

PROJET 6 : CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

LAURA DAINES

MENTOR:
BENJAMIN TARDY

Rappel de la problématique & présentation du jeu de données

Explication des prétraitements & des résultats du clustering

Conclusion & recommandations pour la création éventuelle du moteur de classification

Questions - Réponses

SOMMAIRE

RAPPEL DE LA PROBLÉMATIQUE ET PRÉSENTATION DU JEU DE DONNÉES

RAPPEL DE LA PROBLÉMATIQUE

Contexte :

- Lancement d'une marketplace e-commerce par l'entreprise "Place de marché".
- Classification manuelle des articles par les vendeurs.

Mission:

- Étude de faisabilité : automatiser l'attribution de la catégorie des articles.

PRÉSENTATION DU JEU DE DONNÉES

- 1050 lignes (articles) et 15 colonnes. Par article: photo et informations (texte) : nom, description, prix, spécifications.
- 7 catégories, chacune contenant 150 articles:
 - Baby Care
 - Watches
 - Beauty and Personal Care
 - Home Decor Festive Needs, Home Furnishing
 - Computers
 - Kitchen Dining

EXPLICATION DES PRÉTRAITEMENTS ET DES RÉSULTATS DU CLUSTERING

FEATURE ENGINEERING

Création de variables:

- description + product name + product specifications
 - description + product specifications
 - product name + product specifications

TEXTE: BAG OF WORDS + UMAP + KMEANS

- description + product name + product specifications
- Features : preprocessing, bag of words
- Réduction de dimensions: UMAP
- Clustering: K-means – 7 clusters

EXPLICATION DES PRÉTRAITEMENTS : TEXTE

Nettoyage et transformation du texte:

- Conversion en minuscules
- Suppression des chiffres
 - Tokenisation
- Stem / Lemmatisation des mots
- Suppression des stopwords et de la ponctuation

Explication des prétraitements: Texte

	count	percentage
buy	1	50.0
cat	2	100.0
coffe	1	50.0
drank	1	50.0
go	1	50.0
green	1	50.0
pink	2	100.0
stuck	1	50.0
tabl	1	50.0
turn	1	50.0
underneath	1	50.0

Deux phrases d'exemple pour illustrer et tester la bonne réalisation des prétraitements:

- "My green cat turned pink after it drank coffee!? What should I do with my cat? - (/)"
- "I am not going to buy the pink table. But my cat is stuck underneath it!! § :,,"

	buy	cat	coffe	drank	go	green	pink	stuck	tabl	turn	underneath
0	0	0	2	1	1	0	1	1	0	0	0
1	1	1	1	0	0	1	0	1	1	1	1

TEXTE: BAG OF WORDS + UMAP + KMEANS

Gridsearch pour maximiser le ARI (catégories réelles vs étiquettes du Kmeans) sur :

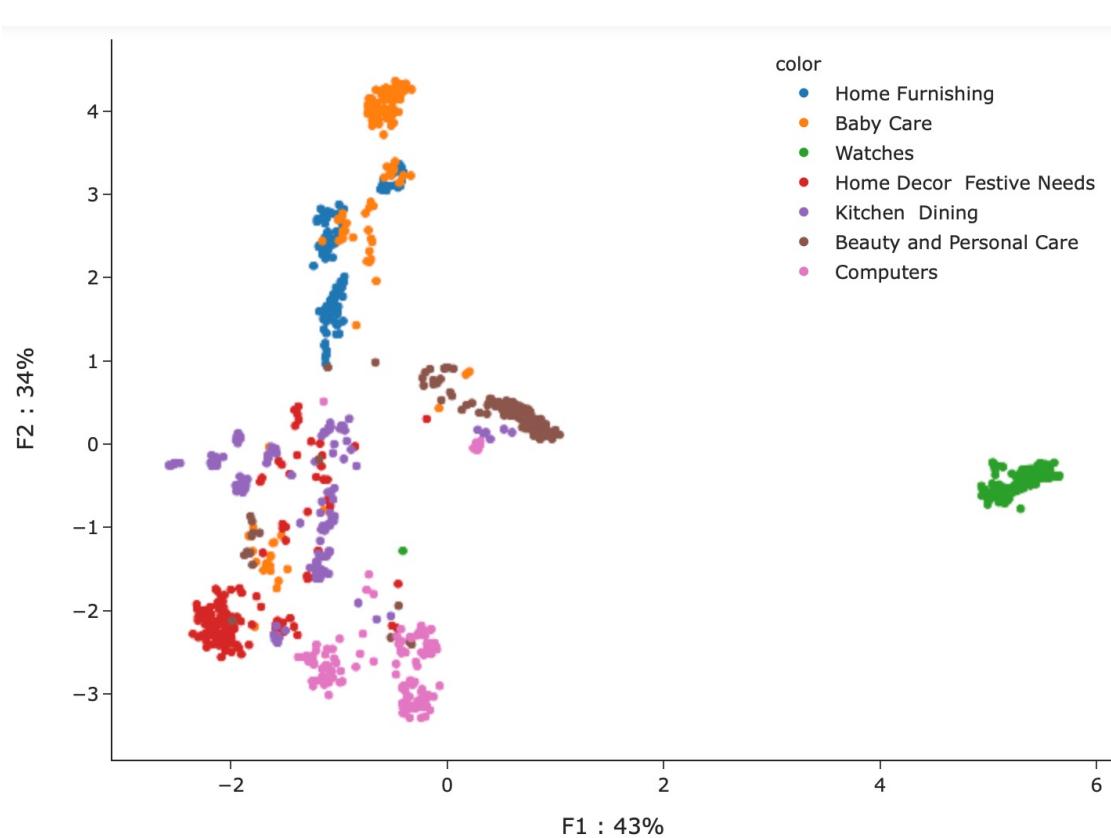
- notre fonction de preprocessing du texte
- TfidfVectorizer()
- UMAP()

Output du gridsearch :

- Stemmer
- Stopwords utilisés : mots présents dans plus de 30% et moins de 1% des individus
- Pas d'utilisation du TfidfVectorizer()
- UMAP : min_dist 0.1, n_components 25, n_neighbors 100

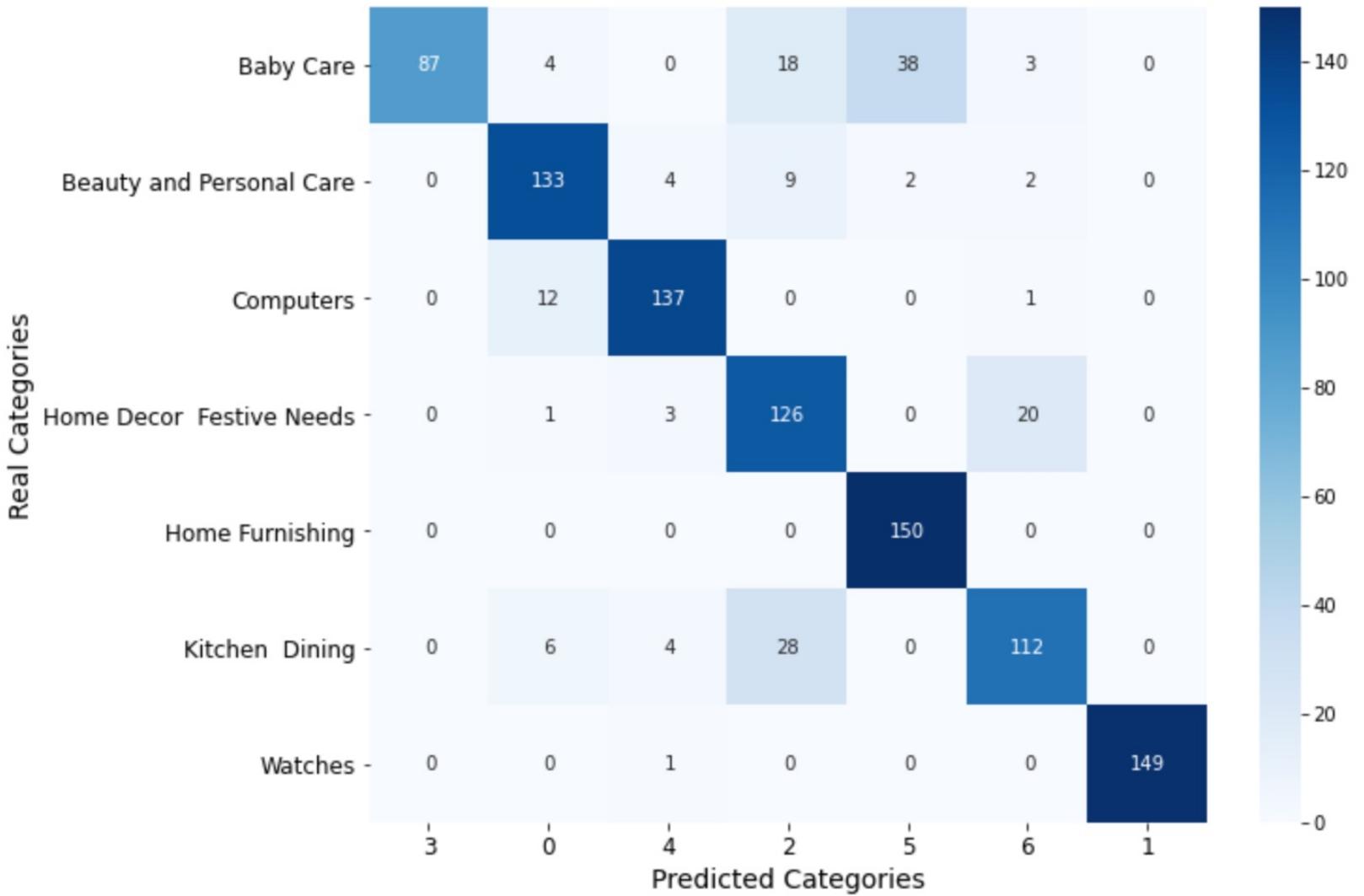
TEXTE : BAG OF WORDS + UMAP + KMEANS

ACP & T-SNE des données après BoW et UMAP



BAG OF WORDS + UMAP + KMEANS

■ Adjusted Rand Score: 0.704



TEXTE : USE + UMAP + KMEANS

- description + product name + product specifications
- Features : Extraction avec réseau de neurones
- Réduction dimension: UMAP
- Clustering : K-means sur matrice de données après réduction

Output du gridsearch :

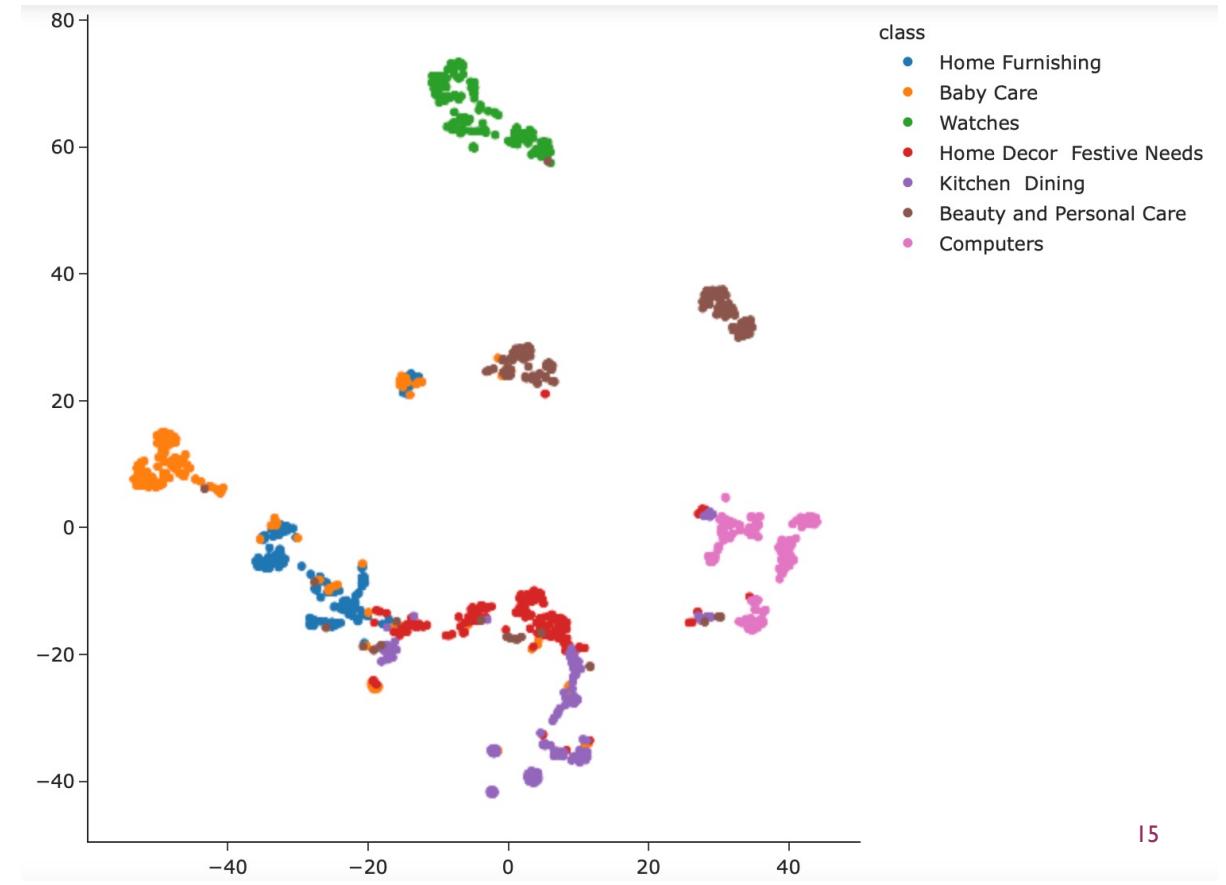
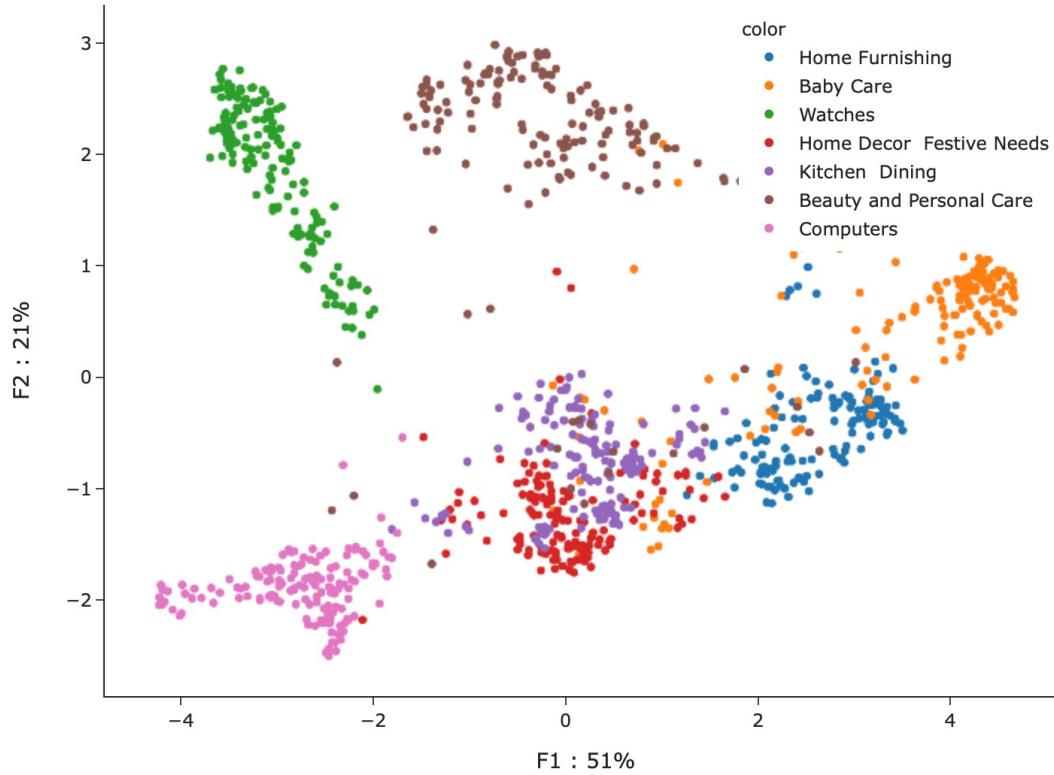
- UMAP : min_dist 0.5, n_components 50, n_neighbors 50

TEXTE: USE (GOOGLE UNIVERSAL SENTENCE ENCODER)

- Universal Sentence Encoder est obtenu sur le site TF Hub: <https://www.tensorflow.org/hub/modules/google/universal-sentence-encoder-large/>
- Le modèle est pré-entraîné avec un DAN (Deep Averaging Network) encoder. Modèle optimisé pour les textes d'une longueur supérieure à un mot, tels que des phrases, des expressions ou des courts paragraphes.
- Embeddings pour des phrases (vs pour des mots, comme avec Word2Vec par exemple)
- Output: vecteur de 512 dimensions.

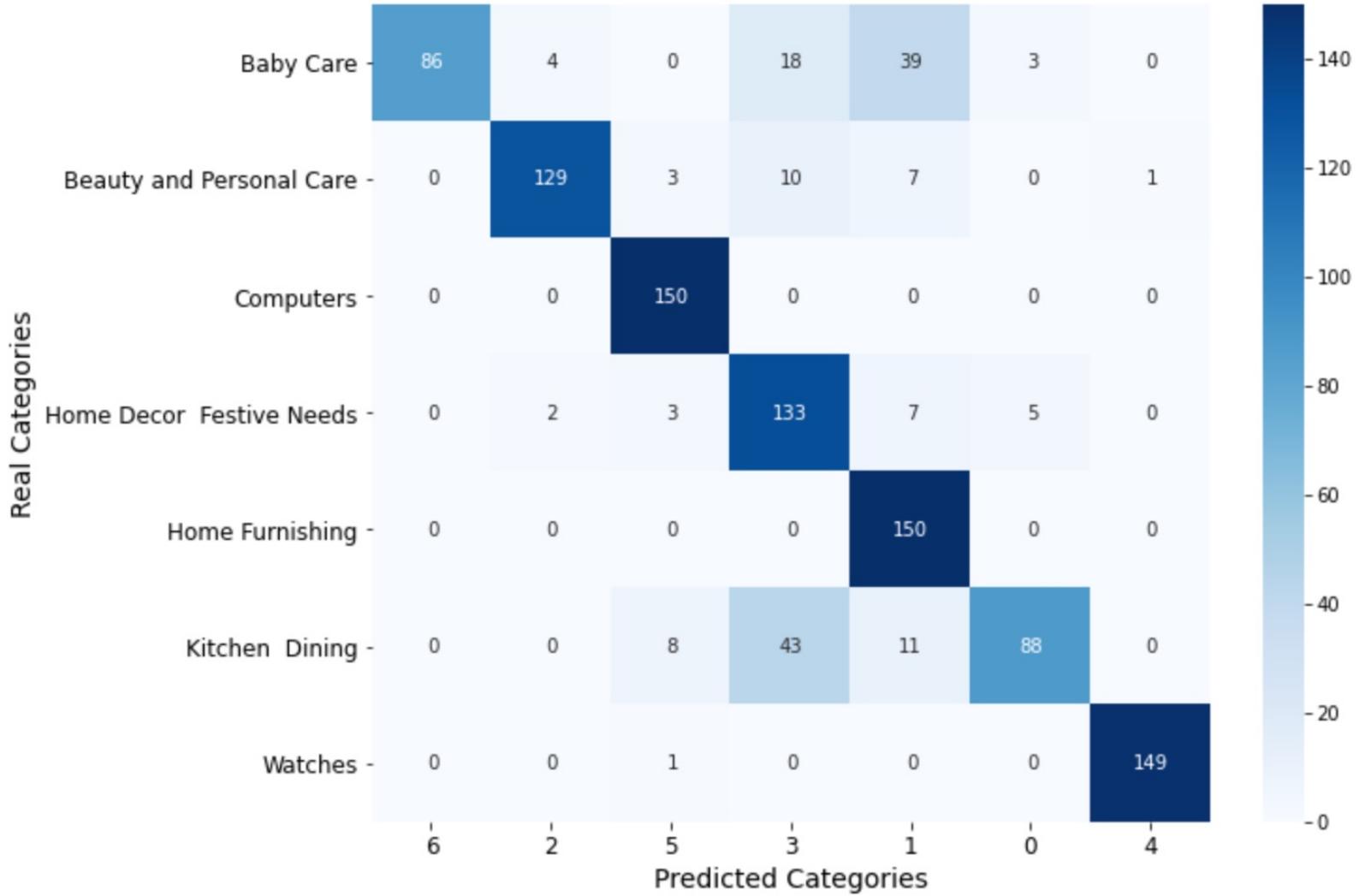
TEXTE : USE + UMAP + KMEANS

ACP / T-SNE des données après USE et UMAP



USE (GOOGLE UNIVERSAL SENTENCE ENCODER) + UMAP + KMEANS

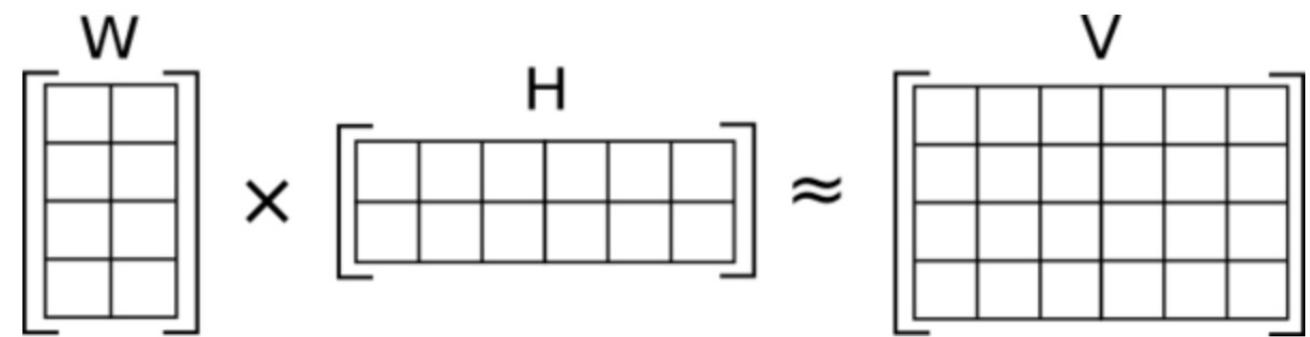
■ Adjusted Rand Score: 0.68



TEXTE:

NMF (NON NEGATIVE MATRIX FACTORISATION)

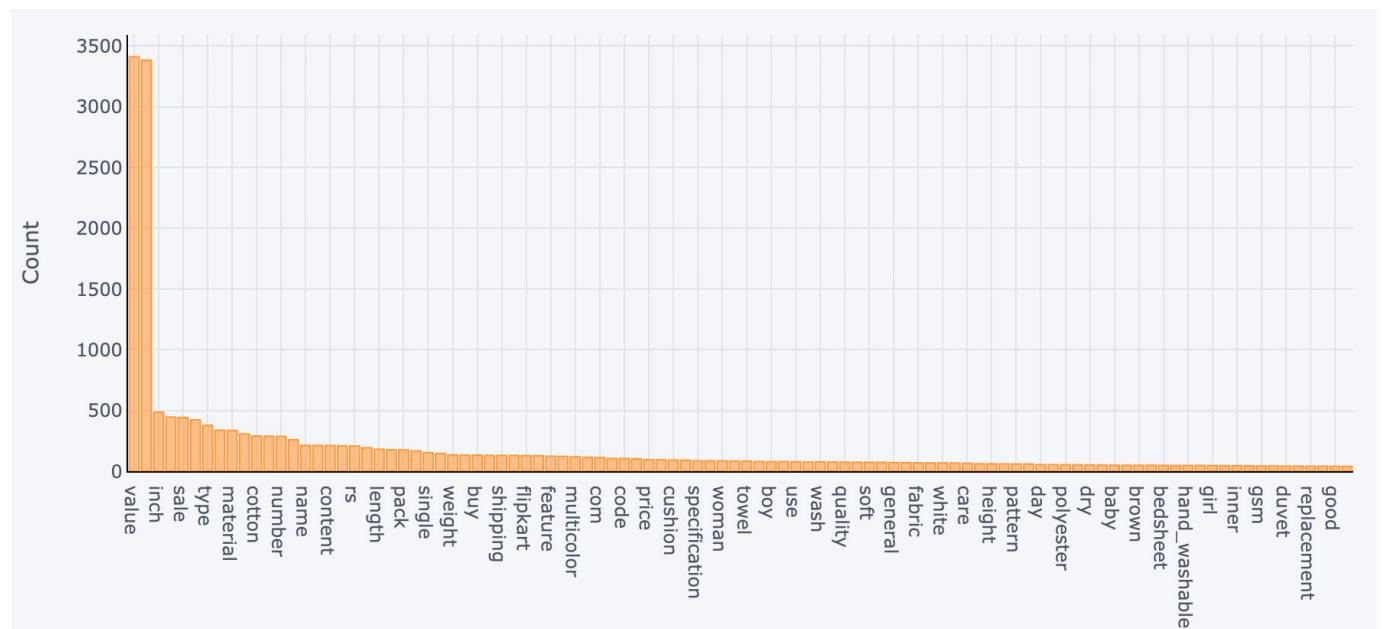
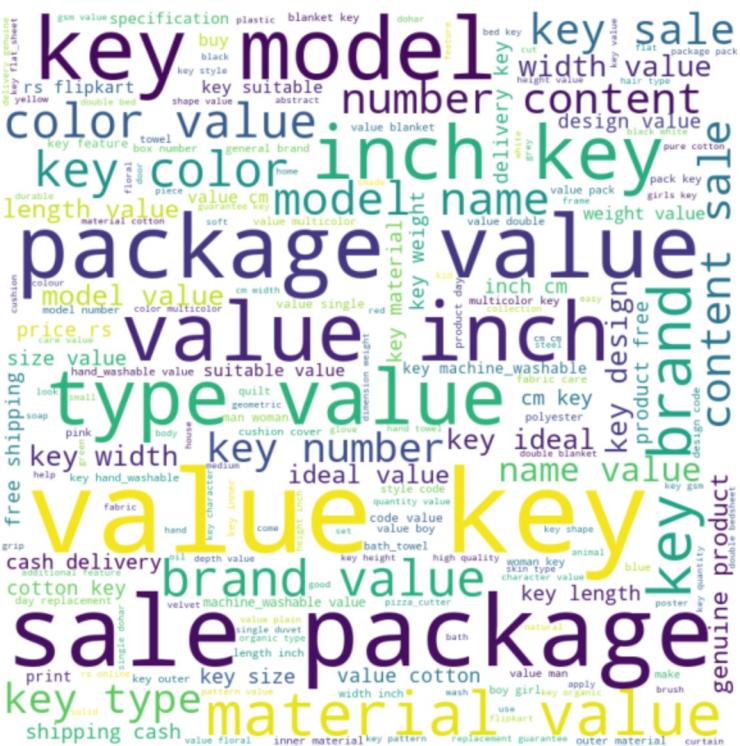
- Factorisation de matrice non négative:

$$\begin{matrix} W \\ \times \\ H \end{matrix} \approx \begin{matrix} V \end{matrix}$$


Daniel D. Lee, H. Sebastian Seung, Wikidata ID Q10843505

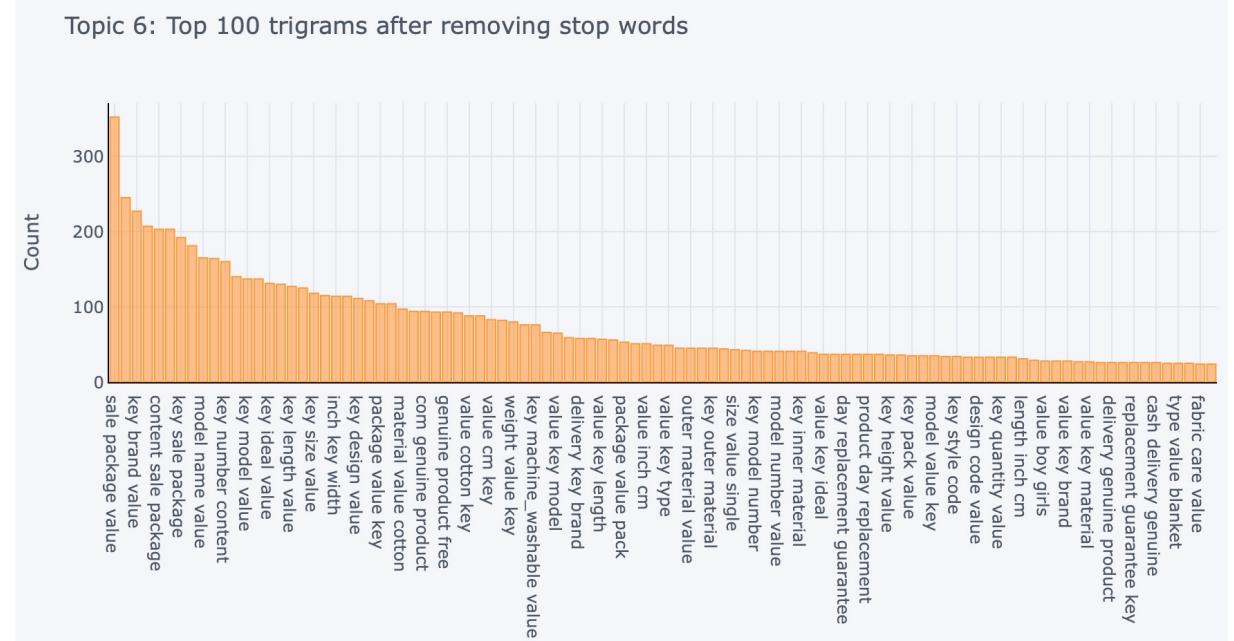
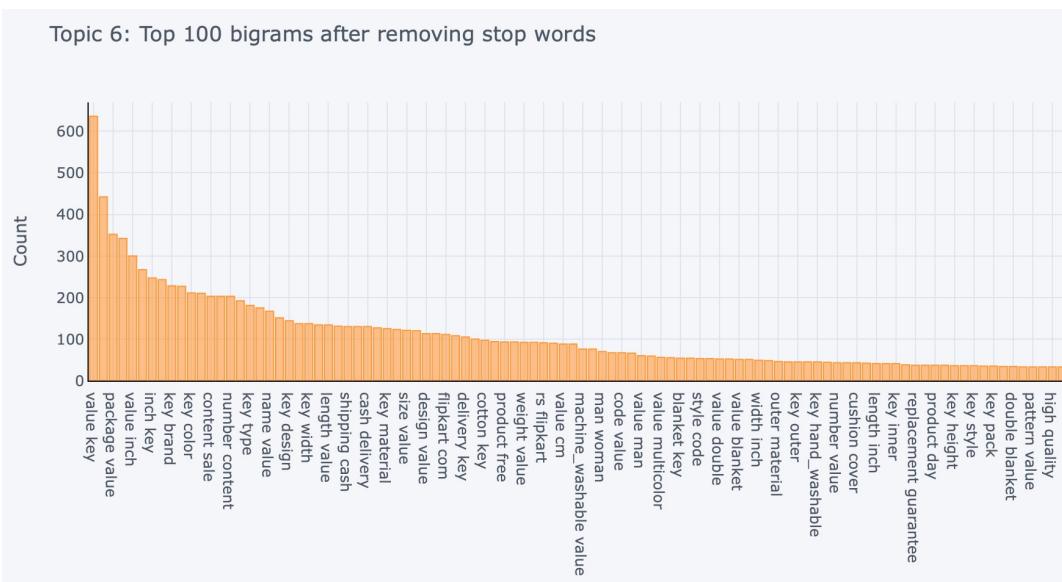
TEXTE : WORD CLOUD ET FRÉQUENCE DES MOTS

Classe estimée par NMF : Baby Care / Home Furnishing



TEXTE : FRÉQUENCE DES BIGRAMS ET TRIGRAMS

Classe estimée par NMF : Baby Care /
Home Furnishing



TEXTE : COMPARAISON DES MODÈLES

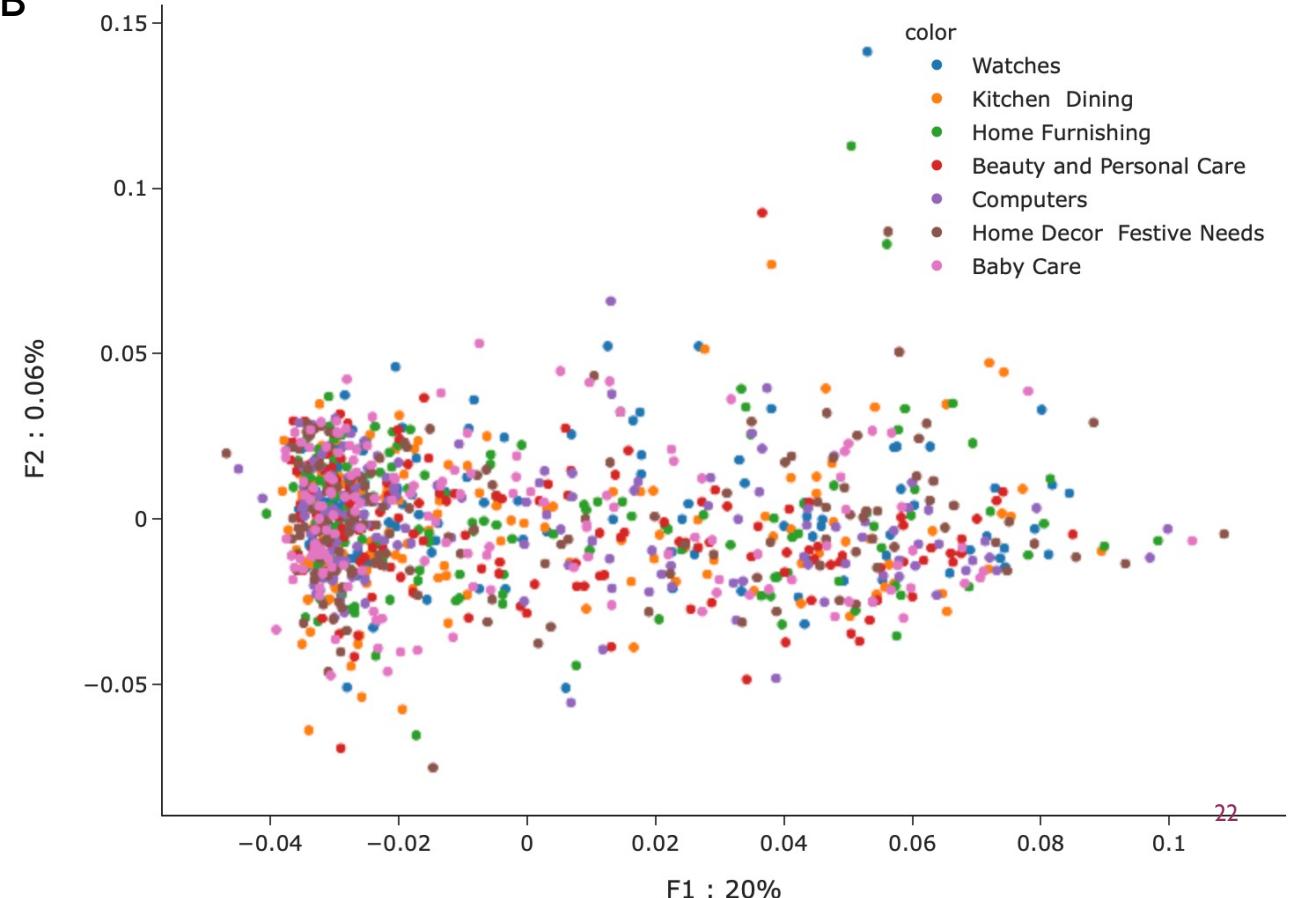
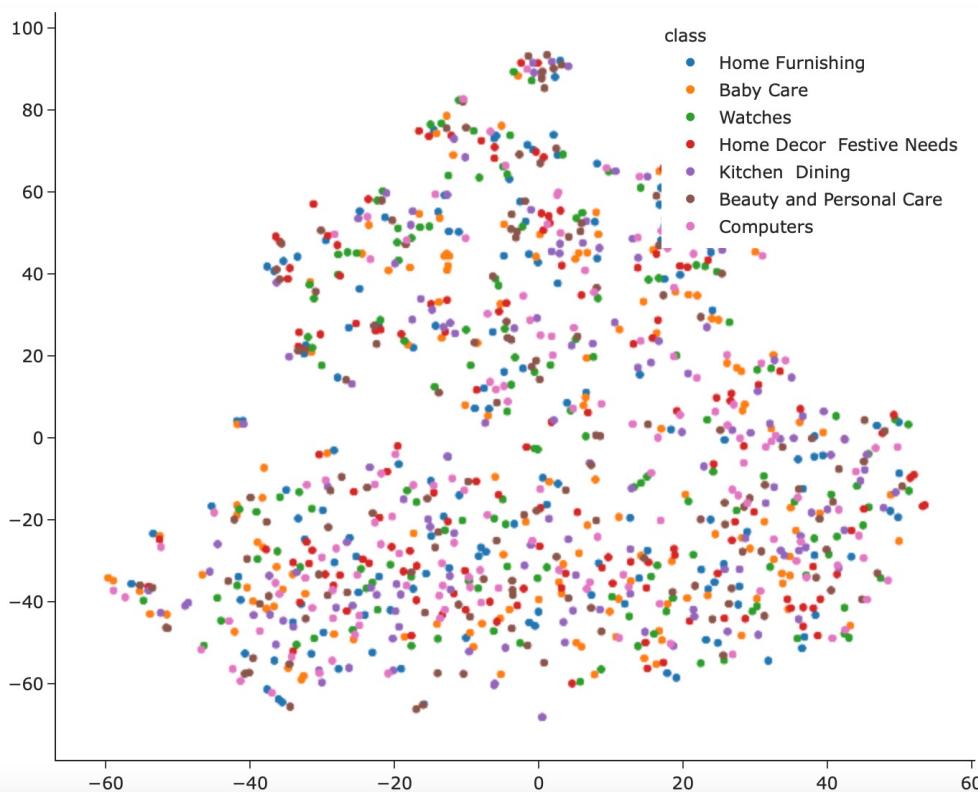
Text: Methods Used	ARI
BoW / Truncated SVD / Kmeans	0.61
BoW / UMAP / Kmeans	0.7
USE / PCA / Kmeans	0.68
USE / UMAP / Kmeans	0.685
LDA	0.38
NMF	0.54

IMAGES : DESCRIPTEURS ORB (ORIENTED FAST AND ROTATED BRIEF)

- Features : Descripteurs ORB
- Réduction dimension: K-means clustering sur descripteurs, projection dans l'espace des centroïdes
- Clustering: K-means sur matrice de données après réduction – 7 clusters

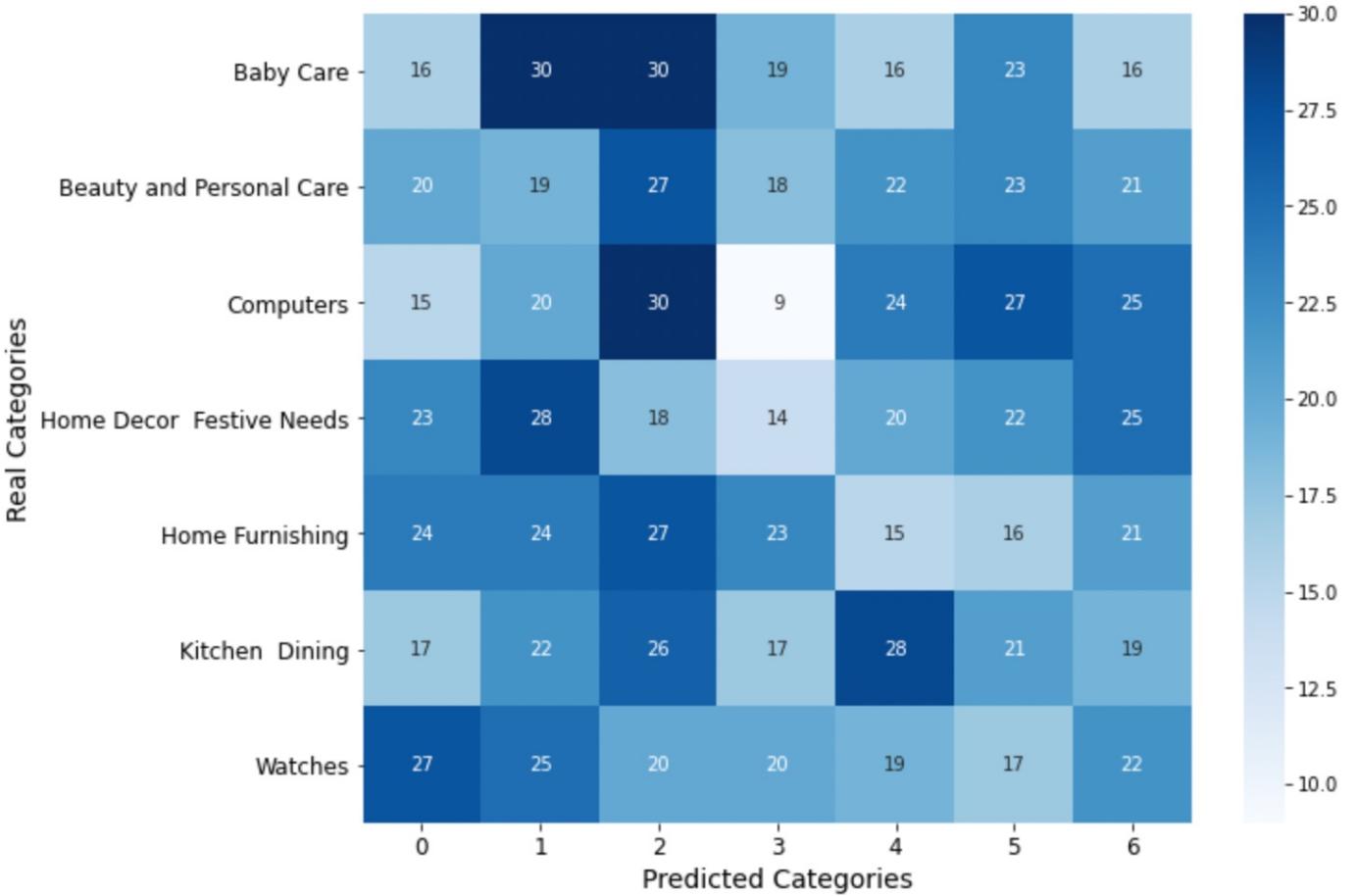
IMAGES : DESCRIPTEURS ORB

ACP / T-SNE des features extraits - descripteurs ORB



DESCRIPTEURS ORB (ORIENTED FAST AND ROTATED BRIEF)

■ Adjusted Rand Score: -0.00



IMAGES: TRANSFER LEARNING RESNET-50 + UMAP

- Features: Extraction avec réseau de neurones convolutifs ResNet-50 (suppression de la dernière couche)
- Réduction de dimension: UMAP - $\text{min_dist}' = 0.1$, $n_{\text{components}} = 50$, $n_{\text{neighbors}} = 100$
- Clustering: K-means - 7 clusters

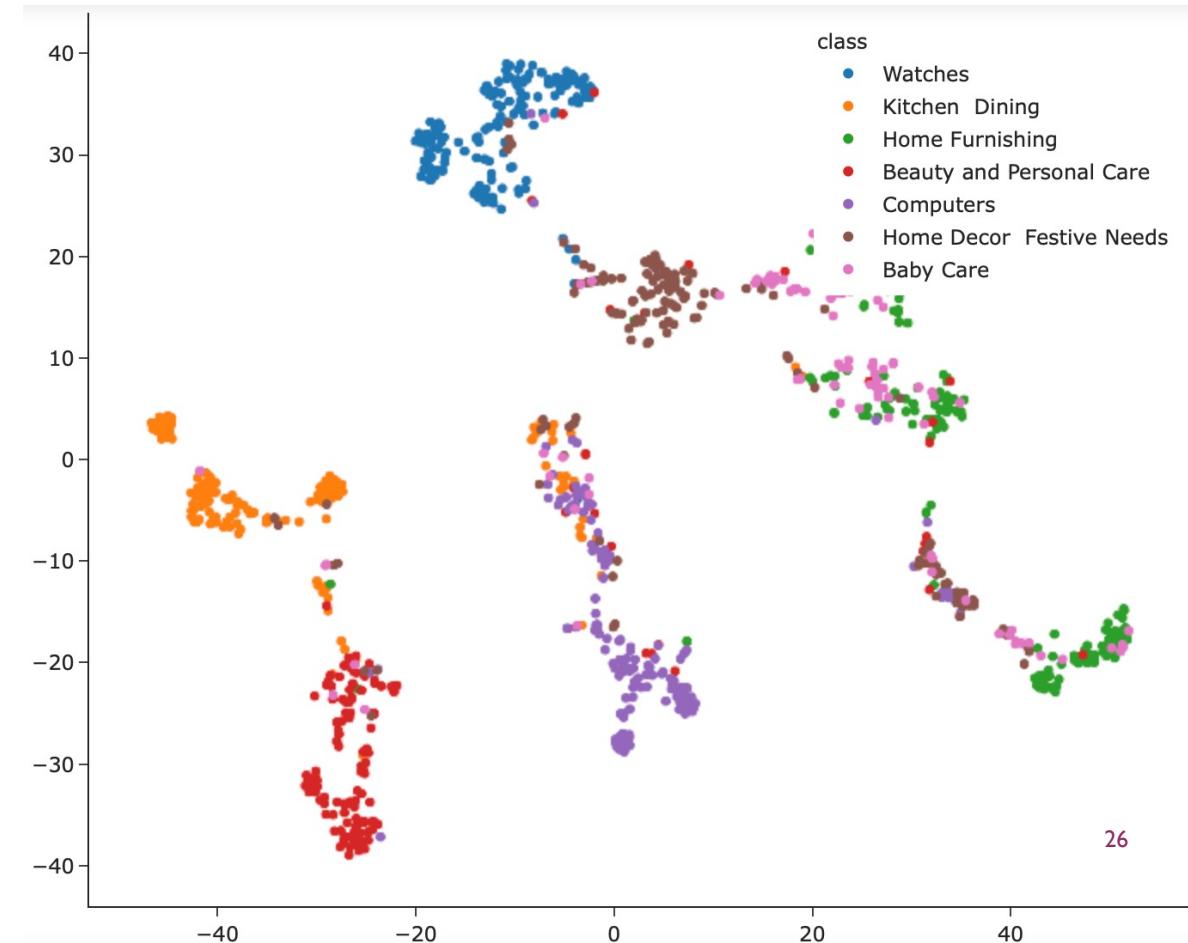
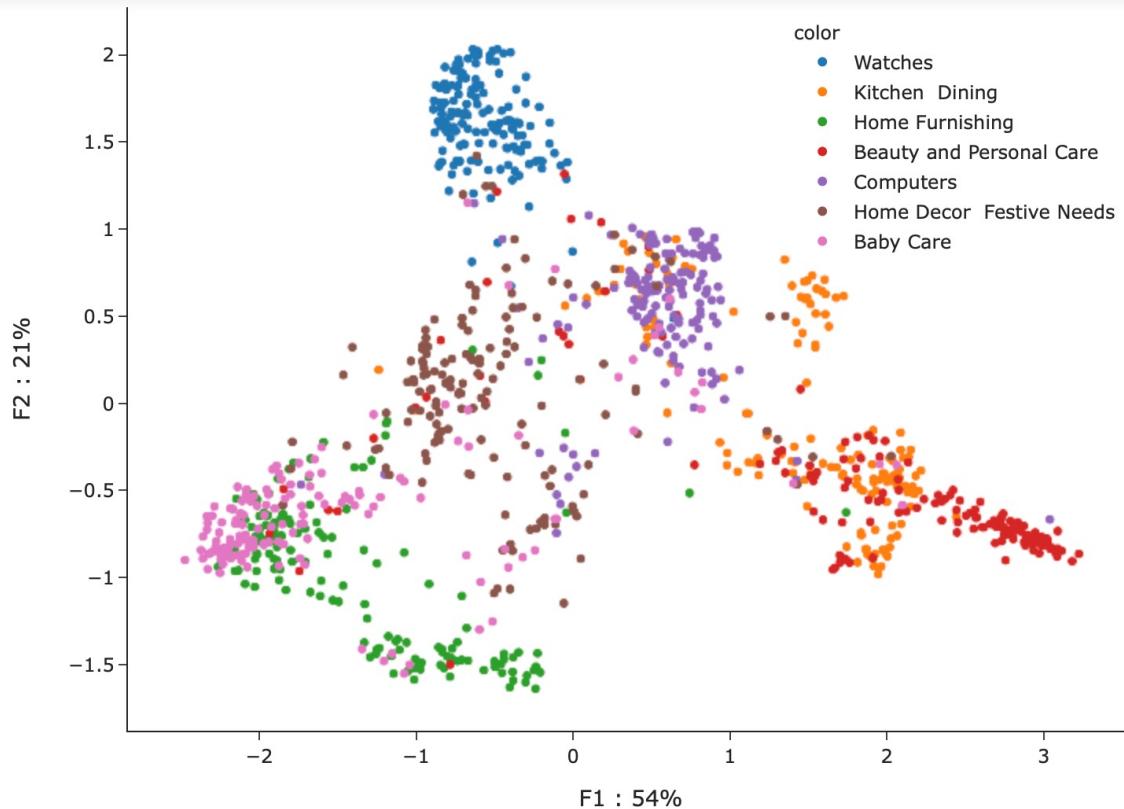
IMAGES: TRANSFER LEARNING RESNET-50 + UMAP

- ResNet-50 (residual network - 50) est un réseau neuronal convolutif de 50 couches, formé sur plus d'un million d'images de la base de données ImageNet.

conv5_block3_3_bn (BatchNormalization)	(None, 7, 7, 2048)	8192
conv5_block3_add (Add)	(None, 7, 7, 2048)	0
conv5_block3_out (Activation)	(None, 7, 7, 2048)	0
avg_pool (GlobalAveragePooling2D)	(None, 2048)	0
predictions (Dense)	(None, 1000)	2049000
=====		

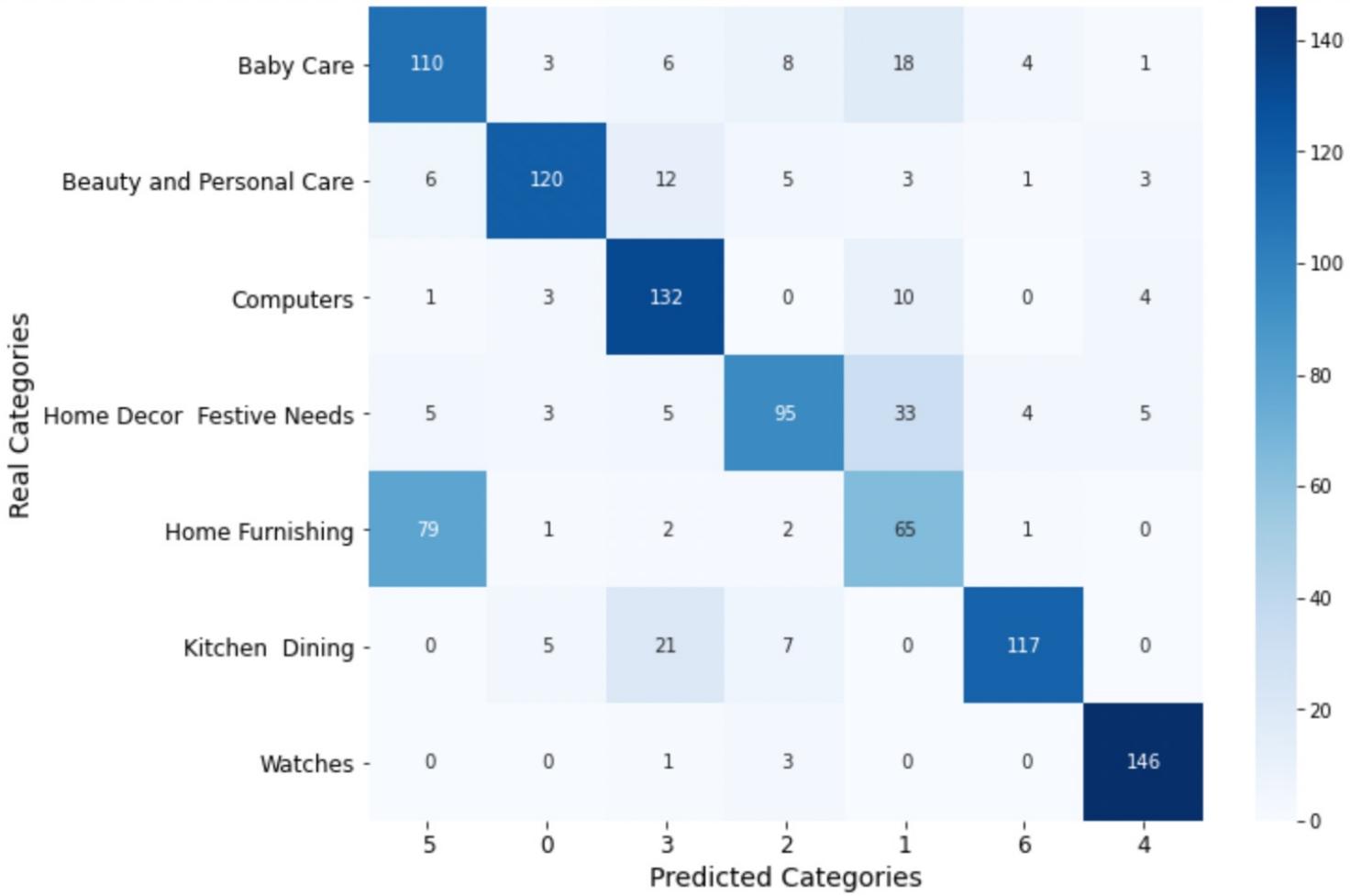
TRANSFER LEARNING RESNET-50 + UMAP

ACP / T-SNE des données après ResNet-50 et UMAP



TRANSFER LEARNING RESNET-50 + UMAP

■ Adjusted Rand Score: 0.56

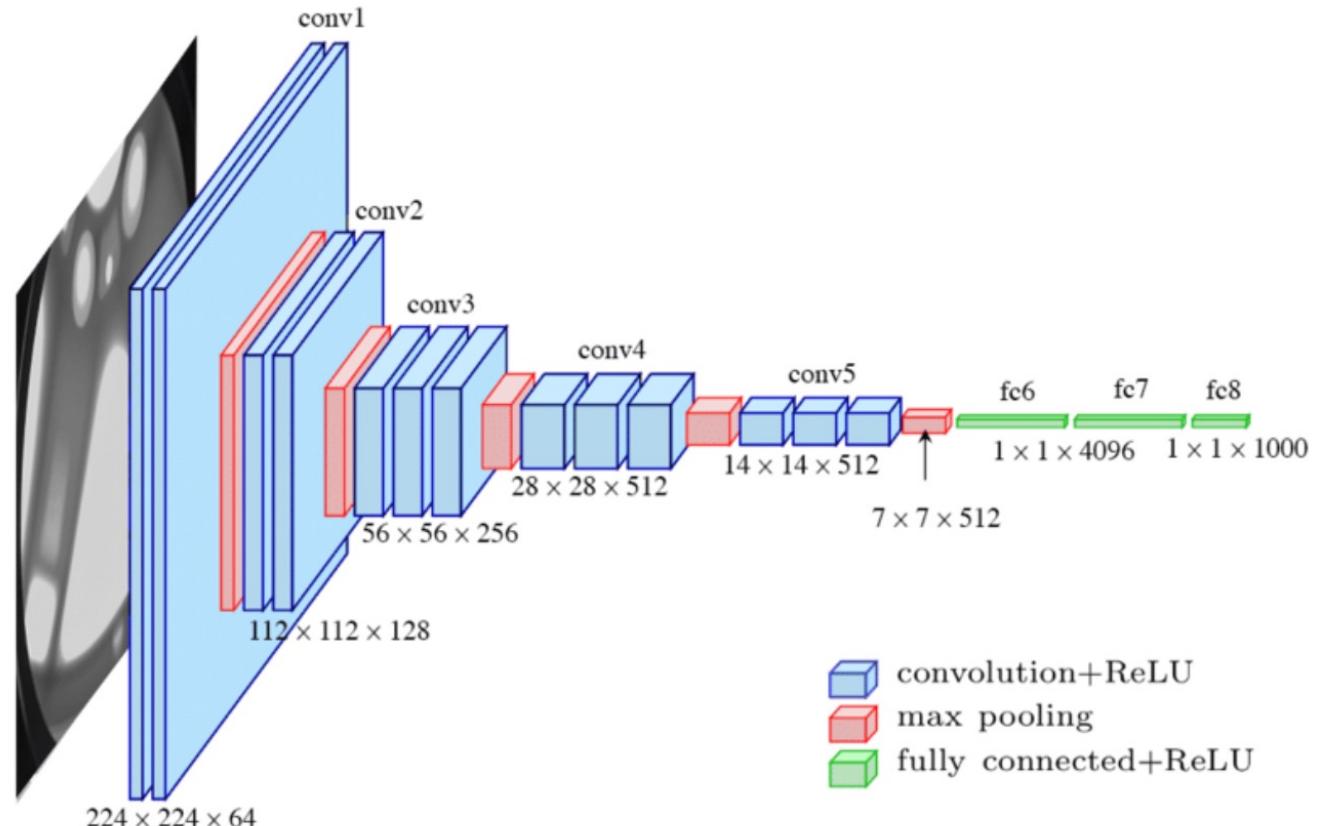


IMAGES: TRANSFER LEARNING - VGG-16 + NMF

- Features: extraction avec réseau de neurones convolutifs VGG-16 (suppression de la dernière couche)
- Réduction dimension: NMF (factorisation de matrice non négative)
- Clustering : Rattachement à l'axe de meilleure projection

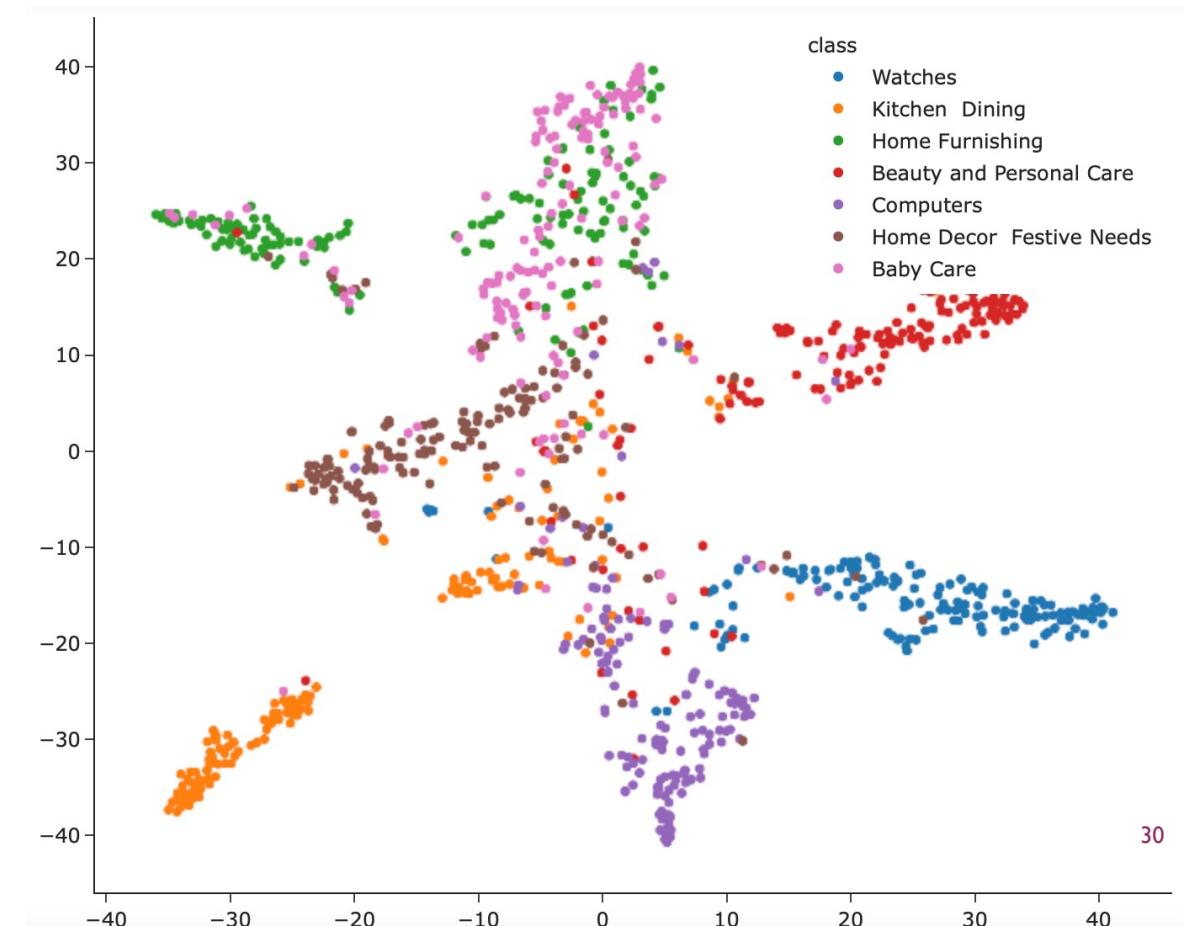
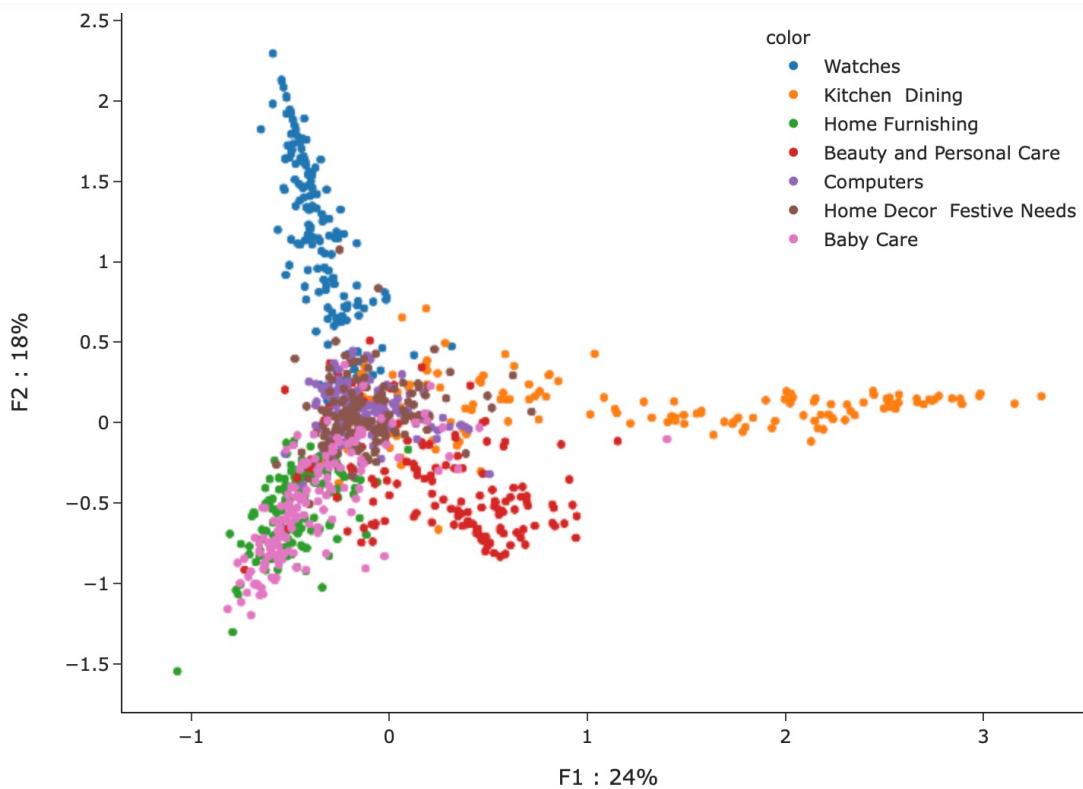
IMAGES: TRANSFER LEARNING VGG-16 + NMF

- Features: extraction avec réseau de neurones convolutifs VGG-16



TRANSFER LEARNING VGG-16 + NMF

ACP / T-SNE des données après VGG-16 et NMF



TRANSFER LEARNING VGG-16 + NMF

■ Adjusted Rand Score: 0.50

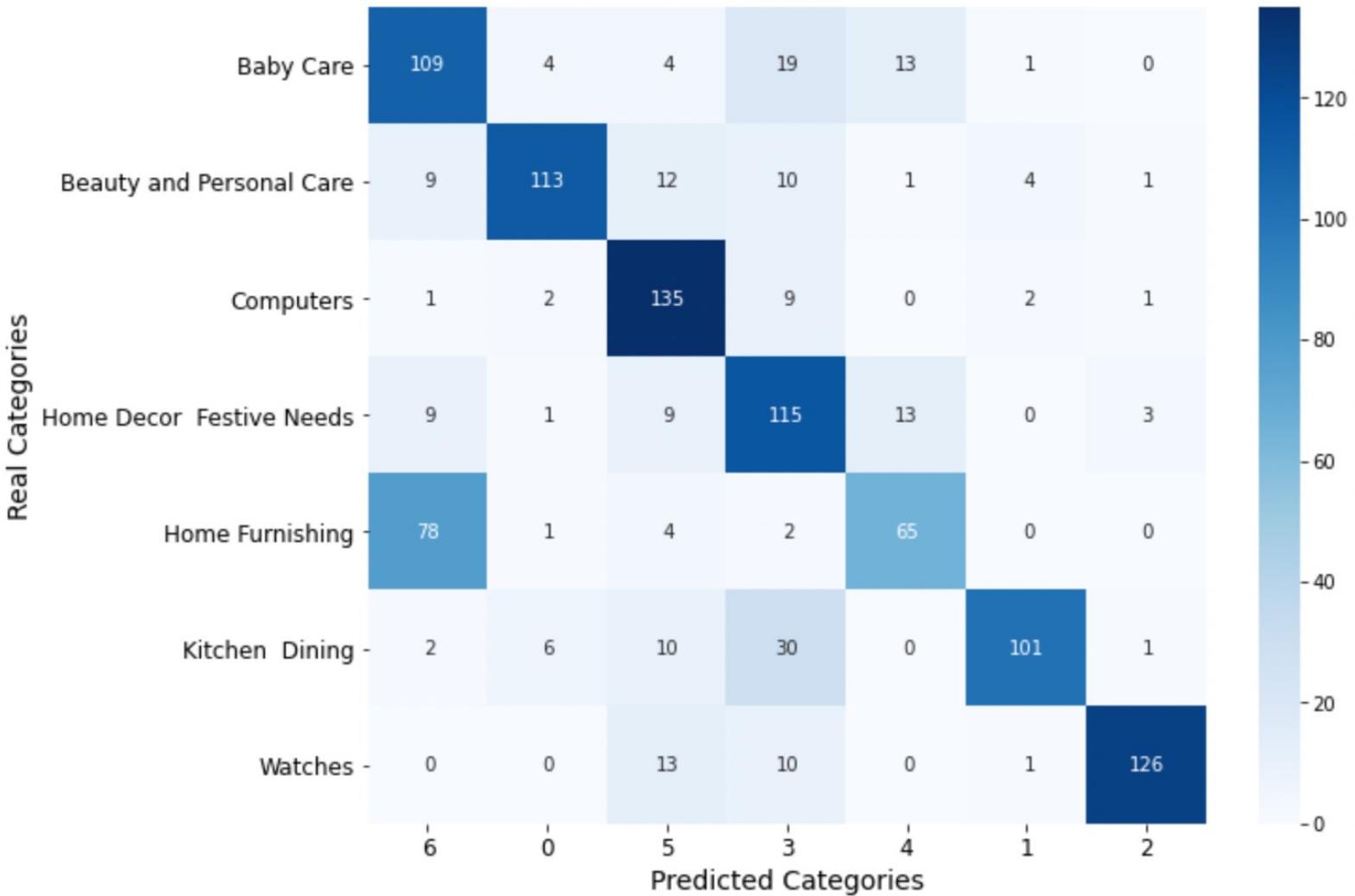


IMAGE : COMPARAISON DES MODÈLES

Images: Methods Used	ARI
ORB	0
VGG16 / PCA	0.44
VGG16 / NMF	0.5
VGG16 / UMAP	0.47
ResNet50 / PCA	0.53
ResNet50 / NMF	0.46
ResNet50 / UMAP	0.56

COMBINAISON : TEXTE & IMAGE

Texte:

- Bag of words: stemmer, stopwords utilisés - mots présents dans plus de 30% et moins de 1% des individus, pas d'utilisation du TfidfVectorizer()
- UMAP : min_dist 0.1, n_components 25, n_neighbors 100

Image:

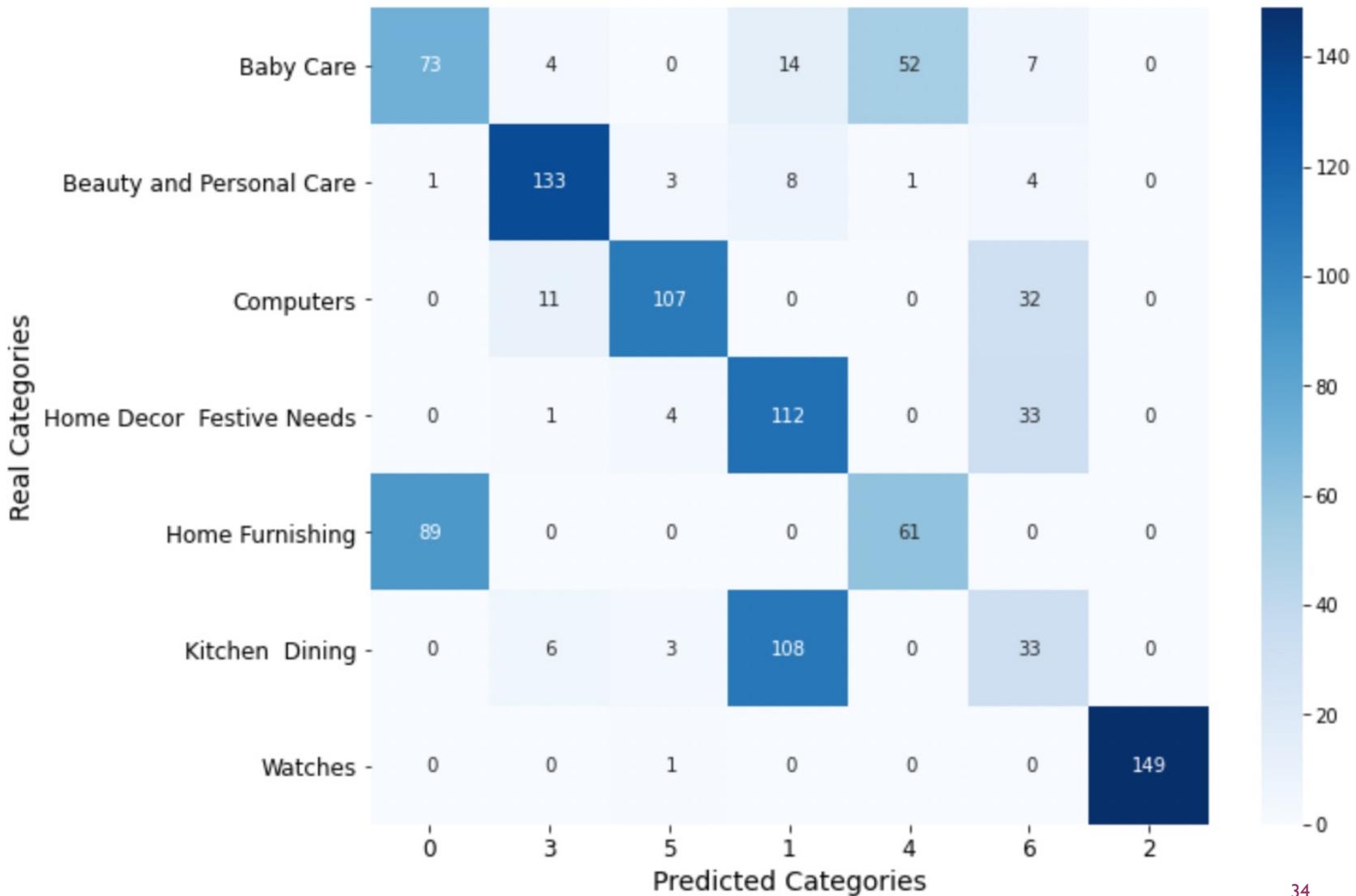
- ResNet-50
- UMAP - min_dist' = 0.1, n_components = 50, n_neighbors = 100

Fusion des matrices issues de UMAP puis K-means avec 7 clusters.

COMBINAISON TEXTE & IMAGE

■ Adjusted Rand Score
Combinaison: 0.53

- ARI Image : 0.56
- ARI Texte : 0.704

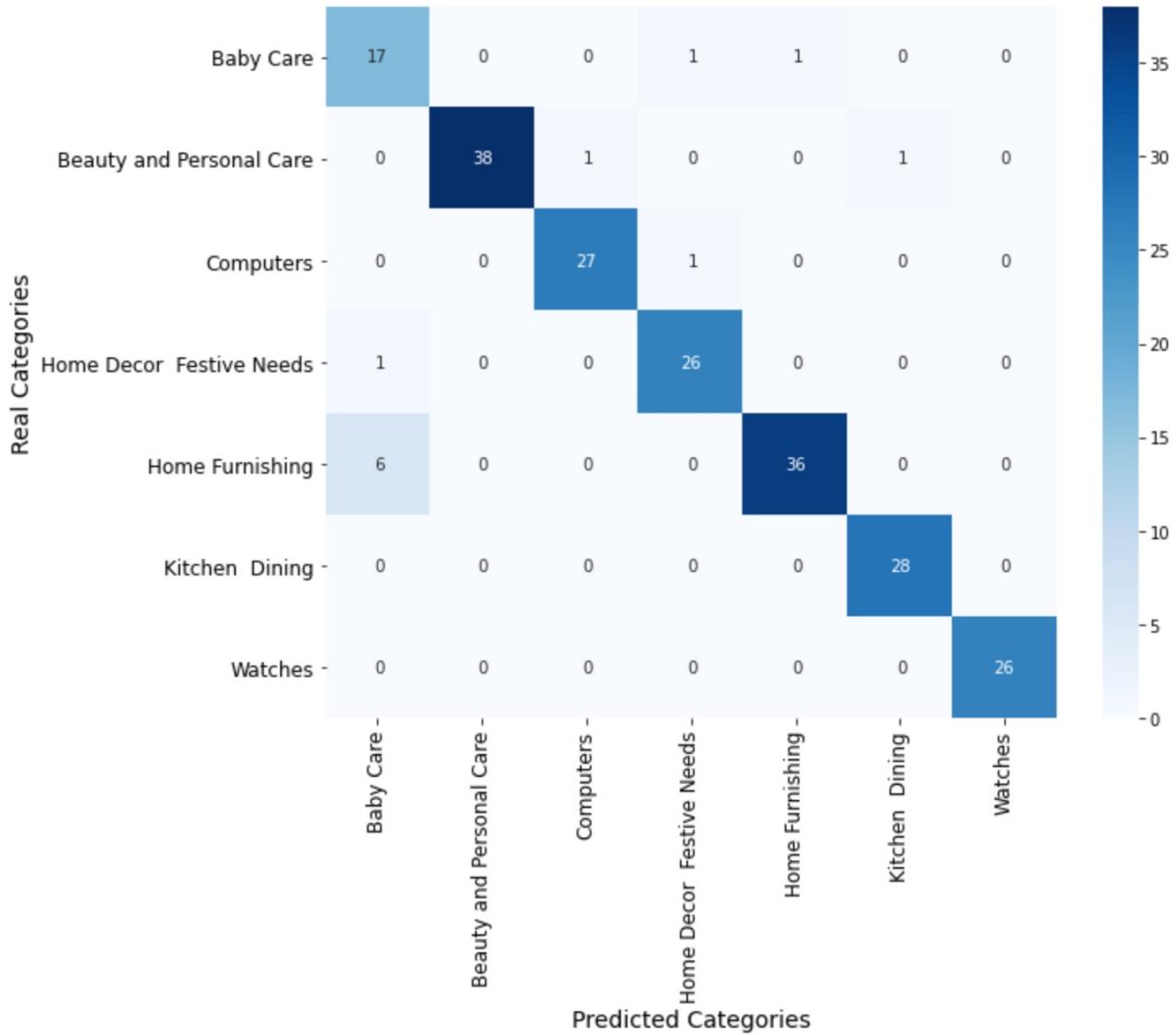


CLASSIFICATION SUPERVISÉE : TEXTE

- Stemmer
- Stopwords utilisés : mots présents dans plus de 30% et moins de 1% des individus
- Pas d'utilisation du TfidfVectorizer()
- Réduction de dimension : UMAP : min_dist = 0.1, n_components = 25, n_neighbors = 100
- Classification supervisée : RandomForestClassifier : max_depth = 30, max_features = 4, n_estimators = 200

CLASSIFICATION SUPERVISEÉ : TEXTE

- Adjusted Rand Score sur le jeu de test : 0.88

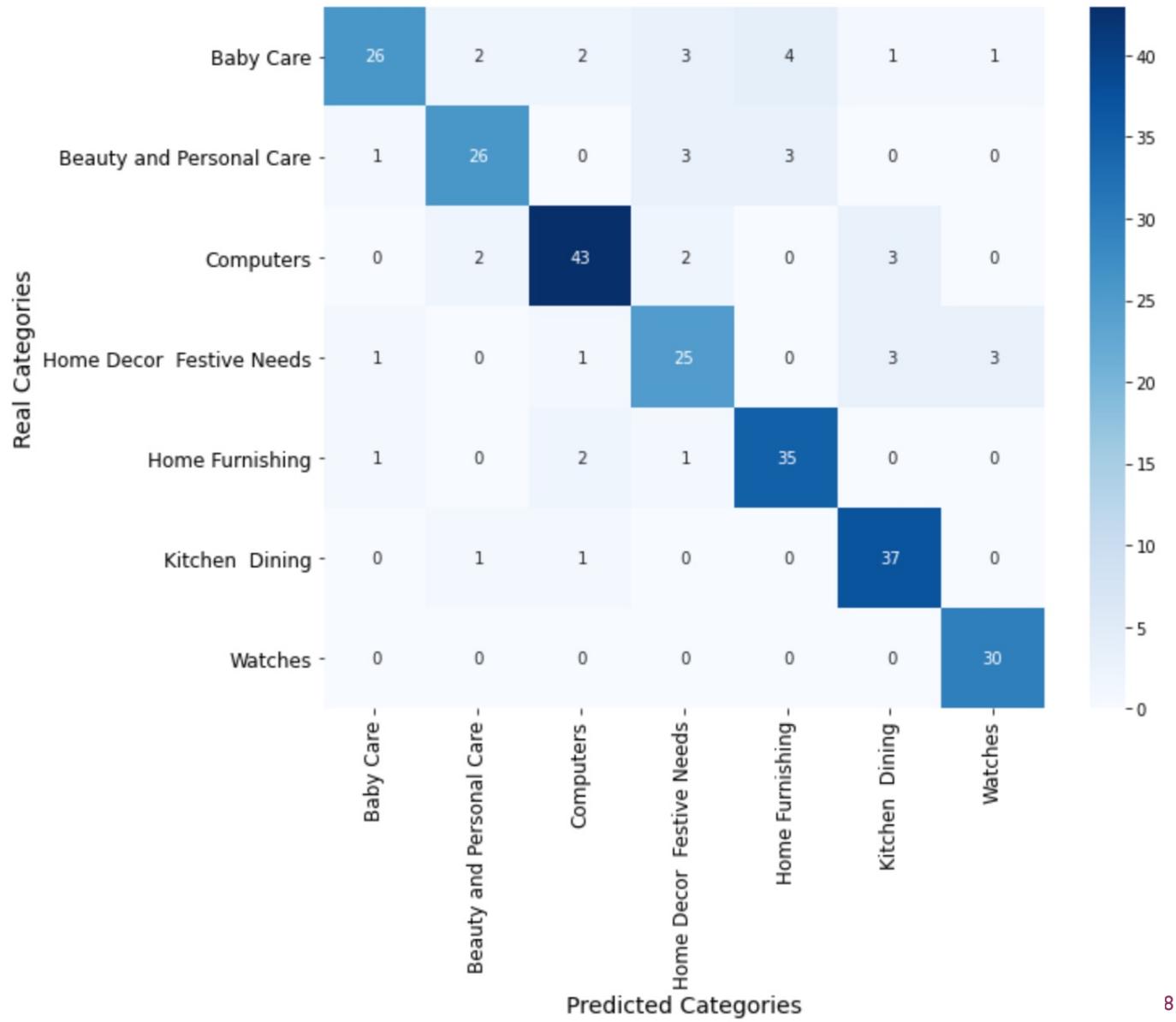


CLASSIFICATION SUPERVISÉE : IMAGE

- Features : Extraction avec réseau de neurones convolutifs ResNet-50
- Réduction de dimension: UMAP - `min_dist' = 0.1, n_components = 50, n_neighbors = 100`
- Classification supervisée : `RandomForestClassifier - max_depth = 10, max_features = 8, n_estimators = 200`

CLASSIFICATION SUPERVISEE : IMAGE

- Adjusted Rand Score sur le jeu de test : 0.67



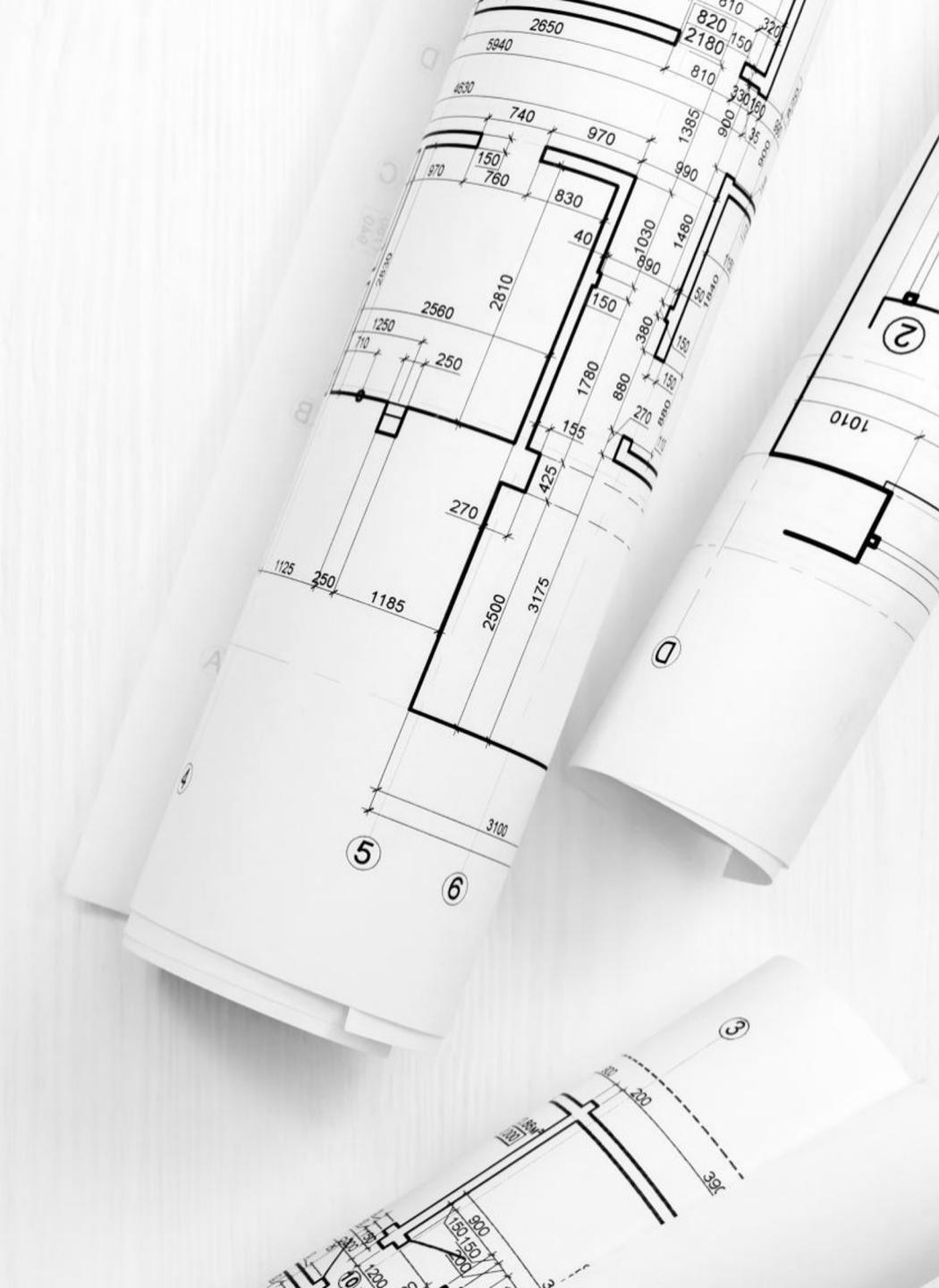
CONCLUSION SUR LA FAISABILITÉ DU MOTEUR DE CLASSIFICATION ET VOS RECOMMANDATIONS POUR SA CRÉATION ÉVENTUELLE

RECOMMANDATIONS POUR SA CRÉATION ÉVENTUELLE

- Choix / évaluation des stopwords
- Augmenter le volume de données d'entraînement, et entraîner un réseau de neurones sur des images correspondant aux 7 catégories
- Combiner les données provenant des images et du texte de façon plus efficace: stacking par exemple.

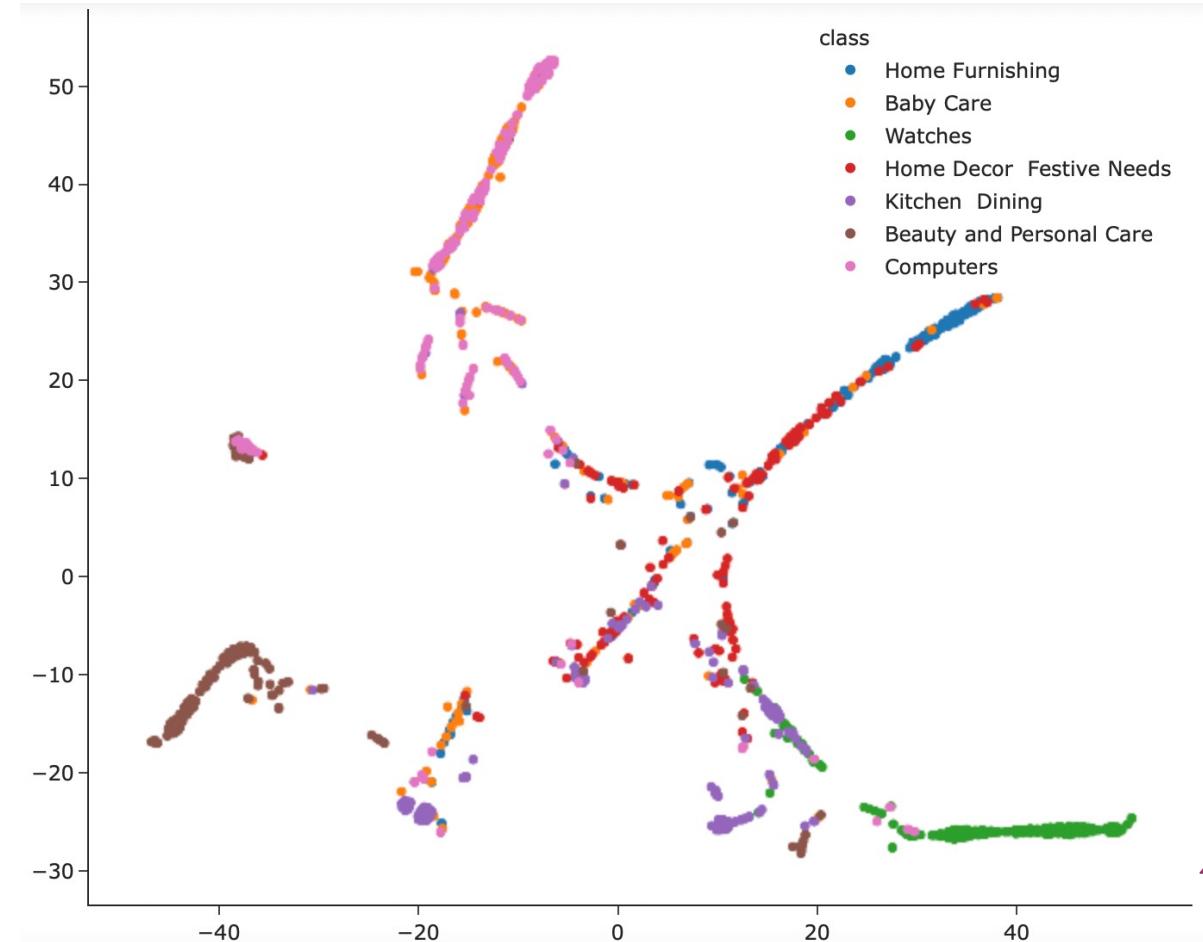
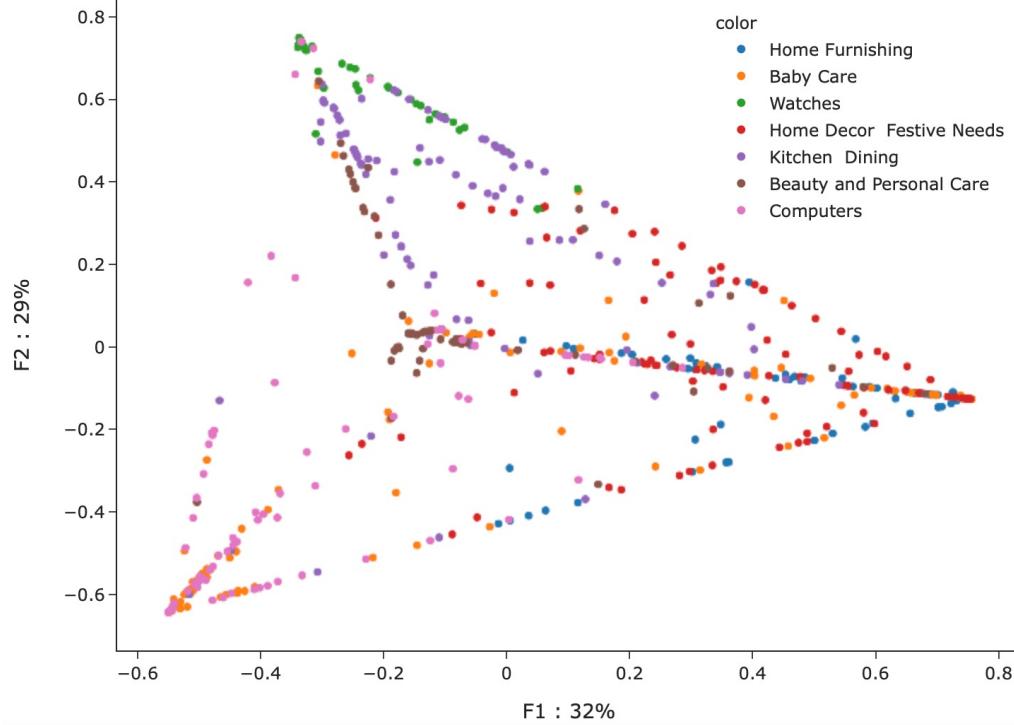
QUESTIONS - RÉPONSES

ANNEXES



ANNEXES : LDA (LATENT DIRICHLET ALLOCATION)

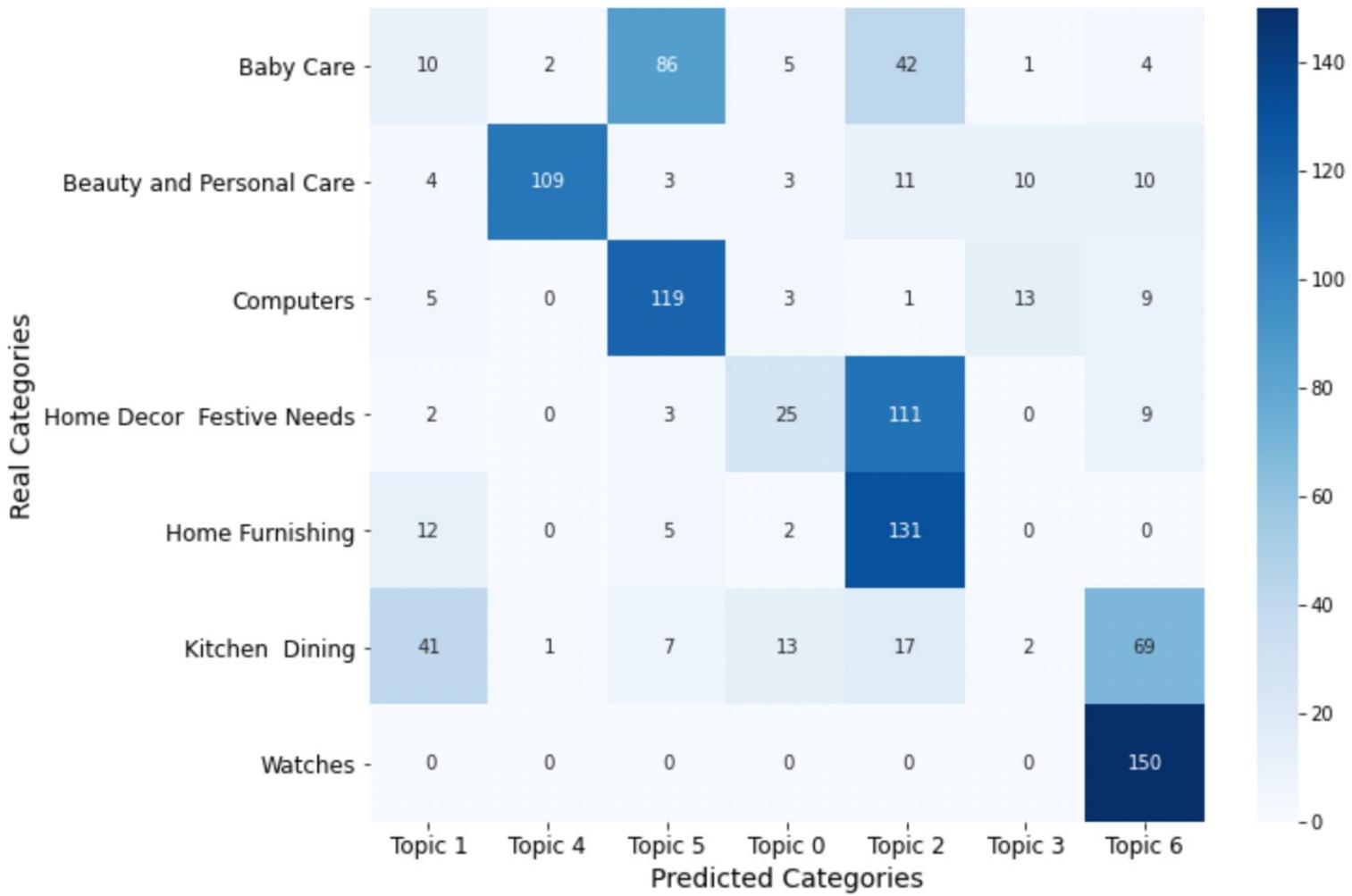
ACP / T-SNE des données après BoW et LDA



ANNEXES:

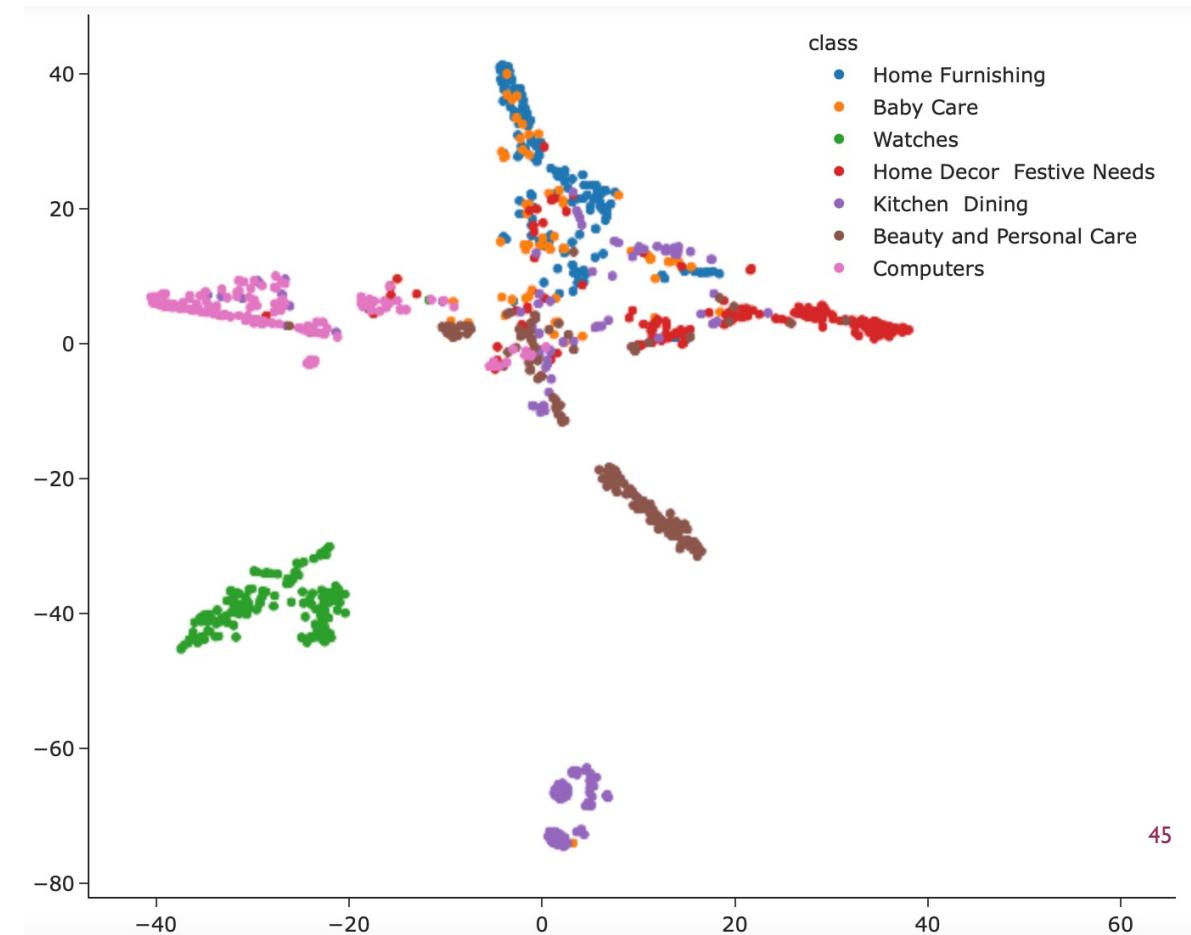
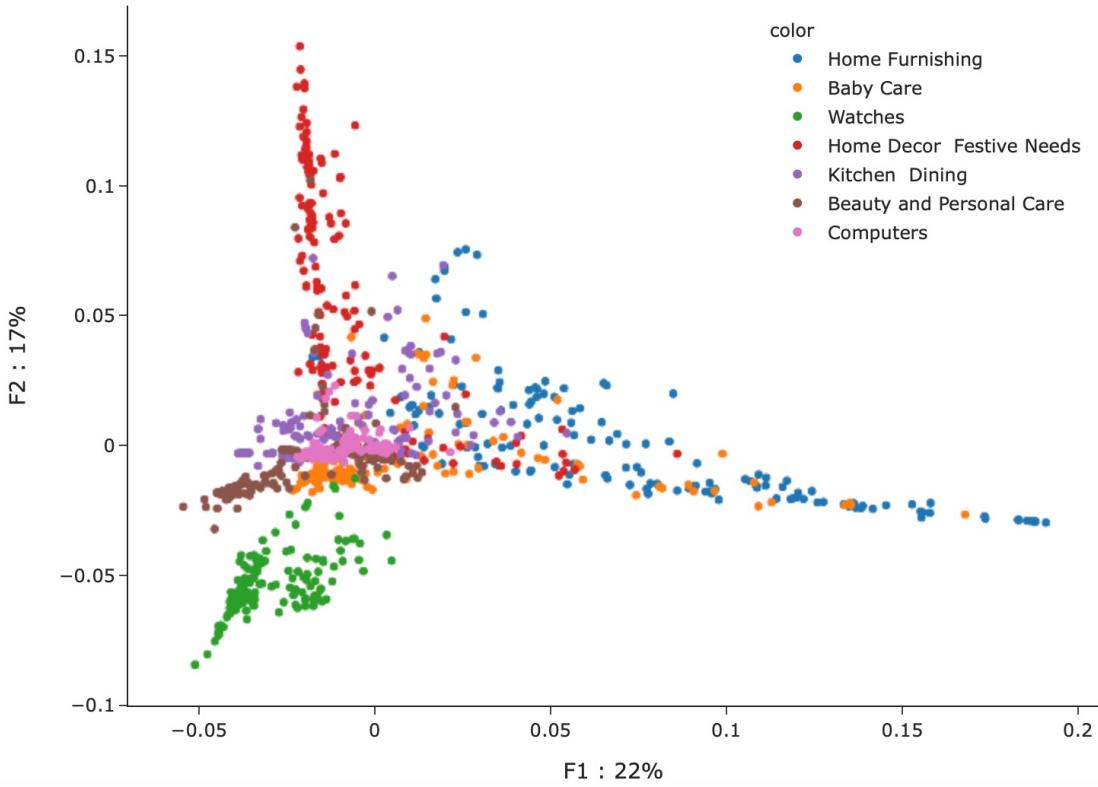
LDA (LATENT DIRICHLET ALLOCATION)

■ Adjusted Rand Score: 0.38



ANNEXES : NMF (NON-NEGATIVE MATRIX FACTORIZATION)

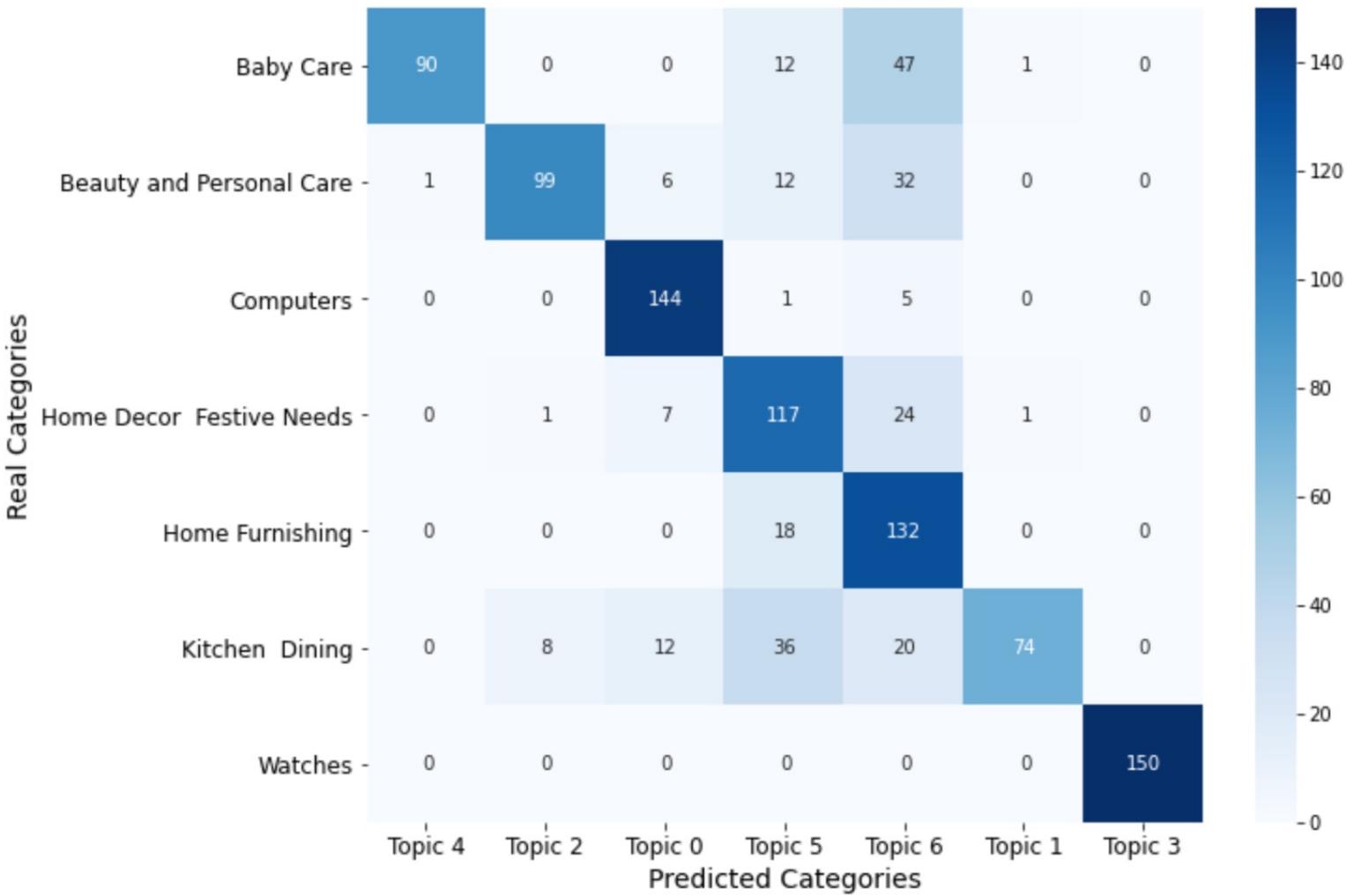
ACP / T-SNE des données après BoW et NMF



ANNEXES:

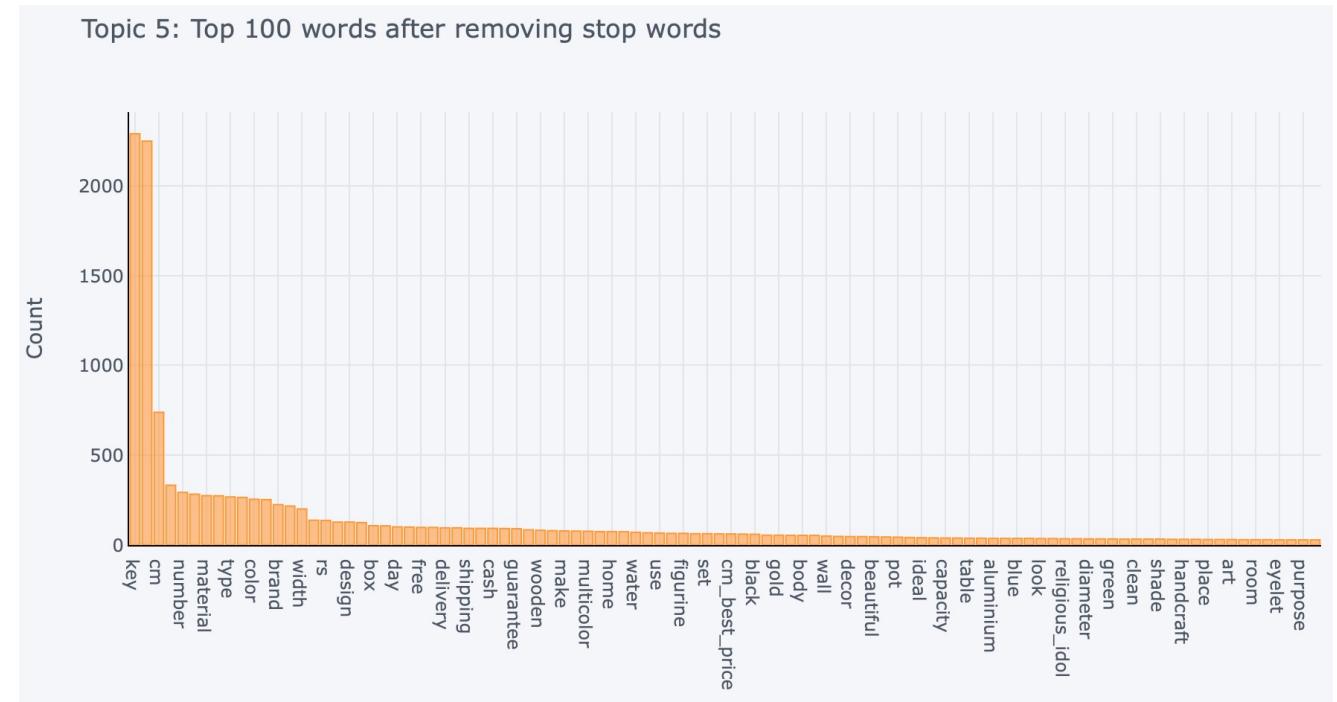
NMF (NON-NEGATIVE MATRIX FACTORIZATION)

■ Adjusted Rand Score: 0.54



ANNEXES TEXTE : WORD CLOUD ET FRÉQUENCE DES MOTS

Classe estimée par NMF: Home Decor Festive Needs / Kitchen Dining



ANNEXES TEXTE : FRÉQUENCE DES BIGRAMS ET TRIGRAMS

Classe estimée par NMF: Home Decor Festive
Needs / Kitchen Dining

