

Classifiez automatiquement des biens de consommation



*By Maria BOUCHEHBOUN
For Openclassrooms
Mentor : Mustafa Ankarali*



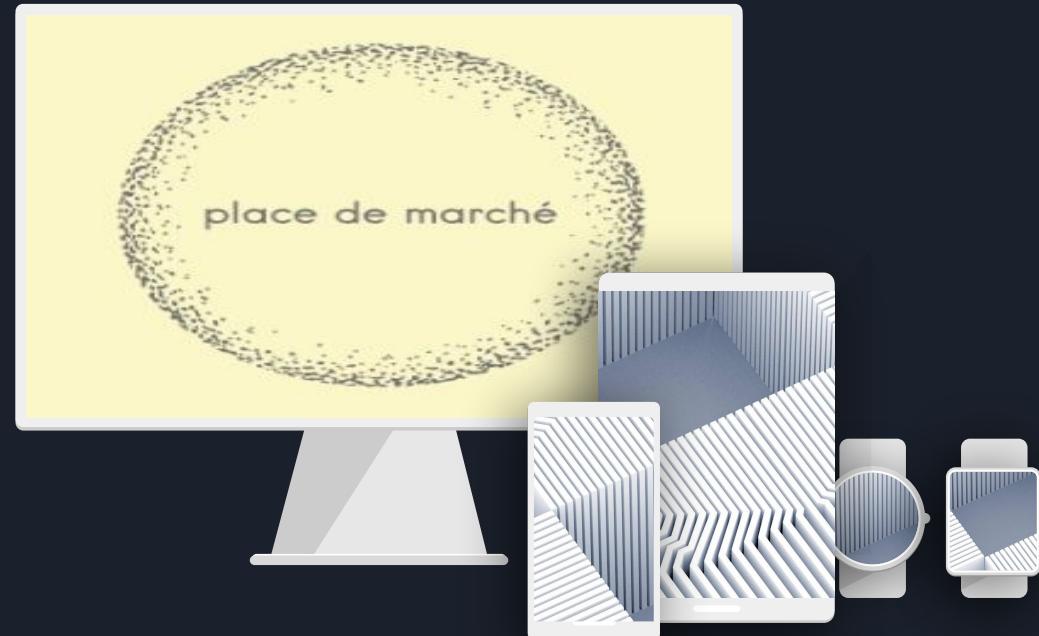
Plan

- 
1. Contexte & Mission
 2. Exploration et traitement Texte
 3. Exploration et traitement Image
 4. RGPD
 5. API
 6. Conclusion

Contexte & Mission

Vous êtes Data Scientist au sein de l'entreprise "**Place de marché**", qui souhaite lancer une marketplace e-commerce

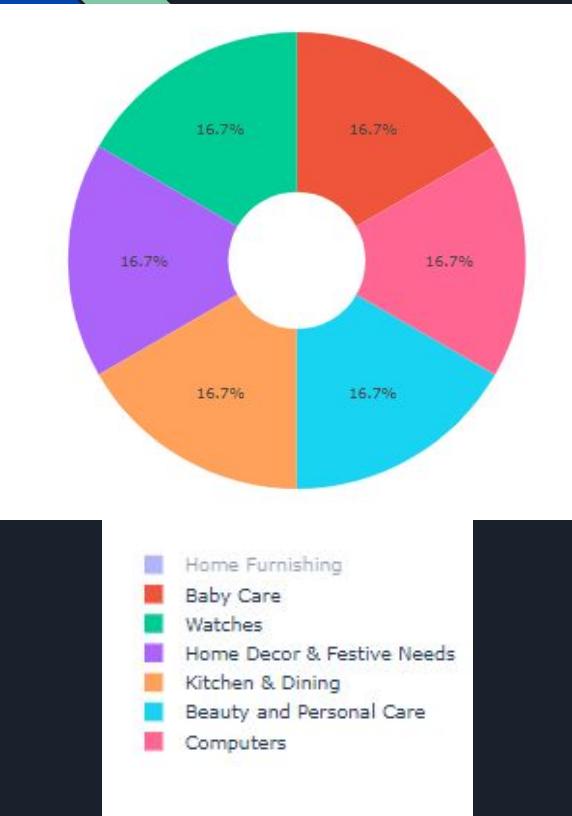
Pour l'instant, l'attribution de la catégorie d'un article est effectuée manuellement par les vendeurs, et est donc peu fiable. De plus, le volume des articles est pour l'instant très petit. Pour rendre l'expérience utilisateur des vendeurs (faciliter la mise en ligne de nouveaux articles) et des acheteurs (faciliter la recherche de produits) la plus fluide possible, et dans l'optique d'un passage à l'échelle, **il devient nécessaire d'automatiser cette tâche.**



Méthodologie



Avant, explorons tout le dataset



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1050 entries, 0 to 1049
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   uniq_id          1050 non-null   object 
 1   crawl_timestamp  1050 non-null   object 
 2   product_url      1050 non-null   object 
 3   product_name     1050 non-null   object 
 4   product_category_tree  1050 non-null   object 
 5   pid               1050 non-null   object 
 6   retail_price     1049 non-null   float64
 7   discounted_price 1049 non-null   float64
 8   image             1050 non-null   object 
 9   is_FK_Advantage_product  1050 non-null   bool   
 10  description       1050 non-null   object 
 11  product_rating    1050 non-null   object 
 12  overall_rating   1050 non-null   object 
 13  brand             712 non-null   object 
 14  product_specifications 1049 non-null   object 
dtypes: bool(1), float64(2), object(12)
memory usage: 116.0+ KB
```

```
cat_niveau_1 : 7
cat_niveau_2 : 62
cat_niveau_3 : 242
```



Exploration et traitement texte

Exemples de descriptions:

0 Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester ...

1 Specifications of Sathiya's Cotton Bath Towel (3 Bath Towel, Red, Yellow, Blue) Bath Towel Features Machine Washable Yes Material Cotton Design Self Design General Brand Sathiya's Type Bath Towel GSM 500 Model Name Sathiya's cotton bath towel Ideal For Men, Women, Boys, Girls Model ID asvbl322 Col...

2 Key Features of Eurospa Cotton Terry Face Towel Set Size: small Height: 9 inch GSM: 360,Eurospa Cotton Terry Face Towel Set (20 PIECE FACE TOWEL SET, Assorted) Price: Rs. 299 Eurospa brings to you an exclusively designed, 100% soft cotton towels of export quality. All our products have soft text...
Name: description, dtype: object

Après traitement : mettre au minuscule; contractions, noise removal, ponctuation

Exploration et traitement texte

1420

avant:

Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester fabric.it features an eyelet style stitch with Metal Ring.It makes the room environment romantic and loving.This curtain is ant- wrinkle and anti shrinkage and have elegant appearance.Give your home a bright and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sunlight.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyester Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In the Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material Polyester

1397

Après traitement : mettre au minuscule; contractions, noise removal, ponctuation...

après:

key features of elegance polyester multicolor abstract eyelet door curtain floral curtain,elegance polyester multicolor abstract eyelet door curtain 213 cm in height pack of 2 price rs 899 this curtain enhances the look of the interiorsthis curtain is made from 100 high quality polyester fabricit features an eyelet style stitch with metal ringit makes the room environment romantic and lovingthis curtain is ant wrinkle and anti shrinkage and have elegant appearancegive your home a bright and modernistic appeal with these designs the surreal attention is sure to steal hearts these contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening you create the most special moments of joyous beauty given by the soothing prints bring home the elegant curtain that softly filters light in your room so that you get the right amount of sunlight specifications of elegance polyester multicolor abstract eyelet door curtain 213 cm in height pack of 2 general brand elegance designed for door type eyelet model name abstract polyester door curtain set of 2 model id duster25 color multicolor dimensions length 213 cm in the box number of contents in sales package pack of 2 sales package 2 curtains body & design material polyester

Exploration et traitement texte

	'tokenized'	'tokenize_stem'	lemmatizer	stemmer
mots de la colonne	52256	53554	362321	695908
mots uniques	4776	4011	27	30

Home Furnishing

product shipping free package cushion delivery buy flipkartcom polyester towel color genuine print
design cover pack cotton

Baby Care

sale print wash neck boy fabric general feature detail cotton baby girl color type sleeve ideal content number

Watches

flipkartcom woman men buy dial product online genuine discount delivery great cash guarantee day analog replacement shipping india

Home Decor & Festive Needs

best cash color replacement guarantee delivery online showpiece inch feature buy genuine home free wall make price day shipping

Kitchen & Dining

coffee day love best product quality rockmantra design price one material perfect give pack mug safe ceramic make feature

Beauty and Personal Care

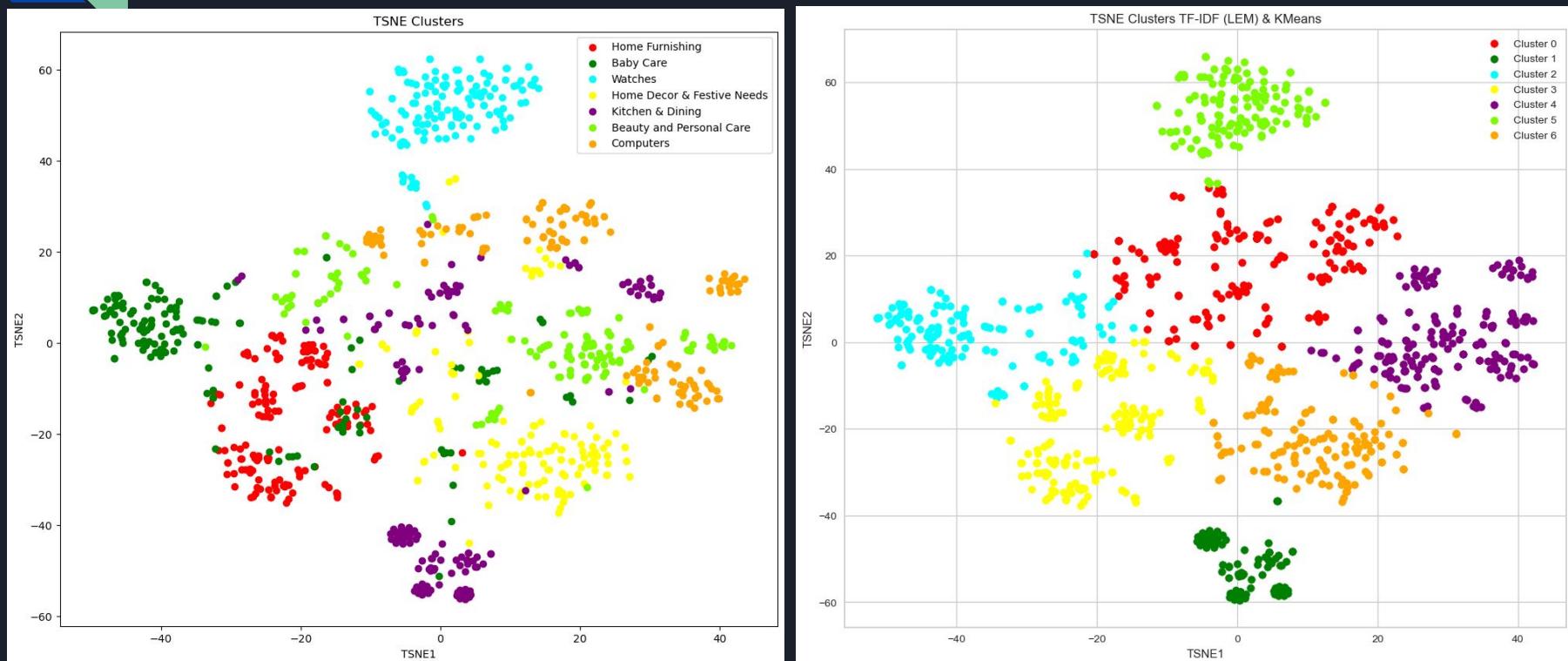
genuine skin free combo set hair shipping delivery flipkartcom box day color type online buy price guarantee replacement

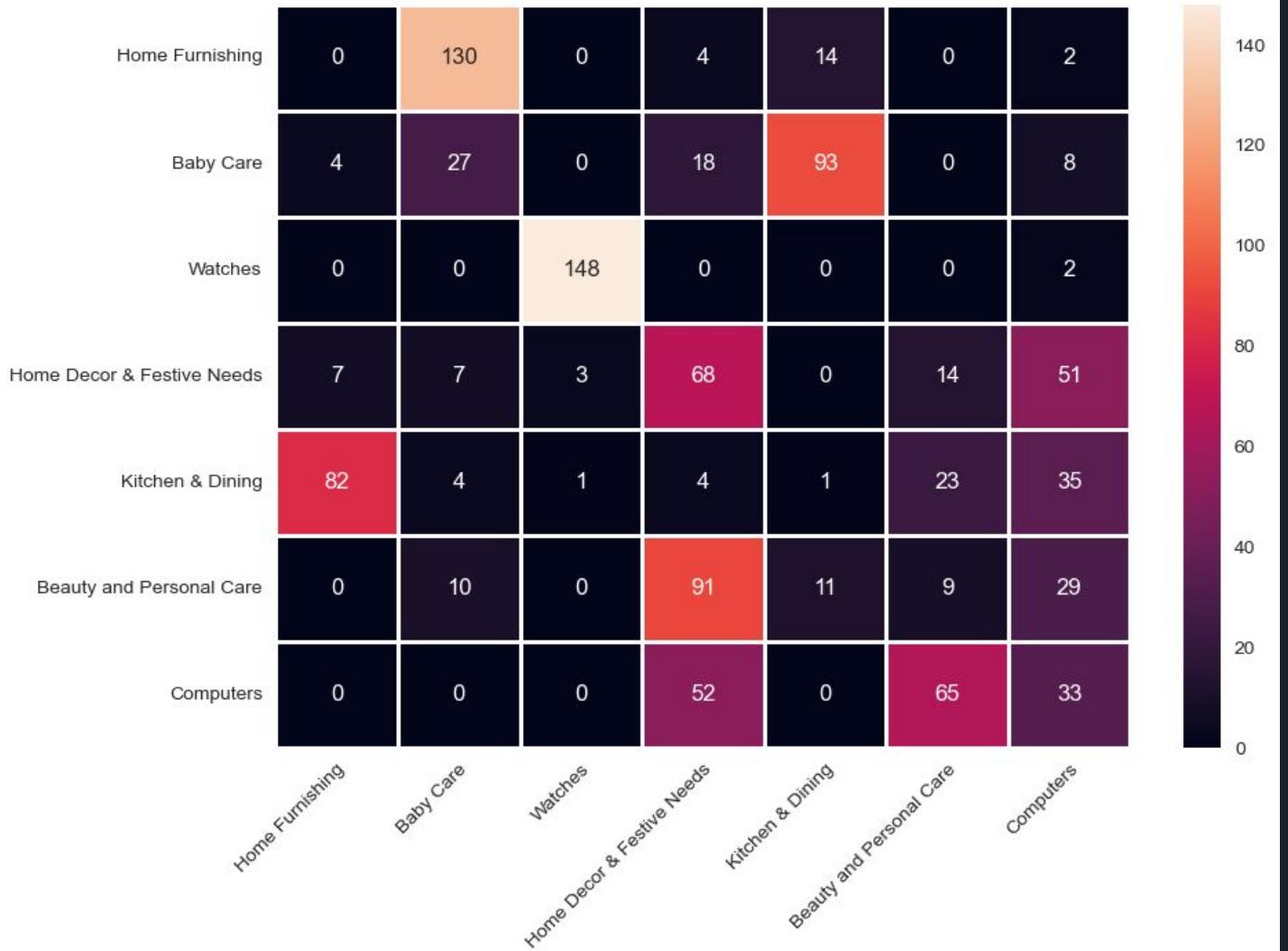
Computers

guarantee price day adapter cash power delivery skin buy battery quality genuine laptop usb replacement product warranty shipping flipkartcom

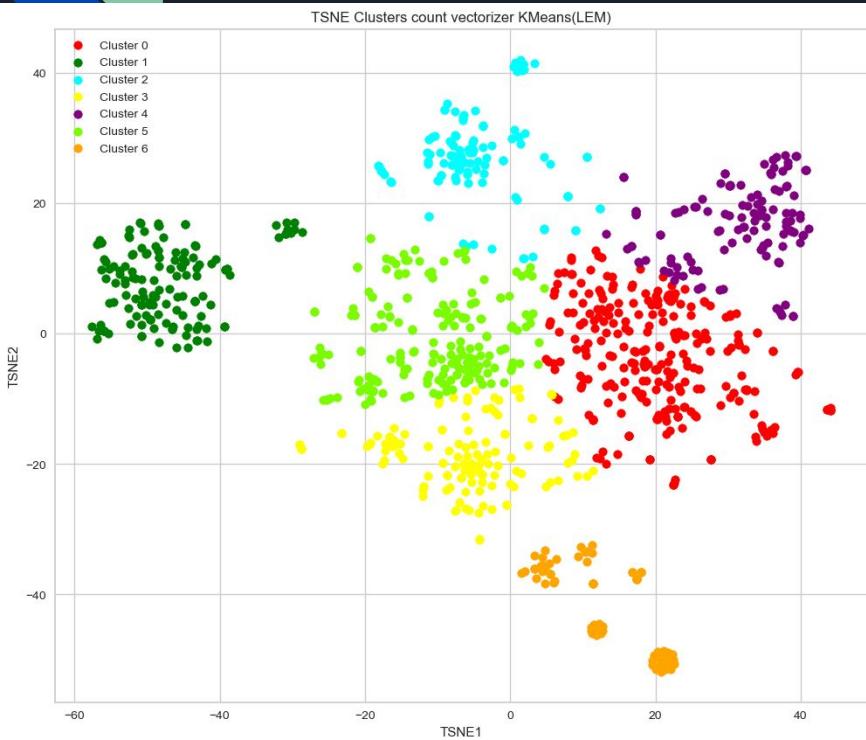
TF-IDF

ARI score suite à la lemmatisation & TF-IDF : 0.4691505881059818

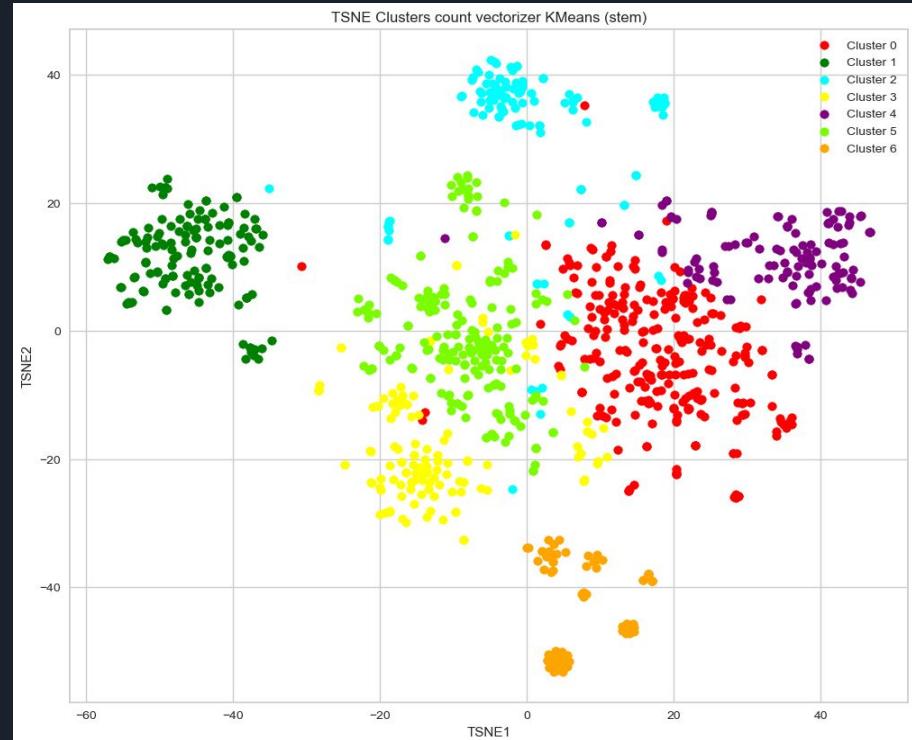




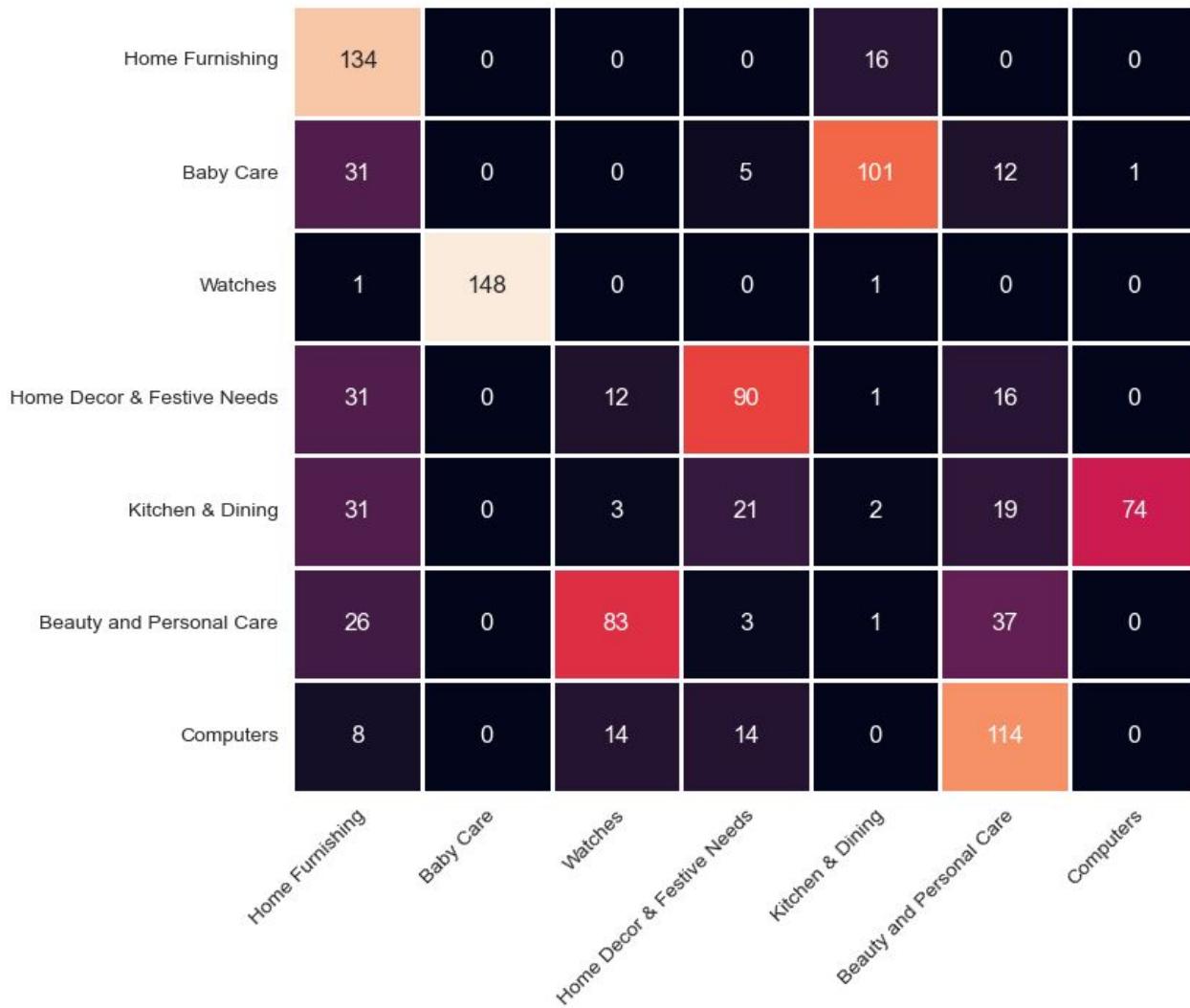
Count vectorizer



ARI score CountVectorizer (Lemmatizer): 0.449

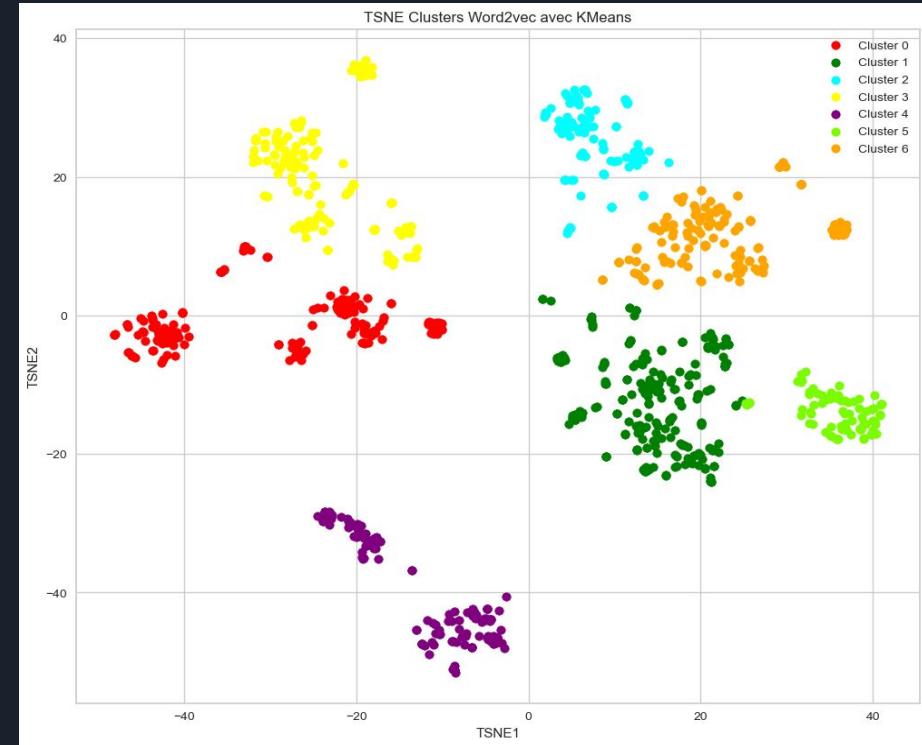
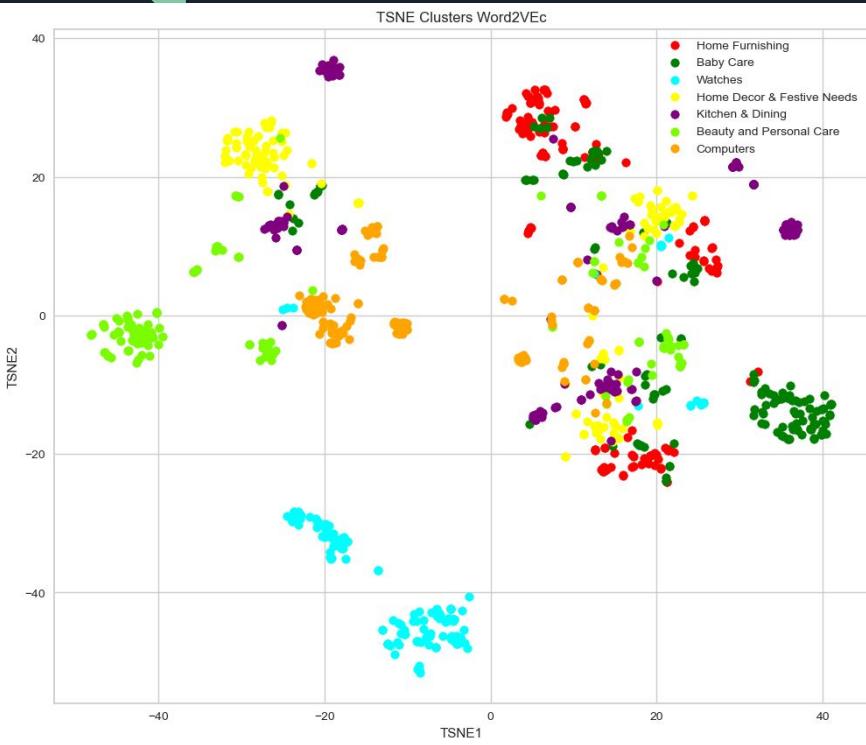


ARI score CountVectorizer (Stemming): 0.443



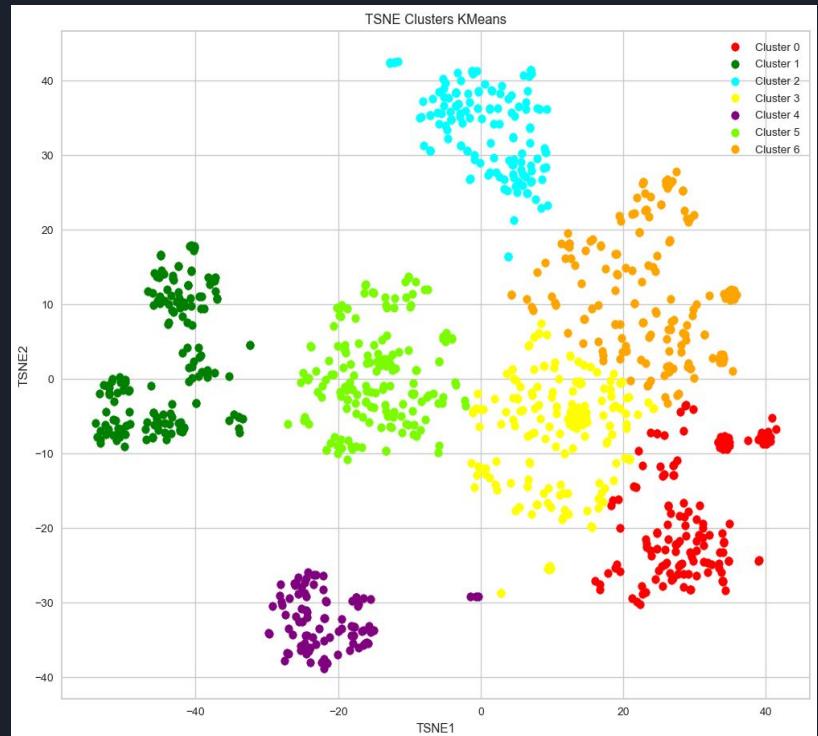
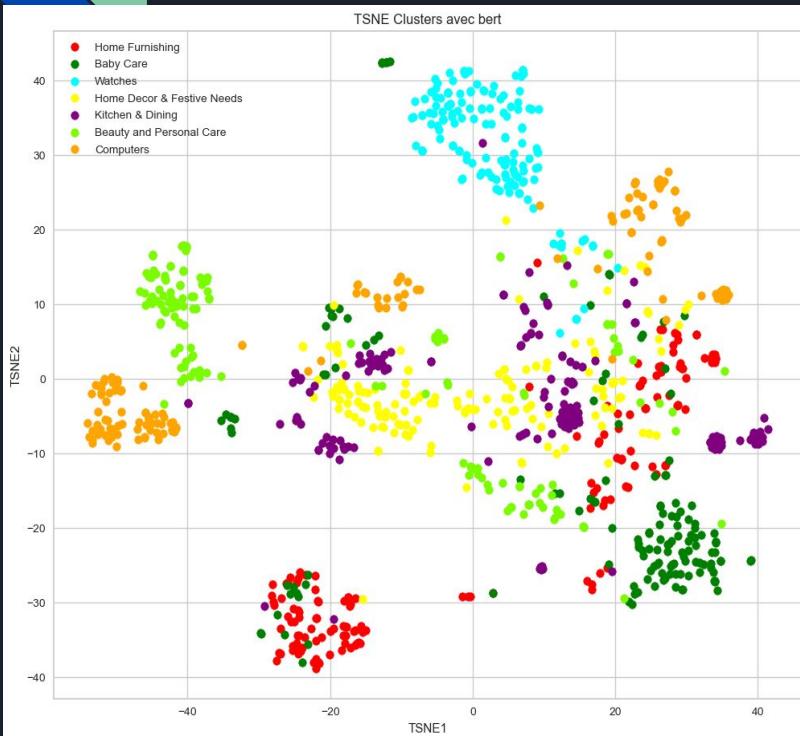
Word2vec

ARI score Word2Vec (LEM) : 0.301



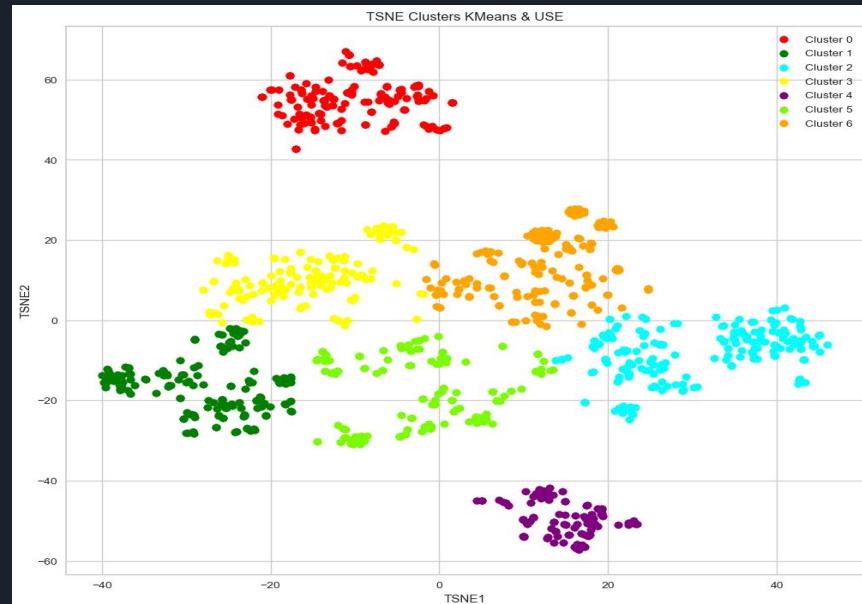
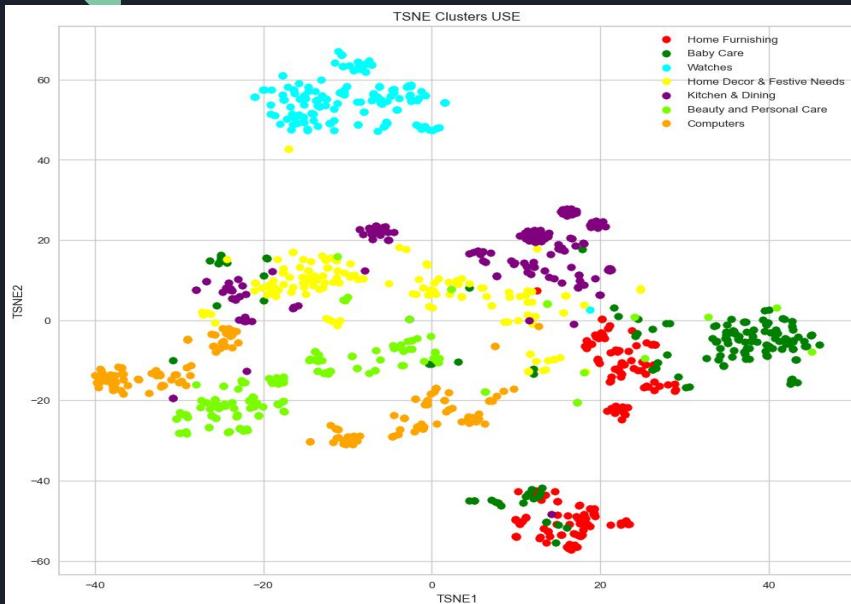
BERT

ARI score BERT (LEM) : 0.31178707171002695



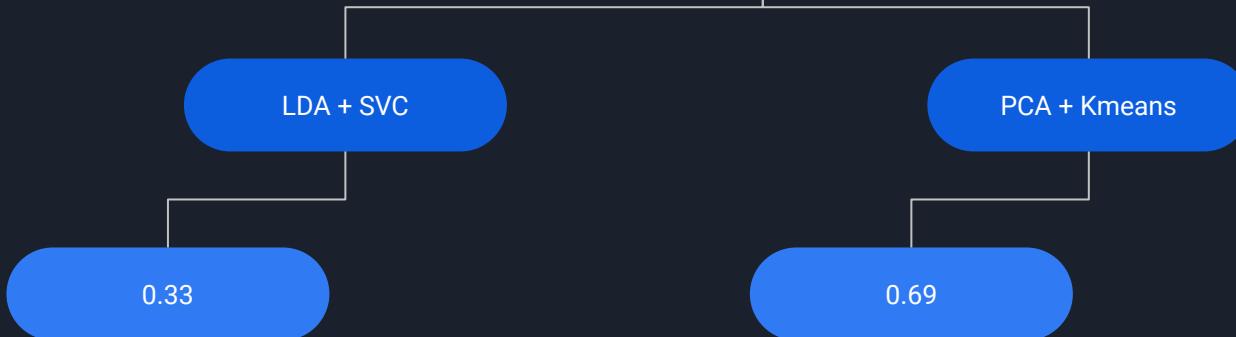
USE

ARI score suite à la lémmatisation : 0.463





	TF-IDF	Word2vec	BERT	USE
ARI Score	0.48	0.43	0.31	0.46



K-means est notre meilleur modèle de classification avec un score de 0.69



Exploration et traitement Image

Chargement

Exploration

Traitement

Faisabilité?

Classification

Chargement des
Librairies
chargement du dataset

Lecture du data set
Création Train/test

Resize
Réduction du bruit
contraste
SIFT

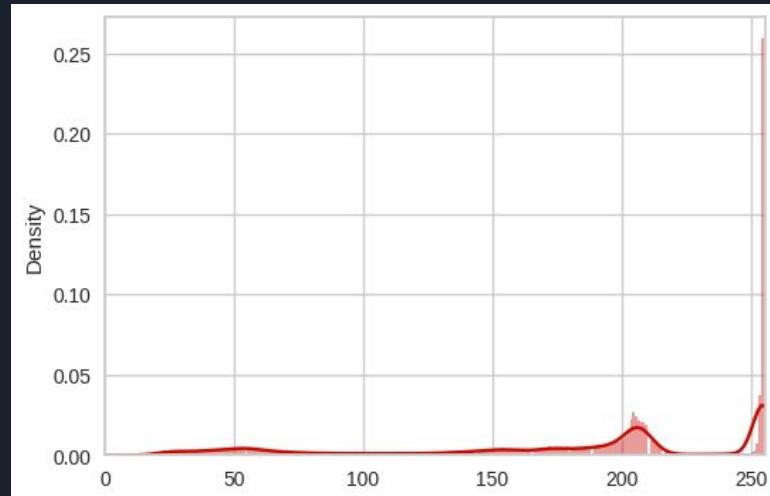
CNN
TSNE

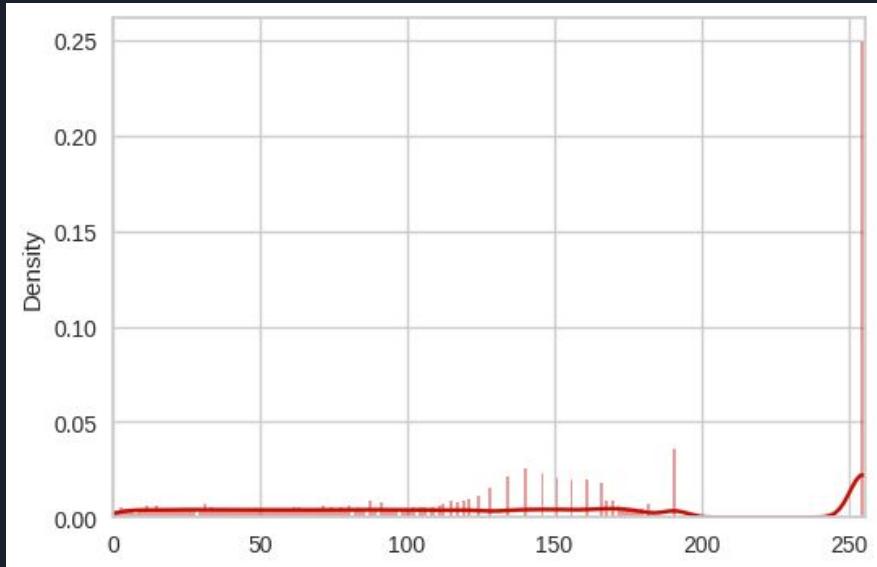
TSNE
PCA
Kmeans

```
Image('/content/drive/MyDrive/OC: Data Science/Datasets/Projet 6/Flipkart/Images/1e8741b5ae27a513546c94b3f3312aee.jpg', width=200)
```

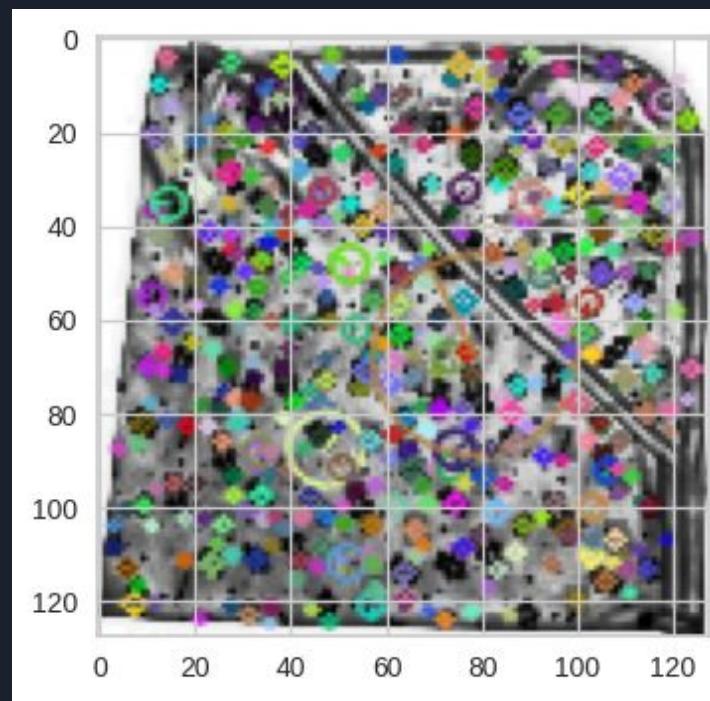
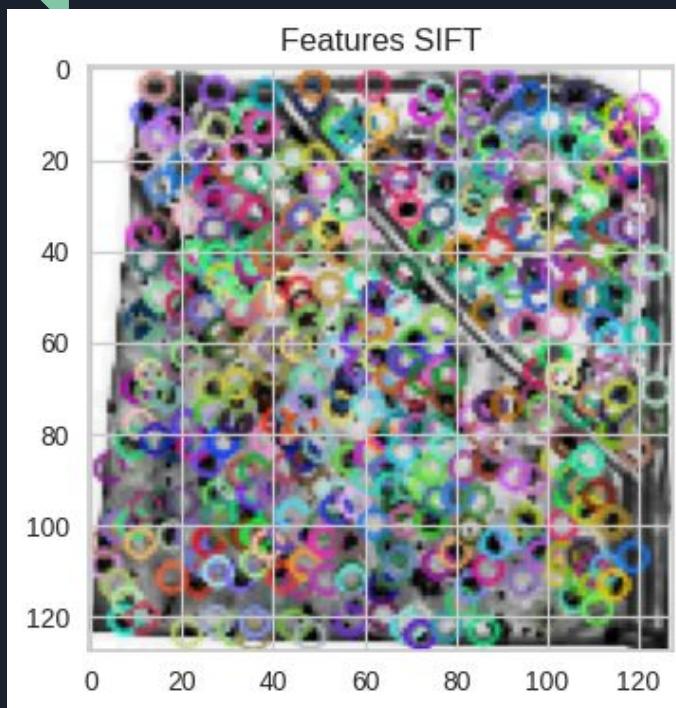








SIFT



CNN

```
[ ] predictions_test.shape
```

```
[ ] (263, 7)
```

```
[ ] from sklearn.metrics import accuracy_score
[ ] print("\033[1;34;41mAccuracy score")
[ ] accuracy_score(np.asarray(predictions_test).round(2).reshape(-1,1),np.asarray(test_array_cats).round(2).reshape(-1,1))
```

```
[ ] Accuracy score
```

```
[ ] 0.7555676262900598
```

```
[ ] ohe.categories_[0][1]
```

```
[ ] 'Beauty and Personal Care'
```

```
[ ] ohe.categories_[0][5]
```

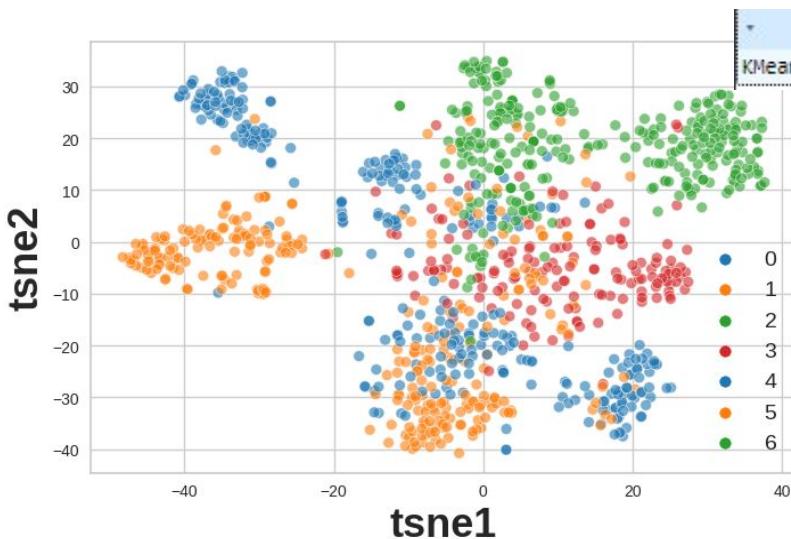
```
[ ] 'Kitchen & Dining'
```

```
[ ] ohe.categories_[0][6]
```

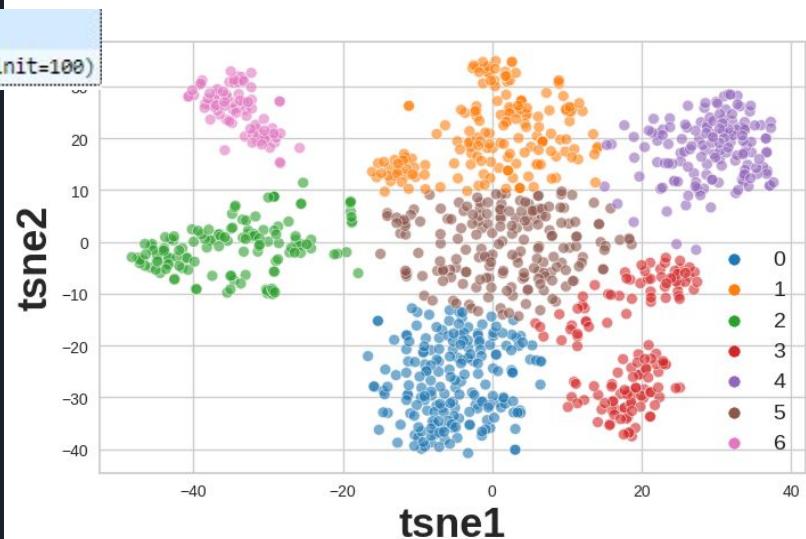
```
[ ] 'Watches'
```

Classification non supervisée

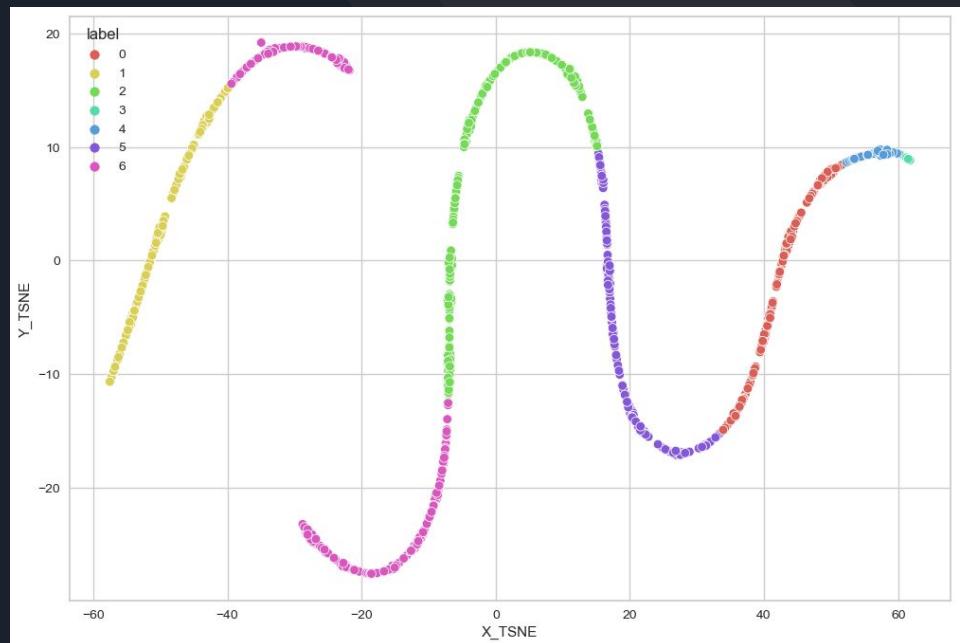
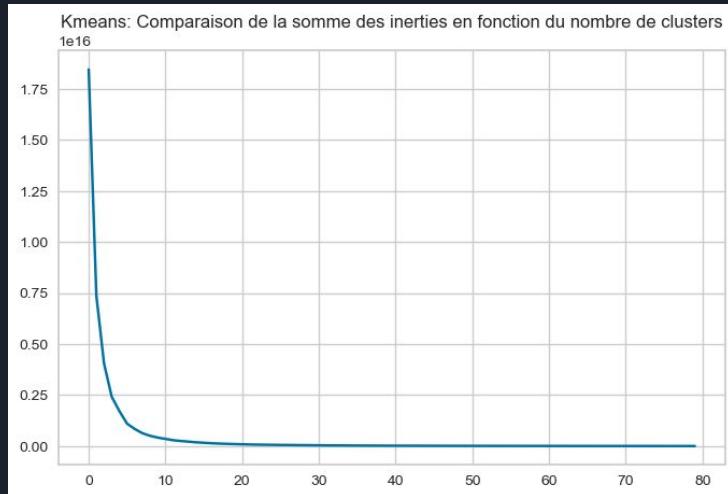
TSNE selon les vraies classes



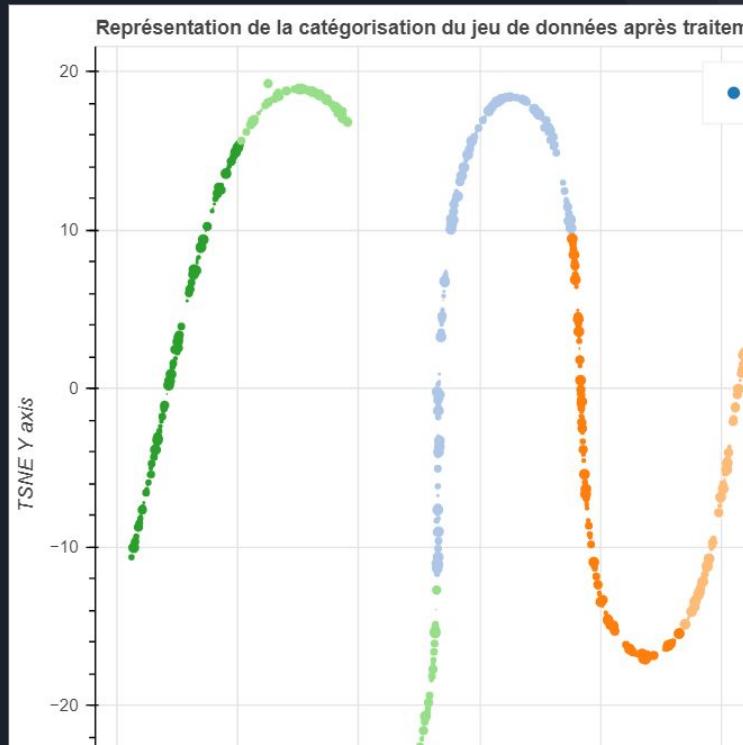
TSNE selon les clusters



Classification Supervisée



Graphe prévisions



RGPD : Règlement général sur la protection des données

PRINCIPE 1 FINALITÉ	LE PRINCIPE DE FINALITÉ Les données sont collectées pour un but bien déterminé et légitime et ne sont pas traitées ultérieurement de façon incompatible avec cet objectif initial. Ce principe limite la manière dont le responsable de traitement pourra utiliser ou réutiliser ces données dans le futur. Exemple : <i>Un maire ne pourra pas se servir du fichier des inscriptions scolaires pour faire de la communication politique. La liste électorale pourra en revanche être utilisée à une telle fin.</i>
PRINCIPE 2 PERTINENCE	LE PRINCIPE DE PERTINENCE Seules les données strictement nécessaires à la réalisation de l'objectif poursuivi doivent être collectées. Il s'agit donc de minimiser autant que possible la collecte des données. Exemple : <i>Seule la mention « Personne en fauteuil roulant » doit être enregistrée si la précision du handicap ou de la maladie en cause n'est pas nécessaire pour assurer une prise en charge adéquate de l'intéressé.</i>
PRINCIPE 3 DURÉE LIMITÉE	LE PRINCIPE DE DURÉE LIMITÉE DE CONSERVATION Les données ne doivent être conservées sous une forme identifiante et en « base active » que le temps nécessaire à la réalisation de l'objectif poursuivi et doivent être par la suite détruites, anonymisées ou archivées dans le respect des obligations légales applicables en matière de conservation des archives publiques. Exemple : <i>Les données sont conservées en base active par la collectivité tant que l'administré bénéficie d'une prestation publique. Au-delà, elles sont supprimées de cette base mais peuvent être archivées.</i>



PRINCIPE 4 SÉCURITÉ

LE PRINCIPE DE SÉCURITÉ

Le responsable de traitement de la collectivité doit prendre toutes les mesures utiles pour garantir l'intégrité et la confidentialité de ces données, en s'assurant notamment que des tiers non autorisés n'y auront pas accès. Ces mesures seront déterminées en fonction des risques (sensibilité des données, objectif du traitement) et seront à la fois d'ordre physique, logique, technique et organisationnel (sécurisation des locaux, armoires et postes de travail, gestion stricte des habilitations et droits d'accès informatiques, encadrement des opérations sous-traitées).

Exemple : *Les agents doivent disposer d'un mot de passe individuel régulièrement changé et leurs droits d'accès aux fichiers sont définis en fonction de leurs besoins réels en lien avec l'exercice de leur mission.*

PRINCIPE 5 DROITS DES PERSONNES

LE PRINCIPE DES DROITS DES PERSONNES

Les personnes concernées par les traitements doivent conserver la maîtrise des données qui les concernent. Ainsi, la loi prévoit que les données ne peuvent être collectées à l'insu des personnes concernées, qui doivent avoir été informées au préalable de cette opération, de sa finalité, des destinataires des données et des modalités d'exercice de leurs droits. Ces droits « Informatique et Libertés », qu'elles peuvent exercer auprès de la collectivité qui détient ces informations sont :

- le droit d'accéder à leurs données et d'en obtenir une copie ;
- le droit de les rectifier
- le droit de s'opposer à leur utilisation, sauf si le traitement répond à une obligation légale (par exemple, un administré ne peut s'opposer à figurer dans un fichier d'état civil).

La loi pour une République numérique d'octobre 2016 est venue renforcer ces droits en prévoyant notamment la possibilité pour les personnes concernées de les exercer par voie électronique, ainsi que de donner des directives relatives à la conservation, à l'effacement et à la communication de leurs données après leur décès.

API

```
import requests

url = "https://edamam-food-and-grocery-database.p.rapidapi.com/api/food-database/v2/parser"
querystring = {"ingr": "champagne"}

headers = {
    "X-RapidAPI-Key": "eea546e3c0msh22a68ee5e0e8e57p1d5922jsne487409fe309",
    "X-RapidAPI-Host": "edamam-food-and-grocery-database.p.rapidapi.com"
}
```

	foodId	label	category	foodContentsLabel	image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	None	https://www.edamam.com/food-img/a71/a718cf3c52...
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	None
2	food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
3	food_a9e0ghsamvoc45bw2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	None
4	food_an4ijueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	None
5	food_bmu5dmkazwuwpa5prh1daa8xs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	https://www.edamam.com/food-img/ab2/ab2459fc2a...
6	food_apl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne;...	None
7	food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	None
8	food_am5egz6aq3fpjaf8xpdkbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	None
9	food_bcz8rhiajk1fuva0vkimeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	None



Conclusion

- 
1. la classification supervisée nous montre qu'on peut bien classer les images selon 7 classes
 2. On devrait peut être utiliser l'API pour mieux alimenter la base des photos et mieux affiner la classification
 3. comme on est limite en terme de slides, nous ne voyons que les heatmap de deux outils (dans le notebook ya toutes les heatmap) permettant de voir la distribution selon les classes de départs et clusters apres modèle.
 4. une rotation a été ajoutée à la fonction permettant le traitement des images

Merci !

