PARCOURS: DATA SCIENTIST

PROJET 6: CLASSIFICATION AUTOMATIQUE DES BIENS DE CONSOMMATION

Etudiant : Fatma Aidi
Mentor : Calliane You
Evaluateur :Bertrand Beaufils
Date :18/05/2021

TABLE

OBJECTIVES

Flipkart

MARKET PLACE EN INDE

Sur la place de marché, des vendeurs proposent des articles à des acheteurs en postant une photo et une description. L'attribution de la catégorie d'un article est effectuée **manuellement** par les vendeurs et est donc peu fiable

Mission

AUTOMATISER L'ATTRIBUTION DES CATÉGORIES.

Etude de la faisabilité d'un moteur de classification des articles en différentes catégories, avec un niveau de précision suffisant.

01
DONNÉES

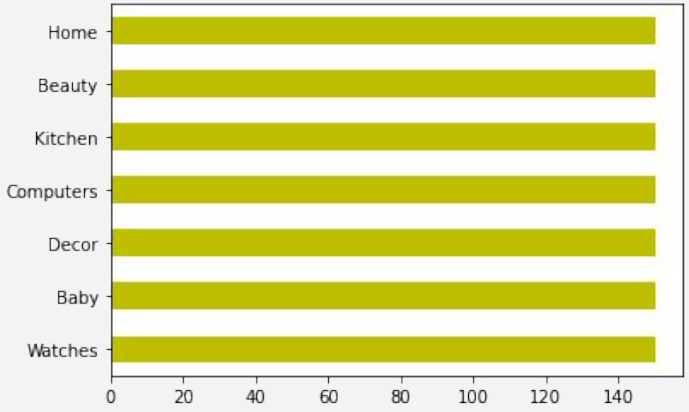
JEUX DE DONNÉES

DESCRIPTION

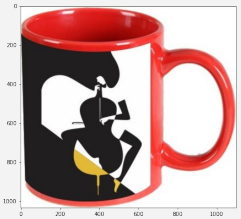
Les variables de la data sont des informations sur les **Produit**: image, prix,nom, description..

Taille:1050 produits et 15 colonnes
Variables qualitatives:11
Variables quantitatives:4
Doublon :0

CATÉGORIES



EXEMPLES

	Index1: 415	Index2: 177	Index3: 628
Image			
Product_name	Ajmal Titanium and Expedition Combo Set	Printland PMR1834 Ceramic Mug	Intex Kids Inflatable Air Chair
Catégorie	Beauty	Kitchen	Baby
Description	Flipkart.com: Buy Ajmal Titanium and Expedition Combo Set online only for Rs. 400 from Flipkart.com. Only Genuine Products. 30 Day Replacement Guarantee. Free Shipping. Cash On Delivery!	Printland PMR1834 Ceramic Mug (350 ml)\r\n Price: Rs. 299\r\n\t\t\t\t\r\n\t\t\tPrintland coffee mug is an adorable and a fantastic coffee mug. One can enjoy their morning coffee/tea in this huge mug. It is made of ceramic material. It is a perfect add on to your kitchen wardrobe. It looks very stylish & elegant to serve tea/coffee in this mug during a casual get together at home. It is also a perfect gift to be presented to your loved one.\r\nPrintland coffee mug is an adorable and a fantastic coffee mug. One can enjoy their morning coffee/tea in this huge mug. It is made of ceramic material. It is a perfect add on to your kitchen wardrobe. It looks very stylish & elegant to serve tea/coffee in this mug during a casual get together at home. It is also a perfect gift to be presented to your loved one.	Buy Intex Kids Inflatable Air Chair for Rs.429 online. Intex Kids Inflatable Air Chair at best prices with FREE shipping & cash on delivery. Only Genuine Products. 30 Day Replacement Guarantee.

DÉMARCHE

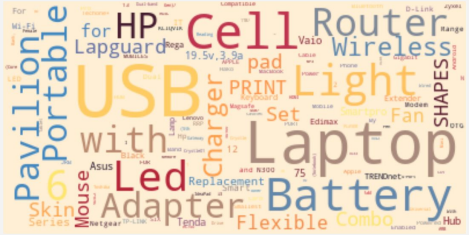
DONNÉES	PRÉ-TRAITEMENT	MODÉLISATION	RÉSULTATS
-Textes -Images	Traitement Features extraction Réduction de dimension	Clustering	Evaluation

02
SEGMENTATION TEXTE

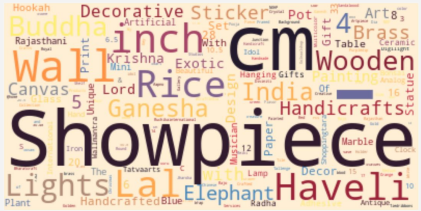
NUAGE DES MOTS PAR CATÉGORIE

Baby

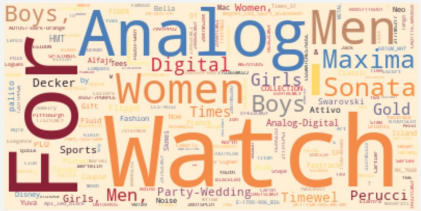
Computers



Decor



Watches



Home

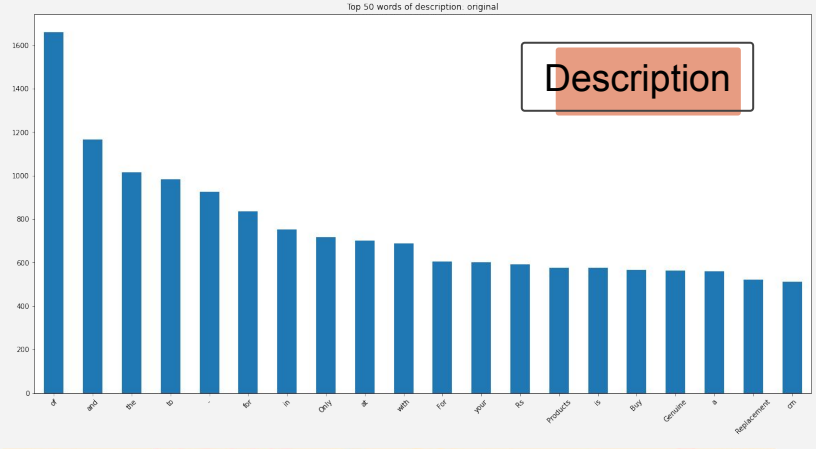
Kitchen

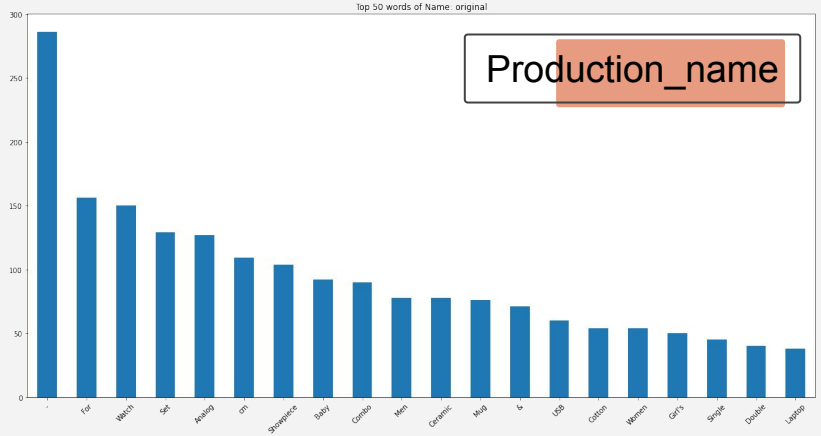
Beauty

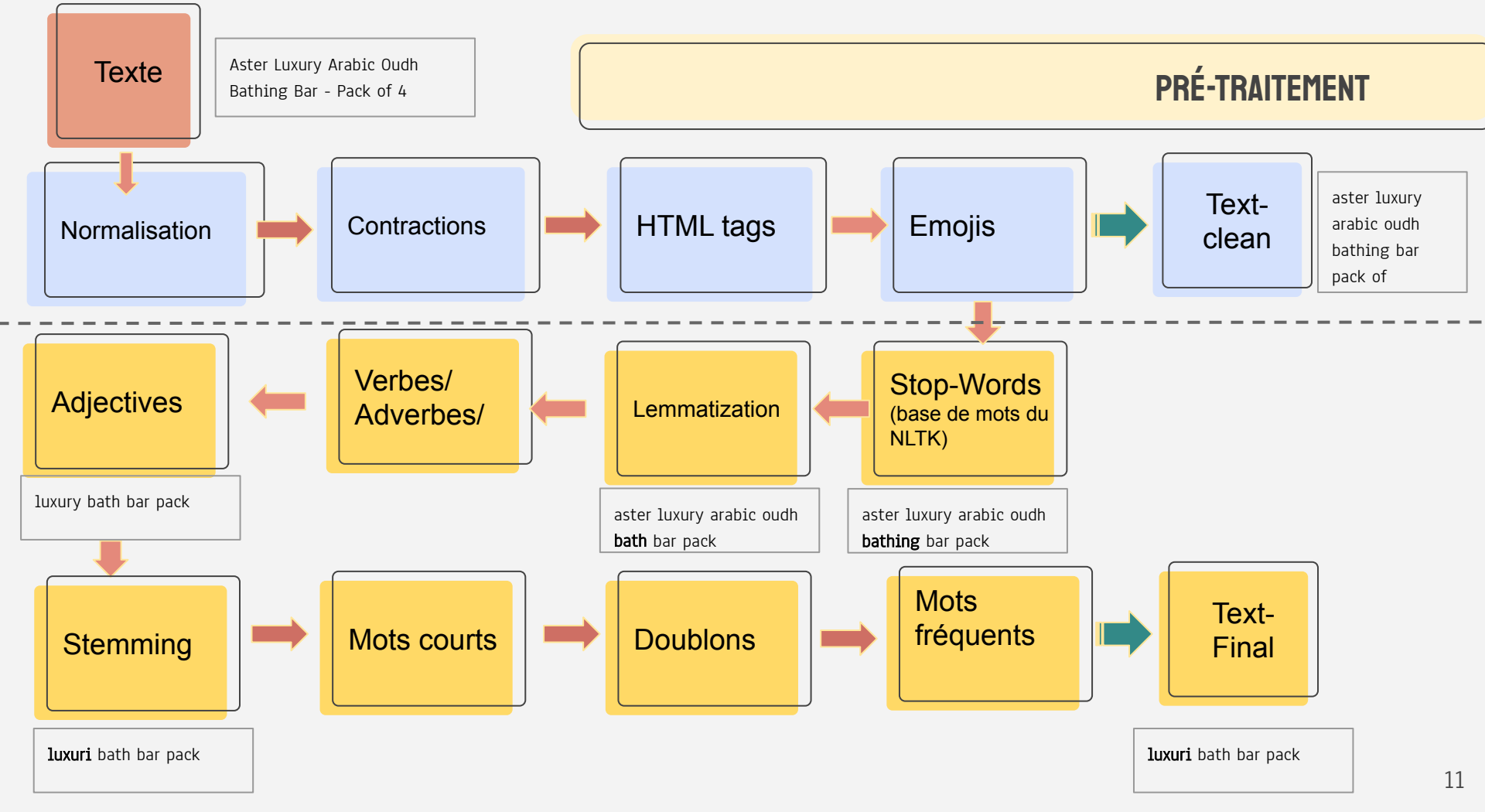
Mots clès: Baby, USB, Showpiece, Watch, Cotton, Ceramic, Combo.
Mots à supprimer: Set, cotton, boy,Men..

Parmis les 50 premier mots on trouve des:Ponctuation, nombre, déterminants..

VARIABLES





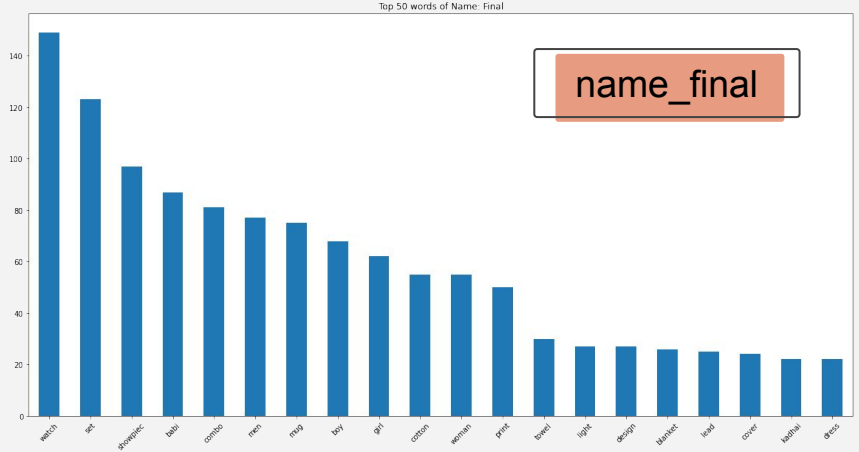

Texte
Aster Luxury Arabic Oudh Bathing Bar - Pack of 4
PRÉ-TRAITEMENT
Normalisation
Contractions
HTML tags
Emojis
Text-clean
aster luxury arabic oudh bathing bar pack of
Stop-Words (base de mots du NLTK)
aster luxury arabic oudh bathing bar pack
Lemmatization
aster luxury arabic oudh bath bar pack
Verbes/ Adverbes/
Adjectives
luxury bath bar pack
Stemming
luxuri bath bar pack
Mots courts
Doublons
Mots fréquents
Text-Final
luxuri bath bar pack

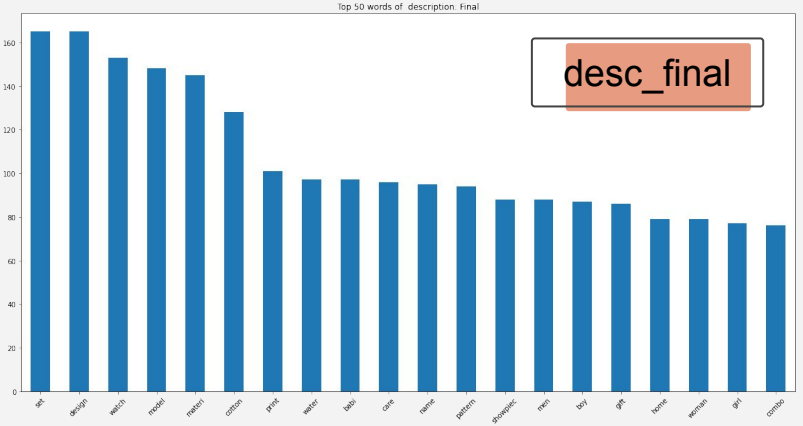
EXEMPLE

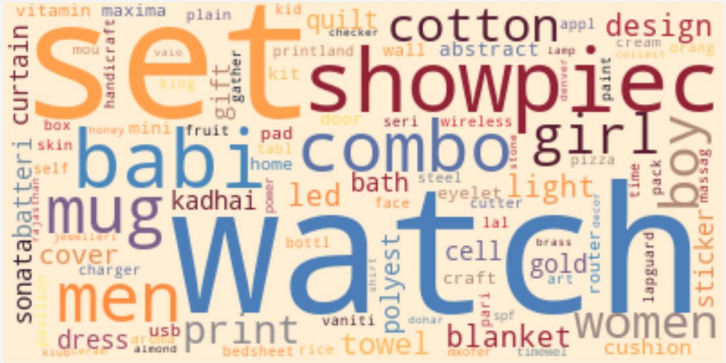
	original	clean	final
Name	Upside Down Sleeveless Applique Baby Girl's, Baby Boy's Jacket	upside down sleeveless applique baby girls baby boys jacket	sleeveless appliqu babi girl boy jacket
Description	Specifications of Upside Down Sleeveless Applique Baby Girl's, Baby Boy's Jacket General Details Ideal For Baby Girl's, Baby Boy's Pattern Applique Jacket Details Fabric Poly Cotton Reversible No Hooded No Closure Buttons Sleeve Sleeveless Lining Cotton Fabric Care Dont Wash With Other Garments, Hand Wash With Mild Detergent,Dont Tumble Dry, Dry In Shade. Additional Details Style Code LIGHT PINK SMILEY FACE JACKET Other Details Sleeveless In the Box 1 Jacket	specifications of upside down sleeveless applique baby girls baby boys jacket general details ideal for baby girls baby boys pattern applique jacket details fabric poly cotton reversible no hooded no closure buttons sleeve sleeveless lining cotton fabric care do not wash with other garments hand wash with mild detergentdo not tumble dry dry in shade additional details style code light pink smiley face jacket other details sleeveless in the box jacket	appliqu babi girl boy jacket poli cotton hood closur button sleev line care wash garment hand detergentdo shade code face

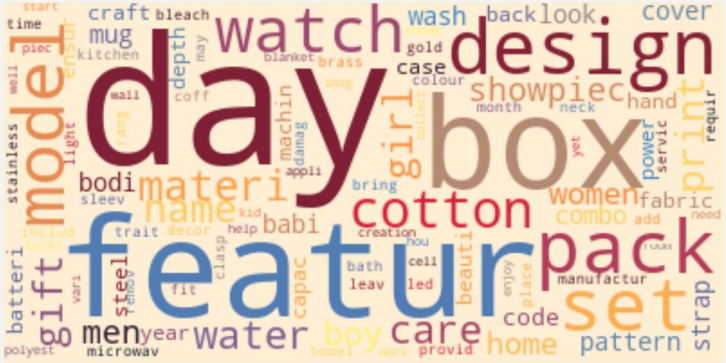
On distingue mieux les mots clés comme: watch,babi,combo,mug avec les noms que les descriptions

VARIABLES APRÈS PRÉ-TRAITEMENT









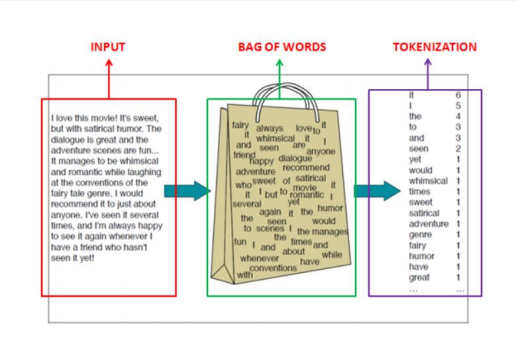
FEATURE_EXTRACTION

Feature extraction

Bag of Words

Tf-IdF: Term Frequency-Inverse Document Frequency

Embedding Word2vec



$$TFIDF_{t,d,D} = TF_{t,d} \times IDF_{t,D}$$

Importance d'un terme t dans un document d

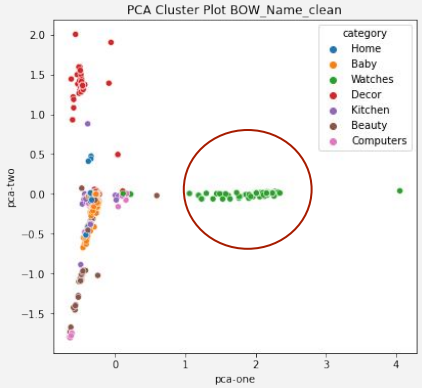
Fréquence d'un terme t dans un document d

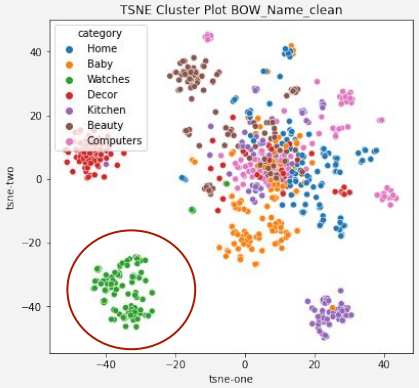
Importance du terme t dans l'ensemble des documents D

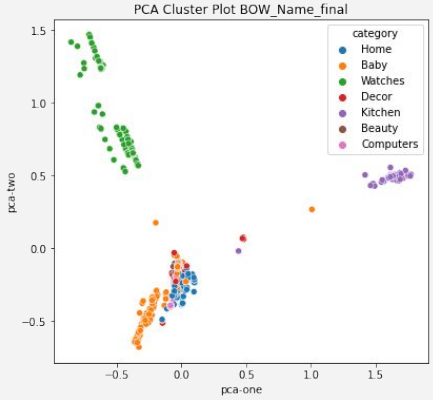
La séparations est claire pour les catg watches et décor.

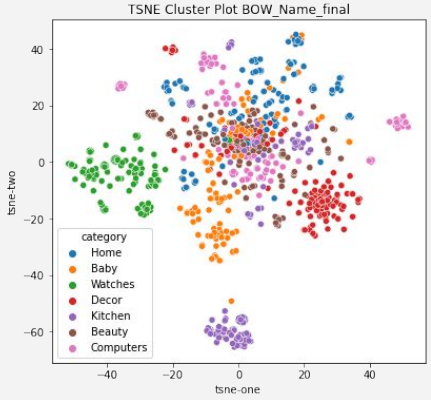
RÉDUCTION DES DIMENSIONS

BOW

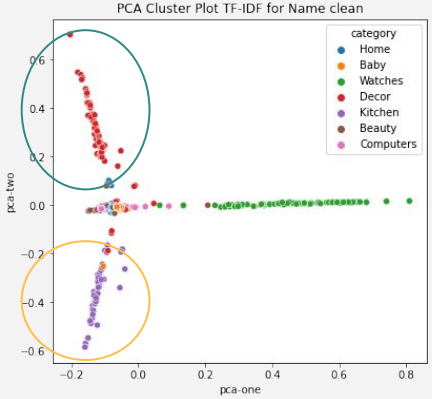


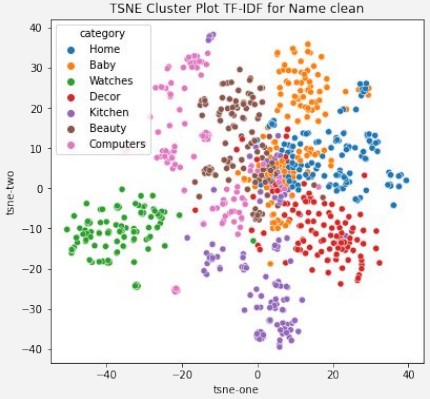


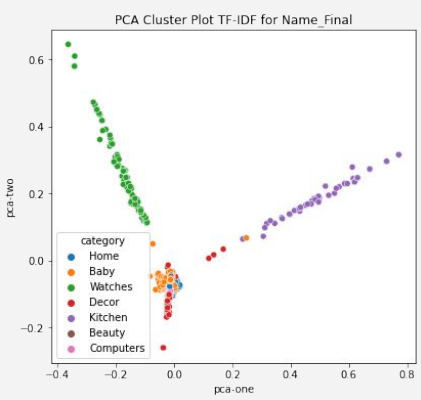


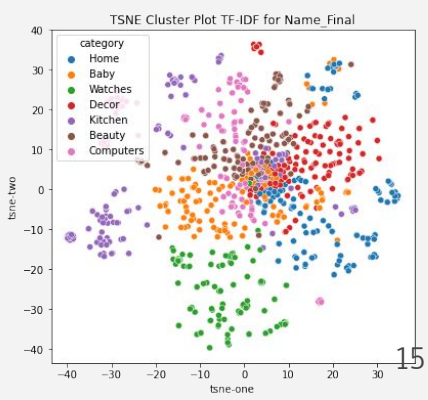


TF-IDF

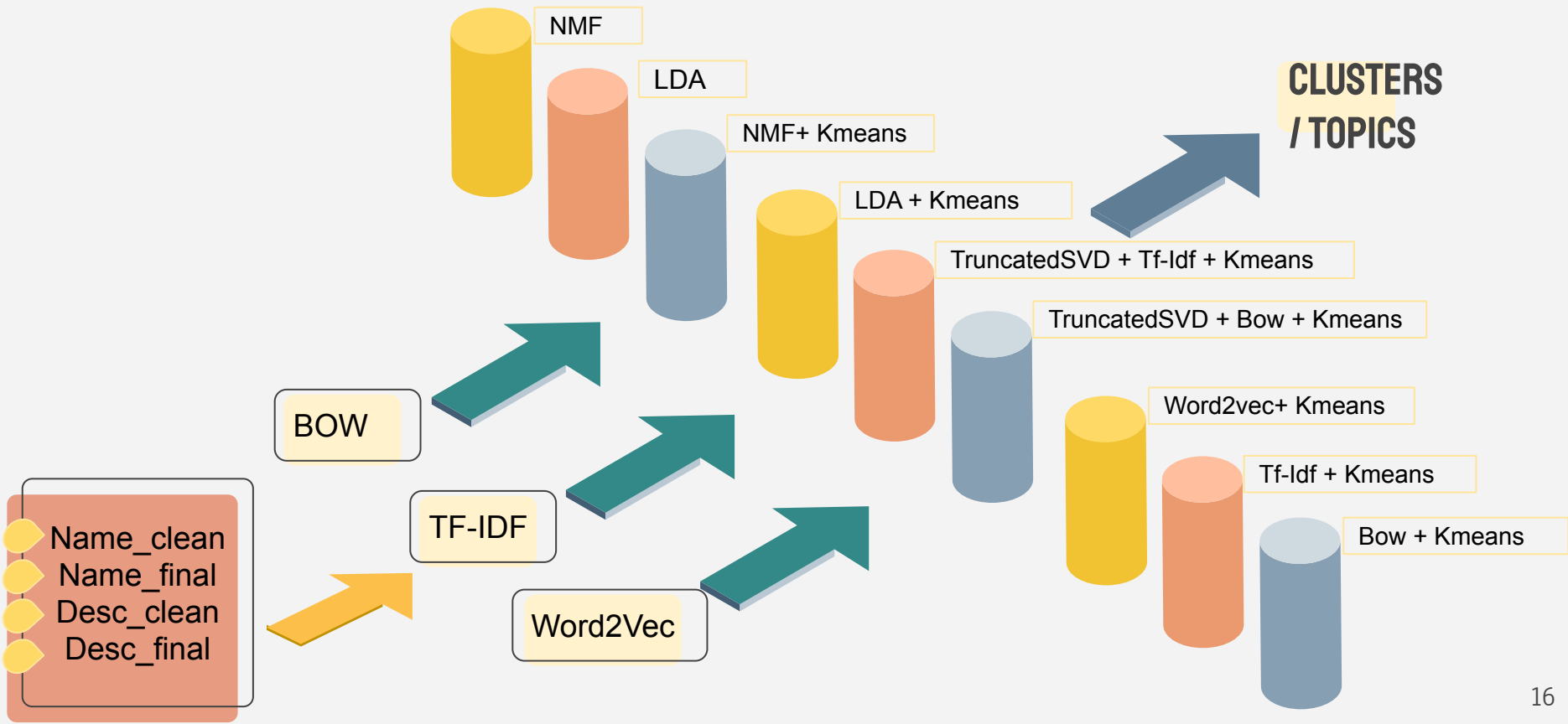




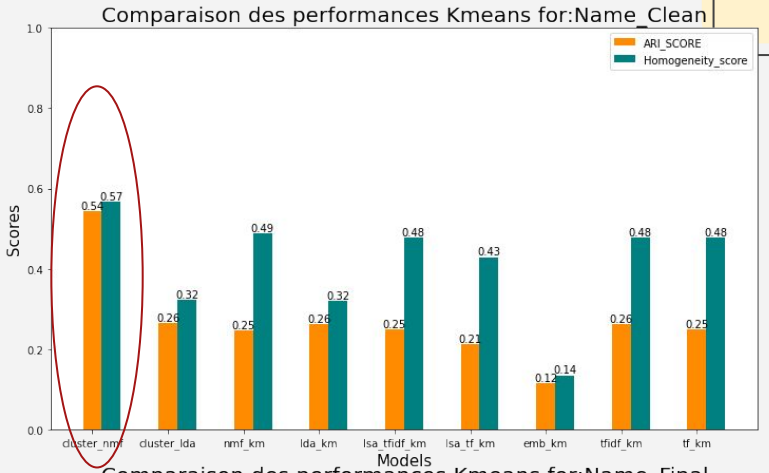


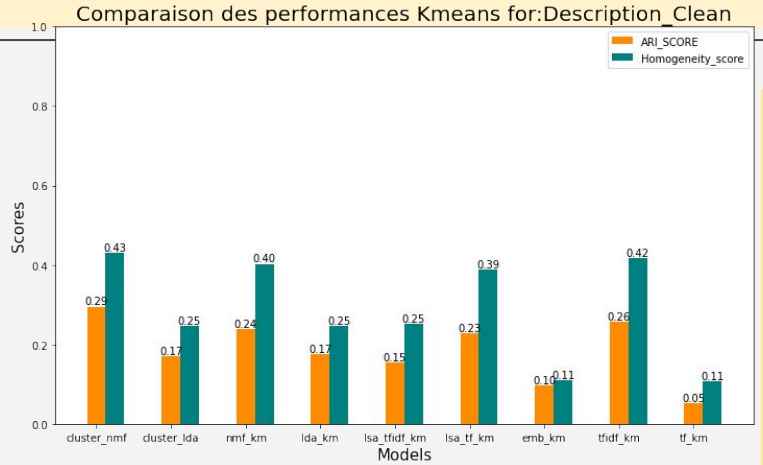


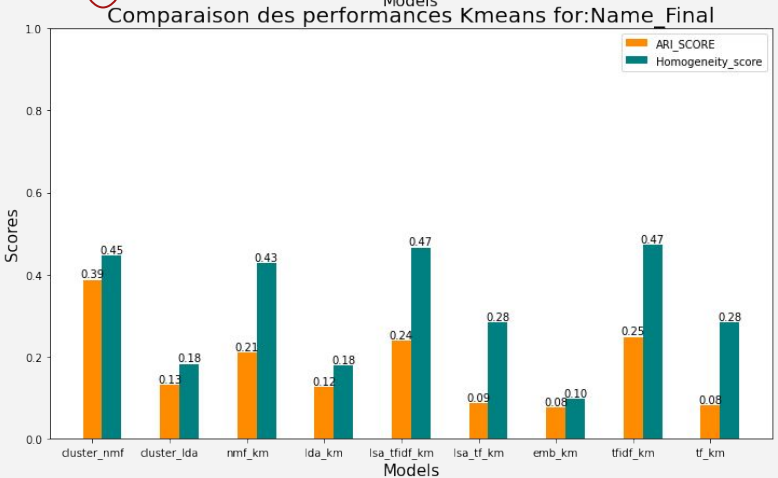
MODÉLISATION NON SUPERVISÉE: TOPIC MODELING

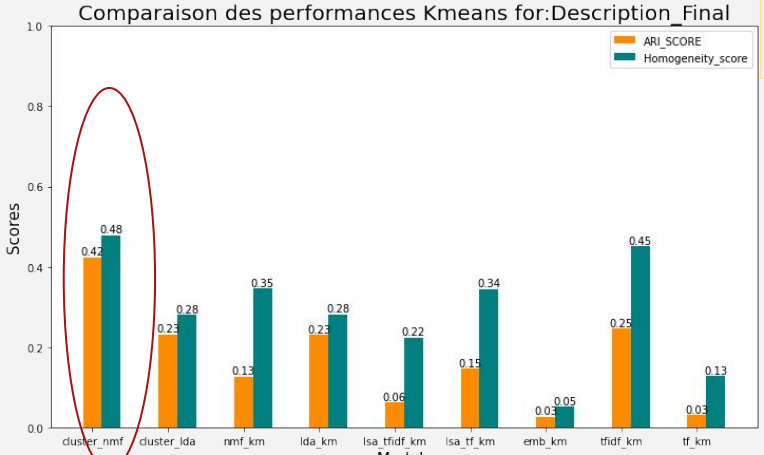


CLUSTERING - EVALUATION







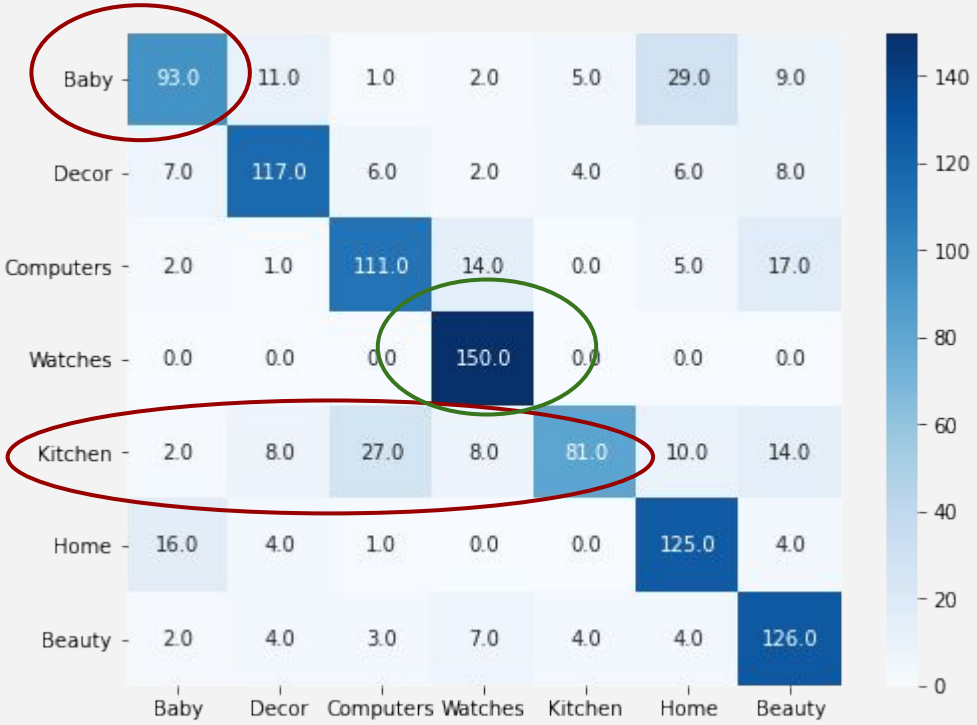


-Meilleur score pour la variable nom_clean(juste ponctuation et caractère spéciaux)

-Amélioration du modèle pour les descriptions avec toutes les étapes du prétraitement

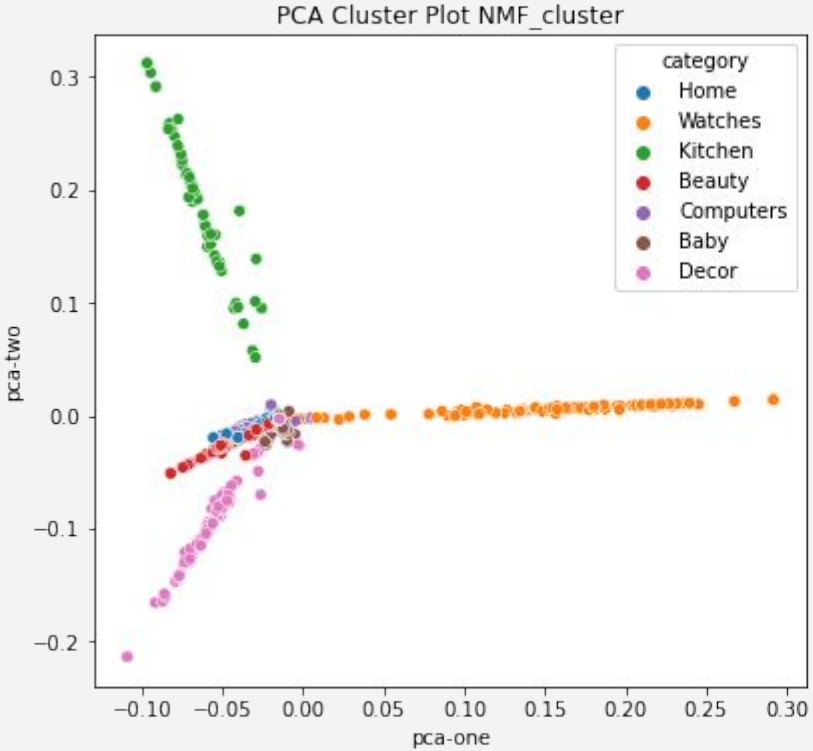
RESULTATS

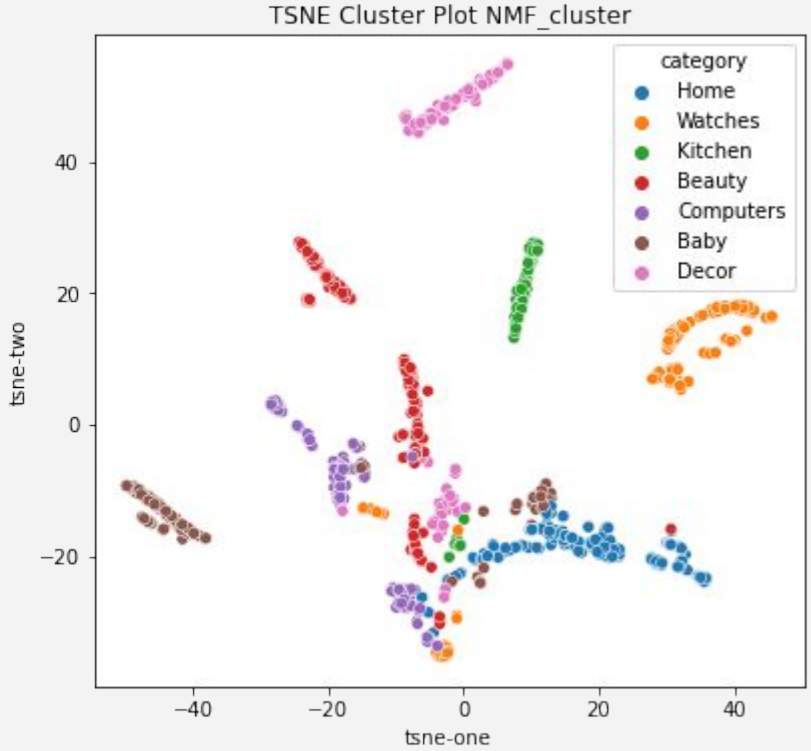
Matrice de confusion



Très Bien classé:
Watches,
Moins bien classés: Baby, Kitchen

CLUSTERS


PCA Cluster Plot NMF_cluster
category
Home
Watches
Kitchen
Beauty
Computers
Baby
Decor
pca-two
pca-one
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30

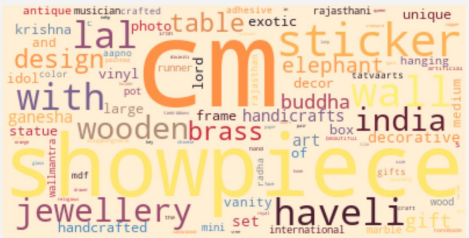

TSNE Cluster Plot NMF_cluster
category
Home
Watches
Kitchen
Beauty
Computers
Baby
Decor
tsne-two
tsne-one
40
20
0
−20
−40
−20
0
20
40

NUAGE DES MOTS PAR CLUSTER

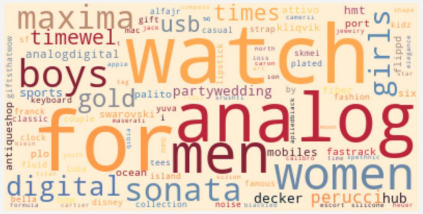
Baby

Computers

Decor

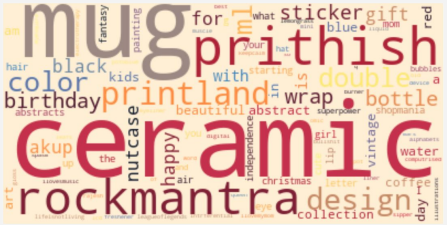


Watches



Home

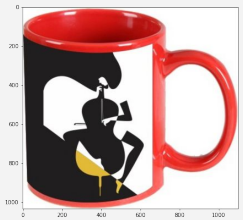
Kitchen



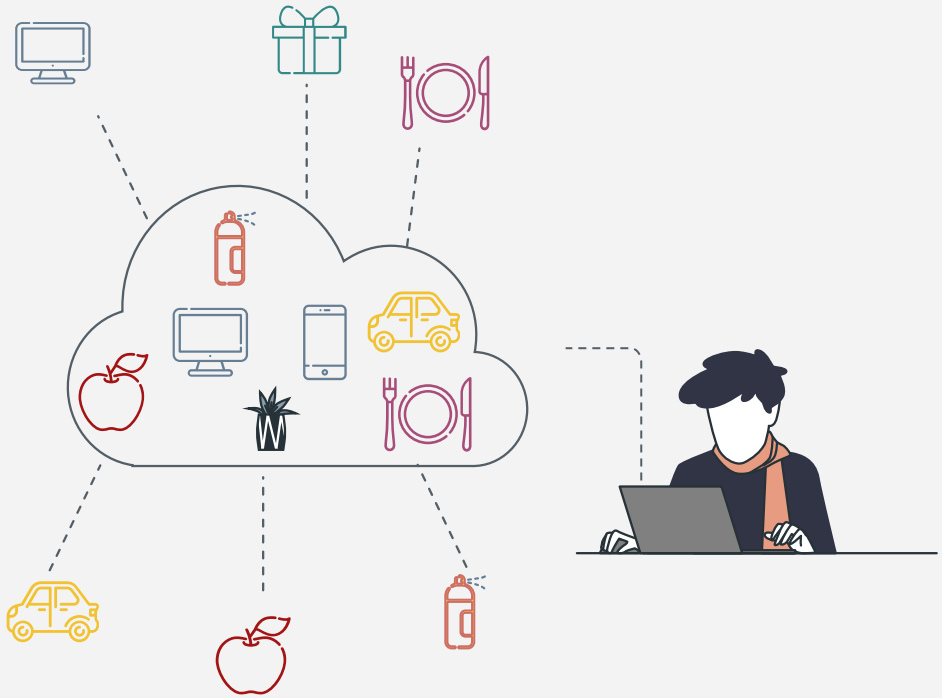
Beauty

On retrouve bien les mots clés pour chaque catégorie sur les clusters

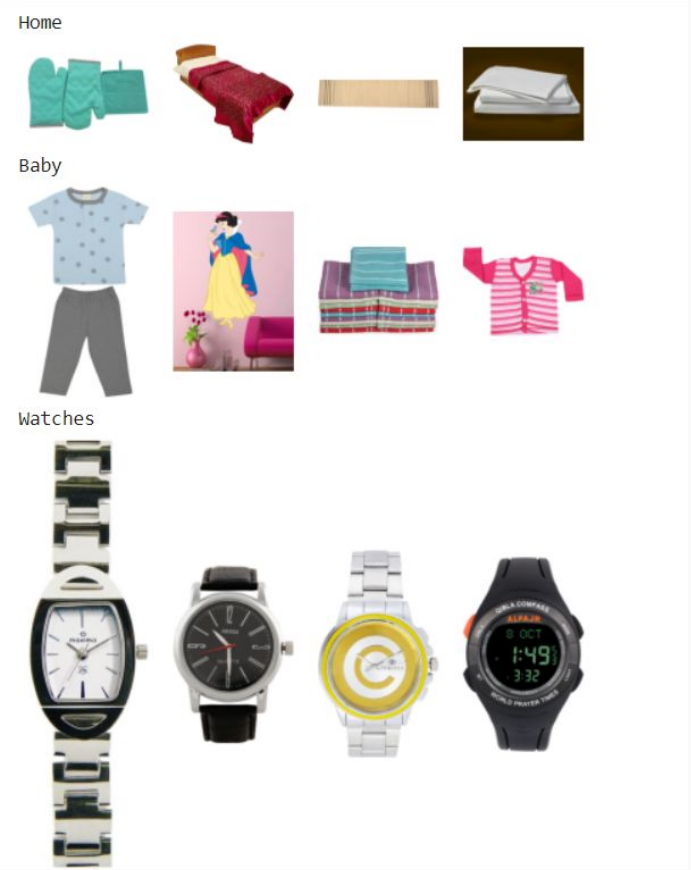
EVALUATION

	Index1: 415	Index2: 177	Index3: 889
Image			
Product_name	Ajmal Titanium and Expedition Combo Set	Printland PMR1834 Ceramic Mug	Intex Kids Inflatable Air Chair
Catégorie	Beauty	Kitchen	Baby
Cluster_text	Beauty	Kitchen	Kitchen

03
SEGMENTATION IMAGE



IMAGE



-Dimension des images sont differents

Original Image

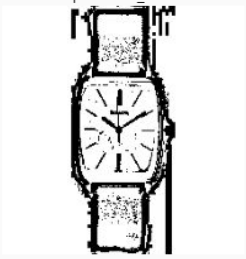
Stratégie 1

process1_PIL

- -Filtre Gaussian blur
- -auto-contrast
- -Égalisation d'histogramme
- -Lissage pour éliminer le bruit
- -Redimension des images

Stratégie 2

process2_CV2



- -Passage au gris
- -Redimension des images
- -Seuil adaptatif
- -Egalisation adaptative d'histogramme à contraste limité

PRÉTRAITEMENT DES IMAGE

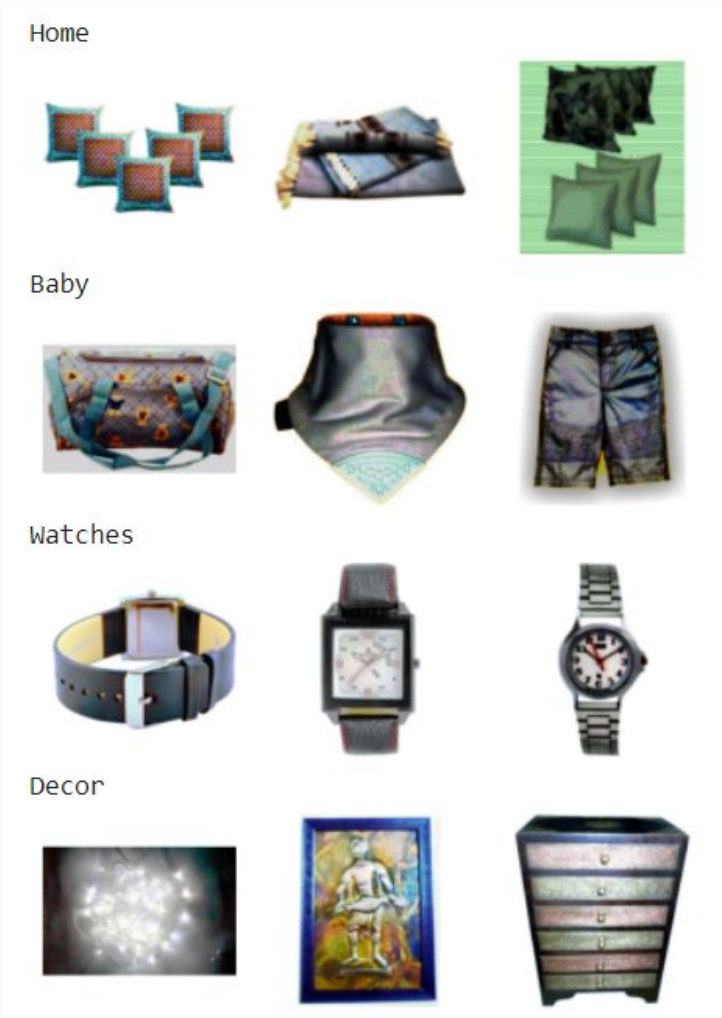
Stratégie 1 | Stratégie 2

FEATURES EXTRACTION

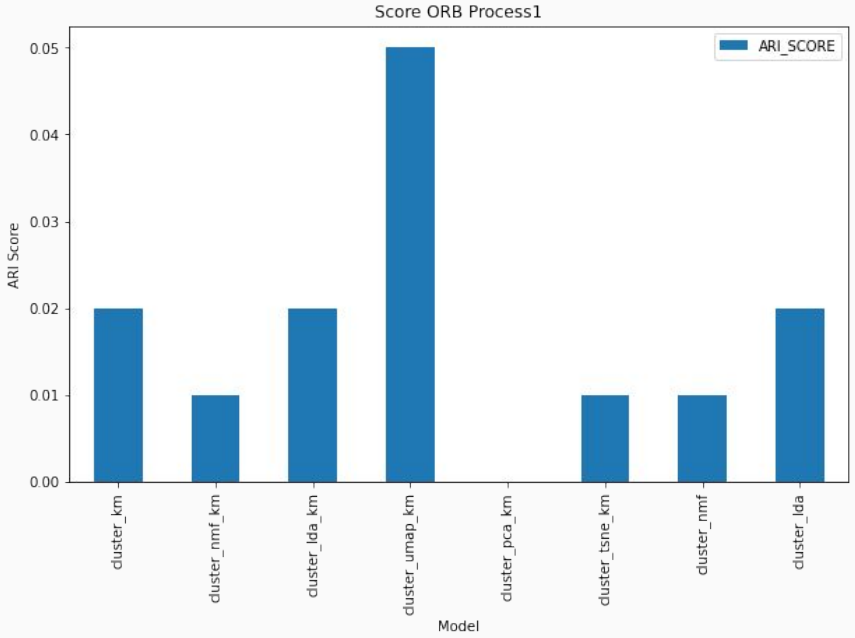
- Pre-trained model: VGG16, ResNet50V2, Xception..
- Bag Of Visual Words de OpenCV: ORB (Oriented FAST and Rotated BRIEF)

CLUSTERING: KMEANS

Evaluation ARI

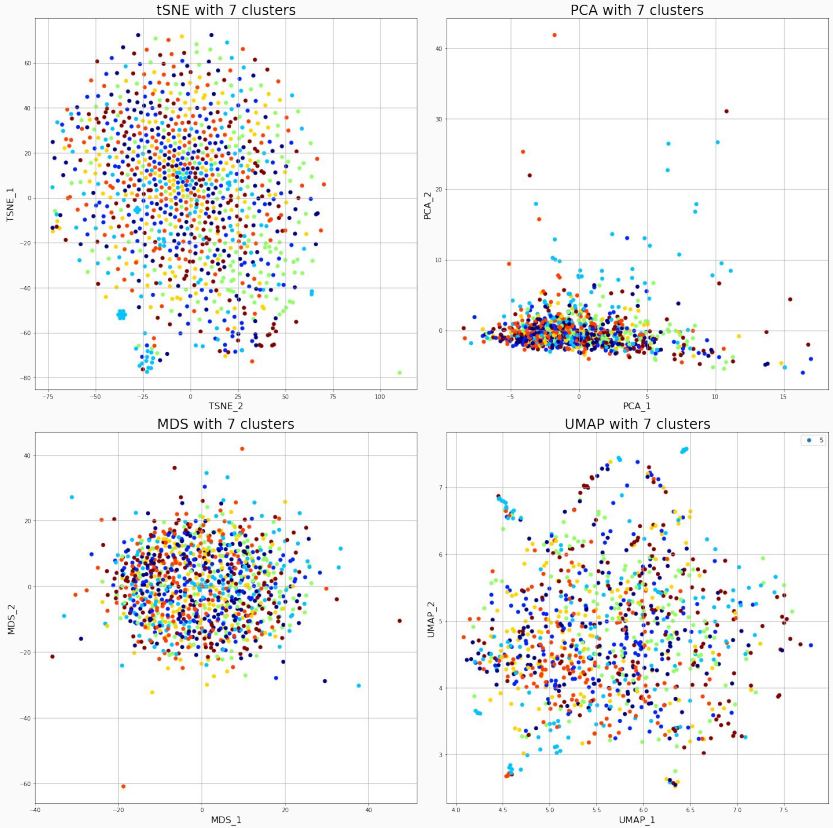


I ERE STRATÉGIE - ORB



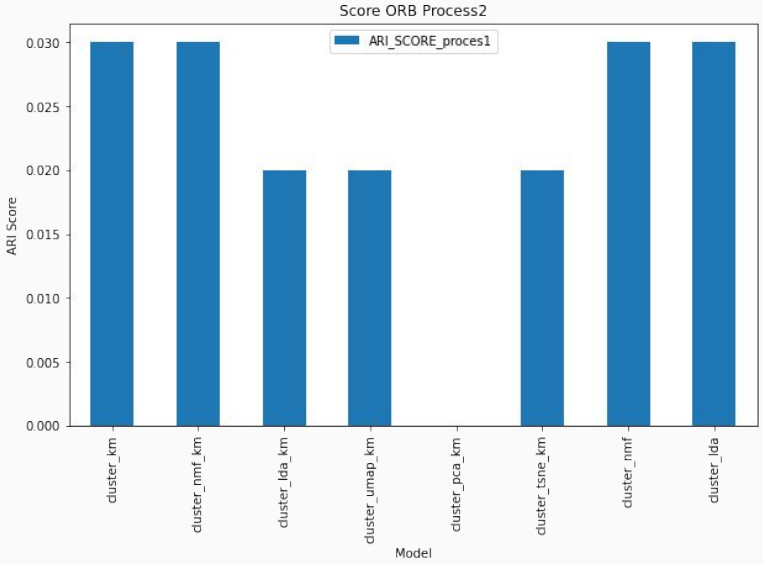
Meilleur score d'ARI: 5%

IERE STRATÉGIE

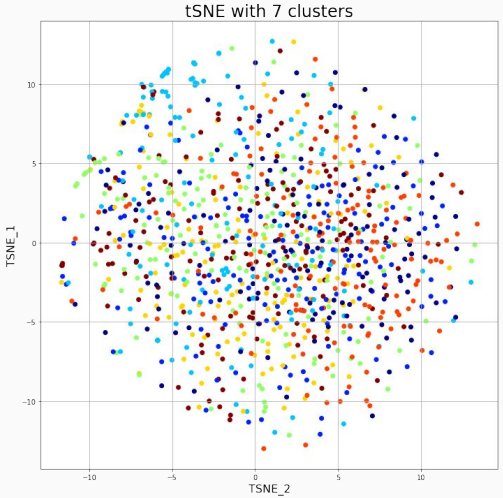


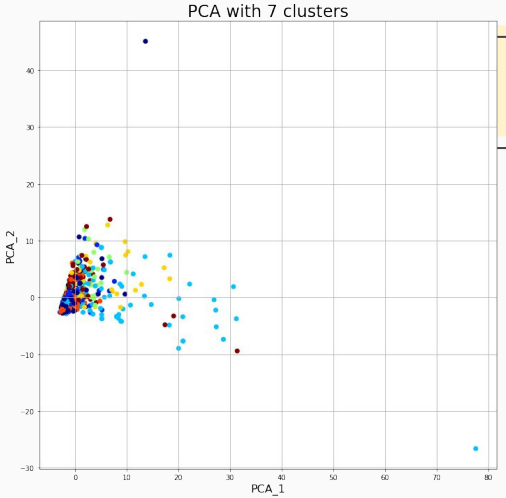
- Projection en réduction de dimension non satisfaisante

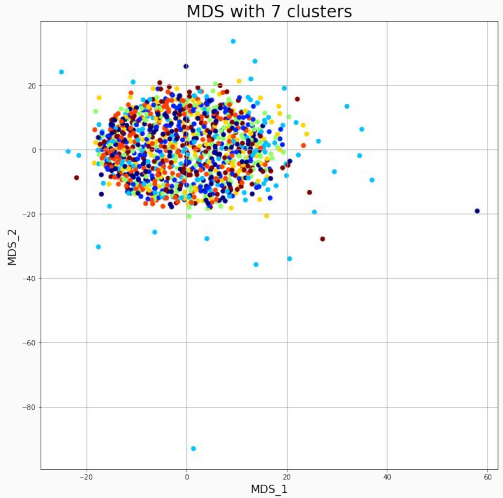
2 EME STRATÉGIE - ORB

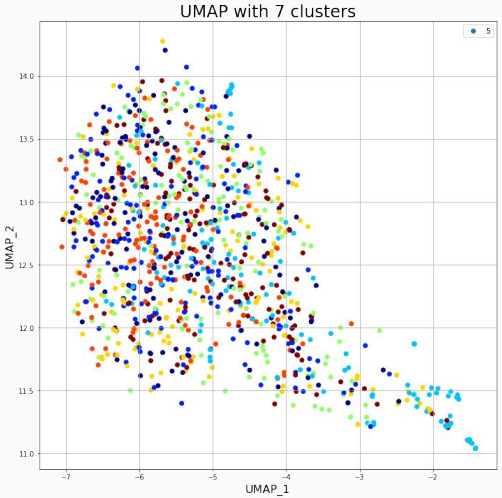


Meilleur score d’ARI: 3%









2 EME STRATÉGIE

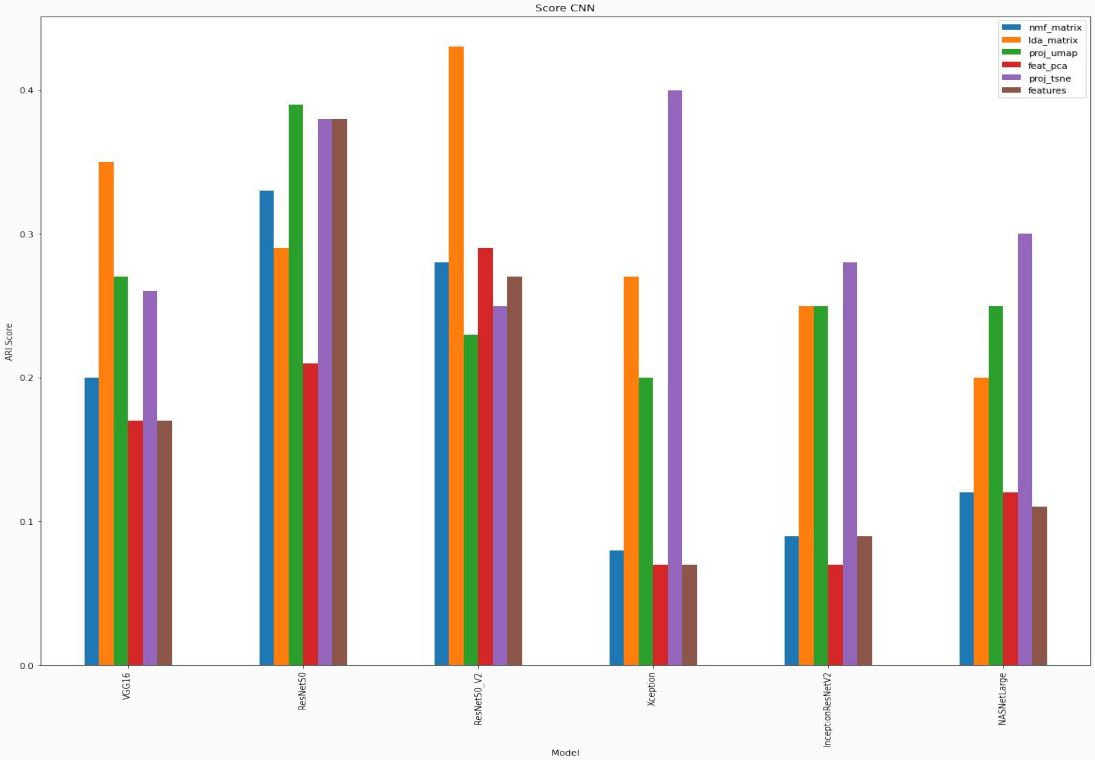
- Projection en réduction de dimension non satisfaisante

TRANSFER LEARNING- MODEL KERAS

```
model_vgg16 = VGG16(include_top=False,
                    weights="imagenet",
                    input_shape=(224,224,3))
model_restnet_v2 = ResNet50V2(include_top=False,
                          weights="imagenet",
                          input_shape=(224,224,3),
                          input_tensor=None,
                          classes=1000,
                          classifier_activation='softmax')
model_restnet = ResNet50(include_top=False,
                          weights="imagenet",
                          input_shape=(224,224,3))
model_xception=Xception(include_top=True,
                        weights="imagenet",
                        input_shape=(299,299,3),
                        classifier_activation="softmax")
model_nasnet=NASNetLarge(input_shape=(331,331,3),
                          weights="imagenet",)
model_inception=InceptionResNetV2(include_top=True,
                                  weights="imagenet",
                                  classifier_activation="softmax")
```

- Choix du 6 Modèles de deep learning pré-entraînés pour feature extraction de la bibliothèque de keras
- Inconvénient temps de calcul très important

TRANSFER LEARNING

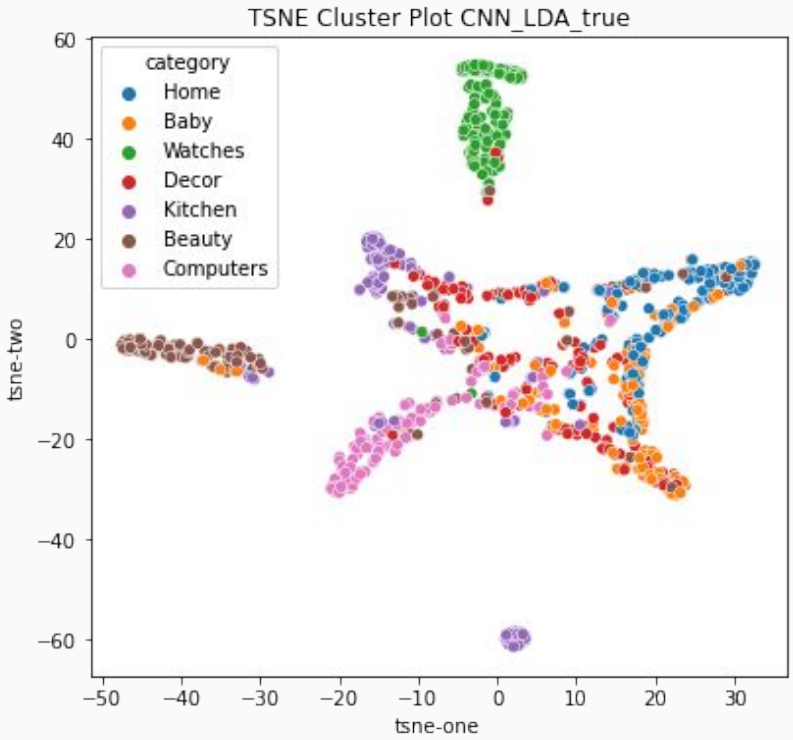


- Résultat du score ARI varie entre 5% jusqu'à 43%
- ResNet50V2 donne les meilleurs score de segmentation sur LDA

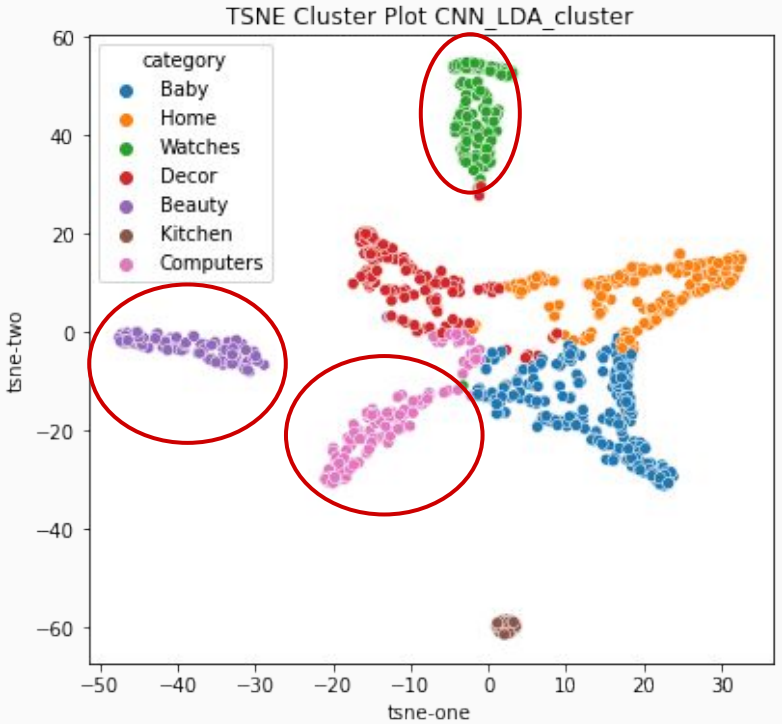
3 catégories on arrive à les séparer: watches, Beauty et computers

TRANSFER LEARNING

Catégories

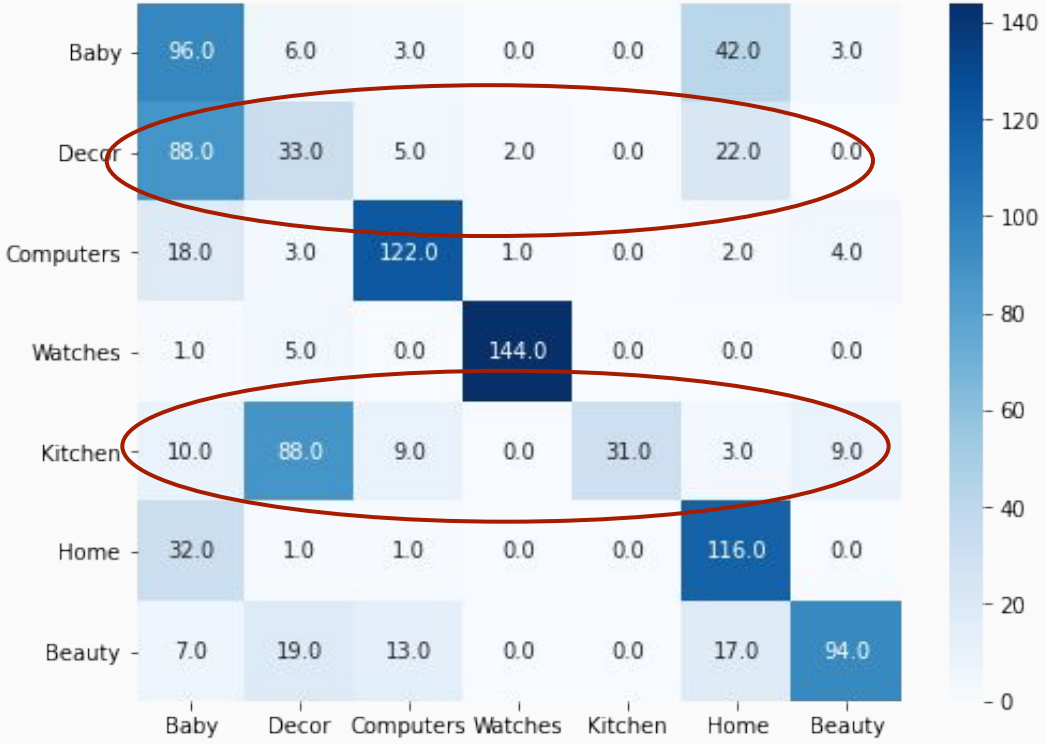


Clusters



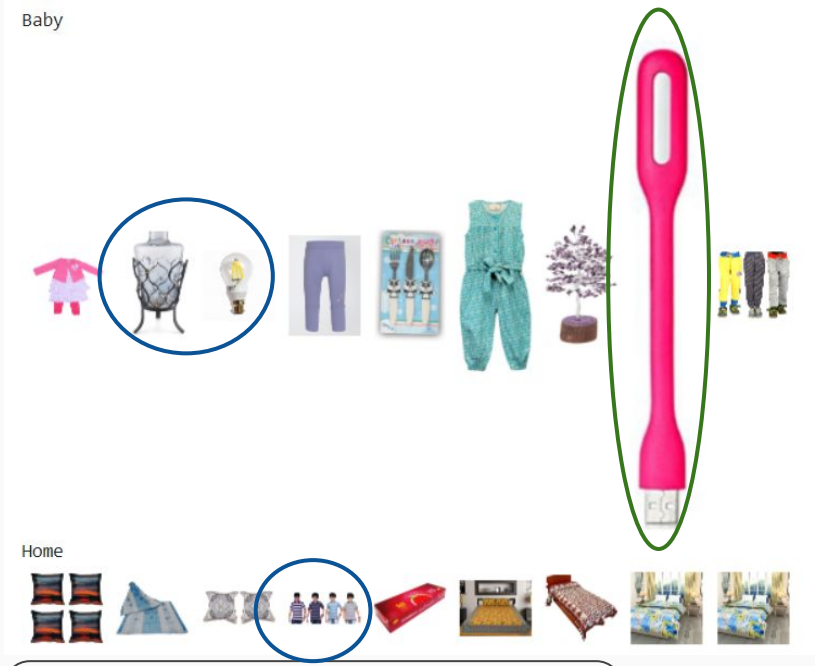
TRANSFER LEARNING

Matrice de confusion



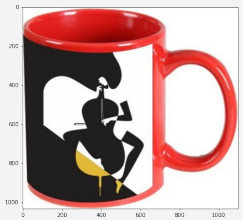
Décor et Kitchen sont les catégories les moins bien classée

TRANSFER LEARNING: EVALUATION



Les mugs sont très mal classés , selon l'orientation des anses aussi les bouteilles .

EVALUATION

	Index1: 415	Index2: 177	Index3: 889
Image			
Product_name	Ajmal Titanium and Expedition Combo Set	Printland PMR1834 Ceramic Mug	Intex Kids Inflatable Air Chair
Catégorie	Beauty	Kitchen	Baby
Cluster_text	Beauty	Kitchen	Kitchen
Cluster_Image	Beauty	Decor	Baby

CONCLUSIONS

- Selon les résultats du clustering, la faisabilité du moteur de classification est possible et satisfaisante avec les données textuelles, moin avec les données visuelles.

- Pour les images si on donne aux vendeurs quelques recommandation pour la prise des images de leurs produits, on peut améliorer les segmentation.(comme pour les mugs avec les anses)

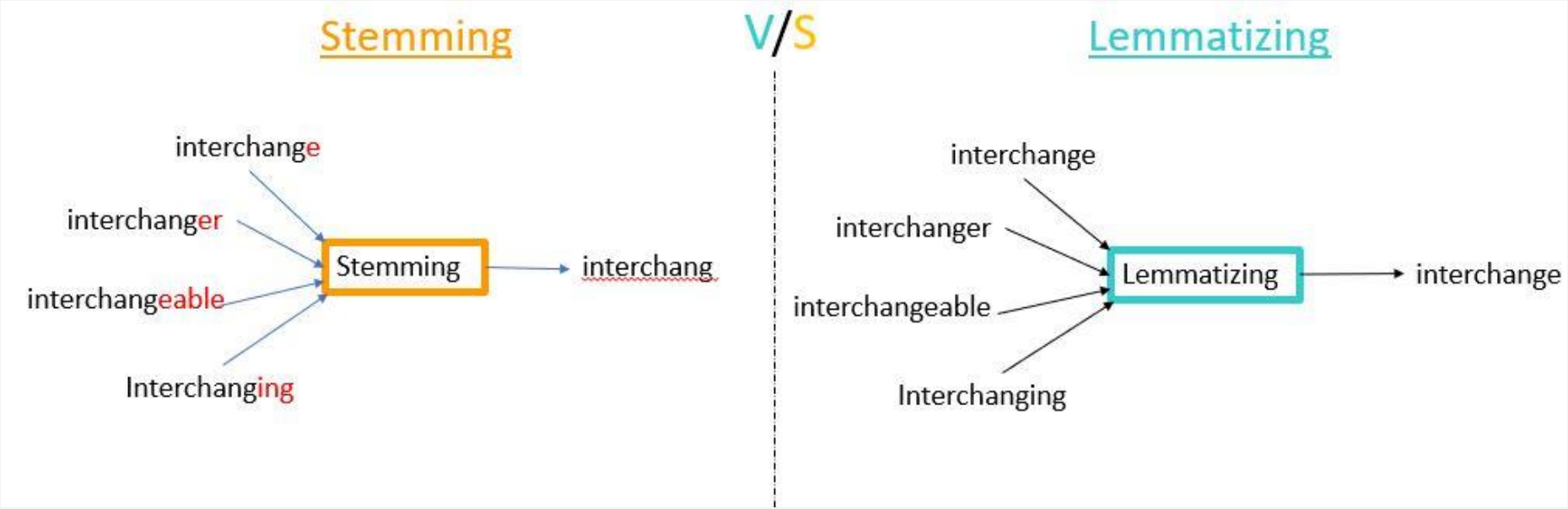
- Les montres et produits de beauté sont les mieux classés dans les deux cas (Texte ou image)

- Tester une segmentation avec les deux données ensembles.

RÉFERENCE

- https://www.programmersought.com/article/4304366575/
- https://autoveille.info/2018/02/19/les-mesures-de-statistiques-textuelles-tf-idf-rappel-precision-vues-par-des-experts-en-tal-interview-n1-damien-nouvel/
- https://www.liksi.tech/2018/07/16/nlp-et-extraction-dintents/
- https://docs.opencv.org/master/d7/d4d/tutorial_py_thresholding.html
- https://medium.com/@pierre_guillou/nlp-fastai-topic-modeling-af2687b5c276
- slidesgo

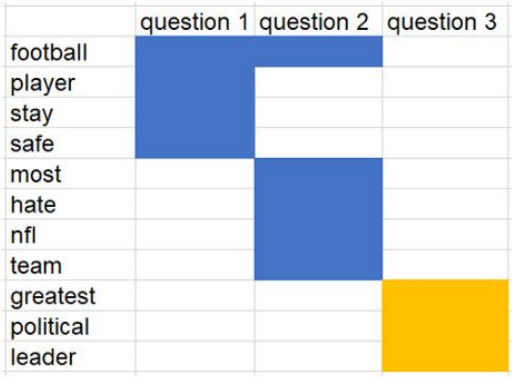
ANNEXE



https://www.pluralsight.com/guides/importance-of-text-pre-processing

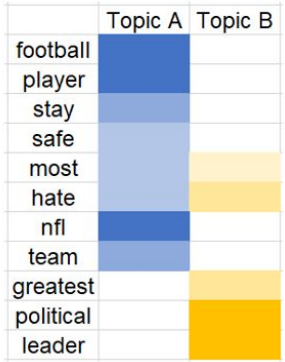
NMF :NON-NEGATIVE MATRIX FACTORIZATION

- How do football players stay safe?
- What is the most hated NFL football team?
- Who is the greatest political leader?



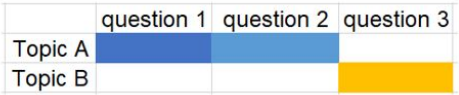
=

W
Weights for terms



X

H
Weights for questions



NMF est un algorithme **déterministe** qui aboutit à une représentation unique du corpus en termes de sujets latents.

En substance, étant donné une matrice de documents par mots (A), NMF nous donnera deux matrices: une matrice **W avec thèmes par mots**, et la matrice de coefficients **H avec documents par thèmes.**

https://towardsdatascience.com/topic-modeling-quora-questions-with-lda-nmf-aff8dce5e1dd

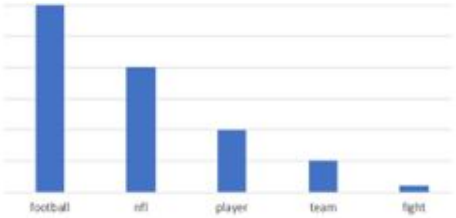
LDA: LATENT DIRICHLET ALLOCATION

Every **question** consists of a mix of **topics**

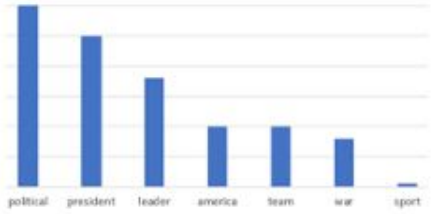
question	question	question	question
How Do Football Players Stay Safe?	What is the most hated NFL football team of all time	Who is the greatest political leader in the world and why?	Why do people treat politics like it's a football team or some kind of sport?
100% Topic A	90% Topic A 10% Topic C	100% Topic B	40% Topic A 60% Topic B

Every **topic** consists of a mix of **words**

Topic: Sport



Topic: Politics



LDA, est un modèle statistique génératif probabiliste: Une question est une distribution de probabilité de sujets, et chaque sujet est une distribution de probabilité de mots.

ANNEXE

. *The sky is blue.*
. *The sun is bright today.*
. *The sun in the sky is bright.*
. *We can see the shining sun, the bright sun.*

$f_{t,d}$

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0

TF

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}}$$

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

- TF-IDF: Multiply TF and IDF scores, use to rank importance of words within documents
  - Most important word for each document is highlighted

x

$$\mathrm{idf}(t,D) = \log_{10}\frac{N}{n_t}$$

$N = 4$

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

$\log_{10}\frac{4}{1} = 0.602$ $\log_{10}\frac{4}{3} = 0.125$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

	blue	bright	can	see	shining	sky	sun	today
1	**0.301**	0	0	0	0	0.151	0	0
2	0	0.0417	0	0	0	0	0.0417	**0.201**
3	0	0.0417	0	0	0	**0.100**	0.0417	0
4	0	0.0209	**0.100**	**0.100**	**0.100**	0	0.0417	0

Tf-Idf:
Term Frequency-Inverse Document Frequency

https://github.com/HeshamAsem/NLP/blob/main/PDF%20Files/4.6%20TF-IDF.pdf

14