Projet P6

Classification automatique de bien de consommation

OPENCLASSROOMS

Introduction

• DataScientist pour Place de Marche -> Faire une étude de faisabilité d'un moteur de classification d'articles en différentes catégories.

• **Objectifs :**

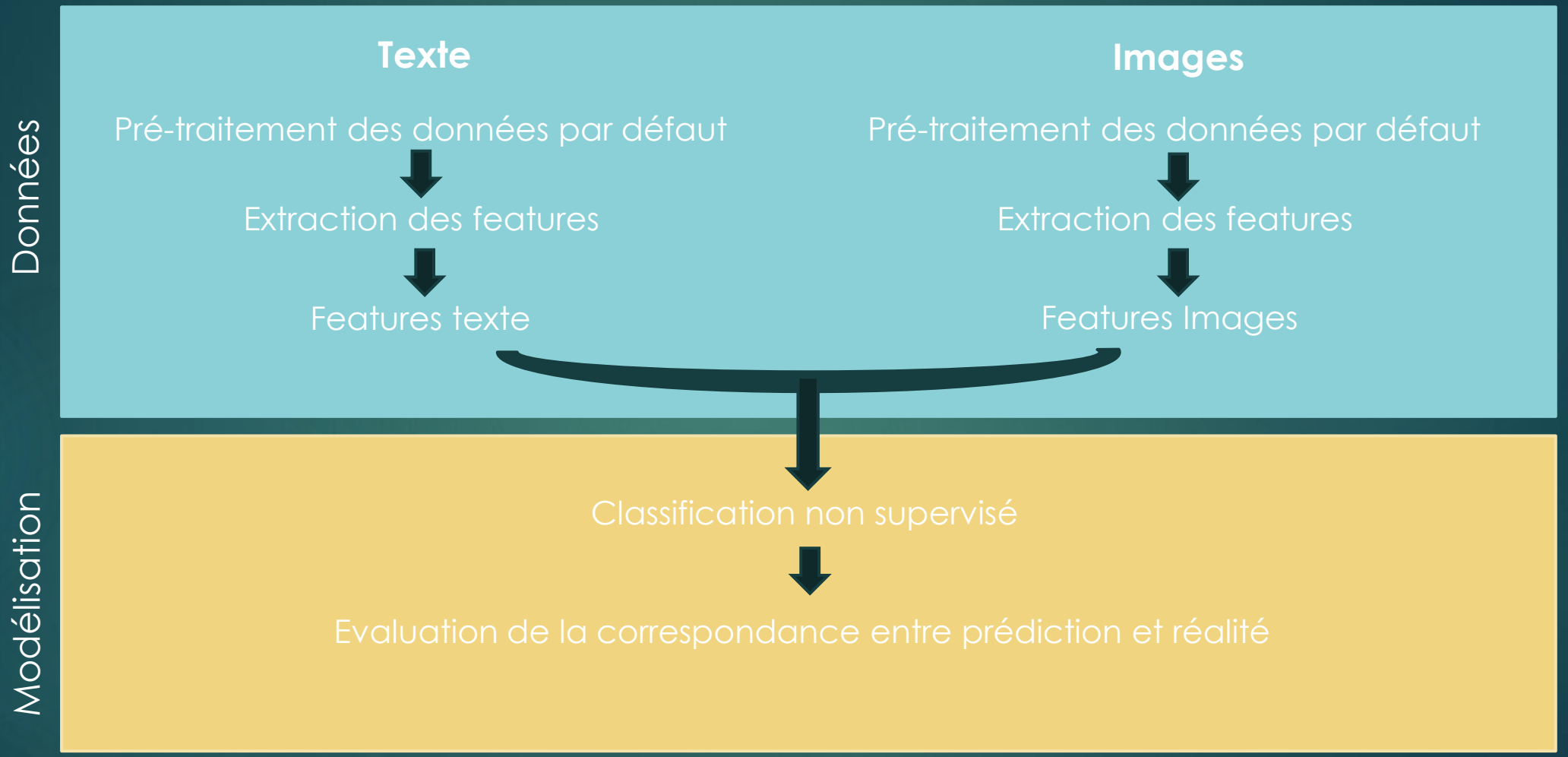
- Travail sur une base de données limitée de 1050 produits

- Extraction des features texte

- Extraction des features images

- - Obtenir une classification pertinente des produits de manière non-supervisée

Démarche


Données
Texte
Pré-traitement des données par défaut
Extraction des features
Features texte
Images
Pré-traitement des données par défaut
Extraction des features
Features Images
Modélisation
Classification non supervisé
Evaluation de la correspondance entre prédiction et réalité

Données

Jeux de données de 1050 lignes et 15 colonnes :

- **Description du produit**
- **Categories du produit**
- **Images**

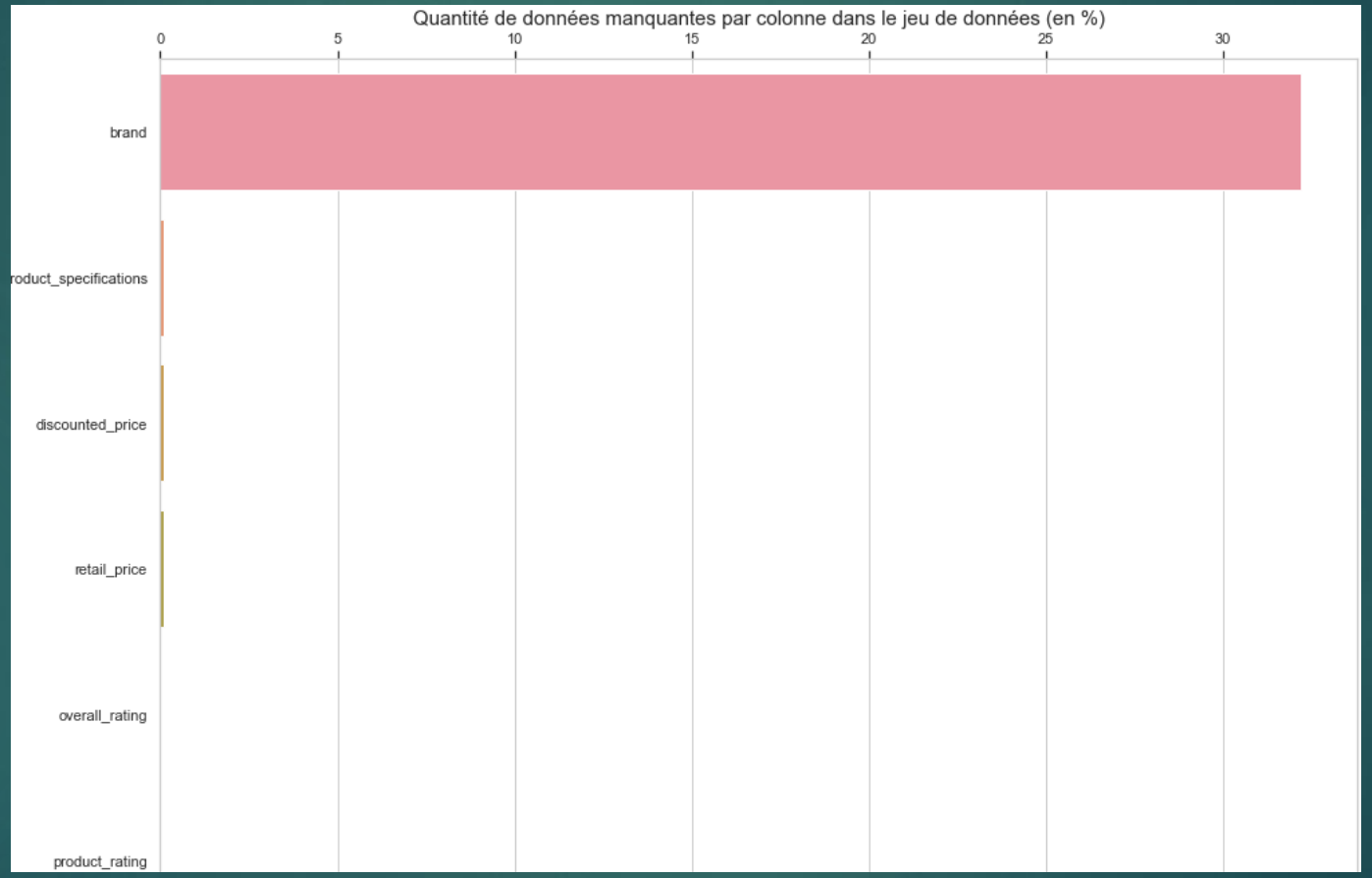
	product_category_tree	description	image
0	["Home Furnishing >> Curtains & Accessories >>...	Key Features of Elegance Polyester Multicolor ...	55b85ea15a1536d46b7190ad6fff8ce7.jpg
1	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	Specifications of Sathiyas Cotton Bath Towel (...	7b72c92c2f6c40268628ec5f14c6d590.jpg
2	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	Key Features of Eurospa Cotton Terry Face Towe...	64d5d4a258243731dc7bbb1eef49ad74.jpg
3	["Home Furnishing >> Bed Linen >> Bedsheets >>...	Key Features of SANTOSH ROYAL FASHION Cotton P...	d4684dcdc759dd9cdf41504698d737d8.jpg
4	["Home Furnishing >> Bed Linen >> Bedsheets >>...	Key Features of Jaipur Print Cotton Floral Kin...	6325b6870c54cd47be6ebfbffa620ec7.jpg

Sommaire

1. Nettoyage des données
2. Text Processing
3. Image Processing
4. Concaténation des features Image & Texte
5. Modélisation & Evaluation
6. Conclusion

Nettoyage des données

Suppression des colonnes inutiles



0 valeurs dupliqué

Nettoyage des données

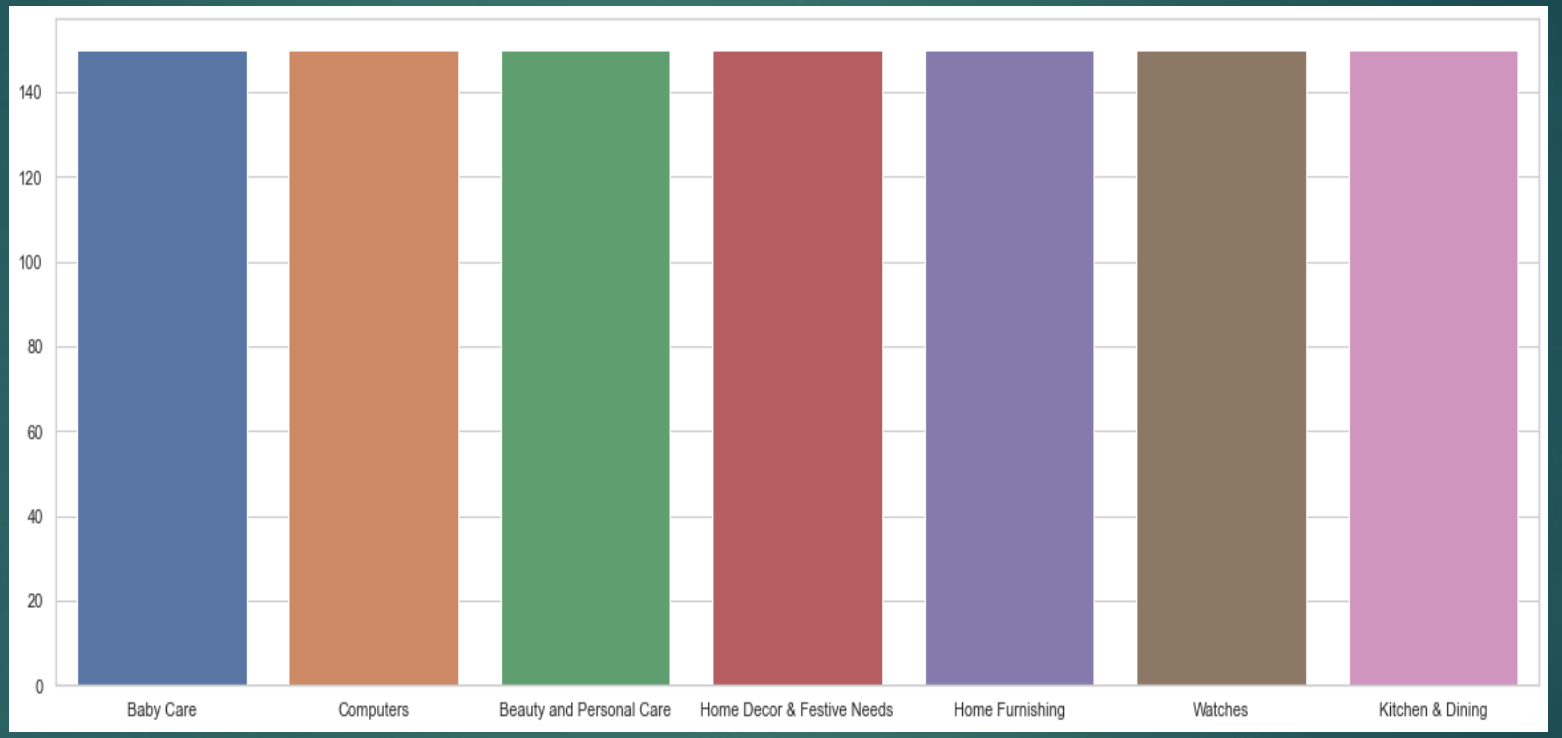
Formatage de la colonne catégorie

```
'["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Ey
elet Do..."]'
```

	cat_lvl_1	cat_lvl_2	cat_lvl_3
0	Home Furnishing	Curtains & Accessories	Curtains

Nettoyage des données

Formatage de la colonne catégorie



7 catégories avec 150 articles chacunes

Text Processing

- Pré-traitement de la colonne description (tokenisation, leminisation)
- Extraction des features en utilisant divers methodes :

  - Bag of Words (comptage simple et TF-IDF)
  - Word2Vec
  - Bert
  - USE

- Clustering pour évaluer les performances de chaque méthodes

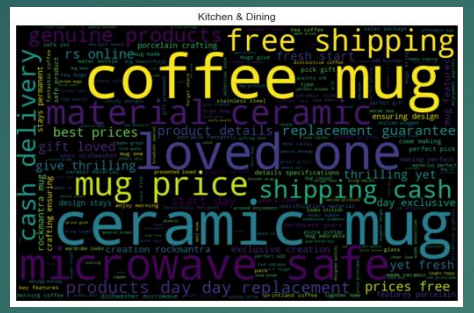
Text Processing

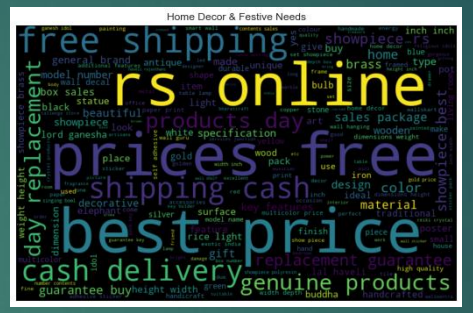
Tokenisation & leminisation

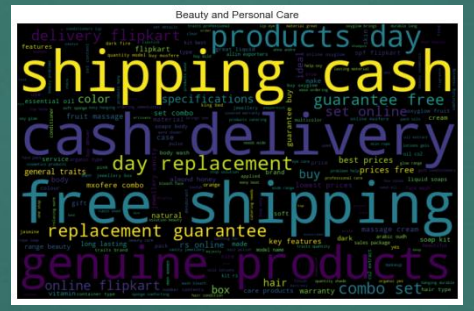
- Elimination des caractères non-alphabétiques
- Mise en minuscule
- Elimination des stopwords et des lettres isolées
- Lemmatisation

Text Processing

Analyse des mots clés par catégories







- Certaines catégories ont des mots-clés plus représentatifs que d'autres
- Mots-clés communs à certaines catégories

Text Processing

Extraction des features

- On va comparer 5 méthode d'extraction de features textuelles :
  - Comptage simple avec CountVectorizers
  - Comptage avec TF-IDF
  - Word2VEC (sentence embedding)
  - Bert
  - USE
- Clusterisation avec Kmeans
- Evaluation des modèles (ari, accuracy)
- Representation sous tsne

Text Processing

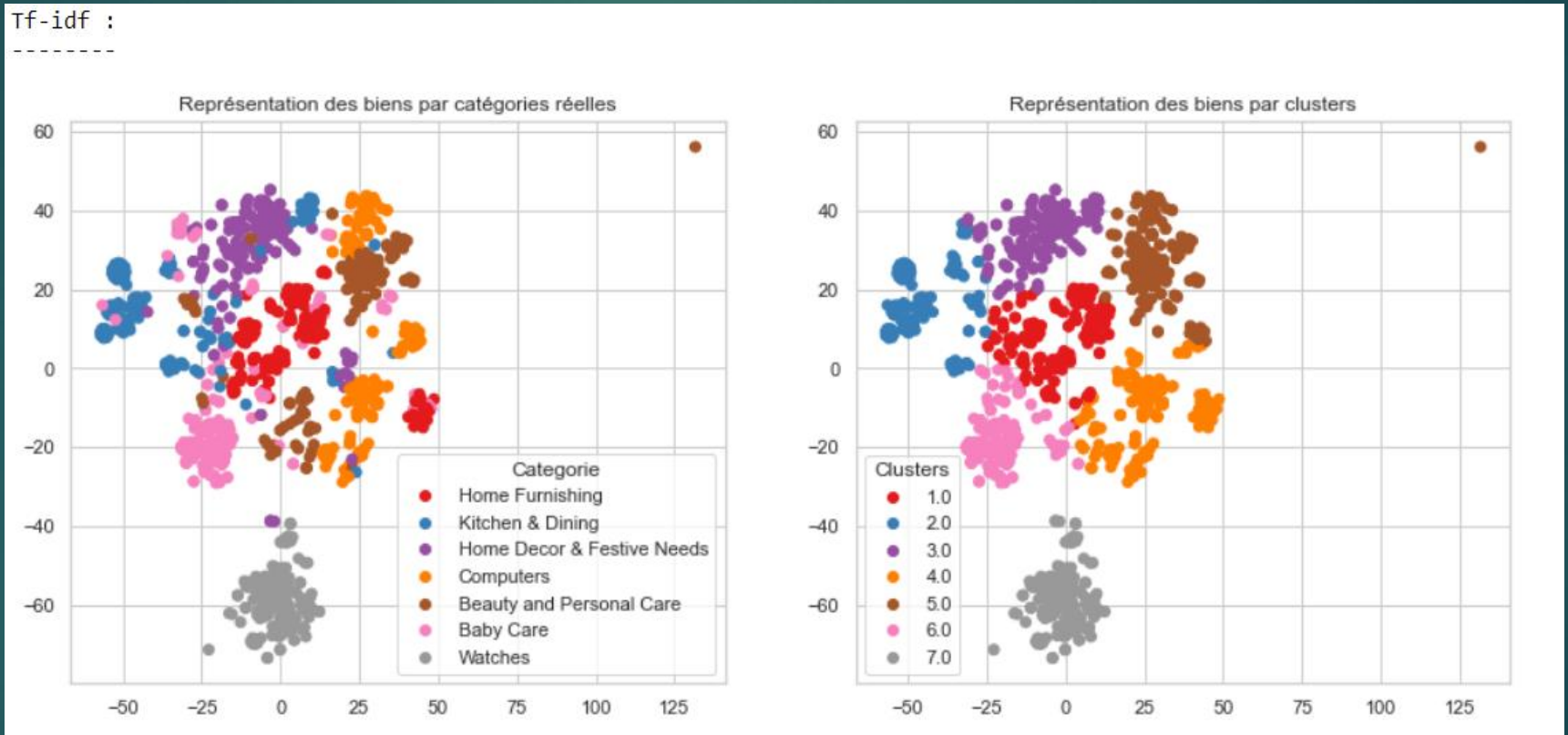
Extraction des features

	CV	TF-IDF	W2V	BERT	USE
accuracy	0.62	0.721905	0.497143	0.569524	0.613333
ari	0.367824	0.511265	0.277111	0.326547	0.422782

TF-IDF et USE nous donne de meilleurs performances
On choisi ces 2 méthodes pour l'extraction des features textes

Text Processing

Visualisation T-sne



Text Processing

Visualisation T-sne

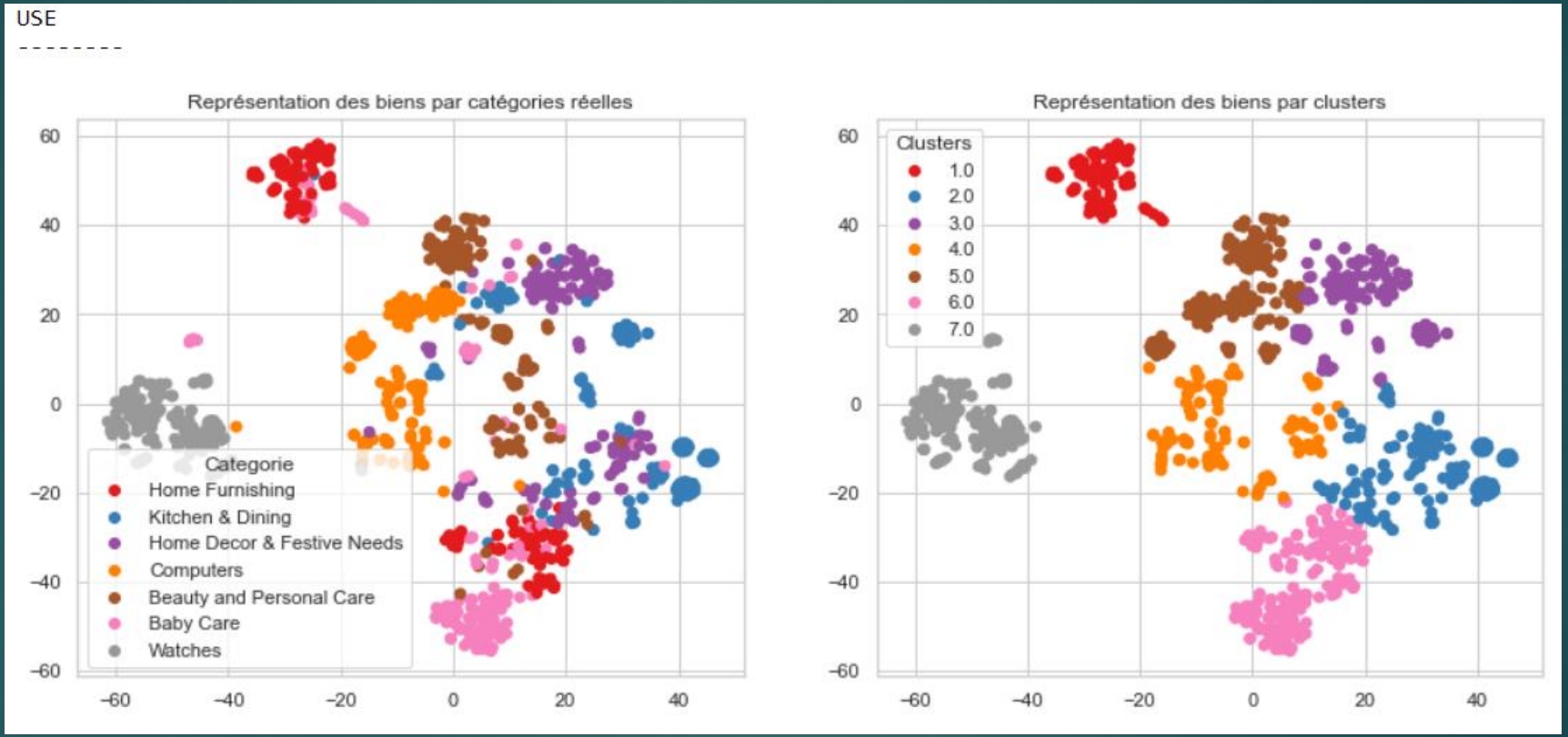


Image Processing

Extraction des features

- On va comparer 3 méthode d'extraction de features images :
  - SIFT
  - ORB
  - CNN
- Clusterisation avec Kmeans
- Evaluation des modèles (ari, accuracy)
- Representation sous tsne

Image Processing

Extraction des features

	CNN	SIFT	ORB
accuracy	0.5086	0.28092	0.233941
ari	0.344047	0.0645153	0.0296804

CNN donne des performances correct

Performances faibles pour ORB & SIFT

Image Processing

Visualisation T-sne

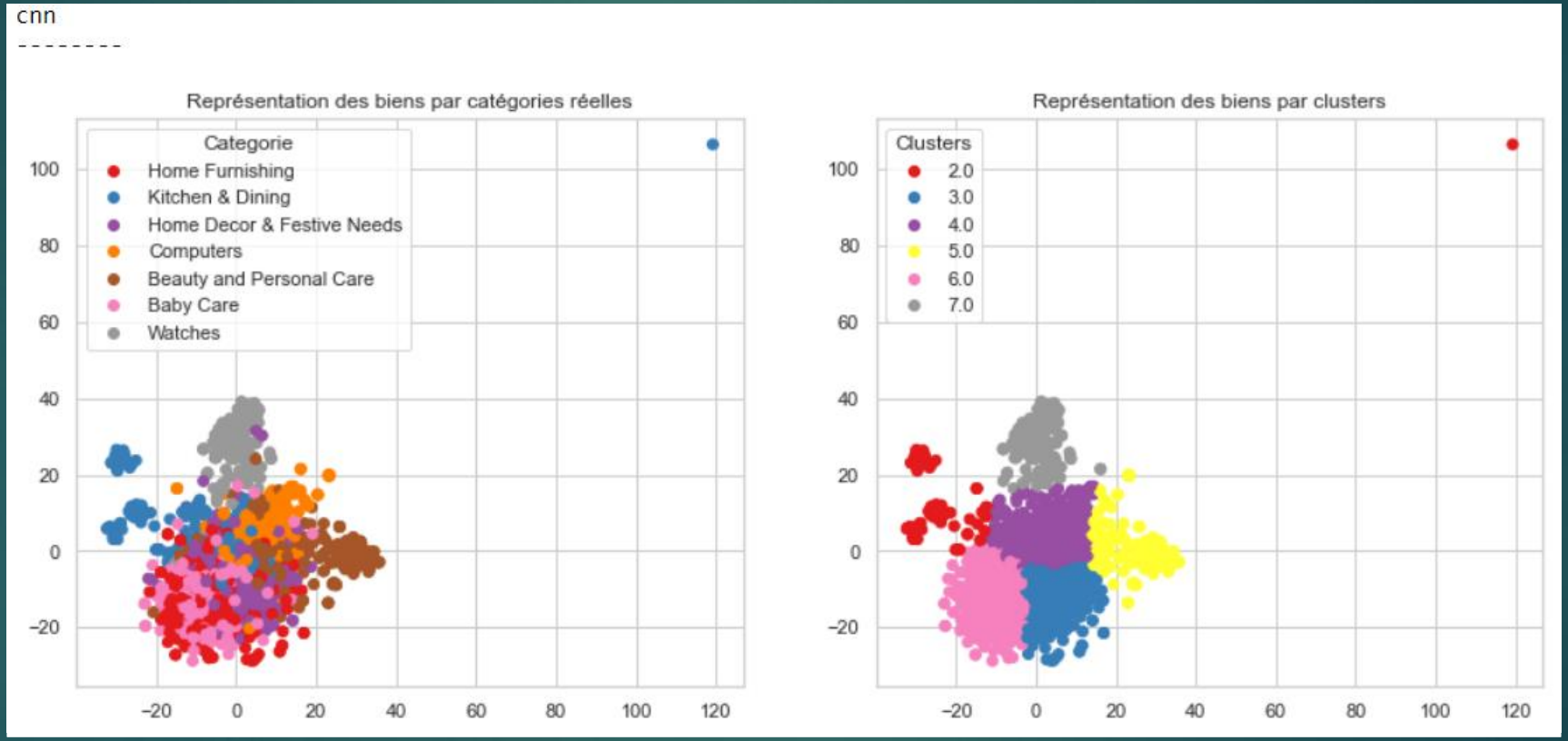
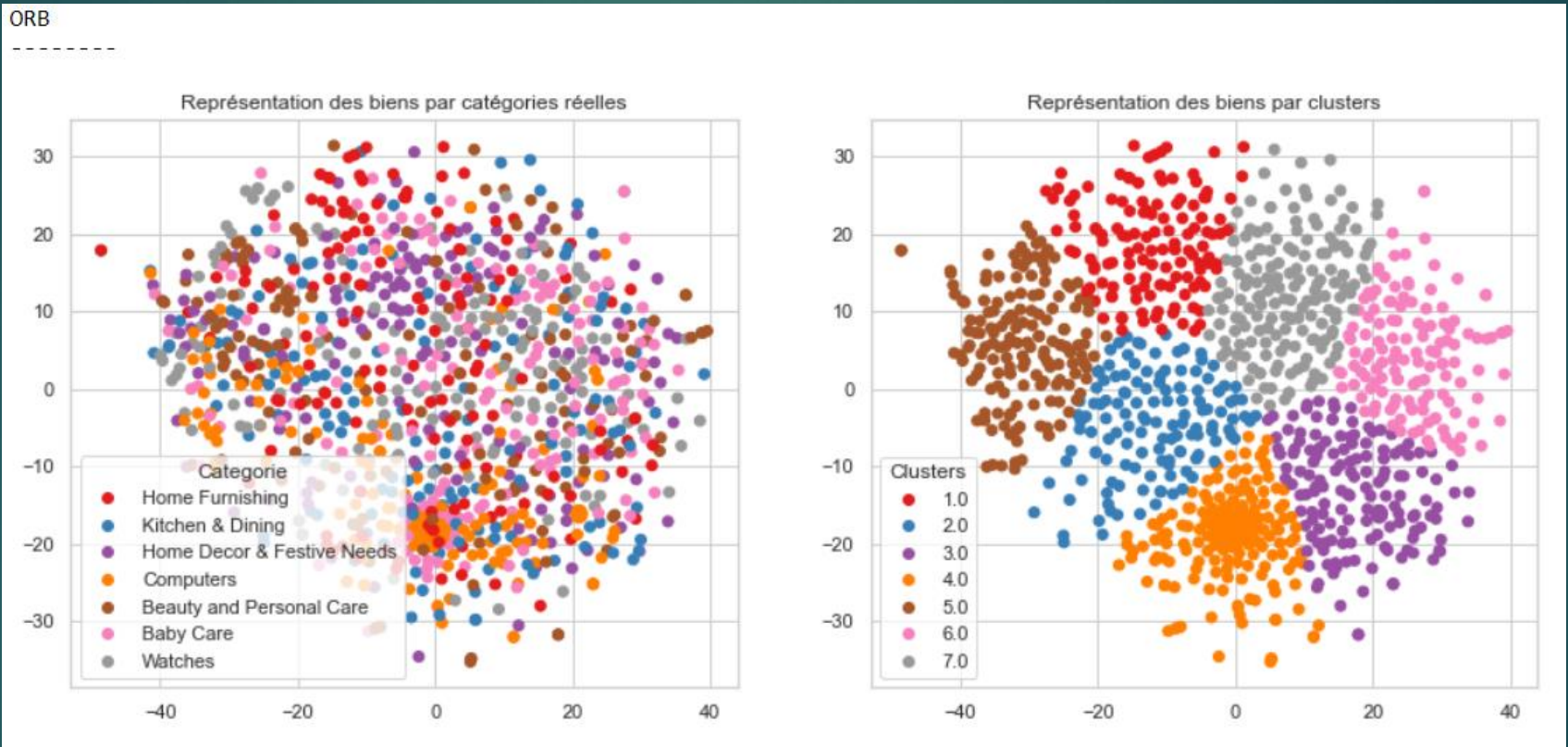


Image Processing

Visualisation T-sne

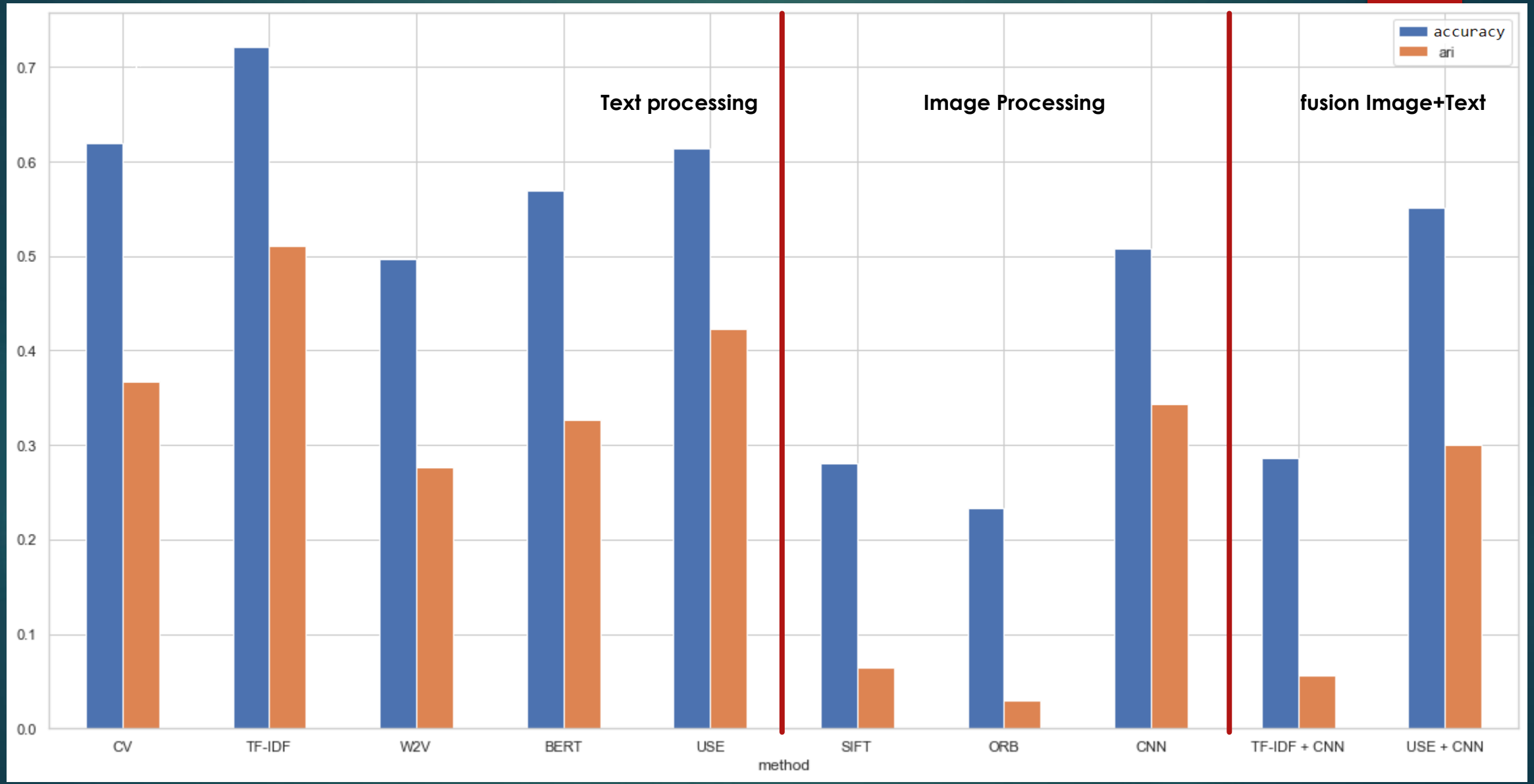


Fusion textes et images

On utilise l'approche de la concatenation des meilleurs features

- Pré traitement des features
- Clusterisation avec kmeans
- Comparaison des performances

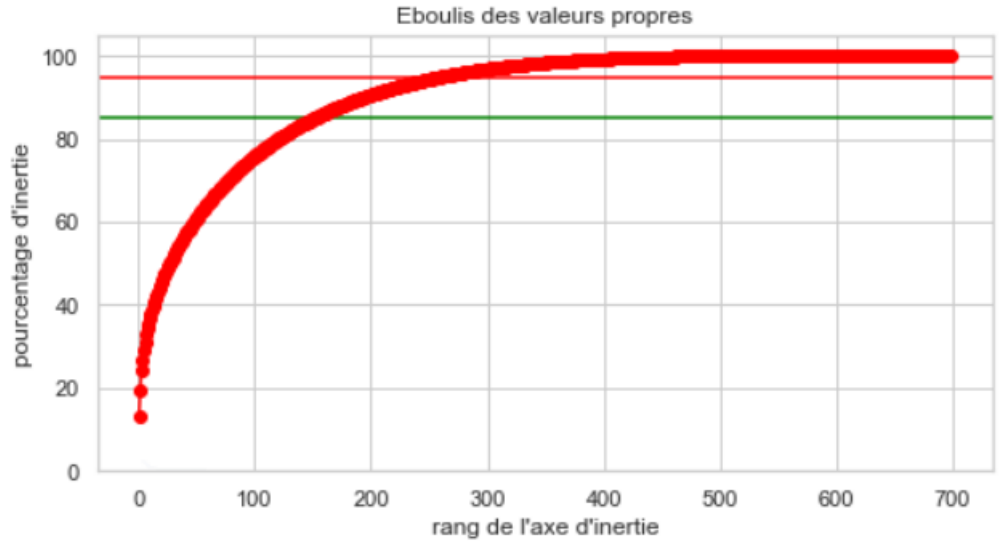
Fusion textes et images



Fusion textes et images

Apres ACP

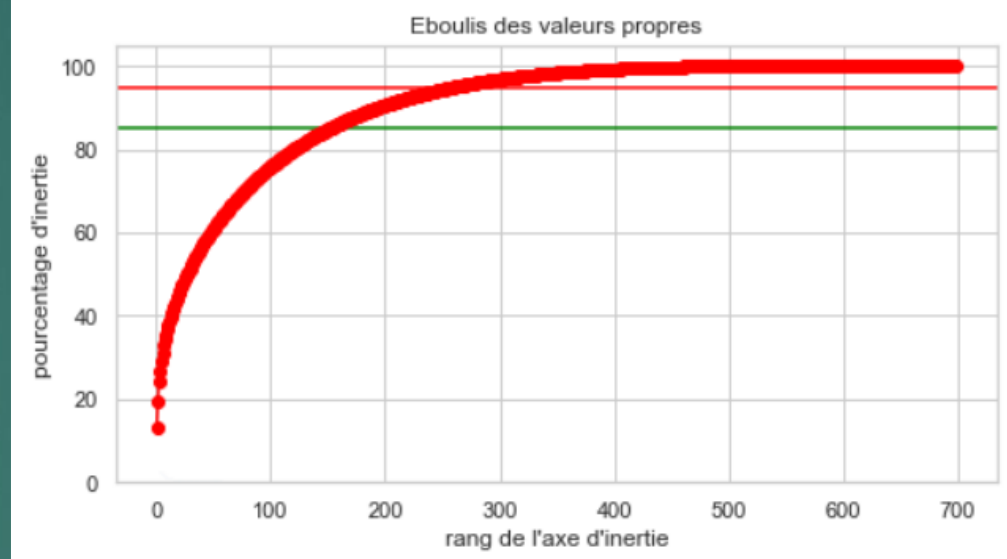
```
Association APRES ACP:  TF-IDF  &  CNN
AVANT ACP, nb de composante =  698
```




```
(698,)
(698, 260)
APRES ACP, nb de var =  260
accuracy =  0.4579710144927536  & ari =  0.19619400263424036
```

Avant Reduction de Dimension	accuracy	ari
TF-IDF + CNN	0.286957	0.0561738

```
Association APRES ACP:  USE  &  CNN
AVANT ACP, nb de composante =  698
```




```
(698,)
(698, 260)
APRES ACP, nb de var =  260
accuracy =  0.4579710144927536  & ari =  0.19619400263424036
```

Avant Reduction de Dimension	accuracy	ari
USE + CNN	0.550725	0.300869

Conclusion

- Text processing plus efficace que l'Image Processing

- Solution : Plus de données afin d'augmenter l'Image Processing

Ex : L'espace des phase visuelle pour une montre est plus vaste que l'espace des phases textuelle

- Pour la fusion texte et image, une autre technique peut etre utilisé : la pondération des prédiction