Analyse

Place de marché

Classifiez automatiquement des biens de consommation

Sofiane Mouhab

13 juillet 2021

1 - Généralités

1.1 - Problématique

Pour l'instant, l'attribution de la catégorie d'un article est effectuée manuellement par les vendeurs et est donc peu fiable. De plus, le volume des articles est pour l'instant très petit.

Pour rendre l'expérience utilisateur des vendeurs (faciliter la mise en ligne de nouveaux articles) et des acheteurs (faciliter la recherche de produits) la plus fluide possible et dans l'optique d'un passage à l'échelle, **il devient nécessaire d'automatiser cette tâche.**

1.2 - Objectif

- Analyser le jeu de données en réalisant un prétraitement des images et des descriptions des produits
- Appliquer une réduction de dimension, puis un clustering
- Réaliser une représentation graphique
- Réaliser une première étude de faisabilité d'un moteur de classification

1.3 - Condition de mise en oeuvre

Pour pouvoir sereinement réaliser ses objectifs, il nous faut donc diverses informations qui pourrait se trouver dans notre base de données.

À nous donc, d'examiner celle-ci, de déterminer à quel point les informations sont viables, ou perfectible.

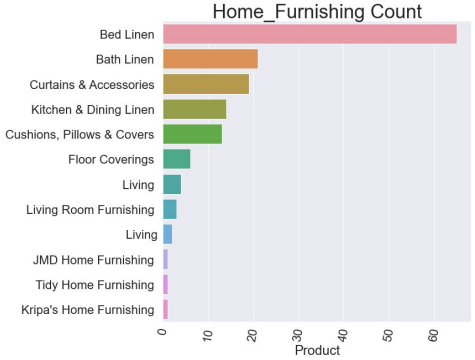
Il y a donc 3 grandes interrogations :

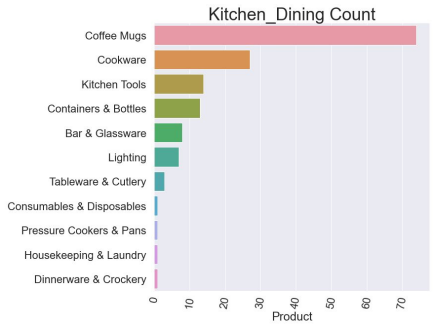
- A-t-on assez de données ?
- Peut-on faire une classification cohérente ?
- Peut-on obtenir une classification pertinente des produits de manière non-supervisée
- Déterminer les Niveaux de précision

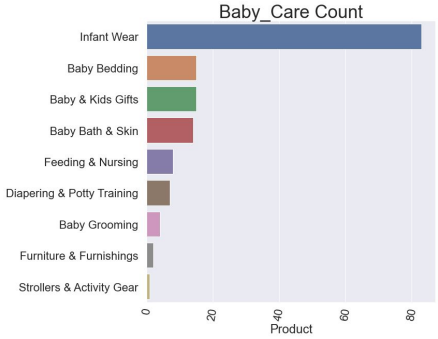
Passons de suite à ce travail, en commençant par rapidement prendre connaissance des données en présence...

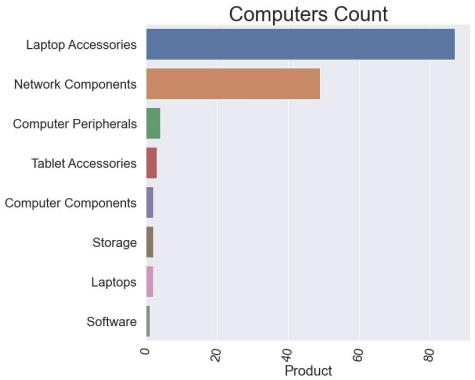
2 - Les données

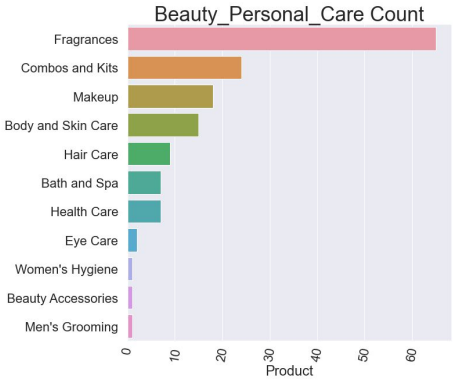
2.1 - Description

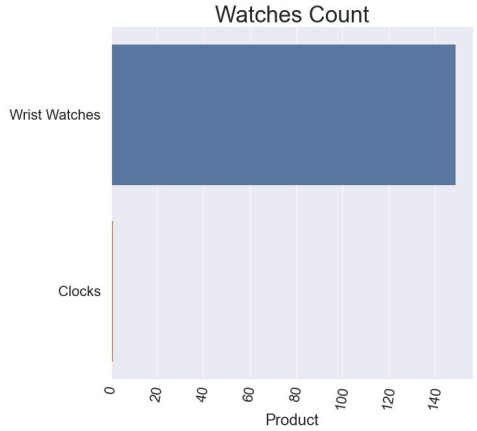








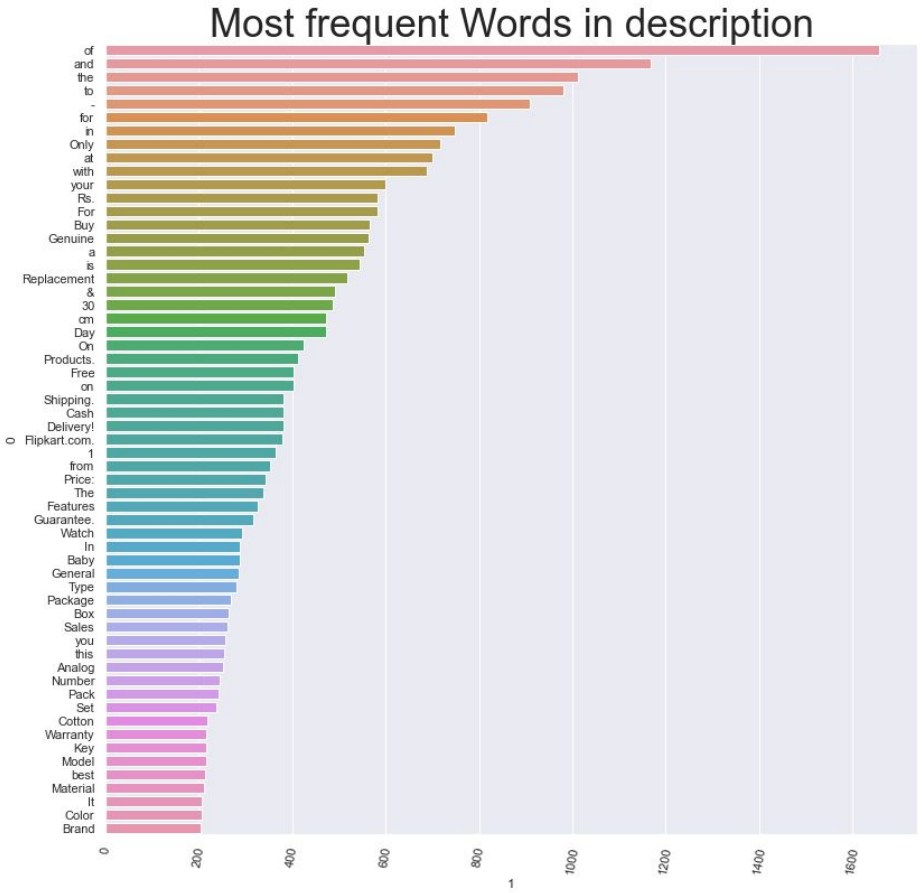


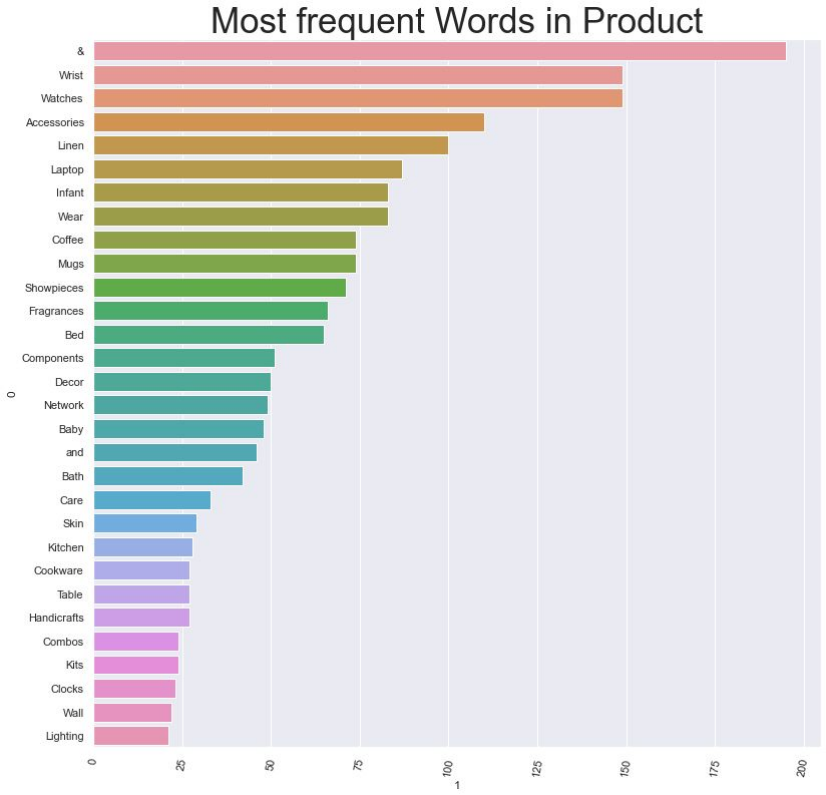


Chiffres clés :

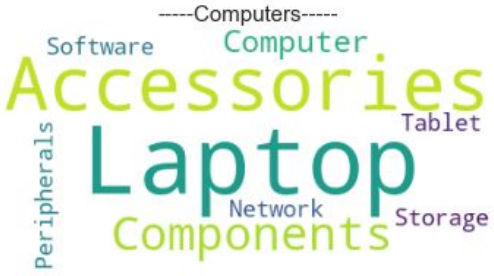
- 1050 produits
- 1050 photos
- 1050 descriptions
- 7 Catégories
- 63 Produits

2.2 - Description



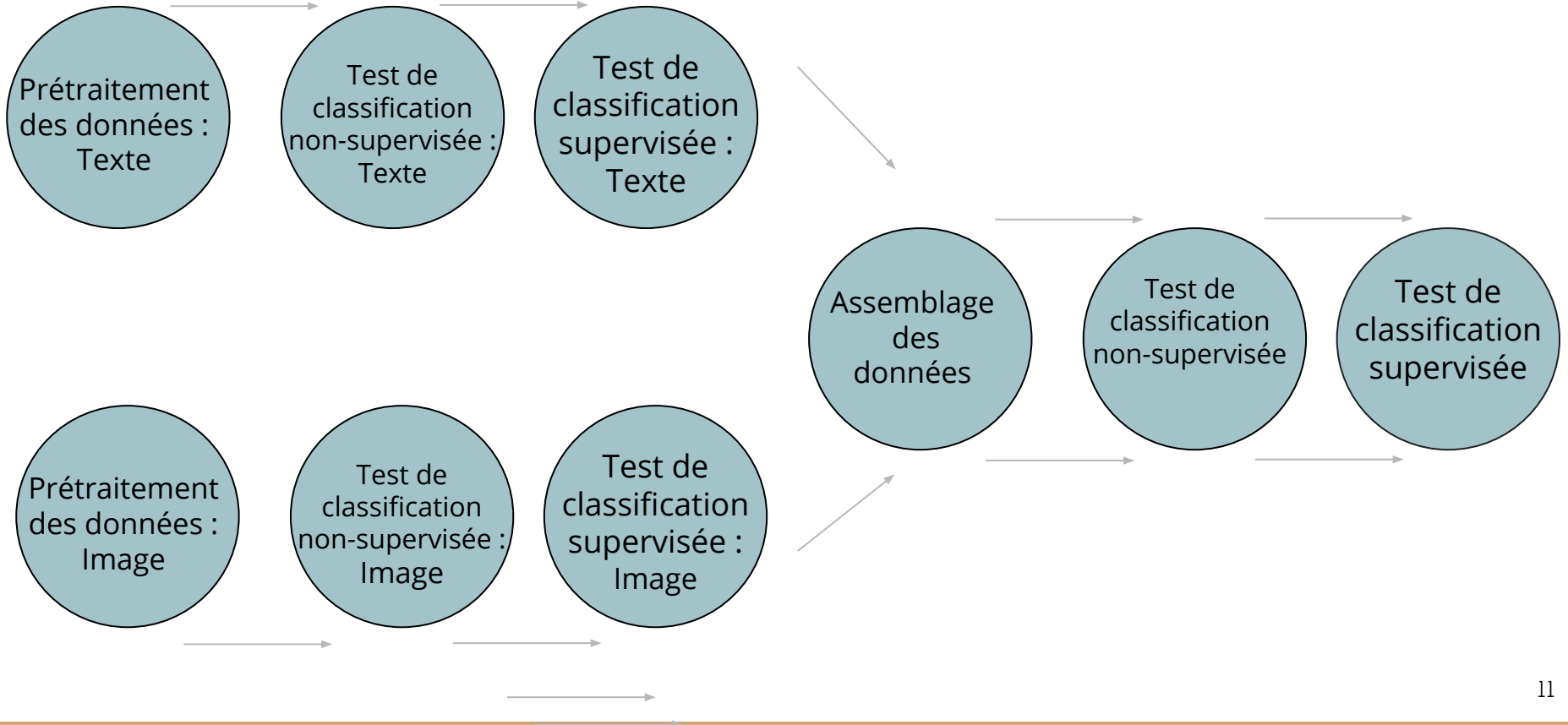


2.3 - Description



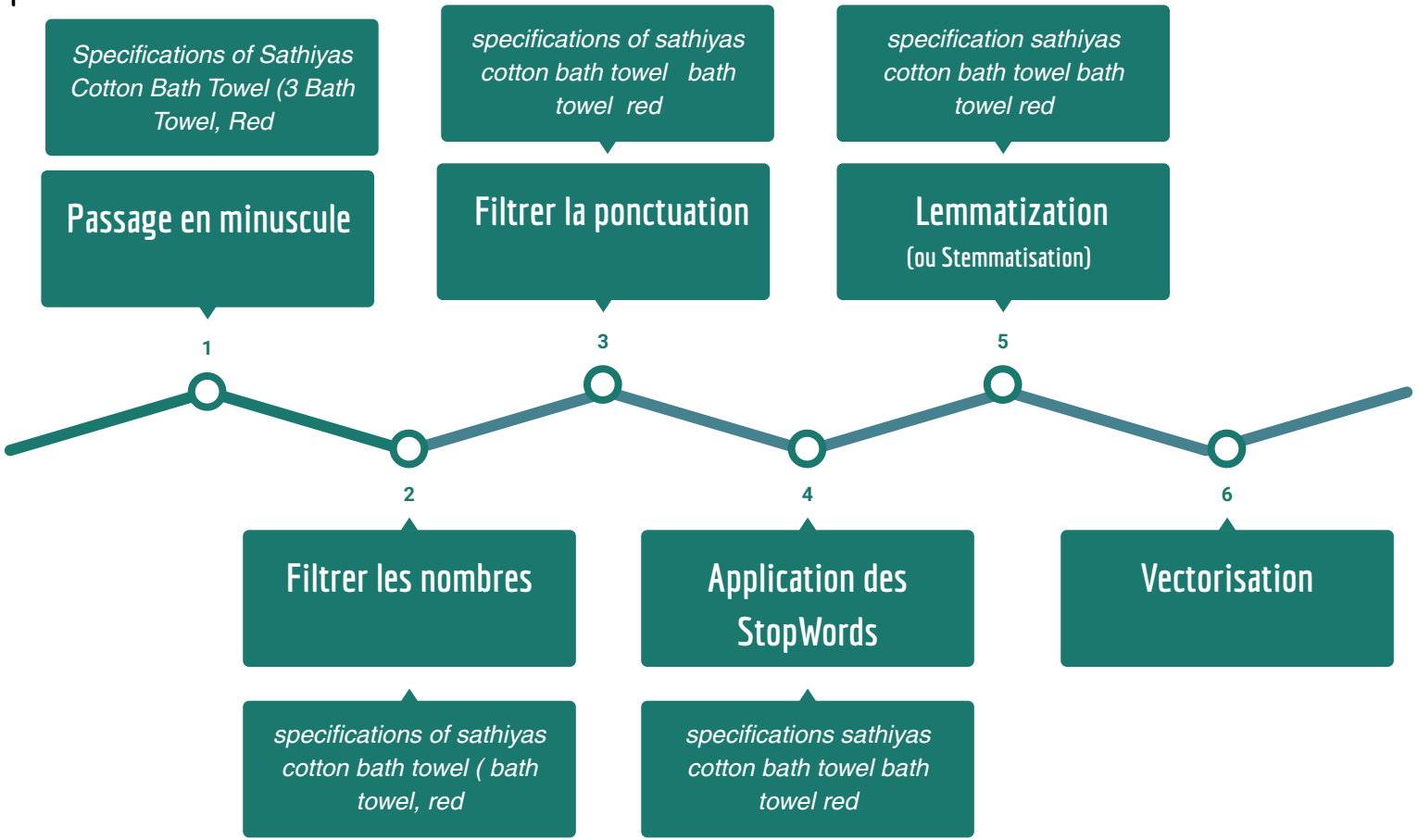
2.4 - Description

2.5 - Étape du processus de faisabilité



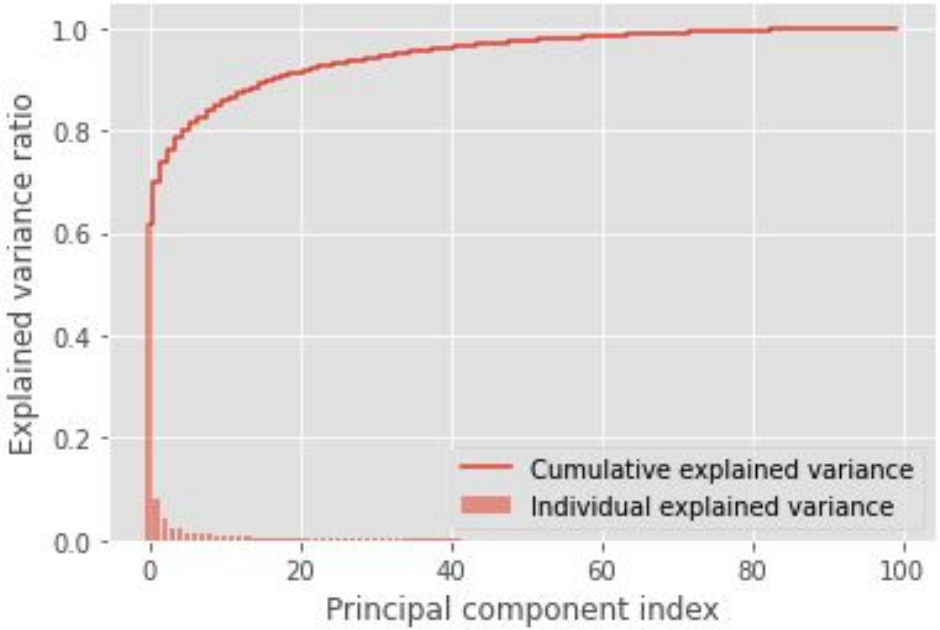
Texte

3.1 - Étape du traitement de texte



3.2 - Texte - Travail préalable

3.2.1 - Réduction dimension : PCA




```
Dimensions dataset avant réduction PCA :  (1048, 100)
Dimensions dataset après réduction PCA :  (1048, 71)
```

3.2.2 - Réduction dimension : T-Sne (2D)

```
TNSE - Texte
```

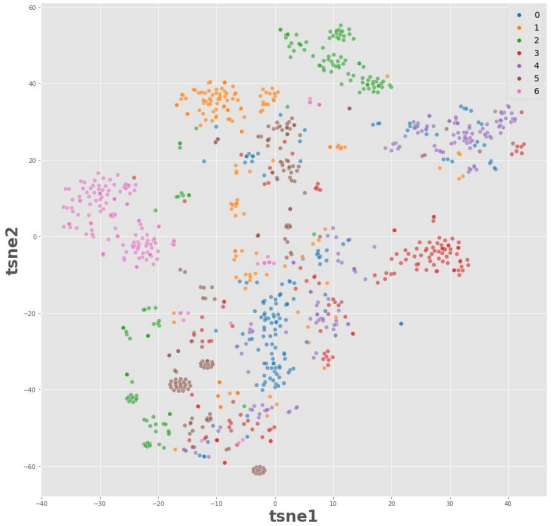
:

	tsne1	tsne2
0	-2.749500	-47.303238
1	-4.353122	-26.918577
2	-10.603967	-49.480179
3	3.624078	-44.831573
4	-11.343846	-57.593109
...	...	...
1043	-15.264066	-56.015659
1044	-11.480761	-53.720688
1045	-2.882393	21.645700
1046	-4.873415	20.905138
1047	-3.781827	20.923018

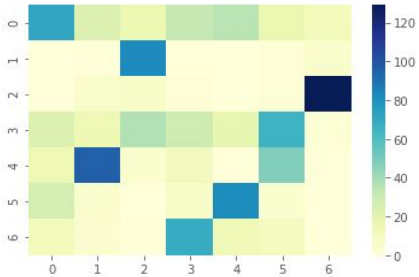
1048 rows × 2 columns

3.3 - Texte - non SuperVisé - KMEANS

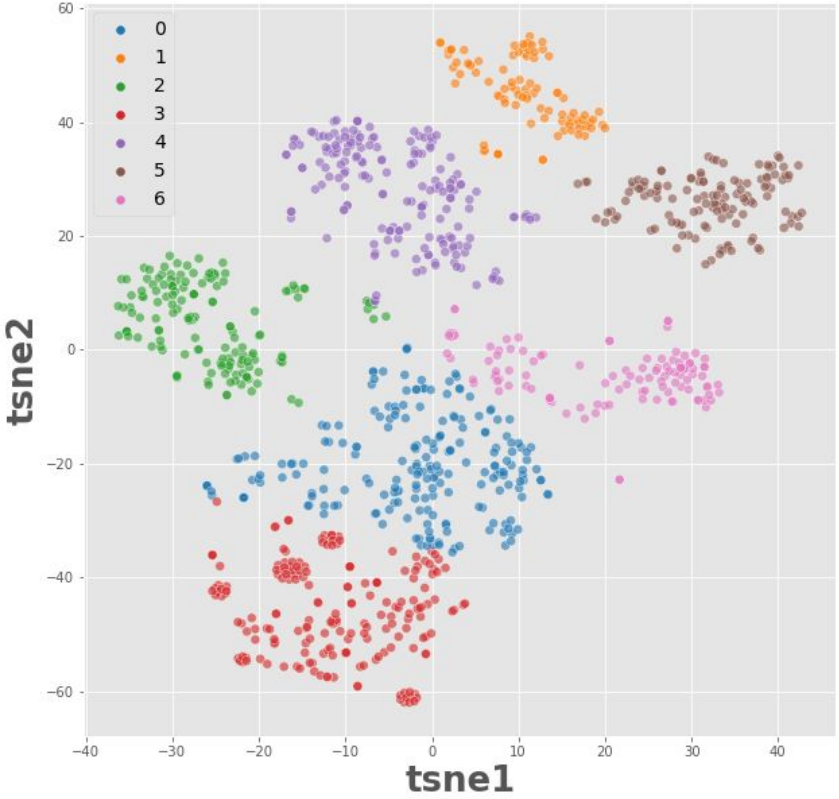
Texte - KMeans - Categories du Dataset



Texte - KMeans - Matrice de confusion

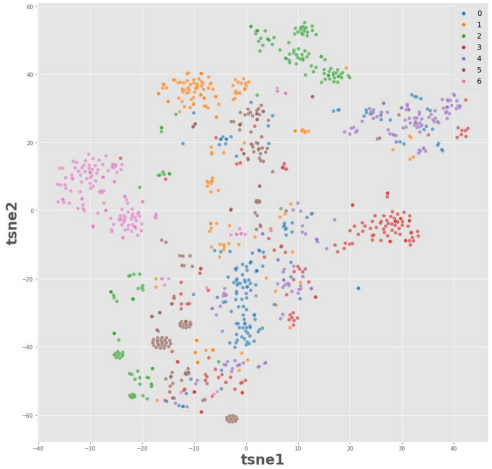


Texte - KMeans - Categories definies

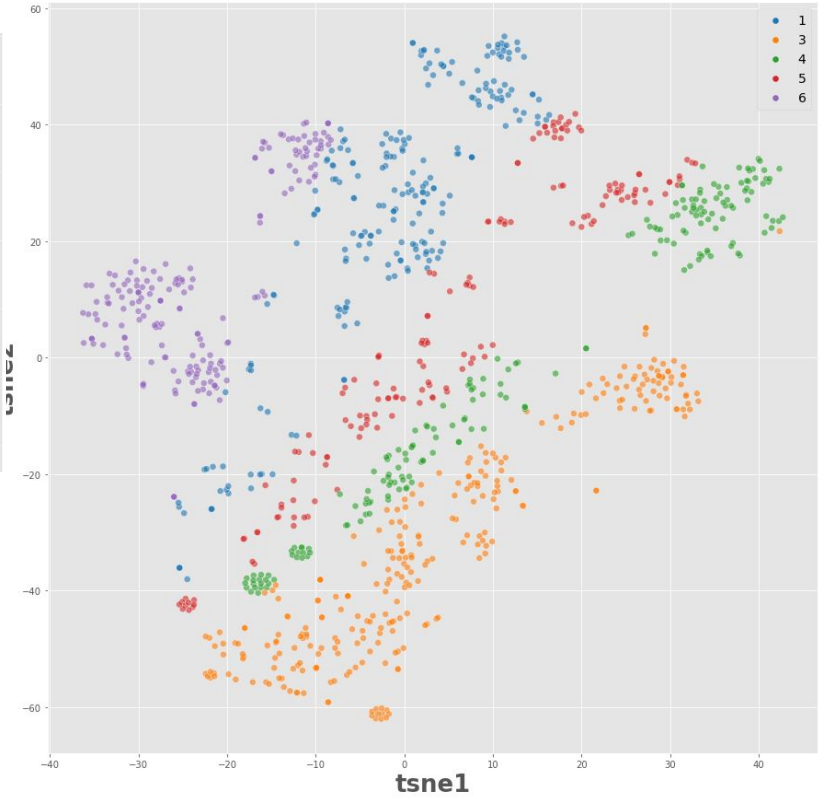


3.4 - Texte - SuperVisé - Naive Bayes

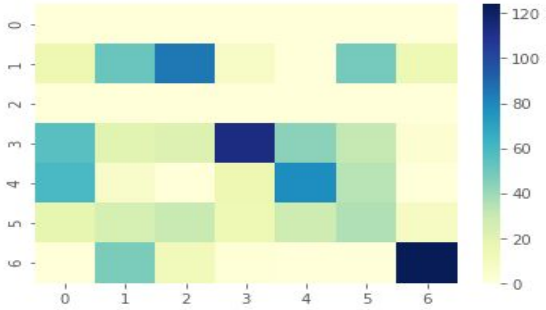
Texte - Naive B. - Categories du DataSet



Images - Naive B. - Categories definies

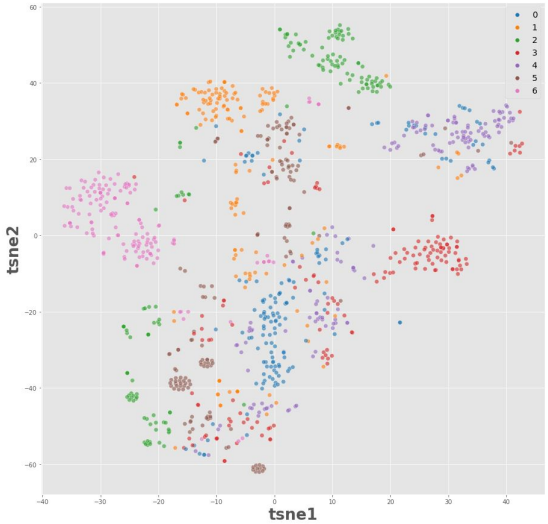


Textes - Naive Bayes - Matrice de confusion

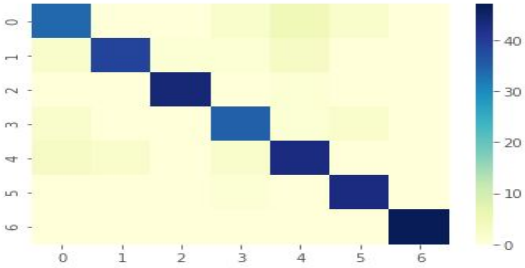


3.5 - Texte - SuperVisé - SVM

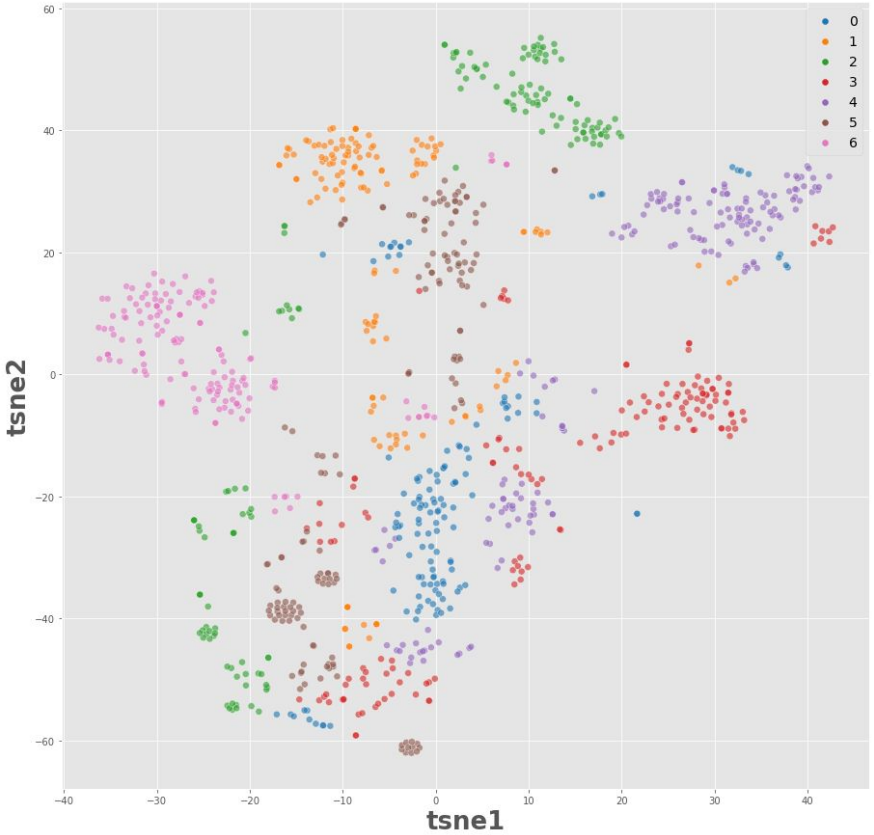
Texte - SVM - Categories du Dataset



Texte - SVM - Avec HP - Matrice



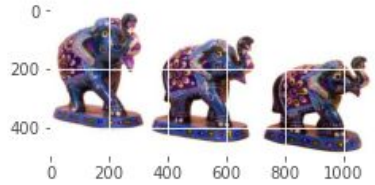
Texte - SVM - Categories definies

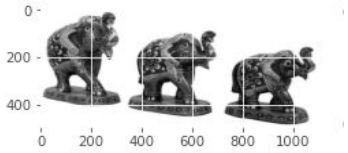


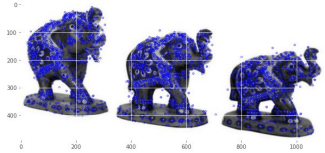
3.6 - Texte - Récapitulatif

Texte		Accuracy Score
	Modele	
Non-Supervisé	K-Means	11
Supervisé	Naive Bayes	28
	Support Vector Machine	88

Images







Importation

Changer le contraste

Trouver les descripteurs

1

2

3

4

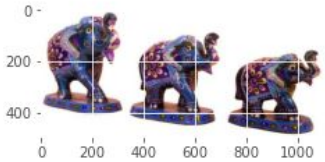
5

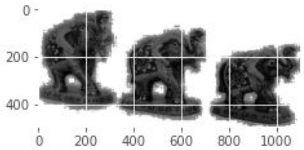
6

Redimension

Normalisation

Vectorisation

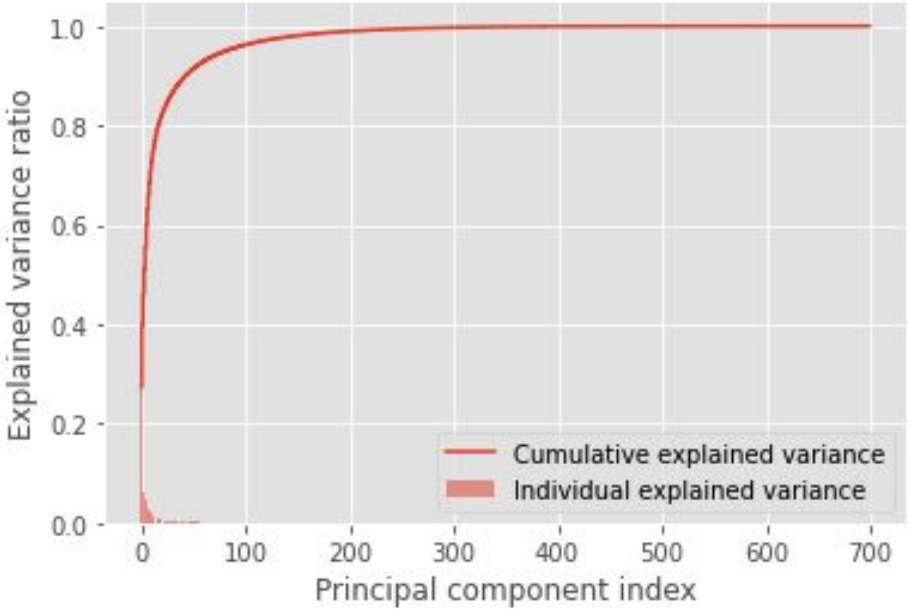




4.1 - Image - Processus

4.2 - Images - Travail préalable

4.2.1 - Réduction dimension : PCA




```
Dimensions dataset avant réduction PCA :  (1048, 700)
Dimensions dataset après réduction PCA :  (1048, 44)
```

4.2.2 - Réduction dimension : T-Sne (2D)

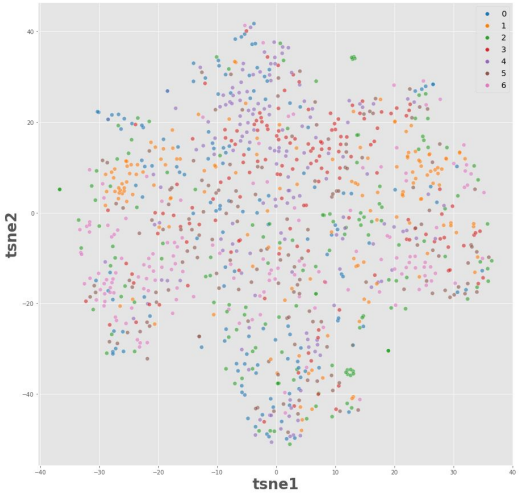
TNSE - Image

	tsne1	tsne2
0	-12.440010	21.075386
1	3.874655	-41.413559
2	-20.821356	8.433664
3	-26.129625	6.386205
4	-28.081547	-11.710289
...	...	...
1043	-3.154574	10.160899
1044	1.427474	-5.849862
1045	25.629892	22.901733
1046	-6.140348	24.798140
1047	-11.114718	-9.849862

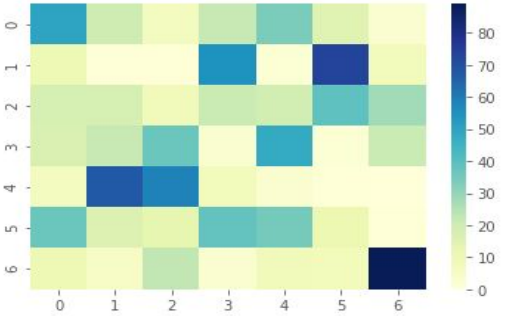
1048 rows × 2 columns

4.3 - Images- non SuperVisé - KMEANS

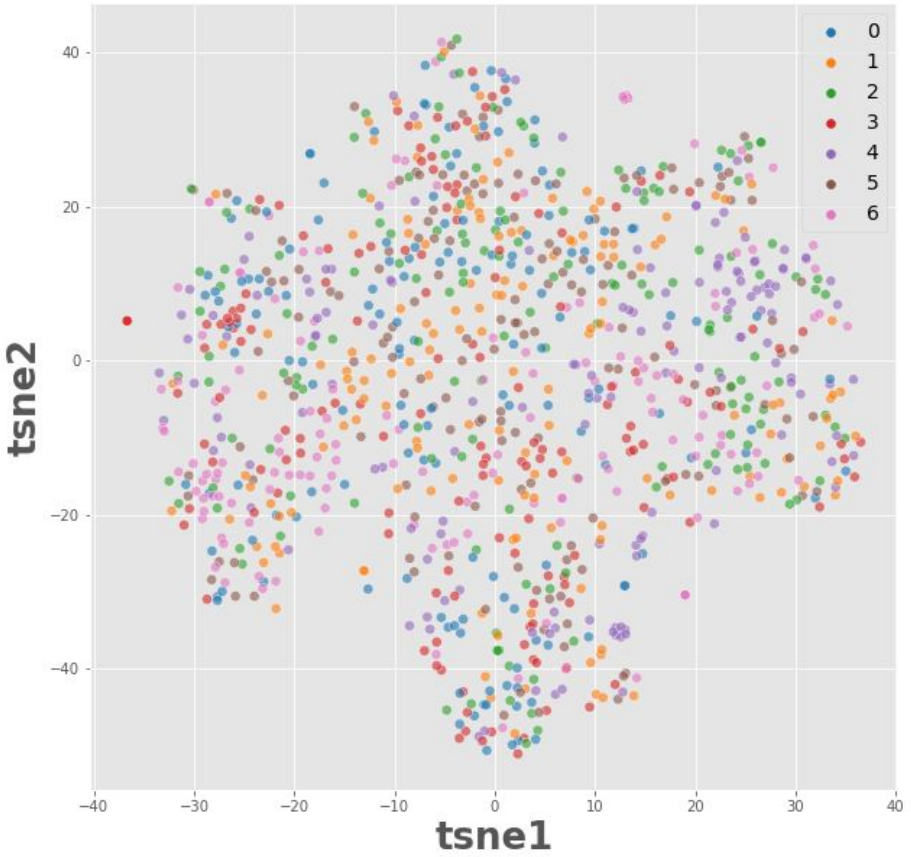
Images - KMeans - Categories du Dataset



Images - KMeans - Matrice de confusion

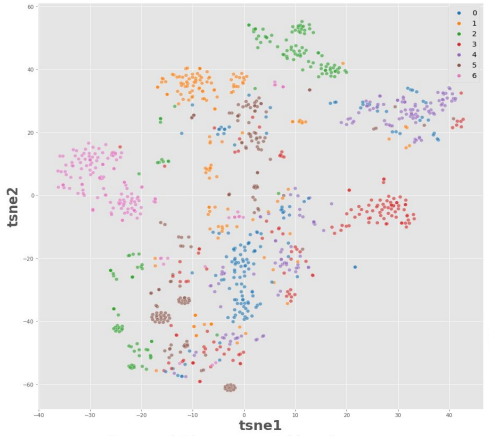


Images - KMeans - Categories definies

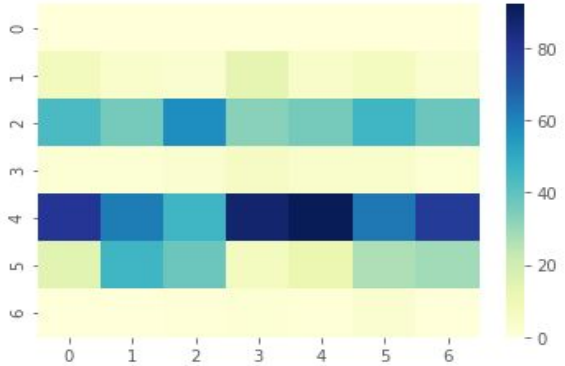


4.4 - Images - SuperVisé - Naive Bayes

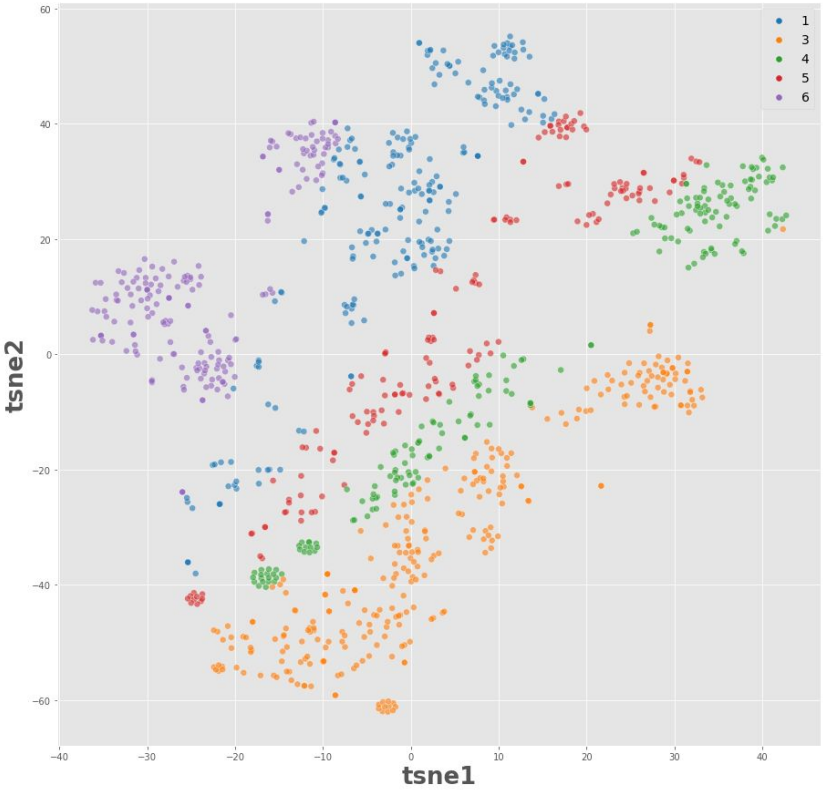
Texte - Naive B. - Categories du DataSet



Images - Naive Bayes - Matrice de confusion

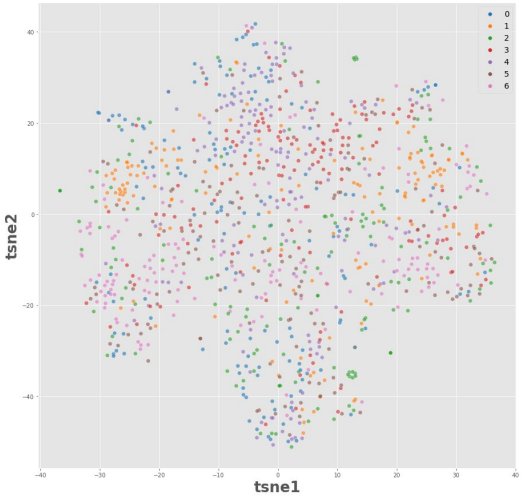


Images - Naive B. - Categories definies

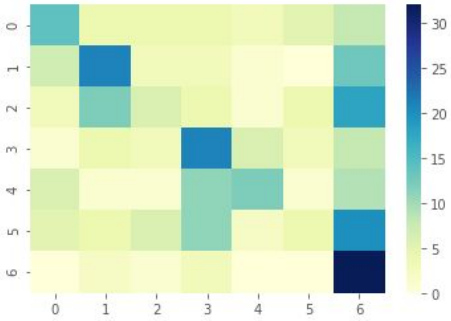


4.5 - Images - SuperVisé - SVM

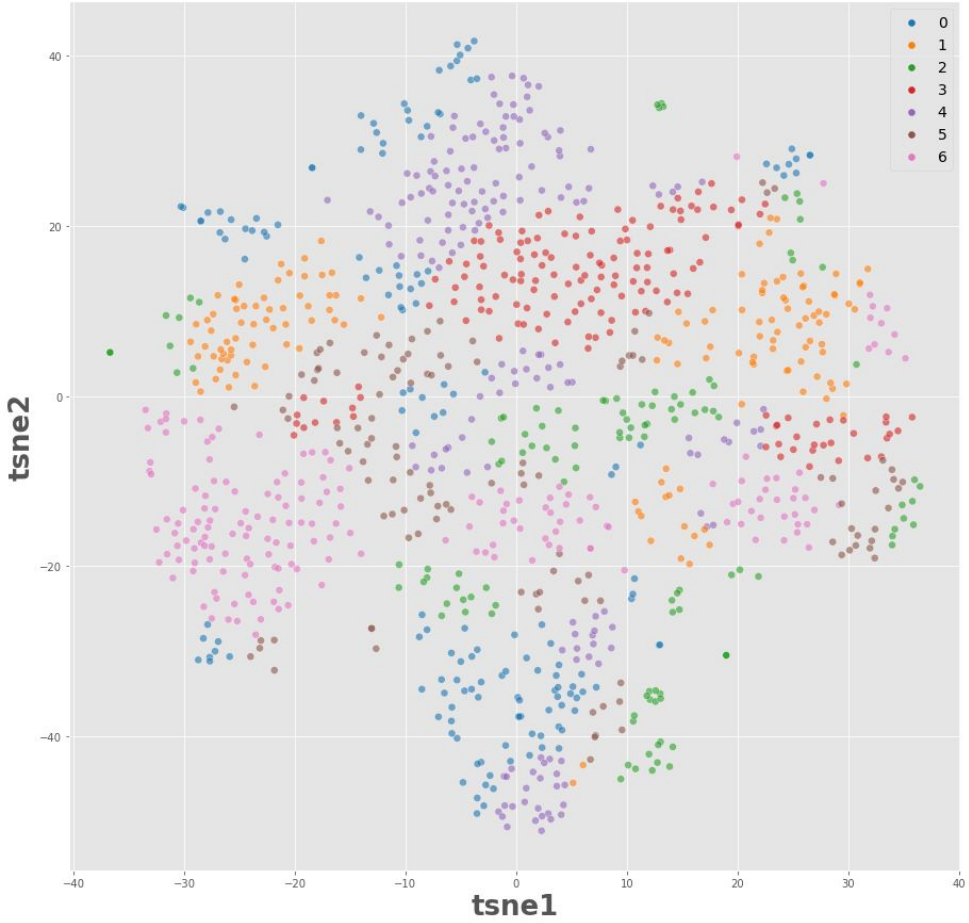
Texte - SVM - Categories du DataSet



Images - SMV - Sans HP - Matrice de confusion



Images - SVM - Categories definies



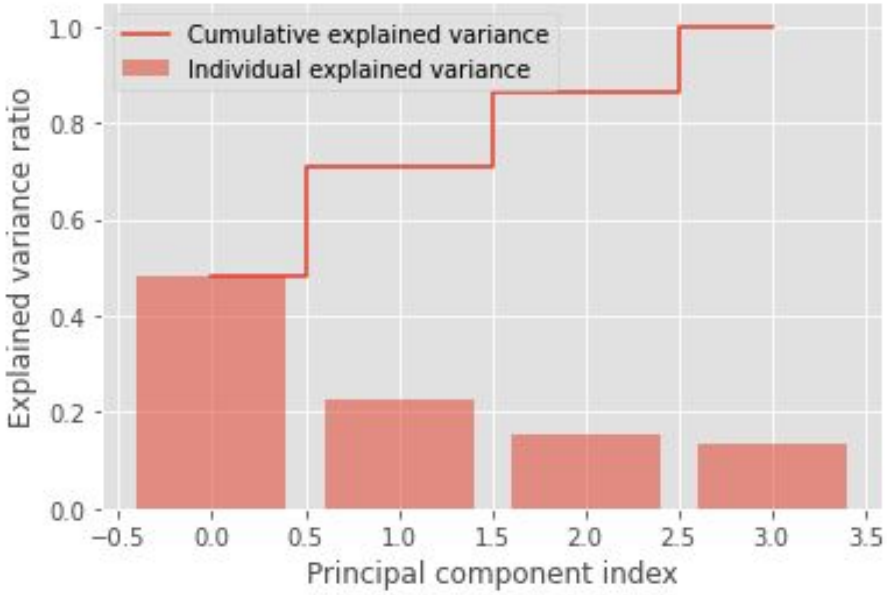
4.6 - Images - Récapitulatif

Image		Accuracy Score
	Modele	
Non-Supervisé	K-Means	15
Supervisé	Naive Bayes	16
	Support Vector Machine	36

Reunion

5.1 - Reunion - Travail préalable

5.1.1 - Réduction dimension : PCA




```
Dimensions dataset avant réduction PCA :  (1048, 4)
Dimensions dataset après réduction PCA :  (1048, 4)
```

5.1.2 - Réduction dimension : T-Sne (2D)

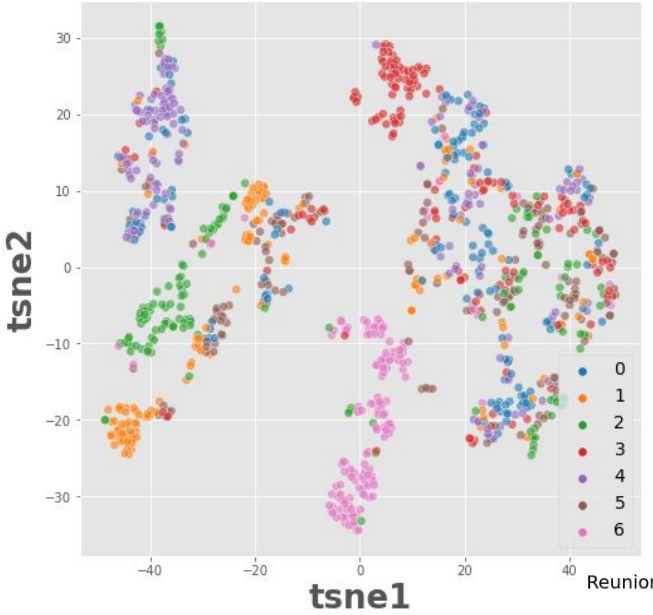
TNSE - Reunion

	tsne1	tsne2
0	-20.276886	9.728692
1	-28.746336	-6.249609
2	-20.151337	10.107433
3	-19.260603	7.550618
4	-19.295708	10.735259
...	...	...
1043	-16.124903	15.069901
1044	-14.605281	21.346445
1045	48.978096	9.353537
1046	43.719349	5.636261
1047	48.660324	9.673699

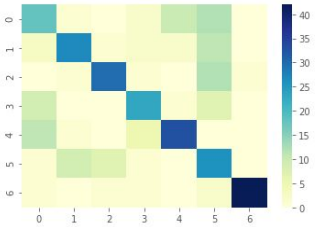
1048 rows × 2 columns

5.2 - Reunion - SuperVisé - SVM

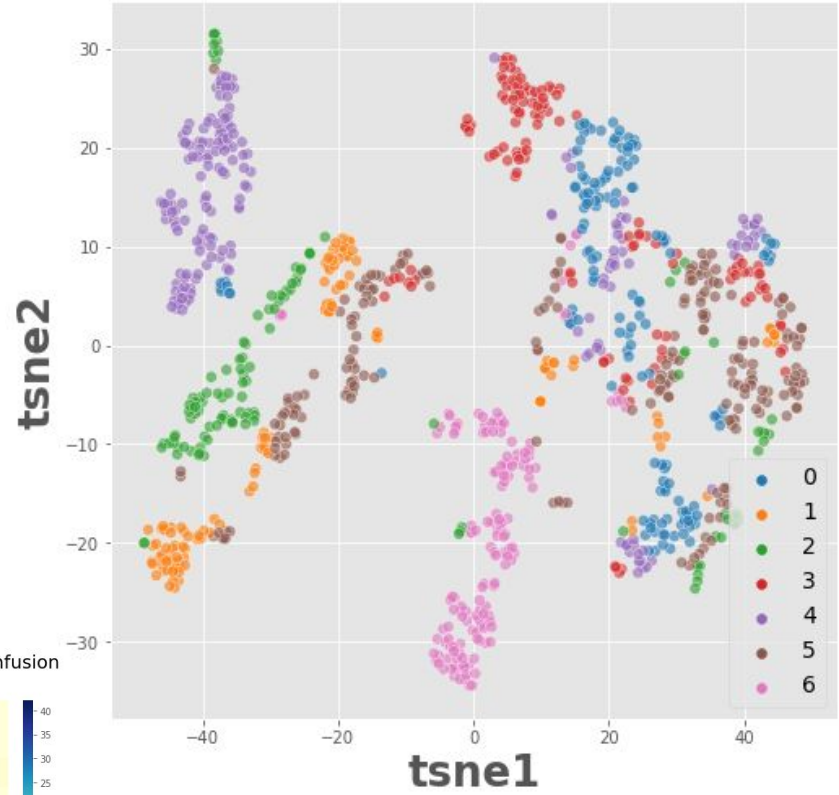
Reunion - SVM - Categories du DataSet



Reunion - SVM - Matrice de confusion

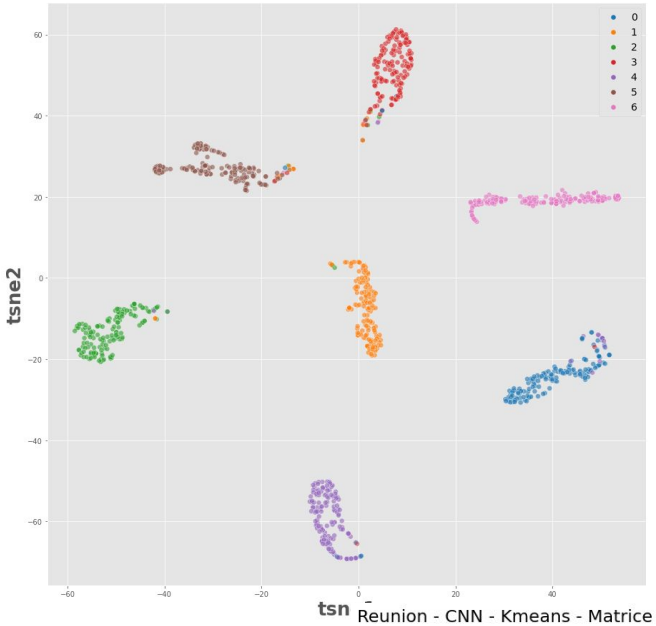


Reunion - SVM - Categories definies

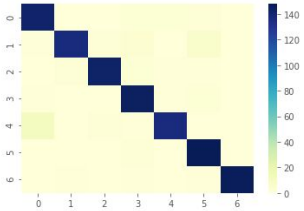


5.3 - Reunion - CNN - Kmeans

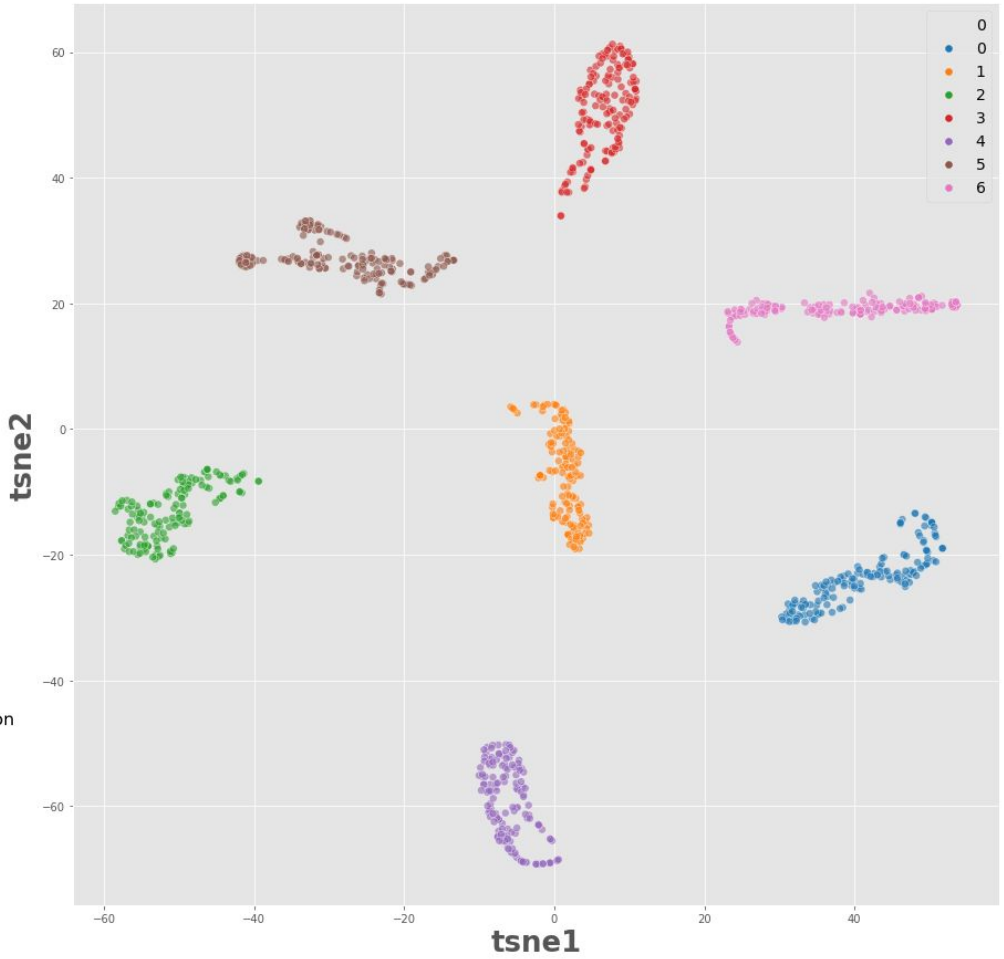
Reunion - CNN - Kmeans - Categories du DataSet



Reunion - CNN - Kmeans - Matrice de confusion



Reunion - CNN - Kmeans - Categories definies



5.3 - Reunion - Récapitulatif

Reunion		Accuracy Score
	Modele	
Supervisé	Naive Bayes	33
	Support Vector Machine	79
	Reseaux de neurones	93

Conclusion

En conclusion, nous pouvons répondre de manière évidente qu'une classification d'articles basé sur un image et une description, est tout à fait possible au sein des 7 catégories en présence.

En particulier grâce aux modèle Support Vector Machine et à un réseau de neurones .

Piste d'amélioration:

- Faire régulièrement une étude pour mettre à jour les données
- Etudier les erreurs des modèles pour améliorer votre score
- Tester d'autres méthodes de pré-processing, du texte en particulier
- Tester un réseau de neurones qui analyse le texte et les images ensemble