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 diplay
# 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]:

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

```
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        17-0  …
7d390121-17e8-43bf-a357-9d06b79d2d47        15-0  …
f234cd84-c8f6-433f-85ec-6e0b6980adc6        33-0  …
e82a8173-b379-41ac-b319-aa058a04fcfb        20-0  …
4b12c47c-84f5-4132-b362-22b864379a67        20-0  …
…                                              …  …
5cde49c6-9e7e-4bd2-b22a-3239f643379d        88-0  …
0273eadc-851a-4b68-8020-8041700a4f3d        17-0  …
ef42a938-2203-446e-8d28-9fd27c6d3146        17-0  …
68f5d81b-7f91-40a0-8504-0ec320a86de4        17-0  …
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a        17-0  …

                                      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                                               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: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                                            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.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.pprodqmdd:secondaryPackagingPhoto.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                                    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.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                                    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.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
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

[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.pprodc: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.
```