open_model

Pierre Massé

May 12, 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)
```

[40]:
```
# 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 PIMIngredientExtractor
# 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 :

[5]:
```
requester = Requester('prd')
requester.fetch_all_from_PIM()
requester.result
```

```
Done
```

[5]:
```
[<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]>]
```

[6]:
```
df = requester.result_to_dataframe(record_path='entries', index='uid')
df
```

[6]:

uid	entity-type	repository	path	type	state	parentRef	isCheckedOut	isVersion	isProxy	changeToken	…	properties.pprodqmdd:manufacturingDiagram.length
afee12c7-177e-4a68-9539-8cbb68442503	document	default	/default-domain/pomSupplierWorkspace/SICO/DEST…	pomProduct	product.waiting.supplier.validation	a58845c0-cab3-492f-b48d-531f146c3777	True	False	False	17-0	…	NaN
7d390121-17e8-43bf-a357-9d06b79d2d47	document	default	/default-domain/pomSupplierWorkspace/UNILEVER_…	pomProduct	product.waiting.supplier.validation	a37abc27-f485-4ae9-921b-f761f16c8c1c	False	False	False	15-0	…	NaN
f234cd84-c8f6-433f-85ec-6e0b6980adc6	document	default	/default-domain/pomSupplierWorkspace/AZTECA_FO…	pomProduct	product.waiting.supplier.validation	3ff7819a-a392-493f-beb8-0b323ac331c7	True	False	False	33-0	…	NaN
…	…	…	…	…	…	…	…	…	…	…	…	…
ef42a938-2203-446e-8d28-9fd27c6d3146	document	default	/default-domain/pomSupplierWorkspace/SICO/DETE…	pomProduct	product.waiting.supplier.validation	a58845c0-cab3-492f-b48d-531f146c3777	True	False	False	17-0	…	NaN
68f5d81b-7f91-40a0-8504-0ec320a86de4	document	default	/default-domain/pomSupplierWorkspace/SICO/NETT…	pomProduct	product.waiting.supplier.validation	a58845c0-cab3-492f-b48d-531f146c3777	True	False	False	17-0	…	NaN
6dfce29e-fd4c-4670-9f9c-5c02a5b4d52a	document	default	/default-domain/pomSupplierWorkspace/SICO/SPRA…	pomProduct	product.waiting.supplier.validation	a58845c0-cab3-492f-b48d-531f146c3777	True	False	False	17-0	…	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
…                                                                             …
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
…                                                                               …
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
…                                                                                 …
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
…                                                                                 …
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
…                                                                                 …
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
…                                                                                 …
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
…                                                                                       …
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
…                                                                                       …
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
…                                                                …
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"

[13]:
```
# 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,')
out_of_scope_df = df.loc[~df.index.isin(scope_df.index)]
print(f'and {len(out_of_scope_df)} records left out.')
```

```
After filters, there are 3412 records in the dataset,
and 9816 records left out.
```

1.4 Entraînement : constitution du vocabulaire

On entraîne le modèle sur les listes d'ingrédients du périmètre. Cela revient à fitter le CountVectorizer sous-jacent.

[17]:
```
model = PIMIngredientExtractor('prd')
model.fit(scope_df['properties.pprodc:ingredientsList'])
```

[17]:
```
<src.pimest.PIMIngredientExtractor at 0x7ff2b555db50>
```

On peut imprimer une partie du vocabulaire qui a été construit :

[48]:
```
print(f'Vocabulary consists in {len(model._count_vect.vocabulary_)} words.\n')
print('Some words examples are :')

for i, word in enumerate(model._count_vect.vocabulary_.keys()):
    print('- ', word)
    if i > 6:
        break
```

```
Vocabulary consists in 2509 words.

Some words examples are :
-  morilles
-  eau
-  de
-  source
-  kombu
-  déshydraté
-  100
-  graines
```

On peut également afficher les mots les plus fréquents dans le corpus de listes d'ingrédients d'entraînement. On constitue d'abord la matrice des textes transformés :

[26]:
```
vectorized = model._count_vect.transform(scope_df['properties.pprodc:ingredientsList'])
vectorized.shape
```

[26]:
```
(3412, 2509)
```

On a bien 3412 documents projetés sur 2509 mots. Si on extrait les plus fréquents, on obtient :

[59]:
```
inverse_voc = {val: key for key, val in model._count_vect.vocabulary_.items()}
word_counts = np.asarray(vectorized.sum(axis=0)).squeeze()
print('Most frequent words in vocabulary are:')
for idx in word_counts.argsort()[::-1][:10]:
    print(f'{inverse_voc[idx].ljust(7)}: {word_counts[idx]:5} occurences')
```

```
Most frequent words in vocabulary are:
de     : 11419 occurences
sucre  :  2057 occurences
sel    :  1669 occurences
eau    :  1288 occurences
acide  :  1241 occurences
lait   :  1215 occurences
huile  :  1214 occurences
poudre :  1100 occurences
en     :   962 occurences
arôme  :   938 occurences
```

1.5 Prédictions

Le wrapper `PIMIngredientExtractor` permet de simplement récupérer les informations du PIM et les pièces jointes associées, et de faire tourner le modèle pour extraire le bloc le plus similaire aux listes d'ingrédients.

[81]:
```
print(len('=============='))
```

```
13
```

[82]:
```
exec_count = 5
uids = list(out_of_scope_df.sample(exec_count, random_state=41).index)

for uid in uids:
    model.compare_uid_data(uid)
```

```
␣
↪print('\n============================================================\n============================================================
```

```
Fetching data from PIM for uid d9b233a6-b455-4af6-afb4-623f1f7f62a6…
Done
------------------------------------------------------------
Ingredient list from PIM is :

Ingrédients: Huile de tournesol, oignon, curry (11,2%) (ail, coriandre, curcuma, gingembre, paprika, poivre,
cumin, poivre de Cayenne, fenouil, cardamome, noix de muscade, canelle, clous de girofle, safran), pomme,
sel, exhausteur de goût (glutamate de sodium), sucre, huile de colza totalement hydrogénée, extrait de
levure, ail.

------------------------------------------------------------
Supplier technical datasheet from PIM for uid d9b233a6-b455-4af6-afb4-623f1f7f62a6 is:
https://produits.groupe-pomona.fr/nuxeo/nxfile/default/d9b233a6-b455-4af6-afb4-623f1f7f62a6/pprodad:technica
lSheet/FT%20-15838201-%20Mise%20en%20Place%20Curry%20KNORR%20mars%202020.pdf?changeToken=58-0
------------------------------------------------------------
Downloading content of technical datasheet file…
Done!
------------------------------------------------------------
Parsing content of technical datasheet file…
Done!
------------------------------------------------------------
Ingredient list extracted from technical datasheet:

Liste d'ingrédients : Huile de tournesol, oignon, curry (11,2%) (ail, coriandre, curcuma, gingembre,
paprika, poivre, cumin,
poivre de Cayenne, fenouil, cardamome, noix de muscade, canelle, clous de girofle, safran), pomme, sel,
exhausteur de goût
(glutamate de sodium), sucre, huile de colza totalement hydrogénée, extrait de levure, ail

------------------------------------------------------------

============================================================
============================================================

Fetching data from PIM for uid 5666235b-9e78-44f2-8e0e-1de53f88fe04…
Done
------------------------------------------------------------
Ingredient list from PIM is :

Ingrédients : Sucre, Gomme base, Sirop de glucose, Arômes, Humectant (E422), Antioxydant (E321), Colorant
(E141).

------------------------------------------------------------
Supplier technical datasheet from PIM for uid 5666235b-9e78-44f2-8e0e-1de53f88fe04 is:
https://produits.groupe-pomona.fr/nuxeo/nxfile/default/5666235b-9e78-44f2-8e0e-1de53f88fe04/pprodad:technica
lSheet/FT-136899_Chewing%20gum%20menthe%20bte%20200U%20Malabar_Carambar.pdf?changeToken=32-0
------------------------------------------------------------
Downloading content of technical datasheet file…
Done!
------------------------------------------------------------
Parsing content of technical datasheet file…
Done!
------------------------------------------------------------
Ingredient list extracted from technical datasheet:

INGRÉDIENTS : Sucre, Gomme base, Sirop de glucose, Arômes, Humectant (E422),
Antioxydant (E321), Colorant (E141).

------------------------------------------------------------

============================================================
============================================================

Fetching data from PIM for uid 6e976147-adeb-4d2d-925a-cb7c58c111a2…
```

Done

Ingredient list from PIM is :

Maltodextrine, amidon de maïs, sel, farine de BLE, colorant : caramel ordinaire ; arômes (BLE,CELERI), huile de palme, épaississant : gomme guar ; oignon, fécule de pomme de terre, extrait de levure, jus de cuisson de viande de boeuf (0,9%), acidifiant : acide citrique ; extrait de vin blanc, extraits d'ail, de thym et de poivre. Peut contenir : LAIT, OEUF.

Supplier technical datasheet from PIM for uid 6e976147-adeb-4d2d-925a-cb7c58c111a2 is:
https://produits.groupe-pomona.fr/nuxeo/nxfile/default/6e976147-adeb-4d2d-925a-cb7c58c111a2/pprodad:technicalSheet/FT%20MAGGI%20Fonds%20Brun%20Li%C3%A9%20-%20Bo%C3%AEte%20de%20750g.pdf?changeToken=142-0

Downloading content of technical datasheet file…
Done!

Parsing content of technical datasheet file…
Done!

Ingredient list extracted from technical datasheet:

Maltodextrine, amidon de maïs, sel, farine de blé, colorant : caramel ordinaire ; arômes (blé,céleri), huile de palme, épaississant : gomme guar ; oignon, fécule
de pomme de terre, extrait de levure, jus de cuisson de viande de bœuf (0,9%), acidifiant : acide citrique ; extrait de vin blanc, extraits d'ail, de thym et de poivre.
Peut contenir : lait, œuf.

==========
==========

Fetching data from PIM for uid db449562-d16d-4f72-b7a5-c0d487bc8206…
Done

Ingredient list from PIM is :

Huile d'ARACHIDE

Supplier technical datasheet from PIM for uid db449562-d16d-4f72-b7a5-c0d487bc8206 is:
https://produits.groupe-pomona.fr/nuxeo/nxfile/default/db449562-d16d-4f72-b7a5-c0d487bc8206/pprodad:technicalSheet/FT-7744_%20Huile%20arachide%20bid%205L_HUILERIES%20DE%20SERIGNAN.pdf?changeToken=22-0

Downloading content of technical datasheet file…
Done!

Parsing content of technical datasheet file…
Done!

Ingredient list extracted from technical datasheet:

*Selon le règlement UE n°1259-2011 / In accordance with regulation UE n°1259-2011

CARACTERISTIQUES MICROBIOLOGIQUES

L’huile étant un milieu anhydre, tout développement bactérien est impossible (cf. ouvrage de référence dans ce domaine "La qualité microbiologique des aliments" CNERMA-CNRS coordonné par Jean-louis Jouve).

ORIGINES/ ORIGIN

- Amérique du Sud majoritairement
- Afrique de l’Ouest

AUTRES INFORMATIONS

```
------------------------------------------------------------

============================================================
============================================================

Fetching data from PIM for uid f2af54a2-6820-4f1b-99e7-d6e64642bdf3…
Done
------------------------------------------------------------
Ingredient list from PIM is :

None

------------------------------------------------------------
Supplier technical datasheet from PIM for uid f2af54a2-6820-4f1b-99e7-d6e64642bdf3 is:
https://produits.groupe-pomona.fr/nuxeo/nxfile/default/f2af54a2-6820-4f1b-99e7-d6e64642bdf3/pprodad:technica
lSheet/FT-163146_Hous%20echel%20GNZ%20PE%2020u%2025U_Barbier.pdf?changeToken=78-0
------------------------------------------------------------
Downloading content of technical datasheet file…
Done!
------------------------------------------------------------
Parsing content of technical datasheet file…
Done!
------------------------------------------------------------
Ingredient list extracted from technical datasheet:

mini 16,56 ( - 10 % )

------------------------------------------------------------

============================================================
============================================================
```