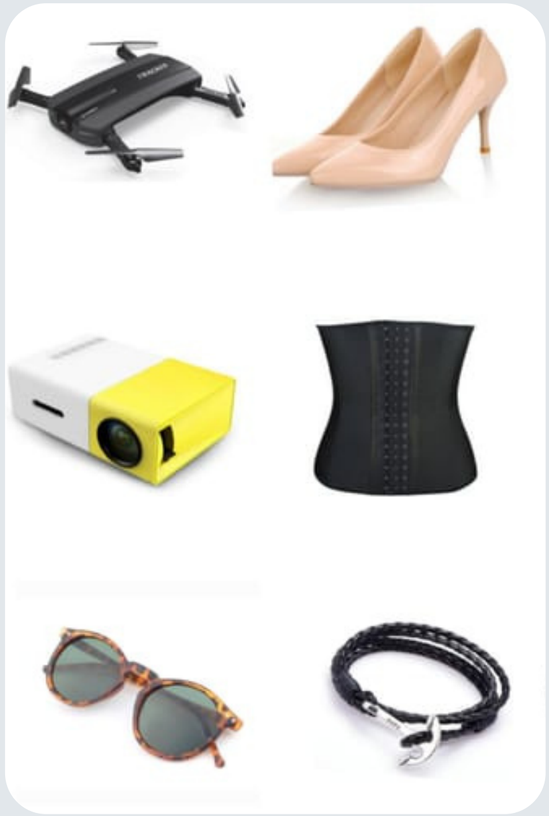
P6: Classifiez automatiquement des biens de consommation

FORMATION DATA SCIENTIST
OPENCLASSROOMS
20 JUILLET 2021 - 18 MAI 2022

Romain Vaillant
Février 2022



Classifiez automatiquement des biens de consommation

Sommaire

1. Contexte & Objectifs

Etudier la faisabilité d'un moteur de classification des articles en différentes catégories, avec un niveau de précision suffisant

Plateforme e-commerce

Etiquetage manuel de la catégorie des articles
Article: Photo & description

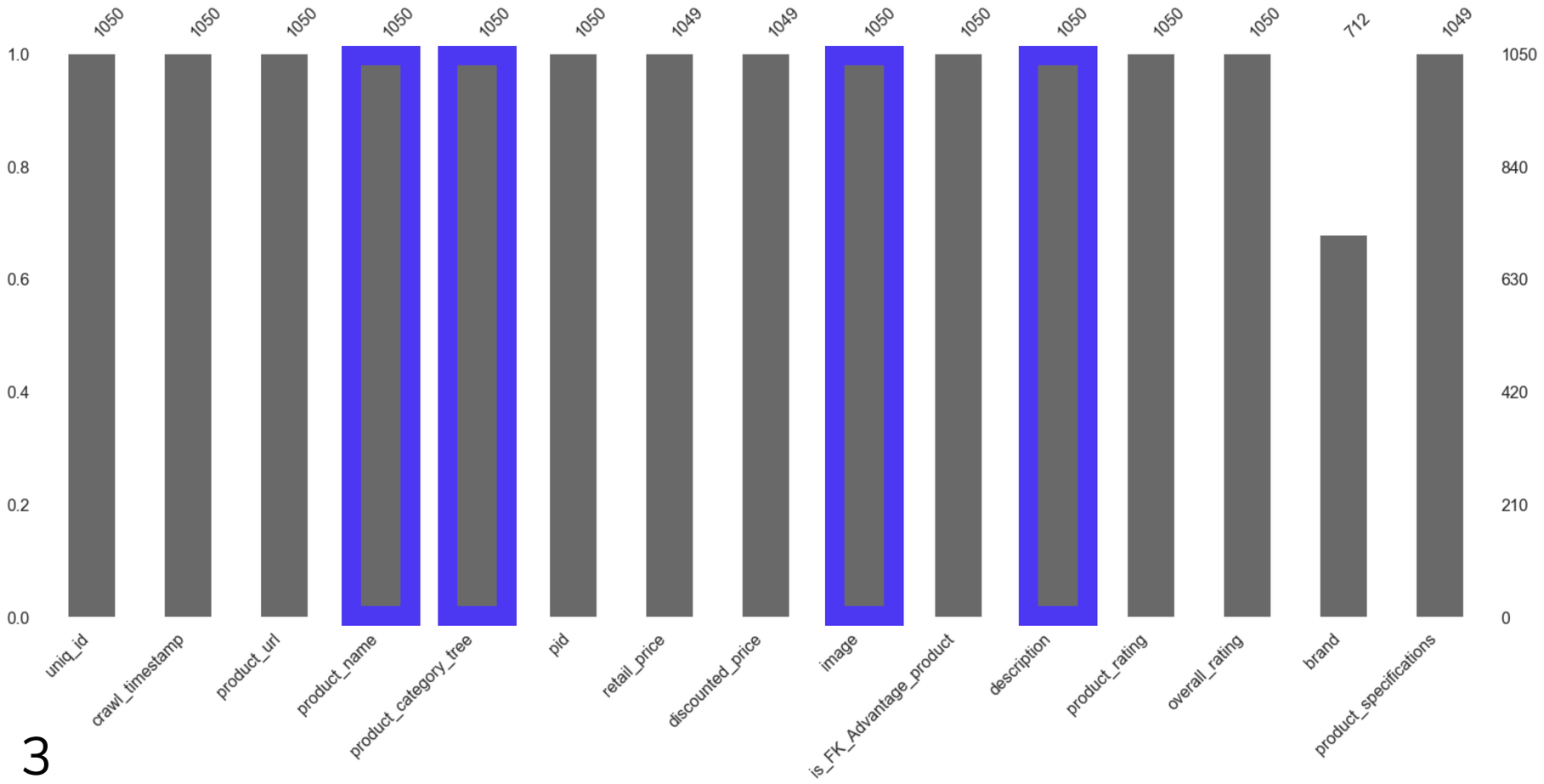
- Peu fiable
- Volume des articles est pour l’instant très petit

Faciliter l'expérience utilisateur des vendeurs et des acheteurs

- Mise en ligne de nouveaux articles (vendeur)
- Recherche de produits (acheteur)

Optique d'un passage à l'échelle: Automatisation nécessaire

2. Jeu de données



1050 lignes
15 colonnes

2. Jeu de données

a. Nettoyage du jeu de données

product_category_tree - (str)

["Home Furnishing >> Curtains & Accessories >> Curtains >> Elegance Polyester Multicolor Abstract Eyelet Do..."]

.split(">>")

product_category_tree - list (str)

[Home Furnishing, Curtains & Accessories, Curtains, Elegance Polyester Multicolor Abstract Eyelet Do...]

depth

product_category - (str)

Home Furnishing

Exploration des données textuelles

b. Exploration des données textuelles

Latent Dirichlet Allocation

Top 10 des mots clés par topic

7 catégories produit:

Home Furnishing, Baby Care, Watches, Home Decor & Festive Needs, Kitchen & Dining, Beauty and Personal Care , Computers

Topic #1	Topic #2	Topic #3	Topic #4	Topic #5	Topic #6	Topic #7
skin	baby	com	cm	mug	sticker	usb
laptop	cotton	flipkart	polyester	ceramic	vinyl	warranty
shape	girl	shipping	design	coffee	medium	adapter
pad	detail	genuine	eyelet	prithish	large	light
mouse	fabric	cash	comfort	perfect	wall	power
print	dress	delivery	curtain	rockmantra	paper	led
warranty	boy	buy	aroma	gift	small	laptop
hair	sleeve	free	height	one	store	flexible
set	neck	product	brown	safe	apply	charger
combo	pack	guarantee	model	loved	surface	portable
✗	✓	✗	✓	✓	✓	✓
-	Baby Care	-	Home Furnishing	Kitchen & Dining	Home Decor & Festive Needs	Computers

3. Approche de modélisation

Description

Pre-processing
- Mise en minuscules
- Suppression de la ponctuation
- Tokenization
- Suppression des stopwords
- Stemmisation & Lemmisation

Vectorisation
- Sklearn: TF-IDF
- Gensim: Doc2Vec()

PCA

Clustering
- ARI
- Silhouette
- Running Time

Image

Pre-processing
- Optimisation de l'exposition
- Optimisation du contraste
- Conversion de la couleur
- Bruit gaussien/Lissage
- Redimensionnement

Vectorisation
- OpenCV: SIFT
- Keras: VGG-16

PCA

Clustering
- ARI
- Silhouette
- Running Time

Description + Image

Pre-processing
- Normalisation (MinMaxScaler())
- Concaténation

Vectorisation
- Sklearn: TF-IDF
- OpenCV: SIFT

PCA

Clustering
- ARI
- Silhouette
- Running Time

a. Prétraitements

Texte

Nettoyage	Exemple	Fonction (NLTK)
Mise en minuscules	… (K)key (F)features of (E)elegance (P)polyester …	.lower()
Suppression de la ponctuation	… bath towel (3 bath towel, red, yellow, blue) …	.RegexpTokenizer(r'[A-Za-z]+')
Tokenization	… key | features | of | elegance | polyester …	.tokenize(text)
Suppression des stopwords	… trendy cap for only Rs. 900 online in india…	from list (179 words)
Stemmisation & Lemmisation	… key features elegance … (Stemmisation) … key features elegance … **(Lemmisation)**	.PorterStemmer() .WordNetLemmatizer()

a. Prétraitements

Image

Photographies

Départ

Contratse

Exposition

Bruit gaussien

Lissage

Redimensionnement

Optimisation du contraste
(PIL.ImageOps)
.autocontrast(img)

Optimisation de l'exposition
(PIL.ImageOps)
.equalize(img)

Conversion de la couleur
(OpenCv)
.cvtColor(img, color)

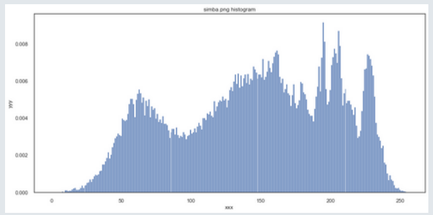
Bruit gaussien/Lissage
(PIL.ImageOps)
.fromarray(img+ GaussianNoise, 'RGB')
(PIL.ImageOps)
.filter(ImageFilter.BoxBlur(1))

Redimensionnement
(OpenCv)
.resize(img, (length, height))

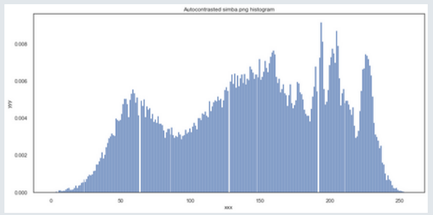
a. Prétraitements

Image

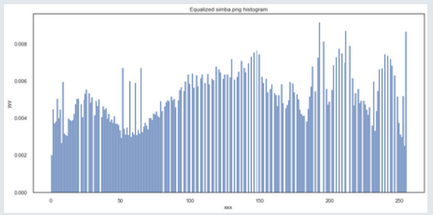
Histogrammes

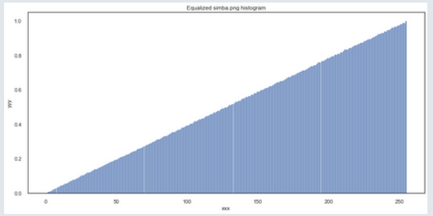


Départ

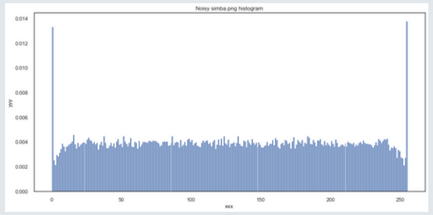


Contratse

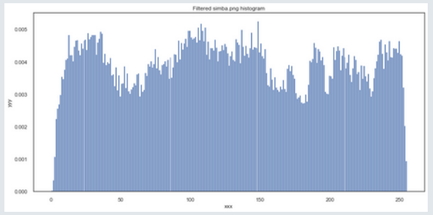




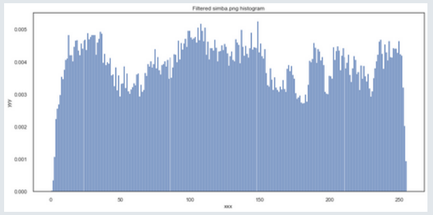
(cumulé)
Exposition



Bruit gaussien



Lissage



Redimensionnement

Optimisation du contraste
(PIL.ImageOps)
.autocontrast(img)

Optimisation de l'exposition
(PIL.ImageOps)
.equalize(img)

Conversion de la couleur
(OpenCv)
.cvtColor(img, color)

Bruit gaussien/Lissage
(PIL.ImageOps)
.fromarray(img+ GaussianNoise, 'RGB')
(PIL.ImageOps)
.filter(ImageFilter.BoxBlur(1))

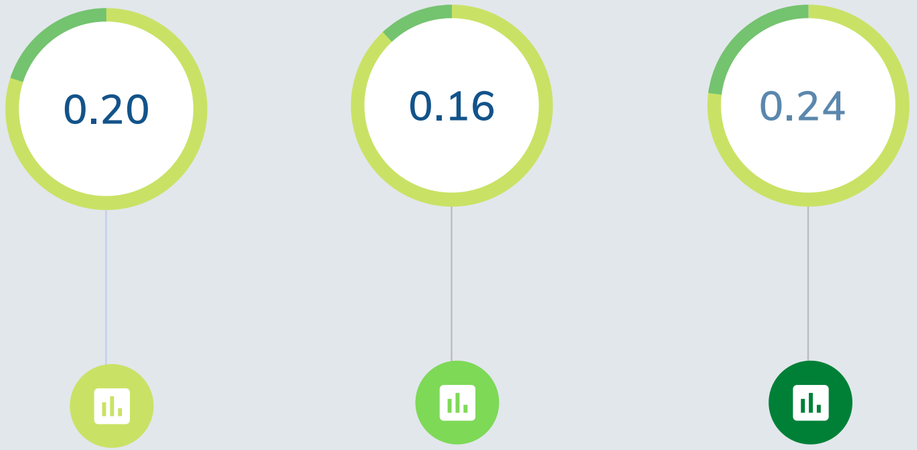
Redimensionnement
(OpenCv)
.resize(img, (length, height))

b. Clustering

Performance des cas d'utilisation

Silhouette

Résultats obtenus pour la meilleure configuration



b. Clustering

Performance des cas d'utilisation

Général

Résultats obtenus pour la meilleure configuration

	ARI	SILHOUETTE	RUNNING TIME
Description (TFIDF PCA)	0.27	0.20	0.30
Image (SIFT PCA)	0.00	0.16	0.09
Description (TFIDF PCA) + Image (SIFT PCA)	0.14	0.24	0.16

b. Clustering

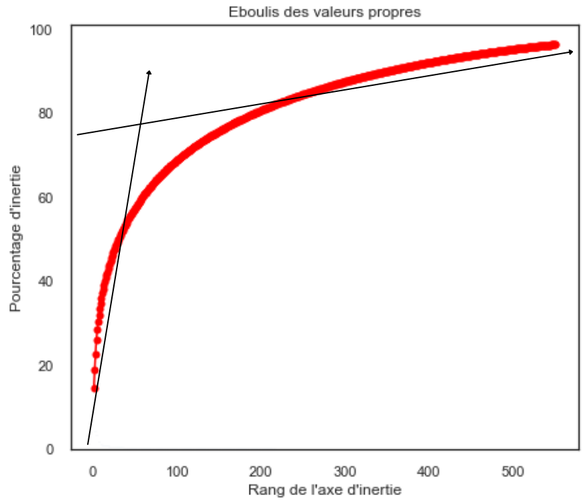
Texte clustering

TFIDF PCA

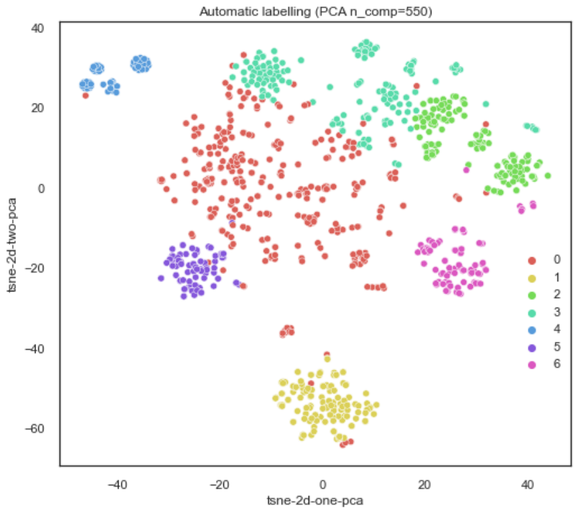
PCA

550 composantes

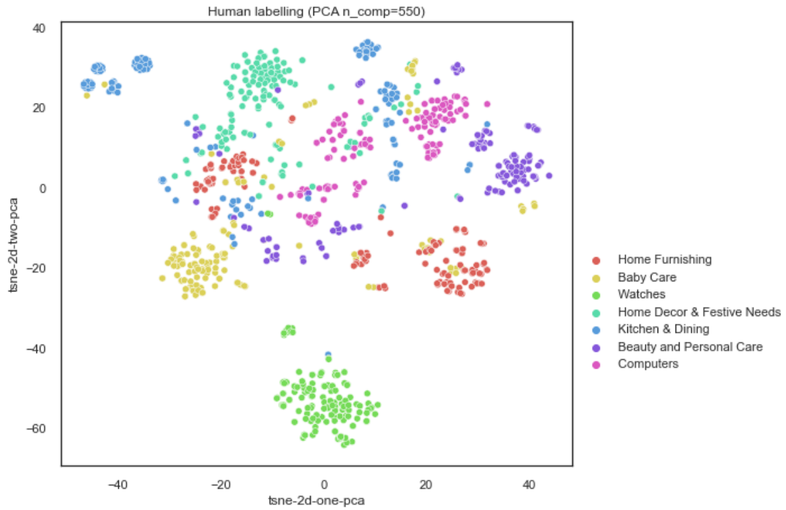
Variance expliquée > 95%



Visualisation 2D - TSNE



Attribution automatique



Attribution manuelle

b. Clustering

Image clustering

VGG-16

Transfer Learning

Suppression des couches **fully connected** et **softmax (clustering)**

Ajout d'une couche de **prédiction(dense)** à 7 **classes (classification)**

ARI

0.03

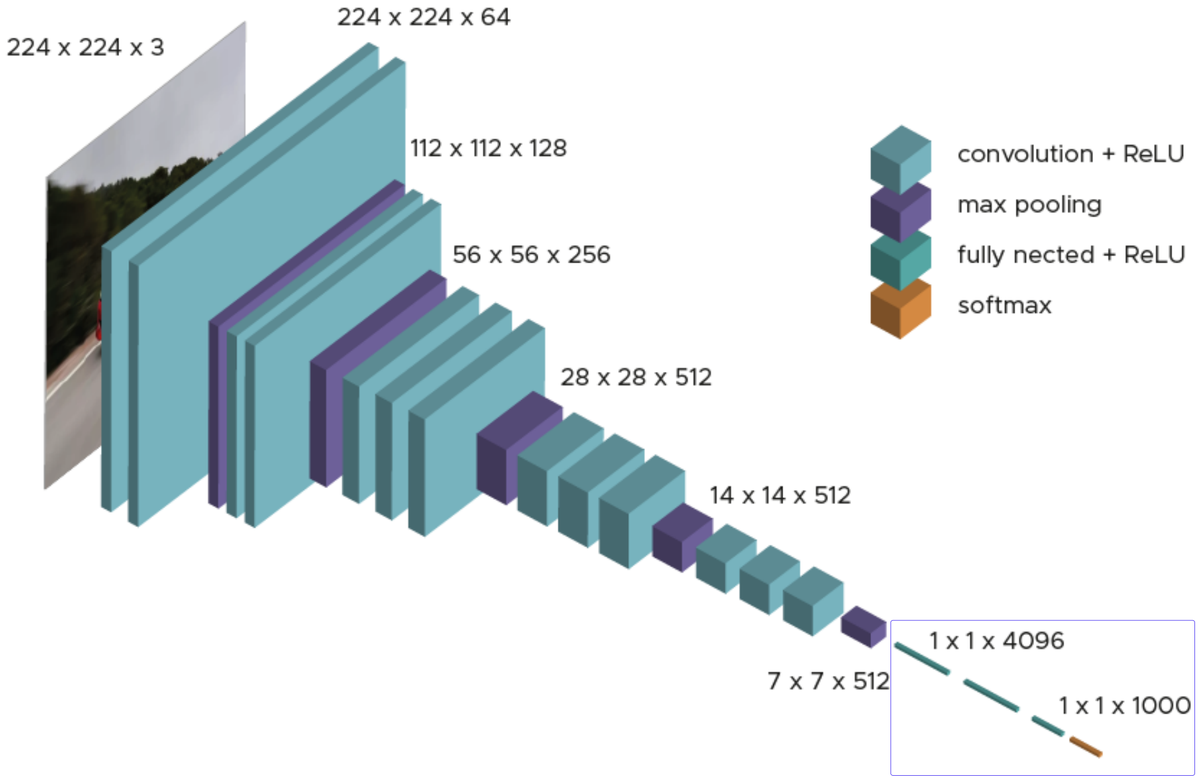
Silhouette

0.10

Running time (secondes)

0.40

Axe d'amélioration:
Tuning nécessaire (Hyperopt, Gridsearch...)



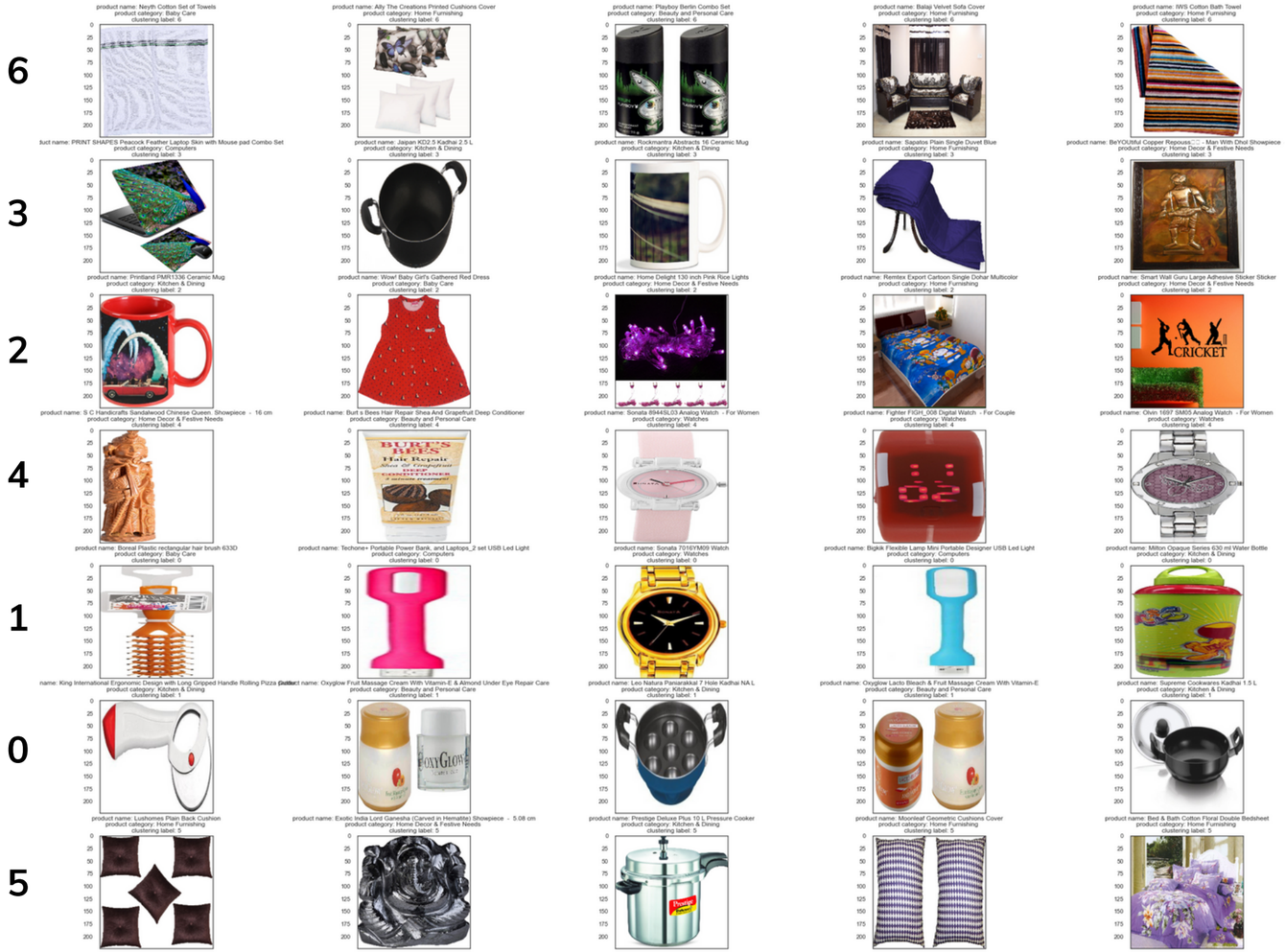
Représentation 3D de l'architecture de VGG-16

b. Clustering

Image clustering

Matrice de confusion

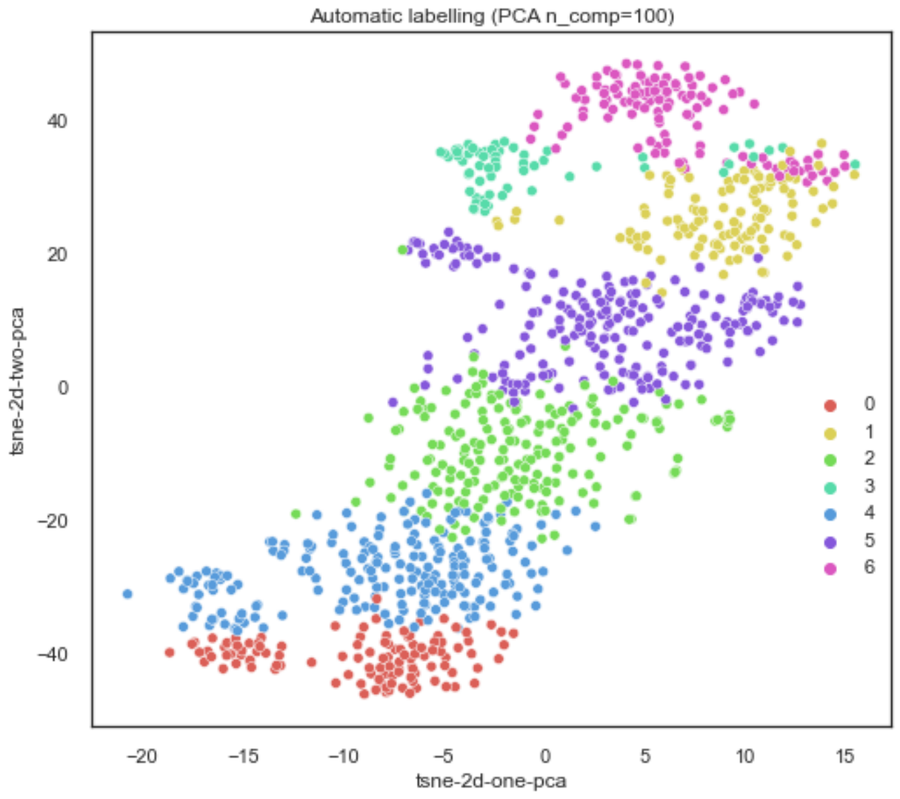
Echantillons d'images: VGG-16



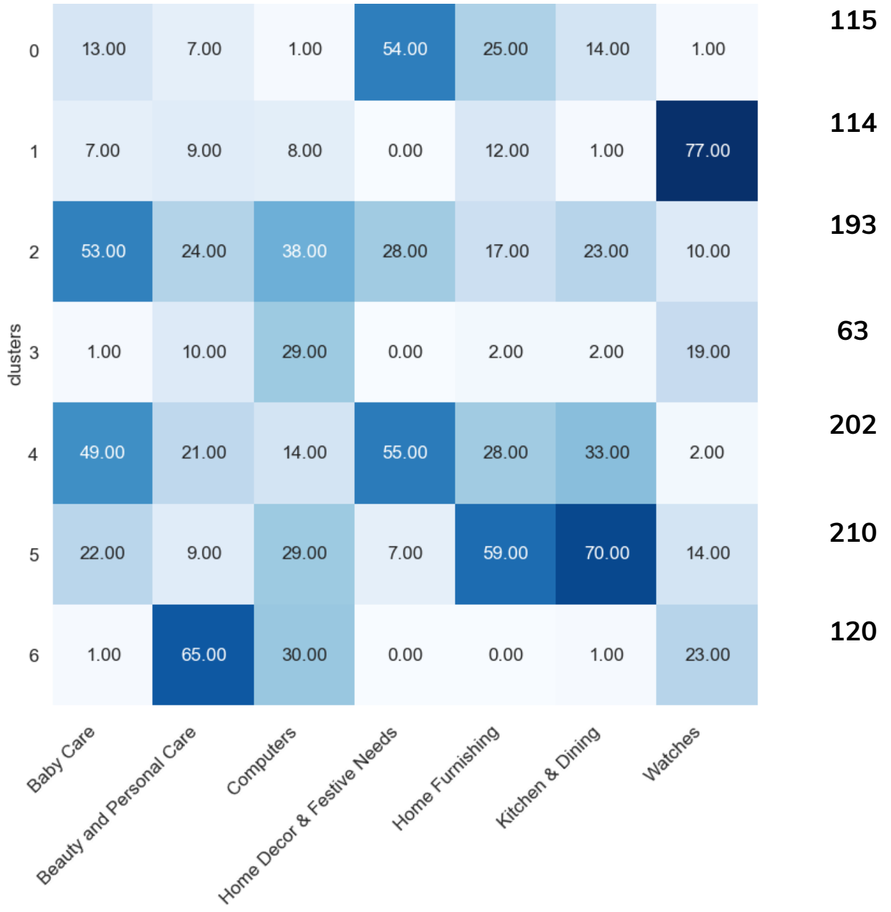
b. Clustering

Texte & Image clustering

Graphique 2D



Matrice de confusion



4. Conclusion

L'étude de faisabilité du moteur de classification donne des **résultats encourageants**

Combiner **texte et image** donne de **meilleurs résultats**:
Silhouette de **0.24** avec **1050** échantillons

La **réduction de dimension** a permis d'**améliorer la silhouette** et la **scalabilité** sur les données textuelles

Le **SIFT** a donné de **meilleurs résultats** que le VGG-16

Classification des images avec VGG-16, résultats médiocres

Axes d'amélioration:
Augmenter le volume de données
Tuning des hyperparamètres

Merci de votre écoute

Séance de questions