

June 11, 2020

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

```
[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
# 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

[3]: # monkeypatch _repr_latex_ for better inclusion of dataframes output in report
def _repr_latex_(self, size='scriptsize',):
    return(f"\\resizebox{{\\linewidth}}{{!}}{{\\begin{{{size}}}}\\centering{{{self.to_latex()}}}\\end{{{size}}}}")
pd.DataFrame._repr_latex_ = _repr_latex_
```

On extrait les données depuis le PIM :

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

Done

[illegible]

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

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

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"

```
[6]: # 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,')
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 3407 records in the dataset,
and 9893 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.

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

```
[7]: <src.pimest.PIMIngredientExtractor at 0x7ff230e1fb20>
```

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

```
[8]: 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 2514 words.

Some words examples are :

- morilles
- kombu
- déshydraté
- 100
- eau
- graines
- de
- moutarde

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 :

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

```
[9]: (3407, 2514)
```

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

```
[10]: 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][1:10]:
    print(f'{inverse_voc[idx].ljust(7)}: {word_counts[idx]:5} occurrences')
```

Most frequent words in vocabulary are:

```
de      : 11544 occurrences
sucre   : 2069  occurrences
sel     : 1647  occurrences
eau     : 1266  occurrences
acide   : 1245  occurrences
huile   : 1241  occurrences
lait    : 1228  occurrences
poudre  : 1099  occurrences
en      : 972   occurrences
arôme   : 940   occurrences
```

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.

```
[20]: exec_count = 5
uids = list(out_of_scope_df.sample(exec_count, ).index)
i = 0

for uid in uids:
    try:
        model.compare_uid_data(uid)
        print('\n=====')
        print('\n=====')
        i += 1
    except:
        pass
    if i > exec_count:
        break
```

Fetching data from PIM for uid 78dd3d32-5c23-495e-a4f9-71f6b3c692d7...

Done

Ingredient list from PIM is :

Sens gourmet Sarl - 15/17 rue du travy ZI sénia 715 94657 Thiais T 01 49 79 98 29 F 01 48 85 36 32
info@sensgourmet.com Code 58050000 (500 gr) Code EAN 8414933570004 RSIPAC N° 31-04482/CAT-31.01506B
DESCRIPTION : Ingrédient d'origine naturel. Agent gélifiant pour des gelées complètement transparentes à base d'eau. Origine : Europe APPLICATIONS : La gélatine est un ingrédient d'origine naturelle qui a une grande capacité d'absorption des molécules d'eau. Il s'agit d'une gélatine très élastique avec une bonne absorption de l'eau. Une température de gélification : 60°C. Dosage : 50 g/kg COMPOSITION : Maltodextrine, agent épaississant: carrageenan (E407), dextrose, chlorure de potassium (E508), acidifiant: trisodium citrate (E331iii), agent épaississant: gomme de caroube (E410), saccharose.

Supplier technical datasheet from PIM for uid 78dd3d32-5c23-495e-a4f9-71f6b3c692d7 is:

https://produits.groupe-pomona.fr/nuxeo/nxfile/default/78dd3d32-5c23-495e-a4f9-71f6b3c692d7/pprodad:technicalSheet/FT-163464_Gelatine%20vegetale%20pot%20500Gx6_SENS%20GOURMET.pdf?changeToken=25-0

Downloading content of technical datasheet file...

Done!

Parsing content of technical datasheet file...

Done!

Ingredient list extracted from technical datasheet:

Code
Code EAN
RSIPAC N°

DESCRIPTION :

Ingrédient d'origine naturel. Agent gélifiant pour des gelées complètement transparentes à base d'eau.
Origine : Europe

APPLICATIONS :

La gélatine est un ingrédient d'origine naturelle qui a une grande capacité d'absorption des molécules d'eau.
Il s'agit d'une gélatine très élastique avec une bonne absorption de l'eau. Une température de gélification :
60°C. Dosage : 50 g/kg

COMPOSITION :

Maltodextrine, agent épaississant: carrageenan (E407), dextrose, chlorure de potassium (E508),
acidifiant: trisodium citrate (E331iii), agent épaississant: gomme de caroube (E410), saccharose.

PARAMETRES ORGANOLEPTIQUES

Aspect : poudre fine, couleur blanche
Goût : neutre, doux
Odeur : neutre

PROPRIETES PHYSIQUES ET CHIMIQUES

Humidité : maximum
ASH

VALEURS NUTRITIONNELLES

Energie
Graisses
Saturées
Glucides
Sucres
Protéines
Sel
Fibres

PROPRIETES MICROBIOLOGIQUES :

microorganismes aérobies mésophiles (cfu/g) Max 5000/g
E.coli en 0.1g
Salmonelle spp (cfu/25gr)
Moisissures (cfu/g)
Levures (cfu/g)
Total coliforms

Sens gourmet Sarl - 15/17 rue du travy - ZI sénia 715 - 94657 Thiais
T 01 49 79 98 29 - F 01 48 85 36 32 - info@sensgourmet.com

=====
=====
Fetching data from PIM for uid 49f82f07-94af-4437-b433-351aaa6837d8...
Done

Ingredient list from PIM is :

None

Supplier technical datasheet from PIM for uid 49f82f07-94af-4437-b433-351aaa6837d8 is:
https://produits.groupe-pomona.fr/nuxeo/nxfile/default/49f82f07-94af-4437-b433-351aaa6837d8/pprodad:technicalSheet/FT-182074_Bisc%20avoine%20myrt%20Bio%20bte%2045G%20TBegin_Mononaturel.docx.pdf?changeToken=21-0

Downloading content of technical datasheet file...
Done!

Parsing content of technical datasheet file...
Done!

Ingredient list extracted from technical datasheet:

45 g
50 g
50 X 40 X H 85 mm
4751018890904

24 cookies
1008 g
1200 g
170X160X H 190 mm

800 X 1000 X H 1400 mm
4704 units
196 boxes
28 boxes
215 kg
235 kg

=====

Fetching data from PIM for uid e6fdec57-8df2-477a-942a-7d47bba41fd4...
Done

Ingredient list from PIM is :

None

Supplier technical datasheet from PIM for uid e6fdec57-8df2-477a-942a-7d47bba41fd4 is:
https://produits.groupe-pomona.fr/nuxeo/nxfile/default/e6fdec57-8df2-477a-942a-7d47bba41fd4/pprodad:technicalSheet/FT-178340_Sabots%20securite%20T36%20blc_Mutexil.pdf?changeToken=25-0

Downloading content of technical datasheet file...
Done!

Parsing content of technical datasheet file...
Fetching data from PIM for uid c771677a-b017-43ff-a5de-0d8be410a830...
Done

Ingredient list from PIM is :

Huile de colza, huile d'olive (24 %), vinaigre de vin affiné (conservateur : E222), eau, sel, épice.

Supplier technical datasheet from PIM for uid c771677a-b017-43ff-a5de-0d8be410a830 is:
https://produits.groupe-pomona.fr/nuxeo/nxfile/default/c771677a-b017-43ff-a5de-0d8be410a830/pprodad:technicalSheet/FT-67277_Vinaigrette%20huile%20olive%20col%2024G_Gyma.pdf?changeToken=34-0

Downloading content of technical datasheet file...
Done!

Parsing content of technical datasheet file...
Done!

Ingredient list extracted from technical datasheet:

Huile de colza, huile d'olive (24 %), vinaigre de vin affiné (conservateur : E222*), eau, sel, épice.
*Contient :
SULFITES.
Rapeseed oil, olive oil (24%), refined wine vinegar (preservative: E222*), water, salt, spice. *Contains:
SULFITES.

=====

Fetching data from PIM for uid cbbd03b7-9cf4-47b3-bae9-28a9af1263ba...
Done

Ingredient list from PIM is :

Potiron 49%, graisse de palme, amidon de pomme de terre, pomme de terre 7,7%, sel, arômes (dont BLÉ, ORGE),
huile de maïs, sucre, LACTOSE, oignon grillé 1,9%, extrait de levure, maltodextrine, protéines de LAIT,
épices (poivre, noix de muscade).

Supplier technical datasheet from PIM for uid cbbd03b7-9cf4-47b3-bae9-28a9af1263ba is:
https://produits.groupe-pomona.fr/nuxeo/nxfile/default/cbbd03b7-9cf4-47b3-bae9-28a9af1263ba/pprodad:technicalSheet/FT-86667_Creme%20potiron%20bte%201,155K_Knorr.pdf?changeToken=26-0

Downloading content of technical datasheet file...
Done!

Parsing content of technical datasheet file...
Done!

Ingredient list extracted from technical datasheet:

Liste d'ingrédients: Potiron 49%, graisse de palme, amidon de pomme de terre, pomme de terre 7,7%, sel, arômes (dont BLÉ, ORGE), huile de maïs, sucre, LACTOSE, oignon grillé 1,9%, extrait de levure, maltodextrine, protéines de LAIT, épices (poivre, noix de muscade). Peut contenir : œuf, céleri et moutarde.

=====

[]: