



# Classifiez automatiquement des biens de consommation

PROJET 6

EMANUELE PARTENZA

# Etude de faisabilité d'un moteur de classification

Texte :

---

CountVectorizer

---

TfidfVectorizer

---

Word2Vec

---

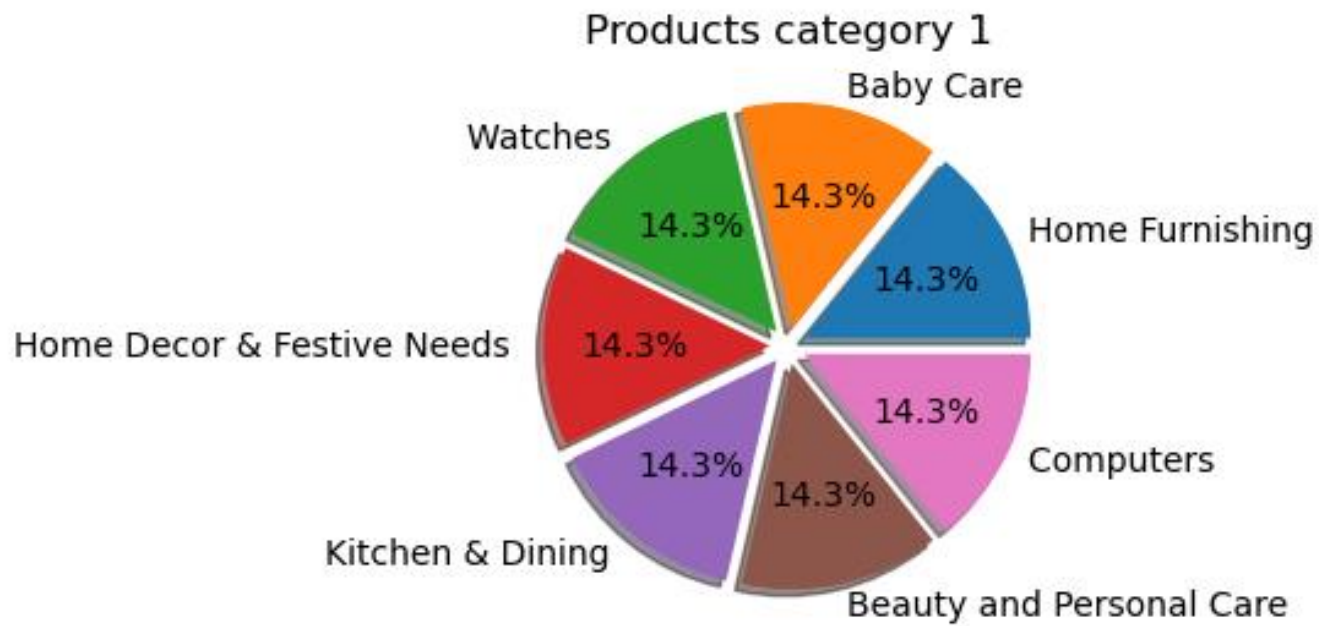
BERT

---

USE

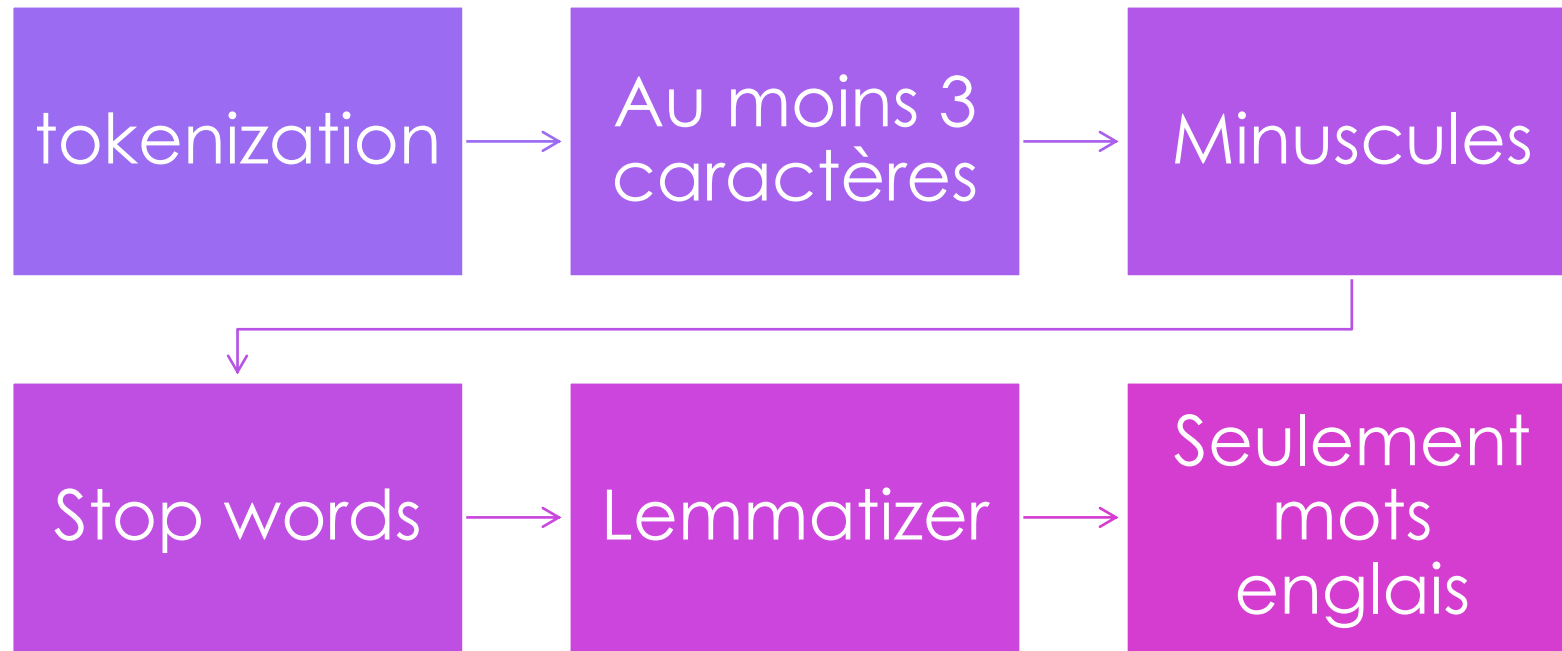
Image :

- ▶ ORB
- ▶ ResNet50
- ▶ VGG16



# Data Frame

# Nettoyage du texte



# Bag of words

## CountVectorizer

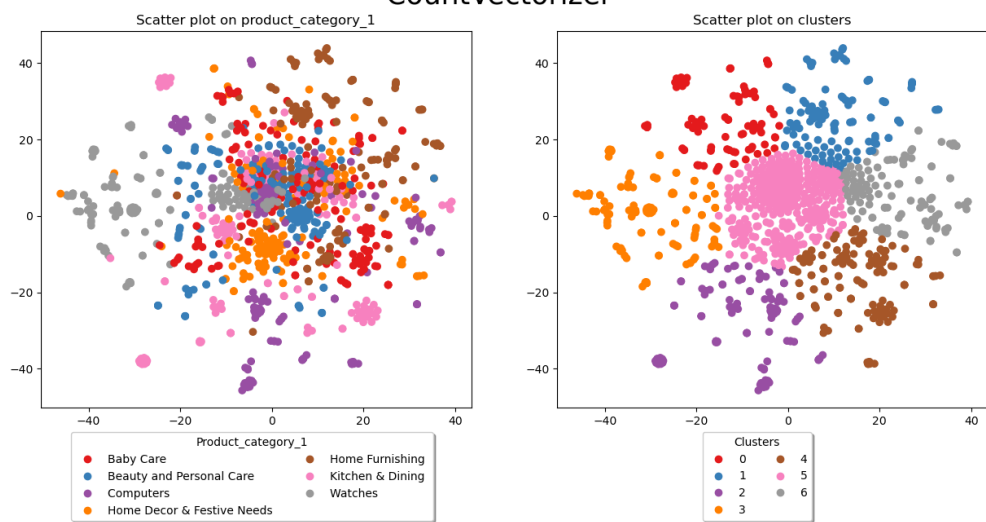
Transforme le corpus en une matrice de comptage des mots

## TfidfVectorizer

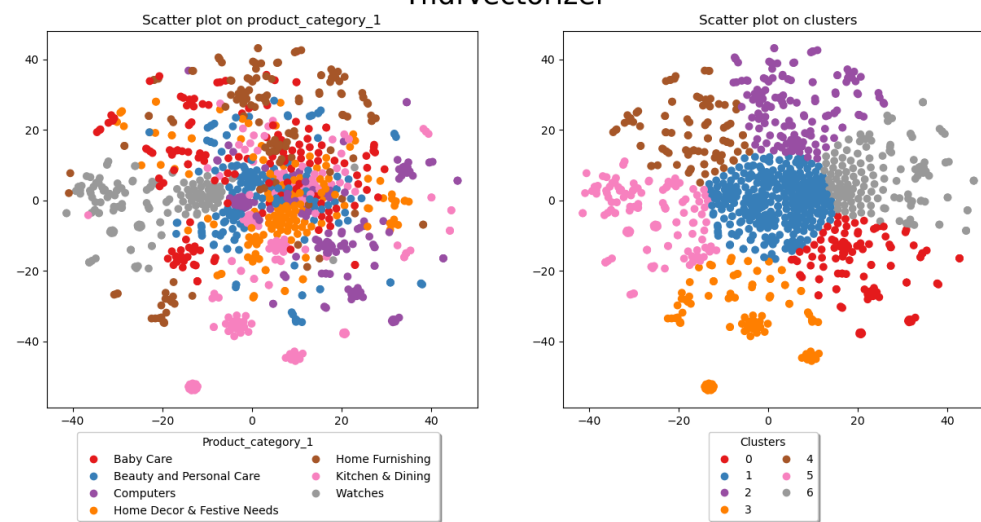
Transforme le corpus en une matrice dans laquelle à chaque mot est associé, non plus sa fréquence, mais un poids.

Ce poids correspond à la fréquence fois l'indicateur idf (inverse document frequency).

## CountVectorizer



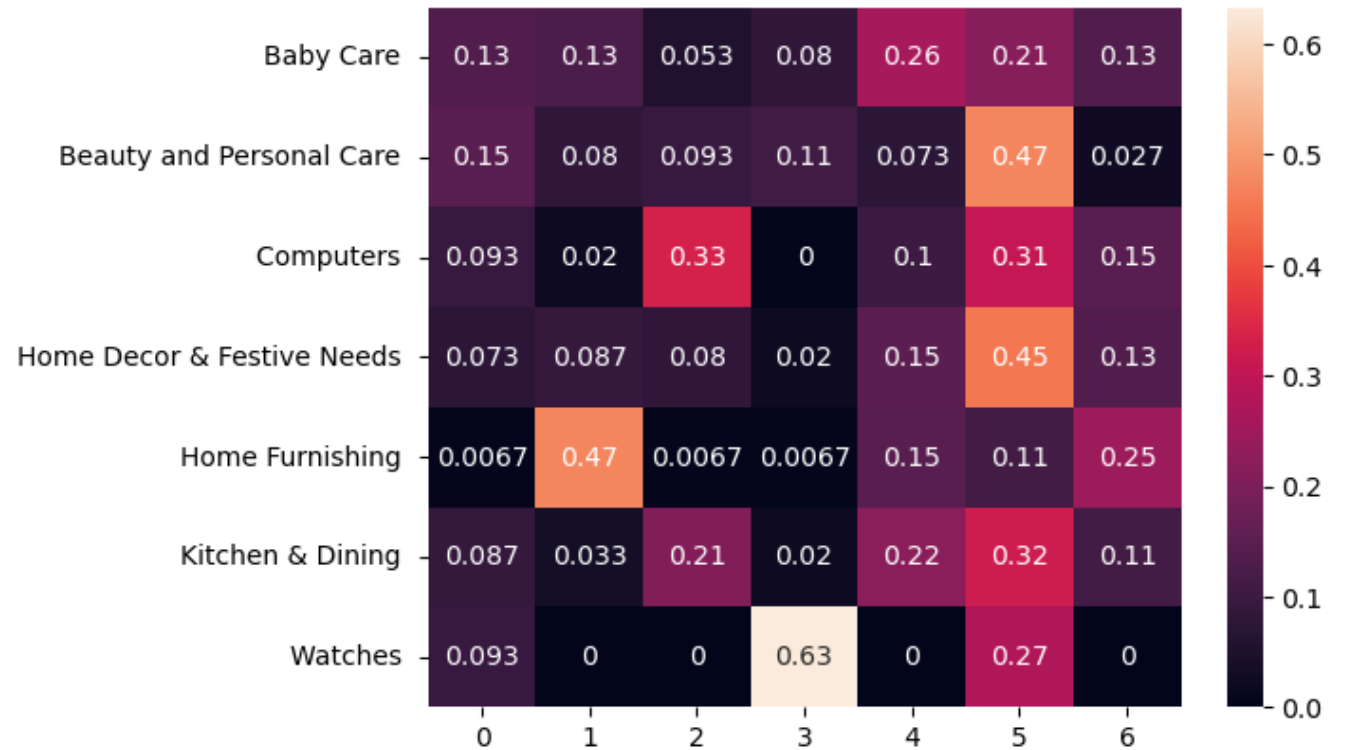
## TfidfVectorizer



# CountVectorizer

SCORES :

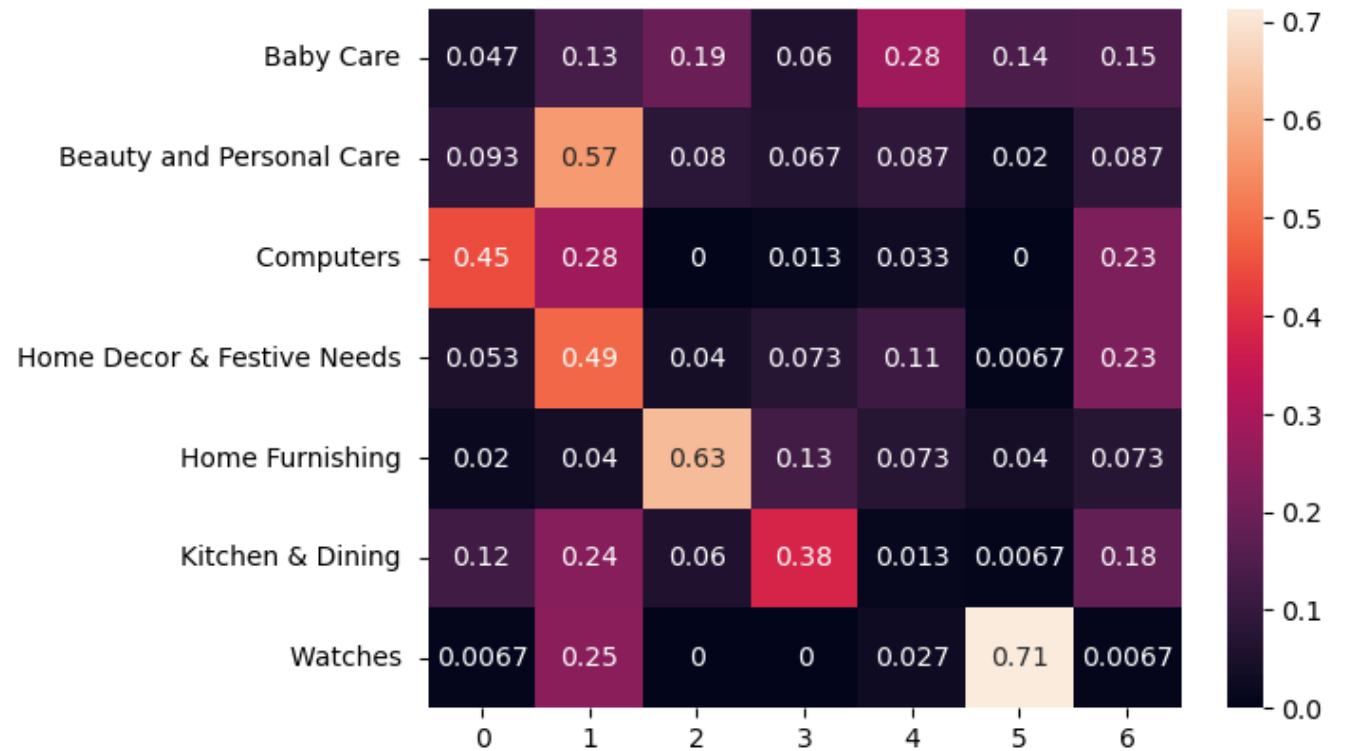
- ▶ ARI : 0.110
- ▶ F1 : 0.143
- ▶ Precision : 0.152
- ▶ Recall : 0.147



# TfidfVectorizer

SCORES :

- ▶ ARI : 0.188
- ▶ F1 : 0.088
- ▶ Precision : 0.082
- ▶ Recall : 0.110





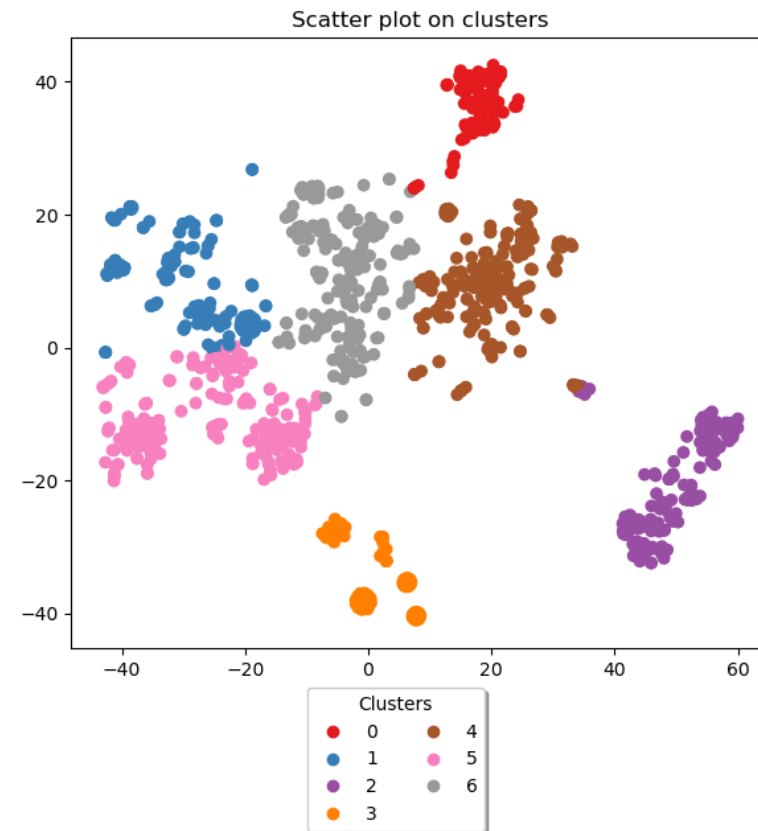
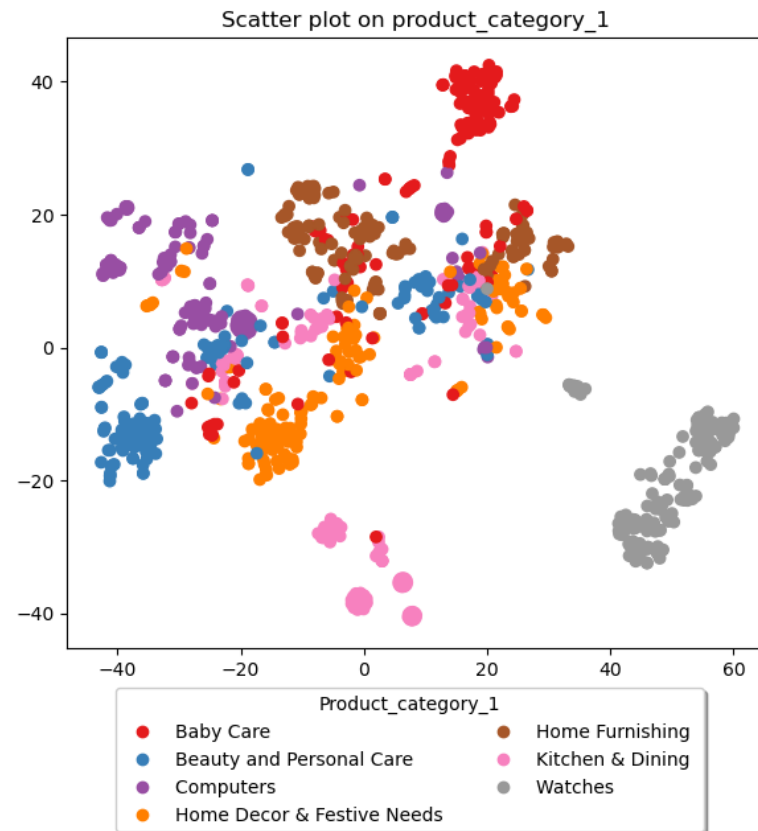
# Word2Vec

Ce modèle essaye de prendre en compte le contexte dans lequel le mot a été trouvé.

Il le fait en combinant deux techniques:

- ▶ CBOW (Continuous Bag of Words) qui entraîne le réseau de neurones pour prédire un mot en fonction du contexte;
- ▶ Skip-gram qui essaye de prédire le contexte en fonction du mot.

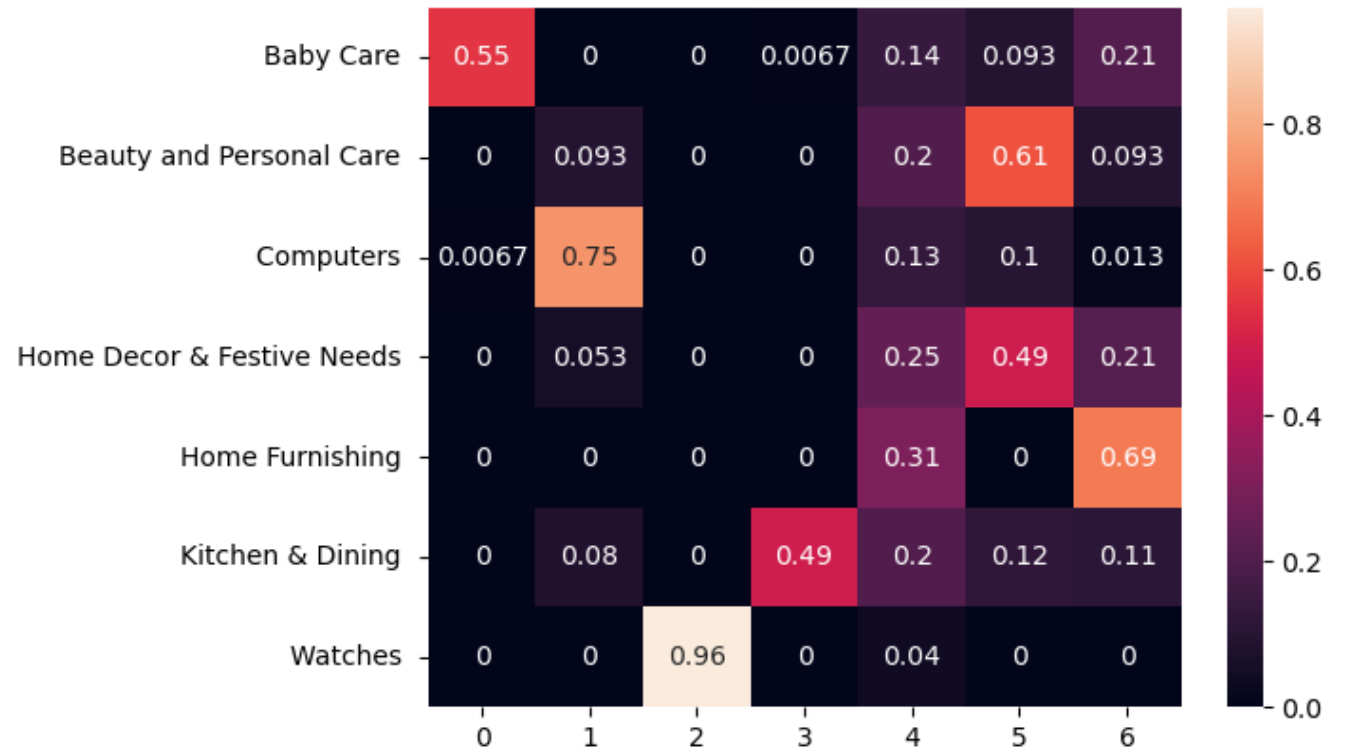
## Word2Vec



# Word2Vec

SCORES :

- ▶ ARI : 0.385
- ▶ F1 : 0.167
- ▶ Precision : 0.201
- ▶ Recall : 0.153

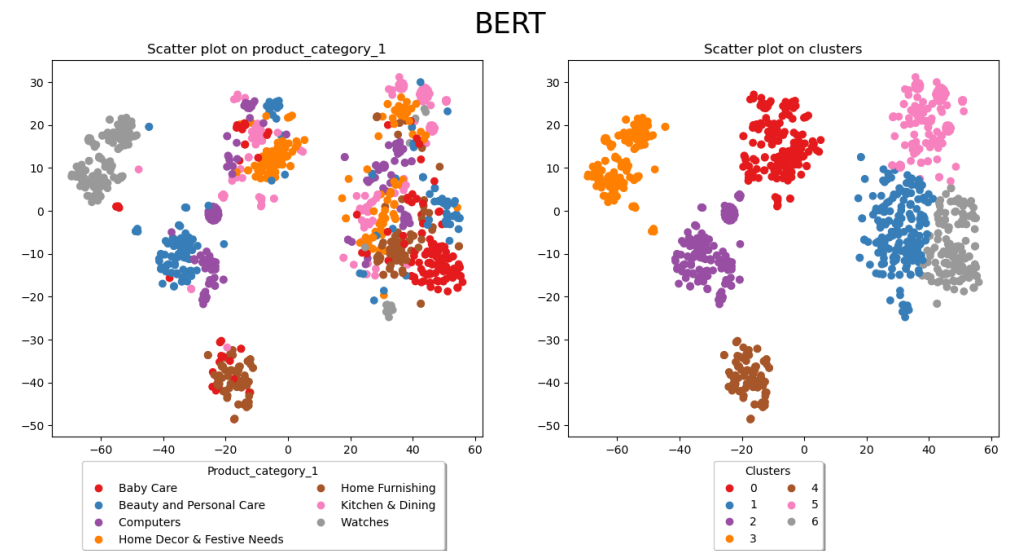
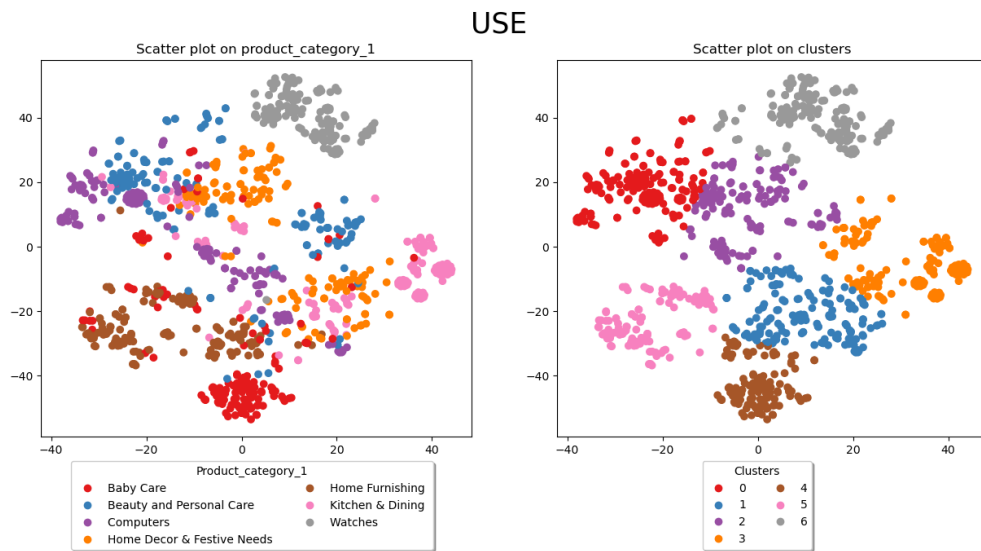


# BERT & USE

BERT (Bidirectional  
Encored  
Représentations  
from Transformers)

USE (Universal  
Sentence  
Encoder)

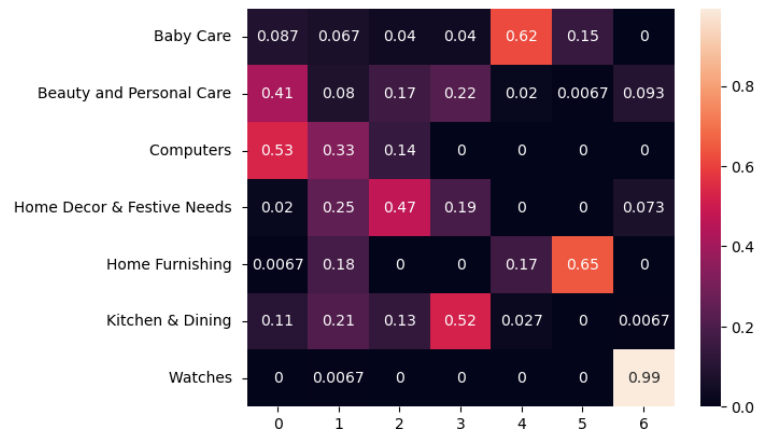
Ce sont deux  
modèles pré-  
entraînés



# Scores

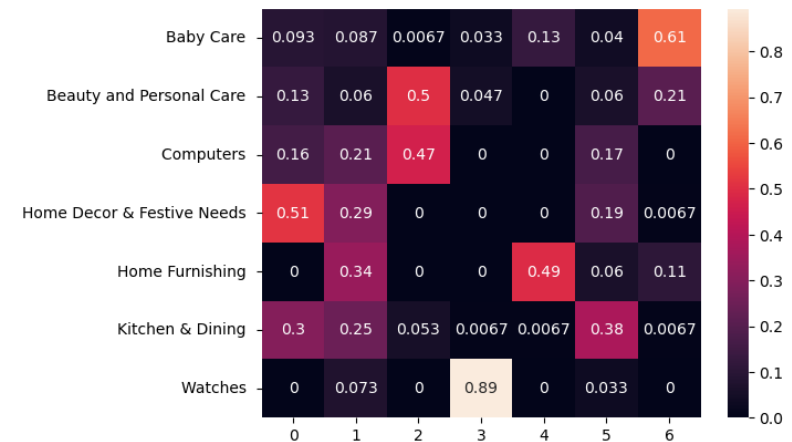
## USE

- ▶ ARI : 0.361
- ▶ F1 : 0.226
- ▶ Precision : 0.219
- ▶ Recall : 0.236



## BERT

- ▶ ARI : 0.301
- ▶ F1 : 0.228
- ▶ Precision : 0.253
- ▶ Recall : 0.213

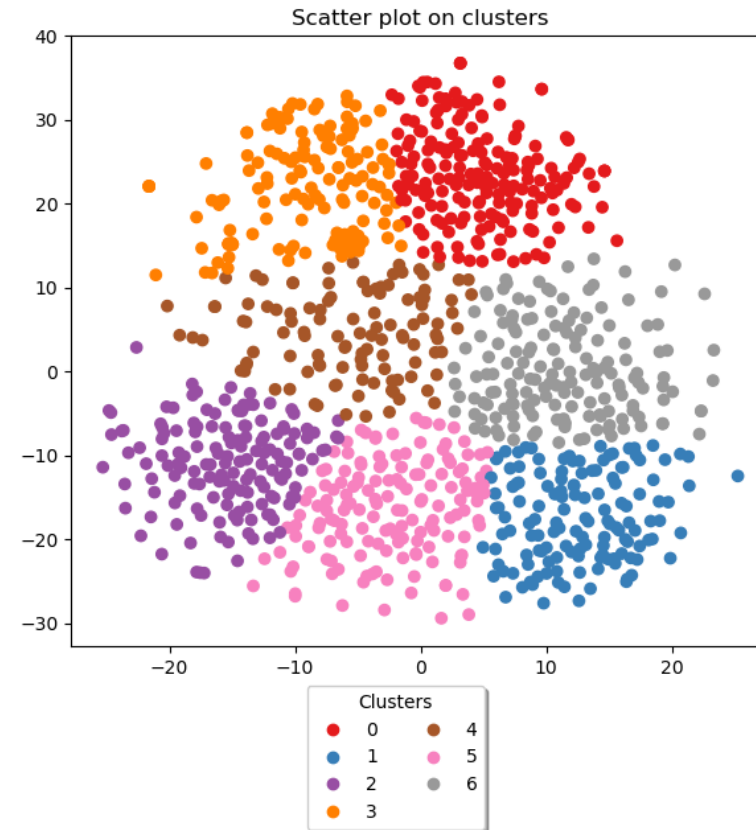
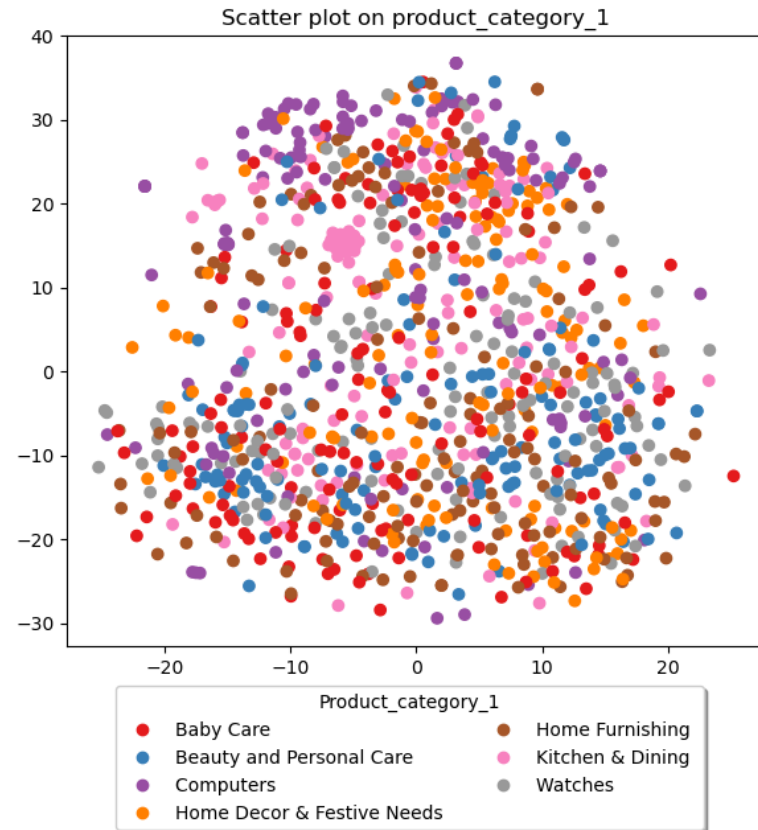


# ORB

L'algorithme ORB (Oriented FAST and Rotated BRIEF) nous permet d'extraire des features des images et de calculer leurs descripteurs.

Cette méthode est la fusion de FAST, pour la détection des keypoints, et de BRIEF pour les descripteurs.

## ORB

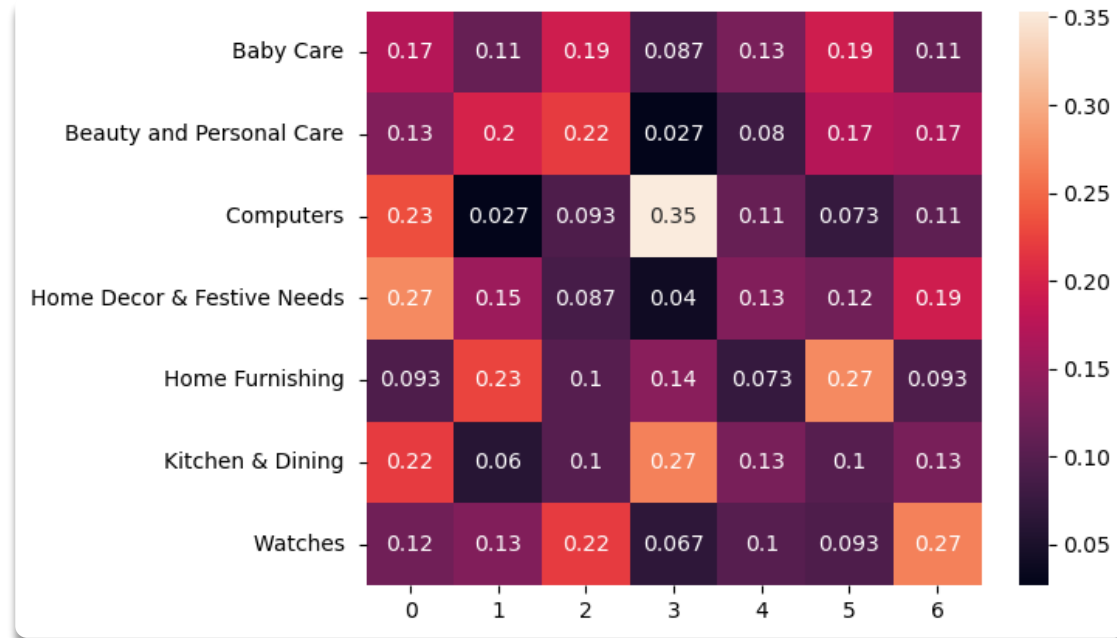




# ORB

## SCORES :

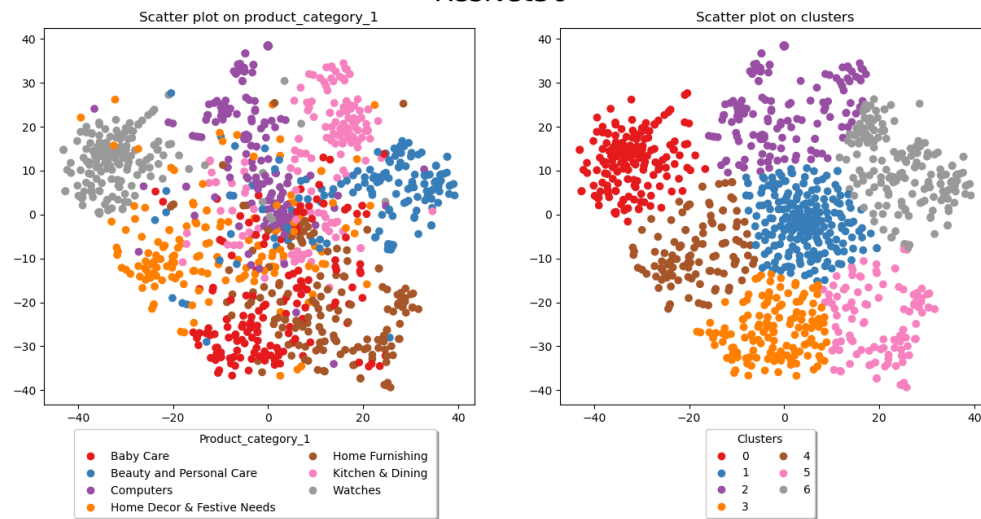
- ▶ ARI : 0.032
- ▶ F1 : 0.133
- ▶ Precision : 0.133
- ▶ Recall : 0.135



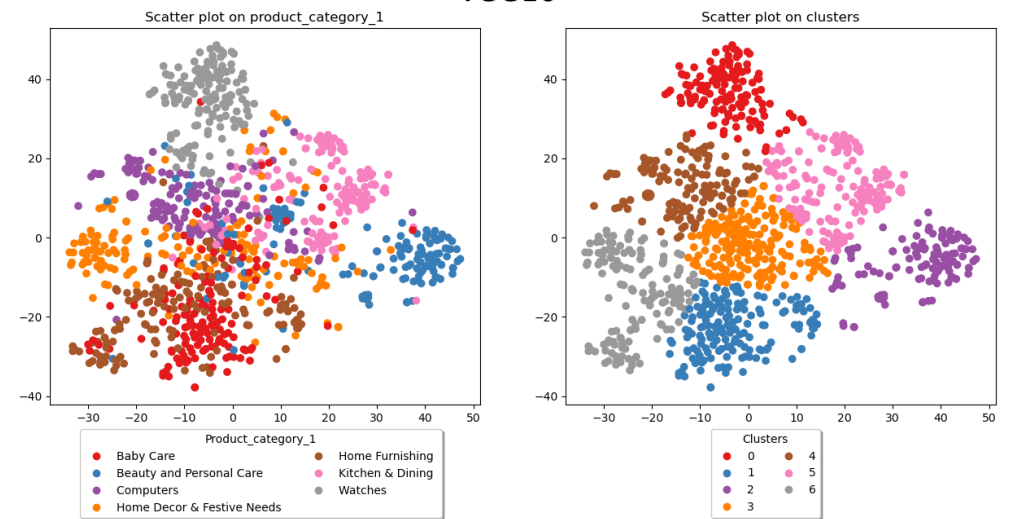
# ResNet50 & VGG16

ResNet50 (deep Residual Networks) et VGG16 (Visual Geometry Group) sont deux modèles pré-entraînés utilisant un réseau de neurones convolutif (CNN)

ResNet50



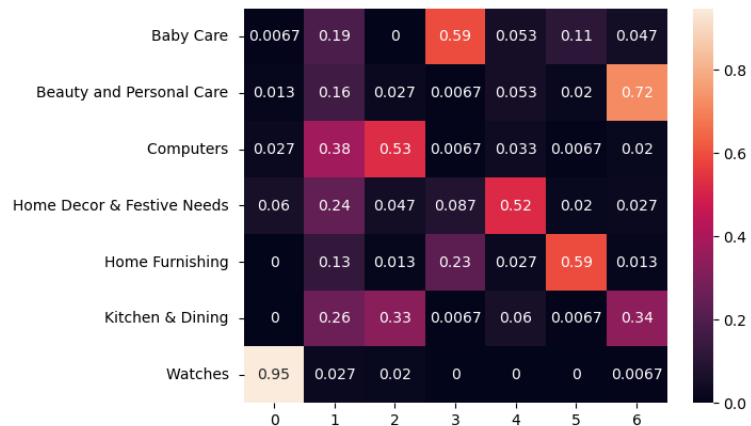
VGG16



# Scores

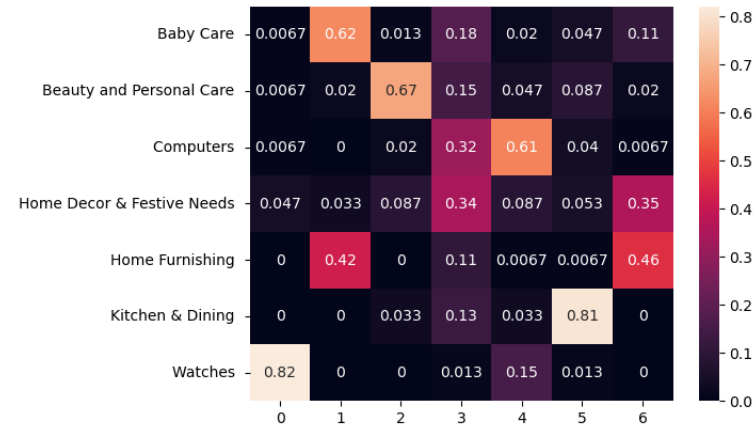
## ResNet50

- ▶ ARI : 0.371
- ▶ F1 : 0.116
- ▶ Precision : 0.116
- ▶ Recall : 0.117



## VGG16

- ▶ ARI : 0.394
- ▶ F1 : 0.164
- ▶ Precision : 0.257
- ▶ Recall : 0.171



Merci pour votre  
attention

---