# P6 - CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

21/08/2021

Etudiant : Luc Rogers
Mentor : Etienne Sanchez

# Sommaire

# Problématique

# Problématique




- Etude de faisabilité d'un moteur de classification d'articles

- Objectifs :
  - Classification à partir des descriptions objets
  - Classification à partir des images

Base de données open source :

https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours_data_scientist/Projet+-+Textimage+DAS+V2/Dataset+projet+pre%CC%81traitement+textes+images.zip

# Exploration

# Exploration



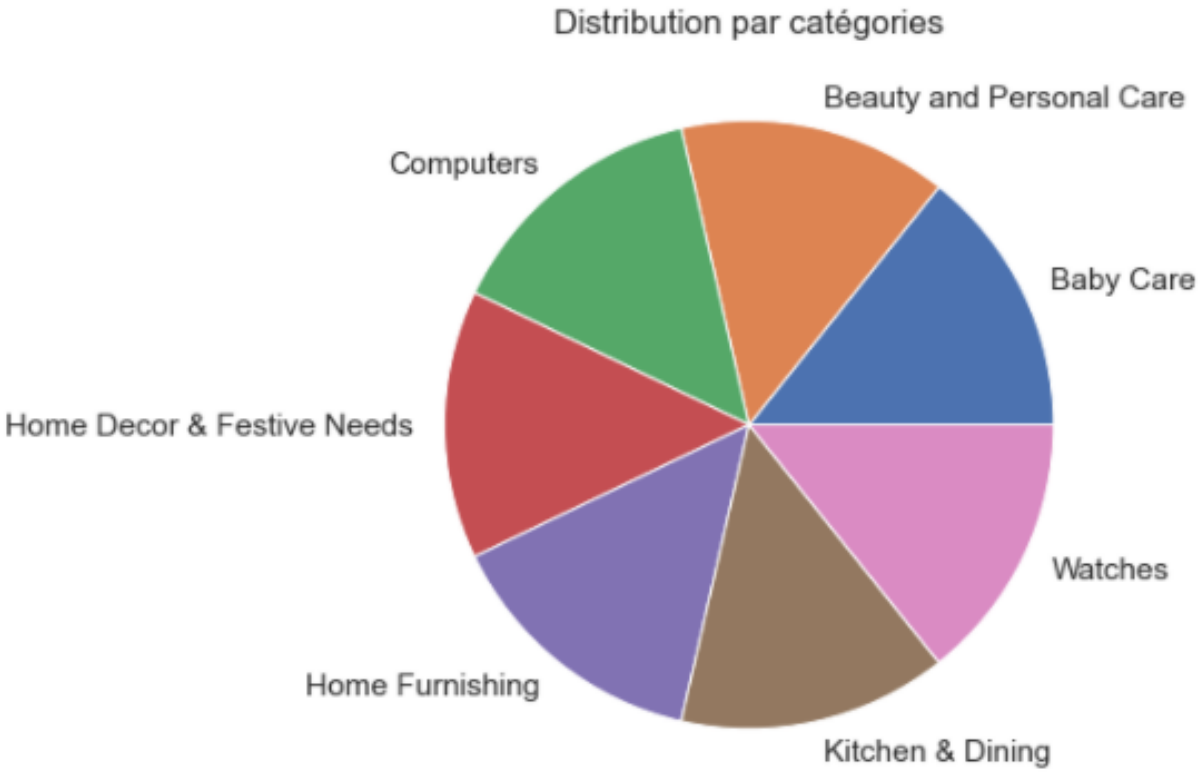
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1050 entries, 0 to 1049
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   uniq_id                 1050 non-null   object
 1   crawl_timestamp         1050 non-null   object
 2   product_url             1050 non-null   object
 3   product_name            1050 non-null   object
 4   product_category_tree   1050 non-null   object
 5   pid                     1050 non-null   object
 6   retail_price            1049 non-null   float64
 7   discounted_price        1049 non-null   float64
 8   image                   1050 non-null   object
 9   is_FK_Advantage_product 1050 non-null   bool
 10  description             1050 non-null   object
 11  product_rating          1050 non-null   object
 12  overall_rating          1050 non-null   object
 13  brand                   712 non-null    object
 14  product_specifications  1049 non-null   object
dtypes: bool(1), float64(2), object(12)
memory usage: 116.0+ KB
```

➔ Pas de valeurs manquantes sur nos variables d'intérêt

# Exploration

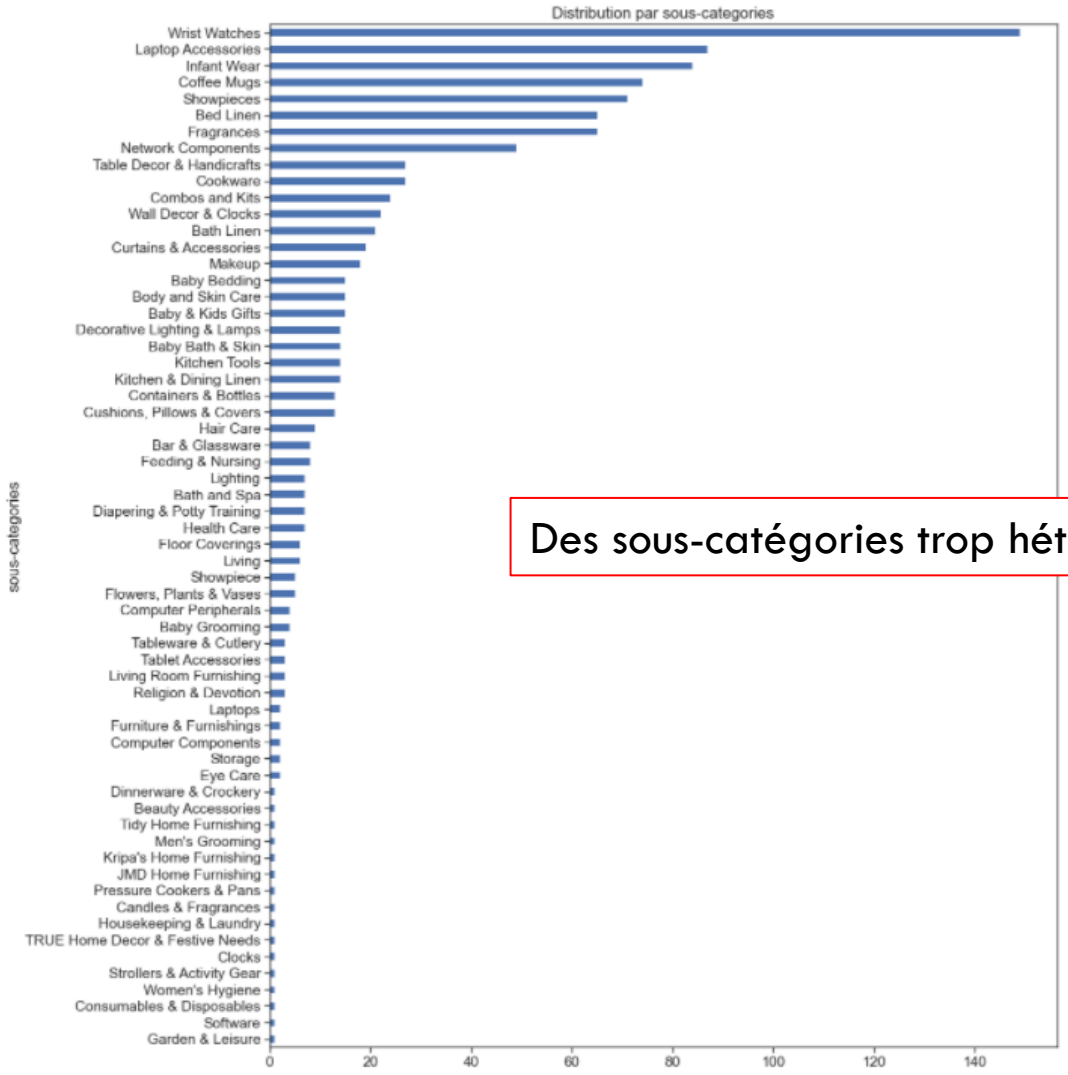





- Catégories de premier niveau bien distribuées

# Exploration






Des sous-catégories trop hétérogènes !

# Données texte

# Pre-processing



Exemple de description d'un objet:

```
0    Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegan
ce Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 Thi
s curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester fa
bric.It features an eyelet style stitch with Metal Ring.It makes the room environment romantic and lo
ving.This curtain is ant- wrinkle and anti shrinkage and have elegant apparance.Give your home a brig
ht and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These co
ntemporary 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 soo
thing prints. Bring home the elegant curtain that softly filters light in your room so that you get t
he right amount of sunlight.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Cur
tain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Ab
stract Polyester Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In
the Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material
Polyester
Name: description, dtype: object
```

# Pre-processing



- Nettoyage du texte
  - Passage en minuscules
  - Nettoyage ponctuation
  - Suppression des stop words
- Racination ou Lématisation
- Tokenisation du texte
- Extraction de features:
  - Vectorizer de type tf-idf

# Problématique



- Corpus vectorisé avec tf-idf:

| | 00 | 00 flipkartcom | 001 | 001 material | 004 | 004 kg | 006 | 006 analog | 006 online | 01 | ... | zone printed | zone uv | zoom | zoom type | zora | zora laptop | zyxel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

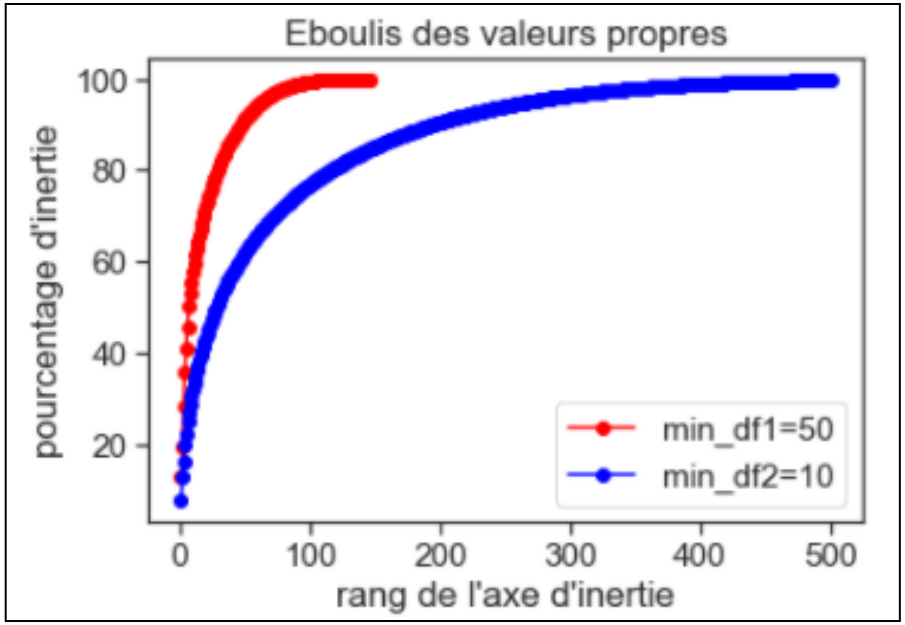
5 rows × 26746 columns

- sans seuil, tous les n-grams sont tokénisés (environ 27000 tokens)
- On fixe:
  - min_df = 0,02
  - Max_df = 0,15
- Cela nous ramène à un vocabulaire de 347 tokens

# Problématique



- Réduction de dimension




- ➔ Le choix du nombre de composantes est à adapter en fonction des seuils choisis (~90% de la variance expliquée)

# Problématique



- Tokens les plus courants (moins utiles) à travers le corpus:

| | tfidf |
|---|---|
| watch | 2.920545 |
| dimension | 2.953551 |
| model | 2.953551 |
| content | 2.967065 |
| design | 2.973891 |
| ideal | 2.980764 |
| number content | 3.015855 |
| specification | 3.052223 |
| discount | 3.052223 |
| discount genuine | 3.052223 |
| content sale | 3.052223 |
| india flipkartcom | 3.052223 |
| great discount | 3.052223 |
| general brand | 3.067148 |
| cotton | 3.097685 |
| analog | 3.105467 |
| made | 3.113310 |
| analog watch | 3.137216 |
| fabric | 3.212565 |
| detail | 3.230110 |
| package pack | 3.238998 |
| size | 3.266149 |

On prend soin de garder le terme « watch ».

On peut créer un vocabulaire plus restreint dénué de ces tokens peu utiles pour essayer d'améliorer La qualité du clustering.

# Problématique



- Tokens les plus singuliers:

| | tfidf |
|---|---|
| yet fresh | 4.866455 |
| porcelain | 4.866455 |
| please | 4.866455 |
| pick gift | 4.866455 |
| come making | 4.866455 |
| permanent | 4.866455 |
| one toodishwasher | 4.866455 |
| mug feature | 4.866455 |
| mug 55 | 4.866455 |
| making perfect | 4.866455 |

Ces tokens sont suffisamment uniques pour représenter avec efficacité un document.

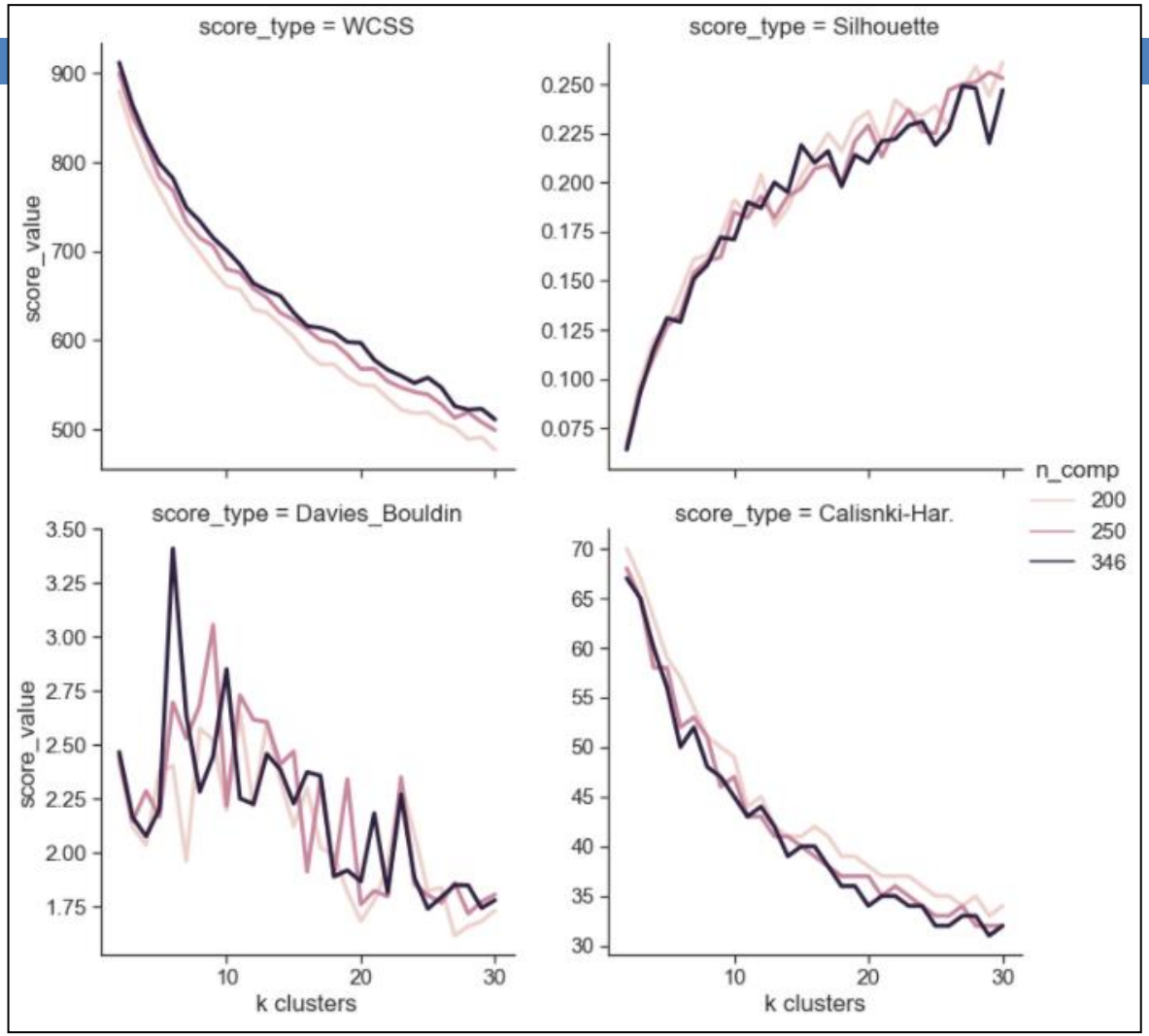
# Clustering



Kmeans:

Optimal:
- WCSS: au niveau du coude
- silhouette: le maximum
- Davies Bouldin: le minimum
- Calinski et Harabasz: le maximum

Difficile de prédire un nombre de clusters optimal sur la base des scores de validation.



# Clustering



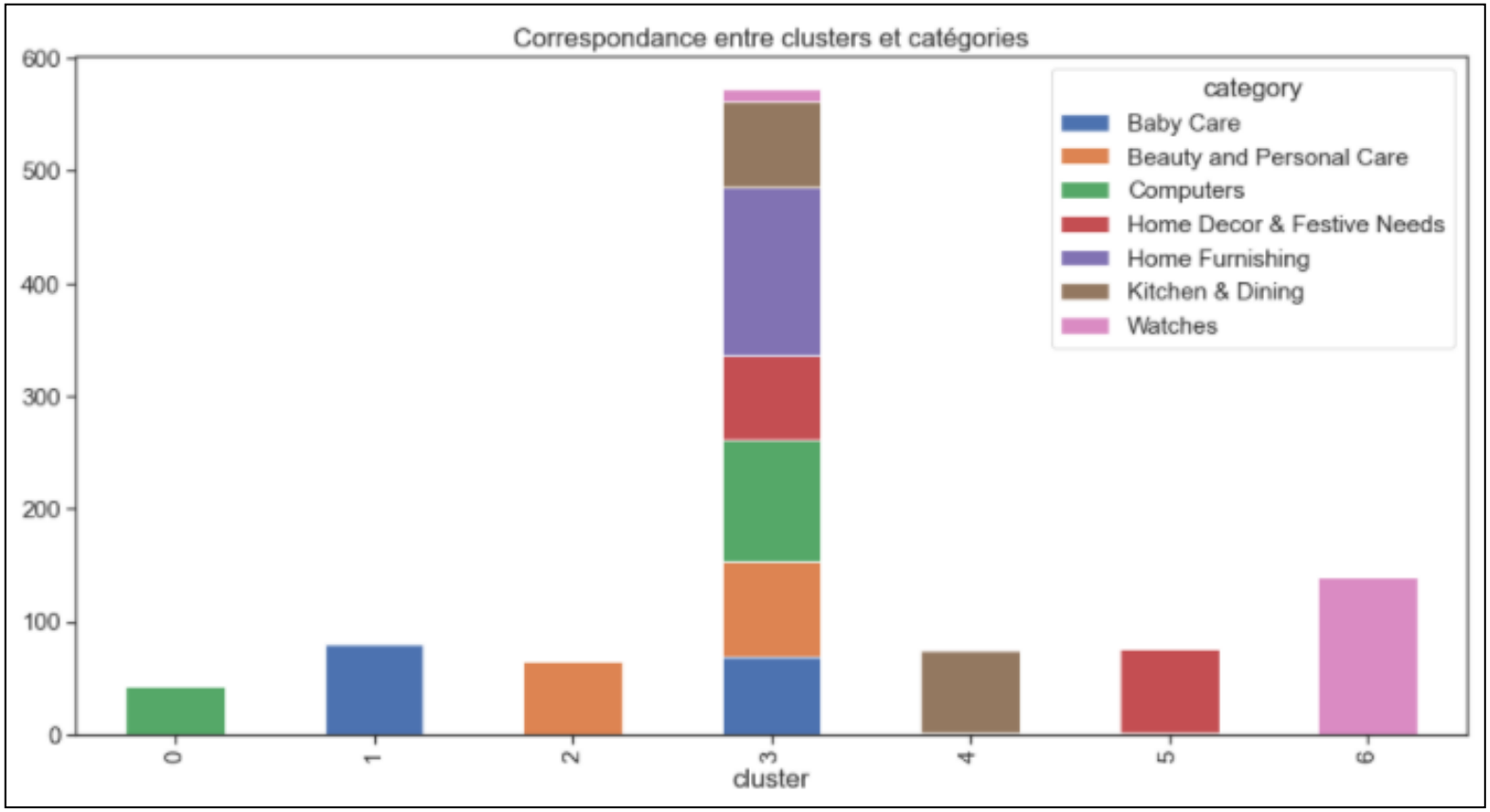
- On choisit 7 clusters
- Tokens les plus représentatifs par cluster:

```
Cluster 0:
usb | light | led | led light | usb led | fan | hub | flexible | usb usb | usb hub |

Cluster 1:
baby | baby girl | girl | detail | fabric | dress | baby boy | cotton | sleeve | boy |

Cluster 2:
set online | combo set | combo | flipkartcom buy | buy denver | denver | online 350 | 350 flipkartcom
| adidas | set combo |

Cluster 3:
product free | towel | kadhai | cotton | inch | multicolor | single | laptop | warranty | cover |

Cluster 4:
mug | ceramic | ceramic mug | prithish | coffee | perfect | rockmantra | one | mug best | coffee mug
|

Cluster 5:
showpiece | cm best | showpiece cm | handicraft | buddha | statue | brass | ganesha | gift | exotic i
ndia |

Cluster 6:
watch | analog | analog watch | men | watch men | discount genuine | discount | great discount | indi
a flipkartcom | watch woman |
```

# Clustering



- Correspondance avec les catégories:

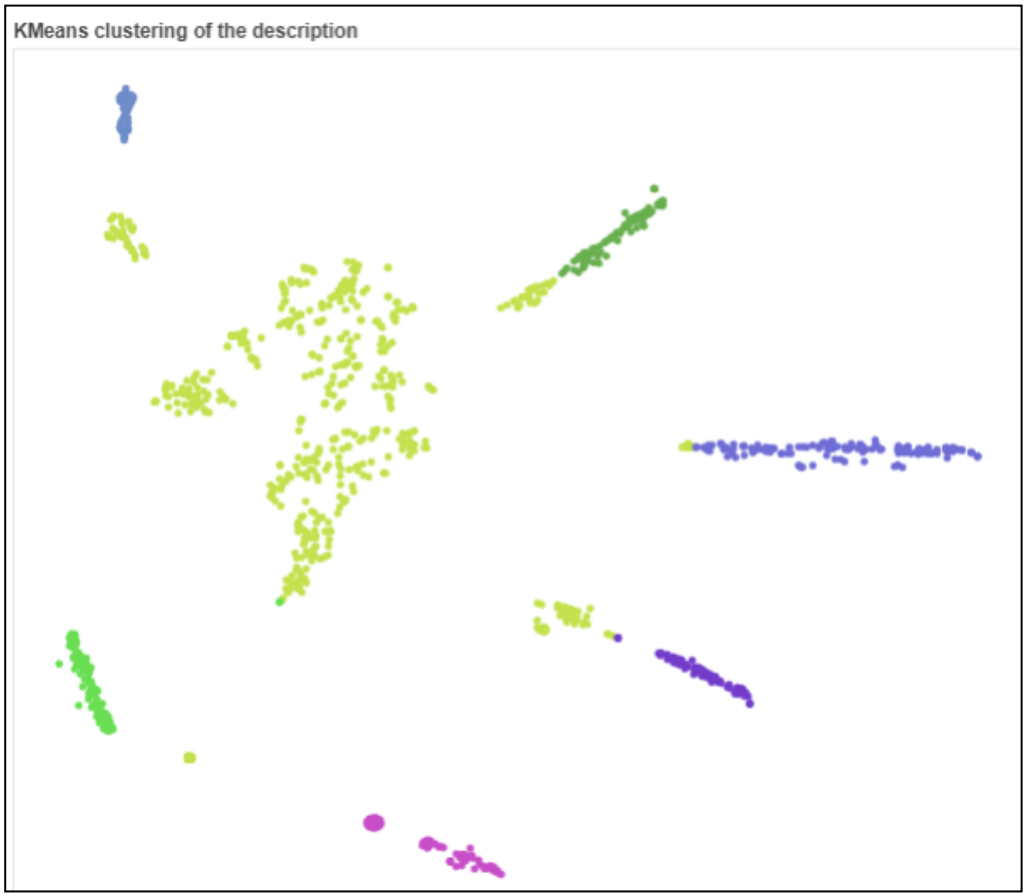



- ➔ Les items du cluster 3 n'arrivent pas à être départagés

# Clustering



- Représentation 2D avec t-SNE:

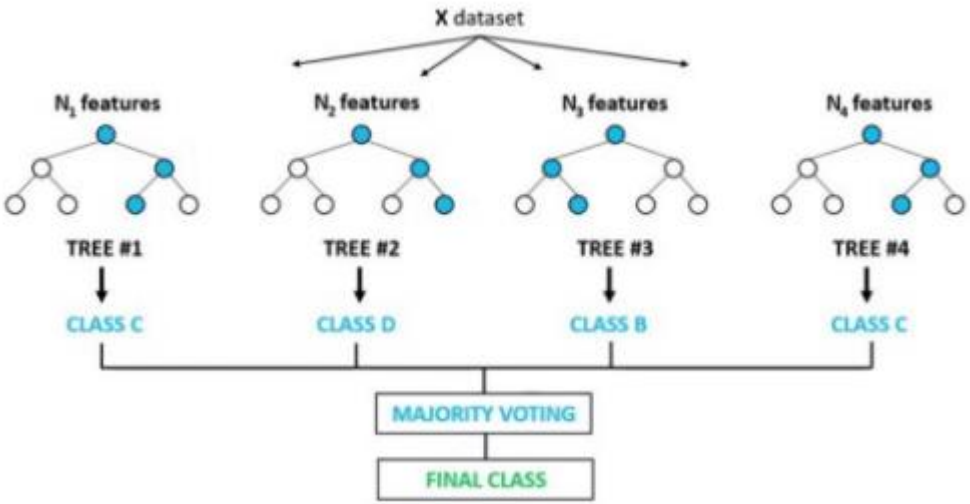



- ➔ On retrouve un ensemble d'items qui ne peuvent être facilement départagés

# Clustering



- Approche supervisée avec un classifieur Random Forest




```
Paramètres optimaux: {'max_depth': 55, 'n_estimators': 35}
Accuracy: 0.93
Temps de calcul: 5.32s
```

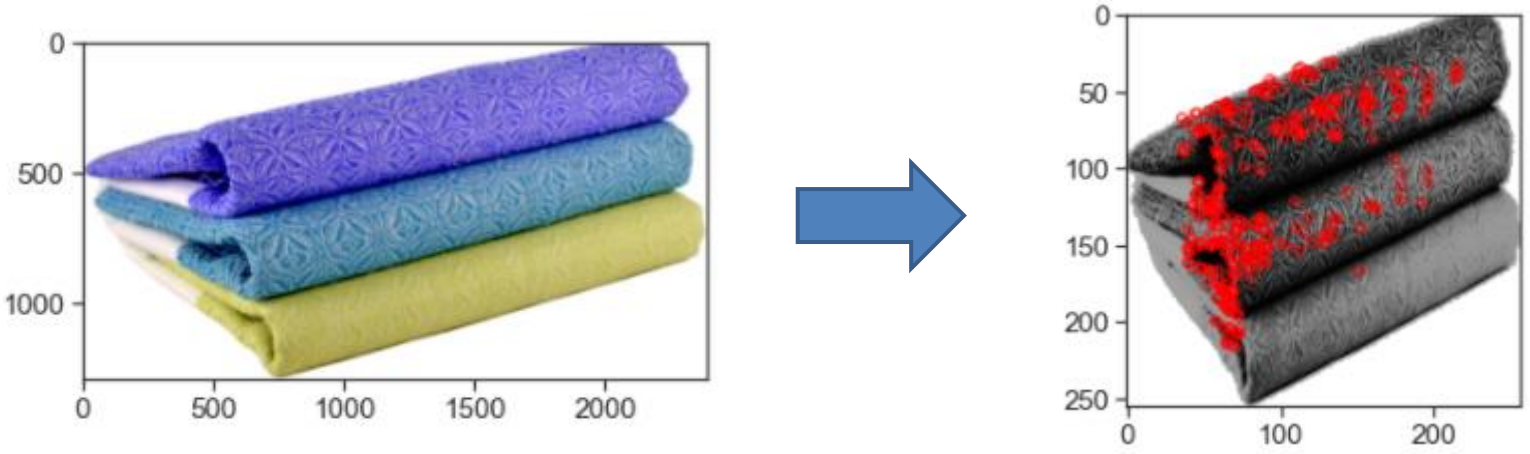
➔ Une classification supervisée permet d'atteindre 93% de précision sur le jeu de données test.

# Données images

# ORB



- Pre-processing:
  - Grayscale
  - Resizing: 256*256
  - Egalisateur histogramme
- Détection des keypoints avec ORB:

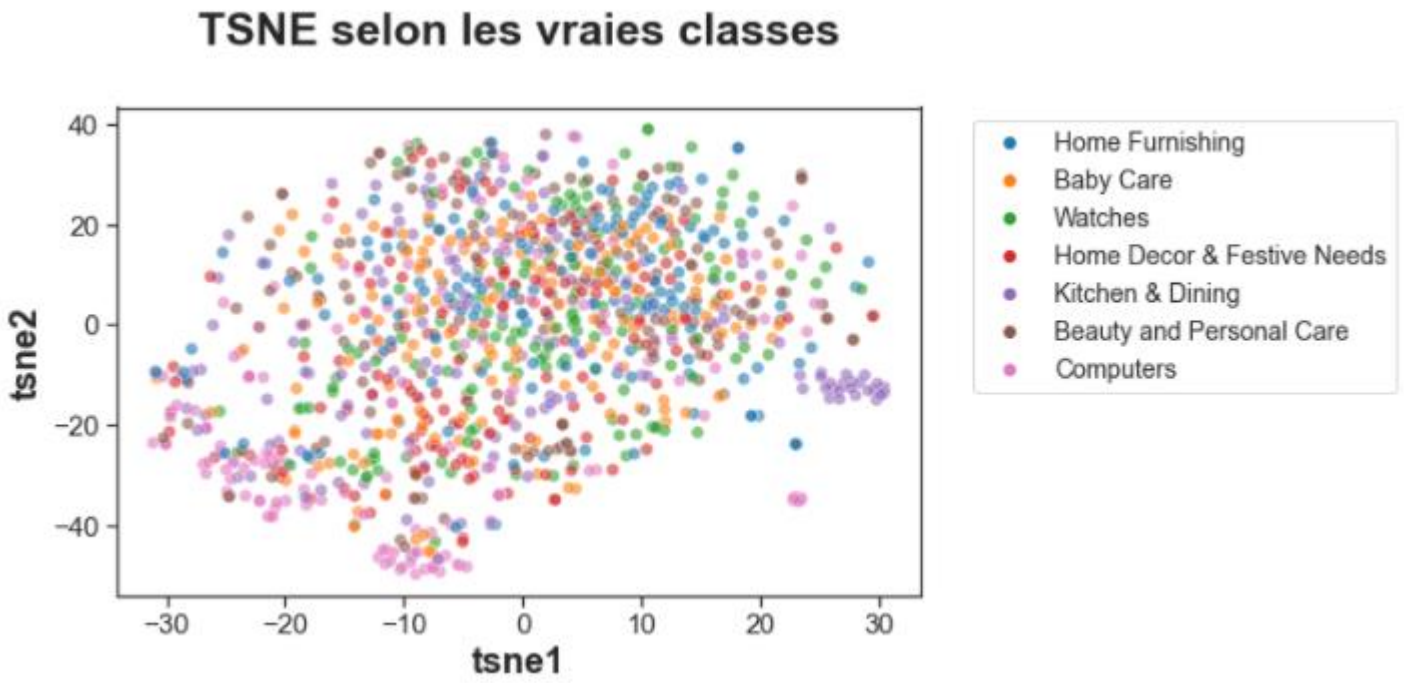



- Liste de descripteurs (par défaut 500 descripteurs par image)
- Kmeans sur la liste de descripteurs

# ORB



- Réduction de dimensions PCA: 652 ➔ 532 composantes
- Réduction de dimensions t-SNE: 2 dimensions
- Représentation 2D:

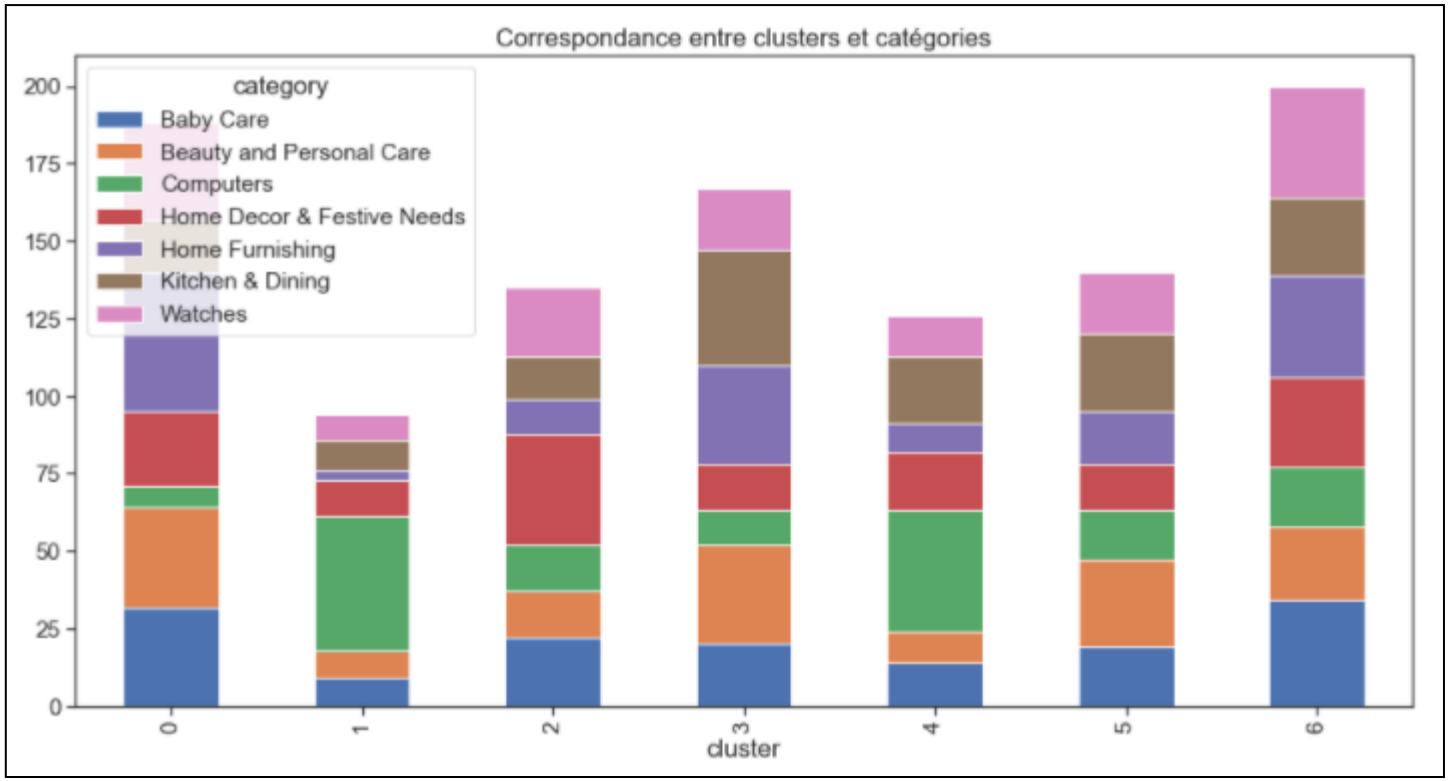



➔ Les clusters obtenus via ORB sont peu concluants…

# ORB



- Correspondance avec les catégories:

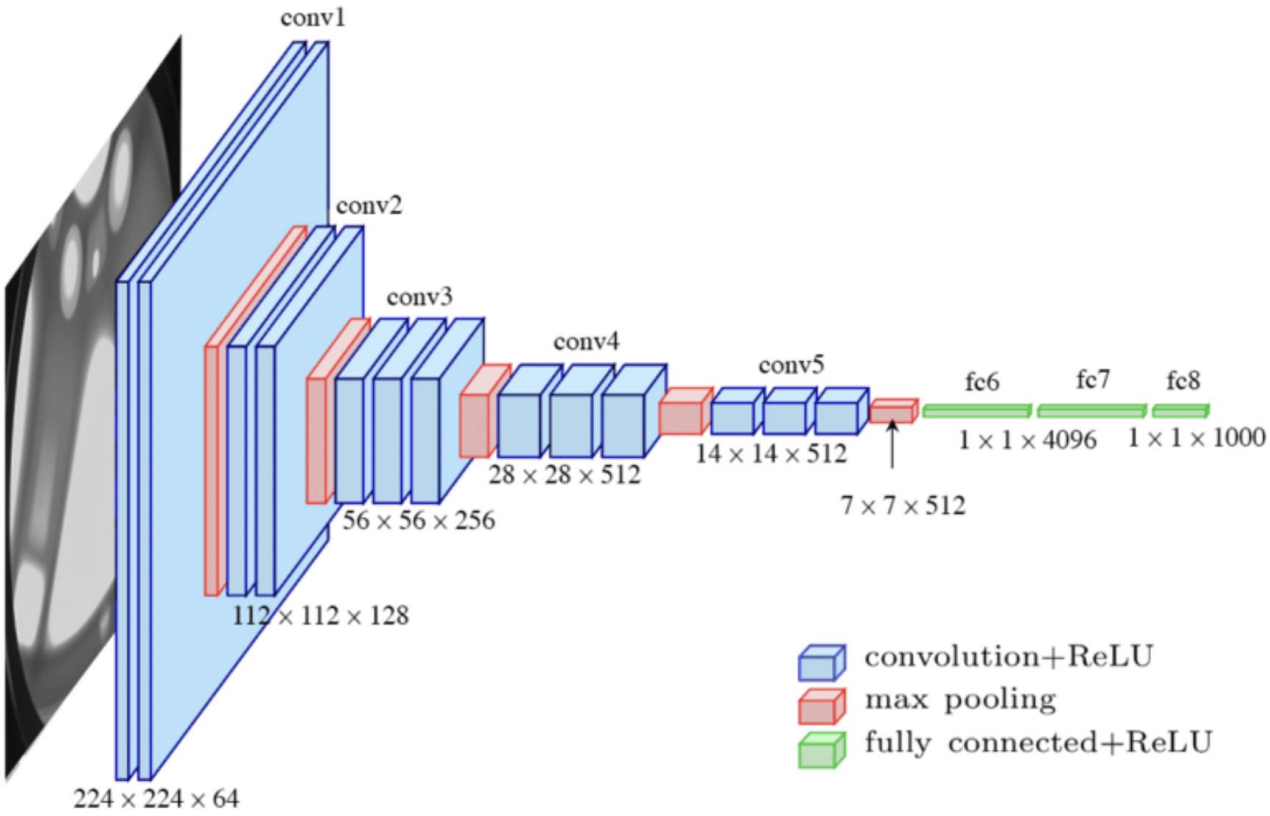



➔ Les clusters obtenus via ORB sont peu concluants…

# Transfer learning



- Réseau de neurones convolutionnel pré-entraîné:




On ne modifie que la couche fully-connected, que l'on entraîne pour notre problème de classification sur notre jeu de données.

# Transfer learning



- Le transfer learning est très efficace et permet de bénéficier d'un réseau de neurones convolutionnel entraîné sur une très longue période.

```
model.evaluate(X_scaled_test,y_test)
5/5 [==============================] - 7s 1s/step - loss: 0.6630 - acc: 0.8038
[0.6629988551139832, 0.8037974834442139]
```

➔ On atteint les 80% de précision avec un temps de calcul minimal.

# Conclusion



- Première approche en clustering non supervisé peu efficace sur ce jeu de données
- Nécessité de bien paramétrer les hyperparamètres:
  - seuils min_df et max_df
  - n-grammes
- t-SNE permet de visualiser les résultats sous forme 2D
- Approche supervisée plus efficace:
  - 93% de précision avec les données textes
  - 80% de précision avec les données images
- Transfer learning permet de bénéficier d'un modèle longuement entraîné et s'avère très efficace pour notre problème

*Merci de votre attention*

# Annexe