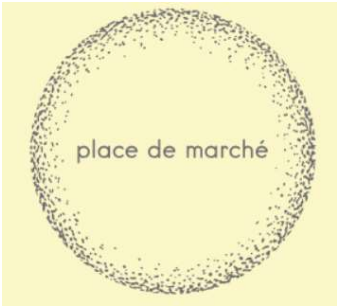


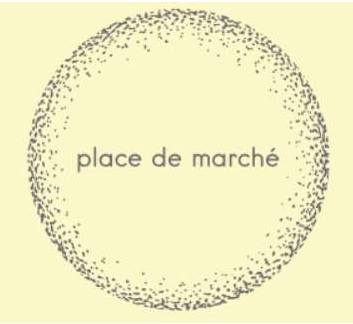
Classification automatique de biens de consommation

MISE EN ŒUVRE DE METHODES DE CLUSTERING

BASE DE DONNÉES D'ARTICLES flipkart_com-ecommerce_sample_1050.csv

+ PHOTOS

Plan



Présentation du jeu de données

Analyse exploratoire

Méthodologie et Feature Engineering

Benchmark des méthodes de Clustering

Fusion des modèles NLP + Image Processing

Conclusion

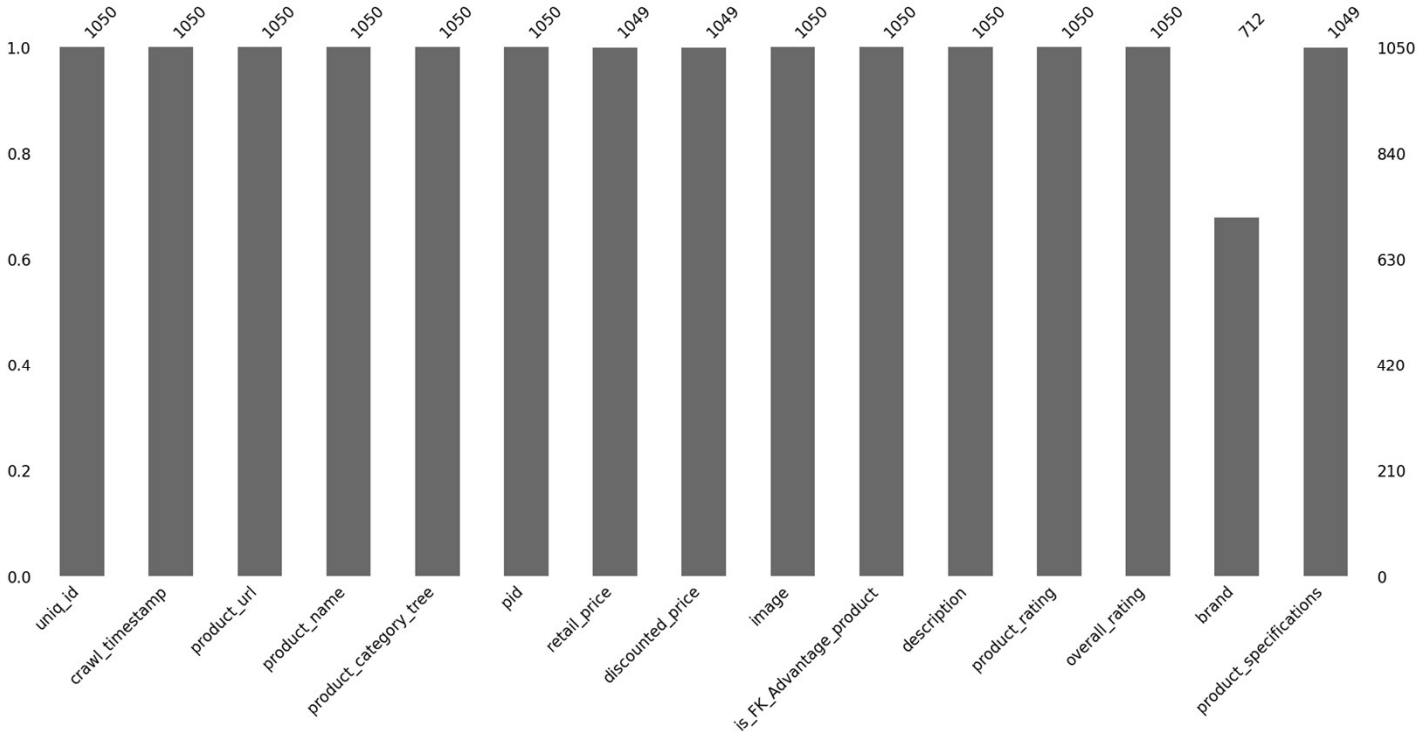
Analyse exploratoire

Présentation du jeu de données

BASE D'ARTICLES
1050 ENTREES x 15 COLONNES
1050 PHOTOS

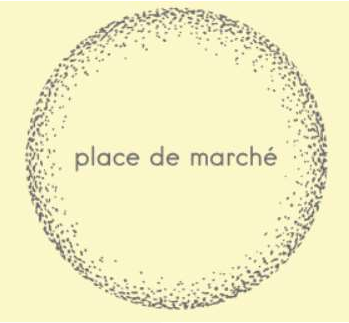
colonnes TEXTE d'intérêts :
- **DESCRIPTION**
- **NAME**
- **SPECIFICATIONS**

Classification:
- **Catégories : 7 classes**

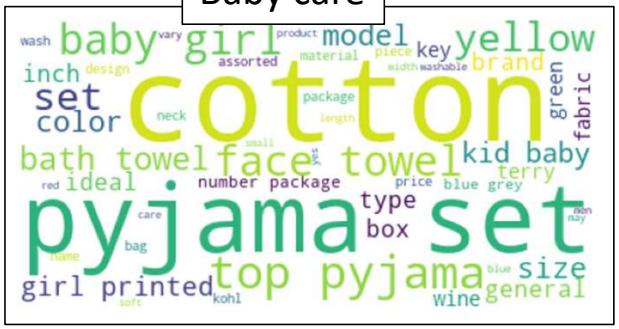


Analyse exploratoire

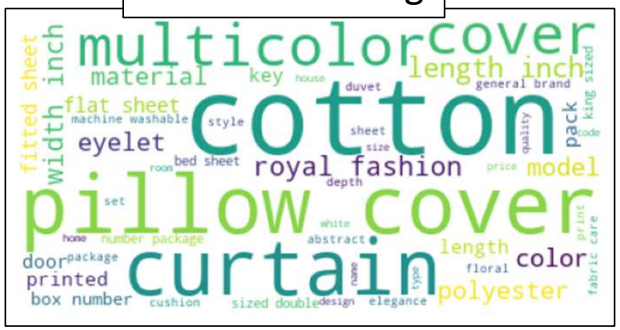
Wordcloud



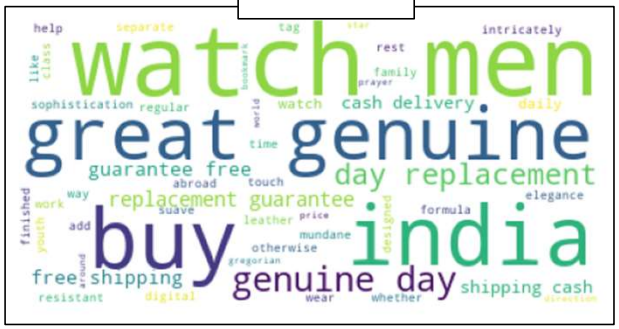
Baby care



Home furnishing

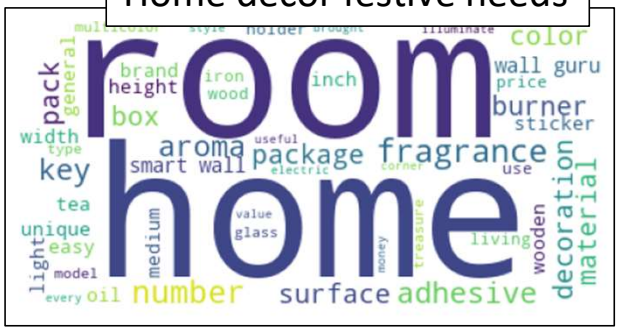


watches

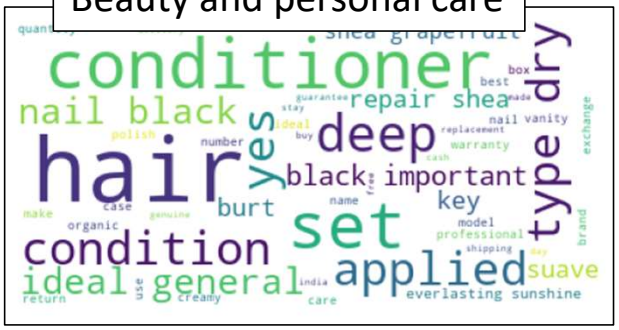


computers

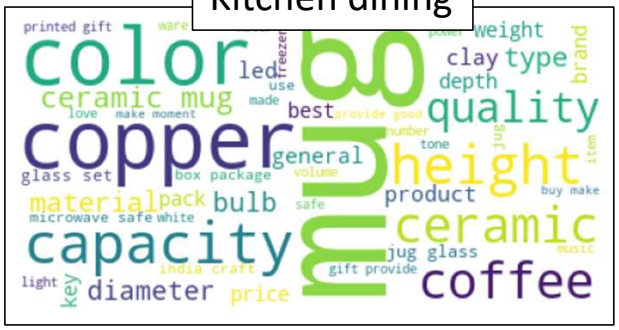
Home decor festive needs



Beauty and personal care

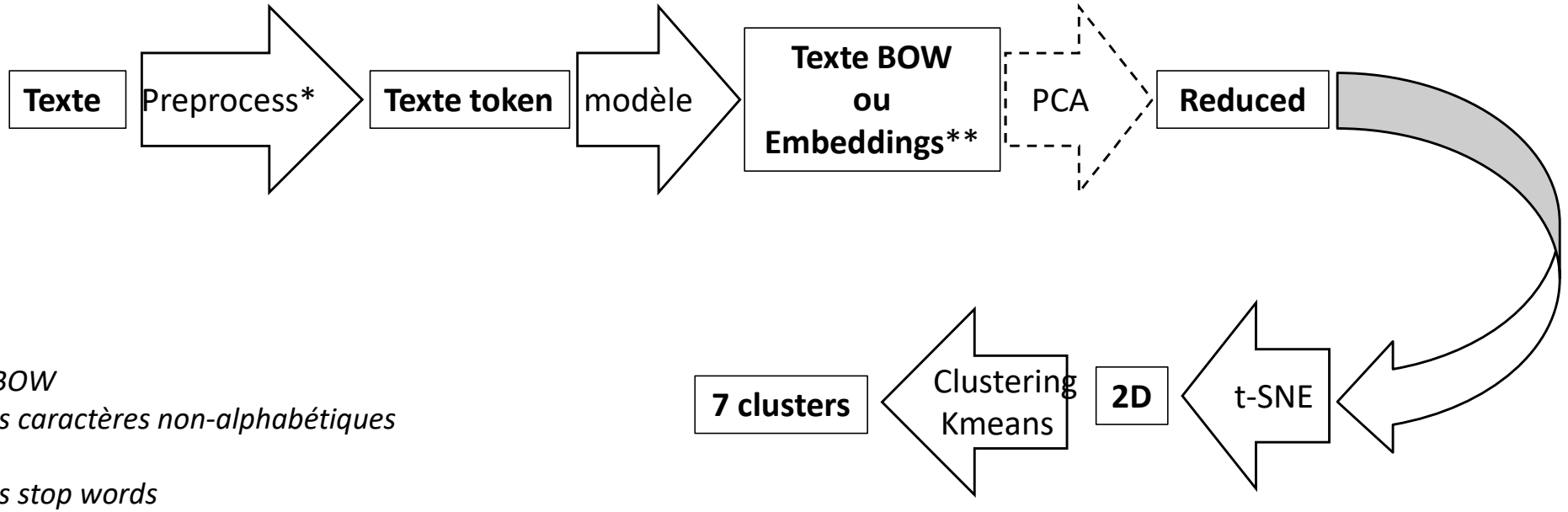


Kitchen dining



Méthodologie modèles NLP

Feature Engineering & clustering



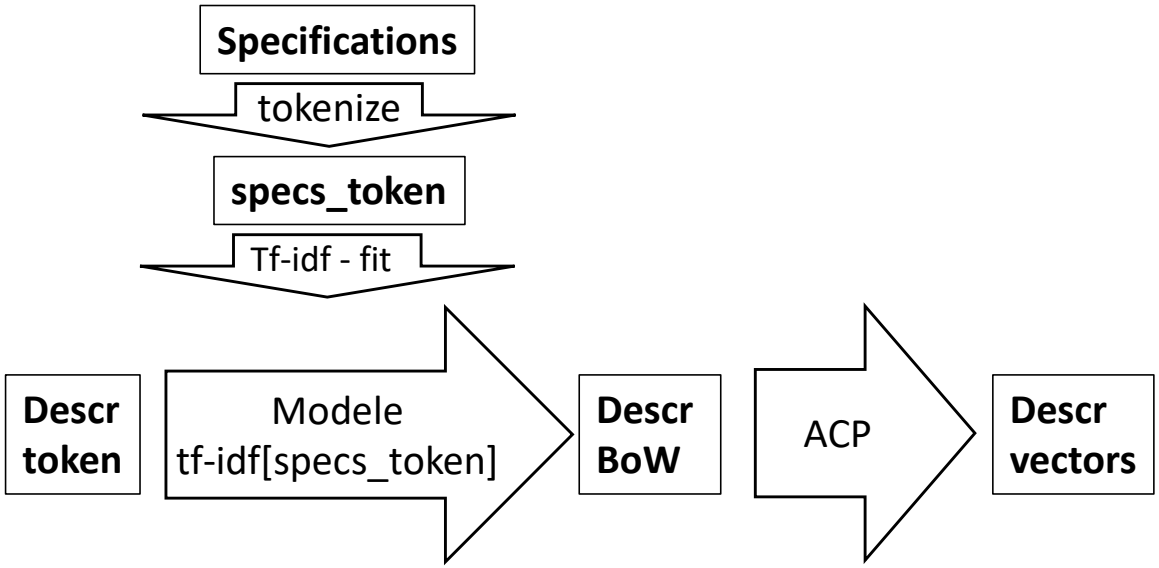
**Preprocess pour BOW*

- *Suppression des caractères non-alphabétiques*
- *Tokenization*
- *Suppression des stop words*
- *Vocabulaire restreint au dictionnaire anglais*
- *Lemmatization*

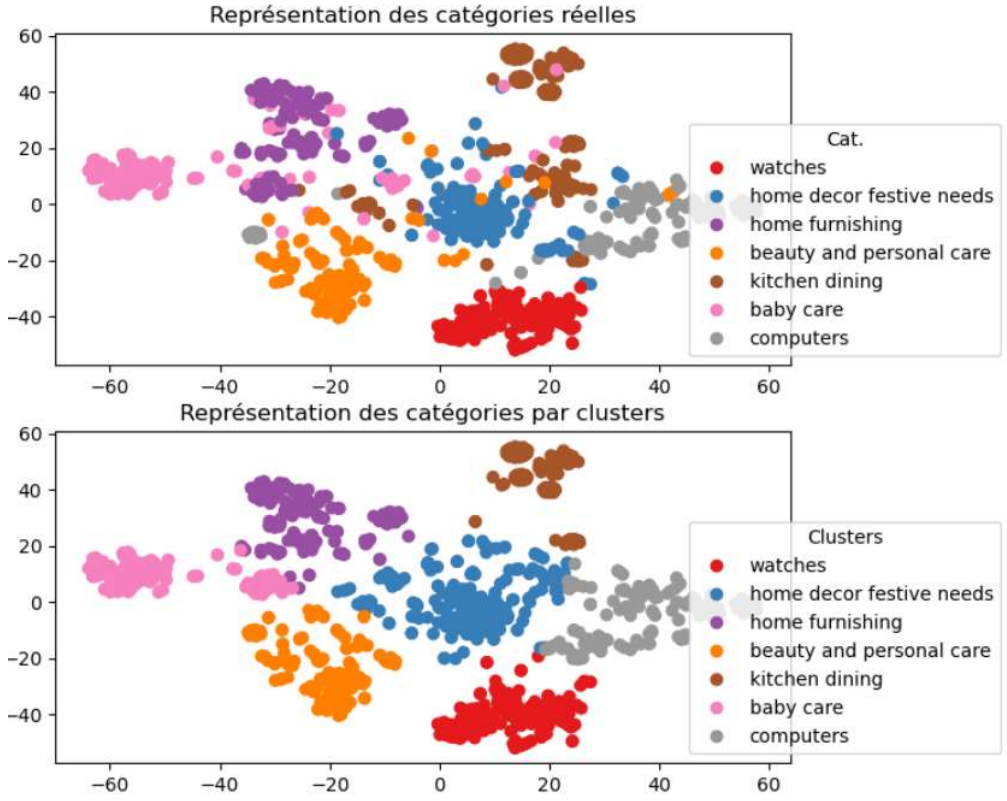
*** word embeddings(word2vec), sentence embeddings (Transformers)*

NLP Clustering benchmark

Tf-idf + ACP (TruncatedSVD)



Tf-idf	max_df	min_df	n_ACP	ARI
Fit(Specs) – Transform(Specs+Descr)	0.95	2	955	**0.66**



NLP Clustering benchmark

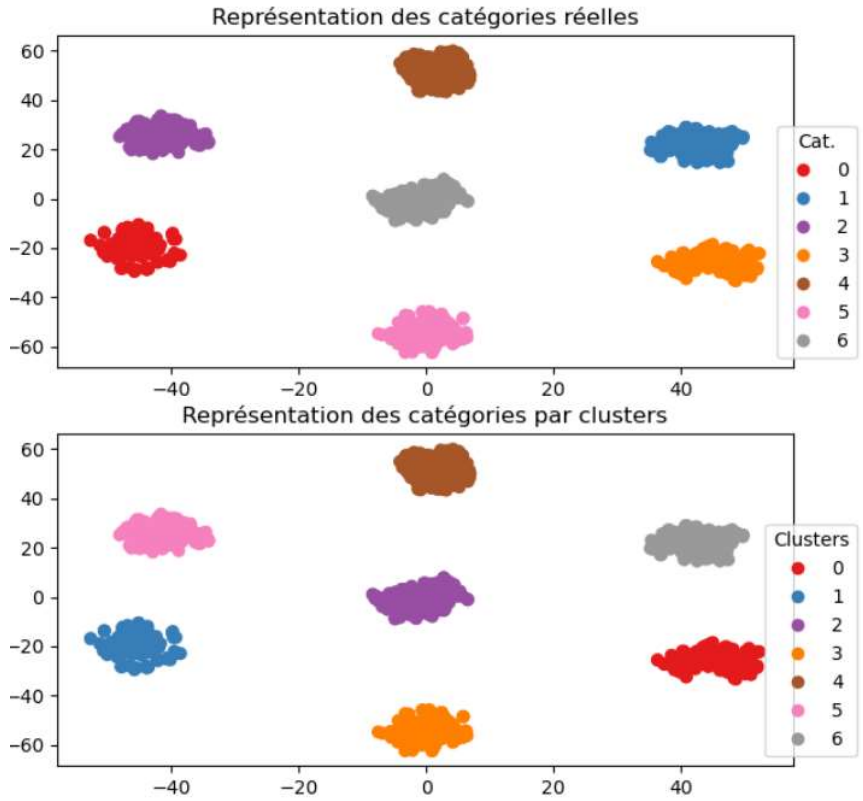
Tf-idf + LDA Linear Discriminant Analysis

Descr Vectors
(Modèle Tf-idf optimisé)

LDA

Clusters_LDA

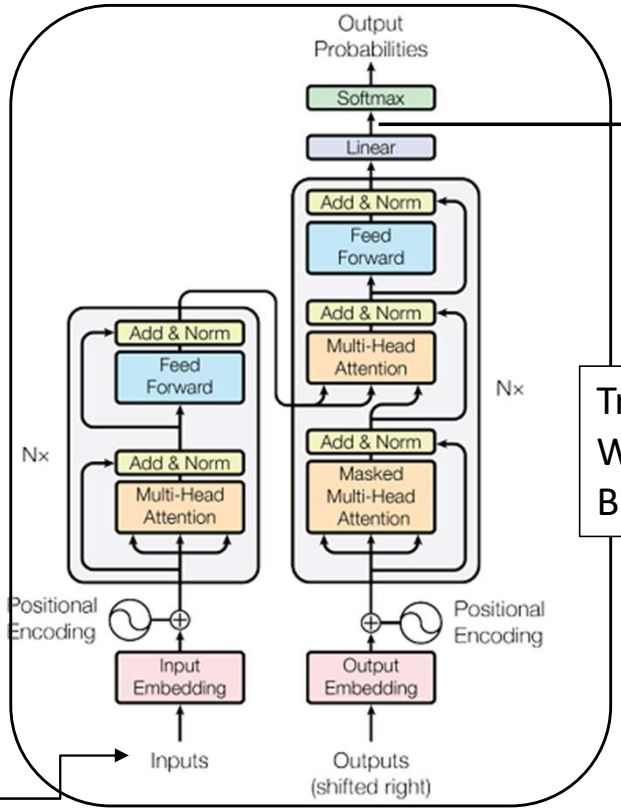
Tf-idf	ARI
LDA (6 composantes)	1



NLP Clustering benchmark

BERT : bert_en_uncased_L-12_H-768_A-12

inference



description

Trained on Wikipedia and BookCorpus

Description_embeddings
= last_hidden_states

1050x64x768

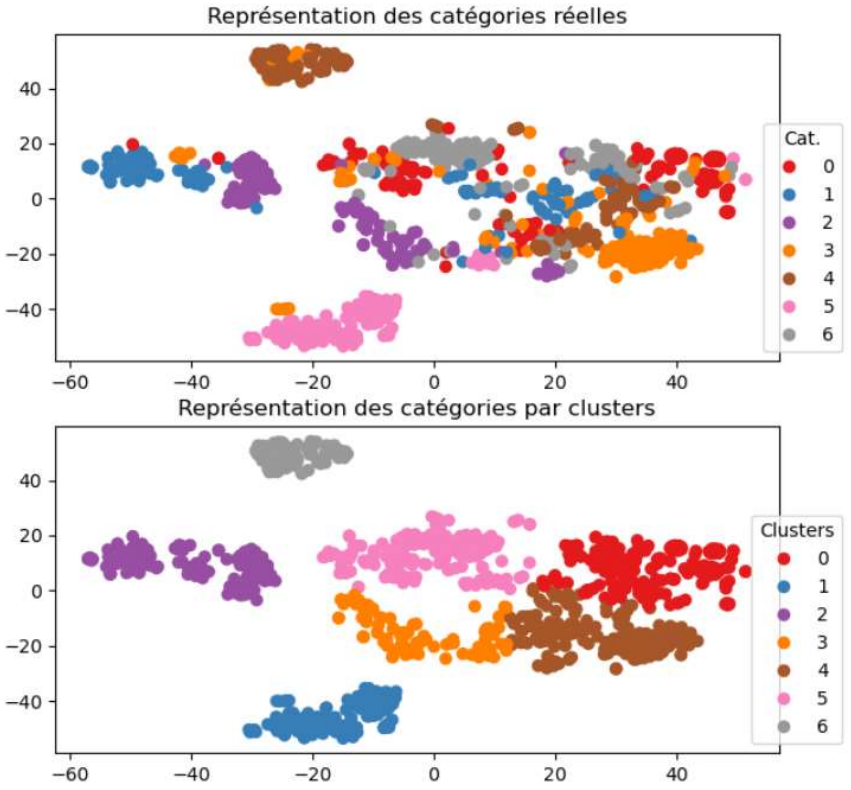
Clustering

NLP Clustering benchmark

Transformers : bert_en_uncased_L-12_H-768_A-12 (BERT)

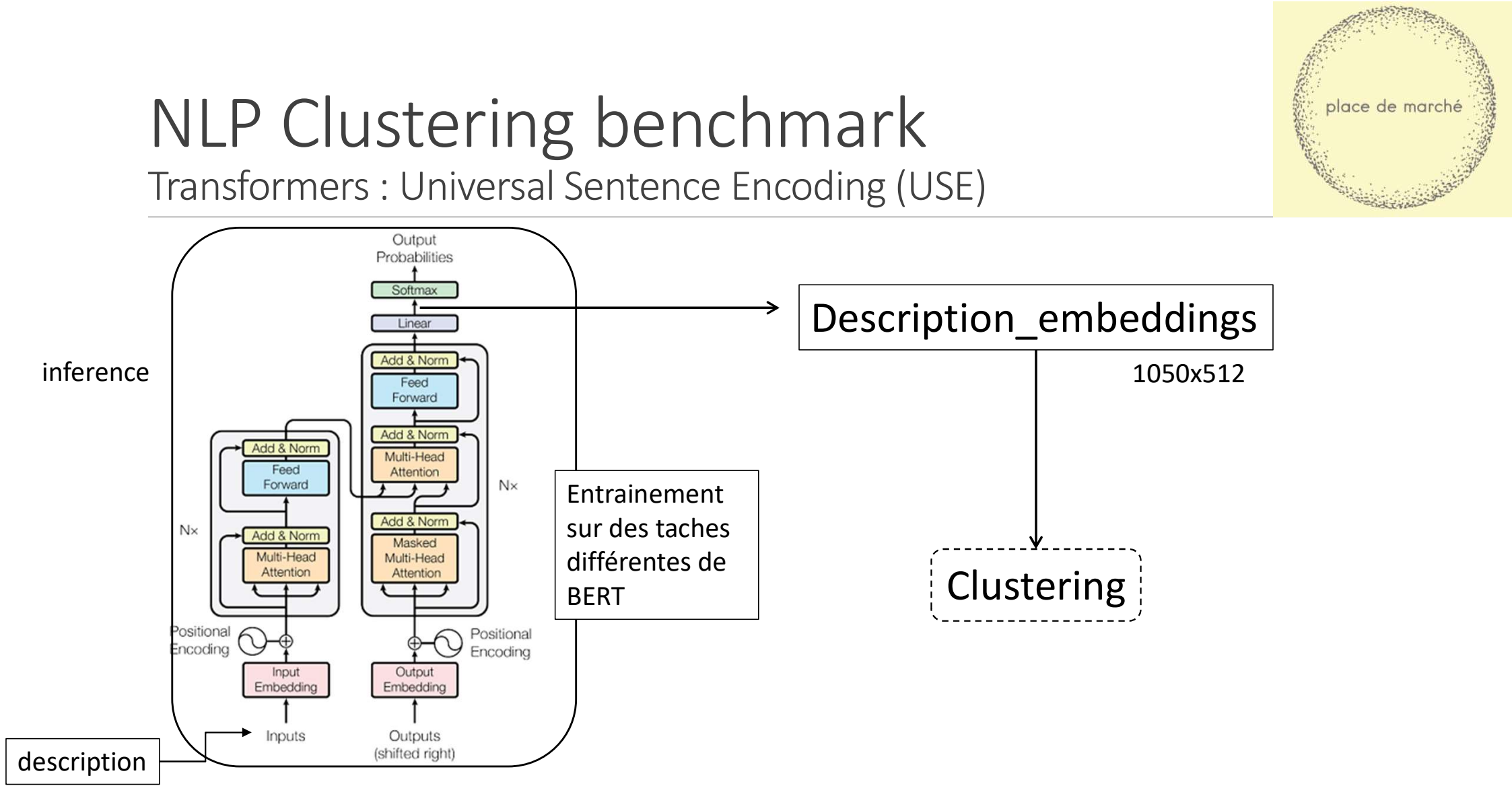
Différentes Stratégie de traitement de last_hidden_states

Entrées	Stratégie	Nb_token	score
descr_dl	average	64	**0.34**
descr_dl	topic	64	0.33
descr_dl	ACP	64	0.29
descr_dl	average	128	0.29
name_dl	average	64	0.06
specs_dl	average	64	0.28



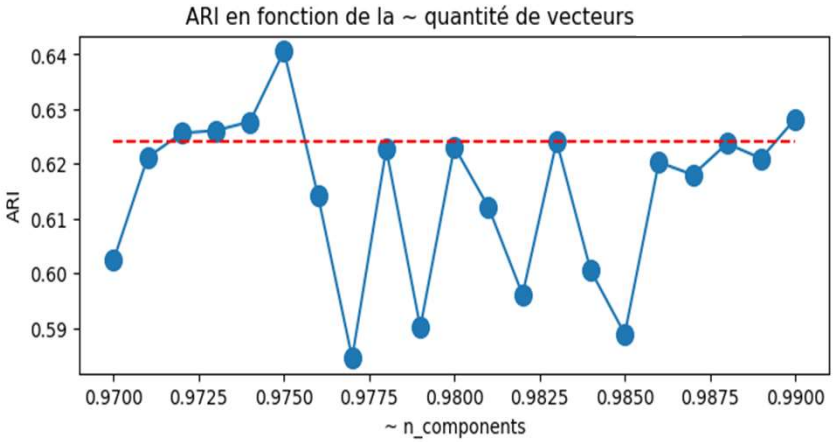
NLP Clustering benchmark

Transformers : Universal Sentence Encoding (USE)

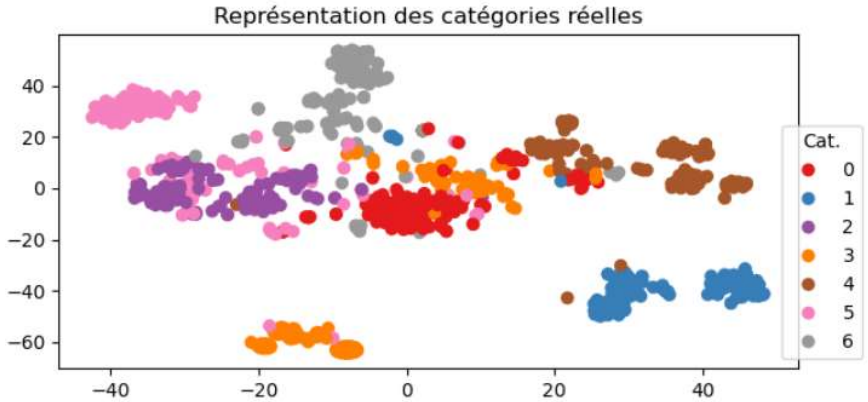


NLP Clustering benchmark

transformers : Universal Sentence Encoding (USE)



Entrée	ARI
descr_dl	0.44
name_dl	0.32
specs_df	**0.64**



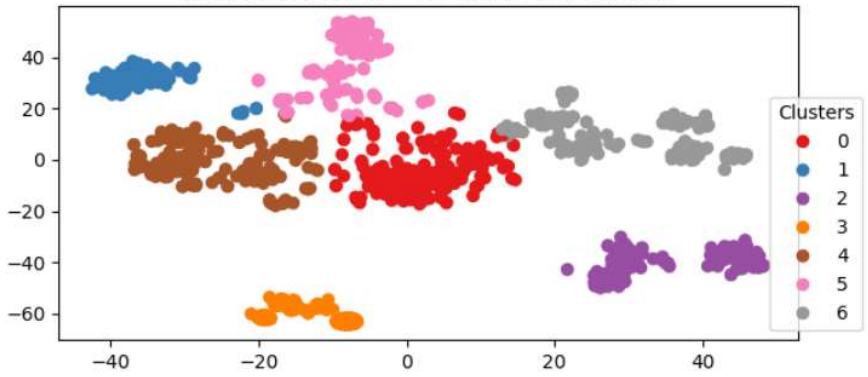
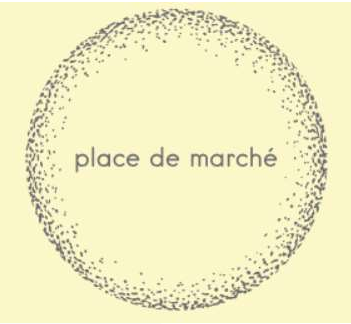
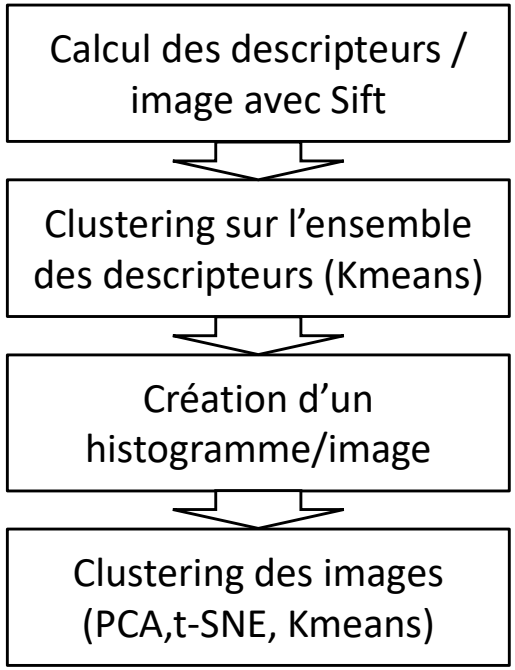


Image Proc. Clustering benchmark

Sift



Méthodologie



Clustering k	PCA N-composant	t-SNE perplexity	t-SNE N-iter	ARI
500	NA	30	5000	0.059
idem	349	idem	idem	0.051
1000	593	Idem	Idem	0.041
1000	NA	idem	Idem	**0.083**
1000	NA	50	Idem	0.062
1500	NA	30	Idem	0.05
1500	723	30	Idem	0.065
900	NA	30	Idem	0.072

Image Proc. Clustering benchmark

CNN : VGG16…

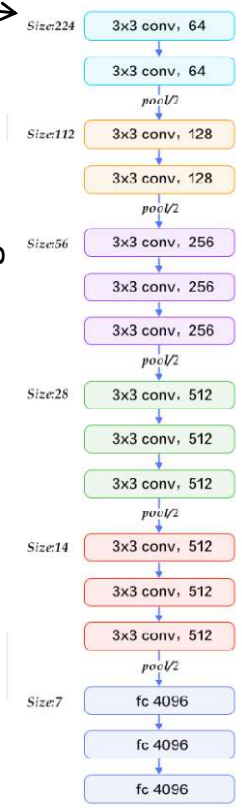
1050 photos RGB avec différentes définitions

Crop & resize

1050 photos RGB 256x256

inférence

Pré-Entrainé sur ImageNet



Min-Max Scaler

1050 vecteurs dim=4096

ACP

1050 vecteurs dim

t-SNE + clustering

7 clusters

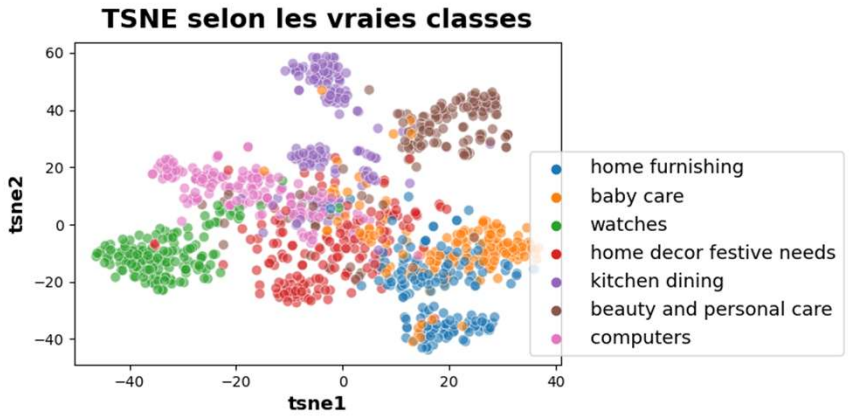
Image Proc. Clustering benchmark

CNN: VGG16

Sans tuning des paramètres du CNN

N_components PCA	ARI*
770	0.47
803	**0.5**
840	0.49
885	0.48
4096	0.43

**t-SNE : perplexity {30, 50}, Iterations {2000, 5000}*



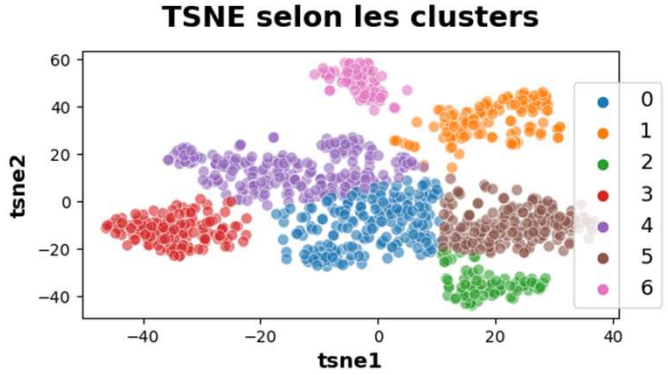
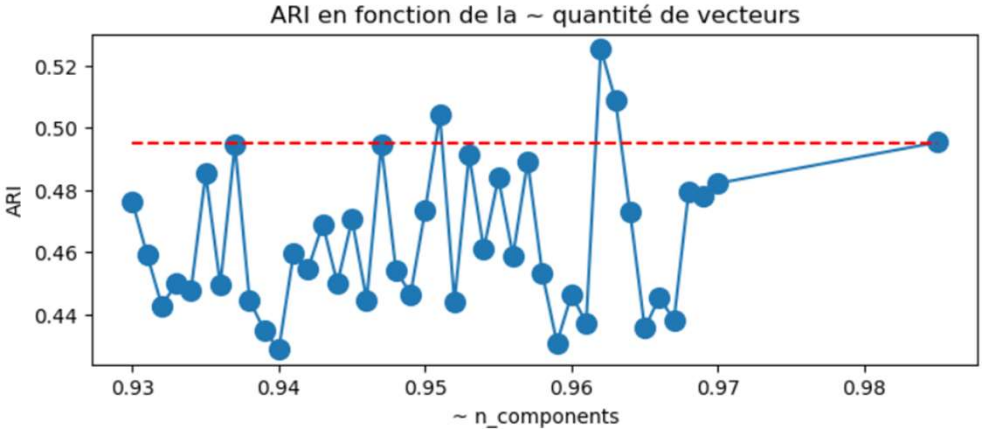


Image Proc. Clustering benchmark

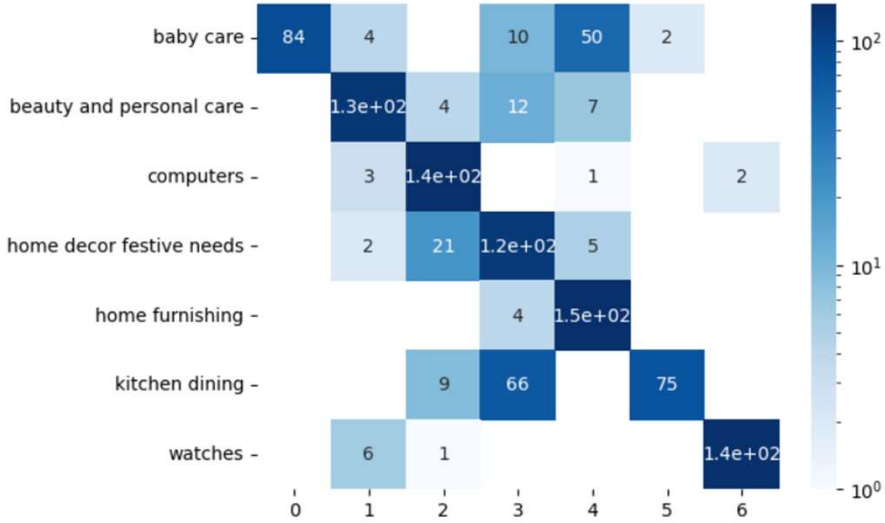
VGG16, Resnet50, MobileNetV2*



Explained variance PCA	ARI*
96.2%	**0.53**

Sans réapprentissage du CNN

Matrice de confusion

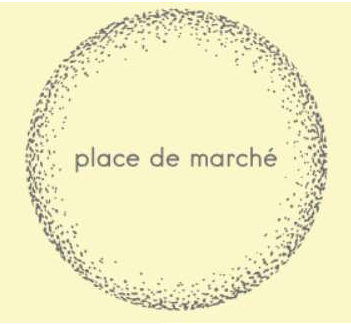


accuracy = 0.80

* **Resnet50 (ARI=0.55)** et **MobileNetV2 (ARI=0.56)** testés *a posteriori* surpassent VGG16 en rapidité de calcul et en précision.

Fusion des modèles

5 stratégies testées



Stratégie 1 : ACP

Features_USE 1050x200	Features_ctf 1050x955	Features_VGG16 1050x556

1050x1711

ACP

Features 1050x21

t-SNE + clustering

ARI :0.53

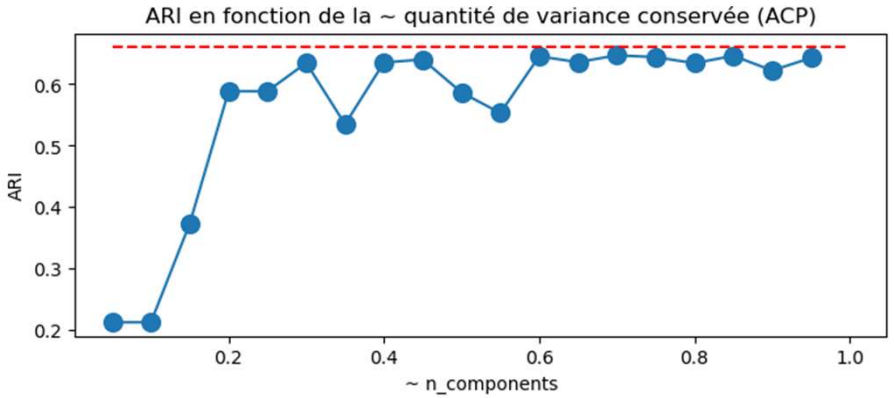
Même méthode

Stratégie 1B : Normalisation minmaxscaler avant ACP

Features_USE 1050x200	Features_ctf 1050x955	Features_VGG16 1050x**21**

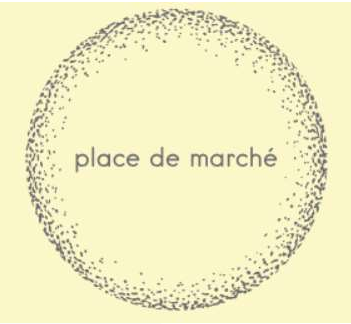
1050x1176

ACP…

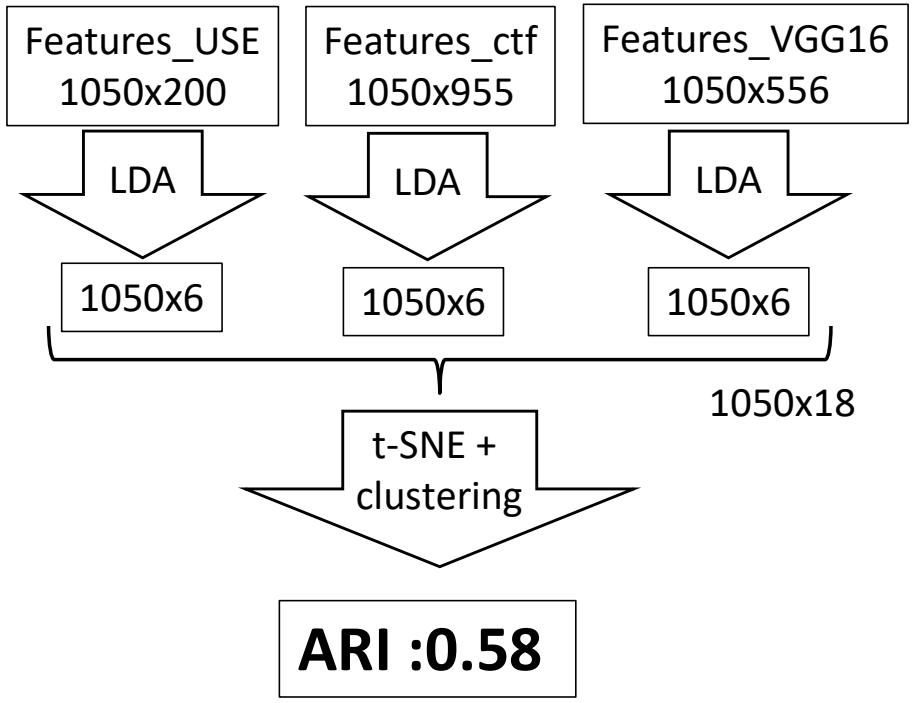


Fusion des modèles

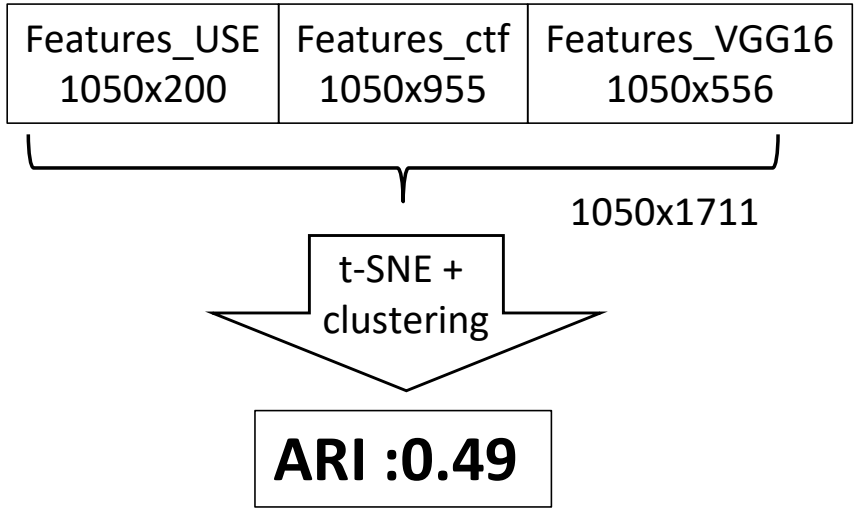
5 stratégies testées



Stratégie 2 : LDA

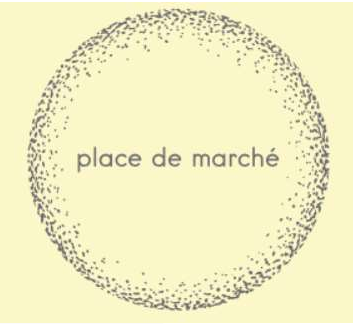


Stratégie 3 : brut

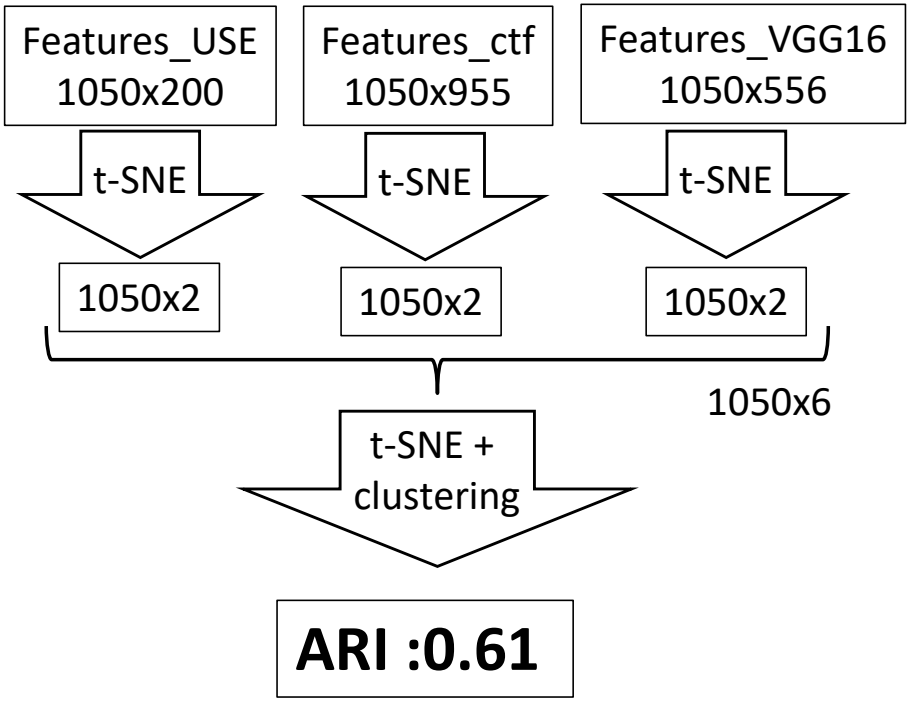


Fusion des modèles

5 stratégies testées

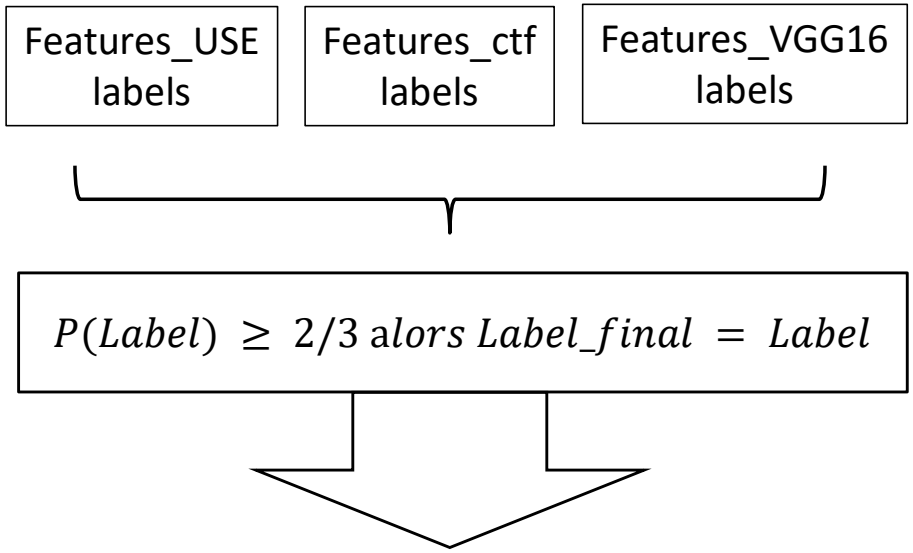


Stratégie 4 : t-SNE



ARI :0.61

Stratégie 5 - probabiliste



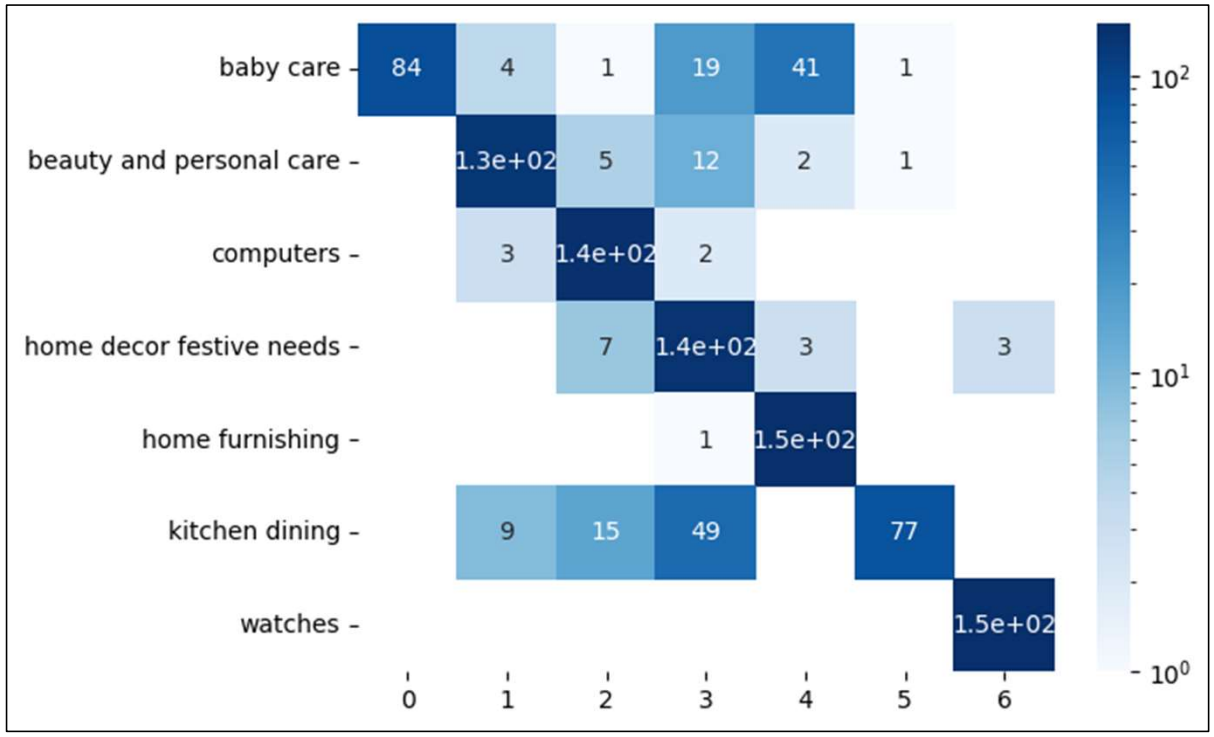
ARI (label_final, classe): 0.67

Fusion des modèles

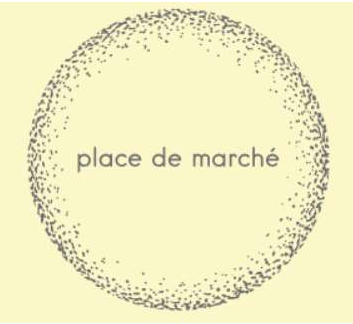
Stratégie probabiliste

ARI = 0.67

Accuracy = 0.83



Conclusion



- Différentes modèle NLP et Traitement d'images ont été testées:
  - Tf-idf, countervectorizer, Lattent Dirichlet Allocation, Word2Vec, Transformer: BERT, USE
  - Sift, CNN : VGG16
- Le traitement le plus efficace est Tf-idf sur Description et Spécifications
- Sans classification : clustering des différents articles avec une ARI de 67% -
  - utilisation de LDA permet de créer des clusters et d'obtenir un ARI de 100%
- Sur les données fournis il est donc possible de réaliser une classification de tous les articles.