.3.1 Règles de gestion pour l'étiquetage

#	REGLES DE GESTION
REGLES GENERALES SUR L'INTERPRETATION DU DOCUMENT	
020	Si la liste d'ingrédients n'est pas présente sur le document, on laisse le champ vide. Ex : 21233a00-bc20-40fc-acb9-ee2e2321cac2
026	Si un document est visiblement corrompu (ex : ne contient rien), on le garde dans le jeu de test, avec aucune liste d'ingrédients en cible. Ex : 8e8bbc12-7e4b-4fff-a111-f7c84e35129a
027	En règle générale, on ne récupère que ce qui est mentionné dans le bloc 'Ingrédients' ou 'Composition' de la fiche technique. Et ce, même si on aurait gardé le texte d'un autre bloc s'il avait été avec le texte de la liste d'ingrédients. Ex : Composition typique (Données inappropriées pour la demande des restitutions) pâte de cacao Tanzanie 69,5% ; sucre 21,5% ; beurre de cacao 9,0% ; vanille naturelle en poudre <1% Contact croisé d'allergènes possible au cours du process Peut contenir : Lait => on ne garde que : pâte de cacao Tanzanie 69,5% ; sucre 21,5% ; beurre de cacao 9,0% ; vanille naturelle en poudre <1% Ingrédients: SEMOULE DE BLE' DUR de qualité supérieure Origine des matières premières Blé cultivé en UE et Hors UE, moulu en Italie => on ne garde que : SEMOULE DE BLE' DUR de qualité supérieure
028	Si la fiche ne contient qu'un tableau récapitulatif des ingrédients, et pas la liste des ingrédients telle qu'elle doit être imprimée sur l'étiquette, on ne récupère rien. Ex : bc48d583-edf0-460e-a536-58f347ba6823
032	Si la liste d'ingrédients ne concerne qu'un seul composant et qu'elle est préfixée par le nom du produit, on ne garde que la liste d'ingrédients en tant que telle. Ex : CARACTERISTIQUES TECHNIQUES Ingrédients Sauce tomate, fumée et pimentée :Vinaigre d'alcool, eau, purée de de tomates concentré (17%), sucre, purée de datte, pâte de piment Chipotle (5%, purée de tomates, piments Jalapeno fumé, eau, oignon, vinaigre d'acétol, sucre, poivres, sel, ail, épices, persil, extrait d'épice), amidon modifié, sel, jus de pomme concentré, poudre de Chili, arôme de fumée, épice, poudre d'ail, poudre d'oignon, conservateur (sorbate de potassium), extrait d'épice. On ne garde que : Vinaigre d'alcool, eau, purée de de tomates concentré (17%), ...
039	Si la liste d'ingrédient mentionne une interprétation de la liste (un synonyme plus parlant), on ne reprend pas cette mention. Ex : 100% tourteau de cacao maigre = poudre de cacao pure Peut contenir des traces d'arachide et de lait devient : 100% tourteau de cacao maigre Peut contenir des traces d'arachide et de lait
LANGUES	
029	Si la fiche technique est en anglais, on ne récupère rien. Ex : 45aaafdde-9ca7-42ab-973f-70d6fa402e17
037	Si la liste d'ingrédients est présente en plusieurs langues, on ne garde que le français. Y compris si les textes sont "mêlés". Ex : Ingrédients - Ingredients: semoule de blé dur de qualité supérieure, issu de la filière Alpina durum wheat premium semolina from Alpina durum wheat channel Peut contenir des traces d'œufs - May contain traces of eggs devient : semoule de blé dur de qualité supérieure, issu de la filière Alpina Peut contenir des traces d'œufs
MISE EN PAGE ET FORMAT	
001	On conserve les retours à la ligne qui correspondent à un retour chariot volontaire. On garde sur une unique ligne les longues suites de mots qui sont visuellement retournées à la ligne parce qu'on arrive "au bout de la zone". Ex : Aubergine 60,5% (aubergine, huile de tournesol), eau, oignon, huile de tournesol, jus de citron, concentré de tomate, huile d'olive vierge extra (2%), ail, sel, persil, basilic, poivre, thym, romarin. devient : Aubergine 60,5% (aubergine, huile de tournesol), eau, oignon, huile de tournesol, jus de citron, concentré de tomate, huile d'olive vierge extra (2%), ail, sel, persil, basilic, poivre, thym, romarin.
002	Si des tirets, des puces ou des listes numérotées sont présentes, on garde le marqueur de début de ligne tel quel. Ex : - 100% Semoule de BLE dur de qualité supérieure - Contient du gluten Si le numéro de lot contient la lettre N : peu contenir de l'œuf
003	La casse et les accents sont conservés tels quels. Si le copier / coller depuis pdf ajoute des espaces ou autres artefacts, on les corrige pour avoir un texte qui correspond à ce qui est lu par un utilisateur Ex : A u b e r g i n e 6 0 , 5 % (aubergine, huile de tournesol), ... devient: Aubergine 60,5% (aubergine, huile de tournesol),
005	La ponctuation est conservée telle quelle, en particulier on garde le point final s'il est présent.
008	On garde les fautes d'orthographe
010	On remplace l'accent aigu sans lettre ' par une apostrophe normale : '

	011	Remplacer l'e dans l'o (œ) par oe
	012	Pas de multiples retours à la ligne successifs (même s'ils apparaissent à la lecture)
014		<p>Si les ingrédients sont séparés par des retours chariot, on conserve cette mise en forme. Ex : eau sucre étrait de thé noir (1,4g/l) acidifiant (acide citrique) jus de citron à base de concentré (0,1%) arôme correcteur d'acidité (citrate trisodique) antioxydant (acide ascorbique)</p>
		PREFIXES A LA LISTE
036	006	<p>On ne garde pas les préfixes du type "Ingrédients :" ou "Composition :"</p> <p>Si le bloc d'ingrédient fait référence au règlement INCO, on ne conserve pas cette mention. Ex : Déclaration d'étiquetage Selon le règlement 1169/2011 UE : ananas, jus d'ananas, antioxydant : acide ascorbique, acidifiant : acide citrique</p> <p>On ne garde que : ananas, jus d'ananas, antioxydant : acide ascorbique, acidifiant : acide citrique</p>
		MENTION DES ALLERGENES
007	007	<p>On garde les précisions relatives aux traces d'allergènes, aux risques de contamination croisées, ... Ex : Fabriqué dans une usine utilisant des FRUITS à COQUE, SOJA et ARACHIDES => On garde.</p>
017b		<p>Si une note de bas de page précise la présence d'un composant dans un ingrédient, on le mentionne. Ex : Eau, huile de colza, vinaigre d'alcool, MOUTARDE de Dijon (eau, graines de MOUTARDE, vinaigre d'alcool, sel, correcteur d'acidité : E330, conservateur : E224*), sel, sucre*, acidifiant : E330, épice, stabilisants : E412 et E415. *Contient : SULFITES. <=> On garde cette ligne</p>
035	008	<p>On ne garde pas de mention précisant l'absence d'allergènes : Ex : Filets de maquereaux scomber scombrus 60% (Pêchés en Atlantique Nord-Est), eau, concentré de tomates (30% de la sauce), huile de colza, vinaigre, sel, épaississant : gomme xanthane, arômes naturels. Allergènes majeurs (hors poisson) : absence. ==> on ne garde pas cette mention</p>
035b	009	<p>On conserve les mentions précisant la présence d'un allergène si jamais la liste d'ingrédients ne permet pas de le déduire (sulfites uniquement ?) Ex : Huile de colza, à Aceto Balsamico di Modena IGP » (25%) (vinaigre de vin, moût de raisin cuit, colorant E-150d), huile d'olive vierge extra. Contient sulfites. <=> on garde cette ligne.</p>
		PRECISIONS SUR LA COMPOSITION - PRESENCE DE COMPOSANTS SPECIFIQUES HORS ALLERGENES
017	017	<p>On mentionne les composants et les recommandations qui peuvent être présentes. Ex : Maltodextrine, carraghénane E407a, amidon de blé, colorant : jus concentré de betterave E162, édulcorant : aspartame E951* (0,37%), arôme goût framboise (substances aromatisantes, substances aromatisantes naturelles). *Contient une source de phénylalanine, ne convient pas aux femmes enceintes. ==> On garde cette ligne</p>
		PRECISIONS SUR LA COMPOSITION - PROPORTIONS
009	009	<p>On garde les précisions relatives au détail de la composition (en particulier le taux de cacao) Ex : Taux de coumarine compris entre 1 et 3,5 % ==> on garde ----- sucre, pâte de cacao, beurre de cacao, émulsifiant : lécithine de tournesol (E322), arôme vanille cacao : 40%minimum ==> on garde ----- chocolat supérieur au lait 39% (sucre, lait en poudre, beurre de cacao, pâte de cacao, émulsifiants : lécithines [soja], vanilline), sucre, lait écrémé en poudre, huile de palme, beurre concentré, émulsifiants : lécithines [soja], vanilline. Sur le total : produits laitiers 33 % (lait écrémé en poudre, lait en poudre : 27,4 %, beurre concentré 5,6 %) - cacao 12,5 %. Le chocolat utilisé est un chocolat pur beurre de cacao. <=> On garde cette seconde partie ----- eau, sirop de glucose-fructose, sucre, épaississant : amidon modifié, graisse végétale non hydrogénée de noix de coco, jus concentré de citron 2%, gélifiant : pectines de fruits, arôme naturel de citron, émulsifiants E472C et E473, arôme, conservateur : sorbate de potassium, sel, colorant : E161b. Teneur en Lutéine : moins de 5 mg/kg <=> On garde cette partie</p>
009b	009b	<p>Y compris lorsqu'accompagnée de mise en garde particulière : TENEUR ÉLEVÉE EN CAFÉINE, DÉCONSEILLÉ AUX ENFANTS ET AUX FEMMES ENCEINTES OU ALLAITANTES OU AUX PERSONNES SENSIBLES À LA CAFÉINE (21mg/100ml). CONSOMMER AVEC MODÉRATION. <=> On garde cette ligne</p>
		MENTIONS CONDITIONNELLES
034	034	<p>Si une mention précise une condition pour la présence d'un ingrédient, on conserve cette mention. Ex : Pommes 21%, poires 39%, eau, sucre, arôme, antioxydant : acide ascorbique, acidifiant : acide citrique (si le fruit n'en contient pas suffisamment). On garde la mention : (si le fruit n'en contient pas suffisamment)</p>
034b	034b	<p>Si une mention précise une condition pour la présence d'une trace d'allergène, on conserve cette mention. Ex : - 100% Semoule de BLE dur de qualité supérieure - Contient du gluten Si le numéro de lot contient la lettre N : peu contenir de l'œuf ==> on conserve cette mention.</p>
034c	034c	<p>Si une mention précise une condition pour la présence d'additif, on conserve cette mention. Ex : INGRÉDIENTS : Segments de pamplemousse, eau, sucre, acidifiant : acide citrique*, agent de fermeté : chlorure de calcium*. * En fonction des origines, des additifs peuvent être ajoutés. <=> On garde cette mention</p>
009c	009c	<p>Si une note de bas de page précise la variabilité du taux dans la liste d'ingrédients, on conserve cette note. Ex : pulpe de pommes (50%)*, sucre, sirop de glucose, gélifiant : pectines, acidifiant : acide citrique, arômes, colorants : E100 - E163 - E160c - E141. * Pourcentage à la mise en œuvre pour chaque parfum. <=> On conserve cette note.</p>
		LISTES D'INGRÉDIENTS MULTIPLES POUR UNE MÊME FICHE TECHNIQUE

	Dans le cas d'assortiments, s'il y a plusieurs listes d'ingrédients successives, on les listes toutes en gardant les entêtes qui définissent le produit. Ex : Confiture de myrtilles et de cassis Ingrédients : fruits (myrtilles 41%, cassis 9%), sucre, sucre roux de canne, jus de citrons concentré, gélifiant : pectine de fruits. Confiture de fraises et de groseilles Ingrédients : fruits (fraises 27 %, groseilles 23 %), sucre, sucre roux de canne, jus de citrons concentré, gélifiant : pectine de fruits. Confiture d'abricots et de pêches fruits (abricots 34%, pêches 16%), sucre, sucre roux de canne, jus de citrons concentré, gélifiant : pectine de fruits. Malgré tous nos soins, cette confiture peut contenir des noyaux. Marmelade d'oranges douces et de mandarines Ingrédients : fruits (oranges douces 37%, mandarines 3%), sucre, sucre roux de canne, jus de citrons concentré, gélifiant : pectine de fruits.
013b	S'il existe plusieurs manières différentes de présenter le produit, chacune avec sa liste d'ingrédients, on conserve les 2 de la même manière que lorsqu'il y a plusieurs composants. Ex: Vinaigre de vin rouge au jus d'échalote 7% d'acidité Vinaigre de vin rouge (sulfites), jus d'échalote (0,5%), arôme*, conservateur: sulfite acide de sodium Vinaigre de vin rouge aromatisé à l'échalote 7% d'acidité Vinaigre de vin rouge (sulfites), arômes échalote (1%), conservateur: sulfite acide de sodium
040	Si une liste d'ingrédient possède une variante en fonction du conditionnement, on conserve les informations relatives à ces précisions. Ex : Eau de source*/Eau**; jus de fruits à base de concentrés 12.4% (orange 11.4%, ananas 1%); sucre; acidifiants: acide citrique, acide malique; extraits d'orange; arômes; antioxydant: acide ascorbique; stabilisant: gomme arabique. *PET **CAN => On conserve les 2 dernières lignes qui définissent pour quel type de conditionnement la variante s'applique.
	ORIGINES Si une mention de l'origine de transformation est faite, on NE récupère PAS. Ex : Sucre, amidon de FROMENT, farine de FROMENT, amidon modifié, fibres solubles, poudre à lever : (E450, E500, E341), poudre de LAIT écrémé, émulsifiants : (E471, E472e), amidon, farine de FROMENT pré gélatinisée, sel, arôme naturel. Fabriqué en France. => on ne garde pas le "Fabriqué en France"
015	Si une mention de l'origine des ingrédients est faite, on la récupère. Ex : Pomme 77% (origine 100% France), abricot 20% (origine 100% France), sucre (origine 100% France), jus concentré de carotte, antioxydant: acide ascorbique. Certains ingrédients de ce produit de proviennent pas de France. => On garde cette ligne.
016	En particulier, pour les produits qui contiennent du poisson, si la zone de pêche est présente on la récupère. Ex : Sardines (Sardina pilchardus) 70 %, huile de tournesol, sel. Zone de pêche : Océan Atlantique Centre Est => On conserve cette mention
016b	Si la liste d'ingrédients précise l'origine végétale ou animale des ingrédients, on conserve cette information. Ex : 100% thé noir Origine végétale
038	DETAIL PAR PHASE Si une partie de la recette est détaillée, on conserve également la description de cette sous-partie. Ex : 018 Légumes (pommes de terre (14%), carottes (10%), champignons (9%), haricots verts, persil), sauce, thon* (22%). Composition de la sauce : Eau, huile de tournesol, vinaigre, moutarde de Dijon (eau, graines de moutarde, vinaigre, sel), sel, arômes, amidons transformés de maïs, épaississants : gomme guar et gomme xanthane. => On garde la composition de la sauce
	PRECISION SUR L'IDENTIFICATION D'UN COMPOSANT Si une note de bas de paragraphe précise le nom latin du poisson, on conserve cette note de bas de page. Ex : 019 Légumes (pommes de terre (14%), carottes (10%), champignons (9%), haricots verts, persil), sauce, thon* (22%). *Euthynnus (Katsuwonus) pelamis
019'	Si le nom latin d'un ingrédient hors poisson est présent dans la liste des ingrédients, on le conserve. Ex : semoule de blé dur blanche biologique (Triticum durum) <= On conserve Triticum durum
	LABELS DES INGREDIENTS Si une note de bas de paragraphe précise les ingrédients d'origine biologique, on conserve cette mention. Ex : 019b Pur thé vert* (100%). (*) Ingrediénts issus de l'Agriculture Biologique => on garde cette mention
019b	Si la note de bas de page mentionne également les références du certificat bio, on NE garde PAS ces références. Ex : 019b' Semoule de blé dur précurte* (gluten) (44,5%), semoule de blé dur complète précurte* (gluten) (29,5%), flocons de soja* (15%), flocons de céréales* (gluten) (blé* (3%), orge*, avoine*, seigle* (2,2%), riz* (0,7%). * Ingrediénts issus de l'Agriculture Biologique => On garde cette mention Produit issu de l'agriculture biologique certifié bio par Ecocert FR-BIO 01 – N/N identification ECOCERT : 44/20349 => On NE garde PAS cette mention
019c	Si une note de bas de paragraphe précise un label sur un des ingrédients, on conserve cette mention. Ex : infusion de thé noir 94% (eau, infusion intense de thé noir*), sucre, jus de pêche à base de concentré 1%, arômes naturels de thé, acidifiants: acide malique et acide citrique, arômes naturels, correcteur d'acidité: citrate de sodium, antioxydant: acide ascorbique. *Vérifié Rainforest Alliance => on garde cette mention
	ALLEGATIONS ET RESERVES On NE mentionne PAS les réserves relatives au traitement physique. Ex : On NE garde PAS ceci Malgré tous les soins apportés à nos produits, leur caractère naturel n'exclut pas la présence éventuelle de noyaux, de pépins, de restes de peaux ou de parties fibreuses.
021	On NE mentionne PAS l'absence d'OGM ou d'ingrédients ionisés. Ex : On NE garde PAS ceci Le produit ne contient ni OGM, ni ingrédients ionisés.
022	

024	On NE récupère PAS d'information de composition qui ne sont pas spécifiques à un ingrédient. Ex : Préparée avec 35g de fruits pour 100g de produit fini. ==> on ne récupère pas
025	On NE récupère PAS d'information en lien avec les données nutritionnelles. Ex : Teneur totale en sucres : 60g pour 100g. ==> on ne récupère pas.
030	On NE conserve PAS les informations relatives au processus de fabrication ou de conditionnement. Ex : Conditionné sous atmosphère protectrice.
031	On NE conserve PAS les consignes de conservation. Ex : A conserver à l'abri de la chaleur et de l'humidité.
033	On ne conserve pas les précisions d'utilisation positionnées dans la liste d'ingrédients. Ex : Ingrediénts : Ingrédients : Colorant : Caramel de sulfite caustique E150b (contient des sulfites), eau. Pour denrées alimentaires. On ne garde pas "Pour denrées alimentaires."

Annexe C

LES NOTEBOOKS DE CE PROJET

,

Analyse quantitative

Pierre MASSÉ

April 22, 2020

1 Analyse quantitative multibranche

```
[1]: # data analysis
import pandas as pd
pd.options.display.width=108
import numpy as np

# visualization
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib_venn import venn3_unweighted, venn3
import matplotlib as mpl
# mpl.rcParams['text.usetex'] = True
# plt.rcParams['text.latex.preamble'] = [r'\usepackage{lmodern}']

# utils
from pathlib import Path
```

Définition des couleurs :

```
[2]: c_pomona = tuple(val / 255 for val in [0, 92, 132])
c_terreazur = tuple(val / 255 for val in [0, 152, 170])
c_episaveurs = tuple(val / 255 for val in [255, 69, 0])
c_passionfroid = tuple(val / 255 for val in [109, 32, 124])
c_deliceetcreation = tuple(val / 255 for val in [97, 45, 28])
c_saveursdantoine = tuple(val / 255 for val in [156, 34, 63])
```

On charge les données d'un fichier exporté du système de gestion des branches RHD (SAP).

```
[3]: path = Path('..') / 'data' / 'export2020.csv'

types = {
    'material': 'object',
    'branch': 'int',
    'plant': 'object',
    'type': 'object',
    'designation': 'object',
    'del_mand': 'bool',
    'del_plant': 'bool',
    'march_group': 'object',
    'storage_cond': 'object',
    'hier': 'object',
}
df = pd.read_csv(path,
                  sep=';',
                  encoding='latin-1',
                  engine='python',
                  header=0,
                  skipfooter=1, # footer line with totals in export
                  dtype=types,
                  true_values=['X'], # for del_mand and del_plant
                  false_values=['', np.nan], # for del_mand and del_plant
)
```

```
df = df[types.keys()] #filter and reorder columns
```

Parmi les colonnes conservées, on a : - le code article (material)

- le code de branche de création (branch).
 - 1: PassionFroid
 - 2: EpiSaveurs
 - 3: TerreAzur
- le code d'activation sur une branche (plant).
 - 1PPF: PassionFroid
 - 2PES: EpiSaveurs
 - 3PTA: TerreAzur
- le type d'article (type). Seuls ZNEG et ZPRE représententent des articles de marchandises.
 - ZNEG: Négoce
 - ZPRE: Prestation
 - ZENG: Article d'engagement (fictif pour facturation)
 - ZEMB: Article d'emballage (ex: palette)
 - ZSER: Article de service
- le libellé de l'article (designation)
- si l'article est marqué pour suppression pour toutes les branches (del_mand)
- si l'article est marqué pour suppression sur la branche mentionnée dans la colonne plant (del_plant).
- le groupe de marchandises (march_group) :
 - ZSURGE: Surgelés
 - ZFRAIS: Frais (PassionFroid)
 - ZEPI: Epicerie
 - ZBOI: Boissons
 - ZHYG: Hygiène et chimie
 - ZFLF: Fruits et légumes (TerreAzur)
 - ZPMF: Produits de la mer (TerreAzur)
 - ZFP: Fleurs et plantes
 - ZELAB: Produits élaborés (TerreAzur)
- la condition de stockage (storage_cond) :
 - FR: Frais (PassionFroid)
 - SU: Surgelé,
 - EP: Epicerie,
 - AL: Alcool
 - HY: Hygiène et chimie
 - FL: Fruits et légumes (TerreAzur)
 - FP: Fleurs et plantes
 - MA: Marée
 - SA: Saurisserie (produits élaborés de la mer)
 - SE: Articles de Service
 - PL: Articles de publicité
- la hiérarchie produit (hier). Un plan de classement sur 6 niveaux, représentés par 2 caractères numériques chacun.

On crée une nouvelle feature qui correspond au niveau 1 de la hiérarchie produit.

```
[4]: # Creation of first level of product hierarchy
df.loc[:, 'hier1'] = df.hier.str[:2]
```

On définit un dictionnaire permettant de rappeler les libellés long des divers codes présents dans le dataset.

```
[5]: # Label names
lab = {'type': 'Type de produit',
       'march_group': 'Groupe de marchandises',
       'storage_cond': 'Condition de stockage',
       'hier1': 'Niveau 1 hiérarchie produit',
       '1PPF': 'PassionFroid',
       '2PES': 'EpiSaveurs',
       '3PTA': 'TerreAzur',
       'ZNEG': 'Article de négoce',
       'ZPRE': 'Article de prestation',
       'ZSURGE': 'Surgelés',
       'ZFRAIS': 'Frais',
       'ZEPI': 'Epicerie',
       'ZBOI': 'Boissons',
       'ZHYG': 'Hygiène',
```

```

'ZFLF': 'Fruits et Légumes',
'ZPMF': 'Produits de la mer',
'ZELAB': 'Produits élaborés',
'ZFP': 'Fleurs et plantes',
'ZAUTRE': 'Autres',
'SU': 'Surgelés',
'FR': 'Frais',
'EP': 'Epicerie',
'AL': 'Alcool',
'HY': 'Hygiène',
'FL': 'Fruits et légumes',
'MA': 'Marée',
'FP': 'Fleurs et plantes',
'SA': 'Saurisserie',
'PL': 'Publicié',
'10': 'Beurre, oeufs, fromage',
'20': 'Elaborés',
'30': 'Garnitures et fruits',
'40': 'Produits carnés',
'50': 'Produits de la mer',
'60': 'Consommables',
'70': 'Emballage',
'80': 'Publicité sur le lieu de vente',
'83': 'Epicerie',
'85': 'Liquides',
'87': 'Hygiène et entretien',
'90': 'Services',
'92': 'Fruits',
'94': 'Légumes',
'96': 'Produits de la mer Frais',
'98': 'Fleurs - plantes',
}

```

```
[6]: df.loc[[5000, 90000, 100000, 130000, 110000], :]
```

	material	branch	plant	type	designation	del_mand	del_plant	\
5000	15712	2	2PES	ZNEG	PSVNX CERN BRISURE S/AZ SAC 1KGX12 CERNO	True	True	
90000	153086	3	3PTA	ZNEG	MANGUE KENT 351/550G PAD 12F DELIC BR°	False	False	
100000	165387	1	1PPF	ZNEG	SALADE PLT 1KGX12 HAMAL	False	False	
130000	203582	1	1PPF	ZPRE	EFFILOCHE BOEUF BARBACOA (2KGX6)/12KG CS	False	False	
110000	177238	2	2PES	ZNEG	COMP POIRE ALL BIO BTE 5/1X3 STM	False	False	
					hier hier1			
5000	ZEPI		EP	832020500505	83			
90000	ZFLF		FL	920518010405	92			
100000	ZFRAIS		FR	202520150505	20			
130000	ZSURGE		SU	401015051505	40			
110000	ZEPI		EP	832005451505	83			

On va définir deux masques, permettant de filtrer : - les articles actifs (i.e. non supprimé niveau mandant ni branche) - les articles actifs de marchandises (i.e. qui ne sont pas des articles "spéciaux")

```
[7]: active_mask = ~df.del_mand & ~df.del_plant
active_march_mask = active_mask & df.type.isin(['ZNEG', 'ZPRE'])
```

On peut calculer la volumétrie d'articles et la représenter comme un histogramme. Les données de Délice et Création et Saveurs d'Antoine sont issue d'estimations fournies par le métier.

```
[8]: counts = df.groupby('plant')['material'].count().rename('Total')
filtered_counts = df[active_mask].groupby('plant')['material'].count().rename('Actifs')
filtered_counts2 = df[active_march_mask].groupby('plant')['material'].count().rename('Marchandises')

report = pd.concat([counts, filtered_counts, filtered_counts2], axis=1)
report.loc['Délice et Création', :] = [10000, np.nan, np.nan]
report.loc['Saveurs d\'Antoine', :] = [12000, np.nan, np.nan]
report.rename({'1PPF': 'PassionFroid'},
```

```

        '2PES': 'EpiSaveurs',
        '3PTA': 'TerreAzur'},
    inplace=True)
report.index.rename('Branche', inplace=True)
report = report.astype('Int64')
report.to_latex(Path(..) / 'tbls' / 'Articles par branche.tex',
    bold_rows=True,
    column_format='lccc',
    na_rep='_'
)

```

report

[8]:

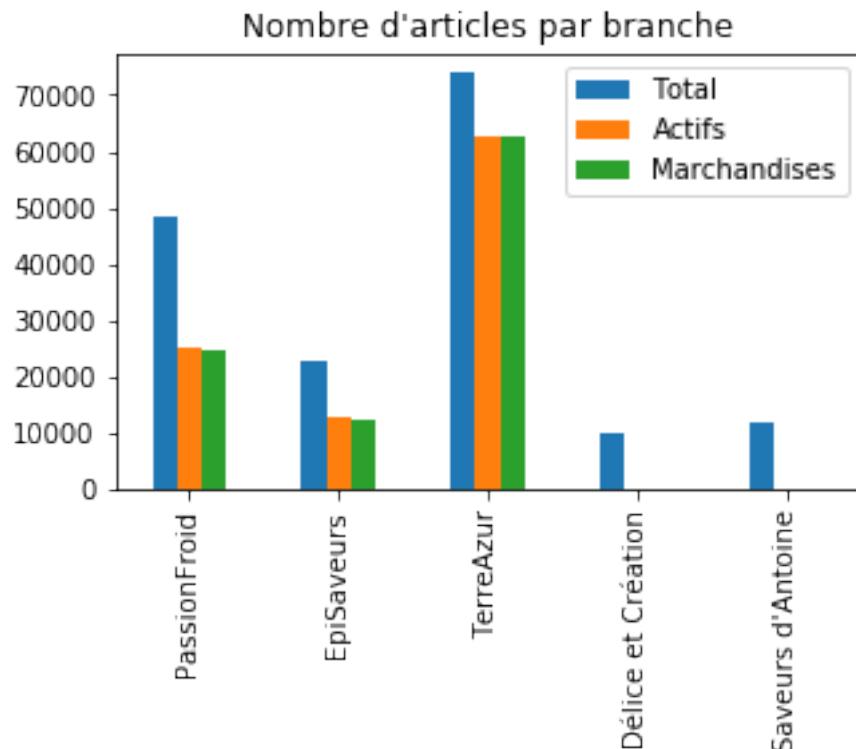
Branche	Total	Actifs	Marchandises
PassionFroid	48478	24898	24554
EpiSaveurs	22498	12798	12241
TerreAzur	73804	62789	62710
Délice et Création	10000	Nan	Nan
Saveurs d'Antoine	12000	Nan	Nan

[9]:

```

fig, ax = plt.subplots(figsize=(5,3))
report.plot(kind='bar', ax=ax)
ax.set_title('Nombre d\'articles par branche')
ax.set_xlabel('')
fig.savefig(Path(..) / 'img' / 'Articles par branche.png', bbox_inches='tight')

```



On peut également construire le diagramme de Venn des articles pour les branches RHD :

[10]:

```

# Filtering the dataset with active materials, and active merchandize materials
branch_sets = [set(df.loc[df.plant == branch_, 'material']) for branch_ in ['1PPF', '2PES', '3PTA']]
filtered_df = df.loc[active_mask]

```

```

filtered_sets = [set(filtered_df.loc[filtered_df.plant == branch_, 'material']) for branch_ in ['1PPF', '2PES', '3PTA']]

filtered_march_df = df.loc[active_march_mask]
filtered_march_sets = [set(filtered_march_df.loc[filtered_march_df.plant == branch_, 'material'])
                      for branch_ in ['1PPF', '2PES', '3PTA']]

```

```

[11]: # This function is used to add label on Venn diagrams axes without showing spines
# (matplotlib-venn disables totally axis's, and spines need to get erased after
# axis's reactivation)
def labelize(ax, label, where='bottom', **kwargs):
    ax.set_axis_on()
    for spine in ['top', 'bottom', 'left', 'right']:
        ax.spines[spine].set_visible(False)
    if where == 'bottom':
        ax.set_xlabel(label, **kwargs)
    elif where == 'left':
        ax.set_ylabel(label, **kwargs)
    else:
        raise ValueError(f"Unexpected 'where' argument: {where}")

```

```

[12]: # Construction of the diagrams
scope = ['Total', 'Actifs', 'Marchandises']
types = ['Non pondéré', 'Pondéré']
nrows, ncols = len(types), len(scope)

fig, axs = plt.subplots(nrows, ncols, sharex='col', sharey='row', figsize=(18, 8))

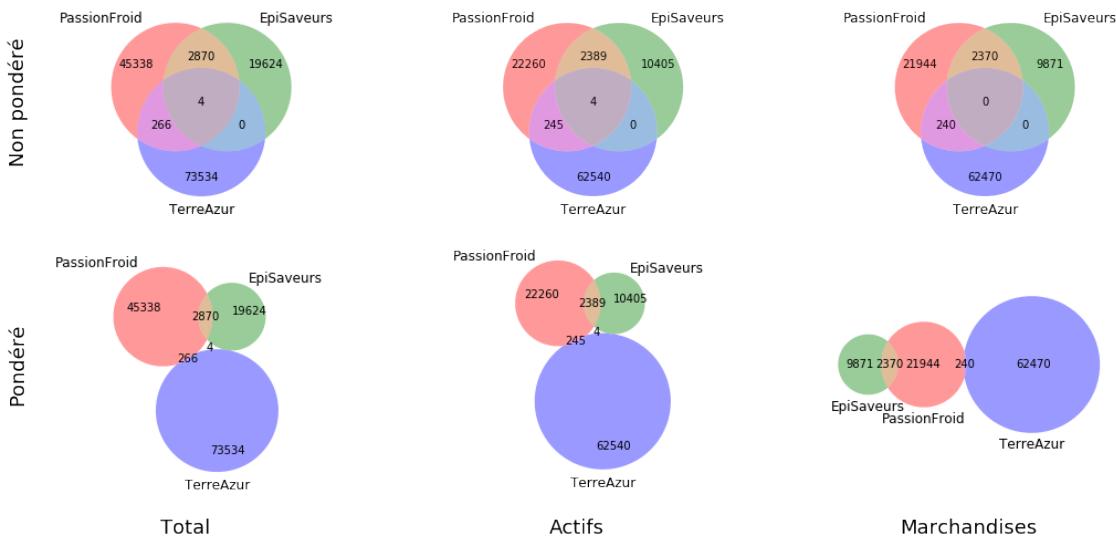
for col, source_df in enumerate([branch_sets, filtered_sets, filtered_march_sets]):
    for row, venn_kind in enumerate([venn3_unweighted, venn3]):
        venn_kind(source_df, set_labels=['PassionFroid', 'EpiSaveurs', 'TerreAzur'], ax=axs[row, col])
        if col == 0:
            labelize(axs[row, col], types[row], where='left', fontsize=18, labelpad=10)
        if row == 1:
            labelize(axs[row, col], scope[col], where='bottom', fontsize=18, labelpad=40)

# Adjusting the min and max of axes lims, as they are not the same by default
xmin = min([axs[row][col].get_xlim()[0] for row in range(nrows) for col in range(ncols)])
xmax = max([axs[row][col].get_xlim()[1] for row in range(nrows) for col in range(ncols)])
ymin = min([axs[row][col].get_ylim()[0] for row in range(nrows) for col in range(ncols)])
ymax = max([axs[row][col].get_ylim()[1] for row in range(nrows) for col in range(ncols)]) + 0.1

for row in range(nrows):
    for col in range(ncols):
        axs[row, col].set_xlim(xmin, xmax)
        axs[row, col].set_ylim(ymin, ymax)

# Saving the file to disk so that it is included in the report
fig.savefig(Path('..') / 'img' / 'Diagrammes de Venn articles.png', bbox_inches='tight')

```



On peut constater que les articles utilisés par les 3 branches RHD sont des articles "spéciaux".

```
[13]: df[df.material.isin(df.material.value_counts()[df.material.value_counts() >= 3].index)]
```

```
[13]:      material branch plant type designation del_mand \
144564    DECOMPTE     1   2PES ZSER ARTICLE DE DECOMPTE CONDITIONS ARRIERES False
144565    DECOMPTE     1   3PTA ZSER ARTICLE DE DECOMPTE CONDITIONS ARRIERES False
144566    DECOMPTE     1   1PPF ZSER ARTICLE DE DECOMPTE CONDITIONS ARRIERES False
144612    FC41849      1   1PPF ZSER RÉGUL SURFACTURATION DÉCONDITIONNEMENT False
144613    FC41849      1   2PES ZSER RÉGUL SURFACTURATION DÉCONDITIONNEMENT False
144614    FC41849      1   3PTA ZSER RÉGUL SURFACTURATION DÉCONDITIONNEMENT False
144642    LOT_ENGT      1   1PPF ZENG LOT ENGAGEMENT False
144643    LOT_ENGT      1   3PTA ZENG LOT ENGAGEMENT False
144644    LOT_ENGT      1   2PES ZENG LOT ENGAGEMENT False
144719 S_PALETTE_PERDUE 3   3PTA ZEMB PALETTE 80X120 PERDUE False
144720 S_PALETTE_PERDUE 3   2PES ZEMB PALETTE 80X120 PERDUE False
144721 S_PALETTE_PERDUE 3   1PPF ZEMB PALETTE 80X120 PERDUE False

      del_planet march_group storage_cond      hier hier1
144564    False     ZAUTRE      NaN 900505050505      90
144565    False     ZAUTRE      NaN 900505050505      90
144566    False     ZAUTRE      NaN 900505050505      90
144612    False     ZAUTRE      NaN 900505050505      90
144613    False     ZAUTRE      NaN 900505050505      90
144614    False     ZAUTRE      NaN 900505050505      90
144642    False       NaN      NaN      NaN      NaN
144643    False       NaN      NaN      NaN      NaN
144644    False       NaN      NaN      NaN      NaN
144719    False     ZAUTRE      NaN 700510050505      70
144720    False     ZAUTRE      NaN 700510050505      70
144721    False     ZAUTRE      NaN 700510050505      70
```

On peut ensuite essayer de représenter les comptages d'articles sur les diverses variables catégorielles.

```
[14]: # Definition of feature and order to show
features = {'type': None,
            'march_group': ['ZFRAIS', 'ZSURGE', 'ZEPI', 'ZHYG', 'ZBOI', 'ZFLF', 'ZPMF', 'ZELAB', 'ZFP'],
            'storage_cond': ['FR', 'SU', 'EP', 'HY', 'FL', 'MA', 'SA', 'FP', 'PL'],
            'hier1': None,
            }
```

```
# Definition of color palette
palette = {'1PPF': c_passionfroid,
           '2PES': c_episaveurs,
           '3PTA': c_terreazur,
          }
```

```
[15]: fig, axs = plt.subplots(nrows=len(features), ncols=2, figsize=(13, 15))
# for each feature, draw counts without and with hue
for idx, (feature, order) in enumerate(features.items()):
    # drawing without hue
    sns.countplot(data=df.loc[active_march_mask],
                  x=feature,
                  order=order,
                  ax=axs[idx][0],
                  color=c_pomona)
    # remove y label, and set x label to full length text
    axs[idx][0].set_ylabel('')
    axs[idx][0].set_xlabel(lab[feature], fontsize=16)
    # drawing with hue
    sns.countplot(data=df.loc[active_march_mask],
                  x=feature,
                  hue='plant',
                  order=order,
                  palette=palette,
                  ax=axs[idx][1],
                  )
    # remove y label, and set x label to full length text
    axs[idx][1].set_ylabel('')
    axs[idx][1].set_xlabel(lab[feature], fontsize=16)
    # hide legend for each axis
    axs[idx][1].legend().set_visible(False)

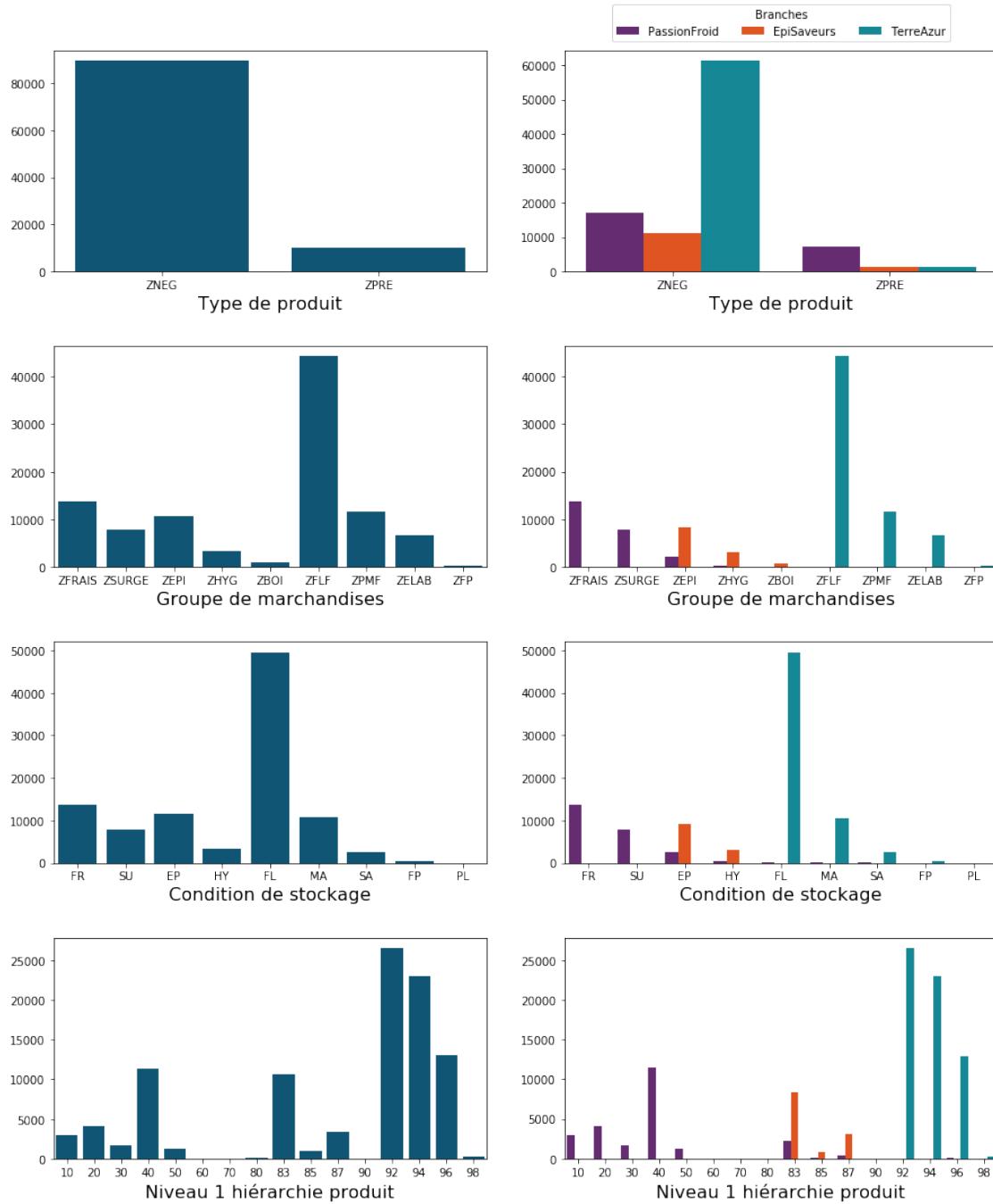
    # redraw legend for the whole figure, above, centered and
    # expanded
    handles, labels = axs[3][1].get_legend_handles_labels()
    fig.legend(handles,
               [lab[label] for label in labels],
               ncol=len(handles),
               title='Branches',
               loc='center',
               bbox_to_anchor=(0, 1, 1, 0.25),
               bbox_transform=axs[0][1].transAxes,
               # mode='expand',
               )

    # adding a title
fig.suptitle('Répartition des articles selon les features catégorielles',
             fontsize=24,
             y=1.025,
             va='bottom',
             )

    # adding padding between plots
fig.tight_layout(pad=3.0)

    # saving to disk
fig.savefig(Path('..') / 'img' / 'Repartition articles categories.png', bbox_inches='tight')
```

Répartition des articles selon les features catégorielles



```
[24]: def long_lab(label):
    if label in lab:
        return(label + ' - ' + lab[label])
    else:
        return(label)

for feature in features.keys():
    # Construct the pivot table for the feature
```

```

piv = pd.pivot_table(df.loc[active_march_mask],
                     columns='plant',
                     index=feature,
                     values='material',
                     aggfunc='count',
                     fill_value=0,
                     )
# Add a 'Total' column
piv['Total'] = piv['1PPF'] + piv['2PES'] + piv['3PTA']

# Changing 0s to '-'
piv = piv.replace(0, '-')

# Reorder indices so that they follow the order defined in
# lab dictionary
if np.all(piv.index.isin(lab.keys())): # check to avoid filtering piv!
    piv = piv.reindex([key for key in lab.keys() if key in piv.index])

# Rename indices, columns and axes for pretty printing
piv = (piv.rename(long_lab, axis=0)
       .rename(lab, axis=1)
       .rename_axis(lab[feature])
       .rename_axis('Branche', axis=1))

print(piv)
print('-----')
# Save to LaTeX format to be included in report
piv.to_latex(Path('..') / 'tbls' / ('Repartition par ' + feature + '.tex'),
             bold_rows=True,
             column_format='lcccc',
             na_rep='-', )
)

```

Branche	PassionFroid	EpiSaveurs	TerreAzur	Total
Type de produit				
ZNEG - Article de négoce	17166	11048	61273	89487
ZPRE - Article de prestation	7388	1193	1437	10018

Branche	PassionFroid	EpiSaveurs	TerreAzur	Total
Groupes de marchandises				
ZSURGE - Surgelés	7756	-	-	7756
ZFRAIS - Frais	13785	6	4	13795
ZEPI - Epicerie	2298	8305	-	10603
ZBOI - Boissons	126	826	-	952
ZHYG - Hygiène	350	3078	-	3428
ZFLF - Fruits et Légumes	4	-	44133	44137
ZPMF - Produits de la mer	142	-	11594	11736
ZELAB - Produits élaborés	91	-	6644	6735
ZFP - Fleurs et plantes	-	-	297	297
ZAUTRE - Autres	2	26	38	66

Branche	PassionFroid	EpiSaveurs	TerreAzur	Total
Condition de stockage				
SU - Surgelés	7758	-	-	7758
FR - Frais	13781	6	3	13790
EP - Epicerie	2430	9155	-	11585
HY - Hygiène	344	3080	-	3424
FL - Fruits et légumes	78	-	49508	49586
MA - Marée	126	-	10501	10627
FP - Fleurs et plantes	-	-	286	286
SA - Saurisserie	34	-	2408	2442
PL - Publicié	2	-	1	3

Branche	PassionFroid	EpiSaveurs	TerreAzur	Total
Niveau 1 hiérarchie produit				
10 - Beurre, oeufs, fromage	3010	6	1	3017
20 - Elaborés	4150	2	6	4158
30 - Garnitures et fruits	1701	-	-	1701

40 - Produits carnés	11413	-	-	11413
50 - Produits de la mer	1214	-	2	1216
60 - Consommables	1	-	-	1
70 - Emballage	-	1	-	1
80 - Publicité sur le lieu de vente	34	25	37	96
83 - Epicerie	2306	8296	-	10602
85 - Liquides	135	836	-	971
87 - Hygiène et entretien	348	3075	-	3423
90 - Services	10	-	-	10
92 - Fruits	35	-	26543	26578
94 - Légumes	37	-	22929	22966
96 - Produits de la mer Frais	160	-	12891	13051
98 - Fleurs - plantes	-	-	301	301

Analyse données du PIM

Pierre MASSÉ

April 25, 2020

1 Analyse des données du PIM

1.1 Extraction des données

1.1.1 Préambule technique

```
[1]: # setting up sys.path for relative imports
from pathlib import Path
import sys
project_root = str(Path(sys.path[0]).parents[1].absolute())
if project_root not in sys.path:
    sys.path.append(project_root)
```

```
[2]: # imports and customization of display
import io
import pandas as pd
pd.options.display.min_rows = 6
pd.options.display.width=108
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker

from src.pimapi import Requester
```

1.1.2 Récupération des données

Le requêtage des données dans le PIM s'appuie sur la classe `Requester` du module `pimapi`.

```
[5]: requester = Requester('prd')
# Let's fetch the full content of PIM system
requester.fetch_all_from_PIM()
requester.result
```

Done

```
[5]: [<Response [200]>,
<Response [200]>]
```

A ce stade, les données sont chargées en mémoire sous la forme de fichiers JSON. La conversion des données récupérées par l'API se fait via la méthode `result_to_dataframe` de la classe `Requester`.

```
[6]: df = requester.result_to_dataframe()
df.head(4)
```

```
entity-type repository \
uid
afee12c7-177e-4a68-9539-8cbb68442503    document    default
7d390121-17e8-43bf-a357-9d06b79d2d47    document    default
f234cd84-c8f6-433f-85ec-6e0b6980adc6    document    default
e82a8173-b379-41ac-b319-aa058a04fcfb    document    default

path          type \
uid
afee12c7-177e-4a68-9539-8cbb68442503 /default-domain/pomSupplierWorkspace/SICO/DEST... pomProduct
7d390121-17e8-43bf-a357-9d06b79d2d47 /default-domain/pomSupplierWorkspace/UNILEVER... pomProduct
f234cd84-c8f6-433f-85ec-6e0b6980adc6 /default-domain/pomSupplierWorkspace/AZTECA_FO... pomProduct
e82a8173-b379-41ac-b319-aa058a04fcfb /default-domain/pomSupplierWorkspace/UVCDR_-C... pomProduct

state \
uid
afee12c7-177e-4a68-9539-8cbb68442503 product.waiting.supplier.validation
7d390121-17e8-43bf-a357-9d06b79d2d47 product.waiting.supplier.validation
f234cd84-c8f6-433f-85ec-6e0b6980adc6 product.waiting.supplier.validation
e82a8173-b379-41ac-b319-aa058a04fcfb product.waiting.sending.supplier

parentRef  isCheckedOut  isVersion \
uid
afee12c7-177e-4a68-9539-8cbb68442503 a58845c0-cab3-492f-b48d-531f146c3777      True      False
7d390121-17e8-43bf-a357-9d06b79d2d47 a37abc27-f485-4ae9-921b-f761f16c8c1c      False     False
f234cd84-c8f6-433f-85ec-6e0b6980adc6 3ff7819a-a392-493f-beb8-0b323ac331c7      True      False
e82a8173-b379-41ac-b319-aa058a04fcfb e4b5167c-ece2-4f7a-83c1-fb884034a1bf      False     False

isProxy  changeToken ... \
uid
afee12c7-177e-4a68-9539-8cbb68442503   False      17-0 ...
7d390121-17e8-43bf-a357-9d06b79d2d47   False      15-0 ...
f234cd84-c8f6-433f-85ec-6e0b6980adc6   False      33-0 ...
e82a8173-b379-41ac-b319-aa058a04fcfb   False      19-0 ...

properties.pprodqmd:manufacturingDiagram.length \
uid
afee12c7-177e-4a68-9539-8cbb68442503                               NaN
7d390121-17e8-43bf-a357-9d06b79d2d47                               NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6                               NaN
e82a8173-b379-41ac-b319-aa058a04fcfb                               NaN

properties.pprodqmd:manufacturingDiagram.data \
uid
afee12c7-177e-4a68-9539-8cbb68442503                               NaN
7d390121-17e8-43bf-a357-9d06b79d2d47                               NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6                               NaN
e82a8173-b379-41ac-b319-aa058a04fcfb                               NaN

properties.pprodq:visualPhoto.name \
uid
afee12c7-177e-4a68-9539-8cbb68442503                               NaN
7d390121-17e8-43bf-a357-9d06b79d2d47                               NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6                               NaN
e82a8173-b379-41ac-b319-aa058a04fcfb                               NaN

properties.pprodq:visualPhoto.mime-type \
uid
afee12c7-177e-4a68-9539-8cbb68442503                               NaN
7d390121-17e8-43bf-a357-9d06b79d2d47                               NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6                               NaN
e82a8173-b379-41ac-b319-aa058a04fcfb                               NaN

properties.pprodq:visualPhoto.encoding \

```

```

uid
afee12c7-177e-4a68-9539-8cbb68442503                               NaN
7d390121-17e8-43bf-a357-9d06b79d2d47                               NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6                               NaN
e82a8173-b379-41ac-b319-aa058a04fcfb                               NaN

properties.pprodq:visualPhoto.digestAlgorithm \
uid
afee12c7-177e-4a68-9539-8cbb68442503                               NaN
7d390121-17e8-43bf-a357-9d06b79d2d47                               NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6                               NaN
e82a8173-b379-41ac-b319-aa058a04fcfb                               NaN

properties.pprodq:visualPhoto.digest \
uid
afee12c7-177e-4a68-9539-8cbb68442503                               NaN
7d390121-17e8-43bf-a357-9d06b79d2d47                               NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6                               NaN
e82a8173-b379-41ac-b319-aa058a04fcfb                               NaN

properties.pprodq:visualPhoto.length \
uid
afee12c7-177e-4a68-9539-8cbb68442503                               NaN
7d390121-17e8-43bf-a357-9d06b79d2d47                               NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6                               NaN
e82a8173-b379-41ac-b319-aa058a04fcfb                               NaN

properties.pprodq:visualPhoto.data properties.notif:notifications
uid
afee12c7-177e-4a68-9539-8cbb68442503                               NaN
7d390121-17e8-43bf-a357-9d06b79d2d47                               NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6                               NaN
e82a8173-b379-41ac-b319-aa058a04fcfb                               NaN

[4 rows x 487 columns]

```

1.2 Définitions pour les mises en formes

1.2.1 Descriptifs longs

On définit un dictionnaire permettant de “traduire” les codes de champs en libellés long.

```
[7]: lab = {
    'code': 'Code produit',
    'supplier': 'Code fournisseur',
    'type': 'Type de produit',
    'GTIN': 'GTIN',
    'base_unit': 'Unité de base',
    'net_weight': 'Poids net',
    'gross_weight': 'Poids brut',
    'dry_weight': 'Poids net égoutté',
    'volume': 'Volume',
    'total_life': 'Durée de vie totale',
    'remaining_life': 'Durée minimale restante',
    'type_cons': 'Type de conservation',
    'before_open': 'Conservation avant ouv.',
    'after_open': 'Conservation après ouv.',
    'cons_temp': 'Température',
}
```

1.2.2 Champs intéressants

On liste également les champs intéressants pour un affichage plus court du dataframe.

```
[8]: def_fields = {'properties.vig:code': 'code',
                 'properties.psec:supplierCode': 'supplier',
```

```
'properties.pprodtop:typeOfProduct': 'type',
'properties.pprodi:gtin': 'gtin',
'properties.pprodi:supplierDesignation': 'designation'}
```

1.3 Description des attributs des produit

1.3.1 Volumétrie des attributs

On constate que chaque produit porte un très grand nombre d'attributs :

```
[9]: print('Count of columns in df:', len(df.columns))
print('\nInfo of df:')
df.info()
```

Count of columns in df: 487

```
Info of df:
<class 'pandas.core.frame.DataFrame'>
Index: 13193 entries, afee12c7-177e-4a68-9539-8ccb68442503 to 6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a
Columns: 487 entries, entity-type to properties.notif:notifications
dtypes: bool(12), float64(60), int64(2), object(413)
memory usage: 48.1+ MB
```

De plus, de par la nature hiérarchique du format JSON, certains attributs dits “multivalués” sont parfois stockés sous forme de liste dans le dataframe “à plat”. Par exemple, on peut voir que le pays de transformation, ou les facettes, peuvent être multivalués.

```
[10]: df.loc[['609af223-2f14-4f83-a553-cef276f2eca7',
            'c94013e4-0dca-441a-85c1-0b29ecb54d0a',
            '82d1af25-2bdd-4315-9670-67784b70dfa7'],
            ['properties.pprodg:transfoCountries',
             'facets']]
```

```
[10]: properties.pprodg:transfoCountries \
uid
609af223-2f14-4f83-a553-cef276f2eca7 [PL, FR, ES]
c94013e4-0dca-441a-85c1-0b29ecb54d0a [DE, NO, BE, RU, CH, BG, LT, GR, FR, UA, HU, E...
82d1af25-2bdd-4315-9670-67784b70dfa7 [FR]

facets
uid
609af223-2f14-4f83-a553-cef276f2eca7 [Versionable, Folderish, Commentable, beginnin...
c94013e4-0dca-441a-85c1-0b29ecb54d0a [endMigration, Versionable, Folderish, Comment...
82d1af25-2bdd-4315-9670-67784b70dfa7 [endMigration, Versionable, Folderish, Comment...
```

De plus, certains attributs sont dits “complexes”, car chacune des valeurs de la liste est elle-même un dictionnaire d'attribut. La combinaison des deux, des attributs “complexes multivalués” existe également. On a alors une liste de dictionnaires. On peut comme ceci imbriquer des niveaux jusqu'à n'importe quelle profondeur.

C'est par exemple le cas des labels qui sont multivalués (un produit peut porter plusieurs labels), qui sont des complexes portant : - le type de label (bio, Label Rouge, ...) - la date de fin de validité du label (si applicable) - le fichier de certification du label (si applicable), qui est lui-même un complexe...

```
[11]: multilabel_ds = df.loc[df['properties.pprod:labels'].apply(len) > 1, 'properties.pprod:labels']
for uid, label_list in multilabel_ds.head(3).iteritems():
    print('product uid:', uid)
    for cpt, label in enumerate(label_list):
        print('\n\tlabel', cpt + 1, ':')
        for key, val in label.items():
            print('\t\t', key, ':', val)
    print('-----')
```

```
product uid: 362e6230-ba3a-4396-8a47-728b0a1d56db

label 1 :
labelCertificateEndDate : 2024-12-30T23:00:00.000Z
```

```

        typeOfLabel : 80
        labelCertificateFile : {'name': 'KCC Coleshill Tissue Paper Ecolabel Renewal Certificate
Mar 2020.pdf', 'mime-type': 'application/pdf', 'encoding': None, 'digestAlgorithm': 'MD5', 'digest':
'6615e3027ff2e014fdc3fa37e67851bb', 'length': '425662', 'data': 'https://produits.groupe-pomona.fr/nuxeo/nxf
ile/default/362e6230-ba3a-4396-8a47-728b0a1d56db/pprod1:labels/0/labelCertificateFile/KCC%20Coleshill%20Tiss
ue%20Paper%20Ecolabel%20Renewal%20Certificate%20Mar%202020.pdf?changeToken=36-0'}
-----
label 2 :
    labelCertificateEndDate : 2022-09-16T22:00:00.000Z
    typeOfLabel : NA
    labelCertificateFile : {'name': 'FCC_DoC_Coleshill Mill_PW_blue_Ref13351_Eng V01.pdf',
'mime-type': 'application/pdf', 'encoding': None, 'digestAlgorithm': 'MD5', 'digest':
'78cfcc67b8bf0f693e060088fd97c48', 'length': '84473', 'data': 'https://produits.groupe-pomona.fr/nuxeo/nxfi
le/default/362e6230-ba3a-4396-8a47-728b0a1d56db/pprod1:labels/1/labelCertificateFile/FCC_DoC_Coleshill%20Mil
l_PW_blue_Ref13351_Eng%20V01.pdf?changeToken=36-0'}
-----
product uid: 3c2a8d1a-634d-40bb-9852-81eb8a340114

label 1 :
    labelCertificateEndDate : None
    typeOfLabel : 30
    labelCertificateFile : None

label 2 :
    labelCertificateEndDate : None
    typeOfLabel : 40
    labelCertificateFile : None
-----
product uid: d3681e26-b024-4603-ae0b-0d5630329fa0

label 1 :
    labelCertificateEndDate : 2020-03-30T22:00:00.000Z
    typeOfLabel : 100
    labelCertificateFile : {'name': 'SAS - Certificat AB V2.pdf', 'mime-type':
'application/pdf', 'encoding': None, 'digestAlgorithm': 'MD5', 'digest': '1d424d2d2c9539abca07b8ad9576a339',
'length': '128780', 'data': 'https://produits.groupe-pomona.fr/nuxeo/nxfile/default/d3681e26-b024-4603-ae0b-
0d5630329fa0/pprod1:labels/0/labelCertificateFile/SAS%20-%20Certificat%20AB%20V2.pdf?changeToken=92-0'}
-----
label 2 :
    labelCertificateEndDate : 2020-03-30T22:00:00.000Z
    typeOfLabel : 80
    labelCertificateFile : {'name': 'SAS - Certificat AB V2.pdf', 'mime-type':
'application/pdf', 'encoding': None, 'digestAlgorithm': 'MD5', 'digest': '1d424d2d2c9539abca07b8ad9576a339',
'length': '128780', 'data': 'https://produits.groupe-pomona.fr/nuxeo/nxfile/default/d3681e26-b024-4603-ae0b-
0d5630329fa0/pprod1:labels/1/labelCertificateFile/SAS%20-%20Certificat%20AB%20V2.pdf?changeToken=92-0'}
-----
```

1.3.2 Description des principaux attributs

On commence par déclarer des utilitaires permettant de mettre en forme les représentations.

```
[12]: # Defining main data to explore
mappings = {
    'identification': {
        'properties.vig:code': 'code',
        'properties.psec:supplierCode': 'supplier',
        'properties.pprodtop:typeOfProduct': 'type',
        'properties.pprodःgtin': 'GTIN',
    },
    'dimensions': {
        'properties.pprodtop:baseUnit': 'base_unit',
        'properties.pprodःnetWeight': 'net_weight',
        'properties.pprodःgrossWeight': 'gross_weight',
        'properties.pprodःdryWeight': 'dry_weight',
        'properties.pprodःvolume': 'volume',
    },
    'conservation': {
```

```

        'properties.pprodg:totalLife': 'total_life',
        'properties.pprodg:guaranteedLife': 'remaining_life',
        'properties.pprodg:typeOfConservation': 'type_cons',
        'properties.pprodg:conservationBeforeOpening': 'before_open',
        'properties.pprodg:conservationAfterOpening': 'after_open',
        'properties.pprodg:conservationTemperature': 'cons_temp',
    }
}

# Helper function to transform pandas `to_latex` method output to a tabularx env instead.
def to_tabularx(stringio):
    text = StringIO.getvalue()
    text = text.replace(r'\begin{tabular}', r'\begin{tabularx}{\linewidth}')
    text = text.replace(r'\end{tabular}', r'\end{tabularx}')
    return(text)

# Function that saves dataframe as Latex tabularx files as input
def save_to_disk(df, path, lab=lab, tex_label=None):
    text = io.StringIO()
    c_format = 'l' + 'X' * len(df.columns)
    (df.rename(lab, axis=1)
     .to_latex(text,
                bold_rows=True,
                column_format=c_format,
                na_rep='-' ,
                label=tex_label,
                ))
    with open(path, mode='w') as file:
        file.write(to_tabularx(text))

```

On boucle sur les différents mappings, et on les sauvegarde dans des tableaux latex pour intégration au rapport.

```

[13]: for map_type, mapping in mappings.items():
    cur_df = df.loc[:, list(mapping.keys())].rename(mapping, axis=1).fillna(np.nan)
    desc = cur_df.describe(include='all')
    samp = cur_df.sample(n=5, random_state=42)
    print(map_type)
    print(samp.rename(lab, axis=1))
    print('-----')
    print(desc.rename(lab, axis=1)
          .round(3))
    )
    print('-----')

    # Writing dataframes to .tex files
    text = io.StringIO()
    c_format = 'l' + 'X' * len(cur_df.columns)
    (samp.rename(lab, axis=1)
     .to_latex(text,
                bold_rows=True,
                column_format=c_format,
                na_rep='-' ,
                ))
    with open(Path('..') / 'tbls' / ('Exemple ' + map_type + '.tex'), mode='w') as file:
        file.write(to_tabularx(text))

    text = io.StringIO()
    (desc.rename(lab, axis=1)
     .round(3)
     .to_latex(text,
                bold_rows=True,
                column_format=c_format,
                na_rep='-' ,
                ))
    with open(Path('..') / 'tbls' / ('Desc ' + map_type + '.tex'), mode='w') as file:
        file.write(to_tabularx(text))

```

identification

	Code produit	Code fournisseur	Type de produit	GTIN
uid				
1351c135-0d48-41ae-a568-2f33af6fdae9	PIMP-0000011253	PIMF-0000000416	grocery	3760063337099
ffabf67f-314e-47a8-932e-37da7a3ab1ae	PIMP-0000011972	PIMF-0000000486	hygiene	NaN
7bb3d9a9-d50c-4042-9d28-21b31f5cbbb1	PIMP-0000009973	PIMF-0000000328	alcoholicDrink	NaN
38b95b6e-5603-46bd-ad44-912a926ee0e4	PIMP-0000012507	PIMF-0000000391	chemistry	7615400045495
8738c768-9d5d-4233-b54c-99358fa66411	PIMP-0000000321	PIMF-0000000074	hygiene	3504082216054

	Code produit	Code fournisseur	Type de produit	GTIN
count	13193	13193	13193	12025
unique	13193	605	5	11339
top	PIMP-0000000721	PIMF-0000000179	grocery	
freq	1	370	8756	352

dimensions

	Unité de base	Poids net	Poids brut	Poids net égoutté	Volume
uid					
1351c135-0d48-41ae-a568-2f33af6fdae9	SAC	0.500	0.520	NaN	NaN
ffabf67f-314e-47a8-932e-37da7a3ab1ae	PU	NaN	NaN	NaN	NaN
7bb3d9a9-d50c-4042-9d28-21b31f5cbbb1	BIB	1.500	1.600	NaN	1.5
38b95b6e-5603-46bd-ad44-912a926ee0e4	BT.	1.053	1.150	NaN	NaN
8738c768-9d5d-4233-b54c-99358fa66411	COL	3.000	3.028	NaN	NaN

	Unité de base	Poids net	Poids brut	Poids net égoutté	Volume
count	13193	12827.000	12827.000	1026.000	1915.000
unique	30	NaN	NaN	NaN	NaN
top	BTE	NaN	NaN	NaN	NaN
freq	3051	NaN	NaN	NaN	NaN
mean	NaN	3.112	2.986	1.475	7.750
std	NaN	59.600	42.171	1.133	78.897
min	NaN	0.000	0.000	0.000	0.000
25%	NaN	0.482	0.535	0.480	0.500
50%	NaN	1.000	1.100	1.500	0.946
75%	NaN	3.000	3.298	2.380	2.500
max	NaN	4900.000	4730.500	10.000	3100.000

conservation

	Durée de vie totale	Durée minimale restante	Type de conservation	\
uid				
1351c135-0d48-41ae-a568-2f33af6fdae9	NaN	720.0	AM	
ffabf67f-314e-47a8-932e-37da7a3ab1ae	NaN	NaN	NaN	
7bb3d9a9-d50c-4042-9d28-21b31f5cbbb1	NaN	NaN	AM	
38b95b6e-5603-46bd-ad44-912a926ee0e4	NaN	NaN	AM	
8738c768-9d5d-4233-b54c-99358fa66411	NaN	NaN	AM	

	Conservation avant ouv.	Convervation après ouv.	Température
uid			
1351c135-0d48-41ae-a568-2f33af6fdae9	ambientTemperature	coolAndDryPlace	NaN
ffabf67f-314e-47a8-932e-37da7a3ab1ae	NaN	NaN	NaN
7bb3d9a9-d50c-4042-9d28-21b31f5cbbb1	ambientTemperature	notConcerned	NaN
38b95b6e-5603-46bd-ad44-912a926ee0e4	NaN	NaN	NaN
8738c768-9d5d-4233-b54c-99358fa66411	NaN	NaN	NaN

	Durée de vie totale	Durée minimale restante	Type de conservation	Conservation avant ouv.	\
count	8946.000	9466.000	12806	9977	
unique	NaN	NaN	2	7	
top	NaN	NaN	AM	ambientTemperature	
freq	NaN	NaN	12772	8270	
mean	651.406	351.618	NaN	NaN	
std	487.412	380.255	NaN	NaN	
min	0.000	0.000	NaN	NaN	
25%	360.000	180.000	NaN	NaN	
50%	540.000	300.000	NaN	NaN	
75%	900.000	480.000	NaN	NaN	
max	9999.000	9999.000	NaN	NaN	

Convervation après ouv. Température

count	9947	21
unique	18	9
top	coolAndDryPlace	15
freq	4512	6
mean	NaN	NaN
std	NaN	NaN
min	NaN	NaN
25%	NaN	NaN
50%	NaN	NaN
75%	NaN	NaN
max	NaN	NaN

1.3.3 Analyses spécifiques : GTIN

On peut mettre en évidence les produits qui portent les mêmes GTIN en double. En y jetant un oeil rapide, quelques explications peuvent être trouvées : - il peut s'agir d'un changement de code fournisseur (les 2 premières lignes ne portent pas le même code fournisseur) - il peut s'agir d'un changement de recette côté industriel, qui a décidé de conserver le même GTIN (second couple) - il peut s'agir d'une erreur, et de produits en doublon dans le système (troisième couple) - ...

```
[14]: GTIN_mask = (df['properties.pprod:gtin'] != '') & ~pd.isna(df['properties.pprod:gtin'])
dups_mask = df.loc[GTIN_mask, 'properties.pprod:gtin'].duplicated(keep=False)
examples = (df.loc[GTIN_mask & dups_mask, def_fields.keys()]
    .rename(def_fields, axis=1)
    .sort_values('gtin')
    .head(8))
save_to_disk(examples,
    Path('..') / 'tbls' / 'Duplicated_GTIN.tex',
    lab={},
    tex_label='tab:dup_gtin',
)
print(examples)
```

uid	code	supplier	type	gtin	designation
048712e3-f145-4f40-b8ad-7c0b912983bd	PIMP-0000009515	PIMF-0000000420	grocery	0020176760607	42 QUICHE FEUILL SG 11CM
4de8ce87-8df5-440c-959d-3d77d59bb4f3	PIMP-0000013159	PIMF-0000000182	grocery	0020176760607	QUICHE FEUILLETEE
7e455046-def3-4526-a28b-bc5c0e6e64fc	PIMP-0000011456	PIMF-0000000290	grocery	03344540125906	622028 SAUCE FUEGO SQUEEZE DE 580 G "O'TACOS"
a92c6ac5-d5be-4f92-98b3-9f6c588f7613	PIMP-0000013198	PIMF-0000000290	grocery	03344540125906	66590f04-5eae-4829-b0da-c899a18dd9cb
27e20042-dc53-46b4-874c-f970db554aec	PIMP-0000010839	PIMF-0000000250	grocery	3011360083845	PIMP-0000001494
02803e27-487a-43e3-9324-9ad1660b63b2	PIMP-0000002338	PIMF-0000000348	grocery	3038353024906	27e20042-dc53-46b4-874c-f970db554aec
52d3f309-e402-4931-974c-b6b6fa721aff	PIMP-0000002337	PIMF-0000000348	grocery	3038353024906	02803e27-487a-43e3-9324-9ad1660b63b2
					52d3f309-e402-4931-974c-b6b6fa721aff

Si l'on produit la répartition du nombre de produit portant un GTIN donné dans le système, on obtient :

```
[15]: df2 = (df.pivot_table(values='properties.vig:code',
                           index='properties.pprod:gtin',
                           aggfunc='count')
        .rename({'properties.vig:code': 'code_count'}, axis=1)
        )

df2 = (df2.reset_index()
       .loc[df2.index != '']
       .pivot_table(index='code_count',
                    aggfunc='count',
```

```

        values='code_count')
    .rename({'properties.pprodi:gtin': 'Nb de GTIN'},
           axis=1)
)

df2.index.rename('Pdt portant le GTIN', inplace=True)

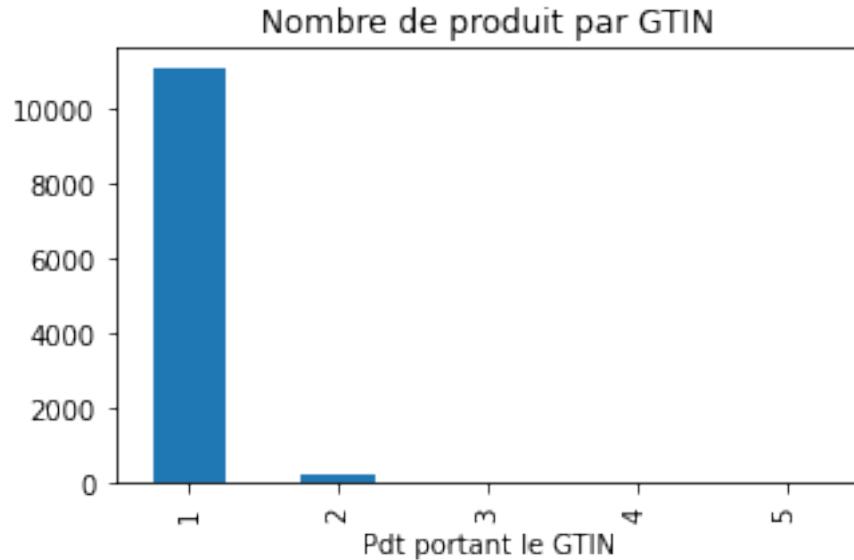
print(df2)

save_to_disk(df2,
             Path('..') / 'tbls' / 'gtin_counts.tex')

fig, ax = plt.subplots(figsize=(5,3))
df2.plot(kind='bar', legend=None, title='Nombre de produit par GTIN', ax=ax)
fig.savefig(Path('..') / 'img' / 'repartition_gtin.png', bbox_inches='tight')

```

	Nb de GTIN
Pdt portant le GTIN	
1	11068
2	227
3	22
4	20
5	1



1.3.4 Analyse spécifique : distribution par fournisseur

On peut représenter la distribution produit, par fournisseur.

```
[16]: # construction the counts
counts = (df.loc[:, list(def_fields.keys())]
          .rename(def_fields, axis=1)
          .pivot_table(values='code',
                      index='supplier',
                      aggfunc='count')
          .reset_index()
          .pivot_table(values=['supplier', 'code'],
                      index='code',
                      aggfunc={'supplier': 'count',
                               'code': 'sum'})
         )
```

```

# aligning index to have it continuous
new_idx = pd.RangeIndex(start=1, stop=max(counts.index) + 1)

counts = (counts.reindex(new_idx)
          .fillna(0)
        )

counts = pd.concat([counts,
                    counts.cumsum().rename({'code': 'cum_code', 'supplier': 'cum_supplier'},
                                           axis=1),
                    ],
                   axis=1)

for feature in ['supplier', 'code']:
    counts['cump_' + feature] = 100 * counts['cum_' + feature] / counts.loc[:, 'cum_' + feature].iloc[-1]

counts

```

```

[16]:   code  supplier  cum_code  cum_supplier  cump_supplier  cump_code
1     79.0      79.0      79.0           79.0      13.057851  0.598802
2    138.0      69.0     217.0          148.0      24.462810  1.644812
3    177.0      59.0     394.0          207.0      34.214876  2.986432
..      ...
368     0.0       0.0     12454.0         603.0      99.669421  94.398545
369   369.0       1.0     12823.0         604.0      99.834711  97.195482
370   370.0       1.0     13193.0         605.0     100.000000 100.000000
[370 rows x 6 columns]

```

```

[17]: fig, axs = plt.subplots(nrows=2,
                           ncols=2,
                           figsize=(14, 6),
                           gridspec_kw= {'width_ratios': [2, 1]})

axs2 = [[ax.twinx() for ax in axrow] for axrow in axs]

for i, feature in enumerate(['supplier', 'code']):
    axs[i][0].bar(data=counts.loc[:, feature].reset_index(), x='index', height=feature)
    axs2[i][0].plot('index', 'cump_' + feature, data=counts.loc[:, 'cump_' + feature].reset_index(),
                    color='red', linestyle='--')
    axs2[i][0].grid(True, axis='y', color='red', alpha=0.5, linestyle='--')

    axs[i][1].bar(data=counts.loc[:, feature].reset_index(), x='index', height=feature)
    axs2[i][1].plot('index', 'cump_' + feature, data=counts.loc[:, 'cump_' + feature].reset_index(),
                    color='red', linestyle='--')
    axs2[i][1].grid(True, axis='y', color='red', alpha=0.5, linestyle='--')

    axs[i][1].set_xlim(0, 30)

for i in range(len(axs)):
    for j in range(len(axs[i])):
        axs2[i][j].set_ylim(0, 100)
        # remove all bottom ticks except for bottom line
        # set_yticks does not work as it removes the grid
        if i < len(axs) - 1:
            axs[i][j].set_xticklabels([])
            for tic in axs[i][j].xaxis.get_major_ticks():
                tic.tick1line.set_visible(False)
                tic.tick2line.set_visible(False)
        # remove all right ticks except for right column
        # set_yticks does not work as it removes the grid
        if j < len(axs[i]) - 1:
            axs2[i][j].set_yticklabels([])
            for tic in axs2[i][j].yaxis.get_major_ticks():
                tic.tick1line.set_visible(False)
                tic.tick2line.set_visible(False)

```

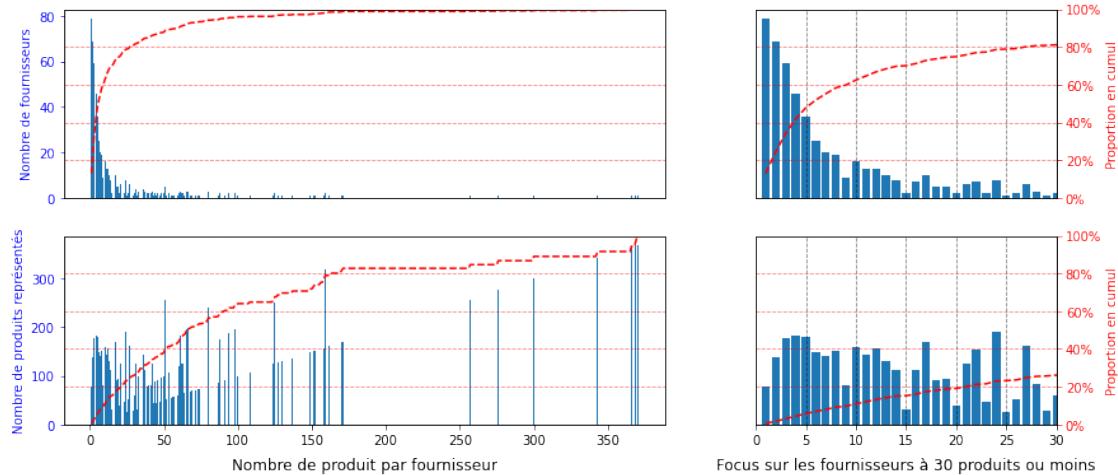
```

# remove all left ticks except for first column
if j > 0:
    axs[i][j].set_yticks([])
if j == len(axs[i]) - 1:
    axs2[i][j].tick_params(axis='y', colors='red')
    axs2[i][j].yaxis.set_major_formatter(ticker.PercentFormatter())
    axs2[i][j].set_ylabel('Proportion en cumul', color='red')
    axs[i][j].grid(True, axis='x', color='k', alpha=0.5, linestyle='--')
if j == 0:
    axs[i][j].tick_params(axis='y', colors='blue')
    if i == 0:
        axs[i][j].set_ylabel('Nombre de fournisseurs', color='blue')
    if i == 1:
        axs[i][j].set_ylabel('Nombre de produits représentés', color='blue')
    axs[1][0].set_xlabel('Nombre de produit par fournisseur',
                         fontsize=12,
                         labelpad=8,
                         )
axs[1][1].set_xlabel('Focus sur les fournisseurs à 30 produits ou moins',
                     fontsize=12,
                     labelpad=8,
                     )

fig.suptitle('Distribution des fournisseurs fonction du nombre de produit par fournisseur',
             fontsize=16,
             )
fig.savefig(Path('..') / 'img' / 'distribution_fournisseurs_par_prd_count.png', bbox_inches='tight')

```

Distribution des fournisseurs fonction du nombre de produit par fournisseur



On peut également représenter le nombre de produits “récupérés” si on prend les fournisseurs par nombre de produits décroissant.

```

[18]: # construction the counts
counts = (df.loc[:, list(def_fields.keys())]
          .rename(def_fields, axis=1)
          .pivot_table(values='code',
                      index='supplier',
                      aggfunc='count')
          .sort_values('code', ascending=False)
)
counts['code_cumsum'] = counts['code'].cumsum()
counts

```

```
[18]:          code  code_cumsum
supplier
PIMF-0000000179    370      370
PIMF-0000000250    369      739
PIMF-0000000283    366     1105
...        ...      ...
PIMF-0000000408     1     13191
PIMF-0000000407     1     13192
PIMF-0000000666     1     13193

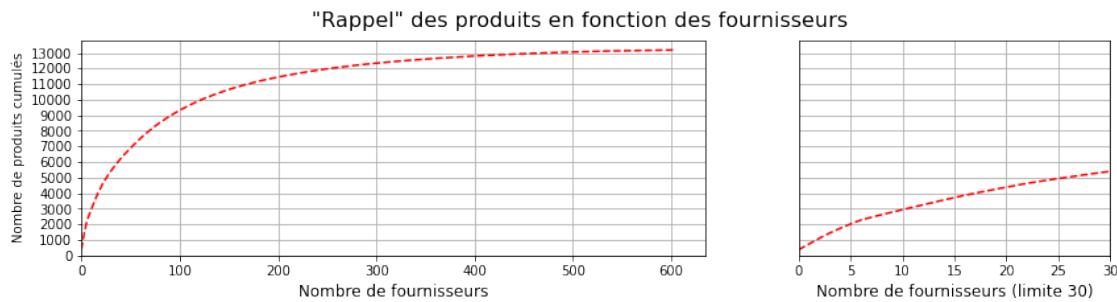
[605 rows x 2 columns]
```

```
[19]: fig, axs = plt.subplots(nrows=1,
                           ncols=2,
                           figsize=(14, 3),
                           gridspec_kw= {'width_ratios': [2, 1]})

for j in range(len(axs)):
    axs[j].plot('index',
                'code_cumsum',
                data=counts.reset_index().reset_index(),
                color='red',
                linestyle='--',
                )
    axs[j].set_xlabel('Nombre de fournisseurs', fontsize=12)
    axs[j].set_ylimit(0)
    axs[j].set_xlim(0)
    axs[j].grid(True)
    axs[j].yaxis.set_ticks(np.arange(0, 14000, 1000))

axs[0].set_ylabel('Nombre de produits cumulés')
axs[1].set_xlim(0, 30)
axs[1].set_xlabel('Nombre de fournisseurs (limite 30)', fontsize=12)
axs[1].set_yticklabels([])
for tic in axs[1].yaxis.get_major_ticks():
    tic.tick1line.set_visible(False)
    tic.tick2line.set_visible(False)

fig.suptitle('"Rappel" des produits en fonction des fournisseurs', fontsize=16)
fig.savefig(Path('..') / 'img' / 'rappel_produit_par_fournisseur.png', bbox_inches='tight')
```



ground_truth_constitution

Pierre MASSÉ

May 3, 2020

1 Constitution de l'échantillon de données étiquetées

L'objet de ce notebook est de produire un échantillon données du PIM, avec les fiches techniques associées. Elles seront ensuite associées manuellement à la liste d'ingrédients qu'elles contiennent.

1.1 Récupération des données

1.1.1 Préambule technique

```
[1]: # setting up sys.path for relative imports
from pathlib import Path
import sys
project_root = str(Path(sys.path[0]).parents[1].absolute())
if project_root not in sys.path:
    sys.path.append(project_root)
```

```
[2]: # imports and customization of display
import os
import pandas as pd
pd.options.display.min_rows = 6
pd.options.display.width=108
from sklearn.model_selection import train_test_split

from src.pimapi import Requester
```

1.1.2 Récupération des données, et de la présences de fiches techniques

Pour constituer l'échantillon, on va d'abord extraire quelques informations du PIM, et particulièrement le type de produit. On récupèrera aussi le fait que les produits ont ou non une fiche technique fournisseur associée.

```
[7]: requester = Requester('prd')
# Let's fetch the full content of PIM system
requester.fetch_all_from_PIM()
requester.result
```

Done

```
[7]: [<Response [200]>,
<Response [200]>]
```

```
[8]: mapping = {'uid': 'uid',
   'designation': 'title',
   'state': 'state',
   'ingredients': 'properties.pprodct:ingredientsList',
   'type': 'properties.pprodtop:typeOfProduct'}
df = requester.file_report_from_result(mapping=mapping, index='uid') # , record_path='entries')
df
```

```
[8]: designation \
uid
afee12c7-177e-4a68-9539-8cbb68442503 DESTR D'ODEURS AIR&TEXTILES 750CCX6 DESODOR U2
7d390121-17e8-43bf-a357-9d06b79d2d47 THÉ VERT AGRUME BTE 25S FRAICH LIPTON
f234cd84-c8f6-433f-85ec-6e0b6980adc6 T WHEAT 30 A 18X6 52C MISSION 1620
...
ef42a938-2203-446e-8d28-9fd27c6d3146 3D VENT FRAIS 5LX4 DESODOR U2
68f5d81b-7f91-40a0-8504-0ec320a86de4 NETTOYANT INOX 500ML LOT 2X6 KING
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a DESINFECTANT 3D+ 750CCX6 DESODOR U2

state \
uid
afee12c7-177e-4a68-9539-8cbb68442503 product.waiting.supplier.validation
7d390121-17e8-43bf-a357-9d06b79d2d47 product.waiting.supplier.validation
f234cd84-c8f6-433f-85ec-6e0b6980adc6 product.waiting.supplier.validation
...
ef42a938-2203-446e-8d28-9fd27c6d3146 product.waiting.supplier.validation
68f5d81b-7f91-40a0-8504-0ec320a86de4 product.waiting.supplier.validation
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a product.waiting.supplier.validation

ingredients type \
uid
afee12c7-177e-4a68-9539-8cbb68442503 None chemistry
7d390121-17e8-43bf-a357-9d06b79d2d47 None grocery
f234cd84-c8f6-433f-85ec-6e0b6980adc6 WHEAT flour (55%), water, vegetable fat (palm)... grocery
...
ef42a938-2203-446e-8d28-9fd27c6d3146 ...
68f5d81b-7f91-40a0-8504-0ec320a86de4 ...
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a ...

has_supplierdatasheet has_supplierlabel
uid
afee12c7-177e-4a68-9539-8cbb68442503 False False
7d390121-17e8-43bf-a357-9d06b79d2d47 False False
f234cd84-c8f6-433f-85ec-6e0b6980adc6 True True
...
ef42a938-2203-446e-8d28-9fd27c6d3146 ...
68f5d81b-7f91-40a0-8504-0ec320a86de4 ...
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a ...

[13212 rows x 6 columns]
```

1.2 Constitution de l'échantillon

On va constituer l'échantillon en appliquant les règles suivantes : - on construit un échantillon de 500 produits - on conserve les produits de type Epicerie et Boisson non alcoolisée - on conserve les produits qui possèdent une fiche technique fournisseur - on fait un échantillon stratifié par type de produit (Epicerie / Boisson)

```
[12]: filtered_df = df.loc[(df.type.isin(['grocery', 'nonAlcoholicDrink']))
   & (df.has_supplierdatasheet)]
train, ground_truth_df = train_test_split(filtered_df,
                                           test_size=500,
                                           random_state=42,
                                           stratify=filtered_df.type)
ground_truth_df
```

```
[12]: designation \
uid
```

```

49428283-104a-4092-966f-07b974112836 Jus de pomme Granny Smith en bouteille verre 1...
c679c923-cf39-4e40-b072-7474928450c6 Purée de pomme de terre en sac 5 kg LUTOSA
ef8afbbb-efbc-4fdf-aa1c-6f664f2c1073 CÂPRE FINE AU VINAIGRE EN BOÎTE 4/4 VITALY'S
...
aff93d0f-a94c-4e9b-a6ca-d69115e3d2eb Thon Listao en morceaux au naturel en boîte 4...
a01039be-eb79-454f-a384-c142d1d80d0c Tajine végétarienne en barquette 2,3 kg CHRIST
97c80844-08c7-45c6-82ac-43ab8a45f2e4 Spécialité pomme-fraise en gourde 90 g ANDROS

state \
uid
49428283-104a-4092-966f-07b974112836 product.validate
c679c923-cf39-4e40-b072-7474928450c6 product.validate
ef8afbbb-efbc-4fdf-aa1c-6f664f2c1073 product.validate
...
aff93d0f-a94c-4e9b-a6ca-d69115e3d2eb product.validate
a01039be-eb79-454f-a384-c142d1d80d0c product.validate
97c80844-08c7-45c6-82ac-43ab8a45f2e4 product.waiting.pomona.validation.niv3

ingredients \
uid
49428283-104a-4092-966f-07b974112836 100 % pur jus de Pommes Granny Smith.
c679c923-cf39-4e40-b072-7474928450c6 Pommes de terre déshydratées (99 %), émulsifia...
ef8afbbb-efbc-4fdf-aa1c-6f664f2c1073 Câpres, eau, vinaigre, sel
...
aff93d0f-a94c-4e9b-a6ca-d69115e3d2eb THON listao, eau, sel
a01039be-eb79-454f-a384-c142d1d80d0c Légumes 48,1% : pommes de terre 18,2%, carotte...
97c80844-08c7-45c6-82ac-43ab8a45f2e4 Ingrédients : Pommes 74%, fraises 20%, sucre, ...

type has_supplierdatasheet has_supplierlabel
uid
49428283-104a-4092-966f-07b974112836 nonAlcoholicDrink True False
c679c923-cf39-4e40-b072-7474928450c6 grocery True True
ef8afbbb-efbc-4fdf-aa1c-6f664f2c1073 grocery True False
...
aff93d0f-a94c-4e9b-a6ca-d69115e3d2eb grocery True True
a01039be-eb79-454f-a384-c142d1d80d0c grocery True True
97c80844-08c7-45c6-82ac-43ab8a45f2e4 grocery True True

```

[500 rows x 6 columns]

Remarque : malgré l'utilisation d'un `random_state` fixé, l'échantillon généré n'est pas toujours le même à chaque exécution. En effet, comme la liste de produits varie au fil du temps (nouveaux référencements, périmètre des filtres qui change), le résultat du `train_test_split` peut varier.

Il s'agit ici seulement d'illustrer la méthode utilisée.

1.3 Export des pièces jointes du PIM et constitution du fichier d'étiquettes

On exporte ensuite le contenu du PIM sur le disque, afin d'avoir les fiches techniques simplement à disposition.

```
[14]: requester.fetch_list_from_PIM(ground_truth_df.index, batch_size=20)
requester.dump_data_from_result(update_directory=False, root_path=os.path.join('.', 'ground_truth_to_del'))
requester.dump_files_from_result(update_directory=False, root_path=os.path.join('.', 'ground_truth_to_del'))
```

```

Done
Done
Launching 25 threads.
Thread complete!

```

```

Thread complete!
Done

```

On exporte également au format csv les uids des produits et les libellés associés (pour s'assurer qu'il n'y a pas eu de confusion lorsqu'on lit une fiche technique).

```
[15]: ground_truth_df['designation'].to_csv(os.path.join('..', 'ground_truth_to_del', 'uid.csv'),
                                         header=True,
                                         encoding='utf-8-sig')
```

On teste également la possibilité de recharger les données depuis le fichier csv, une fois qu'il a été renseigné à la main dans excel.

```
[20]: pd.read_csv(os.path.join('..', '..', 'ground_truth', 'manually_labelled_ground_truth.csv'),
                 sep=';',
                 encoding='latin-1',
                 index_col='uid')
```

```
[20]:                                            designation \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70    Concentré liquide Asian en bouteille 980 ml CHEF
d183e914-db2f-4e2f-863a-a3b2d054c0b8      Pain burger curry 80 g CREATIV BURGER
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8      Macaroni en sachet 500 g PANZANI
...
e67341d8-350f-46f4-9154-4dbbb8035621      PRÉPARATION POUR CRÈME BRûLÉE BIO 6L
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3      Céréales instantanées en poudre saveur caramel...
0faad739-ea8c-4f03-b62e-51ee592a0546      FARINE DE BLÉ TYPE 45, 10KG

                                                ingredients
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70    Eau, maltodextrine, sel, arômes, sucre, arôme ...
d183e914-db2f-4e2f-863a-a3b2d054c0b8      Farine de blé T65, eau, levure, vinaigre de ci...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8      - 100% Semoule de BLE dur de qualité supérieur...
...
e67341d8-350f-46f4-9154-4dbbb8035621    Sucre roux de canne*° (64%), amidon de mais*, ...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3      Farine 87,1 % (Blé (GLUTEN), Blé hydrolysé (GL...
0faad739-ea8c-4f03-b62e-51ee592a0546      Farine de blé T45

[500 rows x 2 columns]
```

1.4 Résultat de l'étiquetage manuel

Le résultat de l'étiquetage manuel est le suivant :

```
[3]: df_gt = pd.read_csv(os.path.join('..', '..', 'ground_truth', 'manually_labelled_ground_truth.csv'),
                       sep=';',
                       encoding='latin-1',
                       index_col='uid')

def to_latex_newline(text):
    return(text.replace('\n', ' '))
```

```

with pd.option_context("max_colwidth", 1000):
    print(df_gt)
    df_gt.to_latex(Path('..') / 'tbls' / 'ground_truth.tex',
                    index=False,
                    index_names=False,
                    column_format='p{5cm}p{10cm}',
                    formatters=[to_latex_newline, to_latex_newline],
                    longtable=True,
                    na_rep='-' ,
                    escape=True,
                    )

```

designation

\

uid

a0492df6-9c76-4303-8813-65ec5ccbfa70	Concentré liquide Asian en bouteille 980 ml CHEF
d183e914-db2f-4e2f-863a-a3b2d054c0b8	Pain burger curry 80 g CREATIV BURGER
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8	Macaroni en sachet 500 g PANZANI

..

e67341d8-350f-46f4-9154-4dbbb8035621	PRÉPARATION POUR CRÈME BRÛLÉE BIO 6L
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3	Céréales instantanées en poudre saveur caramel en boîte 400 g BLEDINA
0faad739-ea8c-4f03-b62e-51ee592a0546	FARINE DE BLÉ TYPE 45, 10KG

ingredients

uid

a0492df6-9c76-4303-8813-65ec5ccbfa70	Eau, maltodextrine, sel, arômes, sucre, arôme naturel de citronnelle, amidon modifié, ail en poudre, épices (combava, curcuma), extraits d'épices (gingembre, poivre), stabilisant (gomme xanthane).
d183e914-db2f-4e2f-863a-a3b2d054c0b8	Farine de blé T65, eau, levure, vinaigre de cidre, huile de colza, assaisonnement poudre de curry, sel, acide ascorbique, émulsifiant : E471.
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8	- 100% Semoule de BLE dur de qualité supérieure\n- Contient du gluten\nSi le numéro de lot contient la lettre N : peu contenir de l'oeuf

..

..

e67341d8-350f-46f4-9154-4dbbb8035621	Sucre roux de canne*° (64%), amidon de maïs*, poudre de LAIT écrémé*, poudre d'ŒUFS entiers*, gélifiants : carraghénanes, agaragar* ; arôme naturel de vanille* et autres arômes naturels*, poudre de gousses de vanille*, curcuma*. \n* Produits issus de l'Agriculture Biologique.\n° Ingrediént issu du commerce équitable. 65.1% des ingrédients d'origine agricole sont issus du commerce équitable (Sucre : Paraguay).
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3	Farine 87,1 % (Blé (GLUTEN), Blé hydrolysé (GLUTEN)) - Sucre - Caramel 5,00 % - Arôme - Vitamines (C, B1) - Diposphate ferrique
0faad739-ea8c-4f03-b62e-51ee592a0546	Farine de blé T45

[500 rows x 2 columns]

open_model

Pierre MASSÉ

May 8, 2020

1 Modèle “ouvert”

L'objet de ce notebook est de démontrer la faisabilité de prédire les listes d'ingrédients depuis des fiches techniques

1.1 Préambule technique

```
[1]: # setting up sys.path for relative imports
from pathlib import Path
import sys
project_root = str(Path(sys.path[0]).parents[1].absolute())
if project_root not in sys.path:
    sys.path.append(project_root)

[12]: # imports and customization of display
# import os
# from functools import partial
# import numpy as np
import pandas as pd
pd.options.display.min_rows = 6
pd.options.display.width=108
# from sklearn.feature_extraction.text import CountVectorizer
# from sklearn.model_selection import train_test_split
# from sklearn.model_selection import cross_val_score, cross_validate
# from sklearn.pipeline import Pipeline
# from matplotlib import pyplot as plt

from src.pimapi import Requester
# from src.pimest import ContentGetter
# from src.pimest import PathGetter
# from src.pimest import PDFContentParser
# from src.pimest import BlockSplitter
# from src.pimest import SimilaritySelector
# from src.pimest import custom_accuracy
```

1.2 Extraction des données

On extrait les données depuis le PIM :

```
[6]: requester = Requester('prd')
requester.fetch_all_from_PIM()
requester.result
```

Done

```
-----
NameError                                 Traceback (most recent call last)

<ipython-input-6-d2e4623a16cf> in <module>
      1 requester = Requester('prd')
      2 requester.fetch_all_from_PIM(page_size=1000, max_page=-1, nx_properties='*')
```

```

----> 3 df = requester.result_to_dataframe(record_path='entries', mapping=mapping, index='uid')
4 df

NameError: name 'mapping' is not defined

[9]: df = requester.result_to_dataframe(record_path='entries', index='uid')
df

[9]:
entity-type repository \
uid
afee12c7-177e-4a68-9539-8cbb68442503 document default
7d390121-17e8-43bf-a357-9d06b79d2d47 document default
f234cd84-c8f6-433f-85ec-6e0b6980adc6 document default
e82a8173-b379-41ac-b319-aa058a04fcfb document default
4b12c47c-84f5-4132-b362-22b864379a67 document default
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d document ...
0273eadc-851a-4b68-8020-8041700a4f3d document default
ef42a938-2203-446e-8d28-9fd27c6d3146 document default
68f5d81b-7f91-40a0-8504-0ec320a86de4 document default
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a document default

path \
uid
afee12c7-177e-4a68-9539-8cbb68442503 /default-domain/pomSupplierWorkspace/SICO/DEST...
7d390121-17e8-43bf-a357-9d06b79d2d47 /default-domain/pomSupplierWorkspace/UNILEVER...
f234cd84-c8f6-433f-85ec-6e0b6980adc6 /default-domain/pomSupplierWorkspace/AZTECA_FO...
e82a8173-b379-41ac-b319-aa058a04fcfb /default-domain/pomSupplierWorkspace/UVCDR_-_C...
4b12c47c-84f5-4132-b362-22b864379a67 /default-domain/pomSupplierWorkspace/UVCDR__C...
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d /default-domain/pomSupplierWorkspace/CGMP/NAPP...
0273eadc-851a-4b68-8020-8041700a4f3d /default-domain/pomSupplierWorkspace/SICO/DETE...
ef42a938-2203-446e-8d28-9fd27c6d3146 /default-domain/pomSupplierWorkspace/SICO/DETE...
68f5d81b-7f91-40a0-8504-0ec320a86de4 /default-domain/pomSupplierWorkspace/SICO/NETT...
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a /default-domain/pomSupplierWorkspace/SICO/SPRA...

type \
uid
afee12c7-177e-4a68-9539-8cbb68442503 pomProduct
7d390121-17e8-43bf-a357-9d06b79d2d47 pomProduct
f234cd84-c8f6-433f-85ec-6e0b6980adc6 pomProduct
e82a8173-b379-41ac-b319-aa058a04fcfb pomProduct
4b12c47c-84f5-4132-b362-22b864379a67 pomProduct
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d pomProduct
0273eadc-851a-4b68-8020-8041700a4f3d pomProduct
ef42a938-2203-446e-8d28-9fd27c6d3146 pomProduct
68f5d81b-7f91-40a0-8504-0ec320a86de4 pomProduct
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a pomProduct

state \
uid
afee12c7-177e-4a68-9539-8cbb68442503 product.waiting.supplier.validation
7d390121-17e8-43bf-a357-9d06b79d2d47 product.waiting.supplier.validation
f234cd84-c8f6-433f-85ec-6e0b6980adc6 product.waiting.supplier.validation
e82a8173-b379-41ac-b319-aa058a04fcfb product.waiting.sending.supplier
4b12c47c-84f5-4132-b362-22b864379a67 product.waiting.sending.supplier
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d product.waiting.supplier.validation
0273eadc-851a-4b68-8020-8041700a4f3d product.waiting.supplier.validation
ef42a938-2203-446e-8d28-9fd27c6d3146 product.waiting.supplier.validation
68f5d81b-7f91-40a0-8504-0ec320a86de4 product.waiting.supplier.validation
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a product.waiting.supplier.validation

parentRef \

```

```

uid
afee12c7-177e-4a68-9539-8cbb68442503 a58845c0-cab3-492f-b48d-531f146c3777
7d390121-17e8-43bf-a357-9d06b79d2d47 a37abc27-f485-4ae9-921b-f761f16c8c1c
f234cd84-c8f6-433f-85ec-6e0b6980adc6 3ff7819a-a392-493f-beb8-0b323ac331c7
e82a8173-b379-41ac-b319-aa058a04fcfb e4b5167c-ece2-4f7a-83c1-fb884034a1bf
4b12c47c-84f5-4132-b362-22b864379a67 e4b5167c-ece2-4f7a-83c1-fb884034a1bf
...
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d 0f182b14-e794-4a1a-af96-84d976ea9453
0273eadc-851a-4b68-8020-8041700a4f3d a58845c0-cab3-492f-b48d-531f146c3777
ef42a938-2203-446e-8d28-9fd27c6d3146 a58845c0-cab3-492f-b48d-531f146c3777
68f5d81b-7f91-40a0-8504-0ec320a86de4 a58845c0-cab3-492f-b48d-531f146c3777
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a a58845c0-cab3-492f-b48d-531f146c3777

isCheckedOut isVersion isProxy \
uid
afee12c7-177e-4a68-9539-8cbb68442503 True False False
7d390121-17e8-43bf-a357-9d06b79d2d47 False False False
f234cd84-c8f6-433f-85ec-6e0b6980adc6 True False False
e82a8173-b379-41ac-b319-aa058a04fcfb False False False
4b12c47c-84f5-4132-b362-22b864379a67 False False False
...
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d False False False
0273eadc-851a-4b68-8020-8041700a4f3d True False False
ef42a938-2203-446e-8d28-9fd27c6d3146 True False False
68f5d81b-7f91-40a0-8504-0ec320a86de4 True False False
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a True False False

changeToken ... \
uid
afee12c7-177e-4a68-9539-8cbb68442503 ... ...
7d390121-17e8-43bf-a357-9d06b79d2d47 ... ...
f234cd84-c8f6-433f-85ec-6e0b6980adc6 ... ...
e82a8173-b379-41ac-b319-aa058a04fcfb ... ...
4b12c47c-84f5-4132-b362-22b864379a67 ... ...
...
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d ... ...
0273eadc-851a-4b68-8020-8041700a4f3d ... ...
ef42a938-2203-446e-8d28-9fd27c6d3146 ... ...
68f5d81b-7f91-40a0-8504-0ec320a86de4 ... ...
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a ... ...

properties.pprodqmdd:manufacturingDiagram.length \
uid
afee12c7-177e-4a68-9539-8cbb68442503 NaN
7d390121-17e8-43bf-a357-9d06b79d2d47 NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6 NaN
e82a8173-b379-41ac-b319-aa058a04fcfb NaN
4b12c47c-84f5-4132-b362-22b864379a67 NaN
...
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d ... ...
0273eadc-851a-4b68-8020-8041700a4f3d ... ...
ef42a938-2203-446e-8d28-9fd27c6d3146 ... ...
68f5d81b-7f91-40a0-8504-0ec320a86de4 ... ...
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a ... ...

properties.pprodqmdd:manufacturingDiagram.data \
uid
afee12c7-177e-4a68-9539-8cbb68442503 NaN
7d390121-17e8-43bf-a357-9d06b79d2d47 NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6 NaN
e82a8173-b379-41ac-b319-aa058a04fcfb NaN
4b12c47c-84f5-4132-b362-22b864379a67 NaN
...
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d ... ...
0273eadc-851a-4b68-8020-8041700a4f3d ... ...
ef42a938-2203-446e-8d28-9fd27c6d3146 ... ...
68f5d81b-7f91-40a0-8504-0ec320a86de4 ... ...
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a ... 
```

6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a	NaN
properties.pprodqmdd:secondaryPackagingPhoto.name \	
uid	
afee12c7-177e-4a68-9539-8cbb68442503	NaN
7d390121-17e8-43bf-a357-9d06b79d2d47	NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6	NaN
e82a8173-b379-41ac-b319-aa058a04fcfb	NaN
4b12c47c-84f5-4132-b362-22b864379a67	NaN
...	...
5cde49c6-9e7e-4bd2-b22a-3239f643379d	NaN
0273eadc-851a-4b68-8020-8041700a4f3d	NaN
ef42a938-2203-446e-8d28-9fd27c6d3146	NaN
68f5d81b-7f91-40a0-8504-0ec320a86de4	NaN
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a	NaN
properties.pprodqmdd:secondaryPackagingPhoto.mime-type \	
uid	
afee12c7-177e-4a68-9539-8cbb68442503	NaN
7d390121-17e8-43bf-a357-9d06b79d2d47	NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6	NaN
e82a8173-b379-41ac-b319-aa058a04fcfb	NaN
4b12c47c-84f5-4132-b362-22b864379a67	NaN
...	...
5cde49c6-9e7e-4bd2-b22a-3239f643379d	NaN
0273eadc-851a-4b68-8020-8041700a4f3d	NaN
ef42a938-2203-446e-8d28-9fd27c6d3146	NaN
68f5d81b-7f91-40a0-8504-0ec320a86de4	NaN
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a	NaN
properties.pprodqmdd:secondaryPackagingPhoto.encoding \	
uid	
afee12c7-177e-4a68-9539-8cbb68442503	NaN
7d390121-17e8-43bf-a357-9d06b79d2d47	NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6	NaN
e82a8173-b379-41ac-b319-aa058a04fcfb	NaN
4b12c47c-84f5-4132-b362-22b864379a67	NaN
...	...
5cde49c6-9e7e-4bd2-b22a-3239f643379d	NaN
0273eadc-851a-4b68-8020-8041700a4f3d	NaN
ef42a938-2203-446e-8d28-9fd27c6d3146	NaN
68f5d81b-7f91-40a0-8504-0ec320a86de4	NaN
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a	NaN
properties.pprodqmdd:secondaryPackagingPhoto.digestAlgorithm \	
uid	
afee12c7-177e-4a68-9539-8cbb68442503	NaN
7d390121-17e8-43bf-a357-9d06b79d2d47	NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6	NaN
e82a8173-b379-41ac-b319-aa058a04fcfb	NaN
4b12c47c-84f5-4132-b362-22b864379a67	NaN
...	...
5cde49c6-9e7e-4bd2-b22a-3239f643379d	NaN
0273eadc-851a-4b68-8020-8041700a4f3d	NaN
ef42a938-2203-446e-8d28-9fd27c6d3146	NaN
68f5d81b-7f91-40a0-8504-0ec320a86de4	NaN
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a	NaN
properties.pprodqmdd:secondaryPackagingPhoto.digest \	
uid	
afee12c7-177e-4a68-9539-8cbb68442503	NaN
7d390121-17e8-43bf-a357-9d06b79d2d47	NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6	NaN
e82a8173-b379-41ac-b319-aa058a04fcfb	NaN
4b12c47c-84f5-4132-b362-22b864379a67	NaN
...	...
5cde49c6-9e7e-4bd2-b22a-3239f643379d	NaN

```

0273eadc-851a-4b68-8020-8041700a4f3d           NaN
ef42a938-2203-446e-8d28-9fd27c6d3146           NaN
68f5d81b-7f91-40a0-8504-0ec320a86de4           NaN
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a           NaN

properties.pprodqmd:secondaryPackagingPhoto.length \
uid
afee12c7-177e-4a68-9539-8ccb68442503           NaN
7d390121-17e8-43bf-a357-9d06b79d2d47           NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6           NaN
e82a8173-b379-41ac-b319-aa058a04fcfb           NaN
4b12c47c-84f5-4132-b362-22b864379a67           NaN
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d           ...
0273eadc-851a-4b68-8020-8041700a4f3d           NaN
ef42a938-2203-446e-8d28-9fd27c6d3146           NaN
68f5d81b-7f91-40a0-8504-0ec320a86de4           NaN
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a           NaN

properties.pprodqmd:secondaryPackagingPhoto.data \
uid
afee12c7-177e-4a68-9539-8ccb68442503           NaN
7d390121-17e8-43bf-a357-9d06b79d2d47           NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6           NaN
e82a8173-b379-41ac-b319-aa058a04fcfb           NaN
4b12c47c-84f5-4132-b362-22b864379a67           NaN
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d           ...
0273eadc-851a-4b68-8020-8041700a4f3d           NaN
ef42a938-2203-446e-8d28-9fd27c6d3146           NaN
68f5d81b-7f91-40a0-8504-0ec320a86de4           NaN
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a           NaN

properties.notif:notifications
uid
afee12c7-177e-4a68-9539-8ccb68442503           NaN
7d390121-17e8-43bf-a357-9d06b79d2d47           NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6           NaN
e82a8173-b379-41ac-b319-aa058a04fcfb           NaN
4b12c47c-84f5-4132-b362-22b864379a67           NaN
...
5cde49c6-9e7e-4bd2-b22a-3239f643379d           ...
0273eadc-851a-4b68-8020-8041700a4f3d           NaN
ef42a938-2203-446e-8d28-9fd27c6d3146           NaN
68f5d81b-7f91-40a0-8504-0ec320a86de4           NaN
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a           NaN

[13228 rows x 487 columns]

```

1.3 Constitution du périmètre

On conserve les produits qui : - sont de type Epicerie ou Boisson non alcoolisée - portent une liste d'ingrédients - sont en qualité : - soit ont terminé le processus de migration, soit ont été créés après la reprise initiale - et ont le statut "Validé"

```
[18]: # filter by product type
type_mask = df['properties.pprodtop:typeOfProduct'].isin(['grocery', 'nonAlcoholicDrink'])

# keep only those who have ingredients
ingredient_mask = pd.notna(df['properties.pprod:ingredientsList'])

# filter out those who have not finished migration
df['begin_mig'] = df['facets'].apply(lambda x: 'beginningMigration' in x)
df['end_mig'] = df['facets'].apply(lambda x: 'endMigration' in x)
migration_mask = df.loc[:, 'end_mig'] | ~df.loc[:, 'begin_mig']

# filter out those who are not validated
status_mask = (df.loc[:, 'state'] == 'product.validate')
```

```
scope_mask = type_mask & ingredient_mask & migration_mask & status_mask  
scope_df = df.loc[scope_mask]  
print(f'After filters, there are {len(scope_df)} records in the dataset.')
```

After filters, there are 3412 records in the dataset.

gt_based_model

Pierre MASSÉ

May 7, 2020

1 Modèle basé sur les données manuellement étiquetées

L'objet de ce notebook est de mettre en place le modèle basé sur les données manuellement étiquetées.

1.1 Récupération des données

1.1.1 Préambule technique

```
[1]: # setting up sys.path for relative imports
from pathlib import Path
import sys
project_root = str(Path(sys.path[0]).parents[1].absolute())
if project_root not in sys.path:
    sys.path.append(project_root)
```

```
[2]: # imports and customization of display
import os
import pandas as pd
pd.options.display.min_rows = 6
pd.options.display.width=108
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline

from src.pimapi import Requester
from src.pimest import ContentGetter
from src.pimest import PathGetter
from src.pimest import PDFContentParser
from src.pimest import BlockSplitter
from src.pimest import SimilaritySelector
```

1.1.2 Chargement du fichier des données manuellement étiquetées

On commence par charger le fichier csv contenant les données manuellement étiquetées.

```
[3]: ground_truth_df = pd.read_csv(Path('..') / '..' / 'ground_truth' / 'manually_labelled_ground_truth.csv',
                                    sep=';',
                                    encoding='latin-1',
                                    index_col='uid')
ground_truth_df.head()
```

```
[3]:                                     designation \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Concentré liquide Asian en bouteille 980 ml CHEF
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Pain burger curry 80 g CREATIV BURGER
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 Macaroni en sachet 500 g PANZANI
528d4be3-425c-4f8b-8a87-12f1bc645ddd Fève de Tonka en sachet 100 g COMPTOIR COLONIAL
51b38427-b2ea-4c56-93e8-4242361ef31b Caviar d'aubergine en pot 500 g PUGET RESTAURA...
                                                ingredients
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Eau, maltodextrine, sel, arômes, sucre, arôme ...
```

```
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Farine de blé T65, eau, levure, vinaigre de ci...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 - 100% Semoule de BLE dur de qualité supérieur...
528d4be3-425c-4f8b-8a87-12f1bc645ddd fève de tonka (graines ridées de 25 à 50mm de ...
51b38427-b2ea-4c56-93e8-4242361ef31b Aubergine 60,5% (aubergine, huile de tournesol...
```

```
[4]: ground_truth_uids = list(ground_truth_df.index)
ground_truth_uids[:5]
```

```
[4]: ['a0492df6-9c76-4303-8813-65ec5ccbfa70',
'd183e914-db2f-4e2f-863a-a3b2d054c0b8',
'ab48a1ed-7a3d-4686-bb6d-ab4f367cada8',
'528d4be3-425c-4f8b-8a87-12f1bc645ddd',
'51b38427-b2ea-4c56-93e8-4242361ef31b']
```

1.1.3 Pipeline d'acquisition du contenu des données

On commence par construire un premier pipeline d'acquisition des données. Il fonctionne en 3 étapes : - détermination du chemin vers lequel aller chercher les fiches techniques - récupération du contenu binaire du fichier - conversion de ce contenu binaire en texte

```
[5]: acqui_pipe = Pipeline([('PathGetter', PathGetter(ground_truth_uids=ground_truth_uids,
                                                    train_set_path=Path('..') / '..' / 'ground_truth',
                                                    ground_truth_path=Path('..') / '..' / 'ground_truth',
                                                    )),
                        ('ContentGetter', ContentGetter(missing_file='to_nan')),
                        ('ContentParser', PDFContentParser(None_content='to_empty')),
                       ],
                      verbose=True)
```

```
[6]: texts_df = acqui_pipe.fit_transform(ground_truth_df)
texts_df
```

```
[Pipeline] ... (step 1 of 3) Processing PathGetter, total= 0.1s
[Pipeline] ... (step 2 of 3) Processing ContentGetter, total= 0.6s
Launching 8 processes.
[Pipeline] ... (step 3 of 3) Processing ContentParser, total= 37.1s
```

```
[6]: designation \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Concentré liquide Asian en bouteille 980 ml CHEF
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Pain burger curry 80 g CREATIV BURGER
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 Macaroni en sachet 500 g PANZANI
...
e67341d8-350f-46f4-9154-4dbbb8035621 PRÉPARATION POUR CRÈME BRÛLÉE BIO 6L
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 Céréales instantanées en poudre saveur caramel...
0faad739-ea8c-4f03-b62e-51ee592a0546 FARINE DE BLÉ TYPE 45, 10KG

ingredients \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Eau, maltodextrine, sel, arômes, sucre, arôme ...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Farine de blé T65, eau, levure, vinaigre de ci...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 - 100% Semoule de BLE dur de qualité supérieur...
...
e67341d8-350f-46f4-9154-4dbbb8035621 Sucre roux de canne* (64%), amidon de maïs*, ...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 Farine 87,1 % (Blé (GLUTEN), Blé hydrolysé (GL...
0faad739-ea8c-4f03-b62e-51ee592a0546 Farine de blé T45

path \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 ../../ground_truth/a0492df6-9c76-4303-8813-65e...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 ../../ground_truth/d183e914-db2f-4e2f-863a-a3b...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 ../../ground_truth/ab48a1ed-7a3d-4686-bb6d-ab4...
...
e67341d8-350f-46f4-9154-4dbbb8035621 ../../ground_truth/e67341d8-350f-46f4-9154-4db...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 ../../ground_truth/a8f6f672-20ac-4ff8-a8f2-3bc...
```

```

Ofaad739-ea8c-4f03-b62e-51ee592a0546  ../../ground_truth/Ofaad739-ea8c-4f03-b62e-51e...
                                         content  \
uid
a0492df6-9c76-4303-8813-65ec5ccbf70 b'%PDF-1.5\r\n%\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 b'%PDF-1.5\r%\xe2\xe3\xcf\xd3\r\n4 0 obj\r</L...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 b'%PDF-1.4\n%\xc7\xec\x8f\x a2\n5 0 obj\n</Len...
...
e67341d8-350f-46f4-9154-4dbbb8035621 b'%PDF-1.7\r\n%\xb5\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 b'%PDF-1.5\r\n%\xb5\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 b'%PDF-1.5\r\n%\xb5\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...

                                         text
uid
a0492df6-9c76-4303-8813-65ec5ccbf70 Concentré Liquide Asian CHEF® \n\nBouteille de...
d183e914-db2f-4e2f-863a-a3b2d054c0b8

ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 Direction Qualité \n\n \n\n \n\nPATES ALIMENTA...
...
e67341d8-350f-46f4-9154-4dbbb8035621 FICHE TECHNIQUE \n\nCREME BRÛLÉE 6L \n\nREF : ...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 81 rue de Sans Souci - CS13754 - 69576 Limones...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 \n1050/10502066400 \n\n10502055300/1050202520...

[500 rows x 5 columns]

```

On peut afficher quelques textes récupérés par le pipeline :

```
[7]: with pd.option_context("max_colwidth", 1000):
    print(texts_df.sample(3, random_state=42)[['text']])
#   (texts_df.sample(3, random_state=42)[['text']]
#     .to_latex(Path('..') / 'tbls' / 'processed_FT.tex',
#     index=False,
#     index_names=False,
#     column_format='p{\linewidth}',
#     na_rep='-', 
#     escape=True,
#     )
#   )
```

```

uid
2892dd68-e3a6-474c-b543-3ebfd3490658 NESCAFÉ® SPÉCIAL FILTRE\n\nDose individuelle de 2 g\nTechnologie
micro-grains\nnCODE EAN (UC)\n\n3033710076017\nnDENOMINATION LEGALE DU PRODUIT\nnDESCRIPTION DU
PRODUIT\nnCafé instantané et café torréfié moulu.\n\nUne dominante Arabica pour l'arôme et une pointe de
Robusta pour le ncorsé, associés à une torréfaction légère pour un café équilibré et peu \namer.\nSachet
dose pour une tasse.\n\nDOSAGE PRECONISÉ\nnMODE OPERATOIRE\nnPour obtenir\nn1 café Court (DA)\n\n1 café
Long (DA)\n\nEau\nn7\nn12\nncl\nncl\nnNESCAFÉ®\n\n SPÉCIAL FILTRE\nn2\nn2\nnng\nnng\nnA reconstituer
avec de l'eau. \nTempérature de l'eau : 75°C\nPour une qualité optimale, utilisez de l'eau
filtrée.\n\nIngrédients : Café instantané, café torréfié moulu (3%).\n\nINGRÉDIENTS\nnPROFIL
GUSTATIF\nnIntensité\nnConditionné sous atmosphère protectrice.\n\nENGAGEMENT QUALITÉ\nn- NESTLÉ a un
système de management de la qualité, le NMS (NESTLÉ \nManagement System), en cohérence avec les systèmes ISO
9001 ...
a57c1561-b88e-4694-8bd8-55623f2afa17 LENTILLES BLONDES 4mm\nnRéférence PQG007-3.22.1\nnVersion\nnDate
d'application :nPage 1/2\nnG\nn15/10/2019\nnPrésentation\nnCaractéristi -\n\nques
\n\nphysico-\nchimiques\nnDéfinition\nnOrigine\nnDenomination \nlégale\nnLentilles de couleur brun clair.
Elles sont de forme biconvexe et \npos possèdent une peau assez épaisse. Leur diamètre est compris \nentre 4mm
et 5mm\nnChine, Canada, France, Italie, USA, Turquie\nnLentilles blondes\nnProcess\nnNettoyage,
épierrage, triages\nnConservation\nn36 mois à l'abri de la chaleur et de l'humidité\nnCritères
d'analyses\nnMoyenne/Tolérance\nnMéthodes\nnHumidité\nnMatières minérales étrangères\nnMatières végétales
étrangères\nGraines\nnImpropres\nnBrisées\nnGermées\nnCalibre 4-5 mm\nn11,5% / 16%\nmax\n0,05% / 1%\nmax\n0,15%
/ 0,5%\nmax\nn0,5% / 1%\nmax\nn0,4% / 1%\nmax\nn0,05% / 1%\nmax\nn95% / 90%\nmin\nnNF V03707\nnMicrobiologie\nnIl
n'existe pas de réglementation concernant les exigences microbiologiques \npour ce
produit.\n\nPesticides\nnMét...
3634fb1e-ee79-41d1-8aaa-084c1fae5bd5 FICHE TECHNIQUE \n\nPRODUIT FINI\nn000100\nnPurée de Poire Sans
Sucres Ajoutés\nnDate d'application: 05/05/2014\nnPage: 1/2\nnCoupelles Aluminium 120 x 95
g\nn\nDéfinition\nnCe produit est une purée de fruits obtenue à partir des parties comestibles des fruits
(après broyage et sans \nconcentration notable).\nCe produit est sans sucres ajoutés: il contient uniquement
les sucres naturellement présents dans les fruits.\nLa purée présente une texture homogène et légèrement
```

granuleuse.\n\nLa stabilité du produit est obtenue par pasteurisation et dosage à chaud.\n\nAspects nutritionnels\nDésignation et liste des ingrédients\nValeurs nutritionnelles (pour 100 g)\nDésignation légale :\nPurée de Poires sans sucres ajoutés *\nContient les sucres naturellement présents dans \nles fruits\nListe des ingrédients :\nPoire 99,9%, antioxydant: acide ascorbique.\nMatières grasses\nEnergie\n65 kcal\n273 kJ\nndont acides gras saturés\nGlucides\nFibres alimentaires\nPro...

Name: text, dtype: object

1.2 Découpage en blocs

On découpe les longs textes en blocs. Chaque texte devient une liste de strings plus court.

```
[8]: def splitter(text):
    return(text.split('\n\n'))
```

```
[9]: split_transfo = BlockSplitter(splitter_func=splitter)
splitted_df = split_transfo.fit_transform(texts_df)
splitted_df
```

Launching 8 processes.

```
[9]: designation \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Concentré liquide Asian en bouteille 980 ml CHEF
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Pain burger curry 80 g CREATIV BURGER
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 Macaroni en sachet 500 g PANZANI
...
e67341d8-350f-46f4-9154-4dbbb8035621 PRÉPARATION POUR CRÈME BRÛLÉE BIO 6L
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 Céréales instantanées en poudre saveur caramel...
0faad739-ea8c-4f03-b62e-51ee592a0546 FARINE DE BLÉ TYPE 45, 10KG

ingredients \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Eau, maltodextrine, sel, arômes, sucre, arôme ...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Farine de blé T65, eau, levure, vinaigre de ci...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 - 100% Semoule de BLE dur de qualité supérieur...
...
e67341d8-350f-46f4-9154-4dbbb8035621 Sucre roux de canne* (64%), amidon de maïs*, ...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 Farine 87,1 % (Blé (GLUTEN), Blé hydrolysé (GL...
0faad739-ea8c-4f03-b62e-51ee592a0546 Farine de blé T45

path \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 ../../ground_truth/a0492df6-9c76-4303-8813-65...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 ../../ground_truth/d183e914-db2f-4e2f-863a-a3b...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 ../../ground_truth/ab48a1ed-7a3d-4686-bb6d-ab4...

...
e67341d8-350f-46f4-9154-4dbbb8035621 ../../ground_truth/e67341d8-350f-46f4-9154-4db...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 ../../ground_truth/a8f6f672-20ac-4ff8-a8f2-3bc...
0faad739-ea8c-4f03-b62e-51ee592a0546 ../../ground_truth/0faad739-ea8c-4f03-b62e-51e...

content \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 b'%PDF-1.5\r\n%\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 b'%PDF-1.5\r%\xe2\xe3\xcf\xd3\r\n4 0 obj\r</L...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 b'%PDF-1.4%\xc7\xec\x8f\xa2\n5 0 obj\n</Len...
...
e67341d8-350f-46f4-9154-4dbbb8035621 b'%PDF-1.7\r\n%\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 b'%PDF-1.5\r\n%\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...
0faad739-ea8c-4f03-b62e-51ee592a0546 b'%PDF-1.5\r\n%\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...

text \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Concentré Liquide Asian CHEF® \n\nBouteille de...
d183e914-db2f-4e2f-863a-a3b2d054c0b8
```

```

ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 Direction Qualité \n\n \n\n \n\nPATES ALIMENTA...
...
e67341d8-350f-46f4-9154-4dbbb8035621 FICHE TECHNIQUE \n\nCREME BRÛLÉE 6L \n\nREF : ...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 81 rue de Sans Souci - CS13754 - 69576 Limones...
0faad739-ea8c-4f03-b62e-51ee592a0546 \n1050/10502066400 \n\n10502055300/1050202520...

blocks
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 [Concentré Liquide Asian CHEF®, Bouteille de ...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 [
]
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 [Direction Qualité , , , PATES ALIMENTAIRES ...
...
e67341d8-350f-46f4-9154-4dbbb8035621 [FICHE TECHNIQUE , CREME BRÛLÉE 6L , REF : NAP...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 [81 rue de Sans Souci - CS13754 - 69576 Limone...
0faad739-ea8c-4f03-b62e-51ee592a0546 [ \n1050/10502066400 , 10502055300/10502025200...

[500 rows x 6 columns]

```

On peut afficher un exemple de texte découpé en blocs :

```
[10]: sep = '\n-----\n'
sample = splitted_df.sample(1, random_state=39)[['blocks']].iloc[0]
print(sep.join(sample))

tex_str = (
    pd.DataFrame(sample, columns=['Bloc'])
    .to_latex(column_format='p{10cm}', 
              index=False,
              index_names=False,
              escape=True,
              )
    .replace(r'\textbackslash n', '\\newline')
)
#with open(Path('..') / 'tbls' / 'block_example.tex', mode='w') as file:
#    file.write(sep.join(sample).replace('\n', r' \newline '))
```

30/12/19

Date d'impression :

Remarque :

Les informations contenues dans cette fiche technique sont données de bonne foi, en l'état actuel de nos connaissances, et selon les indications communiquées par le producteur ou le fournisseur. Il appartient au client de vérifier la conformité de la marchandise par rapport à l'usage qu'il en fait.

Création :

12/06/12

12 rue René Cassin
37390 NOTRE DAME

Tél :
02 47 85 55 00
Fax :02 47 41 33 32

FICHE TECHNIQUE

Mélange du trappeur, 70 g
Trapper blend, 70g

Code article KEREX
Nom latin (si disponible)
/ EAN Code

Code barre

/ KEREX Code

/ (Latin name)

TEEPTRAPPEUR
X
3760063322262

Poids net
Poids brut
Origine

/ net weight
/ gross weight
/ Origin

0,07 Kilogramme
0,125 Kilogramme
CANADA

/ General information

Informations générales
DLUO conseillée / "Best before date" recommended
Nomenclature douanière / Customs code
Conditions idéales de stockage
/ Conditions of storage
Ingrédients :

Conserver dans un endroit frais et sec
Store in a cool dry place

5 ans / 5 years
0910999900

Sucre, poivre noir, coriandre, légumes déshydratés (ail, oignon, poivron rouge), sel de mer, sucre d'érable, arôme d'érable naturel, huile végétale (canola)
Sugar, black pepper, coriander, dehydrated vegetables (garlic, onion, red bell pepper), sea salt, maple sugar, natural maple aroma, vegetable oil (canola)

/ Ingredients

Contaminants / Contaminating
Ionisation / Irradiation

OGM / GMO

Pesticides/ Pesticides

Métaux Lourds

/ Heavy Metals

Allergènes et leurs dérivés (si présents)
/ Allergens (if existing)

Conformité à la directive 1999/2/CE (22/02/99)
Produit non ionisé et ne contenant pas d'ingrédients ionisés.
Not irradiated
accordingly with the Reg 1999/2/CE (22/02/99).
Free from GMO
Ne contient pas d'OGM, est non soumis à l'étiquetage sur les OGM
Conforme à la directive 396/2005 /CE
In accordance with Reg 396/2005 /CE.

Conforme au règlement 1881/2006 /CE
In accordance with Reg 1881/2006 /CE..

Gluten
Crustacés
Oeufs
Poisson
Soja
Lait
Fruits à coque - Arachides
Céleri
Moutarde
Sésame
Sulfites
Lupin
Mollusques

/ Gluten
/ Crustaceans
/ Eggs
/ Fish
/ Soy
/ Milk
/ Peanuts and Treenuts
/ Celery and celeriac
/ Mustarde
/ Sésame
/ Sulphites
/ Lupin
/ Shellfish

Absence
Absence

Caractères microbiologiques

/ Microbiological characteristics

Microorganismes aérobie 30 °C
Escherichia coli
Salmonelles
Levures
Moisisures
Aflatoxine Total
Aflatoxine B1

/ Total plat count (APC)
E. Coli
/
/ Salmonella
/ Yeasts
/ Moulds
/ Total aflatoxin
B1 aflatoxin
/

NF V05-051 < 6 000 000 / g

```

NF V08-053 < 10 / g
NF V08-052 Absence dans 25g
NF V08-059 < 10 000 / g
NF V08-059 < 10 000 / g
Kit Enzymatique < 10 ppb
Kit Enzymatique < 5 ppb
-----
```

1.3 Train / Test split

On procède au découpage en un jeu d'entraînement et un jeu de test en gardant 400 produits pour l'entraînement et 100 produits pour le test :

```
[11]: train, test = train_test_split(splitted_df, train_size=400, random_state=42)
```

1.4 Entraînement sur le jeu d'entraînement

On entraîne un modèle `SimilaritySelector`, sur le set d'entraînement :

```
[13]: model = SimilaritySelector(similarity='projection')
```

```
[15]: model.fit(train['blocks'], train['ingredients'])
```

```
[15]: <src.pimest.SimilaritySelector at 0x7f3cc41371c0>
```

```
[26]: predicted = pd.Series(model.predict(test['blocks']),
                           index=test.index,
                           name='predicted')
predicted = pd.concat([test['ingredients'], predicted], axis=1)
predicted
```

```
[26]: ingredients \
uid
2892dd68-e3a6-474c-b543-3ebfd3490658     Café instantané, café torrefié moulu (3%).
a57c1561-b88e-4694-8bd8-55623f2afa17     Lentilles blondes
3634fb1e-ee79-41d1-8aaa-084c1fae5bd5     Poire 99,9%, antioxydant: acide ascorbique.
...
ebfc9e73-5d91-4b45-8331-8c8f9bed3bb3     Jus d'orange à base de concentré
c33aa83e-a502-4339-a8e0-c56db2e59e69     Farine de BLÉ, sucre, huile de colza,, cacao m...
54f40033-f9cf-411c-81a5-11974f6715aa     Piment rouge fort équeuté* (85%), cumin, ail m...
                                                 ...
predicted
uid
2892dd68-e3a6-474c-b543-3ebfd3490658 - NESTLÉ a un système de management de la qual...
a57c1561-b88e-4694-8bd8-55623f2afa17 Cette fiche technique n'a pas de valeur contra...
3634fb1e-ee79-41d1-8aaa-084c1fae5bd5 Ce produit est une purée de fruits obtenue à p...
...
ebfc9e73-5d91-4b45-8331-8c8f9bed3bb3   \n \nVALEURS NUTRITIONNELLES pour 100mL / NUT...
c33aa83e-a502-4339-a8e0-c56db2e59e69   Ingrédients : Farine de BLÉ, sucre, huile de c...
54f40033-f9cf-411c-81a5-11974f6715aa   A) Ingrédients : \n \nPiment rouge fort équ...
                                                 ...
[100 rows x 2 columns]
```

```
[27]: predicted['pred_len'] = predicted['predicted'].apply(len)
sub_sample = predicted.loc[predicted['pred_len'] <= 500, ['ingredients', 'predicted']]
sub_sample
```

```
[27]: ingredients \
uid
2892dd68-e3a6-474c-b543-3ebfd3490658     Café instantané, café torrefié moulu (3%).
3634fb1e-ee79-41d1-8aaa-084c1fae5bd5     Poire 99,9%, antioxydant: acide ascorbique.
345591f4-d887-4ddc-bb40-21337fa9269d     Gésier de dinde émincé 50%, graisse de canard ...
```

...

ebfc9e73-5d91-4b45-8331-8c8f9bed3bb3	Jus d'orange à base de concentré
c33aa83e-a502-4339-a8e0-c56db2e59e69	Farine de BLÉ, sucre, huile de colza,, cacao m...
54f40033-f9cf-411c-81a5-11974f6715aa	Piment rouge fort équeuté* (85%), cumin, ail m...

predicted

uid

2892dd68-e3a6-474c-b543-3ebfd3490658	- NESTLÉ a un système de management de la qual...
3634fb1e-ee79-41d1-8aaa-084c1fae5bd5	Ce produit est une purée de fruits obtenue à p...
345591f4-d887-4ddc-bb40-21337fa9269d	Gésier de dinde émincé 50%, graisse de canard...

...

ebfc9e73-5d91-4b45-8331-8c8f9bed3bb3	\n \nVALEURS NUTRITIONNELLES pour 100mL / NUT...
c33aa83e-a502-4339-a8e0-c56db2e59e69	Ingrédients : Farine de BLÉ, sucre, huile de c...
54f40033-f9cf-411c-81a5-11974f6715aa	A) Ingrédients : \n \nPiment rouge fort équ...

[76 rows x 2 columns]

[31]: `sub_sample.sample(20, random_state=41).replace(r'^\s*$', np.nan, regex=True)`

[31]:

uid	ingredients \
d1be6f74-1e0e-4631-bb4a-6b16b9fc908f	sucre*, LAIT en poudre*, beurre de cacao*, pât... NaN
49b11281-34ea-44b0-a11c-4ae21d4c58e3	Amidon de maïs* - Lait écrémé* - Sel - Fécule ... NaN
d59d96cb-0230-4090-8220-78ce8496fd91	semoule de blé dur supérieure et de l'eau NaN
5adc7512-6168-4966-ae3f-f6ec133bf56e	Eau, maltodextrine, sel, arômes, sucre, arôme ... Sucre, cacao maigre en poudre (beurre de cacao... Sucre; sirop de glucose; graisse de palme; hum... Gésier de dinde émincé 50%, graisse de canard ... Purée de tomates mi réduite (64%), sucre, vina... NaN
75088d85-f350-4d81-a7f4-954411ba089e	Sirop de glucose-fructose, framboises 35%, suc... Café instantané, café torréfié moulu (3%). sucre, pâte de cacao, beurre de cacao, cacao m... Eau, huile de tournesol, beurre 9,5 %, jaune d... Piment rouge fort équeuté* (85%), cumin, ail m... Sucre, amidon de maïs, arôme vanille Salicornes de culture, eau, sel, acide citrique cèpes 70% (Boletus edulis et respective famill... Eau, haricots verts, sel. NaN
a0492df6-9c76-4303-8813-65ec5ccbfa70	NaN
e521bd01-f2bb-4e00-9ae0-0151a1c7a047	NaN
8dec0469-c9f5-4139-be25-efa258959444	NaN
345591f4-d887-4ddc-bb40-21337fa9269d	NaN
41da4d6f-7e9f-4f95-bf2a-2acdd7138cd9	NaN
21233a00-bc20-40fc-acb9-ee2e2321cac2	NaN
9ef0d351-4982-4a2d-88a9-85573dc396dc	NaN
2892dd68-e3a6-474c-b543-3ebfd3490658	NaN
df1caa23-9714-4659-803b-33501d64eedad	NaN
7f622727-e4ad-45cc-9af4-4509acf91154	NaN
54f40033-f9cf-411c-81a5-11974f6715aa	NaN
536361db-1bbb-4e64-ae53-d970eeac7db2	NaN
a2418174-e16a-41e0-ac14-c87208fb3529	NaN
046cdb1f-1915-4916-8874-902cc5ec73be	NaN
b7d7621a-fcdd-4487-9b38-e07fae698c4a	NaN

predicted

uid	
d1be6f74-1e0e-4631-bb4a-6b16b9fc908f	Liste des Ingrédients:\nsucre*, LAIT en poudr...
49b11281-34ea-44b0-a11c-4ae21d4c58e3	NaN
d59d96cb-0230-4090-8220-78ce8496fd91	Amidon de maïs* - Lait écrémé* - Sel - Fécule ... Ingrédients: semoule de blé dur supérieure et ... Boisson gazeuse aromatisée au jus de fruit à b...
5adc7512-6168-4966-ae3f-f6ec133bf56e	Eau, maltodextrine, sel, arômes, sucre, arôme ... Sucre, cacao maigre en poudre (beurre de cacao... Liste ingrédients : Sirop de glucose-fructose,... - NESTLÉ a un système de management de la qual...
75088d85-f350-4d81-a7f4-954411ba089e	NaN
a0492df6-9c76-4303-8813-65ec5ccbfa70	NaN
e521bd01-f2bb-4e00-9ae0-0151a1c7a047	NaN
8dec0469-c9f5-4139-be25-efa258959444	NaN
345591f4-d887-4ddc-bb40-21337fa9269d	NaN
41da4d6f-7e9f-4f95-bf2a-2acdd7138cd9	NaN
21233a00-bc20-40fc-acb9-ee2e2321cac2	NaN
9ef0d351-4982-4a2d-88a9-85573dc396dc	NaN
2892dd68-e3a6-474c-b543-3ebfd3490658	NaN
df1caa23-9714-4659-803b-33501d64eedad	NaN
7f622727-e4ad-45cc-9af4-4509acf91154	NaN
54f40033-f9cf-411c-81a5-11974f6715aa	NaN
536361db-1bbb-4e64-ae53-d970eeac7db2	NaN
a2418174-e16a-41e0-ac14-c87208fb3529	NaN
046cdb1f-1915-4916-8874-902cc5ec73be	NaN
b7d7621a-fcdd-4487-9b38-e07fae698c4a	NaN

On constitue une table pour intégration dans le rapport :

```
[33]: with pd.option_context("max_colwidth", 100000):
    tex_str = (
        sub_sample.sample(20, random_state=41)
            .replace(r'^\s*$', np.nan, regex=True)
            .to_latex(index=False,
                      index_names=False,
                      column_format='p{7cm}p{7cm}',
                      na_rep='<rien>',
                      longtable=True,
                      header=["Liste d'ingrédients cible", "Liste d'ingrédients prédictive"],
                      label='tbl:GT_prediction_sample',
                      caption="Extrait des résultats de la prédiction",
            )
            .replace(r'\textbackslash n', r' \newline ')
            .replace(r'\\', r'\\ \hline')
    )

with open(Path('..') / 'tbls' / 'GT_prediction_sample.tex', 'w') as file:
    file.write(tex_str)
```

Performance_measurement

Pierre MASSÉ

May 8, 2020

1 Mesure de la performance du modèle

L'objet de ce notebook est d'illustrer la méthodologie de mesure de la performance du modèle.

1.1 Préambule technique

```
[1]: # setting up sys.path for relative imports
from pathlib import Path
import sys
project_root = str(Path(sys.path[0]).parents[1].absolute())
if project_root not in sys.path:
    sys.path.append(project_root)

[46]: # imports and customization of display
import os
from functools import partial
import numpy as np
import pandas as pd
pd.options.display.min_rows = 6
pd.options.display.width=108
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, cross_validate
from sklearn.pipeline import Pipeline
from matplotlib import pyplot as plt

from src.pimest import ContentGetter
from src.pimest import PathGetter
from src.pimest import PDFContentParser
from src.pimest import BlockSplitter
from src.pimest import SimilaritySelector
from src.pimest import custom_accuracy
```

1.2 Acquisition des données

On récupère les données manuellement étiquetées et on les intègre dans un dataframe

```
[3]: ground_truth_df = pd.read_csv(Path('..') / '..' / 'ground_truth' / 'manually_labelled_ground_truth.csv',
                                    sep=';',
                                    encoding='latin-1',
                                    index_col='uid')
ground_truth_uids = list(ground_truth_df.index)

acqui_pipe = Pipeline([('PathGetter', PathGetter(ground_truth_uids=ground_truth_uids,
                                                train_set_path=Path('..') / '..' / 'ground_truth',
                                                ground_truth_path=Path('..') / '..' / 'ground_truth',
                                                )),
                       ('ContentGetter', ContentGetter(missing_file='to_nan')),
                       ('ContentParser', PDFContentParser(none_content='to_empty')),
                      ],
                     verbose=True)
```

```

texts_df = acqui_pipe.fit_transform(ground_truth_df)
texts_df

[Pipeline] ... (step 1 of 3) Processing PathGetter, total= 0.1s
[Pipeline] ... (step 2 of 3) Processing ContentGetter, total= 0.1s
Launching 8 processes.
[Pipeline] ... (step 3 of 3) Processing ContentParser, total= 37.4s

[3]: designation \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Concentré liquide Asian en bouteille 980 ml CHEF
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Pain burger curry 80 g CREATIV BURGER
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 Macaroni en sachet 500 g PANZANI
...
e67341d8-350f-46f4-9154-4dbbb8035621 PRÉPARATION POUR CRÈME BRÛLÉE BIO 6L
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 Céréales instantanées en poudre saveur caramel...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 FARINE DE BLÉ TYPE 45, 10KG

ingredients \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Eau, maltodextrine, sel, arômes, sucre, arôme ...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Farine de blé T65, eau, levure, vinaigre de ci...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 - 100% Semoule de BLE dur de qualité supérieur...
...
e67341d8-350f-46f4-9154-4dbbb8035621 Sucre roux de canne* (64%), amidon de maïs*, ...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 Farine 87,1 % (Blé (GLUTEN)), Blé hydrolysé (GL...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 Farine de blé T45

path \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 ../../ground_truth/a0492df6-9c76-4303-8813-65e...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 ../../ground_truth/d183e914-db2f-4e2f-863a-a3b...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 ../../ground_truth/ab48a1ed-7a3d-4686-bb6d-ab4...

...
e67341d8-350f-46f4-9154-4dbbb8035621 ../../ground_truth/e67341d8-350f-46f4-9154-4db...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 ../../ground_truth/a8f6f672-20ac-4ff8-a8f2-3bc...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 ../../ground_truth/Ofaad739-ea8c-4f03-b62e-51e...

content \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 b'%'PDF-1.5\r\n%\xb5\xb5\xb5\xb5\r\nn1 0 obj\r\n...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 b'%'PDF-1.5\r%\xe2\xe3\xcf\xd3\r\nn4 0 obj\r</...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 b'%'PDF-1.4\r%\xc7\xec\x8f\xa2\n5 0 obj\n</Len...

...
e67341d8-350f-46f4-9154-4dbbb8035621 b'%'PDF-1.7\r\n%\xb5\xb5\xb5\xb5\r\nn1 0 obj\r\...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 b'%'PDF-1.5\r\n%\xb5\xb5\xb5\xb5\r\nn1 0 obj\r\...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 b'%'PDF-1.5\r\n%\xb5\xb5\xb5\xb5\r\nn1 0 obj\r\...

text
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Concentré Liquide Asian CHEF® \n\nBouteille de...
d183e914-db2f-4e2f-863a-a3b2d054c0b8

ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 Direction Qualité \n\n \n\n \nPATES ALIMENTA...
...
e67341d8-350f-46f4-9154-4dbbb8035621 FICHE TECHNIQUE \n\nCREME BRÛLÉE 6L \n\nREF : ...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 81 rue de Sans Souci - CS13754 - 69576 Limones...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 \n1050/10502066400 \n\n10502055300/1050202520...

[500 rows x 5 columns]

```

On splitte les textes en blocs de manière basique.

```
[4]: def splitter(text):
    return(text.split('\n\n'))
```

```

split_transfo = BlockSplitter(splitter_func=splitter)
splitted_df = split_transfo.fit_transform(texts_df)
splitted_df

```

Launching 8 processes.

```

[4]:                                            designation \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Concentré liquide Asian en bouteille 980 ml CHEF
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Pain burger curry 80 g CREATIV BURGER
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 Macaroni en sachet 500 g PANZANI
...
e67341d8-350f-46f4-9154-4dbbb8035621 PRÉPARATION POUR CRÈME BRÛLÉE BIO 6L
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 Céréales instantanées en poudre saveur caramel...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 FARINE DE BLÉ TYPE 45, 10KG

                                            ingredients \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Eau, maltodextrine, sel, arômes, sucre, arôme ...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 Farine de blé T65, eau, levure, vinaigre de ci...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 - 100% Semoule de BLE dur de qualité supérieur...
...
e67341d8-350f-46f4-9154-4dbbb8035621 Sucre roux de canne*° (64%), amidon de maïs*, ...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 Farine 87,1 % (Blé (GLUTEN), Blé hydrolysé (GL...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 Farine de blé T45

                                            path \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 ../../ground_truth/a0492df6-9c76-4303-8813-65e...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 ../../ground_truth/d183e914-db2f-4e2f-863a-a3b...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 ../../ground_truth/ab48a1ed-7a3d-4686-bb6d-ab4...

...
e67341d8-350f-46f4-9154-4dbbb8035621 ../../ground_truth/e67341d8-350f-46f4-9154-4db...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 ../../ground_truth/a8f6f672-20ac-4ff8-a8f2-3bc...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 ../../ground_truth/Ofaad739-ea8c-4f03-b62e-51e...

                                            content \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 b'%PDF-1.5\r\n%\xb5\xb5\xb5\r\n1 0 obj\r\n...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 b'%PDF-1.5\r%\xe2\xe3\xcf\xd3\r\n4 0 obj\r<</L...
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 b'%PDF-1.4\n%\xc7\xec\x8f\x2a\n5 0 obj\r\n<</Len...

...
e67341d8-350f-46f4-9154-4dbbb8035621 b'%PDF-1.7\r\n%\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 b'%PDF-1.5\r\n%\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 b'%PDF-1.5\r\n%\xb5\xb5\xb5\xb5\r\n1 0 obj\r\n...

                                            text \
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 Concentré Liquide Asian CHEF® \n\nBouteille de...
d183e914-db2f-4e2f-863a-a3b2d054c0b8

ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 Direction Qualité \n\n \n\n \nPATES ALIMENTA...
...
e67341d8-350f-46f4-9154-4dbbb8035621 FICHE TECHNIQUE \n\nCREME BRÛLÉE 6L \n\nREF : ...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 81 rue de Sans Souci - CS13754 - 69576 Limones...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 \n1050/10502066400 \n\n10502055300/1050202520...

                                            blocks
uid
a0492df6-9c76-4303-8813-65ec5ccbfa70 [Concentré Liquide Asian CHEF®, Bouteille de ...
d183e914-db2f-4e2f-863a-a3b2d054c0b8 [
]
ab48a1ed-7a3d-4686-bb6d-ab4f367cada8 [Direction Qualité , , , PATES ALIMENTAIRES ...
...
e67341d8-350f-46f4-9154-4dbbb8035621 [FICHE TECHNIQUE , CREME BRÛLÉE 6L , REF : NAP...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3 [81 rue de Sans Souci - CS13754 - 69576 Limone...
Ofaad739-ea8c-4f03-b62e-51ee592a0546 [\n1050/10502066400 , 10502055300/1050202520...

```

```
[500 rows x 6 columns]
```

1.3 Train/Test split, entraînement et transformation

On effectue classiquement les étapes de train/test split, on entraîne le modèle sur le set d'entraînement et on le lance sur le set de test.

```
[5]: train, test = train_test_split(splitted_df, train_size=400, random_state=42)
model = SimilaritySelector(similarity='projection')
model.fit(train['blocks'], train['ingredients'])
predicted = pd.Series(model.predict(test['blocks']),
                       index=test.index,
                       name='predicted')
predicted = pd.concat([test['ingredients'], predicted], axis=1)
predicted
```

```
[5]:                                     ingredients \
uid
2892dd68-e3a6-474c-b543-3ebfd3490658      Café instantané, café torrefié moulu (3%).
a57c1561-b88e-4694-8bd8-55623f2afa17      Lentilles blondes
3634fb1e-ee79-41d1-8aaa-084c1fae5bd5      Poire 99,9%, antioxydant: acide ascorbique.
...
ebfc9e73-5d91-4b45-8331-8c8f9bed3bb3      Jus d'orange à base de concentré
c33aa83e-a502-4339-a8e0-c56db2e59e69      Farine de BLÉ, sucre, huile de colza,, cacao m...
54f40033-f9cf-411c-81a5-11974f6715aa      Piment rouge fort équeuté* (85%), cumin, ail m...
                                                 ...
predicted
uid
2892dd68-e3a6-474c-b543-3ebfd3490658      - NESTLÉ a un système de management de la qual...
a57c1561-b88e-4694-8bd8-55623f2afa17      Cette fiche technique n'a pas de valeur contra...
3634fb1e-ee79-41d1-8aaa-084c1fae5bd5      Ce produit est une purée de fruits obtenue à p...
...
ebfc9e73-5d91-4b45-8331-8c8f9bed3bb3      \n \nVALEURS NUTRITIONNELLES pour 100mL / NUT...
c33aa83e-a502-4339-a8e0-c56db2e59e69      Ingrédients : Farine de BLÉ, sucre, huile de c...
54f40033-f9cf-411c-81a5-11974f6715aa      A) Ingrédients : \n \nPiment rouge fort équ...
```

```
[100 rows x 2 columns]
```

1.4 Mesure de la performance : Précision

1.4.1 Approche naïve

Dans cette première version, on calculera une précision brute, où seuls les strings parfaitement identiques sont considérés comme ok.

```
[6]: predicted['result'] = (predicted['ingredients'].fillna('') == predicted['predicted'].fillna(''))
predicted['result'].value_counts()
```

```
[6]: False    99
True      1
Name: result, dtype: int64
```

On a une précision très faible, 1%. L'unique liste d'ingrédients du set de test correctement prédite est la suivante :

```
[7]: print(predicted[predicted['result']].iloc[0, 0])
```

```
Sirop de glucose, sucre, eau, stabilisants (E440i, E440ii, E415), acidifiants (E330, E450i), conservateur (E202).
```

1.4.2 Cross-validation de l'approche naïve

Pour avoir une vision plus précise de la performance du modèle, on peut effectuer une cross-validation sur le set d'entraînement.

On commence par définir une fonction de scoring, qui pourra être appelée par la fonction standard de cross-validation de scikit-learn. Comme précédemment, il s'agit d'une fonction d'accuracy basique :

```
[8]: def accuracy_scorer(estim, X, y):
        y_pred = estim.predict(X)
        return((y_pred == y).mean())
```

On retrouve évidemment le même score que précédemment lorsqu'on utilise cette fonction sur le set de test :

```
[9]: accuracy_scorer(model, test.reset_index()['blocks'], test.reset_index()['ingredients'])
```

[9]: 0.01

Si on lance la cross-validation avec les paramètres par défaut ($cv=5$), on obtient le résultat suivant :

```
[11]: X = splitted_df.reset_index()['blocks'].copy()
y = splitted_df.reset_index()['ingredients'].copy()

cross_val = cross_validate(model,
                           X=X,
                           y=y,
                           scoring=accuracy_scorer,
                           )
print(f'Strict accuracy yields a result of {np.mean(cross_val["test_score"]):.2%} +/-{np.
    std(cross_val["test_score"]):.2%}')
print(cross_val['test_score'])
```

Strict accuracy yields a result of 2.20% +/-0.75%
[0.03 0.03 0.02 0.01 0.02]

On voit que sur chacun des 5 folds (validation sur 80 produits), l'accuracy varie entre 1 et 3%.

Si on trace l'accuracy et la standard deviation pour plusieurs valeurs de cv, on obtient les résultats suivants :

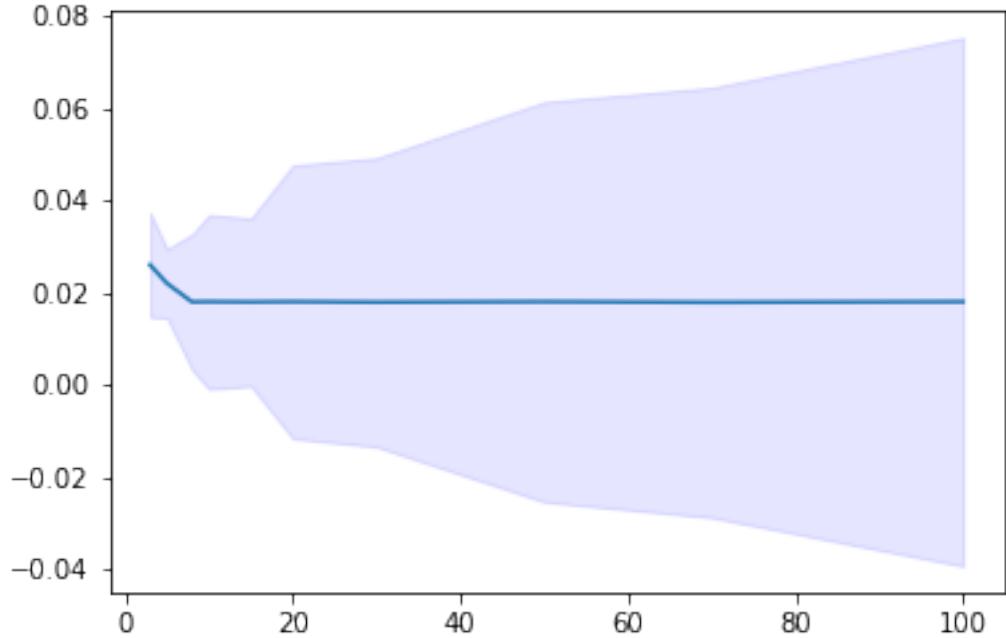
```
[12]: x = [3, 5, 8, 10, 15, 20, 30, 50, 70, 100]
mean = np.array([])
std = np.array([])
for n_cv in x:
    cross_val = cross_validate(model,
                                X=splitted_df['blocks'],
                                y=splitted_df['ingredients'],
                                scoring=accuracy_scorer,
                                cv=n_cv,
                                )
    mean = np.append(mean, [np.mean(cross_val['test_score'])]), axis=0)
    std = np.append(std, [np.std(cross_val['test_score'])]), axis=0)

print('mean:', mean, '\nstandard dev:', std)
```

```
mean: [0.02598418 0.022      0.01795315 0.018      0.01794415 0.018
      0.01789216 0.018      0.01785714 0.018      ]
standard dev: [0.01126571 0.00748331 0.01470225 0.01886796 0.01816919 0.0295973
      0.03127858 0.04331282 0.04656573 0.05723635]
```

```
[13]: fig, ax = plt.subplots()
        ax.plot(x, mean)
        ax.fill_between(x, (mean - std), (mean + std), color='b', alpha=.1)
```

[13]: <matplotlib.collections.PolyCollection at 0x7fa08009e3a0>



Il apparaît que l'accuracy se situe aux alentours de 2%, avec un écart type important si on le compare à cette accuracy.

```
[14]: cross_val = cross_validate(model,
                                X=splitted_df['blocks'],
                                y=splitted_df['ingredients'],
                                scoring=accuracy_scorer,
                                cv=10,
                                )
print(f'Strict accuracy yields a result of {np.mean(cross_val["test_score"]):.2%} +/-{np.
->std(cross_val["test_score"]):.2%}')
print(cross_val['test_score'])
```

```
Strict accuracy yields a result of 1.80% +/-1.89%
[0.04 0. 0. 0.06 0.02 0.02 0. 0.02 0. 0.02]
```

1.4.3 Ajout d'une étape de text-postprocessing

On utilise la fonction `custom_accuracy` définie dans le module pimest pour calculer l'accuracy avec du text processing. Elle prend en paramètre les mêmes attributs que le `CountVectorizer` de scikit-learn, en plus d'un attribut “`tokenize`” qui va tokeniser le résultat (pour prise en compte des whitespace et de la ponctuation).

```
[31]: custom_accuracy(model,
                      test['blocks'].fillna(''),
                      test['ingredients'].fillna(''),
                      tokenize=True,
                      strip_accents='unicode',
                      lowercase=True,
                      )
```

```
[31]: 0.14
```

L'accuracy est maintenant estimée à 14% (vs. 1%) sur le set de test, après entraînement sur le set d'entraînement.

On peut manuellement inspecter les blocks identique, en reproduisant le comportement de la fonction d'accuracy :

```
[49]: def text_processor(text, **kwargs):
    unused_model = CountVectorizer(**kwargs)
```

```

    prepro = unused_model.build_preprocessor()
    token = unused_model.build_tokenizer()
    return(' '.join(token(prepro(text)))))

partial_processor = partial(text_processor, strip_accents='unicode', lowercase=True)

```

```
[61]: prediction = model.predict(test['blocks'].fillna('')).rename('predicted')
processed_prediction = prediction.apply(partial_processor)
processed_prediction.head(3)
```

```
[61]: uid
2892dd68-e3a6-474c-b543-3ebfd3490658      nestle un systeme de management de la qualite ...
a57c1561-b88e-4694-8bd8-55623f2afa17      cette fiche technique pas de valeur contractue...
3634fb1e-ee79-41d1-8aaa-084c1fae5bd5      ce produit est une puree de fruits obtenue par...
Name: predicted, dtype: object
```

```
[62]: processed_ground_truth = test['ingredients'].fillna('').apply(partial_processor)
processed_ground_truth.head(3)
```

```
[62]: uid
2892dd68-e3a6-474c-b543-3ebfd3490658      cafe instantane cafe torrefie moulu
a57c1561-b88e-4694-8bd8-55623f2afa17      lentilles blondes
3634fb1e-ee79-41d1-8aaa-084c1fae5bd5      poire 99 antioxydant acide ascorbique
Name: ingredients, dtype: object
```

```
[63]: corrects = test.join(prediction).loc[processed_prediction == processed_ground_truth , ['ingredients','predicted']]
corrects
```

	ingredients \
uid	
345591f4-d887-4ddc-bb40-21337fa9269d	Gésier de dinde émincé 50%, graisse de canard ...
13980d31-9002-457d-8d49-b451f08f473c	Edulcorants sorbitol, isomalt, sirop de maltit...
c3b6b4df-e586-4f10-8e58-15fb0f0816acb	mini poivrons jaunes, eau, sucre, sel, affermi...
0481d91b-9653-42e7-b525-9dc9b87b06f2	Farine de BLE, huile de colza non hydrogénée, ...
484ac00a-a670-46a9-a9c4-5114174d9e3b	Pommes de terre 59,5 % - Céleris 40 % - Amidon...
49b11281-34ea-44b0-a11c-4ae21d4c58e3	NaN
d59d96cb-0230-4090-8220-78ce8496fd91	Amidon de maïs* - Lait écrémé* - Sel - Fécule ...
b8cbe6f9-71d4-4e51-a169-1c163d49a561	Farine de FROMENT, poudre de LACTOSERUM, sucre...
a0492df6-9c76-4303-8813-65ec5ccbfa70	Eau, maltodextrine, sel, arômes, sucre, arôme ...
09e45b38-4da1-4eb5-888a-3ebd437a2291	ŒUFS, farine de BLE, sucre, amidon de BLE, st...
4f83306f-66de-4545-9b12-7790b57b61ae	Sirop de glucose, sucre, eau, stabilisants (E4...
5cee689e-6fb1-493c-b232-1d8fb1f88a57	Flageolets verts. Jus : eau, sel, affermissant...
63968dc3-6e7c-4056-bd53-820c6cc925be	Carottes, eau, sucre, sel, vinaigre d'alcool, ...
dc536305-82fd-4afe-a472-5056ca0e21ea	Légumes 43,2 % (pomme de terre, oignon, carott...

	predicted
uid	
345591f4-d887-4ddc-bb40-21337fa9269d	Gésier de dinde émincé 50%, graisse de canard...
13980d31-9002-457d-8d49-b451f08f473c	Edulcorants sorbitol, isomalt, sirop de maltit...
c3b6b4df-e586-4f10-8e58-15fb0f0816acb	mini poivrons jaunes, eau, sucre, sel, affermi...
0481d91b-9653-42e7-b525-9dc9b87b06f2	Farine de BLE, huile de colza non hydrogénée, ...
484ac00a-a670-46a9-a9c4-5114174d9e3b	Pommes de terre 59,5 % - Céleris 40 % - Amidon...
49b11281-34ea-44b0-a11c-4ae21d4c58e3	
d59d96cb-0230-4090-8220-78ce8496fd91	Amidon de maïs* - Lait écrémé* - Sel - Fécule ...
b8cbe6f9-71d4-4e51-a169-1c163d49a561	Farine de FROMENT, poudre de LACTOSERUM, sucre...
a0492df6-9c76-4303-8813-65ec5ccbfa70	Eau, maltodextrine, sel, arômes, sucre, arôme ...
09e45b38-4da1-4eb5-888a-3ebd437a2291	ŒUFS, farine de BLE, sucre, amidon de BLE, st...
4f83306f-66de-4545-9b12-7790b57b61ae	Sirop de glucose, sucre, eau, stabilisants (E4...
5cee689e-6fb1-493c-b232-1d8fb1f88a57	Flageolets verts. Jus : eau, sel, affermissant...
63968dc3-6e7c-4056-bd53-820c6cc925be	Carottes, eau, sucre, sel, vinaigre d'alcool, ...
dc536305-82fd-4afe-a472-5056ca0e21ea	Légumes 43,2 % (pomme de terre, oignon, carott...

```
[64]: with pd.option_context("max_colwidth", 100000):
    tex_str = (
        corrects.replace(r'^\s*$', np.nan, regex=True)
        .to_latex(index=False,
                   index_names=False,
                   column_format='p{7cm}p{7cm}',
                   na_rep='<rrien>',
                   longtable=False,
                   header=["Liste d'ingrédients cible", "Liste d'ingrédients prédicté"],
                   # label='tbl:GT_postprocessed_corrects',
                   # caption="Prédictions identifiées comme correctes après postprocessing",
        )
        .replace(r'\textbackslash n', r' \newline ')
        .replace(r'\\', r'\\ \\hline')
    )

    with open(Path('..') / 'tbls' / 'GT_postprocessed_corrects.tex', 'w') as file:
        file.write(tex_str)
```

Annexe D

LE CODE DES DIFFÉRENTS MODULES

D.1 Gestion du fichier de configuration - Module conf

Ce petit module a pour but de permettre de gérer les paramètres du programme dans un fichier de configuration (afin de simplifier la maintenance). Il est utilisé dans l'ensemble des autres modules de ce projet.

```
"""Configuration Module

This module provides a class that enables to fetch configuration stored in the
cfg folder and make it available to other python modules.

"""

import yaml
import os

class Config:

    def __init__(self, env):
        if env not in {'dev', 'int', 'rec', 'qat', 'prd'}:
            raise ValueError(f'Specified env : {env} not expected')

        self.env = env
        self.path = os.path.join(os.path.dirname(__file__), 'cfg', 'config.yaml')
        stream = open(self.path, 'r')
        data = yaml.safe_load(stream)
        config_keys = data['cross-env'].keys()
        for k in config_keys:
            setattr(self, k, data['cross-env'][k])
        config_keys = data[env].keys()
        for k in config_keys:
            setattr(self, k, data[env][k])
```

Un exemple de fichier de configuration (dont certains champs ont été anonymisés pour des raisons de confidentialité) est présenté ci-dessous. TODO !! Mettre le fichier ici.

D.2 Extraction des données du PIM - Module pimapi

```
"""PIM API Module

This module aims to enable to fetch data from PIM system, into local folders.

"""

import requests
import os
import json
import warnings
import pandas as pd
import numpy as np
import copy
import threading
from requests.exceptions import ConnectionError

from . import conf
warnings.simplefilter('always', UserWarning)

class Requester(object):
    """Requester class to retrieve information from PIM
    """
    def __init__(self, env, proxies='default', auth=None):
        self.cfg = conf.Config(env)
        self.session = requests.Session()
        if not auth:
            self.session.auth = (self.cfg.user, self.cfg.password)
        else:
            self.session.auth = auth
        if proxies == 'default':
            try:
                target_proxies = self.cfg.proxies
            except AttributeError:
                print('No proxy conf found for env : \'{}\''.format(self.cfg.env))
                'No proxy will be used.')
                target_proxies = None
        else:
            target_proxies = proxies
        self.session.proxies = target_proxies

    if proxies == 'default':
        # We try the connection with the default proxy conf
        # If it fails, we retry with no proxy configuration
        # (e.g. in the case of working from outside the network)
        try:
            self.check_connection()
        except (ConnectionError):
            self.session.proxies = None
            self.check_connection()
        else:
            # If a proxy configuration has been passed as an argument, we
            # only validate the connection with that configuration.
            self.check_connection()
        self.rlock = threading.RLock()
        self.result = []
        try:
            self._load_directory()
        except FileNotFoundError:
            warnings.warn('No directory found for current env : \'{}\''.format(self.cfg.env))
            'A new directory should be set.')
    def check_connection(self):
        """Checks wether the connection with the environments works

        The methods tries to get content of PIM system homepage. It returns
        True if the request is handled properly, or raises an exception.
        Note: this method does NOT check whether credentials are correct.
        """
        resp = self.session.get(self.cfg.baseurl)
        if resp.status_code != 200:
            raise ConnectionError('Connection could not be validated during '
                    'check_connection method call for '
                    'environment : \'{}\''.format(self.cfg.env))
        return(True)

    def check_credentials(self):
        """Checks wether the credentials provided allow to connect to PIM

        The methods tries to get content of PIM root document. It returns
        True if the request is handled properly, or raises an exception.
        Note: this method does NOT check whether credentials are correct.
        """
        resp = self.session.get(self.cfg.rooturl)
        if resp.status_code != 200:
            raise ConnectionError('Connection could not be validated during '
                    'check_credentials method call for '
                    'environment : \'{}\''.format(self.cfg.env))
        return(True)
```

```

True if the request is handled properly, or raises an exception.
Note: if the connection is invalid for another reason (e.g. incorrect
proxy configuration), an exception will be raised. the
`check_connection` method should be used prior to checking the
credentials.
"""
resp = self.session.get(self.cfg.baseurl +
                       self.cfg.suffixid +
                       self.cfg.rootuid)
if resp.status_code != 200:
    raise ConnectionError('Connection could not be validated during '
                          'check_credentials method call for '
                          f'environment : \'{self.cfg.env}\'')

def fetch_all_from_PIM(self, nx_properties='*', max_page=None,
                      page_size=None):
    """Fetches all product data from PIM into result

This method fetches all product data from PIM. It implements
multithreading to speed up retrieval.
"""
query = (f"SELECT * "
         f"FROM Document "
         f"WHERE ecm:primaryType='pomProduct' "
         f"AND ecm:isVersion=0")
headers = {'Content-Type': 'application/json',
           'X-NProperties': nx_properties}
params = {'query': query}
url = (self.cfg.baseurl +
       self.cfg.suffixid +
       self.cfg.rootuid + '/' +
       '@search')
result_count = self.query_size(headers, params, url)
self.result = []
max_page = max_page if max_page else self.cfg.maxpage
page_size = page_size if page_size else self.cfg.pagesize
params['pageSize'] = page_size
thread_count = result_count // page_size + 1
if max_page != -1 and thread_count > max_page:
    thread_count = max_page
    warnings.warn(f'\nMax size reached ! \n'
                  f'Only {max_page * page_size} results will be '
                  f'fetched out of {result_count} results\n')
threads = []
for page_index in range(thread_count):
    t = threading.Thread(target=self.get_page_from_query,
                         args=(url, headers, params, page_index))
    t.start()
    threads.append(t)
for t in threads:
    t.join()
print('Done')

def get_page_from_query(self, url, headers, params, currentPageIndex=0):
    """Fetches data for a single page of results

This methods fetches data for a prepared query for a single page. It
is used to implement multithreading for `fetch_all_from_PIM`.
"""
try:
    local_params = copy.deepcopy(params)
    local_params['currentPageIndex'] = currentPageIndex
    resp = self.session.get(url,
                           headers=headers,
                           params=local_params)
    with self.rlock:
        self.result.append(resp)
except Exception as e:
    print('An error occurred in this thread!')
    print(e)

def fetch_list_from_PIM(self, iter_uid, batch_size=50, nx_properties='*'):
    """Fetches data from an uid iterable, from PIM

This method fetches the data from PIM based on a iterable of uids.
It creates threads to fetch the complete list, and bases on the
Nuzeo @search API with a WHERE clause. It may therefore be less
fast than a full fetch.
Due to http limitations, it requires that no more than 50 uid be
retrieved at a time in a single thread. Failing to enforce will result
in incomplete responses and missed results.
"""
if batch_size > 50:
    raise ValueError(f'batch_size needs to be < 50. '
                     f'Call with:{batch_size}')
query = (f"SELECT * "
         f"FROM Document "
         f"WHERE ecm:primaryType='pomProduct' "
         f"AND ecm:isVersion=0")
headers = {'Content-Type': 'application/json',
           'X-NProperties': nx_properties}
url = (self.cfg.baseurl +
       self.cfg.suffixid +
       self.cfg.rootuid + '/' +
       '@search')
uid_lists = [iter_uid[i:i+batch_size]
             for i in range(0, len(iter_uid), batch_size)]
thread_list = []
self.result = []
for uid_list in uid_lists:
    uid_string = ', '.join(['"' + str(uid) + '"' for uid in uid_list])
    local_query = query + f" AND ecm:uid in ({uid_string})"
    params = {'query': local_query}
    t = threading.Thread(target=self.get_page_from_query,
                         args=(url, headers, params))
    t.start()
    thread_list.append(t)

for thread in thread_list:
    thread.join()
print('Done')

def query_size(self, headers, params, url):
    """Runs a single result query to get the number of results

This methods takes a query to be sent to Nuzeo in order to count the
number of results.
It then execute the query with page_size=1 and max_page=1 to fetch a
single result, and extracts the total number of results from the
response.
"""
loc_headers = copy.deepcopy(headers)
loc_headers['X-NProperties'] = ''
loc_params = copy.deepcopy(params)
loc_params['pageSize'] = 1
loc_params['currentPageIndex'] = 0
resp = self.session.get(url,
                        headers=loc_headers,
                        params=loc_params)
return resp.json()['resultsCount']

def _root_path(self):
    """Returns the root path for the dumps
    """
    return os.path.join(os.path.dirname(__file__),
                       '..',
                       'dumps',
                       self.cfg.env)

def dump_data_from_result(self, filename='data.json',
                          update_directory=True, root_path=None):
    """Dumps data from result attribute as JSON files

This method dumps data from result as JSON files.
Note that result MUST be an iterable of responses (be it from PIM or
from disk), each response should have an 'entries' list of documents.
"""
if update_directory and not hasattr(self, '_directory'):
    self._load_directory()
now = pd.Timestamp.now(tz='UTC')
threads = []
for single_result in self.result:
    t = threading.Thread(target=self.dump_data_from_single_result,
                         args=(single_result, filename, now),
                         kwargs={'update_directory': update_directory,
                                 'root_path': root_path})
    t.start()
    threads.append(t)
for t in threads:
    t.join()
print('Done')

def dump_data_from_single_result(self, single_result, filename, now,
                                  update_directory=False, root_path=None):
    """Dumps data from a single result

This method dumps data from a single result passed as an argument.
It is used for multithreading the result list.
"""
try:
    doc_list = single_result.json()['entries']
    s_list = []
    if root_path is None:
        root_path = self._root_path()
    for document in doc_list:
        path = os.path.join(root_path, document['uid'])
        if not os.path.exists(path):
            os.makedirs(path)
        full_path = os.path.join(path, filename)
        with open(full_path, 'w') as outfile:
            json.dump(document, outfile)
        s_list.append(pd.Series(now,
                               index=[document['uid']],
                               name='lastFetchedData'))
    df = pd.concat(s_list, axis=0)
    if update_directory:
        with self.rlock:
            self._directory.update(df)
            self._save_directory()
except Exception as e:
    print('An error occurred in this thread!')
    print(e)

def dump_files_from_result(self, update_directory=True, root_path=None):
    """Dumps attached files from result items on disk

This method dumps files from PIM on disk.
It also updates the local directory with current datetime.
Note that result MUST be an iterable of responses (be it from PIM or
from disk), each response should have an 'entries' list of documents,
and each document mention the files attached to it.
Attached files definition MUST be set in the config.yaml file.
"""
if update_directory and not hasattr(self, '_directory'):
    self._load_directory()
now = pd.Timestamp.now(tz='UTC')
threads = []
print(f'Launching {len(self.result)} threads.')
for single_result in self.result:
    t = threading.Thread(target=self.dump_files_from_single_result,
                         args=(single_result, now),
                         kwargs={'update_directory': update_directory,
                                 'root_path': root_path})
    t.start()
    threads.append(t)
for t in threads:
    t.join()

```

```

        t.join()
        print('Done')

def dump_files_from_single_result(self, single_result, now,
                                 update_directory=True, root_path=None):
    """Dumps attached files from items on disk - for a single result

This method dumps files from a single result into the disk. It is used
for multithreading in 'dump_files_from_result' method.
"""
try:
    doc_list = single_result.json()['entries']
    s_list = []
    if root_path is None:
        root_path = self._root_path()
    for document in doc_list:
        path = os.path.join(root_path, document['uid'])
        if not os.path.exists(path):
            os.makedirs(path)
        self.dump_attached_files(document, path)
        s_list.append(pd.Series(now,
                               index=[document['uid']],
                               name='lastFetchedFiles'))
    df = pd.concat(s_list, axis=0)
    if update_directory:
        with self._lock:
            self._directory.update(df)
        self._save_directory()
    print('Thread complete!')
except Exception as e:
    print('An error occurred in this thread!')
    print(e)

def _dump_file(self, file_url, path, filename='file'):
    """Dumps a file on disk from its url"""
    resp = self.session.get(file_url,
                           auth=(self.cfg.user, self.cfg.password),
                           stream=True)
    full_path = os.path.join(path, filename)
    with open(full_path, 'wb') as outfile:
        outfile.write(resp.content)

def dump_attached_files(self, document, path):
    """Dumps attached files from a nuzeo document

This function fetches attached files from a nuzeo document (stored as
a JSON).
Files definitions are stored in the configuration file.
"""
for filekind, filedef in self.cfg.filedefs.items():
    try:
        pointer = document
        for node in filedef['nuxepath']:
            pointer = pointer[node]
        nxfilename = pointer['name']
        url = (self.cfg.baseurl +
               self.cfg.suffixfile +
               self.cfg.nxrepo +
               document['uid'] + '/' +
               filedef['nuxepath'][1])
        ext = Requester.compute_extension(nxfilename)
        filename = filedef['dumpfilename'] + '.' + ext
        self._dump_file(url, path, filename)
    except TypeError:
        pass

def get_directory(self, **kwargs):
    """Get the uid directory for the environment as a pandas DataFrame

This function fetches the uid directory data for the current
environment as a pandas DataFrame.
To fetch all the results, set max_page attribute to -1.
This function returns data as is from PIM, and requires it to be
formatted as a directory later on with '_init_as_directory' or
'_format_as_directory' methods. (else, columns are not consistent with
directory definition)
"""
self.fetch_all_from_PIM(nx_properties='', **kwargs)
df_list = []
for single_result in self.results:
    df_list.append(pd.DataFrame(single_result.json()['entries'])
                  .set_index('uid'))
df = pd.concat(df_list)
df['lastModified'] = pd.to_datetime(df.loc[:, 'lastModified'])
return(df)

@staticmethod
def _directory_headers():
    """Returns the directory headers
"""
headers = ['type',
           'title',
           'lastModified',
           'lastRefreshed',
           'lastFetchedData',
           'lastFetchedFiles']
return(headers)

def _load_directory(self, filename=None):
    """Loads directory from disk into the tool attribute

directory_filename = filename if filename else self.cfg.uiddirectory
full_path = os.path.join(self._root_path(), directory_filename)
self._directory = (pd.read_csv(full_path,
                               encoding='utf-8-sig',
                               parse_dates=['lastModified',
                                            'lastRefreshed',
                                            'lastFetchedData'],
                               index_col='uid'))[headers]
self._directory.set_index('uid')
self._directory = Requester._format_as_directory(self._directory)

@staticmethod
def _init_as_directory(df):
    """Sets initial dates values for dataframes to be used as directories
"""
df['lastRefreshed'] = pd.Timestamp.now(tz='UTC')
df['lastFetchedData'] = np.nan
df['lastFetchedFiles'] = np.nan
return(Requester._format_as_directory(df))

@staticmethod
def _format_as_directory(df):
    """Formats a dataframe as a directory from dtypes point of view

This method formats a dataframe columns as to be consistent with
directory definition. It DOES NOT set initial values (see
'_init_as_directory').
"""
types = {'lastFetchedData': 'datetime64[ns, UTC]',
         'lastFetchedFiles': 'datetime64[ns, UTC]'}
df = df.astype(types)
df = df.loc[:, Requester._directory_headers()]
return(df)

def reset_directory(self, page_size=None, max_page=None, filename=None):
    """COMPLETELY RESETS the directory for the environment

Warning: this function will completely reset the directory for the
current environment. It should only be used when first setting the
directory or if it requires being remade from scratch.
Warning: this function should never be used with max_page parameter
different from -1 as it could result in incomplete data to erase
current directory. Only relevant usage of max_page != -1 is for
debugging purpose.
"""
self._directory = Requester._init_as_directory(
    self._get_directory(max_page=max_page, page_size=page_size))
self._save_directory()

def refresh_directory(self, max_page=None, filename=None, page_size=None):
    """Refreshes the directory for the environment

Warning: this function should never be used with max_page parameter
different from -1 except for debugging purpose. Yet, doing so does
not corrupt or lose data whatsoever.
"""
new_dir = self._get_directory(max_page=max_page, page_size=page_size)
new_dir = Requester._init_as_directory(new_dir)
self._load_directory(filename)
cur_dir = Requester._format_as_directory(self._directory)
cur_dir.update(new_dir)
cur_dir = pd.concat([cur_dir,
                     new_dir[new_dir.index.isin(cur_dir.index)]])
self._directory = cur_dir
self._save_directory()

def _save_directory(self, filename=None):
    """Saves current directory to disk

This methods saves the current directory to disk.
"""
directory_filename = filename if filename else self.cfg.uiddirectory
full_path = os.path.join(self._root_path(), directory_filename)
self._directory.to_csv(full_path, encoding='utf-8-sig')

def modified_items(self, what='any', max_results=None):
    """Returns modified items according to directory

This methods compares the date of last modification stored in the
directory to the date of last fetching of data or attached files.
If what = 'any', all outdated items will be returned
If what = 'data', it will return only items with outdated data
If what = 'files', it will return only items with outdated files
max_results enables to fetch only a limited count of results.
This returns an uid list
"""
if not hasattr(self, '_directory'):
    self._load_directory()
df = self._directory
if what == 'any':
    mask = (df.lastFetchedFiles.isna() |
            (df.lastModified > df.lastFetchedFiles) |
            df.lastFetchedData.isna() |
            (df.lastModified > df.lastFetchedData))
elif what == 'data':
    mask = (df.lastFetchedData.isna() |
            (df.lastModified > df.lastFetchedData))
elif what == 'files':
    mask = (df.lastFetchedFiles.isna() |
            (df.lastModified > df.lastFetchedFiles))
else:
    raise ValueError(f'Unexpected "what" argument: {what}')
uid_list = mask.index[mask].tolist()
if max_results:
    uid_list = uid_list[:max_results]
return(uid_list)

def modification_report(self):
    """Prints some elements about current state

This methods bases on current state of directory.
"""
if not hasattr(self, '_directory'):
    self._load_directory()
print(f'Number of items: {len(self._directory)}')

```

```

print(f'Number of items with outdated data: '
      f'{len(self.modified_items(what="data"))}')
print(f'Number of items with outdated files: '
      f'{len(self.modified_items(what="files"))}')

@staticmethod
def compute_extension(filename):
    """Computes the extension from a filename

    Returns the extension. If the filename has no "." (dot) in it, returns
    an empty string. If the computed extension has strictly more than 4
    characters, returns the empty string."""
    splitted = filename.split('.')
    if len(splitted) == 1:
        return('')
    elif len(splitted[-1]) > 4:
        return('')
    else:
        return(filename.split('.')[1].lower())

def result_to_dataframe(self,
                       record_path='entries',
                       meta=None,
                       mapping=None,
                       index='uid'):
    """Formats result content as a dataframe with defined format

    record_path and meta are pandas json_normalize method arguments
    mapping is a key : address mapping that maps return dataframe keys to
    data address in the JSON. Reminder: json_normalize default separator
    is '.'.
    So, for example if you want to have a record path down to the
    ingredients table (that is, down to
    record > properties > pprod:mainIngredients), with the uid product as
    index, you should use the following parameters:
    record_path['entries', 'properties', 'pprod:mainIngredients'],
    meta=['entries', 'uid'],
    index='entries.uid'
    index is the field identifier(s) for the field(s) to be used as index
"""
    result_json = [result.json() for result in self.result]
    if mapping:
        with CleanJSONDataFrame(result_json,
                               record_path=record_path,
                               meta=meta) as df:
            for key, path in mapping.items():
                try:
                    df[df[key]] = df[df.prefix + path]
                except KeyError:
                    df[df[key]] = np.nan
        df = df[df]
    else:
        df = pd.json_normalize(result_json,
                               record_path=record_path,
                               meta=meta)
    return(df)

def file_report_from_result(self, mapping, index=None, record_path=None):
    """Returns a dataframe about the files contained in the result

    This methods analyzes the content of the result, and generates a
    dataframe which enables the analysis of the files on the products
    Files to report on are based on the content of the config.yaml
    configuration file.
    """
    if index:
        df.set_index(index, inplace=True)
    return(df)

class CleanJSONDataFrame(object):
    """Context manager class for cleanly importing data from JSON

    This class enables to create a context manager that enables to read JSON
    data into a dataframe, by specifying the data to keep (this feature is not
    provided yet by the pandas json_normalize). It does so by following these
    steps:
    - read data from a JSON into a new dataframe
    - enable one to duplicate loaded data into new fields
    - delete the loaded data when exiting, thus keeping only duplicated data
    Complicated prefix is set to avoid duplicates in column names.
    """
    def __init__(self, data, record_path=None, meta=None,
                 prefix='_prev_dup1'):
        self.df = pd.json_normalize(data, record_path=record_path,
                                    meta=meta, record_prefix=prefix,
                                    meta_prefix=prefix)
        self.prefix = prefix
        self.columns = list(self.df.columns.values)

    def __enter__(self):
        return(self)

    def __exit__(self, exc_type, exc_val, exc_tb):
        self.df.drop([c for c in self.columns], axis=1, inplace=True)

```

D.3 Conversion des pièces jointes en textes - Module pimpdf

```

"""PIM PDF Module

This module aims to parse the content of PDF files into text.
"""

import pandas as pd
from multiprocessing import cpu_count
from pathos.multiprocessing import ProcessPool as Pool
from io import StringIO, BytesIO
from functools import partial

from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from pdfminer.pdfdocument import PDFDocument
from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
from pdfminer.pdfpage import PDFPage
from pdfminer.pdfparser import PDFParser

class PDFDecoder(object):
    """Tool that provides basic pdf decoding functionalities

    @staticmethod
    def content_to_text(content,
                        none_content='raise'):
        """Decodes the binary passed as argument

        content arg must be a BytesIO object.
        none_content arg must be raise (if it is not expected to have an empty
        input, the default) or to_empty (which will cause to return an empty
        string)
        """
        if none_content not in {'raise', 'to_empty'}:
            raise ValueError(f'Unexpected value for none_content parameter. '
                            f'Got {none_content} but only \'raise\' or '
                            f'\'to_empty\' are expected.')
        if not content.read():
            if none_content == 'raise':
                raise RuntimeError('PDFMiner got an empty bytesIO object to '
                                   'parse')
            if none_content == 'to_empty':

```

```

@staticmethod
def text_to_blocks_series(text, index=None,
                        split_func=lambda x: x.split('\n\n'),
                        return_type='along_index'):
    """
    Splits text passed as an argument (using splitter func) to a Series

    The return_type can be:
    - 'along_index': the return is a pandas Series of length the number
                      of blocks (with the index provided)
    - 'as_list' : the return is a pandas Series of length 1, with
                  the provided scalar as index, and the value of the
                  Series being the blocks as a list of strings
    If return_type is 'along_index' (the default), the index arg is passed
    as provided to pd.Series constructor, or if it is a scalar it is
    broadcasted as a constant on all values.
    """
    blocks_list = PDFDecoder.text_to_blocks(text,
                                            split_func=split_func,
                                            )
    if return_type not in {'along_index', 'as_list'}:
        raise ValueError(f'Unexpected value for return_type parameter. '
                         f'Got {return_type} but only \'along_index\' or '
                         f'\'as_list\' are expected.')
    if return_type == 'as_list':
        return(pd.Series([blocks_list], index=[index]))
    elif return_type == 'along_index':
        try:
            return(pd.Series(blocks_list, index=index))
        except TypeError:
            index = [index] * len(blocks_list)
            return(pd.Series(blocks_list, index=index))

@staticmethod
def path_to_blocks(path, split_func=lambda x: x.split('\n\n'),
                   missing_file='raise'):
    """
    Decodes file at local path in the form of a list of blocks

    Blocks are part of the original string separated by at least 2
    carriage returns (i.e. with at least a single blank line between them)
    """
    text = PDFDecoder.path_to_text(path, missing_file=missing_file)
    return(PDFDecoder.text_to_blocks(text, split_func=split_func))

@staticmethod
def path_to_blocks_series(path,
                          index=None,
                          split_func=lambda x: x.split('\n\n'),
                          missing_file='raise'):
    """
    Decodes file at local path in the form a pd Series of blocks

    """
    text = PDFDecoder.path_to_text(path, missing_file=missing_file)
    return(PDFDecoder.text_to_blocks_series(text,
                                           split_func=split_func,
                                           index=index))

@staticmethod
def paths_to_blocks(path_series, split_func=lambda x: x.split('\n\n'),
                    missing_file='raise'):
    """
    Decodes files for each path in path list as a blocks Dataseries

    Blocks are part of the original string after the splitting function
    has been applied.
    The input must be a pandas dataseries of paths.
    The output is another pd Series, with the same indexes as the initial
    series (broadcasted to match the block count for each path)
    """
    ds_list = []
    for uid, path in path_series.items():
        ds = (PDFDecoder
              .path_to_blocks_series(path,
                                     split_func=split_func,
                                     index=uid,
                                     missing_file=missing_file))
        ds_list.append(ds)
    return(pd.concat(ds_list, axis=0))

@staticmethod
def threaded_paths_to_blocks(path_series, processes=None,
                            split_func=lambda x: x.split('\n\n'),
                            missing_file='raise',
                            ):
    """
    Threaded version of paths_to_blocks method

    It takes as input a series which index is the uid of the products,
    and the values are the path to the document.
    processes argument is the number of processes to launch. If omitted,
    it defaults to the number of cpu cores on the machine.
    """
    processor = partial(PDFDecoder.path_to_blocks_series,
                        split_func=split_func, missing_file=missing_file)
    processes = processes if processes else cpu_count()
    print(f'Launching {processes} processes.')
    # Pool with context manager do not seem to work due to issue 38501 of
    # standard python library. It hangs when running tests through pytest
    # see: https://bugs.python.org/issue38501
    # Below content should be tested again whenever this issue is closed
    #
    # with Pool(nodes=processes) as pool:
    #     ds_list = pool.map(processor, path_series, path_series.index)
    #
    # End of block
    #
    # This temporary solution should be removed when tests mentioned above
    # are successful.
    # This just closes each pool after execution or exception.

    try:
        pool = Pool(nodes=processes)
        pool.restart(force=True)
        ds_list = pool.map(processor, path_series, path_series.index)
    except Exception:
        pool.close()
        raise
    pool.close()
    # End of block
    #
    # This temporary solution should be removed when tests mentioned above
    # are successful.
    # This just closes each pool after execution or exception.

    return(ds)

@staticmethod
def text_to_blocks(text, index=None,
                  split_func=lambda x: x.split('\n\n'),
                  return_type='along_index'):
    """
    Splits text passed as an argument (using splitter func) to a Series

    The return_type can be:
    - 'along_index': the return is a pandas Series of length the number
                      of blocks (with the index provided)
    - 'as_list' : the return is a pandas Series of length 1, with
                  the provided scalar as index, and the value of the
                  Series being the blocks as a list of strings
    If return_type is 'along_index' (the default), the index arg is passed
    as provided to pd.Series constructor, or if it is a scalar it is
    broadcasted as a constant on all values.
    """
    pool = Pool(nodes=processes)
    pool.restart(force=True)
    ds_list = pool.map(processor,
                       path_series,
                       path_series.index)
    try:
        pool.close()
        raise
    pool.close()
    # End of block
    #
    # This temporary solution should be removed when tests mentioned above
    # are successful.
    # This just closes each pool after execution or exception.

    return(pd.concat(ds_list, axis=0))

@staticmethod
def threaded_contents_to_text(content_series,
                               processes=None,
                               none_content='raise',
                               ):
    """
    Threaded version of content_to_text method

    It takes as input a series which index is the uid of the products,
    and the values are the content (in the form of bytes) of the
    documents.
    processes argument is the number of processes to launch. If omitted,
    it defaults to the number of cpu cores on the machine.
    none_content arg can be 'raise' (default) or to_empty
    """
    processor = partial(PDFDecoder.content_to_text,
                        none_content=none_content,
                        )
    processes = processes if processes else cpu_count()
    print(f'Launching {processes} processes.')
    in_ds = content_series.apply(BytesIO)

    # Pool with context manager do not seem to work due to issue 38501 of
    # standard python library. It hangs when running tests through pytest
    # see: https://bugs.python.org/issue38501
    # Below content should be tested again whenever this issue is closed
    #
    # with Pool(nodes=processes) as pool:
    #     tuples = (list(in_ds.index), pool.map(processor, in_ds))
    #
    # End of block
    #
    # This temporary solution should be removed when tests mentioned above
    # are successful.
    # This just closes each pool after execution or exception.

    try:
        pool = Pool(nodes=processes)
        pool.restart(force=True)
        tuples = (list(in_ds.index), pool.map(processor, in_ds))
    except Exception:
        pool.close()
        raise
    pool.close()
    # End of block
    #
    # This just closes each pool after execution or exception.

    ds = pd.Series(tuples[1], index=tuples[0])
    return(ds)

@staticmethod
def threaded_texts_to_blocks(text_series, processes=None,
                            split_func=lambda x: x.split('\n\n'),
                            return_type='along_index'):
    """
    Threaded version of text_to_blocks_series method

    It takes as input a series which index is the uid of the products,
    and the values are the content (in the form of bytes) of the
    documents..
    processes argument is the number of processes to launch. If omitted,
    it defaults to the number of cpu cores on the machine.
    As for text_to_blocks_series function, return_type can be 'along_axis'
    or 'list_like'.
    """
    processor = partial(PDFDecoder.text_to_blocks_series,
                        split_func=split_func,
                        return_type=return_type)
    processes = processes if processes else cpu_count()
    print(f'Launching {processes} processes.')
    # Pool with context manager do not seem to work due to issue 38501 of
    # standard python library. It hangs when running tests through pytest
    # see: https://bugs.python.org/issue38501
    # Below content should be tested again whenever this issue is closed
    #
    # with Pool(nodes=processes) as pool:
    #     ds_list = pool.map(processor, text_series, text_series.index)
    #
    # End of block
    #
    # This temporary solution should be removed when tests mentioned above
    # are successful.
    # This just closes each pool after execution or exception.

    try:
        pool = Pool(nodes=processes)
        pool.restart(force=True)
        ds_list = pool.map(processor, text_series, text_series.index)
    except Exception:
        pool.close()
        raise
    pool.close()
    # End of block
    #
    # This temporary solution should be removed when tests mentioned above
    # are successful.
    # This just closes each pool after execution or exception.

    ds = pd.concat(ds_list, axis=0)
    return(ds)

```

D.4 Transformateurs et estimateurs spécifiques - Module pimest

```

"""PIN Estimator module

This modules enables to create an estimator to identify which block is the
ingredient list from an iterable of text blocks.
"""

import os
from io import BytesIO

import numpy as np
from scipy.sparse.linalg import norm as sparse_norm
from scipy.sparse import csr_matrix
import pandas as pd
from pathlib import Path
from functools import partial

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.utils.validation import check_is_fitted

from .pimapi import Requester
from .pimpdf import PDFDecoder
from .conf import Config

class CustomTransformer(object):
    """Abstract class for custom transformers
    """
    def __init__(self, source_col, target_col, target_exists):
        self.source_col = source_col
        self.target_col = target_col
        self.target_exists = target_exists

    def raise_if_not_a_df(self, X):
        if not isinstance(X, pd.DataFrame):
            raise TypeError(f'This transformer expects a pandas Dataframe '
                            f'object. Got an object of type \'{type(X)}\' '
                            f'instead')

    def check_target_exists(self):
        if self.target_exists not in {'raise', 'ignore', 'overwrite'}:
            raise ValueError('target_exists parameter should be set to '
                            f'\'{raise}\' or \'{ignore}\' or \'{to_nan}\'. Got '
                            f'\'{self.target_exists}\' instead.')

    def raise_if_target(self, X):
        if self.target_col in X.columns and self.target_exists == 'raise':
            raise RuntimeError(f'Column \'{self.target_col}\' already exists '
                               f'in input DataFrame.')

    def raise_if_no_source(self, X):
        if self.source_col and self.source_col not in X.columns:
            raise KeyError(f'Input DataFrame has no \'{self.source_col}\' '
                           f'column.')

    def fit(self, X, y=None):
        self.check_target_exists()
        self.raise_if_not_a_df(X)
        self.raise_if_target(X)
        self.raise_if_no_source(X)

        self.transform(X, y)

    def fit_transform(self, X, y=None):
        return(self.fit(X).transform(X))

    def get_params(self, deep=True):
        params = dict()
        params['source_col'] = self.source_col
        params['target_col'] = self.target_col
        params['target_exists'] = self.target_exists
        return(params)

    def set_params(self, **parameters):
        for parameter, value in parameters.items():
            setattr(self, parameter, value)
        return(self)

    class IngredientExtractor(object):
        """Estimator that identifies the most 'ingredient like' block from a list
        """
        def __init__(self):
            """Constructor method of ingredient extractor"""
            pass

        def fit(self, X, y=None):
            """Fitter method of ingredient extractor

            X is an iterable of ingredient lists in the form of strings
            y is just here for compatibility in sklearn pipeline usage
            """
            self._count_vect = CountVectorizer()

```

```

            self.vectorized_texts_ = self._count_vect.fit_transform(X)
            self.vocabulary_ = self._count_vect.vocabulary_
            self.mean_corpus_ = self.vectorized_texts_.mean(axis=0)
            return(self)

        def predict(self, X):
            """Predictor method of ingredient extractor

            X is a list of text blocks.
            This methods returns the index of the text block that is most likely
            to hold the ingredient list"""
            X_against_ingred_voc = self._count_vect.transform(X)
            X_norms = sparse_norm(CountVectorizer().fit_transform(X), axis=1)
            X_dot_ingred = np.array(X_against_ingred_voc.sum(axis=1)).squeeze()
            pseudo_cosine_sim = np.divide(X_dot_ingred,
                                         X_norms,
                                         out=np.zeros(X_norms.shape),
                                         where=X_norms != 0)
            self.similarity_ = pseudo_cosine_sim
            return(np.argmax(pseudo_cosine_sim))

        def show_emphasize(self, X):
            """Method that prints strings with words from vocabulary emphasized
            """
            for text in self.emphasize_texts(X):
                print(text)

        def emphasize_texts(self, X):
            """Method that returns strings with words from vocabulary emphasized

            This method shows how some candidates texts are projected on the
            vocabulary that has been provided or gotten from fitting.
            It is useful to see how different blocks compare.
            X argument is an iterable of block candidates.
            """
            check_is_fitted(self)
            preprocessor = self._count_vect.build_preprocessor()
            tokenizer = self._count_vect.build_tokenizer()
            vocabulary = self._count_vect.vocabulary_
            emphasized_texts = []
            for block in X:
                text = self.emphasize_words(block,
                                             preprocessor=preprocessor,
                                             tokenizer=tokenizer,
                                             vocabulary=vocabulary,
                                             )
                emphasized_texts.append(text)
            return(emphasized_texts)

        def emphasize_words(self,
                            text,
                            preprocessor=None,
                            tokenizer=None,
                            vocabulary=None,
                            ansi_color="\033[92m", # green by default
                            ):
            """Method that returns a string with words emhasized

            This methods takes a string and returns a similar string with the words
            emphasized (with color markers)
            """
            check_is_fitted(self)
            ansi_end_block = '\033[0m'
            if not preprocessor:
                preprocessor = self._count_vect.build_preprocessor()
            if not tokenizer:
                tokenizer = self._count_vect.build_tokenizer()
            if not vocabulary:
                vocabulary = self._count_vect.vocabulary_
            preprocessed_text = preprocessor(text)
            tokenized_text = tokenizer(preprocessed_text)
            idx = 0
            emphasized_text = ''
            for token in tokenized_text:
                if token in vocabulary:
                    while preprocessed_text[idx:idx + len(token)] != token:
                        emphasized_text += text[idx]
                        idx += 1
                    emphasized_text += (ansi_color + text[idx:idx + len(token)]) +
                                       ansi_end_block
                    idx += len(token)
            emphasized_text += text[idx:]
            return(emphasized_text)

        def score(self, X, y):
            """Scorer method of ingredient extractor estimator

            X is an iterable of ingredient lists in the form of string
            y is the target as the index of the correct block.
            """
            pass

    class PIMIngredientExtractor(IngredientExtractor):
        """Wrapped estimator that directly extracts the ingredient list from uid
        """
        def __init__(self, env='prd', **kwargs):
            self.requester = Requester(env, **kwargs)
            super().__init__()


```

```

def compare_uid_data(self, uid):
    check_is_fitted(self)
    print(f'Fetching data from PIM for uid {uid}...')
    self.requester.fetch_list_from_PIM([uid])
    try:
        ingredient_list = (self.requester.result[0]
                           .json()['entries'][0]['properties']
                           ['pprod:ingredientsList'])
    except IndexError:
        raise IndexError(f'Fetching data with uid {uid} ')
    print('-----')
    print(f'Ingredient list from PIM is :\n{ingredient_list}')
    print('-----')
    print(f'Supplier technical datasheet from PIM for uid {uid} is:')
    nuxeo_path = (self.requester.cfg.filedefs['supplierdatasheet']
                  ['nuxepath'])
    pointer = self.requester.result[0].json()['entries'][0]
    for node in nuxeo_path:
        pointer = pointer[node]
    file_url = pointer['data']
    print(file_url)
    print('-----')
    print(f'Downloading content of technical datasheet file...')
    self.resp = self.requester.session.get(file_url,
                                            stream=True)
    resp = self.resp
    print('Done!')
    print('-----')
    print(f'Parsing content of technical datasheet file...')
    blocks = (PDFDecoder.content_to_text(BytesIO(resp.content))
              .split('\n\n'))
    idx = self.predict(blocks)
    print('Done!')
    print('-----')
    print(f'Ingredient list extracted from technical datasheet:\n')
    print(blocks[idx])
    print('-----')

def print_blocks(self):
    blocks = (PDFDecoder.content_to_text(BytesIO(self.resp.content))
              .split('\n\n'))
    for i, block in enumerate(blocks):
        print(i, ' | ', block, '\n')

class PathGetter(CustomTransformer):
    """Class that gets path for documents on disk

    This class aims to compute the path to documents, in order to
    fetch documents from the correct folder (depending on whether
    they are from train set or from ground truth)
    All these can be set at initialization, if such is not the case
    then their values is gotten from the configuration file.
    """
    def __init__(self,
                 env='prd',
                 ground_truth_uids=None,
                 train_set_path=None,
                 ground_truth_path=None,
                 path_factory=lambda x: x,
                 filename_factory=lambda x: 'FTF.pdf',
                 source_col=None,
                 target_col='path',
                 target_exists='raise',
                 ):
        self.env = env
        self.ground_truth_uids = ground_truth_uids
        self.train_set_path = train_set_path
        self.ground_truth_path = ground_truth_path
        self.path_factory = path_factory
        self.filename_factory = filename_factory
        super().__init__(source_col=source_col, target_col=target_col,
                         target_exists=target_exists)

    def fit(self, X, y=None):
        """No fit is required for this class.
        """
        super().fit(X)
        self.cfg = Config(self.env)
        if not self.train_set_path:
            self.train_set_path = os.path.join(*self.cfg.trainsetpath)
        if not self.ground_truth_path:
            self.ground_truth_path = os.path.join(*self.cfg.groundtruthpath)
        self.fitted_ = True
        return(self)

    def transform(self, X):
        """Returns the paths for the uids"""
        super().transform(X)
        df = X.copy()
        df[self.target_col] = None
        for uid in X.index:
            if uid in self.ground_truth_uids:
                path = os.path.join(self.ground_truth_path,
                                    self.path_factory(uid),
                                    self.filename_factory(uid),
                                    )
            else:
                path = os.path.join(self.train_set_path,
                                    self.path_factory(uid),
                                    self.filename_factory(uid),
                                    )
        df.loc[uid, self.target_col] = path
        return(df)

    def get_params(self, deep=True):
        params = super().get_params()

```

parms['env'] = self.env
 parms['ground_truth_uids'] = self.ground_truth_uids
 parms['train_set_path'] = self.train_set_path
 parms['ground_truth_path'] = self.ground_truth_path
 parms['path_factory'] = self.path_factory
 parms['filename_factory'] = self.filename_factory
 return(params)

```

class ContentGetter(CustomTransformer):
    """Class that fetches the content of documents on disk

    This class fetches the data from documents on disk as bytes.
    It requires a dataframe with a path column"""
    def __init__(self,
                 missing_file='raise',
                 target_exists='raise',
                 source_col='path',
                 target_col='content',
                 ):
        self.missing_file = missing_file
        super().__init__(source_col=source_col,
                         target_col=target_col,
                         target_exists=target_exists,
                         )

    def fit(self, X, y=None):
        super().fit(X)
        if self.missing_file not in {'raise', 'ignore', 'to_nan'}:
            raise ValueError(f'missing_file parameter should be set to '
                             f'\\'raise\' or \\'ignore\' or \\'to_nan\'. Got '
                             f'\'{self.missing_file}\'' instead.')
        self._raise_if_no_file(X)
        self.fitted_ = True
        return(self)

    def _raise_if_no_file(self, X):
        if self.missing_file == 'raise':
            mask = pd.DataFrame(index=X.index)
            mask['file_exists'] = X['path'].apply(ContentGetter.file_exists)
            if not mask['file_exists'].all():
                example_uid = mask.loc[mask['file_exists']].index[0]
                example_path = X.loc[example_uid, 'path']
                raise RuntimeError(f'No file found for uid \'{example_uid}\'' +
                                   f' at path \'{example_path}\'')

    def transform(self, X):
        super().transform(X)
        self._raise_if_no_file(X)
        X = X.copy()
        mask = pd.DataFrame(index=X.index)
        mask['file_exists'] = X['path'].apply(ContentGetter.file_exists)
        mask['target'] = X['path'].apply(ContentGetter.read_to_bytes)
        if self.missing_file == 'to_nan':
            idx_to_update = mask.index
            else:
                idx_to_update = mask['file_exists']
            X.loc[idx_to_update, 'content'] = mask.loc[idx_to_update, 'target']
            return(X)

    @staticmethod
    def read_to_bytes(path):
        try:
            return(Path(path).read_bytes())
        except FileNotFoundError:
            return(None)

    @staticmethod
    def file_exists(path):
        path = Path(path)
        return(path.is_file())

    def get_params(self, deep=True):
        params = super().get_params()
        params['missing_file'] = self.missing_file
        return(params)

class PDFContentParser(CustomTransformer):
    """Class that parses pdf content to text

    This class converts a file content (in the form of bytes) into text, using
    pimpdf functionalities (based on pdfminer.six)
    """
    def __init__(self,
                 target_exists='raise',
                 source_col='content',
                 target_col='text',
                 none_content='raise',
                 ):
        self.none_content = none_content
        super().__init__(source_col=source_col,
                         target_col=target_col,
                         target_exists=target_exists,
                         )

    def fit(self, X, y=None):
        super().fit(X)
        self.fitted_ = True
        return(self)

    def transform(self, X):
        super().transform(X)
        X = X.copy()
        tran = (PDFDecoder
                .threaded_contents_to_text(X[self.source_col],
                                           none_content=self.none_content))
        X[self.target_col] = tran

```

```

    return(X)

def get_params(self, deep=True):
    parms = super().get_params()
    parms['none_content'] = self.none_content
    return(parms)

class BlockSplitter(CustomTransformer):
    """Class that splits texts into blocks

    This class converts a text string into blocks (a list of string), using
    the splitter function provided
    """
    def __init__(self,
                 target_exists='raise',
                 source_col='text',
                 target_col='blocks',
                 splitter_func=(lambda x: x.split('\n\n'))
                 ):
        self.splitter_func = splitter_func
    super().__init__(target_exists=target_exists,
                    source_col=source_col,
                    target_col=target_col,
                    )

    def fit(self, X, y=None):
        super().fit(X)
        self._check_splitter_callable()
        self.fitted_ = True
        return(self)

    def _check_splitter_callable(self):
        self.splitter_func('')

    def transform(self, X):
        super().transform(X)
        X = X.copy()
        blocks = (PDFDecoder
                  .threaded_texts_to_blocks(X[self.source_col],
                                             split_func=self.splitter_func,
                                             return_type='as_list',
                                             ))
        X[self.target_col] = blocks
        return(X)

    def get_params(self, deep=True):
        parms = super().get_params()
        parms['splitter_func'] = self.splitter_func
        return(parms)

class SimilaritySelector():
    """Class that select the most similar block from a block list

    This class provides functionnalities to fit an estimator on a topic
    specific vocabulary, and to retrieve the best candidate amongst these
    blocks.
    It can append a new column to an input pandas DataFrame with the best
    candidate.
    """
    def __init__(self,
                 count_vect_kwargs=dict(),
                 similarity='projection',
                 source_norm='l2',
                 projected_norm='l1',
                 ):
        self.count_vect_kwargs = count_vect_kwargs
        self.similarity = similarity
        self.source_norm = source_norm
        self.projected_norm = projected_norm

    def get_params(self, deep=True):
        parms = dict()
        parms['count_vect_kwargs'] = self.count_vect_kwargs
        parms['similarity'] = self.similarity
        parms['source_norm'] = self.source_norm
        parms['projected_norm'] = self.projected_norm
        return(parms)

    def set_params(self, **parameters):
        for parameter, value in parameters.items():
            setattr(self, parameter, value)
        return(self)

    def fit(self, X, y):
        self._validate_similarity()
        self._validate_norms()
        try:
            self.count_vect = CountVectorizer(**self.count_vect_kwargs)
        except (TypeError):
            raise ValueError('Unexpected argument at init in '
                            "'count_vect_kwargs.'")
        raise
        try:
            self.count_vect.fit(y.fillna(''))
        except (ValueError):
            raise ValueError('Unexpected argument at fit in '
                            "'count_vect_kwargs.'")
        raise
        self.source_count_vect = CountVectorizer(**self.count_vect_kwargs)
        self.fitted_ = True
        return(self)

    def _validate_similarity(self):
        if self.similarity not in {'projection', 'cosine'}:
            raise ValueError(f'similarity parameter should be set to '
                            f'`projection` or `cosine`. Got '
                            f'{self.similarity} instead.')

```

```

        if self.similarity not in {'projection', 'cosine'}:
            raise ValueError(f'similarity parameter should be set to '
                            f'`projection` or `cosine`. Got '
                            f'{self.similarity} instead.')

def _validate_norms(self):
    norm_dict = {'l1': partial(sparse_norm, axis=1, ord=1),
                'l2': partial(sparse_norm, axis=1, ord=2)}
    try:
        self.source_norm = norm_dict[self.source_norm]
    except KeyError:
        pass
    try:
        self.projected_norm = norm_dict[self.projected_norm]
    except KeyError:
        pass
    test_mat = csr_matrix([[0, 1], [2, 3]])
    try:
        self.projected_norm(test_mat)
    except (TypeError, ValueError) as e:
        print('Incorrect projected norm provided, see full stack for '
              'details')
        raise ValueError(e)
    try:
        self.source_norm(test_mat)
    except (TypeError, ValueError) as e:
        print('Incorrect source norm provided, see full stack for '
              'details')
        raise ValueError(e)

def predict(self, X):
    """ function to predict best candidate

    X : a pandas Series of block lists, or a list.

    """
    check_is_fitted(self)
    docs = [text for block_list in X for text in block_list]
    try:
        self.source_count_vect.fit(docs)
    except ValueError:
        print('No words in blocks. Return defaulted to "")')
        return(np.array([''] * len(X)))
    if self.similarity == 'projection':
        computed_sims = []
        predicted_texts = []
        for block_list in X:
            texts = self.source_count_vect.transform(block_list)
            # project texts on corpus space
            projected_texts = self.count_vect.transform(block_list)
            # Compute norm of source texts
            texts_norms = self.source_norm(texts)
            # Compute norm of projected texts
            projected_norms = self.projected_norm(projected_texts)
            sim = np.divide(projected_norms,
                           texts_norms,
                           out=np.zeros(texts_norms.shape),
                           where=texts_norms != 0)
            computed_sims.append(sim)
            predicted_texts.append(block_list[np.argmax(sim)])
        if isinstance(X, pd.Series):
            return(pd.Series(predicted_texts, index=X.index))
        else: # for example, X is a list
            return(pd.Series(predicted_texts))
    if self.similarity == 'cosine':
        raise NotImplementedError

class DummyEstimator(object):
    """Dummy estimator that predicts y as exactly X

    This estimator has been developped for testing purposes, as pytest
    does not support yet class fixtures.
    """
    def fit(self, X, y=None):
        self.fitted_ = True

    def predict(self, X):
        return(X.copy())

def custom_accuracy(estimator, X, y, tokenize=True, **kwargs):
    """Computes accuracy of estimator, for texts

    This function enables to score an estimator that returns long texts,
    with some text processing.
    It computes an accuracy after text processing
    It is based on sklearn CountVectorizer functionalities.
    tokenize means that the input string will be tokenized as words before
    being glued back with single spaces. It's purpose is to handle
    whitespaces (newlines, tabs, multiple spaces, ...) and punctuation.
    kwargs are directly passed to CountVectorizer constructor, and will
    serve to process the texts. Most useful args are 'strip_accents' and
    'lowercase'.
    """
    preprocessor_countvect = CountVectorizer(**kwargs)
    preprocessor = preprocessor_countvect.build_preprocessor()
    tokenizer = preprocessor_countvect.build_tokenizer()
    if tokenize:
        def transformer(x):
            return(tokenizer(preprocessor(x)))
    else:
        transformer = preprocessor
    y_pred = pd.Series(estimator.predict(X).apply(transformer))
    return((y_pred == y.apply(transformer)).mean())

```