

# 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]:
```

uid	designation \
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...

  

uid	ingredients
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...
```

content \

text

```
[500 rows x 5 columns]
```

```
[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,
      #                 )
      # )
```

2892dd68-e3a6-474c-b543-3ebfd3490658 NESCAFÉ® SPÉCIAL FILTRE\n\nDose individuelle de 2 g\nTechnologie  
 micro-grains\n\nCODE EAN (UC)\n\n3033710076017\n\nDENOMINATION LEGALE DU PRODUIT\n\nDESCRIPTION DU  
 PRODUIT\n\nCafé 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\n\nMODE OPERATOIRE\n\nPour obtenir\n\n1 café Court (DA)\n\n1 café  
 Long (DA)\n\nEau\n\n7\n\n12\n\ncl\n\ncl\n\nNESCAFÉ®\n\nSPÉCIAL FILTRE\n\n2\n\n2\n\nng\n\nng\n\nA reconstituer  
 avec de l'eau.\n\nTempérature de l'eau : 75°C\n\nPour une qualité optimale, utilisez de l'eau  
 filtrée.\n\nIngrédients : Café instantané, café torréfié moulu (3%).\n\nINGRÉDIENTS\n\nPROFIL  
 GUSTATIF\n\nIntensité\n\nConditionné sous atmosphère protectrice.\n\nENGAGEMENT QUALITÉ\n\n- 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\n\nRéférence PQG007-3.22.1\n\nVersion\n\nDate  
 d'application : \nPage 1/2\n\nG\n\n15/10/2019\n\nPrésentation\n\nCaractéristi -\n\nques  
 \n\nphysico-\nchimiques \n\nDéfinition\n\nOrigine\n\nDénomination \nlégale\n\nLentilles de couleur brun clair.  
 Elles sont de forme biconvexe et \nposèdent 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,  
 éperrage, 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\nImpropre\n\nBrisées\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\nmin\n\nNF V03707\n\nMicrobiologie\n\nIl  
 n'existe pas de réglementation concernant les exigences microbiologiques \npour ce  
 produit.\n\nPesticides\n\nMét...  
 3634fb1e-ee79-41d1-8aaa-084c1fae5bd5 FICHE TECHNIQUE \n\nPRODUIT FINI\n\n000100\n\nPurée de Poire Sans  
 Sucres Ajoutés\n\nDate d'application: 05/05/2014\n\nPage: 1/2\n\nCoupelles Aluminium 120 x 95  
 g\n\nDéfinition\n\nCe produit est une purée de fruits obtenue à partir des parties comestibles des fruits  
 (après broyage et sans \nconcentration notable).\n\nCe produit est sans sucres ajoutés: il contient uniquement  
 les sucres naturellement présents dans les fruits.\n\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\n\nDésignation et liste des ingrédients\n\nValeurs nutritionnelles (pour 100 g)\n\nDésignation
légale :\n\nPurée de Poires sans sucres ajoutés *\n* Contient les sucres naturellement présents dans \nles
fruits\n\nListe des ingrédients :\n\nPoire 99,9%, antioxydant: acide ascorbique.\n\nMatières
grasses\n\nEnergie\n\n65 kcal\n\n273 kJ\n\ndont acides gras saturés\n\nGlucides\n\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'))

[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/a0492df6-9c76-4303-8813-65e...
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...
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\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\r%\xc7\xec\x8f\xa2\r\n5 0 obj\r\n<</Len...
...
e67341d8-350f-46f4-9154-4dbbb8035621  b'%PDF-1.7\r\n%\xb5\xb5\xb5\r\n1 0 obj\r\n...
a8f6f672-20ac-4ff8-a8f2-3bc4306c8df3  b'%PDF-1.5\r\n%\xb5\xb5\xb5\r\n1 0 obj\r\n...
0faad739-ea8c-4f03-b62e-51ee592a0546  b'%PDF-1.5\r\n%\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=['Blocs'])
.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  
Absence  
Absence  
Absence  
Absence  
Absence  
Absence  
Absence  
Absence  
Absence  
Absence  
Absence  
Absence  
Absence  
-----

Caractères microbiologiques  
-----

/ Microbiological characteristics  
-----

Microorganismes aérobies 30 °C  
Escherichia coli  
Salmonelles  
Levures  
Moisissures  
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 torréfié 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 torréfié 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]: ingredients \

uid
d1be6f74-1e0e-4631-bb4a-6b16b9fc908f      sucre*, LAIT en poudre*, beurre de cacao*, pât...
49b11281-34ea-44b0-a11c-4ae21d4c58e3      NaN
d59d96cb-0230-4090-8220-78ce8496fd91      Amidon de maïs* - Lait écrémé* - Sel - Fécul...
5adc7512-6168-4966-ae3f-f6ec133bf56e      semoule de blé dur supérieure et de l'eau
75088d85-f350-4d81-a7f4-954411ba089e      NaN
a0492df6-9c76-4303-8813-65ec5ccbfa70      Eau, maltodextrine, sel, arômes, sucre, arôme ...
e521bd01-f2bb-4e00-9ae0-0151a1c7a047      Sucre, cacao maigre en poudre (beurre de cacao...
8dec0469-c9f5-4139-be25-efa258959444      Sucre; sirop de glucose; graisse de palme; hum...
345591f4-d887-4ddc-bb40-21337fa9269d      Gésier de dinde émincé 50%, graisse de canard ...
41da4d6f-7e9f-4f95-bf2a-2acdd7138cd9      Purée de tomates mi réduite (64%), sucre, vina...
21233a00-bc20-40fc-acb9-ee2e2321cac2      NaN
9ef0d351-4982-4a2d-88a9-85573dc396dc      Sirop de glucose-fructose, framboises 35%, suc...
2892dd68-e3a6-474c-b543-3ebfd3490658      Café instantané, café torréfié moulu (3%).
df1caa23-9714-4659-803b-33501d64eead      sucre, pâte de cacao, beurre de cacao, cacao m...
7f622727-e4ad-45cc-9af4-4509acf91154      Eau, huile de tournesol, beurre 9,5 %, jaune d...
54f40033-f9cf-411c-81a5-11974f6715aa      Piment rouge fort équeuté* (85%), cumin, ail m...
536361db-1bbb-4e64-ae53-d970eeac7db2      Sucre, amidon de maïs, arôme vanille
a2418174-e16a-41e0-ac14-c87208fb3529      Salicornes de culture, eau, sel, acide citrique
046cdb1f-1915-4916-8874-902cc5ec73be      cèpes 70% (Boletus edulis et respective famill...
b7d7621a-fcdd-4487-9b38-e07fae698c4a      Eau, haricots verts, sel.

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 - Fécul...
5adc7512-6168-4966-ae3f-f6ec133bf56e      Ingrédients: semoule de blé dur supérieure et ...
75088d85-f350-4d81-a7f4-954411ba089e      Boisson gazeuse aromatisée au jus de fruit à b...
a0492df6-9c76-4303-8813-65ec5ccbfa70      Eau, maltodextrine, sel, arômes, sucre, arôme ...
e521bd01-f2bb-4e00-9ae0-0151a1c7a047      Sucre, cacao maigre en poudre (beurre de cacao...
8dec0469-c9f5-4139-be25-efa258959444      NaN
345591f4-d887-4ddc-bb40-21337fa9269d      Gésier de dinde émincé 50%, graisse de canard...
41da4d6f-7e9f-4f95-bf2a-2acdd7138cd9      Liste des ingrédients : Purée de tomates mi r...
21233a00-bc20-40fc-acb9-ee2e2321cac2      Boisson gazeuse aromatisée au jus de fruit à b...
9ef0d351-4982-4a2d-88a9-85573dc396dc      Liste ingrédients : Sirop de glucose-fructose,...
2892dd68-e3a6-474c-b543-3ebfd3490658      - NESTLÉ a un système de management de la qual...
df1caa23-9714-4659-803b-33501d64eead      Liste des Ingrédients:\nsucre, pâte de cacao, ...
7f622727-e4ad-45cc-9af4-4509acf91154      Eau, huile de tournesol, beurre 9,5 %, jaune d...
54f40033-f9cf-411c-81a5-11974f6715aa      A) Ingrédients : \n \nPiment rouge fort équ...
536361db-1bbb-4e64-ae53-d970eeac7db2      ajouter le produit à la préparation avec les a...
a2418174-e16a-41e0-ac14-c87208fb3529      Se consomment en légumes d'accompagnement avec...
046cdb1f-1915-4916-8874-902cc5ec73be      CODE DU PRODUIT: \nNOME DU PRODUIT: \nFORMAT: ...
b7d7621a-fcdd-4487-9b38-e07fae698c4a      Égoutter, ne pas\nrincer. Faire sauter 3\nminu...

```

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 prédite"],
            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)
```