OPENCLASSROOMS


# P6 : Classer automatiquement des biens de consommation

13/07/2022

DUBART Maxime

# Classification automatique de biens de conso.

OBJETIFS

**Etude de faisabilité sur la classification automatique d’articles**

Fondée sur les descriptions / images des articles (non-supervisé)

# Classification automatique de biens de conso.

DONNEES

**Données pour 1050 articles**

Colonnes d’intérêt :

- Nom de produit
- Description
- Catégories (I et II)
- Lien vers images associées

# Classification automatique de biens de conso.

DONNEES

## Données pour 1050 articles

Colonnes d'intérêt :

- Nom de produit
- Description
- Catégories (I et II) → Conversion de l'arbre en deux nouvelles colonnes
  - Catégorie principale (I)
  - Catégorie secondaire (II)
- Lien vers images associées

# Classification automatique de biens de conso.

DONNEES

## Données pour 1050 articles

Colonnes d'intérêt :

- Nom de produit
- Description
- Catégories (I et II)
- Lien vers images associées

Prétraitement pour NLP

Conversion de l'arbre en deux nouvelles colonnes
- Catégorie principale (I)
- Catégorie secondaire (II)

# Classification des descriptions

# Classification automatique de biens de conso.

METHODES

## Classification des descriptions

### #1 : bag of words (tf, tf-idf)

Fondé sur les fréquences uniquement (contexte non considéré), out of vocab. words (oov), large sparse matrix representation, polysemy

### #2 : words embedding (w2v, fasttext, glove)

Fondé en partie sur le contexte, gestion des oov (fasttext), dense matrix representation, polysemy, texte représentation : moyenne des mots

### #3 : word/sentence embedding (BERT, USE) – transformers based

Fondé sur le contexte (attention), gestion des oov (n-gram based), dense matrix representation, polysemy, texte représentation : moyenne, [CLS], USE vector

# Classification automatique de biens de conso.

METHODES

## Classification des descriptions

**#1 : bag of words (tf, tf-idf)**

Fondé sur les fréquences uniquement (contexte non considéré), out of vocab. words (oov), large sparse matrix representation, polysemy

**#2 : words embedding (w2v, fasttext, glove)**

Fondé en partie sur le contexte, gestion des oov (fasttext), dense matrix representation, polysemy, texte représentation : moyenne des mots

**Préprocessing #1**

**#3 : word/sentence embedding (BERT, USE)**

Fondé sur le contexte (attention), gestion des oov (n-gram based), dense matrix representation, polysemy, texte représentation : moyenne, [CLS], USE vector

**Préprocessing #2**

# Classification automatique de biens de conso.

METHODES

## Préprocessing #1

**Step #1 : passage en minuscules**
**Step #2 : tokenization**
**Step #3 : suppression des stops words / ponctuation**
**Step #4 : Lemmatization (passage à forme canonique)**

# Classification automatique de biens de conso.



## Préprocessing #2

**Step #1 : passage en minuscules**
**Step #2 : tokenization**
**Step #3 : suppression des stops words / ponctuation**
**Step #4 : Lemmatization (passage à forme canonique)**

# Classification automatique de biens de conso.

METHODES

## Préprocessing #2

**Step #1 : passage en minuscules**
**Step #2 : tokenization**
**Step #3 : suppression des stops words / ponctuation**
**Step #4 : Lemmatization (passage à forme canonique)**

**Méthodes d'embedding :** conversion en vecteurs d'entiers de taille fixe (padding)

# #1 bag of words (tf, tf-idf)

Tf - BoW

| | Word #1 | Word #j | … | Word #W |
|---|---|---|---|---|
| Doc. #1 | $tf_{ij}$ | | | |
| … | | | | |
| Doc. #D | | | | |

Fréquence du mot *j* dans le document *i*

# Classification automatique de biens de conso.



## #1 bag of words (tf, tf-idf)

Tf - BoW

| | Word #1 | Word #j | … | Word #W |
|---|---|---|---|---|
| Doc. #1 | $tf_{ij}$ | | | |
| … | | | | |
| Doc. #D | | | | |

Fréquence du mot *j* dans le document *i*

Tf-idf - BoW

| | Word #1 | Word #j | … | Word #W |
|---|---|---|---|---|
| Doc. #1 | $tf - idf_{ij}$ | | | |
| … | | | | |
| Doc. #D | | | | |

Fréquence du mot *j* dans le document *i* pondéré par l'inverse de la fréquence du mot dans les documents

$$\text{Tf-idf} = tf \,.\left[\log\left(\frac{D+1}{df+1}\right) + 1\right]$$

$$df_j = \sum_{i=1}^{D} tf_{ij} > 0$$

# Classification automatique de biens de conso.

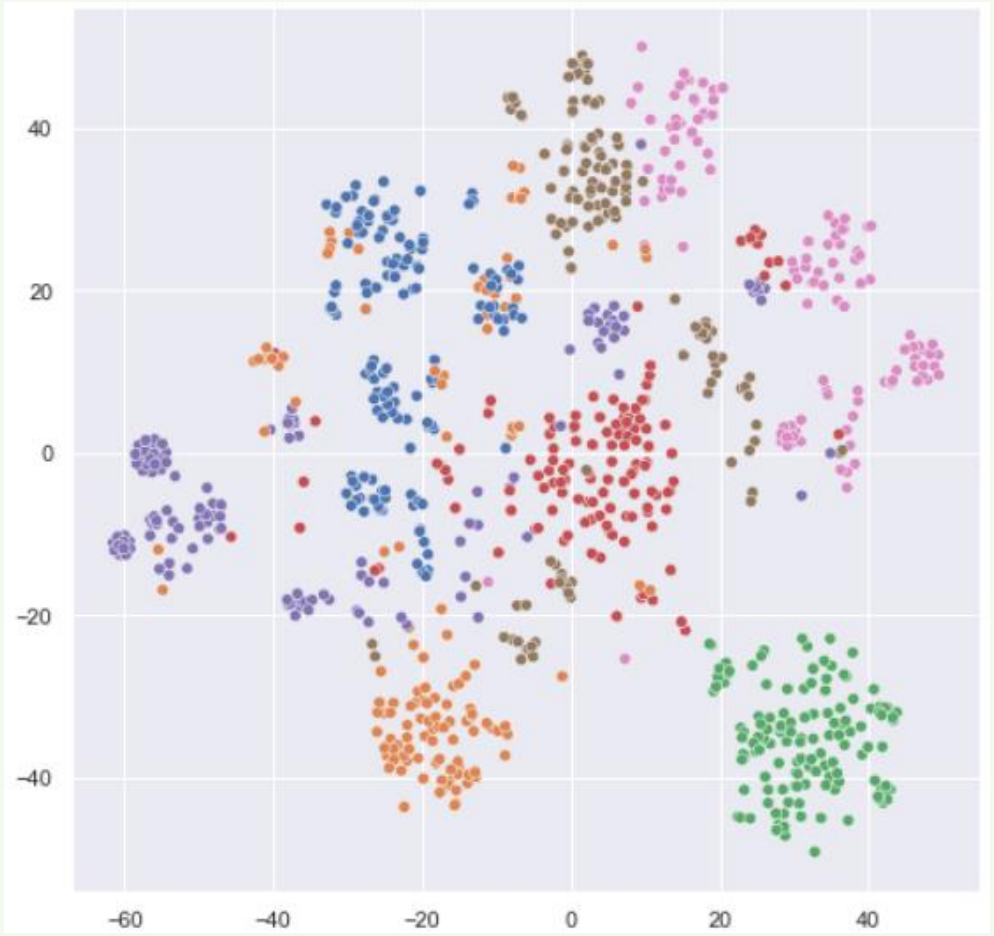
CLUSTERING - #1

t-sne projections (30)

**Terms frequencies**


**ARI = 0.40**

**Tf – inverse doc. frequency**



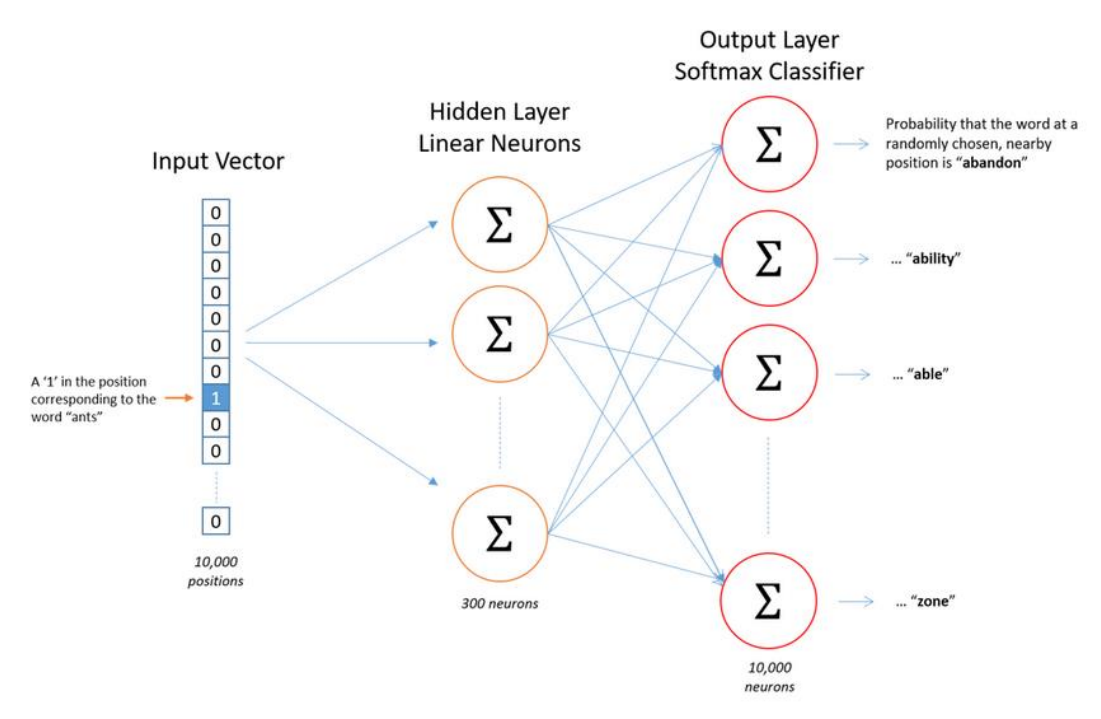

**ARI = 0.52**

# Classification automatique de biens de conso.



## #2 words embedding (w2v, fasttext, glove)

Prédire un mot à partir du contexte (cbow) ou l'inverse (skip-gram)

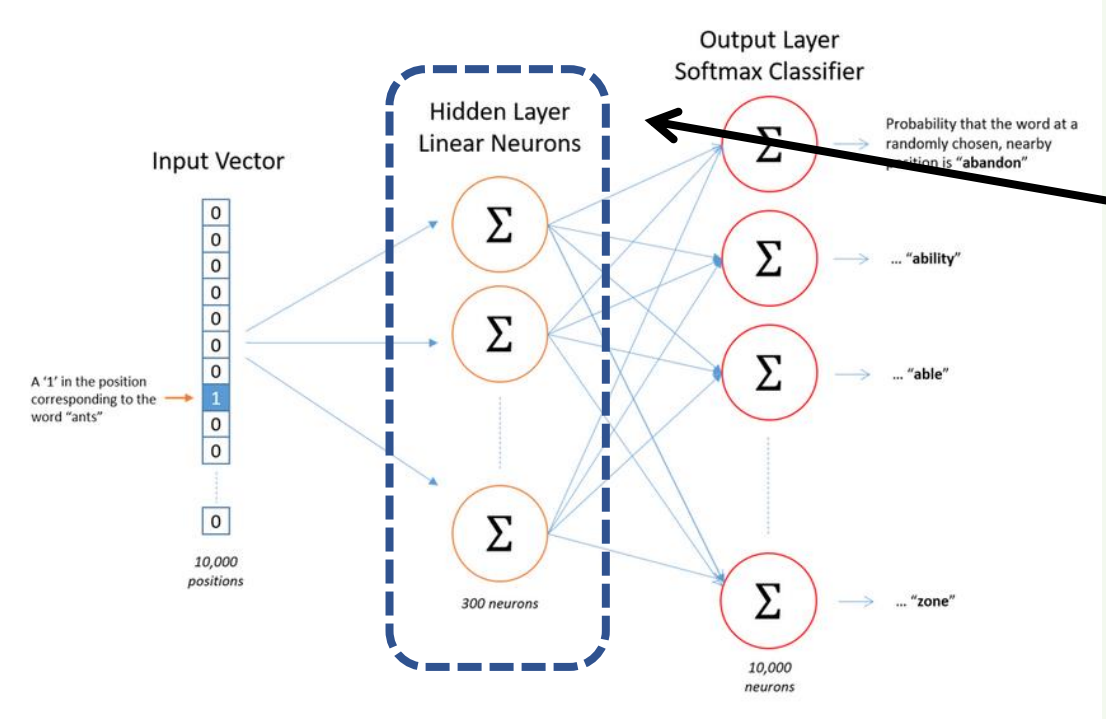



e.g. Skip-gram model, source : McCormickml tutorial

# Classification automatique de biens de conso.

## #2 words embedding (w2v, fasttext, glove)

Prédire un mot à partir du contexte (cbow) ou l'inverse (skip-gram)




Weights matrix (10k x 300) = words embeddings

Représentation du document :
moyenne des vecteurs (représentant mots composant ce document)

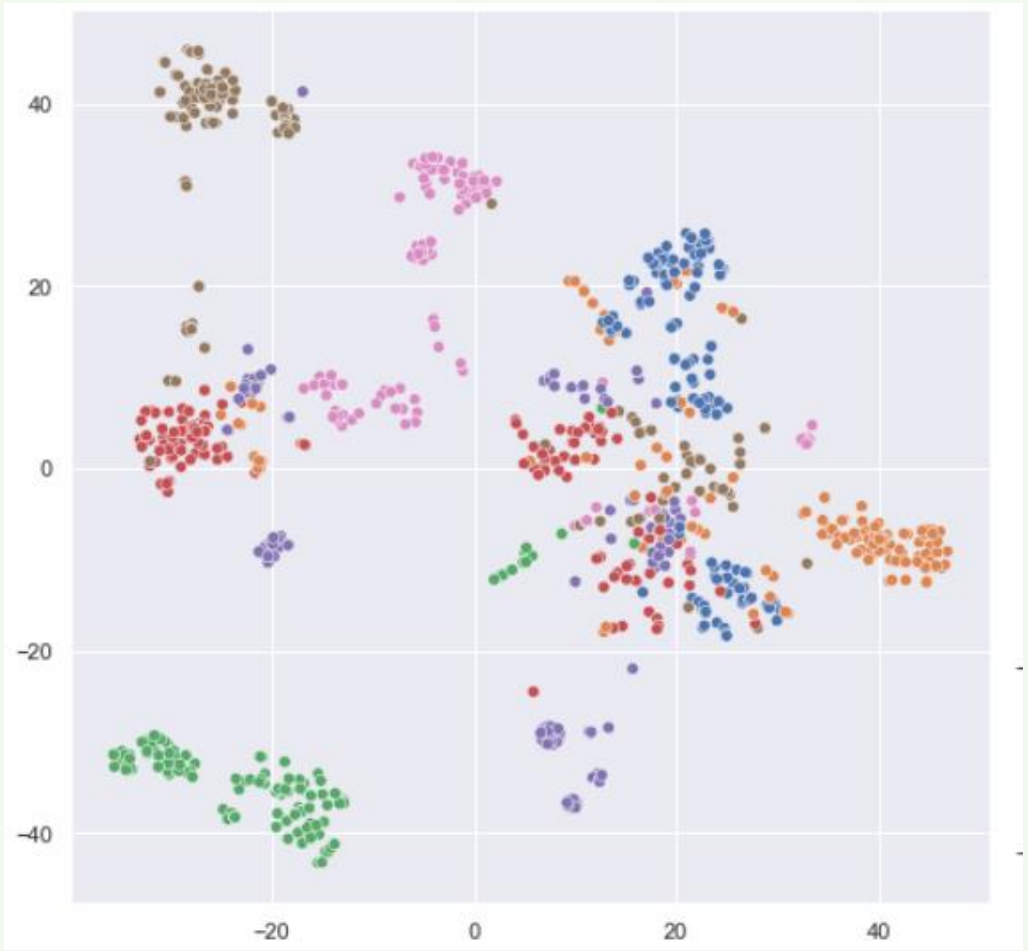
e.g. Skip-gram model, source : McCormickml tutorial

# Classification automatique de biens de conso.

CLUSTERING - #2

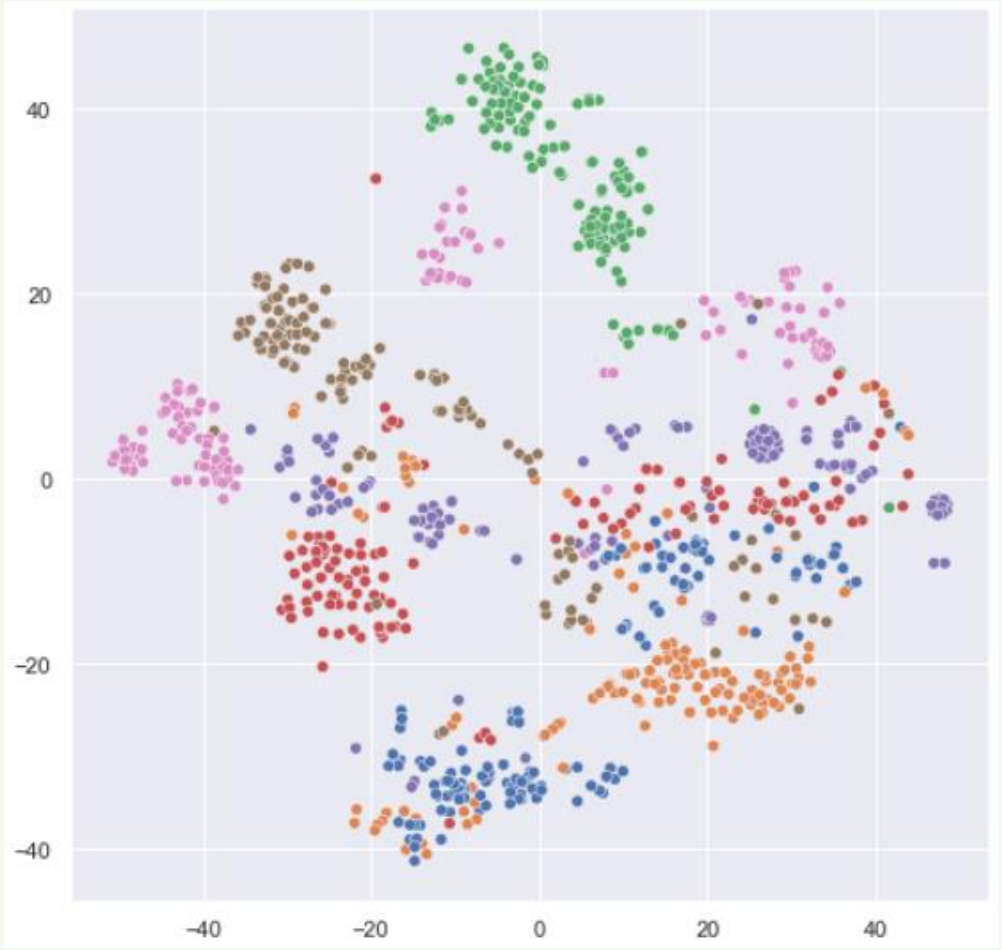
t-sne projections (30)

**w2v – trained on corpus**




**ARI = 0.35**

**w2v – google pretrained**



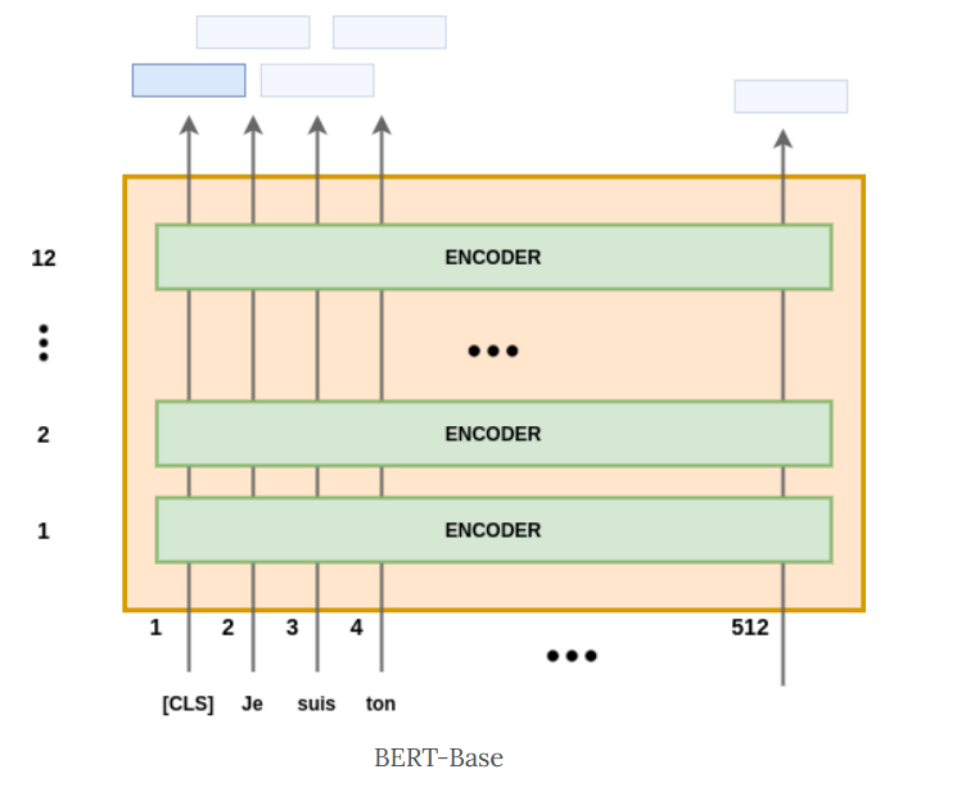

**ARI = 0.33**

# Classification automatique de biens de conso.



## #3 : word/sentence embedding (BERT, USE)

Représenter mots / phrases dans leur contexte (bidirectionnel)



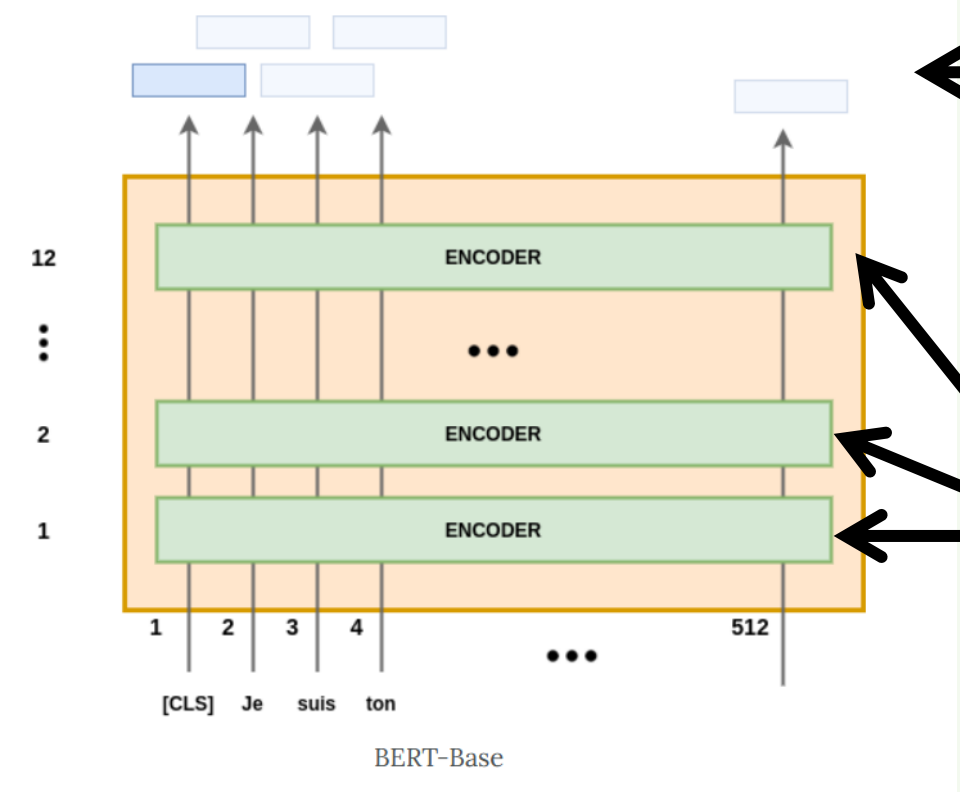

e.g. BERT encoder, source : https://ledatascientist.com/a-la-decouverte-de-bert/

# Classification automatique de biens de conso.



## #3 : word/sentence embedding (BERT, USE)

Représenter mots / phrases dans leur contexte (bidirectionnel)




max_token x 768 matrice
Embedding mots + token speciaux (e.g. [CLS])

Représentation du document:
(i) [CLS] représentation
(ii) Moyenne des représentations des mots

Contient couche d'attention : considération du contexte

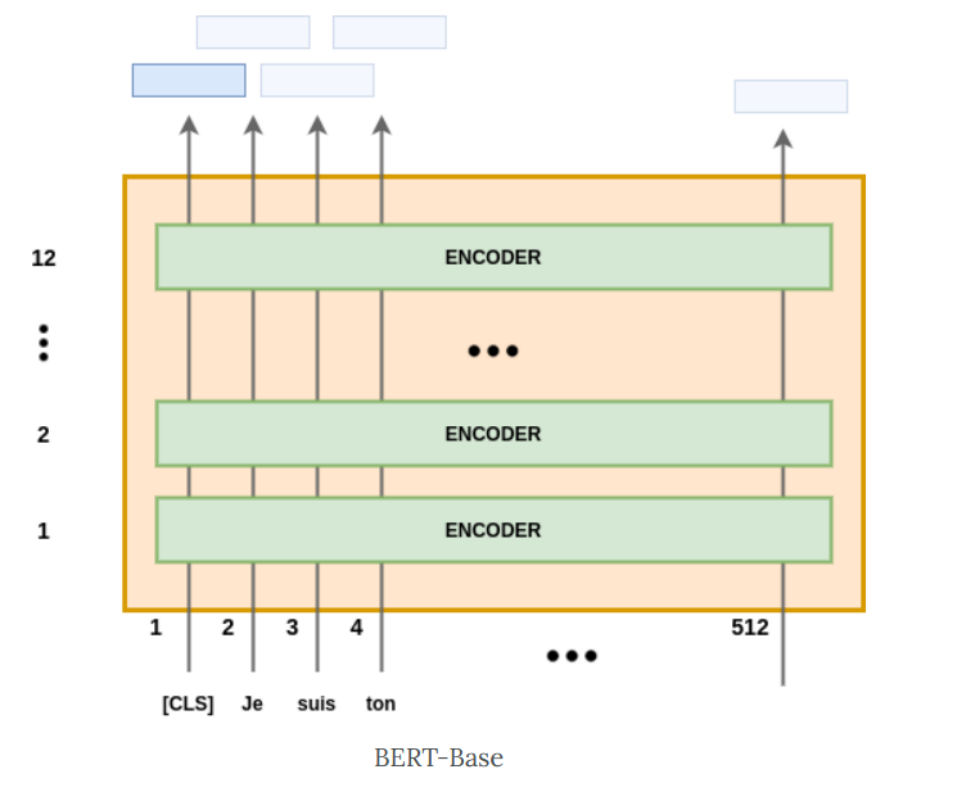
e.g. BERT encoder, source : https://ledatascientist.com/a-la-decouverte-de-bert/

# Classification automatique de biens de conso.



## #3 : word/sentence embedding (BERT, USE)

Représenter mots / phrases dans leur contexte (bidirectionnel)




Pré-entrainement :

(i) Masked Language Modeling
prédire un mot masqué

(ii) Next Sentence Prediction
prédire si une phrase est suivie par une autre

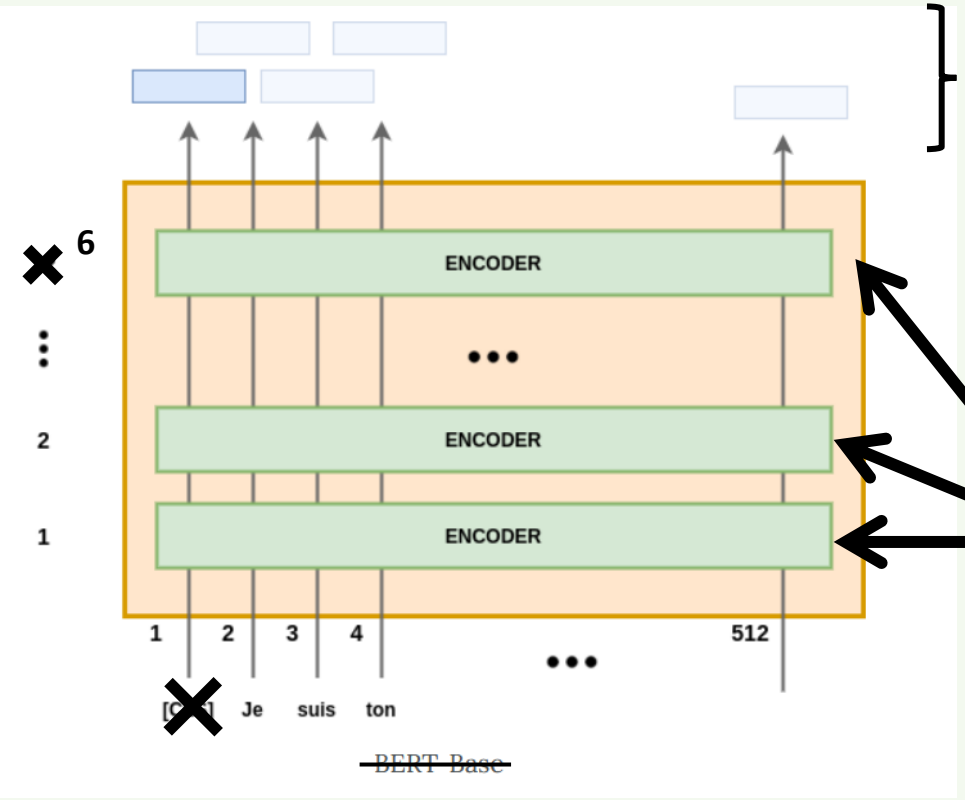
e.g. BERT encoder, source : https://ledatascientist.com/a-la-decouverte-de-bert/

# Classification automatique de biens de conso.



## #3 : word/sentence embedding (BERT, USE)

Représenter mots / phrases dans leur contexte (bidirectionnel)




Token x 512 matrice

Représentation du document:
(i) sommes sur les éléments +
normalisation par sqrt(taille)

Contient couche d'attention : considération du contexte

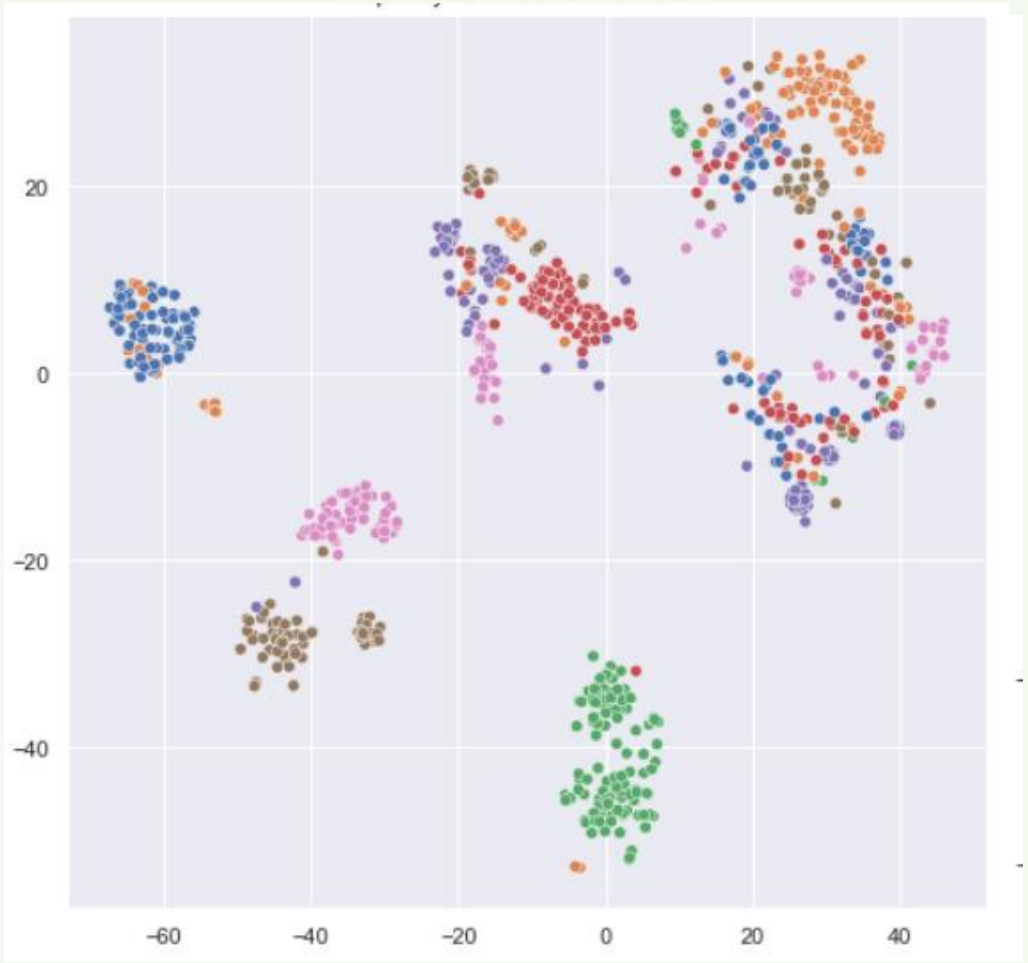
e.g. Universal Sentence Encoder, modifié depuis https://ledatascientist.com/a-la-decouverte-de-bert/

# Classification automatique de biens de conso.

CLUSTERING - #3

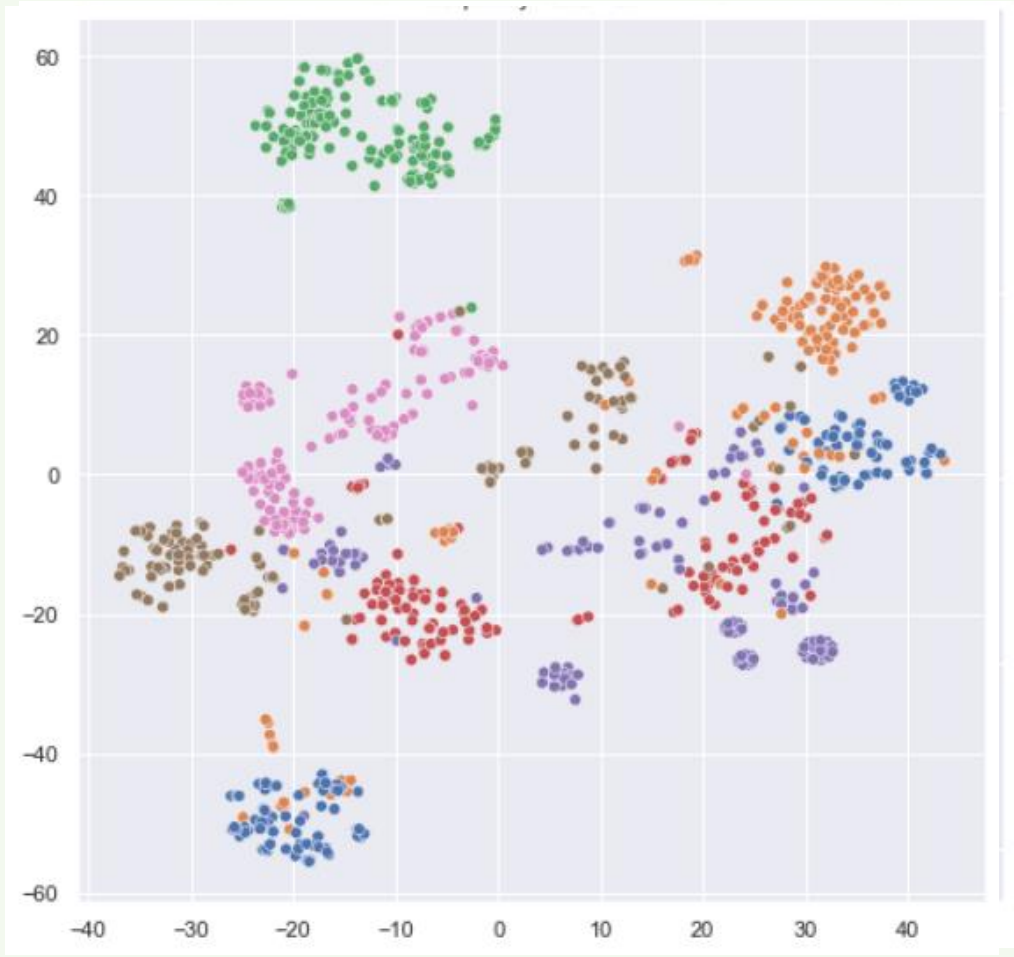
t-sne projections (30)

BERT




ARI = 0.29

USE




ARI = 0.44

# Classification automatique de biens de conso.

CONCLUSION - descriptions

| Embedding model | ARI |
|---|---|
| Tf | 0.40 |
| **Tf-Idf** | **0.52** |
| Word2Vec | 0.35 |
| Word2Vec (pretrained) | 0.33 |
| BERT | 0.29 |
| **USE** | **0.44** |

Meilleure classification obtenue avec :

**Tf-Idf**

Suivi par

**Universal Sentence Encoder**

# Classification des images

# Classification automatique de biens de conso.

METHODES

## Classification des images

### #1 : SIFT (Scale Invariant Feature Transform)

Images en niveaux de gris, utilisation du gradient d'intensité pour détecter points d'intérêt, définir leur orientation, et décrire le point d'intérêt via les orientations/intensités des groupes de pixels environnants (128 bin values) – invariant par changement d'échelle, d'orientation, de contraste et d'intensité.

### #2 : CNN (pré-entrainé VGG16)

Images en couleurs, plusieurs couches (5) de convolution + maxpooling, couches denses (3) pour classification.

# Classification automatique de biens de conso.

METHODES

## Classification des images

### #1 : SIFT (Scale Invariant Feature Transform)

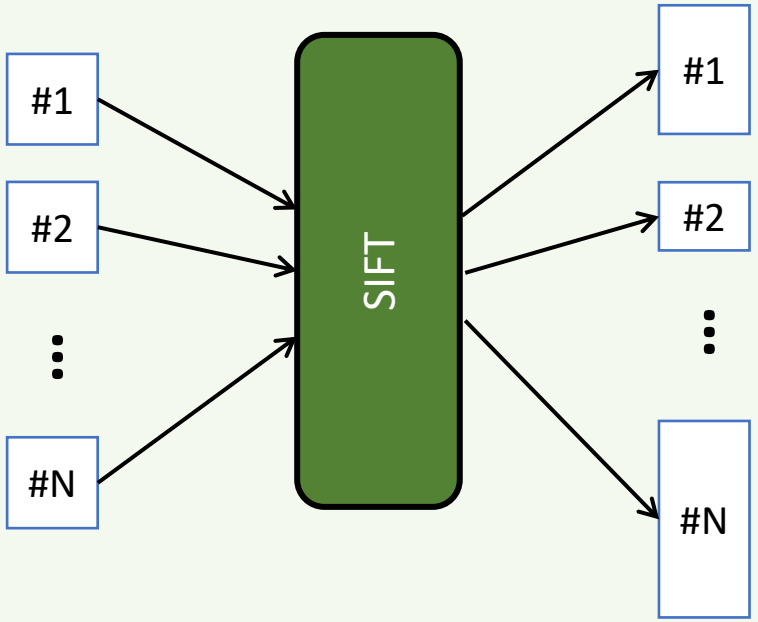
Pré-processing : transformation en niveau de gris, égalisation (correction contraste)

### #2 : CNN features-extraction (pré-entrainé VGG16)

Pré-processing : redimensionnement (224x224), puis identique à celui sur les images d'entrainement (i.e. ImageNet dataset), conversion en BGR et chaque canal est centré.

# Classification automatique de biens de conso.

## SIFT procédure



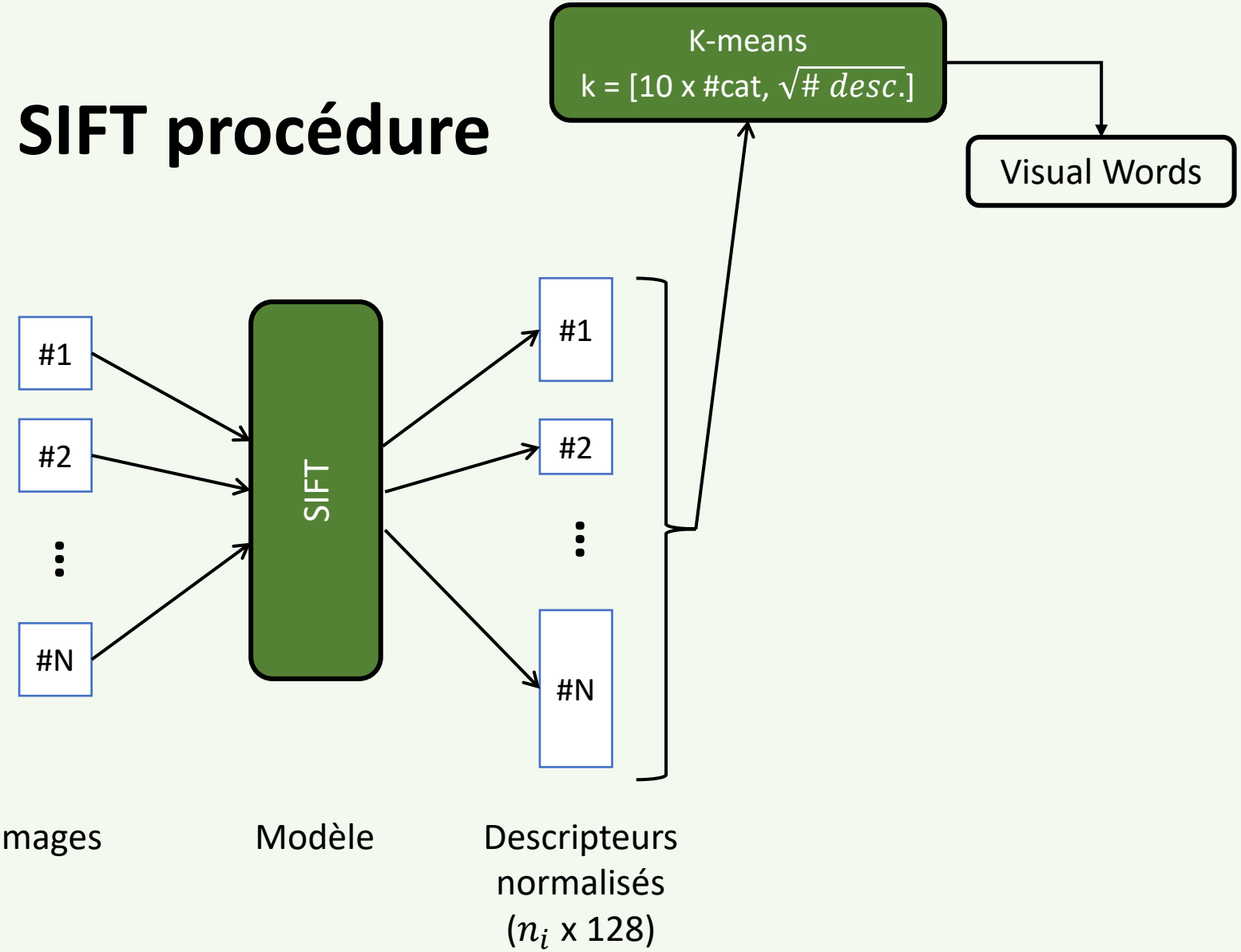

Images

Modèle

Descripteurs normalisés ($n_i$ x 128)

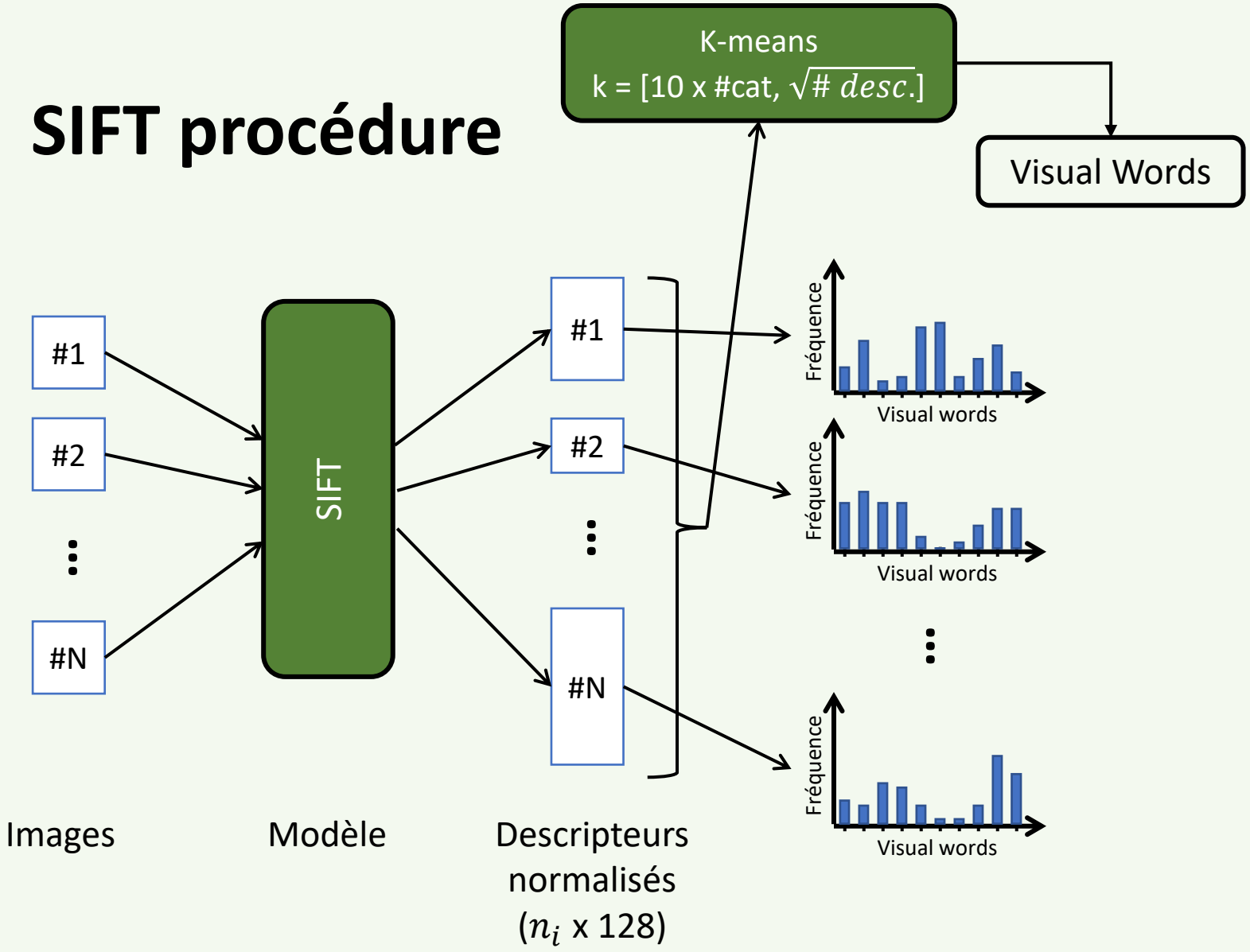
# Classification automatique de biens de conso.

## SIFT procédure



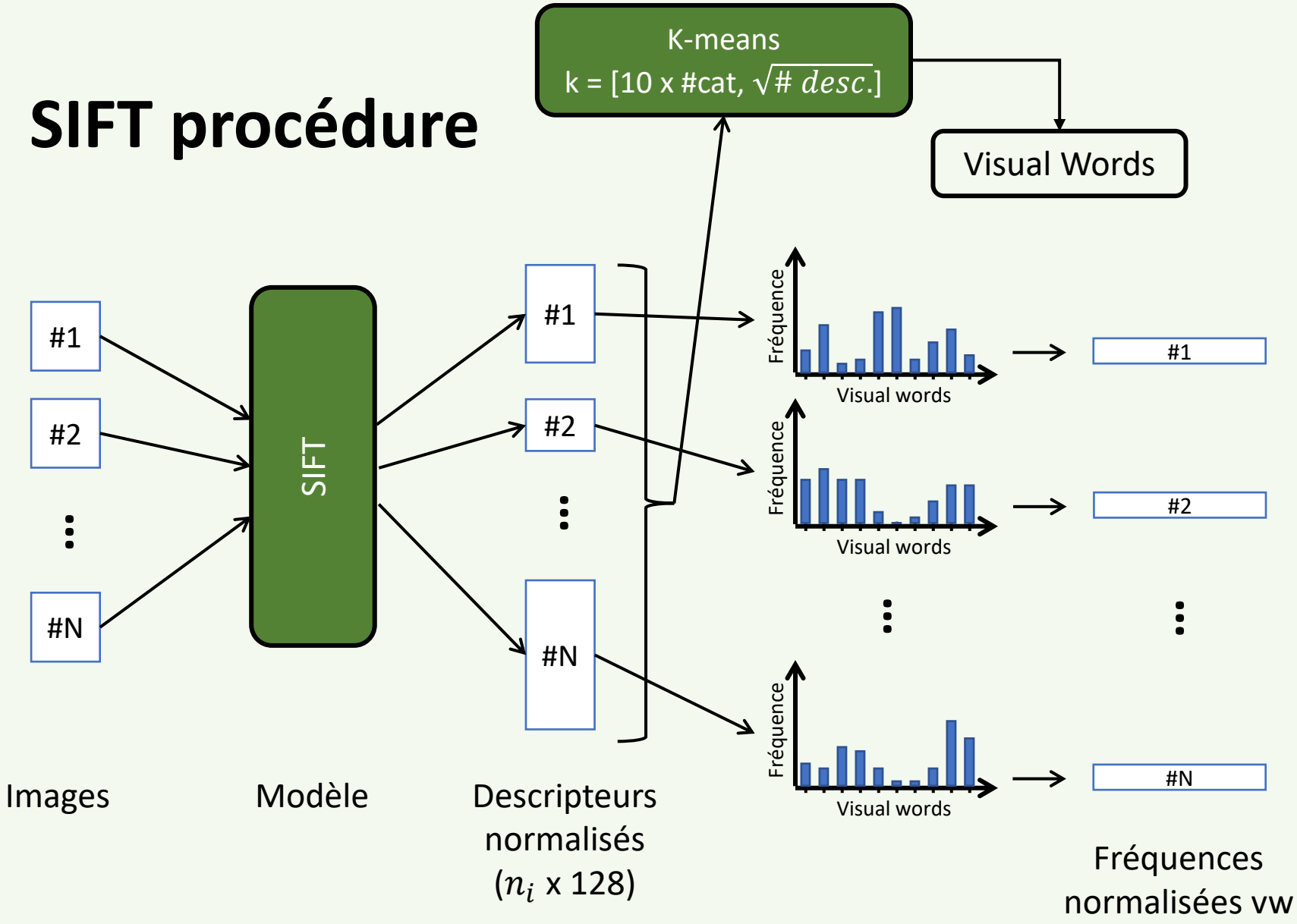
# Classification automatique de biens de conso.

## SIFT procédure



# Classification automatique de biens de conso.

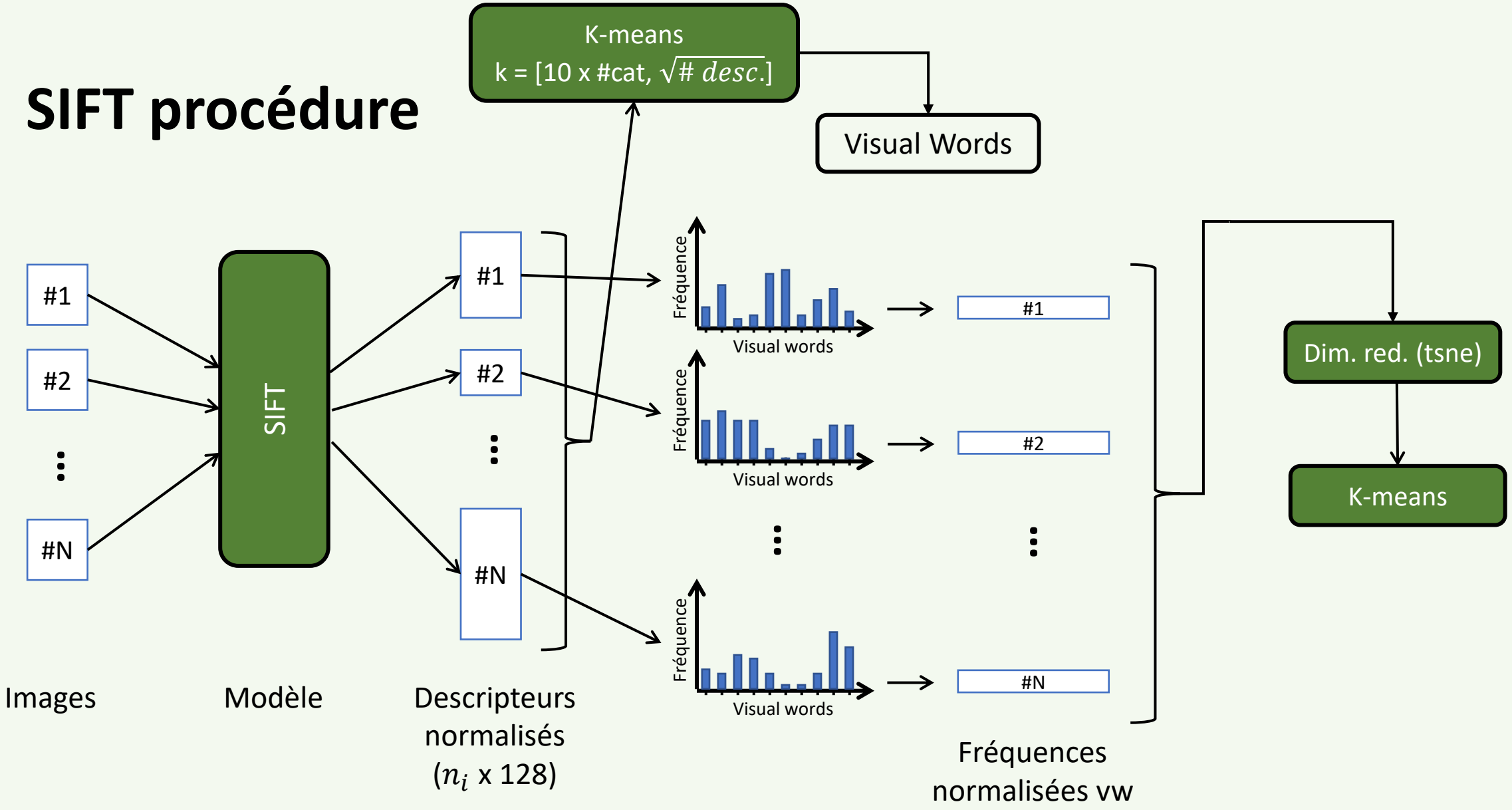
## SIFT procédure



# Classification automatique de biens de conso.

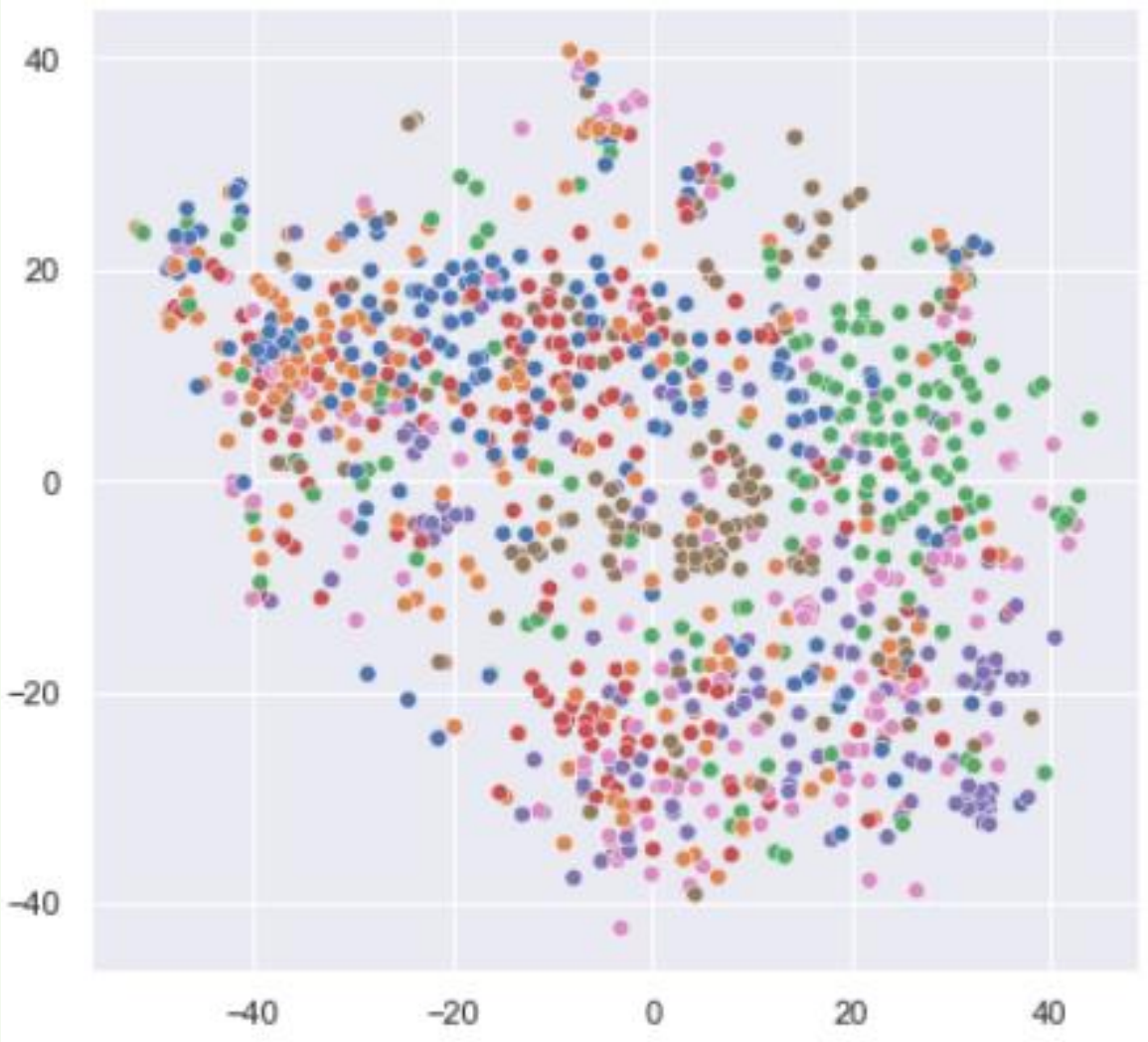


## SIFT procédure



# Classification automatique de biens de conso.

CLUSTERING - #1 SIFT

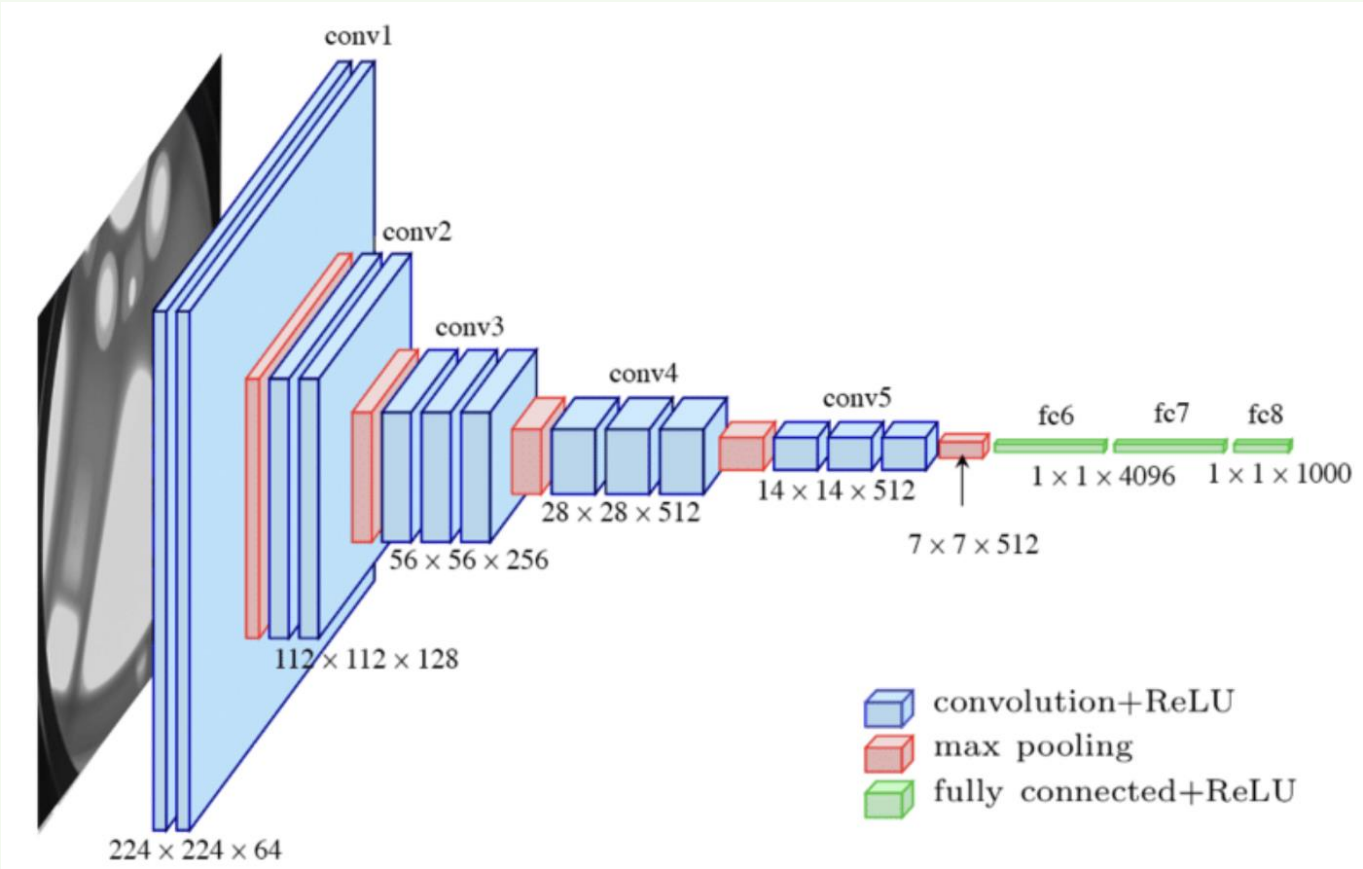



**ARI = 0.08**

# Classification automatique de biens de conso.

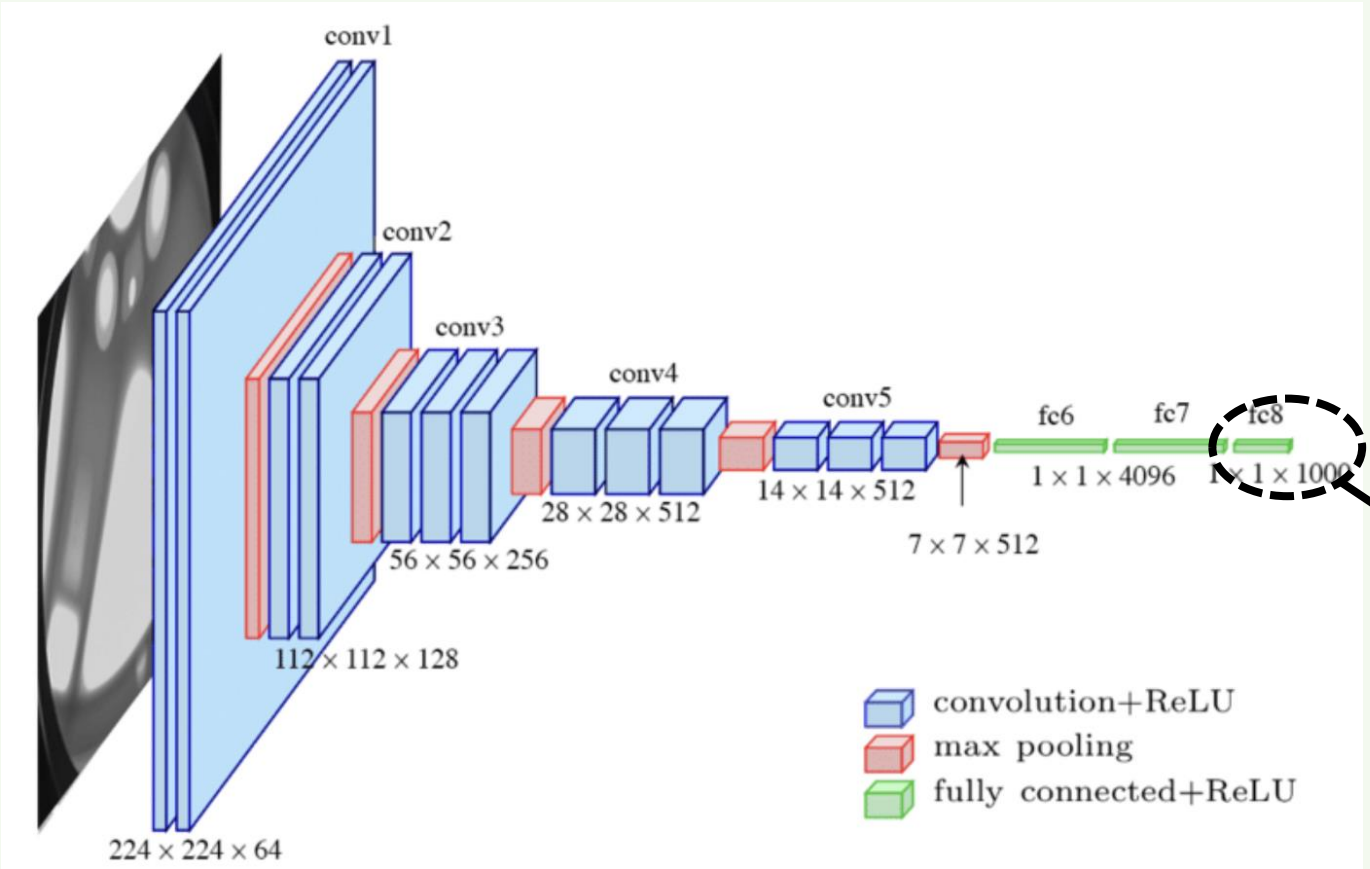


## VGG16 – features extraction



# Classification automatique de biens de conso.

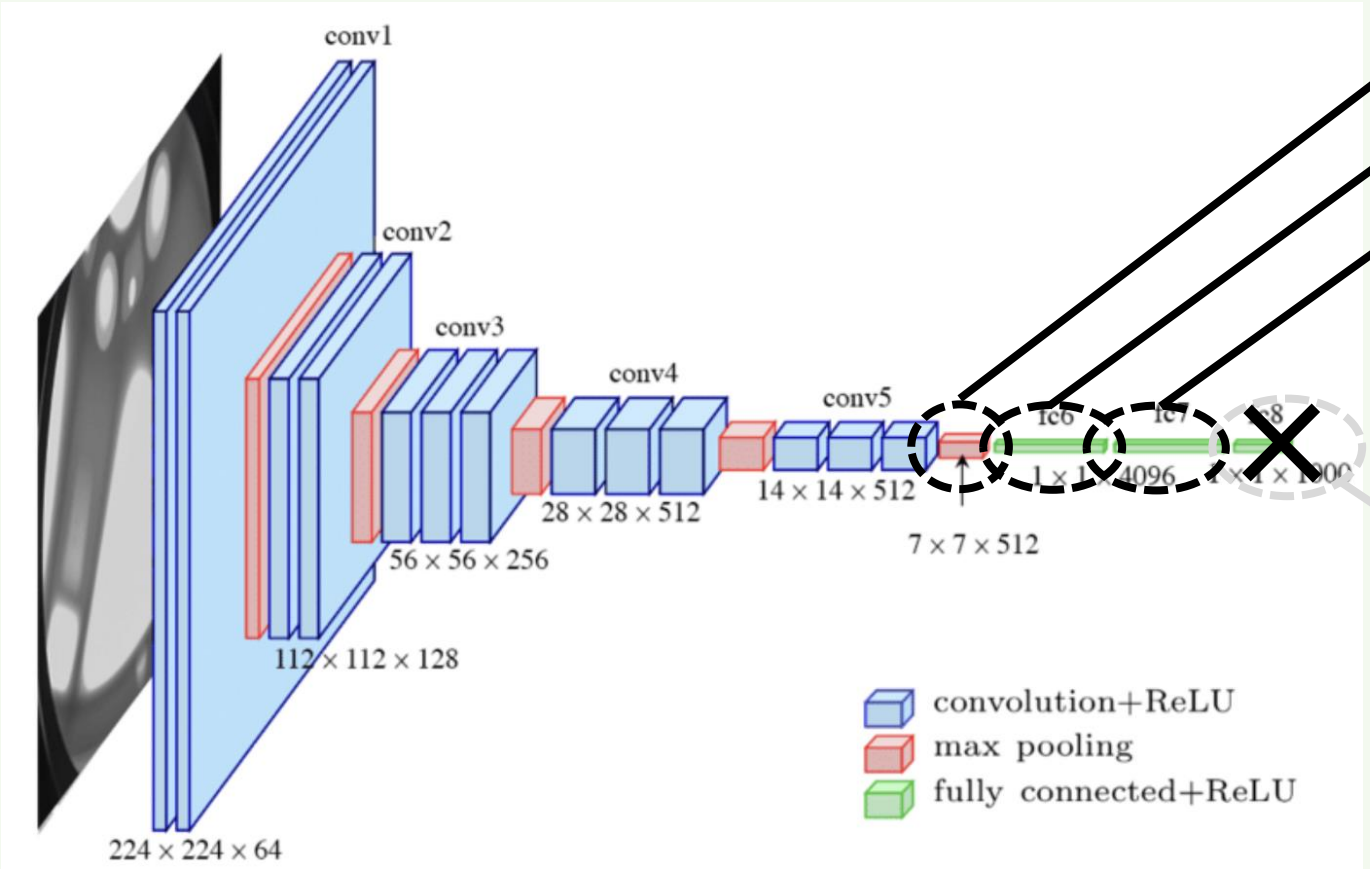
## VGG16 – features extraction




Couche classification (ImageNet)

# Classification automatique de biens de conso.

## VGG16 – features extraction




**Features**

**25088 features ?**

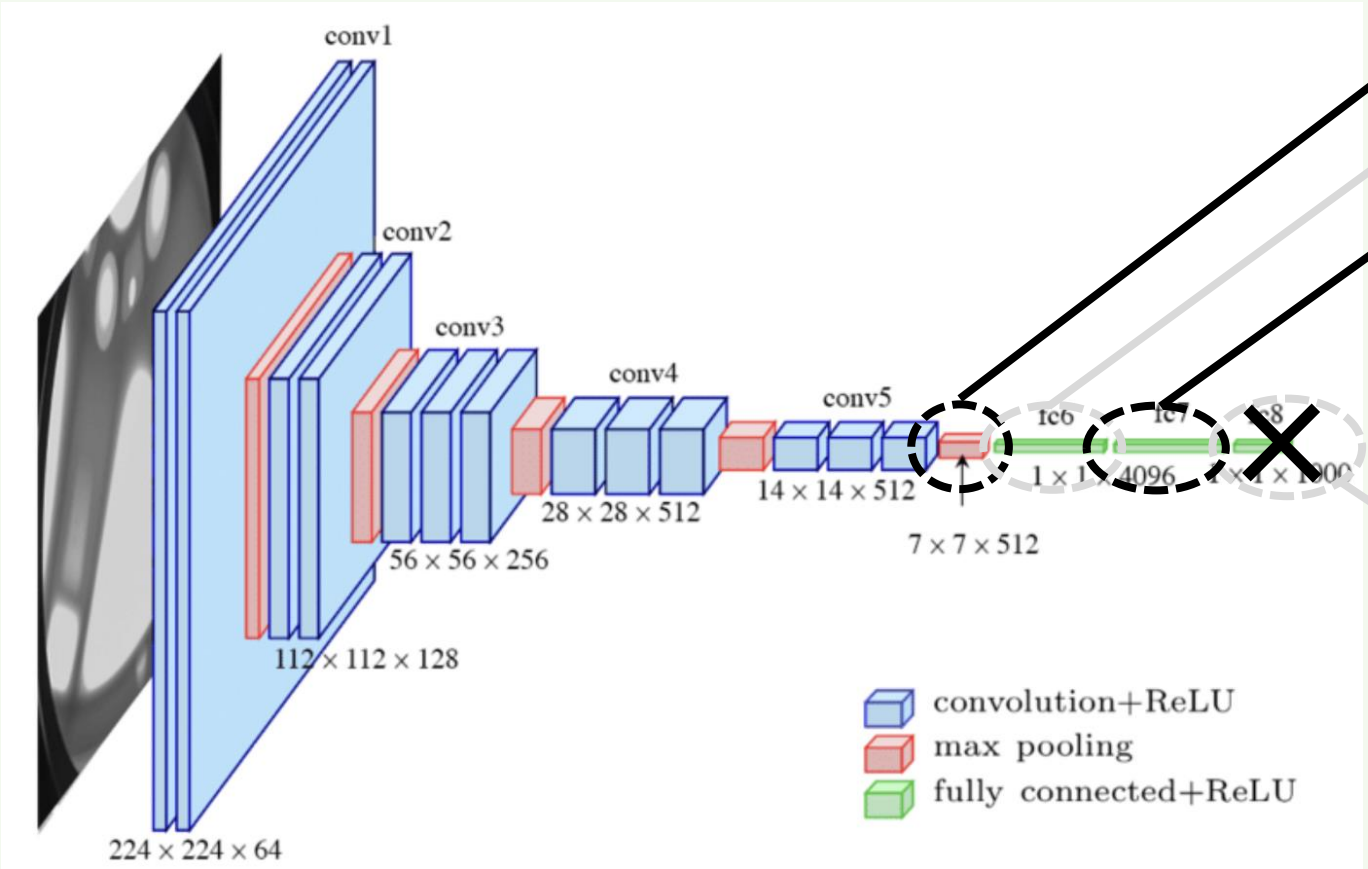
**4096 features ?**

**4096 features ?**

Couche classification (ImageNet)

# Classification automatique de biens de conso.

## VGG16 – features extraction




**Features**

**25088 features (model #1)**

**4096 features ?**

**4096 features (model #2)**

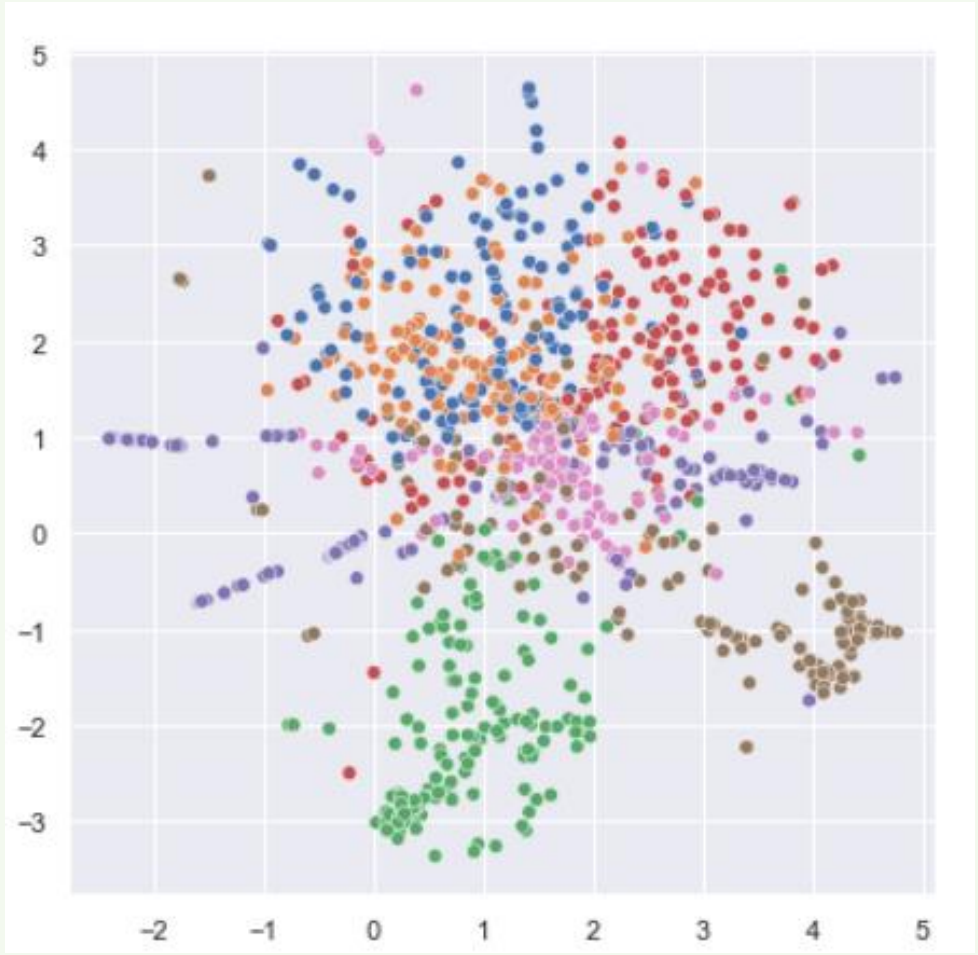
Couche classification (ImageNet)

# Classification automatique de biens de conso.

CLUSTERING - #2 VGG16

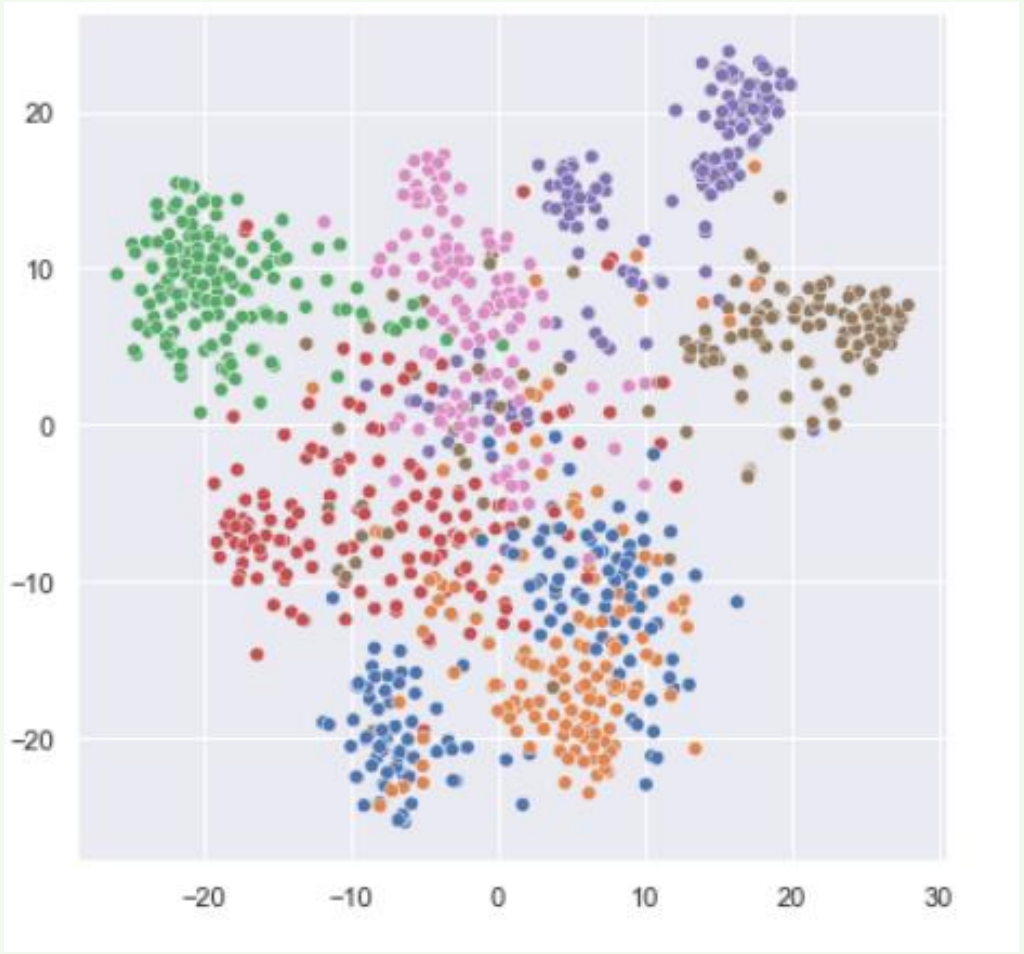
t-sne projections

**Model #1**




**ARI = 0.27**

**Model #2**




**ARI = 0.54**

# Classification automatique de biens de conso.

CONCLUSION - images

| Model | ARI |
|---|---|
| SIFT | 0.08 |
| VGG16 (#1) | 0.27 |
| **VGG16 (#2)** | **0.54** |

Meilleure classification obtenue avec :

**VGG16 (#2)**

# Classification automatique de biens de conso.

CONCLUSION

## Classification non supervisée

### Relativement bon résultats (ca. 50%) sur les descriptions

Utilisation de tf-idf ou Universal Sentence Encoder

**Note :** résultats pourraient être améliorés via une approche supervisée (e.g. avec BERT)

### Relativement bon résultats (ca. 50%) sur les images

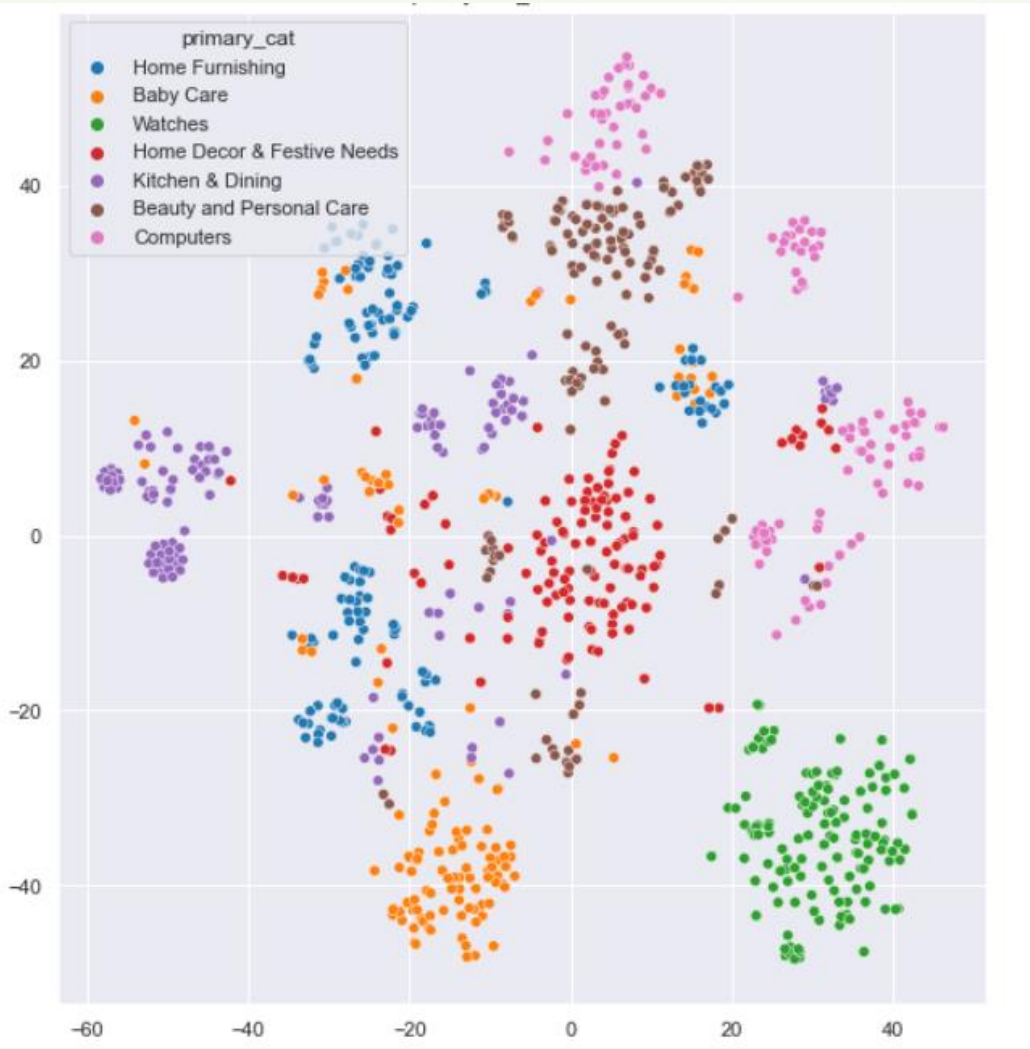
Utilisation de VGG16 (sans la couche de classification)

**Note :** résultats améliorables via approche supervisée (ajout couche de classif + training)

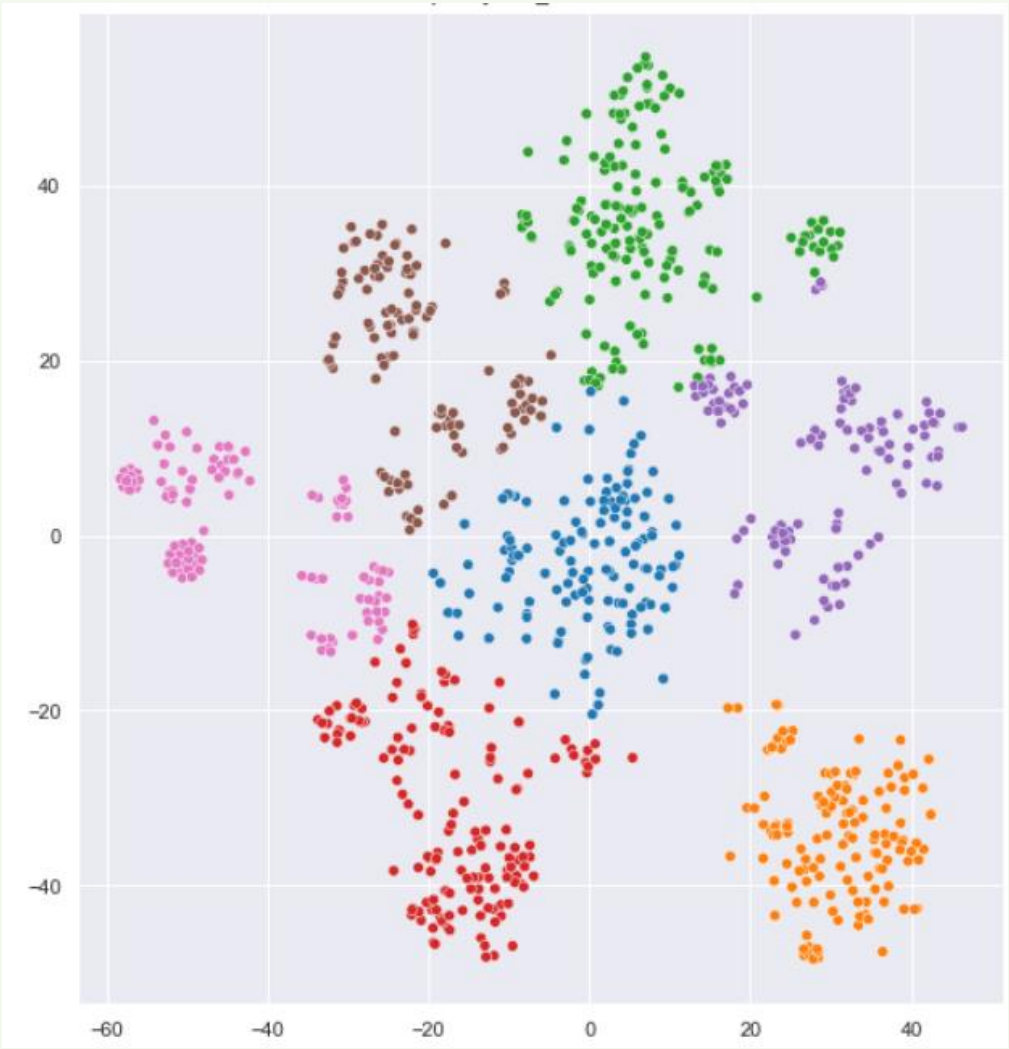
# Classification automatique de biens de conso.

CONCLUSION

## Tf-Idf



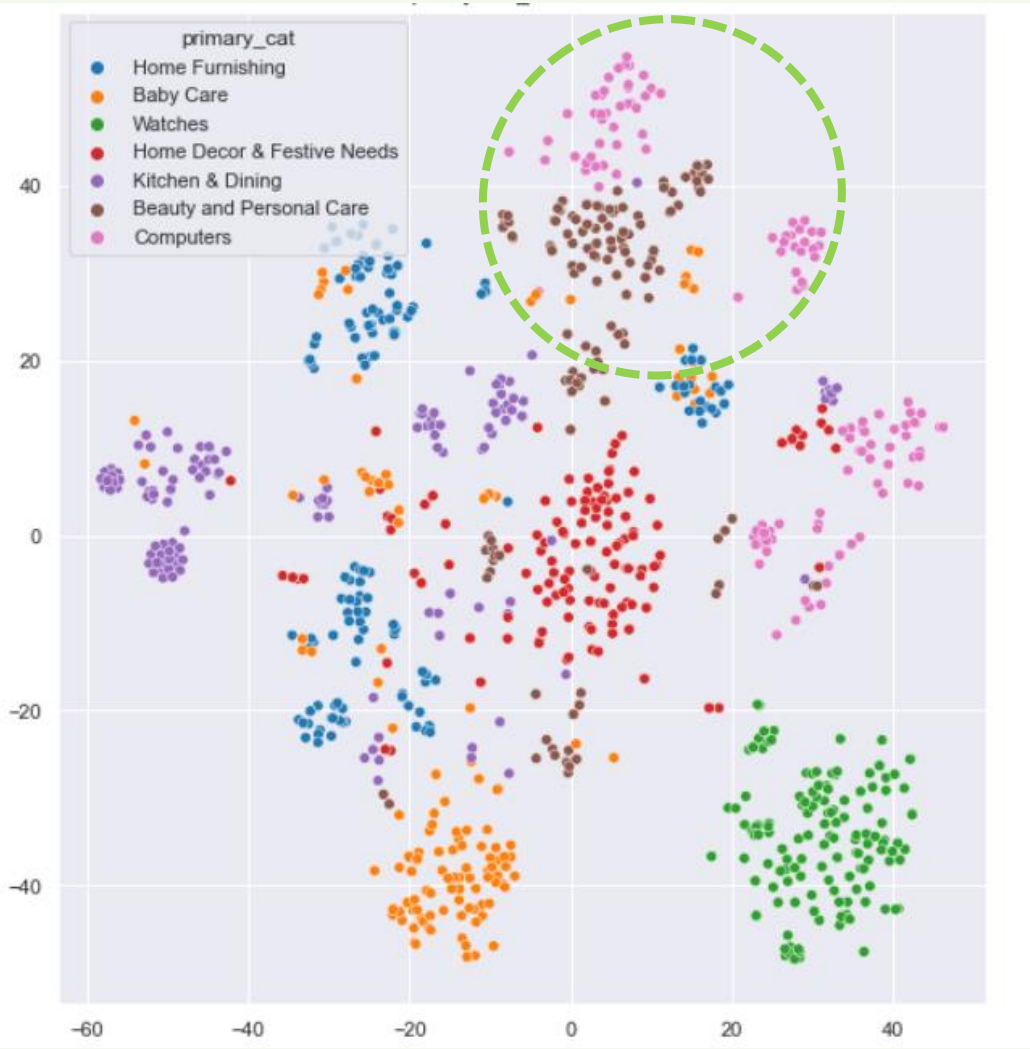

## K-means



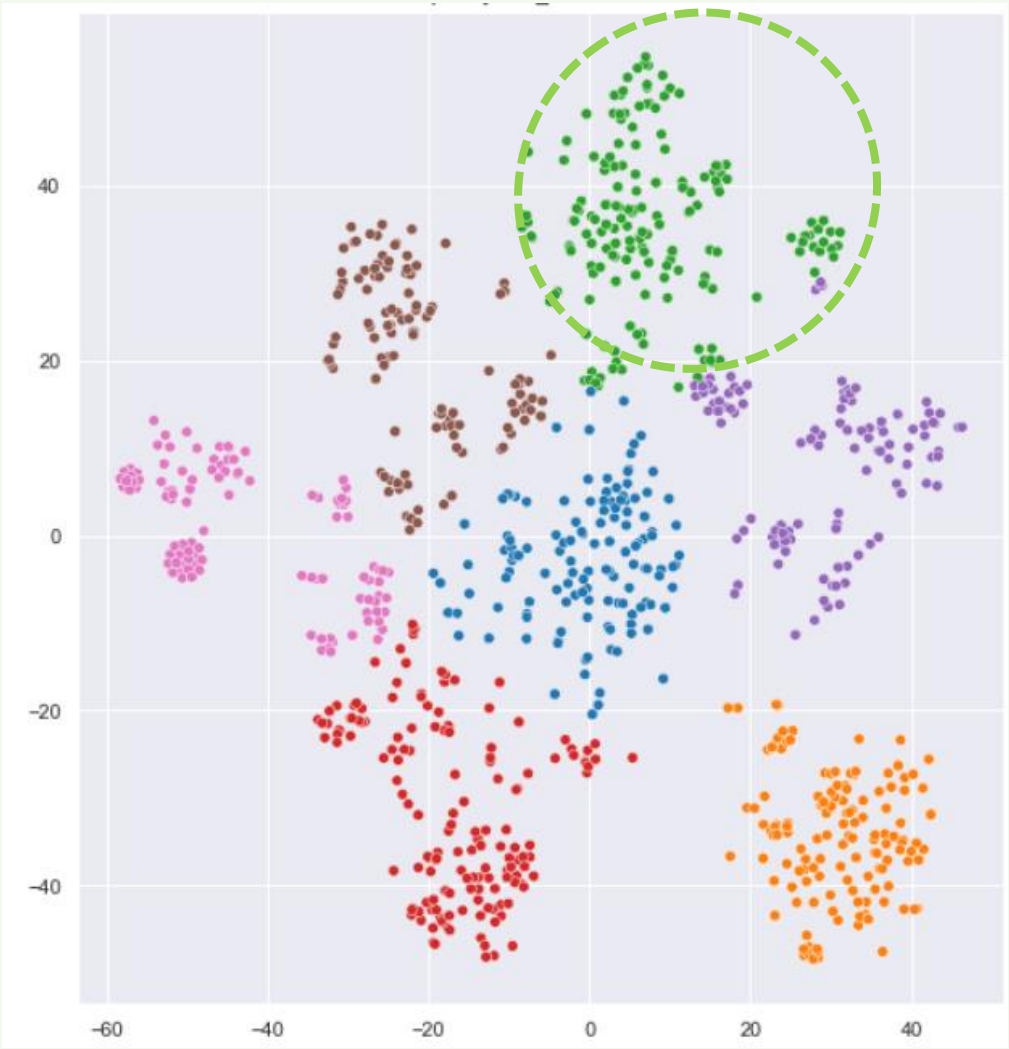
# Classification automatique de biens de conso.

CONCLUSION

## Tf-Idf



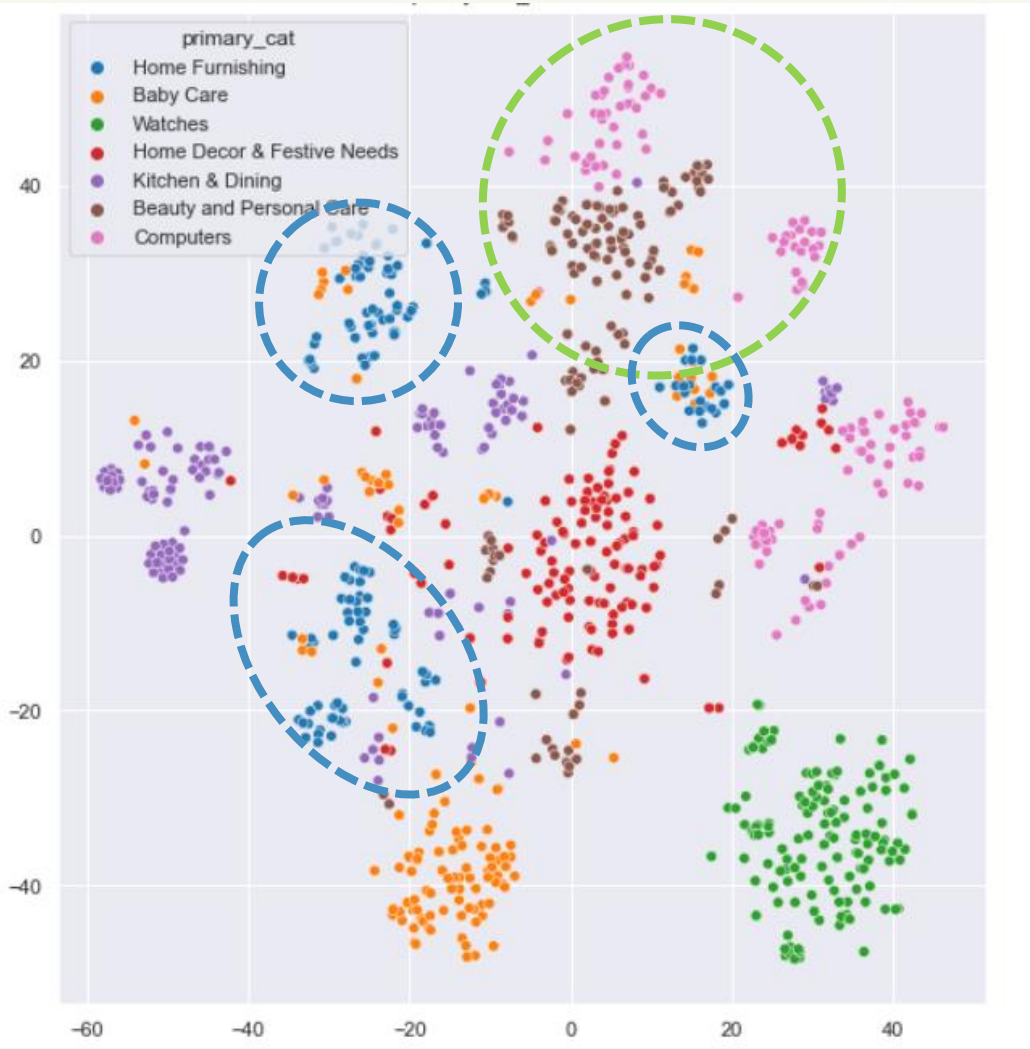

## K-means



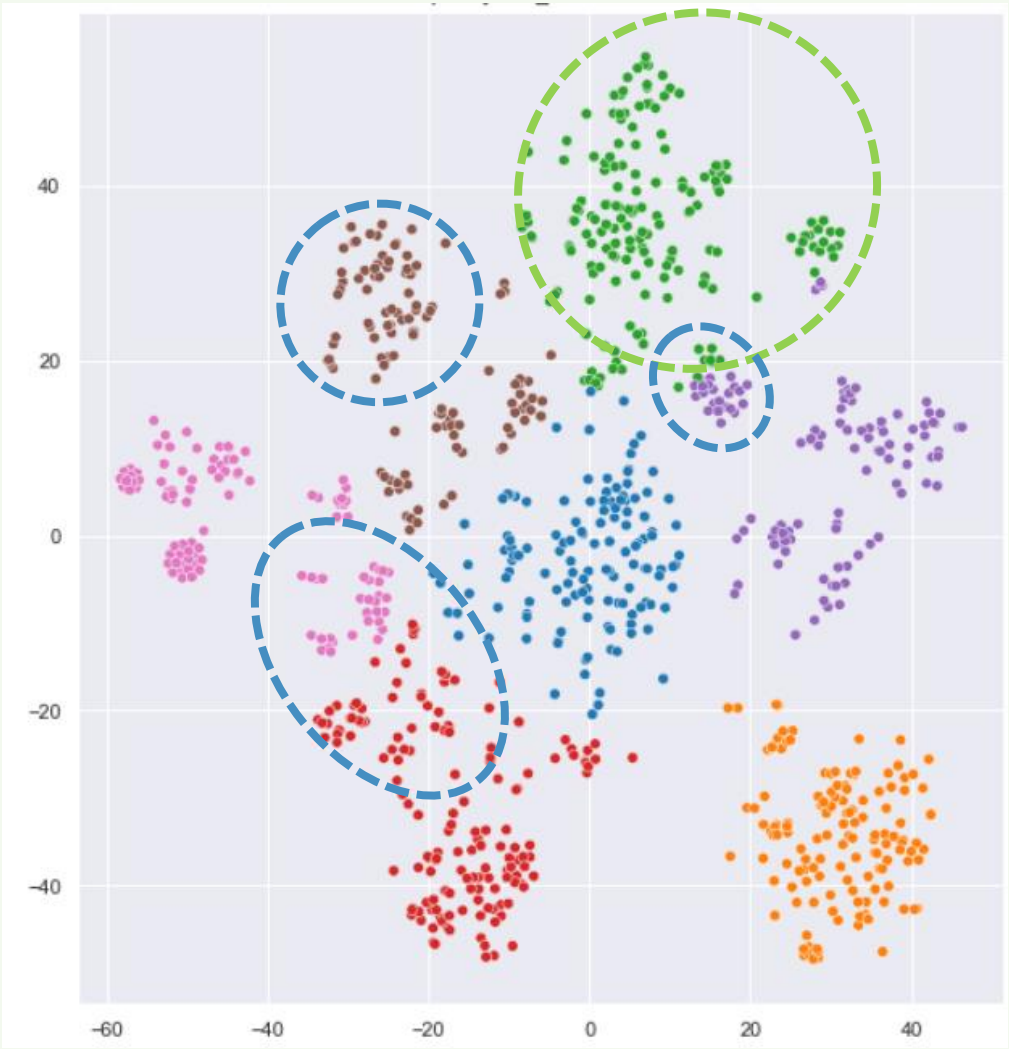
# Classification automatique de biens de conso.

CONCLUSION

## Tf-Idf



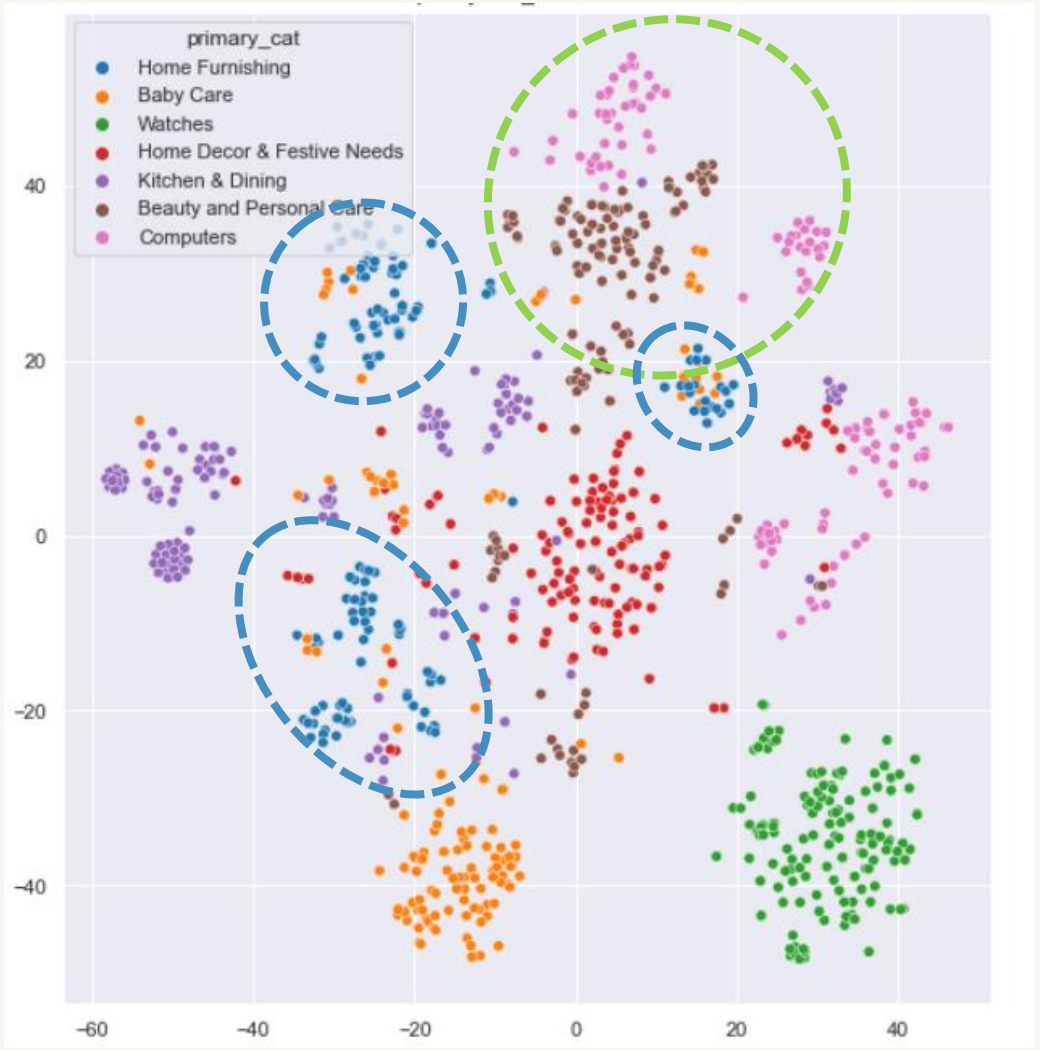

## K-means



# Classification automatique de biens de conso.

CONCLUSION

## Tf-Idf




**Predicted**

**True**

| | Baby care | Beauty & Personal care | Computers | Home dec. & festiv. needs | Home furnishing | Kitchen & Dining | Watches |
|---|---|---|---|---|---|---|---|
| **Baby care** | 100 | 13 | 7 | 5 | 17 | 8 | 0 |
| **Beauty & Personal care** | 13 | 107 | 9 | 21 | 0 | 0 | 0 |
| **Computers** | 0 | 71 | 79 | 0 | 0 | 0 | 0 |
| **Home dec. & festiv. needs** | 6 | 6 | 11 | 114 | 10 | 6 | 3 |
| **Home furnishing** | 39 | 6 | 14 | 1 | 64 | 26 | 0 |
| **Kitchen & Dining** | 11 | 2 | 8 | 12 | 34 | 83 | 0 |
| **Watches** | 0 | 0 | 0 | 0 | 0 | 0 | 150 |

## Confusion matrix

# Classification automatique de biens de conso.

**Merci**