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 diplay
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]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>,
 <Response [200]>]
```

[8]:
```
mapping = {'uid': 'uid',
           'designation': 'title',
           'state': 'state',
           'ingredients': 'properties.pprodc: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                                               None  chemistry
68f5d81b-7f91-40a0-8504-0ec320a86de4                                               None  chemistry
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a                                               None  chemistry

                                      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                  False              False
68f5d81b-7f91-40a0-8504-0ec320a86de4                  False              False
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a                  False              False

[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!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
```

```
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
Thread complete!
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 maïs*, …
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', '  '))
```

```python
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 OEUFS entiers*, gélifiants : carraghénanes, agar-agar* ; arôme naturel de vanille* et autres arômes
naturels*, poudre de gousses de vanille*, curcuma*.\n* Produits issus de l Agriculture Biologique.\n°
Ingrédient 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) -
Diphosphate ferrique
0faad739-ea8c-4f03-b62e-51ee592a0546
Farine de blé T45

[500 rows x 2 columns]
```