

Projet 6 : Classifiez automatiquement des biens de consommation

Problématique :

- Entreprise place de marché : Vente en ligne d'articles divers
- Attribution manuelle des catégories par les vendeurs : fastidieuse et peu fiable
- Objectif : Automatisation de l'attribution de la catégorie de l'article



Plan d'étude

- I. Brève présentation du tableau des données.
- II. Etude des données textuelles.
 1. Bag of words.
 - CountVectorizer / Tf-idf
 2. Word/sentence embedding classique.
 - Word2Vec
 3. Word/sentence embedding.
 - Bert / USE
- III. Etude des données visuelles.
 1. SIFT
 2. CNN Transfer learning
- IV. Conclusions.

Présentation du dataset

- Taille du dataset :
 - 1050 lignes : identifiants des produits
 - 15 colonnes : noms des produits, leurs catégories et sous-catégories, leur description, leur image, les adresse des pages web dédiées à chaque produit (site Flipkart) , etc.
- Pas de doublons
- Quelques valeurs manquantes à la colonne 'brand'

Exemple de produit sur le site Flipkart

Flipkart

Search for products, brands and more

Search

Login

Become a Seller

More

Cart

Electronics

TVs & Appliances

Men

Women

Baby & Kids

Home & Furniture

Sports, Books & More

Flights

Offer Zone

Share

Elegance 213 cm (7 ft) Polyester Door Curtain (Pack Of 2) (Abstract, Multicolor)

Price: Not Available

Currently Unavailable

Designed For

Door

Highlights

- Door (121 cm x 213 cm)
- Material: Polyester
- Pack of: 2
- Closure Type: Eyelet

SuperCoin

For every ₹100 Spent, you earn 2 SuperCoins

Max 50 coins per order

Description

This curtain enhances the look of the interiors. This curtain is made from 100% high quality polyester fabric. It features an eyelet style stitch with Metal Ring. It makes the room environment romantic and loving. This curtain is anti-wrinkle and anti shrinkage and have elegant appearance. Give your home a bright and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sunlight.

Specifications

Model Name

Abstract Polyester Door Curtain Set Of 2

Model Number

Duster25

- 'product_name' : Elegance Polyester Multicolor Abstract Eyelet Door Curtain
- Les 2 données pertinentes 'image' et 'description' présentes dans le dataset et la page web.

Catégories des produits

- Extraction des catégories de la colonne 'product_category_tree':

- 1^{er} élément de la colonne :

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



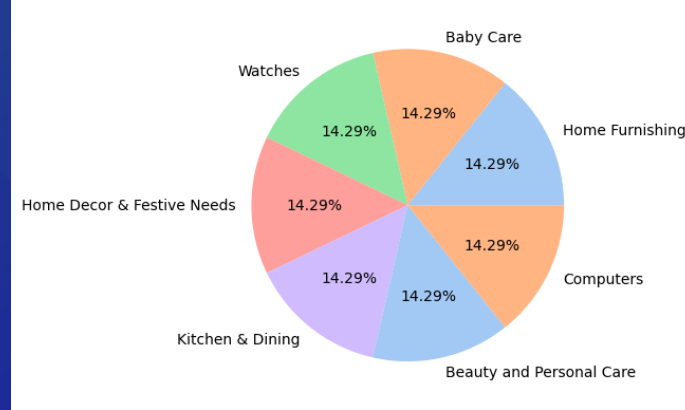
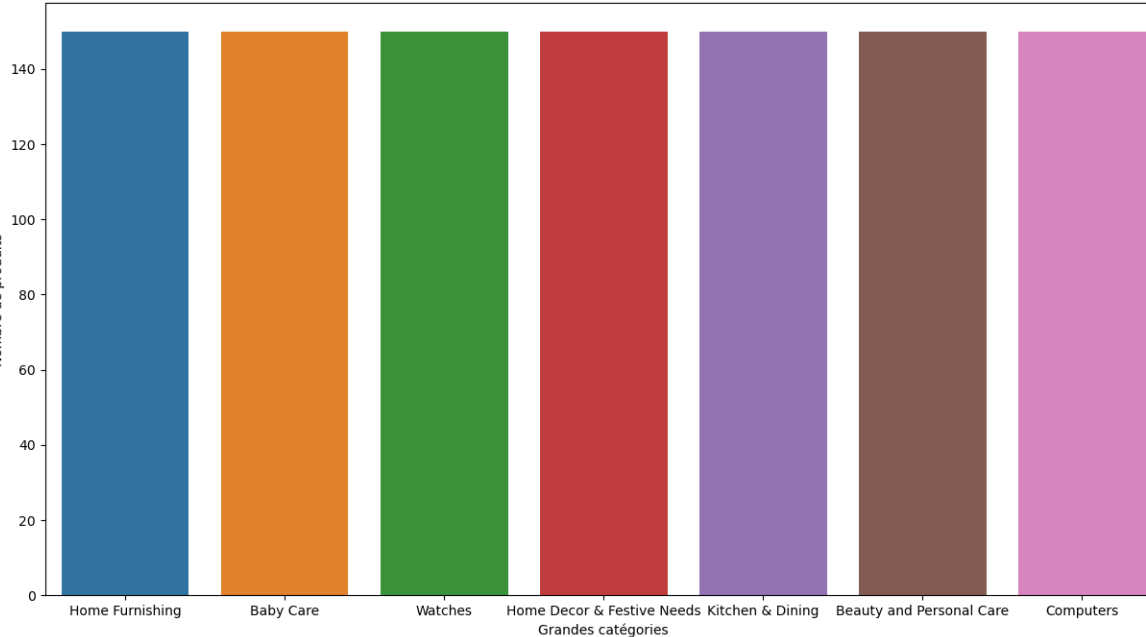
extraction de la catégorie 'Home Furnishing'

- 7 catégories principales : 'Home Furnishing', 'Baby Care', 'Watches', 'Home Decor & Festive Needs', 'Kitchen & Dining', 'Beauty and Personal Care', 'Computers'
- Objectif : Etudier la possibilité de répartir, de manière automatique, les produits en différentes catégories :
 - utilisation de différentes méthodes d'extractions de features au sein de données textuelles et d'images
 - variables pertinentes du jeu de données : 'image', 'description' et la colonne des catégories.

Catégories des produits

- 150 produits dans chacune des 7 catégories principales

Nombre de produits par grande catégorie



Démarche



Etude des données textuelles

- Opérations de pré-traitement :
 - La tokenization = découpage en mots des différents textes de notre corpus.
(la description d'un produit = un document, corpus = l'ensemble des documents)
 - La normalisation = suppression des détails importants au niveau local (ponctuation, majuscules, conjugaison, etc.)
 - suppression des stopwords = mots beaucoup utilisés mais n'apporte pas d'informations pour la compréhension du sens du document et du corpus (les articles, déterminants etc. , exemples en anglais : that, them, the, but etc.)

Etude des données textuelles

(Exemple de pré-traitement)

- Image du produit



- Texte de description du produit

"Anthill Baby Boy's, Baby Girl's Bodysuit - Buy Yellow Anthill Baby Boy's, Baby Girl's Bodysuit For Only Rs. 405 Online in India. Shop Online For Apparels. Huge Collection of Branded Clothes Only at Flipkart.com"

- Liste de mots obtenus suite au traitement du texte

['anthill', 'baby', 'boy', 'baby', 'girl', 'bodysuit', 'buy', 'yellow', 'anthill', 'baby', 'boy', 'baby', 'girl', 'bodysuit', 'rs', '405', 'online', 'india', 'shop', 'online', 'apparels', 'huge', 'collection', 'branded', 'clothes', 'flipkart', 'com']

Bag of words(countVectorizer)

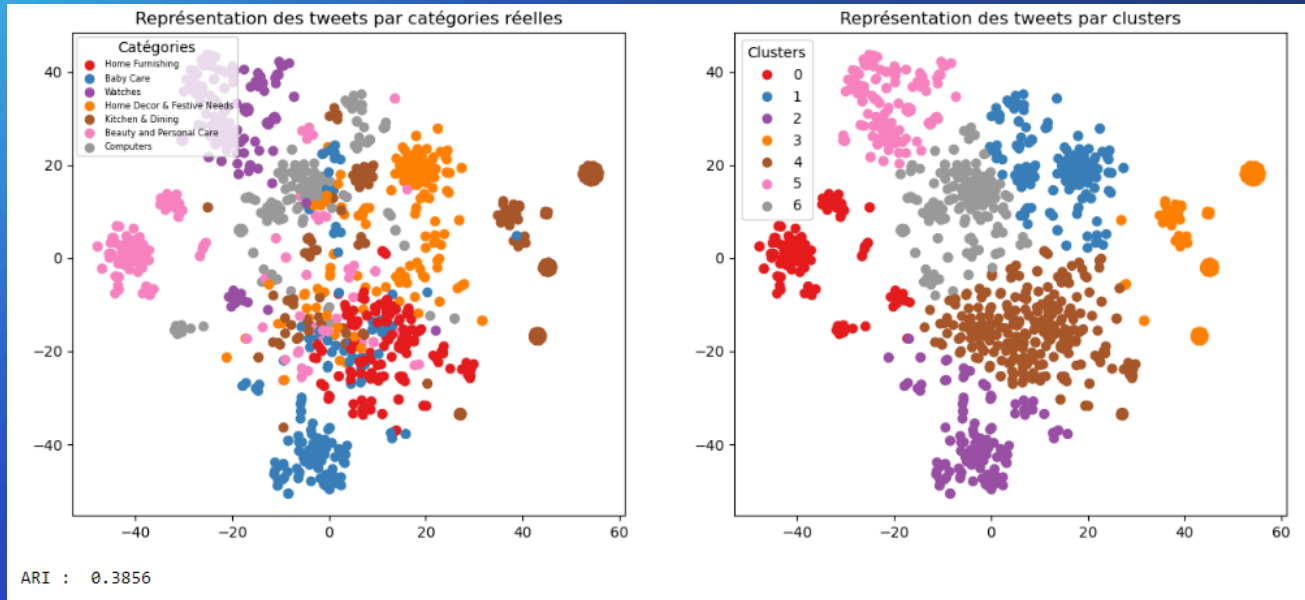
- countVectorizer= comptage simple de mots :
 - Création d'une matrice : chaque mot unique représenté par une colonne et chaque document par une ligne
 - cellule (i, j) : nombre de mot i dans le document j
- Illustration :

```
document = [ "One Geek helps Two Geeks", "Two Geeks help Four Geeks", "Each Geek helps many other Geeks at GeeksforGeeks." ]
```

	at	each	four	geek	geeks	geeksforgeeks	help	helps	many	one	other	two
document[0]	0	0	0	1	1	0	0	1	0	1	0	1
document[1]	0	0	1	0	2	0	1	0	0	0	0	1
document[2]	1	1	0	1	1	1	0	1	1	0	1	0

Bag of words (countVectorizer)

- CountVectorizer

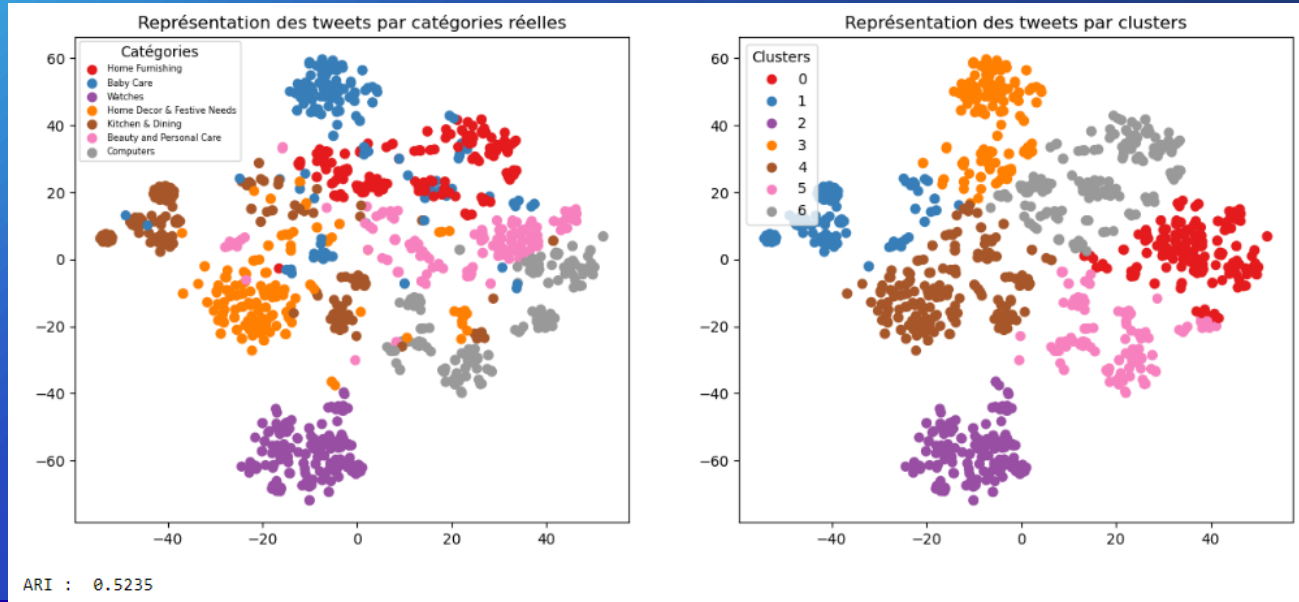


Bag of words (Tf-idf)

- TF-IDF (Term Frequency-Inverse Document Frequency) est un moyen de mesurer la pertinence d'un mot par rapport à un document dans une collection de documents.
- Poids (pertinence) d'un terme t dans le document $d = W_{t,d} = TF_{t,d} \log (N/DF_t)$:
 - $TF_{t,d}$ est le nombre d'occurrences de t dans le document d .
 - DF_t est le nombre de documents contenant le terme t .
 - N est le nombre total de documents dans le corpus.
- TF-IDF est noté entre 0 et 1. Plus la valeur numérique du poids est élevée, plus le terme est rare . Plus le poids est petit, plus le terme est courant .

Bag of words (Tf-idf)

- Tf-idf



Word/sentence embedding classique (Word2Vec)

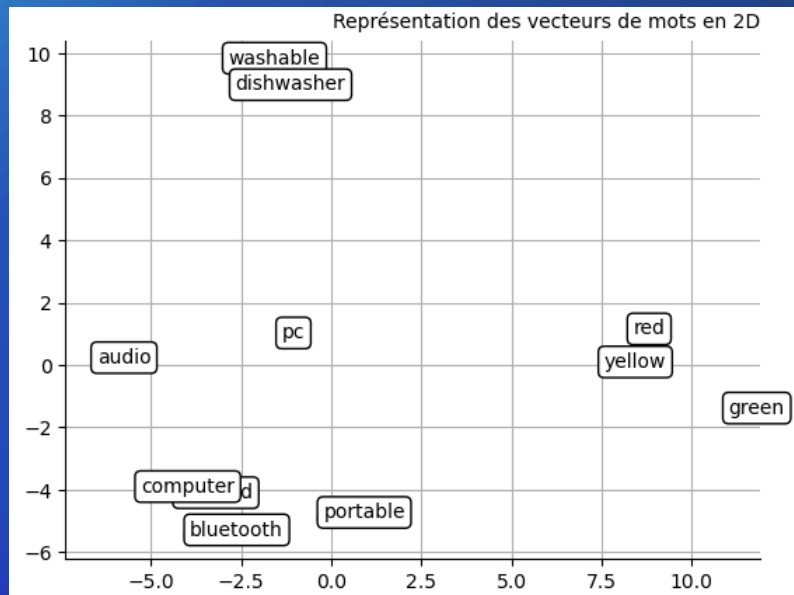
- Word2vec = réseau de neurones à 3 couches (1 couche d'entrée, 1 couche cachée, 1 couche de sortie) de traitement de texte.
- Entrée du modèle : le corpus (ensemble des documents) ; sortie = des vecteurs caractéristiques représentant des mots du corpus.



- La représentation vectorielle prend en compte le contexte du mots (mots adjacents, avant et après)
- word2vec repose sur l' hypothèse distributionnelle (*distributional hypothesis*) = les mots qui ont souvent les mêmes mots voisins ont tendance à être sémantiquement similaires.
- ➡ But et utilité de Word2vec = regrouper dans un espace vectoriel les vecteurs des mots similaires (le modèle détecte mathématiquement les similitudes entre les mots)

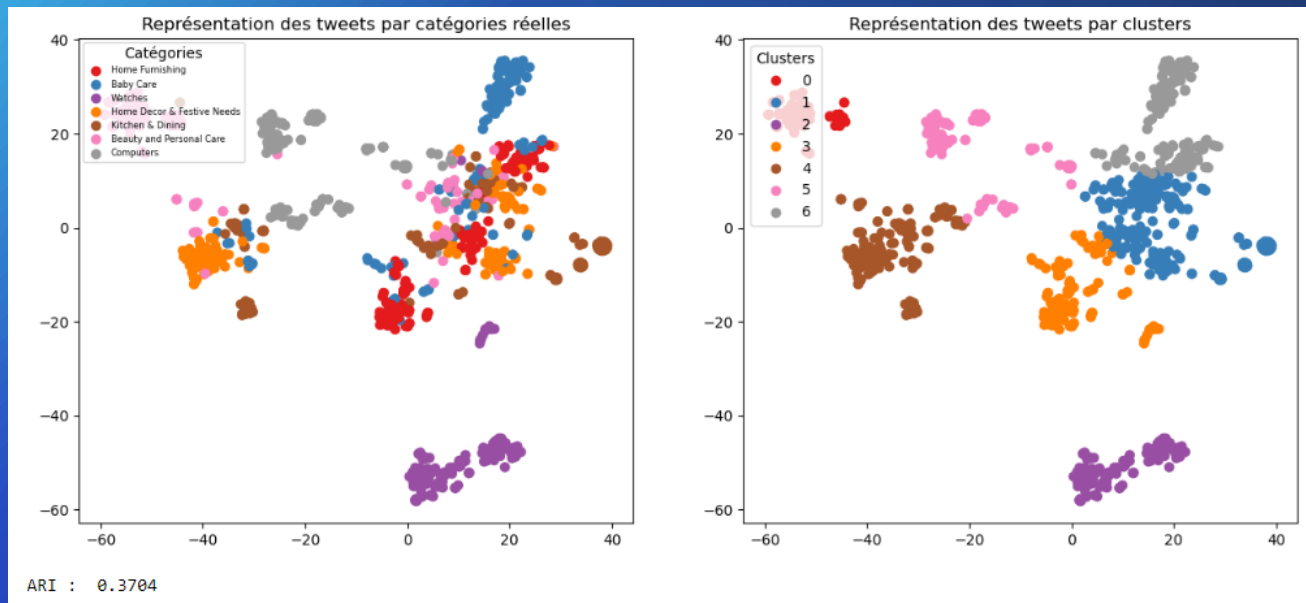
Word/sentence embedding classique (Word2Vec)

- Projection en 2D de certains vecteurs de mots de notre corpus: les mots similaires sont proches les uns des autres



Word/sentence embedding classique (Word2Vec)

- Word2Vec



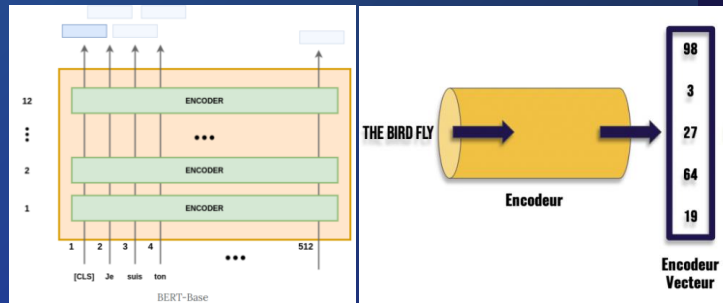
sentence/document embedding (Bert)

- Bert = modèle de type Transformers (d'où le <<T>> de BERT):
« Bidirectional Encoder Representations from Transformers » littéralement
« Représentations d'encodeurs bidirectionnels à partir de transformateurs »

- Transformer = réseau de neurones.
- BERT = superposition d'encodeurs

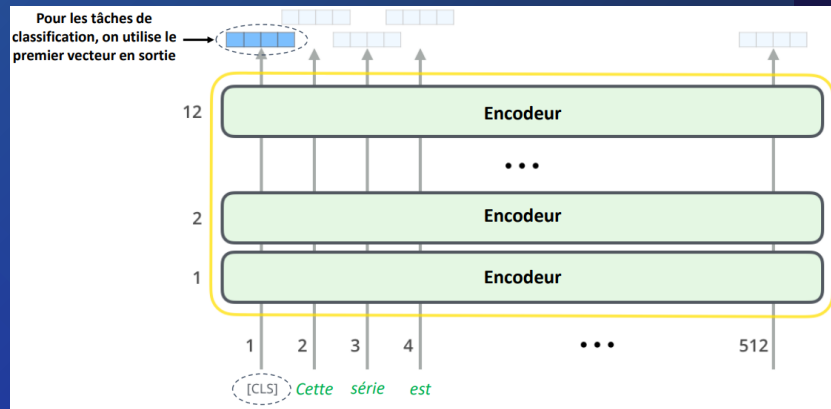
- ➡ le modèle effectue plusieurs étapes, avec application à chaque fois d'un mécanisme d'attention.

- But de ce mécanisme = comprendre les relations entre les mots de la phrase quelles que soit leurs positions
 - Exemple : << Tu as une nouvelle souris pour ton ordinateur >>
 - Pour déterminer le sens du mot *souris*, l'objet et non l'animal, le transformer va prêter attention au mot "ordinateur" et prendre une décision en une étape basée sur ça.
- Cette architecture d'encodeurs est construite pour créer un modèle de représentation du langage (permet de comprendre le langage).



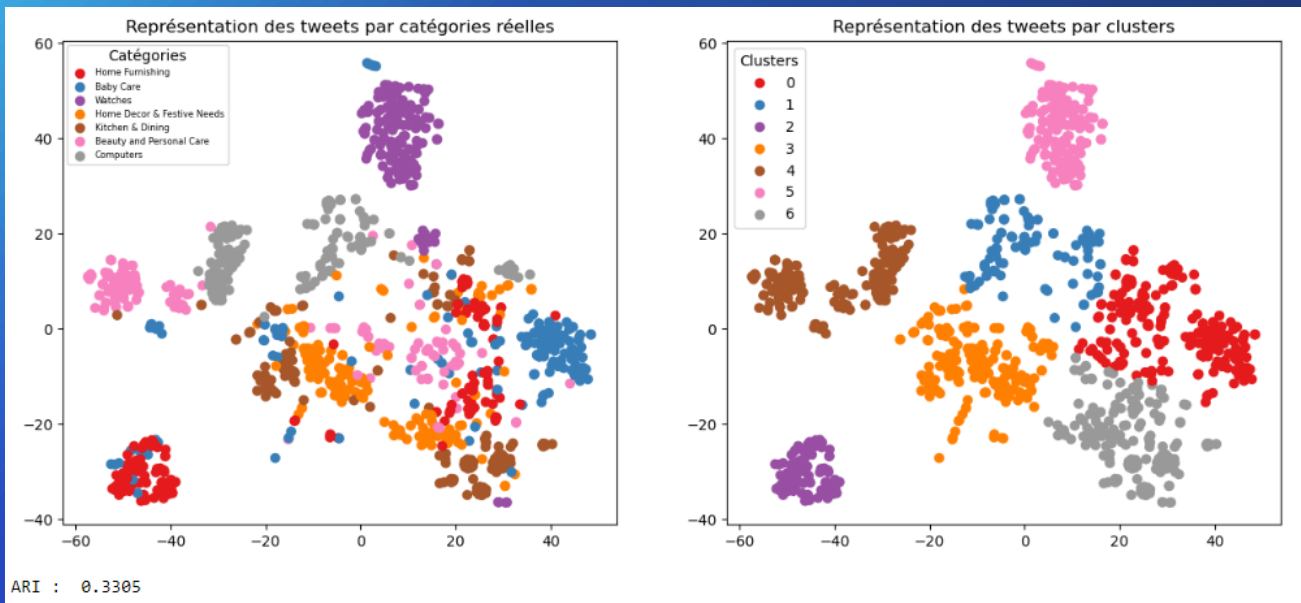
sentence/document embedding (Bert)

- Cas de notre projet : utilisation de BERT pour la classification.
- Récupération du premier vecteur de sortie (sortie du vecteur de classe).
- ➡ résultat : un vecteur par document .
- Application du modèle de classification non supervisé Kmeans
- Comparaison des clusters obtenus avec les catégories réelles.



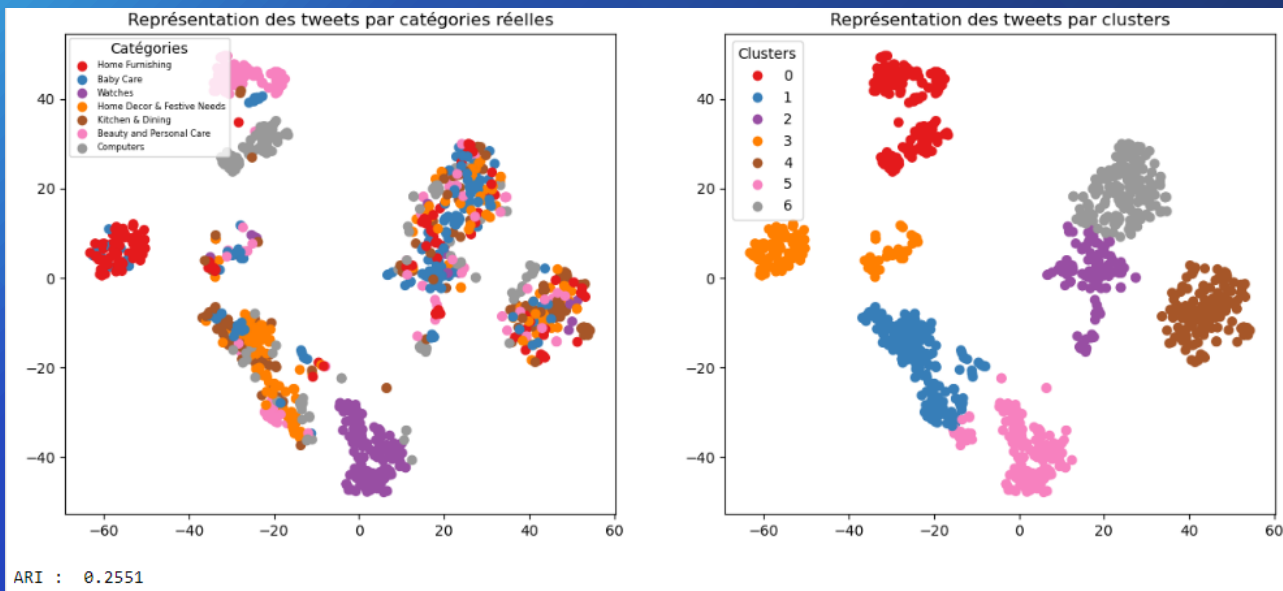
sentence/document embedding (Bert)

- BERT hub Tensorflow
- Modèle de pré-entraînement : « **bert-base-uncased** »



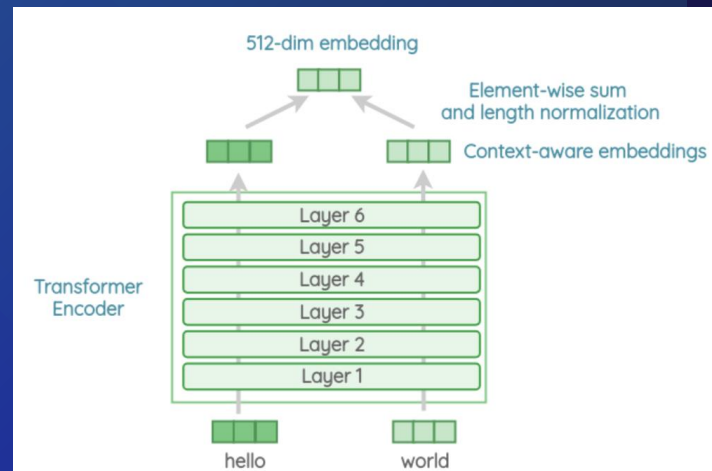
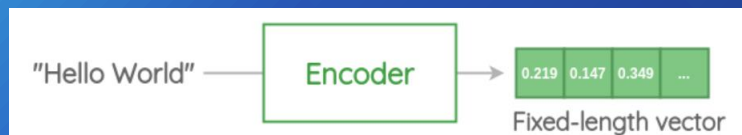
sentence/document embedding (Bert)

- BERT hub Tensorflow
- Modèle de pré-entraînement : « **bert-large-uncased-whole-word-masking-finetuned-squad** »



sentence/document embedding (USE)

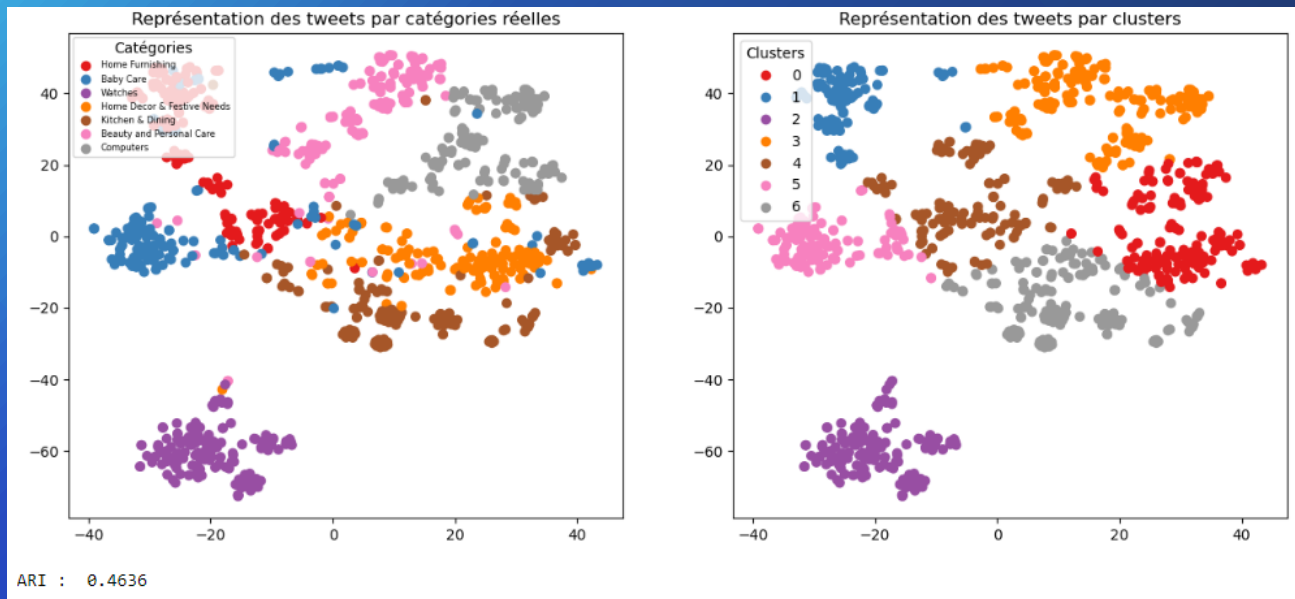
- USE (Universal Sentence Encoder) encode une phrase ou un texte entier en un vecteur de nombres réels (un vecteur par texte)
- Entrée : un texte de longueurs différentes ; sortie un vecteur de dimension 512.



- Utilisation du modèle pour plusieurs tâches : classification de textes, la similarité sémantique et d'autres tâches de langage naturel.

sentence/document embedding (USE)

- USE (Universal Sentence Encoder)



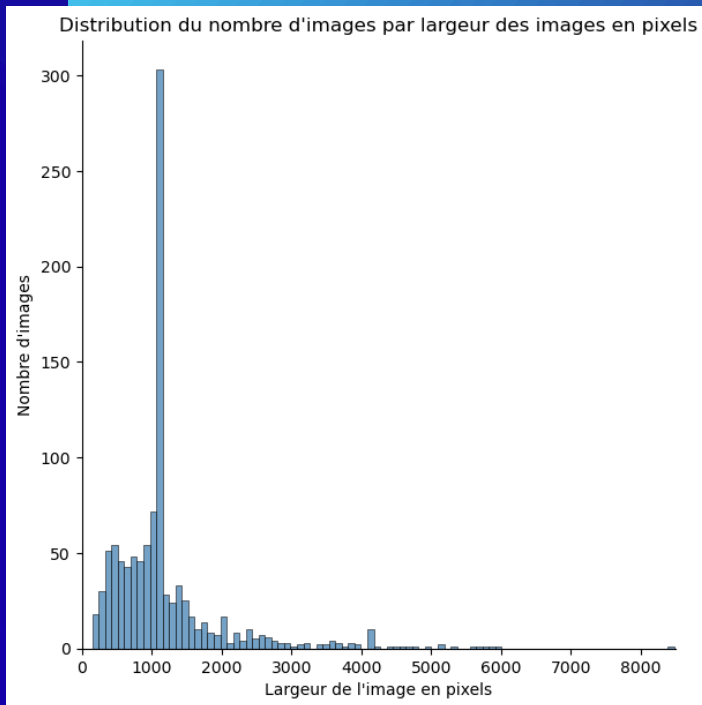
Synthèse sur l'étude des données textuelles

- Possibilité de classer les textes par catégories :
 - obtention de catégories distinctes mais avec des taux de correspondance aux catégories réelles différents suivant les modèles.
- Les indices de similitudes donne une idée de performance des modèles
- Il faut relativiser les comparaisons des modèles avec les ARI obtenus car certains modèles ont une grande marge d'amélioration par l'optimisation de leurs hyperparamètres.

modèle	Counvectorizer	Tf-idf	Word2vec	BERT	USE
ARI	0.3856	0.5235	0.3704	0.3305	0.4636

Etude des données visuelles

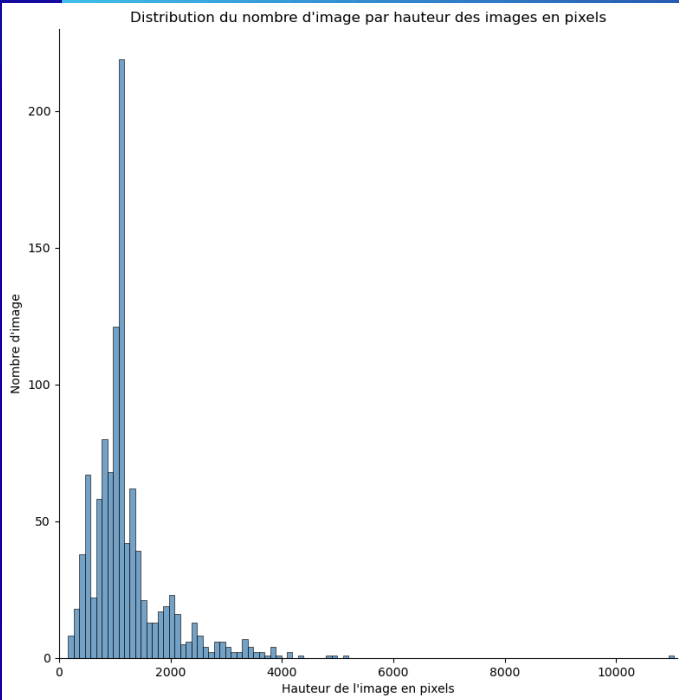
- Analyse exploratoire :



- Largeurs des images non uniformes et réparties de manière inégales.
- ➡ Nécessité de les uniformiser pour les algorithmes de traitement des images.

Etude des données visuelles

- Analyse exploratoire :

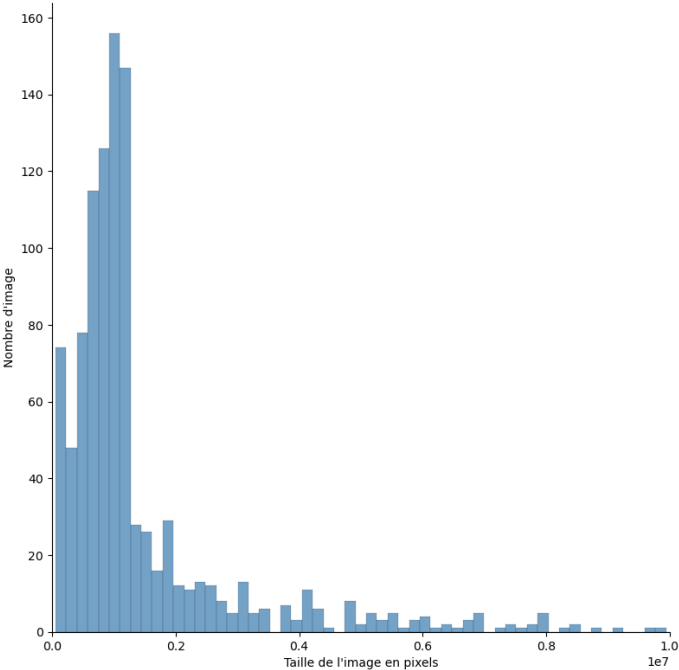


- Hauteurs des images non uniformes et réparties de manière inégales.
- ➡ Nécessité de les uniformiser pour les algorithmes de traitement des images.

Etude des données visuelles

▪ Analyse exploratoire :

Distribution du nombre d'image par taille en pixels



- Tailles des images sont très inégales.
- Taille des images en entrées des réseaux de neurones est 224 X 224
- ➡ Nécessité de les redimensionner pour les réseaux de neurones et le CNN transfert Learning.
- Utilisation de la bibliothèque Python Pillow pour les différents traitements d'image.

Etude des données visuelles

- Amélioration de la luminosité (= exposition) :

image trop sombre (sous-exposée) = majorité des pixels dans les niveaux de gris faibles

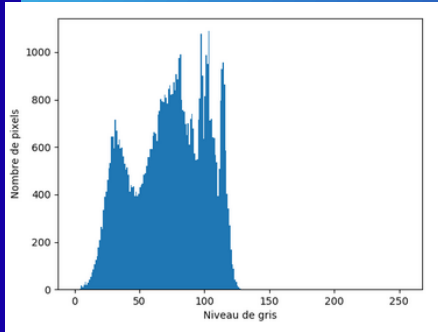
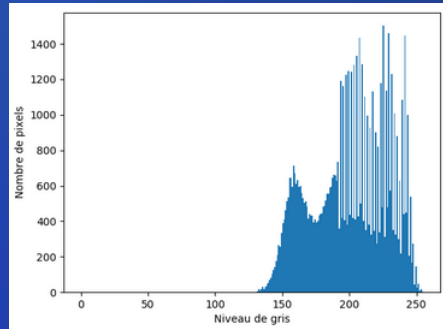
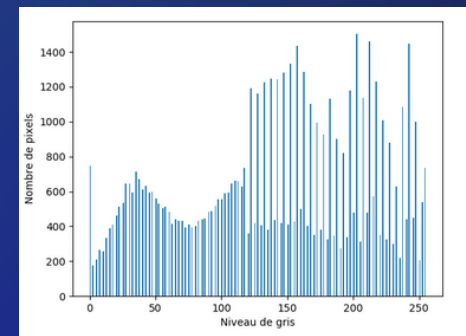


image trop claire (sur-exposée), majorité des pixels dans les niveaux de gris élevés



Amélioration de l'exposition par l'étirement de l'histogramme

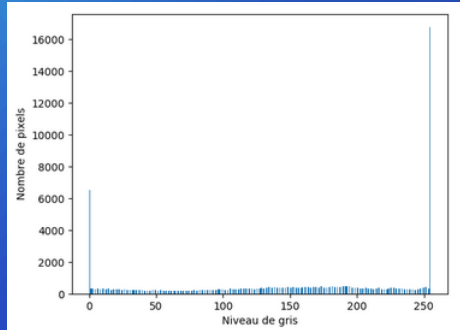


→ répartition des pixels sur tout l'intervalle $[0, 255]$
(image sur-exposée au départ)

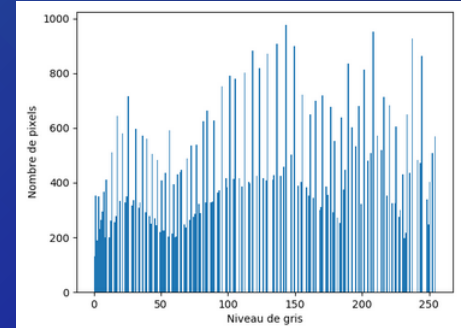
Etude des données visuelles

- Amélioration du contraste (répartition de la lumière dans l'image).
- Une image trop contrastée = grande différence de luminosité entre les zones claires et les zones sombres
- Amélioration du contraste par l'égalisation de l'histogramme.
- Objectif : avoir idéalement le même nombre de pixels à chaque niveau de gris.

Histogramme de l'image trop contrastée



Histogramme après son égalisation (image peu contrastée)



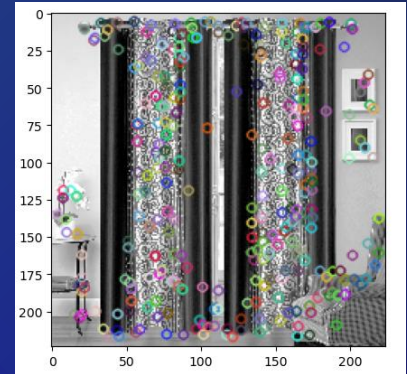
Etude des données visuelles

(SIFT = Scale-invariant feature transform)

- SIFT (« transformation de caractéristiques visuelles invariante à l'échelle ») = algorithme d'extraction des features (ou points d'intérêt) de l'image et de calcul de leurs descripteurs.
- Descripteur = vecteur décrivant le voisinage de la feature à laquelle il est associé.
- Sur une image numérique, Sift détecte une caractéristique locale intéressante et lui attribue ensuite une information quantitative (vecteur) appelées descripteurs



Pré-traitement et application de SIFT




- Image originale

mis en évidence sur l'image
de différents points d'intérêt

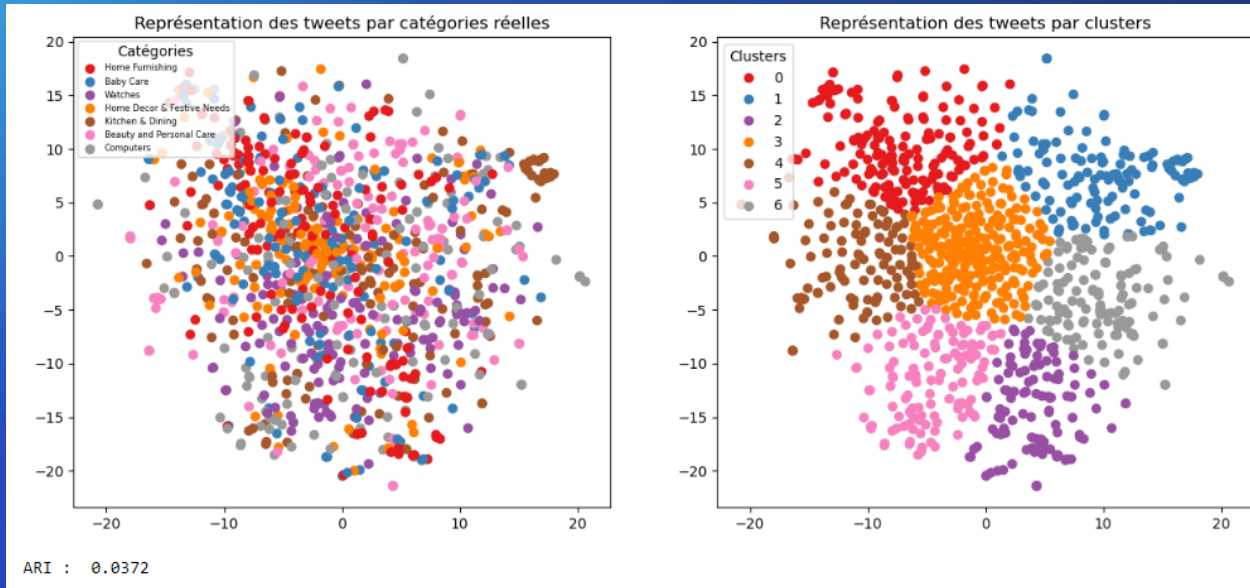
Etude des données visuelles

(SIFT = Scale-invariant feature transform)

- Les descripteurs sont invariants à plusieurs transformations (rotation , échelle, illumination etc.)  un même objet aura des descripteurs similaires dans des images différentes.
- résultat : plusieurs vecteurs par image .
- Application du modèle de classification non supervisé Kmeans.
- Comparaison des clusters obtenus avec les catégories réelles.

Etude des données visuelles (SIFT = Scale-invariant feature transform)

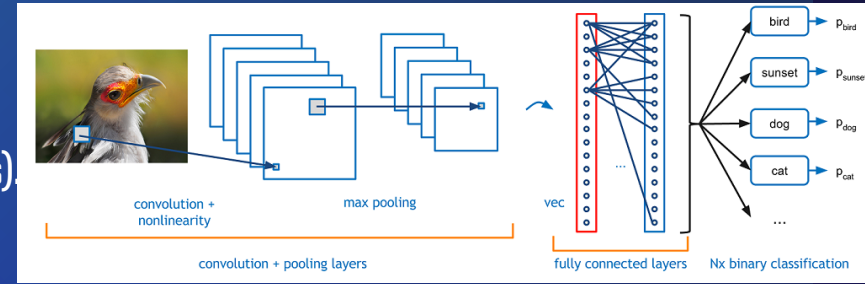
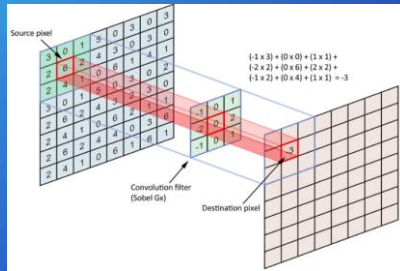
- SIFT



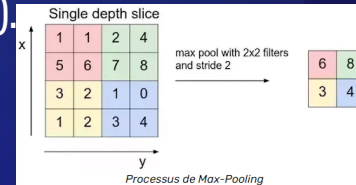
Etude des données visuelles

(CNN = réseau de neurones convolutif)

- CNN = type spécifique de réseau neuronal à plusieurs couches.
- **Etape 1** : la convolutive : extraction des caractéristiques propres à chaque image (utilisation de filtres pour détecter des formes).
- Exemple d'un filtre de convolution.



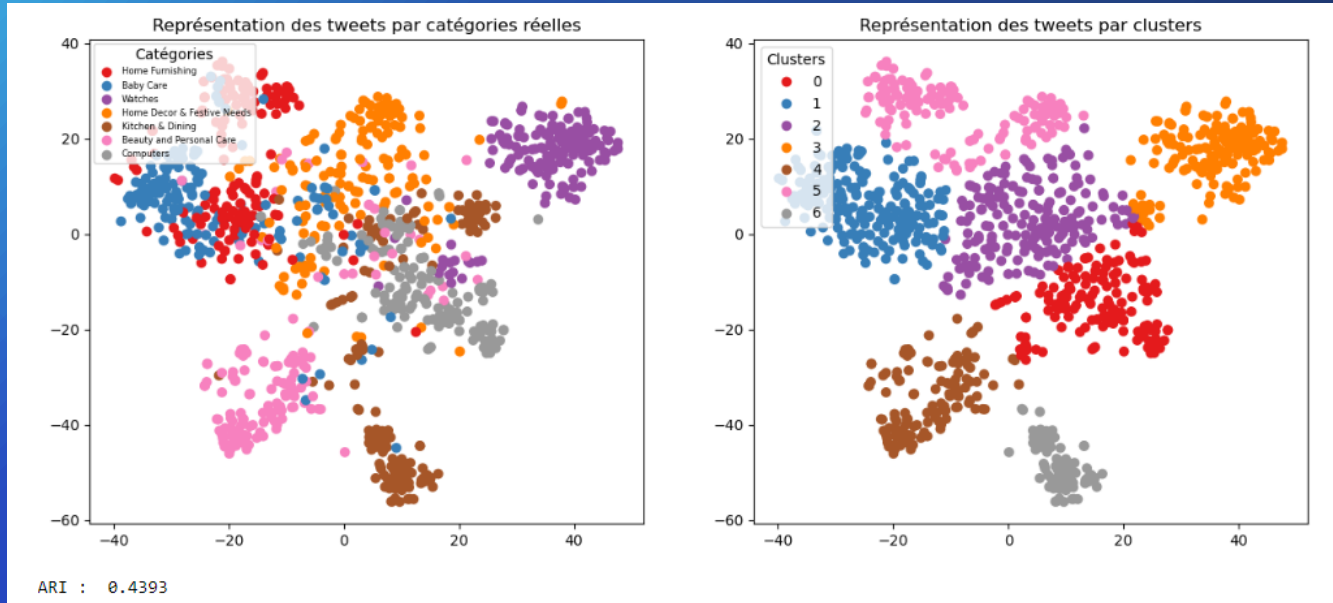
- **Etape 2** : Pooling : réduction de la taille de l'image (réduction des paramètres du réseau et préservation des principales caractéristiques de l'image).



- **Etape 3** : récupération de vecteurs après la convolution pour faire la classification.

Etude des données visuelles (CNN transfert Learning)

- CNN Transfer Learning



Synthèse sur l'étude des images

- Très faible performance de SIFT (peu de correspondance entre les catégories obtenus et les catégories réelles).
- Performance intéressante du CNN qu'on peut encore optimiser.

modèle	SIFT	CNN
ARI	0.0372	0.4393

Conclusions

- Un moteur de classification d'article basé sur une image et une description est faisable.
- Pour avoir un moteur de classification performant, il faut choisir les meilleurs modèles qu'il faudra optimiser par la recherche des meilleurs hyperparamètres et des modèles de pré-entraînement proches de notre sujet
Exemple : - traitement des textes par BERT avec les bons paramètres et un modèle qui a été entraîné sur des données textuelles proches des descriptions des articles.
 - traitement des images par CNN avec les paramètres optimums et également un modèle qui a été entraîné sur des images similaires à celle des articles.
- Choix d'un modèle de classification plus performant que le K-means
- Combinaison des 2 types de données (textes et images)

Conclusions

