Classifiez automatiquement des biens de consommation

Data Science | Projet 6

Fırat Yasar

24/12/2021

Sommaire

Présentation

Présentation de la problématique

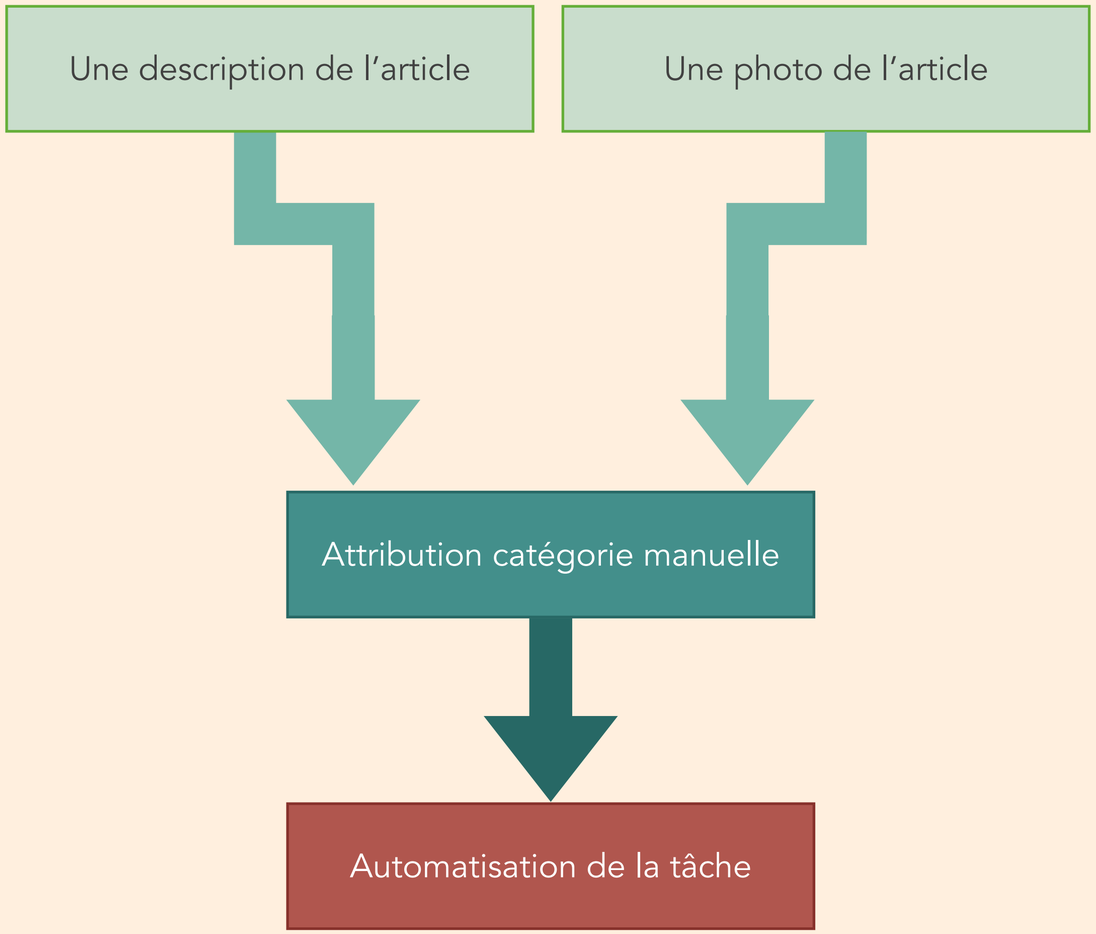
- **L'objectif :**

  Notre objectif est de réaliser **une première étude de faisabilité** d'un moteur de classification d'articles basé sur une image et une description pour l'automatisation de l'attribution de la catégorie de l'article.

- **La base de données :**

  Pour cela, nous avons à notre disposition la base de données :

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



Découverte du jeu de données

DATA

1050 lignes 15 colonnes

#	Column
0	uniq_id
1	crawl_timestamp
2	product_url
3	product_name
4	product_category_tree
5	pid
6	retail_price
7	discounted_price
8	image
9	is_FK_Advantage_product
10	description
11	product_rating
12	overall_rating
13	brand
14	product_specifications

product_category_tree

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

image

description

0. Home Furnishing

Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multi color Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 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 ant- wrinkle and anti shrinkage and h ave elegant apparance.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 fi rst 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 pri nts. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sun light.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract 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

Le premier article dans le jeu de données

Analyse des catégories

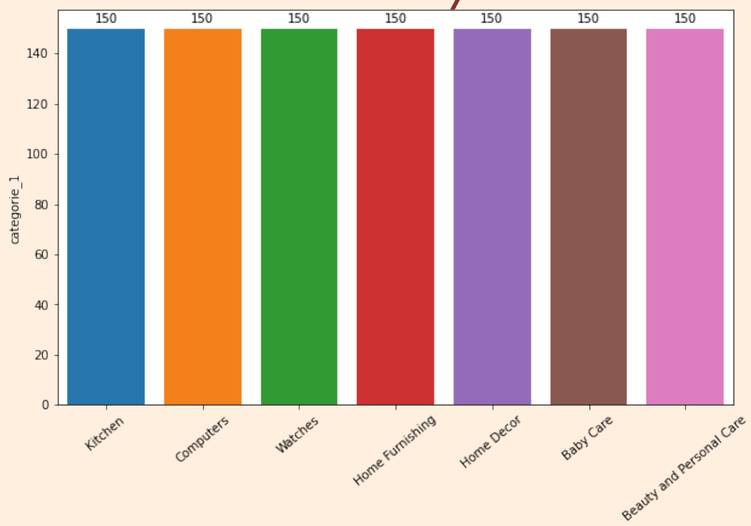
`product_category_tree`

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

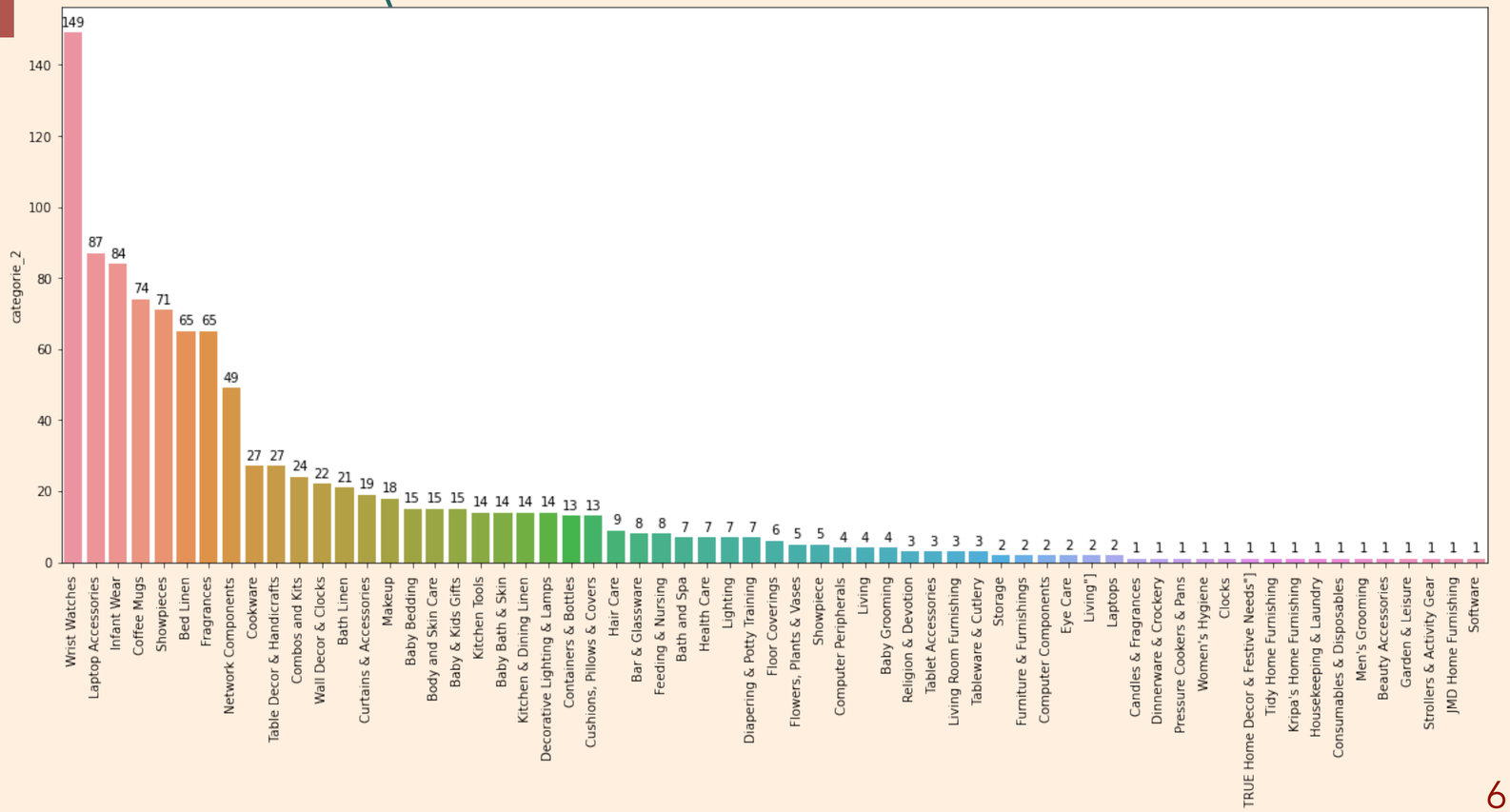
…5 autres sous-catégories détaillées

La catégorie principale
*
cuisine
ordinateurs
montres
ameublement
décoration de maison
soins bébé
beauté

categorie_1



categorie_2



0. Home Furnishing

Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multi color Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 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 ant- wrinkle and anti shrinkage and h ave elegant apparance.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 fi rst 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 pri nts. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sun light.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract 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

---

1. Baby Care

Specifications of Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Yellow, Blue) Bath Towel Features Machine Washabl e Yes Material Cotton Design Self Design General Brand Sathiyas Type Bath Towel GSM 500 Model Name Sathiyas cotton bath towel Ideal For Men, Women, Boys, Girls Model ID asvtwl322 Color Red, Yellow, Blue Size Mediam Dimensions Len gth 30 inch Width 60 inch In the Box Number of Contents in Sales Package 3 Sales Package 3 Bath Towel

---

2. Baby Care

Key Features of Eurospa Cotton Terry Face Towel Set Size: small Height: 9 inch GSM: 360,Eurospa Cotton Terry Face Towel Set (20 PIECE FACE TOWEL SET, Assorted) Price: Rs. 299 Eurospa brings to you an exclusively designed, 100% s oft cotton towels of export quality. All our products have soft texture that takes care of your skin and gives you that enriched feeling you deserve. Eurospa has been exporting its bath towels to lot of renowned brands for last 1 0 years and is famous for its fine prints, absorbency, softness and durability. NOTE: Our product is 100% cotton, so it is susceptible to shrinkage. Product color may vary from the picture. Size may vary by ±3% WASH CARE: Wash i n cold Water, Do not Iron, Do not Bleach, Flat dry, Wash before first use. SIZE- FACE TOWEL - 23 cms X 23 cms.,Spe cifications of Eurospa Cotton Terry Face Towel Set (20 PIECE FACE TOWEL SET, Assorted) Bath Towel Features Materia l Cotton Terry Design SHUVAM General Brand Eurospa GSM 360 Type Face Towel Set Model Name SHUVAM20PCFTSETASSORTED Ideal For Boys, Girls, Men, Women Model ID SHUVAM20PCFTSETASSORTED Size small Color Assorted Dimensions Weight 350 g Length 9 inch Width 9 inch In the Box Number of Contents in Sales Package 20 Sales Package 20 PIECE FACE TOWEL S ET

---

3. Home Furnishing

Key Features of SANTOSH ROYAL FASHION Cotton Printed King sized Double Bedsheet Royal Bedsheet Perfact for Wedding & Gifting,Specifications of SANTOSH ROYAL FASHION Cotton Printed King sized Double Bedsheet (1 Bedsheet,2 Pillow C over, Multicolor) General Brand SANTOSH ROYAL FASHION Machine Washable Yes Type Flat Material Cotton Model Name Go ld Design Royal Cotton Printed Wedding & Gifted Double Bedsheet With 2 Pillow cover Model ID goldbedi-38 Color Mul ticolor Size King Fabric Care Machine Wash, Do Not Bleach Dimensions Flat Sheet Width 90 inch / 230 cm Fitted Shee t Width 228 cm Pillow Cover Width 16 inch / 43 cm Pillow Cover Length 28 inch / 72 cm Fitted Sheet Depth 280 cm Fi tted Sheet Length 278 cm Flat Sheet Depth 282 cm Flat Sheet Length 110 inch / 280 cm In the Box Number of Contents in Sales Package 1 Sales Package 1 Bedsheet,2 Pillow Cover

---

4. Home Furnishing

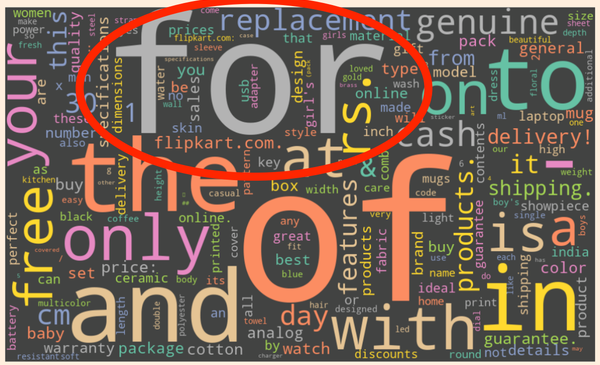
Key Features of Jaipur Print Cotton Floral King sized Double Bedsheet 100% cotton,Jaipur Print Cotton Floral King sized Double Bedsheet (1 bed sheet 2 pillow cover, White) Price: Rs. 998 This nice bed sheet made up of 100% cotto n to give you comfort. This bed sheet is hand printed. This bedsheet gives nice look to your room And its fast col our and good quality gives this bedsheet long life.,Specifications of Jaipur Print Cotton Floral King sized Double Bedsheet (1 bed sheet 2 pillow cover, White) General Machine Washable Yes Brand Jaipur Print Type Flat Model Name jaipur117 Material Cotton Thread Count 140 Model ID jaipur117 Fabric Care machinewash, do not bleach Size King Col or White Warranty waranty of the product only for manufacturing defect only and product will exchange onle when it is not used and returne its origional packing Dimensions Flat Sheet Width 86 inch / 220 cm Fitted Sheet Width 0 cm Pillow Cover Width 17 inch / 45 cm Pillow Cover Length 29 inch / 75 cm Weight 900 g Fitted Sheet Depth 0 cm Fitted Sheet Length 0 cm Flat Sheet Depth 0.2 cm Flat Sheet Length 104 inch / 265 cm In the Box Number of Contents in Sal es Package 1 Sales Package 1 bed sheet 2 pillow cover

---

Partie I : Données textuelles

Les étapes du traitement du corpus - text mining

Nuage de mots dans la colonne tokenized_desc_0 (top 200)



1

Tokenization

Suppression des stop words

2

Nuage de mots dans la colonne clean_tokenized_desc_0 (top 300)

Suppression des mots les plus fréquents

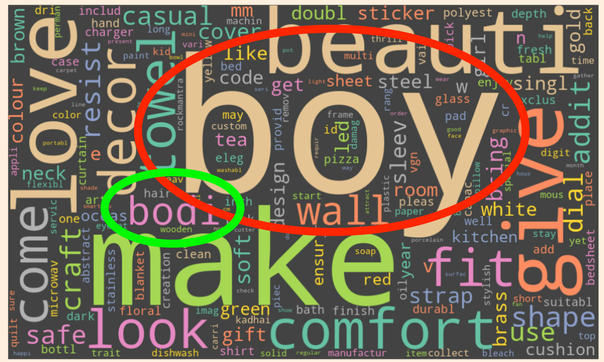
3

Nuage de mots dans la colonne better_clean_tokenized_desc_0 (top 300)

Racinisation (Stemming)

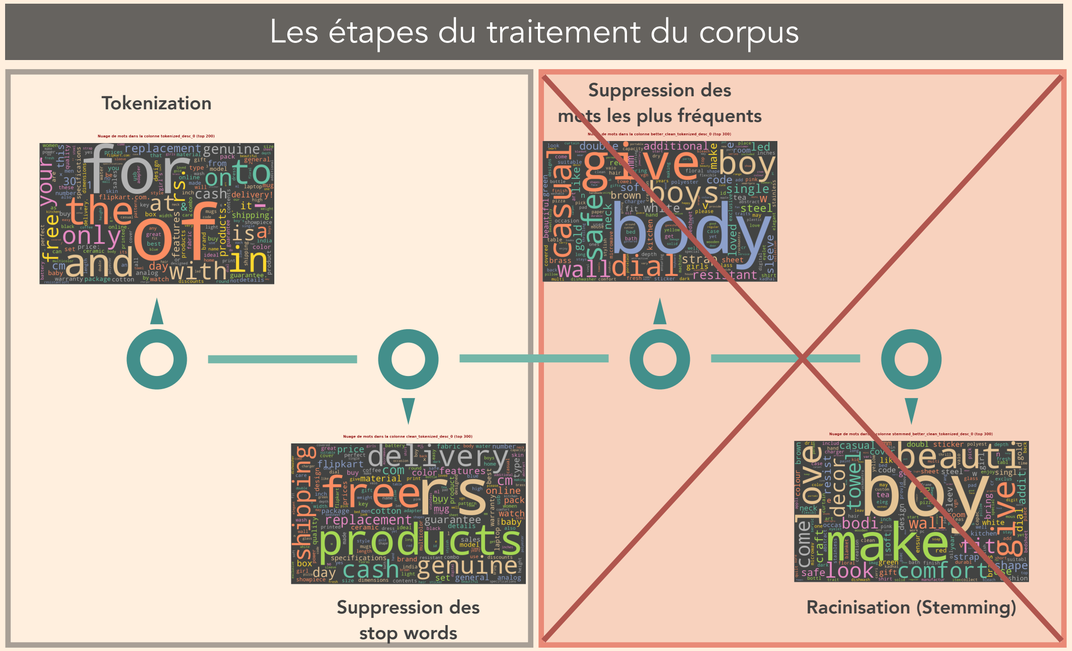
4

Nuage de mots dans la colonne stemmed_better_clean_tokenized_desc_0 (top 300)



Analyse performance

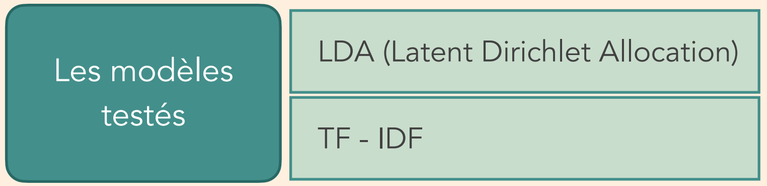
- Utilisation de **la régression logistique** afin de mesurer la performance de l'opération sur le corpus
- Le critère de performance choisi est **accuracy.**
- **Vectorization : TfidfVectorizer**



	L'Opération sur le corpus	Accuracy
1.	Tokenization	57 %
2.	Suppression des stop words	91 %
3.	Suppression des mots les plus fréquents	89 %
4.	Racinisation (Stemming)	89 %

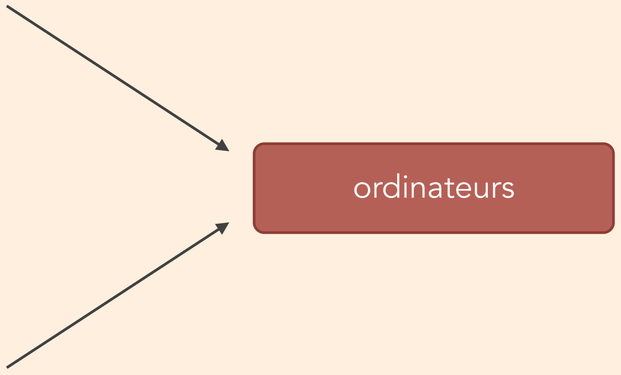
Deux modèles non supervisés

- Le modèle **LDA** est un algorithme d'apprentissage non supervisé.
- **TF-IDF** est une méthode qui prend en compte la fréquence inverse de document (*inverse document frequency*). Cela est une mesure de l'importance du terme dans l'ensemble du corpus.
- Les mots sont regroupés par rapport à leur distribution dans les descriptions.
- Les groupes sont déterminés par des composantes qui forment **le sujet** (topics en dessous) du groupe.



- Les composantes ne peuvent **pas** avoir une signification sémantique.

```
Topic 0:
home hair price showpiece rs cm art towel beautiful brass
Topic 1:
warranty usb adapter power laptop battery replacement light product quality
Topic 2:
mug ceramic coffee perfect mugs gift material safe loved rockmantra
Topic 3:
skin oil traits shampoo wall ml soap type used applied
Topic 4:
products free rs delivery genuine shipping cash buy 30 day
Topic 5:
cm pack baby features color specifications general cotton package number
Topic 6:
laptop skin print shapes pad set mouse warranty combo multicolor
```

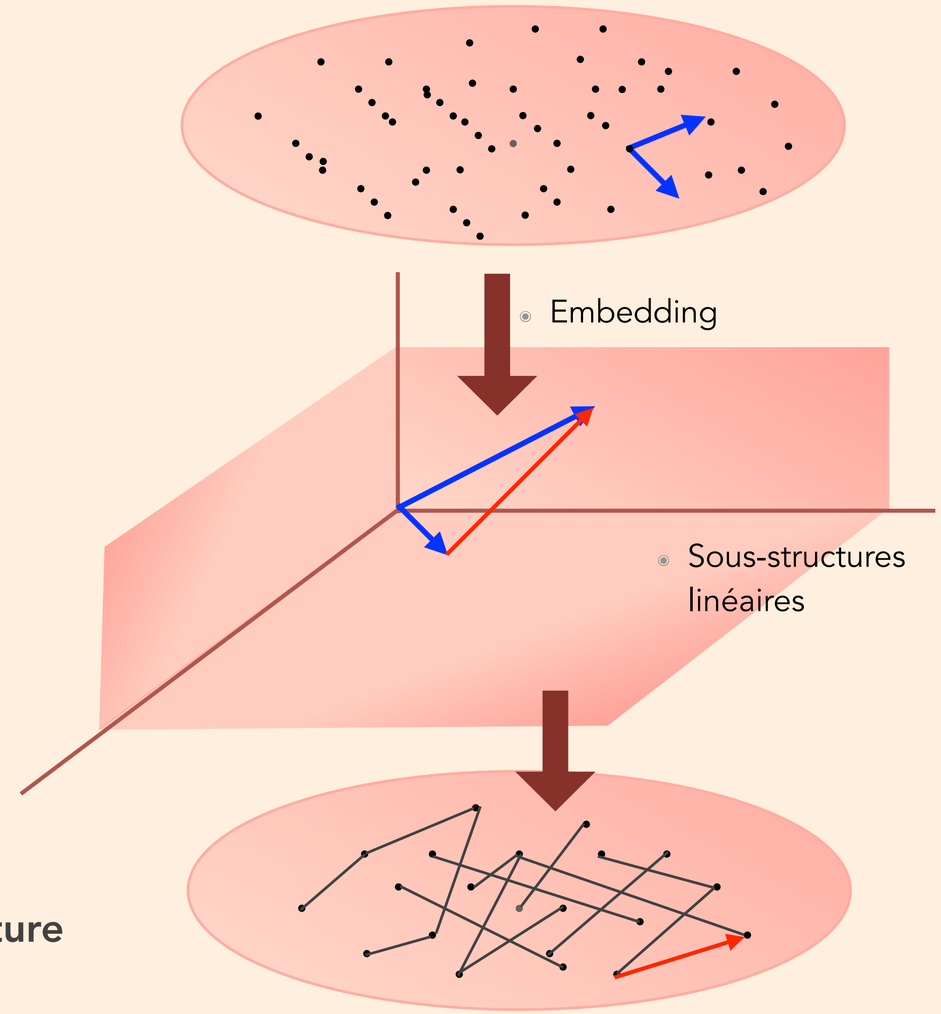


Plongement de mots via GLOVE

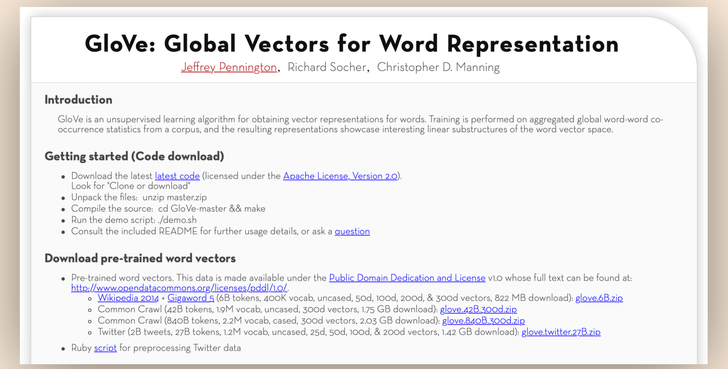
C

(V, m)

C’ avec
une structure



- Word embedding
- Le modèle GloVe https://nlp.stanford.edu/projects/glove/



GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

Introduction

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

Getting started (Code download)

- Download the latest latest code (licensed under the Apache License, Version 2.0). Look for "Clone or download"
- Unpack the files: unzip master.zip
- Compile the source: cd GloVe-master && make
- Run the demo script: ./demo.sh
- Consult the included README for further usage details, or ask a question

Download pre-trained word vectors

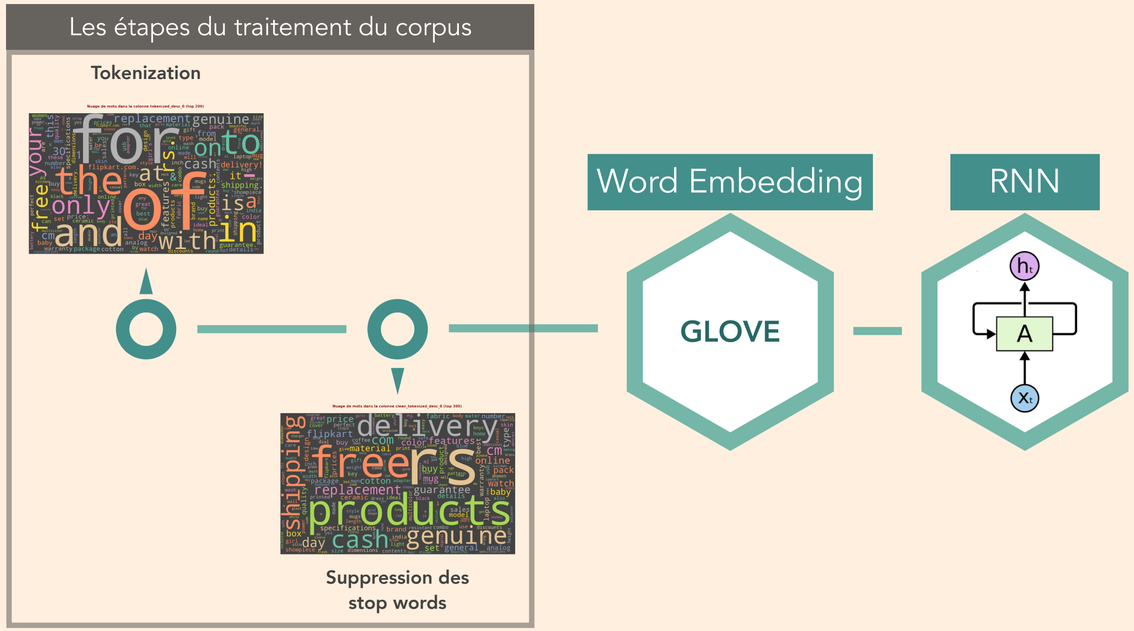
- Pre-trained word vectors. This data is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: http://www.opendatacommons.org/licenses/pddl/1.0/.
  - Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download): glove.6B.zip
  - Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB download): glove.42B.300d.zip
  - Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download): glove.840B.300d.zip
  - Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 25d, 50d, 100d, & 200d vectors, 1.42 GB download): glove.twitter.27B.zip
- Ruby script for preprocessing Twitter data

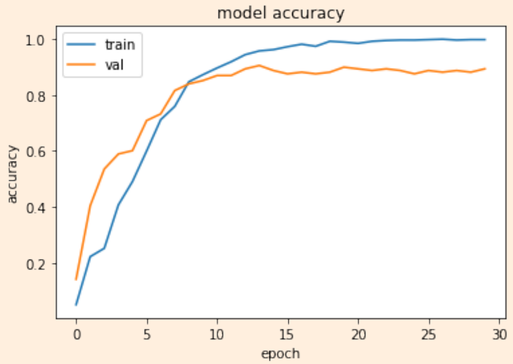
	Dimension	Accuracy
1.	50D	85 %
2.	100D	88 %
3.	200D	90 %
4.	300D	92 %

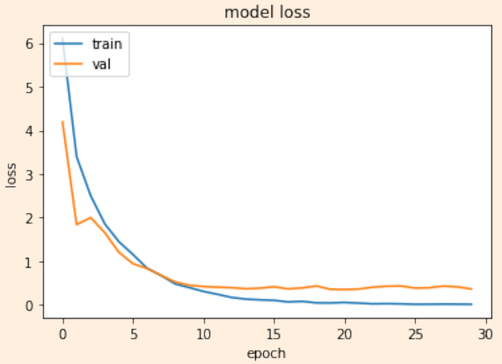
Réseau de neurones convolutifs - RNN (Recurrent Neural Networks)

- Les RNN représentent la famille de réseau de neurones qui traite les données de manière séquentielle.

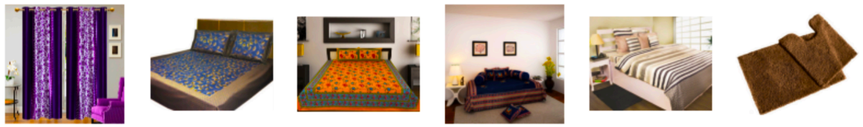
Algorithme	Accuracy
RNN	93,3 %







Home Furnishing



Baby Care

Watches

Home Decor

Kitchen

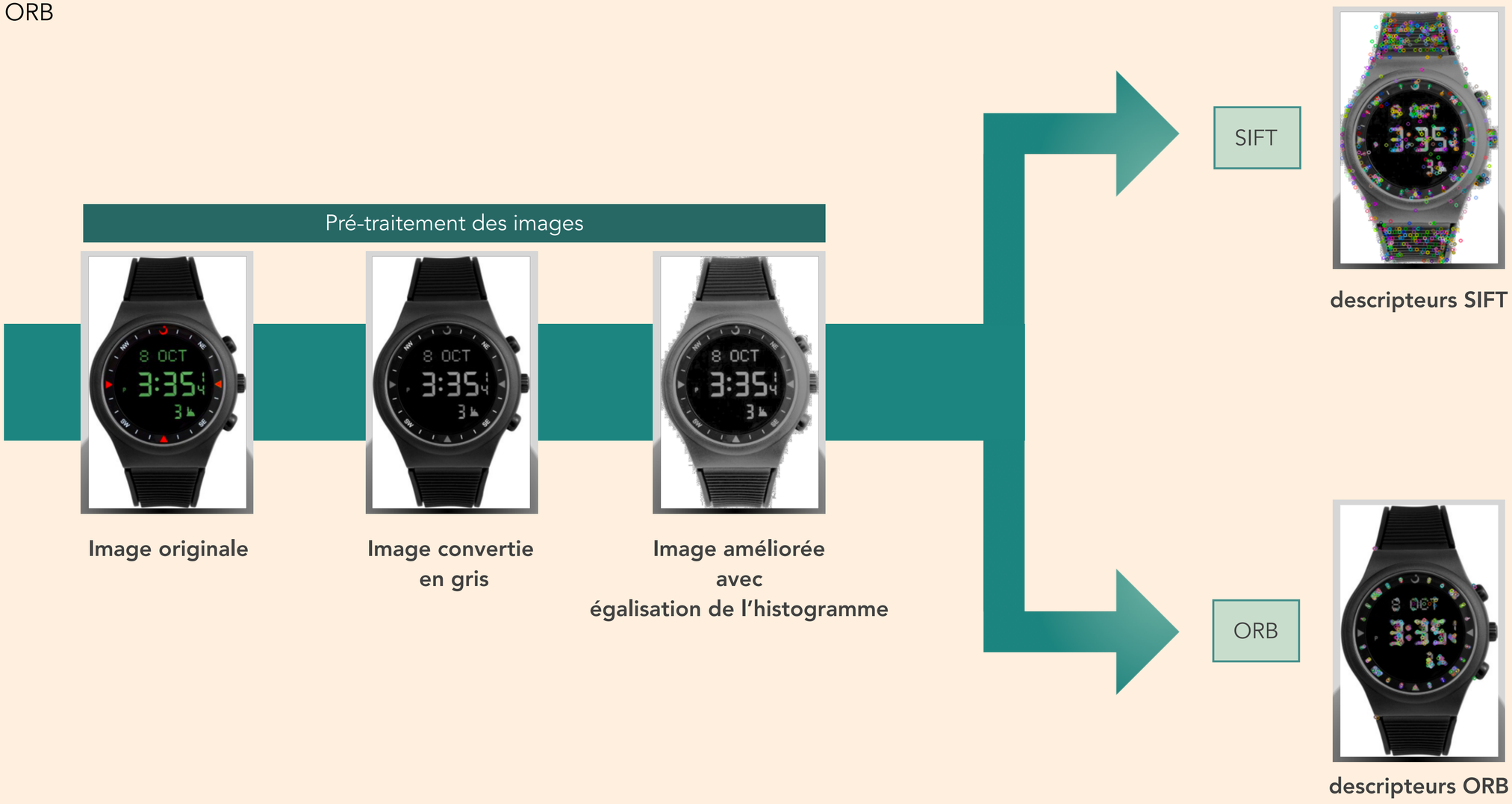
Beauty and Personal Care

Computers

Partie II : Données visuelles

Pré-traitement des images pour les algorithmes non supervisés

- SIFT
- ORB



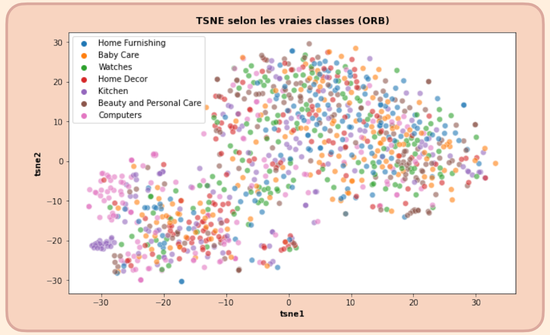
Computer Vision via SIFT et ORB

Pré-traitement des images

Convertir en niveaux de gris
*
Égalisation de l'histogramme
*
Construction des descripteurs **SIFT** (et **ORB**)

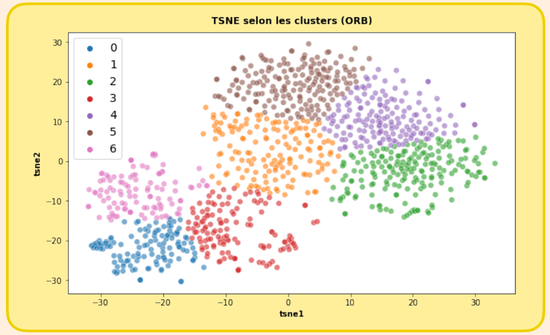
Réduction de dimension

PCA (99 %)
*
t-SNE



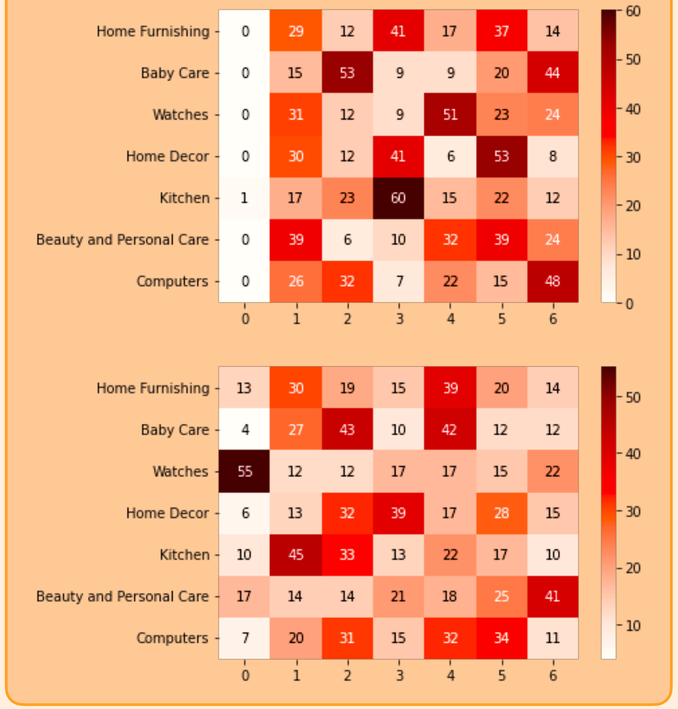
Clustering

Mise en œuvre de l'algorithme **KNN**



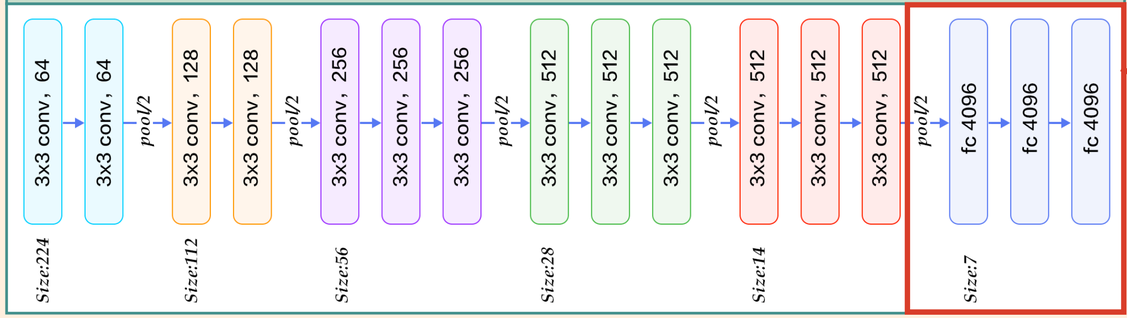
Evaluation des modèles

Analyse des mesures : similarité entre les catégories et les clusters
*
ARI Score
*
Confusion matrice

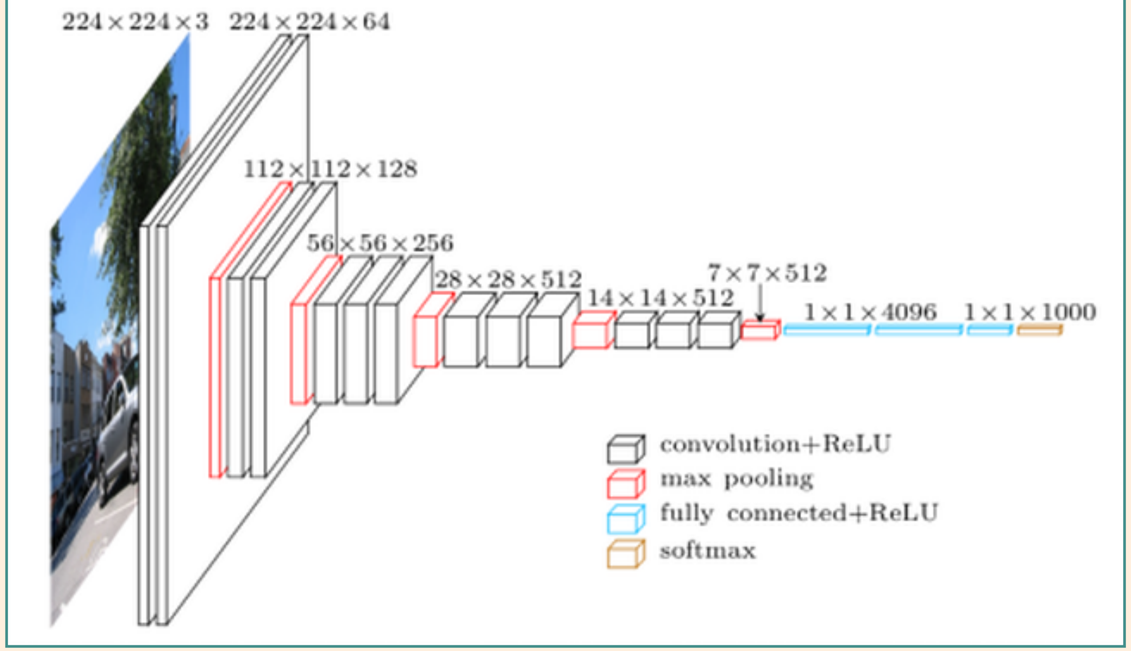


Computer Vision via Deep Learning (Stratégie #1)

Architecture de CNN

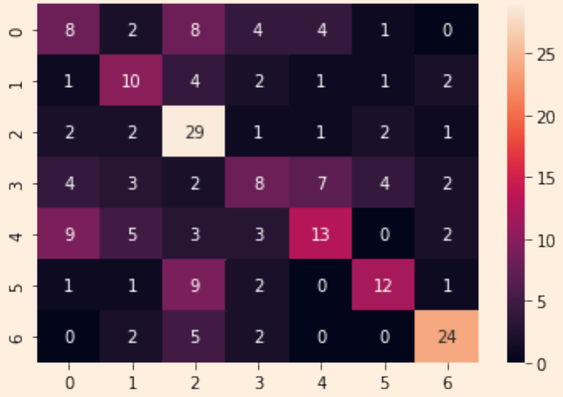


Représentation 3D de l'architecture

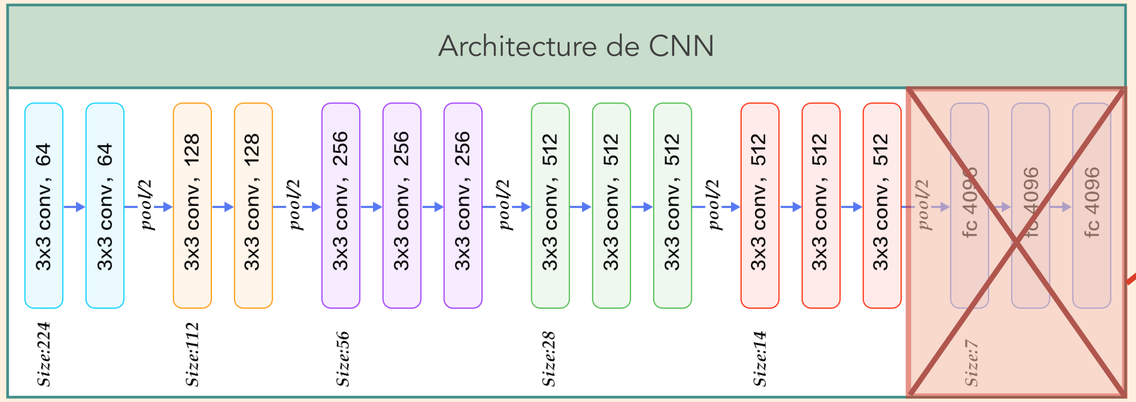


- Transformation des images en taille 224x224 (et 128x128)
- CNN (réseau de neurones convolutif)
- La dernière couche appelée ***fully-connected*** permet de classifier l'image en entrée du réseau

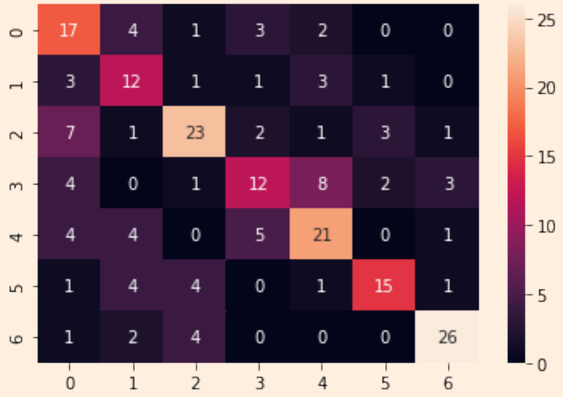
Algorithme	Acc. (taille=224)	Acc. (taille=128)
CNN (Stratégie #1)	50 %	37 %



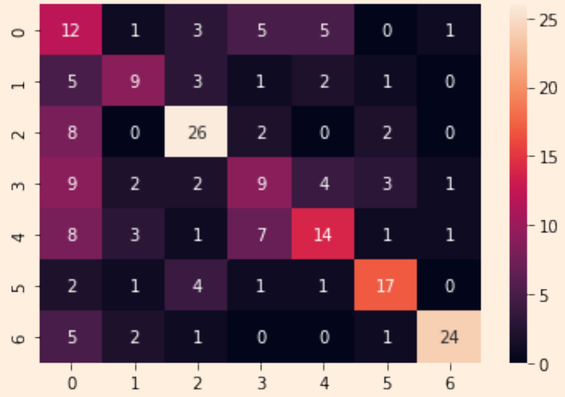
Computer Vision via Deep Learning (Stratégie #2)



- Suppression de la dernière couche
- KNN
- Random Forest Classifier



RF



KNN

Algorithme	Acc. (taille=224)	Acc. (taille=128)
Random Forest	60 %	60 %
KNN	52 %	60 %

Computer Vision via Transfer Learning (Stratégie #3)

Architecture du modèle **VGG-16**

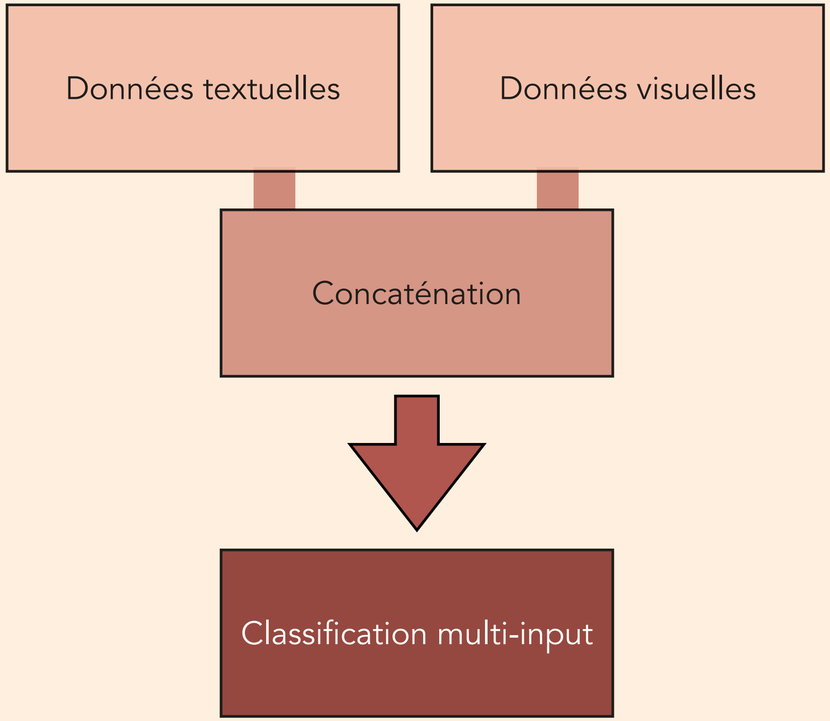
```
Model: "model_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_3 (InputLayer)        [(None, 300, 300, 3)]     0
 block1_conv1 (Conv2D)       (None, 300, 300, 64)      1792
 block1_conv2 (Conv2D)       (None, 300, 300, 64)      36928
 block1_pool (MaxPooling2D)  (None, 150, 150, 64)      0
 block2_conv1 (Conv2D)       (None, 150, 150, 128)     73856
 block2_conv2 (Conv2D)       (None, 150, 150, 128)     147584
 block2_pool (MaxPooling2D)  (None, 75, 75, 128)       0
 block3_conv1 (Conv2D)       (None, 75, 75, 256)       295168
 block3_conv2 (Conv2D)       (None, 75, 75, 256)       590080
 block3_conv3 (Conv2D)       (None, 75, 75, 256)       590080
 block3_pool (MaxPooling2D)  (None, 37, 37, 256)       0
 block4_conv1 (Conv2D)       (None, 37, 37, 512)       1180160
 block4_conv2 (Conv2D)       (None, 37, 37, 512)       2359808
 block4_conv3 (Conv2D)       (None, 37, 37, 512)       2359808
 block4_pool (MaxPooling2D)  (None, 18, 18, 512)       0
 block5_conv1 (Conv2D)       (None, 18, 18, 512)       2359808
 block5_conv2 (Conv2D)       (None, 18, 18, 512)       2359808
 block5_conv3 (Conv2D)       (None, 18, 18, 512)       2359808
 block5_pool (MaxPooling2D)  (None, 9, 9, 512)         0
 global_average_pooling2d_2   (None, 512)              0
 (GlobalAveragePooling2D)
 dense_6 (Dense)             (None, 1024)              525312
 dropout_2 (Dropout)         (None, 1024)              0
 dense_7 (Dense)             (None, 1024)              1049600
 dense_8 (Dense)             (None, 7)                 7175
```

- Création du modèle VGG-16 implémenté par **Keras**
- Pré-entraîné sur **ImageNet.**

Algorithme	Acc. (taille=224)
VGG-16	81 %

Partie III : Multi-inputs modélisation

Classification multi-input



Algorithme	Acc. (taille=224)
Multi modèle	82 %

Conclusion

Modélisations textuelles

TF-IDF
LDA

Word
Embedding

RNN

Modélisations visuelles

Machine Learning : SIFT & ORB

Deep Learning : CNN, Random Forest, KNN

Transfer Learning : VGG-16

- L'utilisation de réseaux neuronaux fournit les meilleurs résultats à la fois pour les données textuelles et visuelles
- Il faut densifier la base de données afin d'améliorer les performances.
- Il existe d'autres modèles de réseaux neuronaux et d'autres algorithmes: Word2Vec, ResNet50, Universal Sentence Encoder de Google.

Merci de votre attention