

Projet 6 : Classifier automatiquement des biens de consommation

Mentor : Amine HADJ-YOUCER

Etudiant : Marin DUCHEMIN



Plan de la Présentation

I ~ Rappel de la Problématique et Apperçu du Jeu de Données

II ~ Données Texte

III ~ Données Images

Conclusions et Perspectives

I ~ Rappel de la Problématique

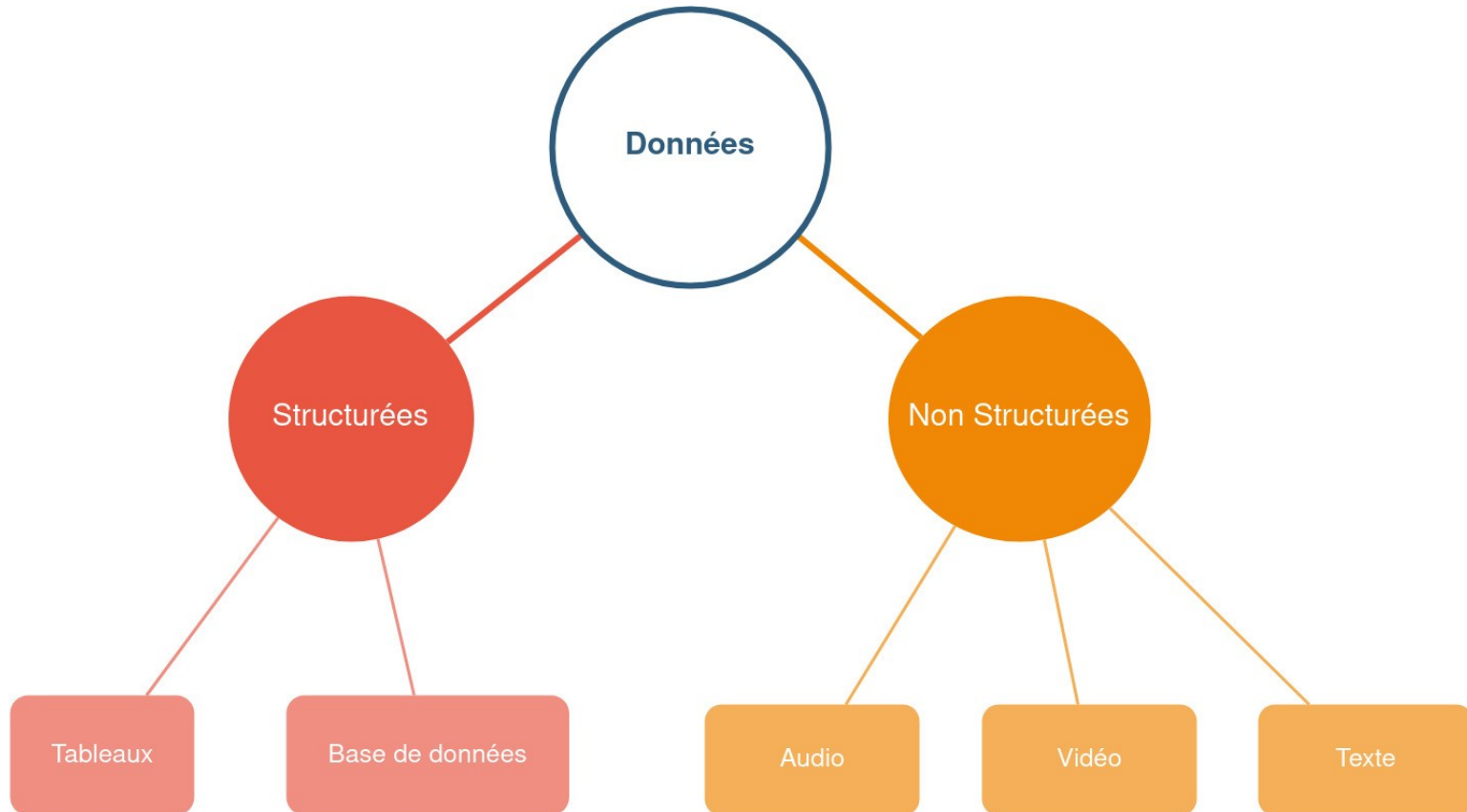
Place de Marché : marketplace e-commerce

Objectifs :

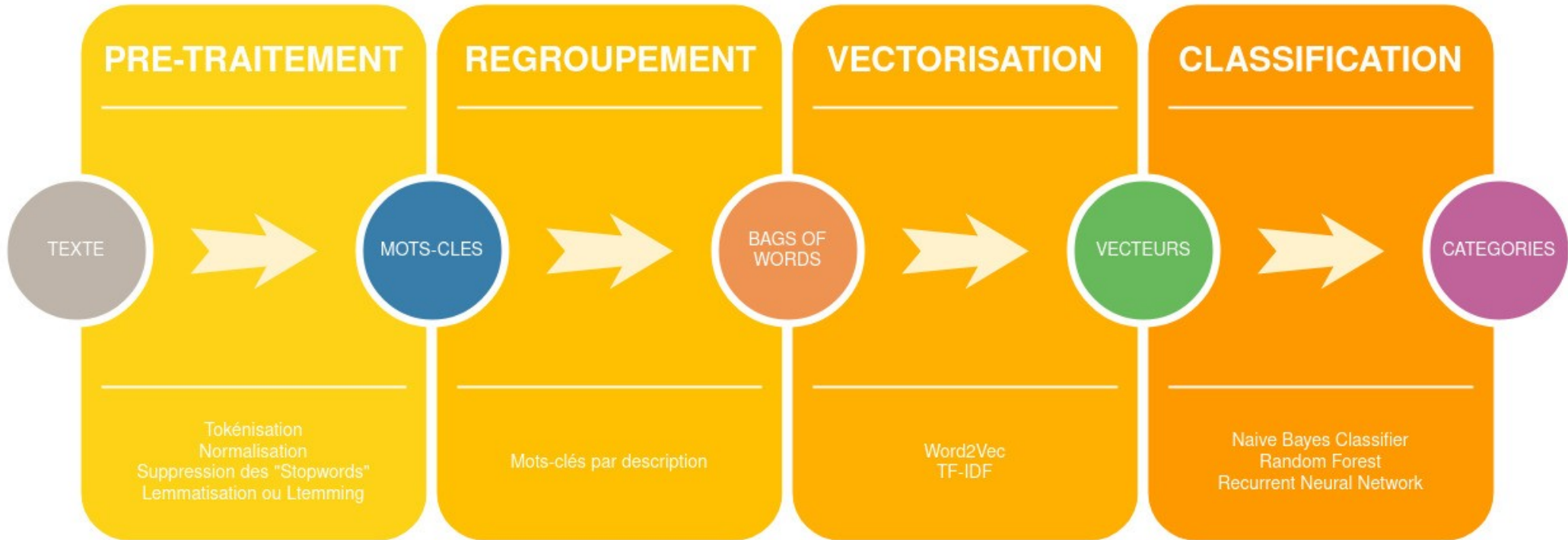
- Etude de faisabilité d'un moteur de classification des biens
- Classification basée sur une description et une image



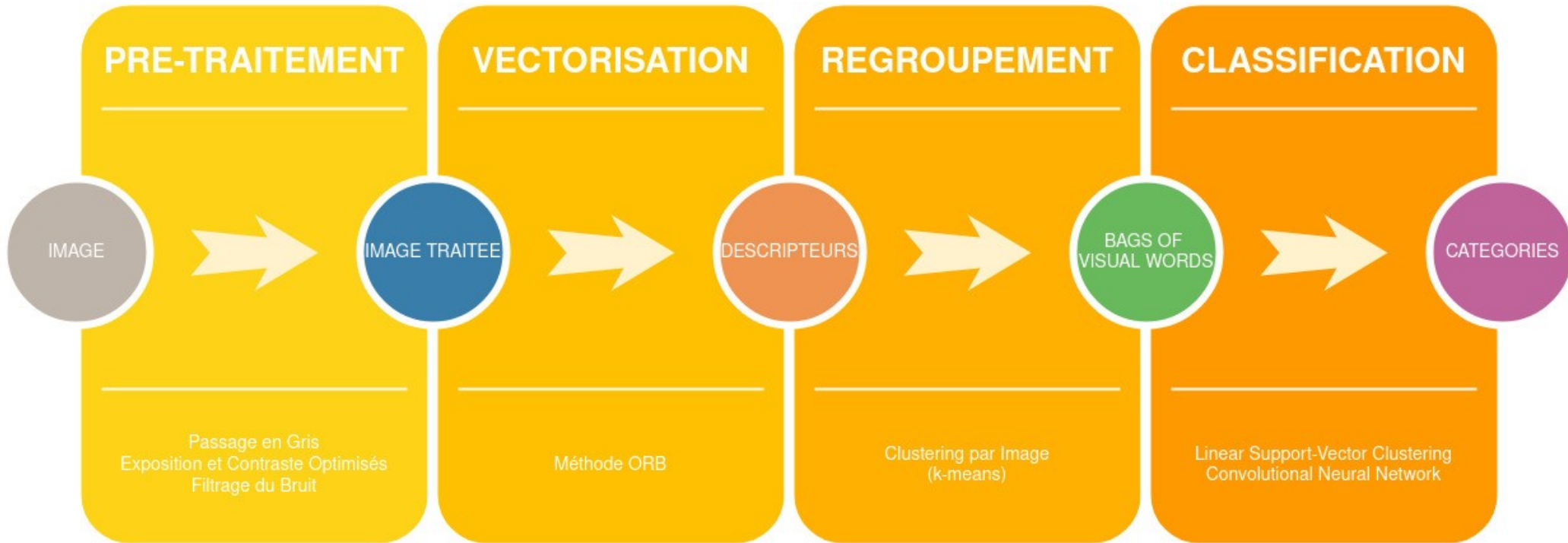
I ~ Rappel de la Problématique



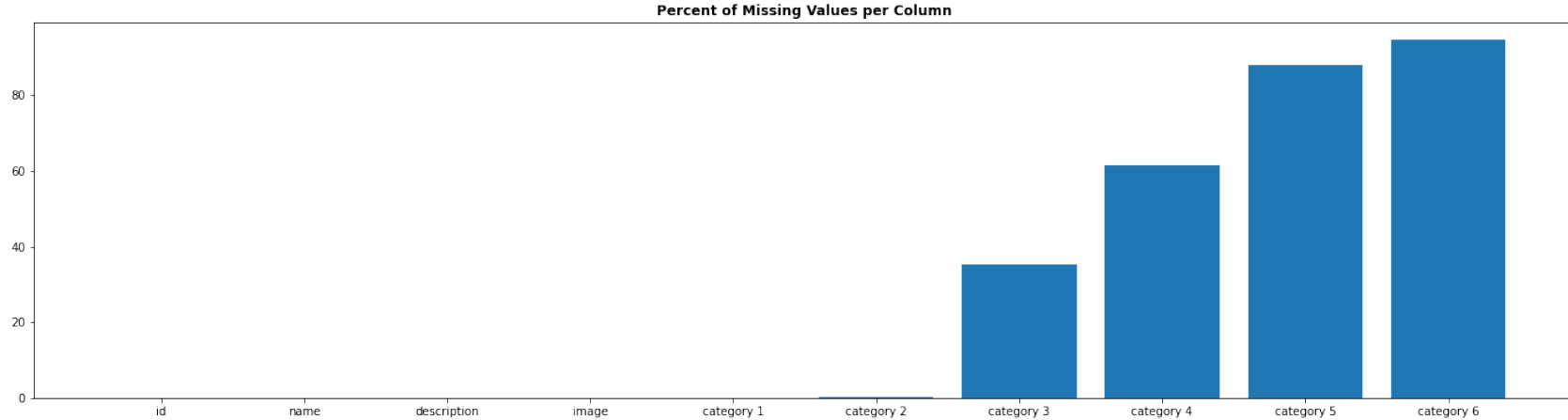
I ~ Rappel de la Problématique



I ~ Rappel de la Problématique



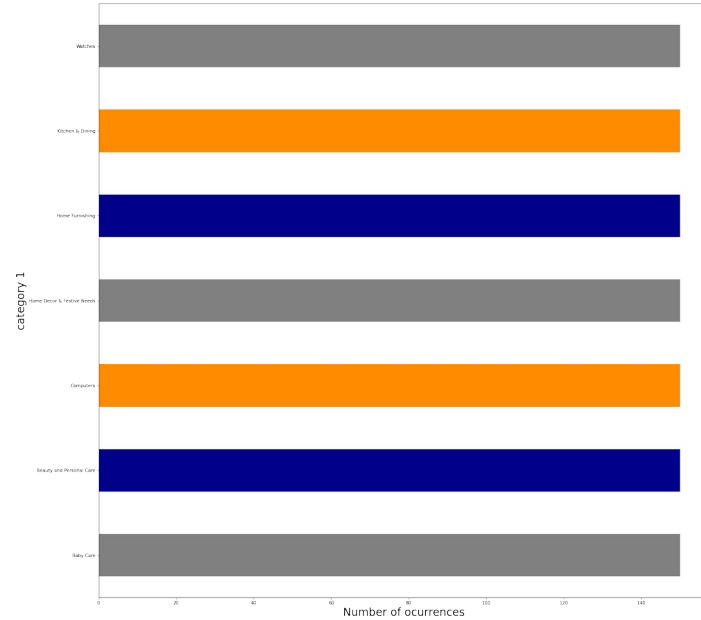
I ~ Rappel de la Problématique



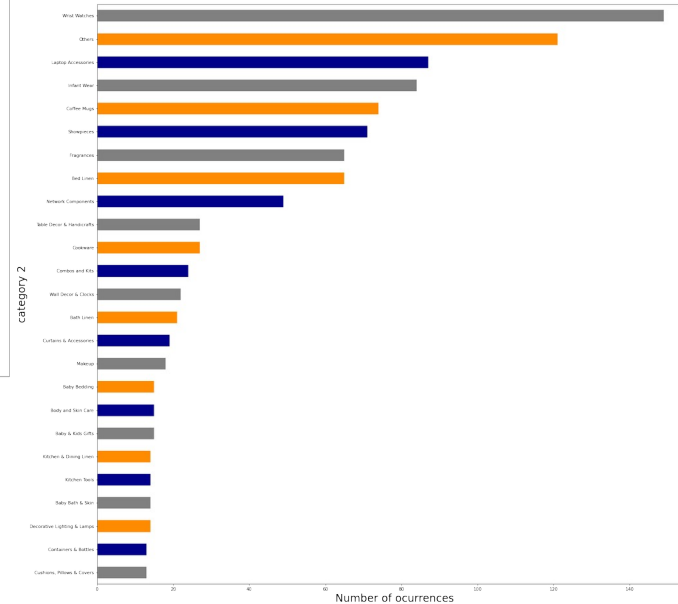
- Aucune description ou image manquante.
- Peu d'informations sur les trois dernières catégories.

I ~ Rappel de la Problématique

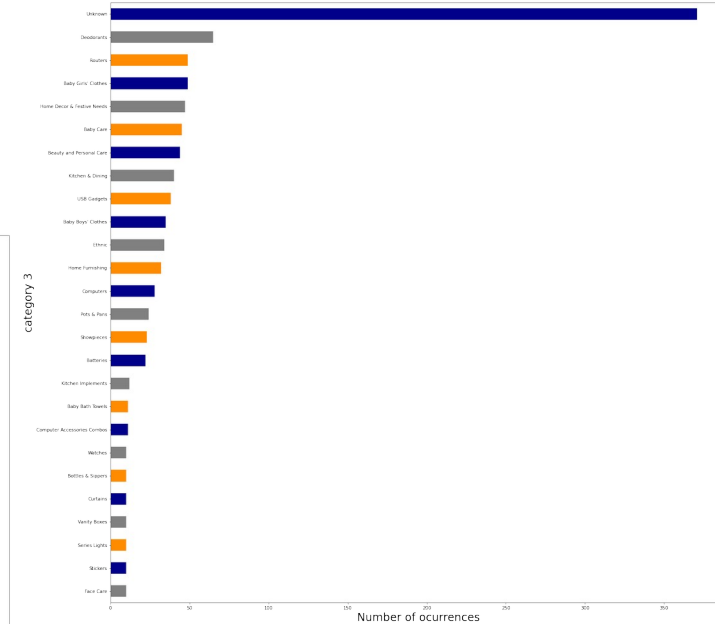
Number of Product in each category



Number of Product in each category



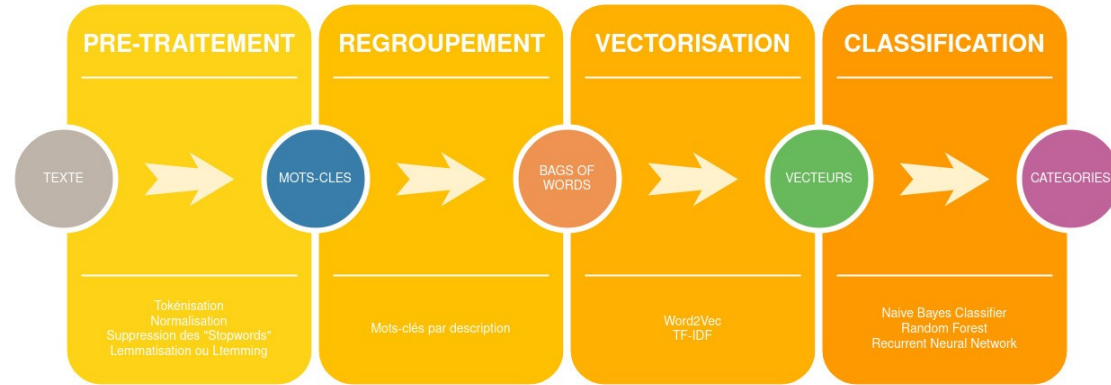
Number of Product in each category



II ~ Pré-traitement du Texte

Pré-traitements (nltk, gensim, wordcloud):

- Tokenisation
- Passage en minuscule
- Suppression des 'stopwords'
- Lemmatisation



We went from this :

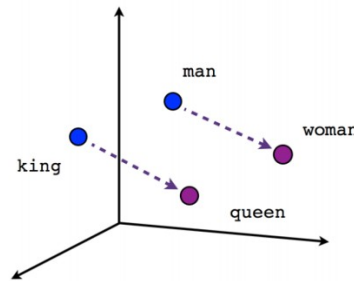
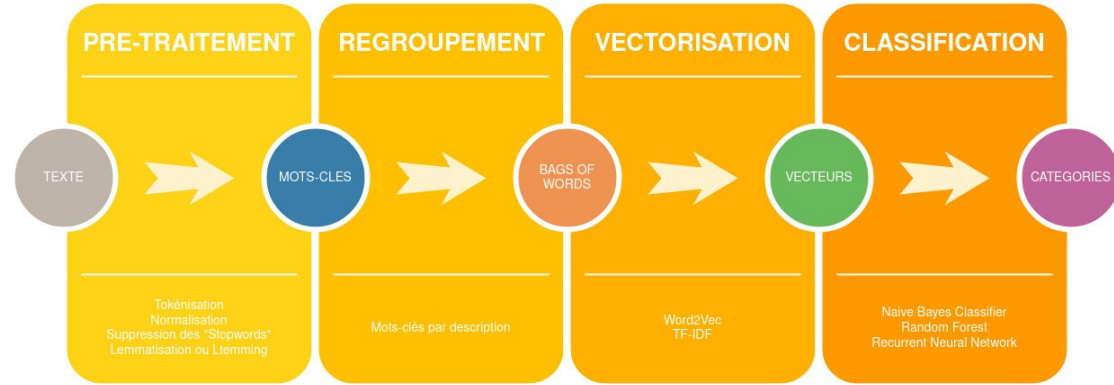
This is a text, made to illustrate the text's transformations processes that will and have occurred.
to this :

```
['text', 'make', 'illustrate', 'text', 'transformation', 'process', 'occur']
```

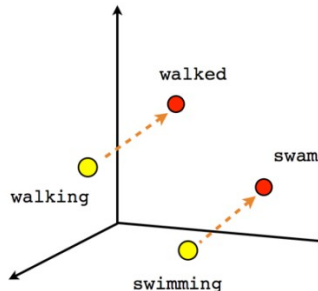

II ~ Pré-traitement du Texte

Vectorisation :

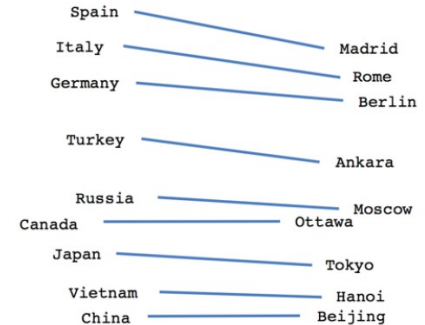
- Term Frequency - Inverse Document Frequency (tf-idf)
- Word2Vec (Continuous Bags of Words, Skip Gram)



Male-Female



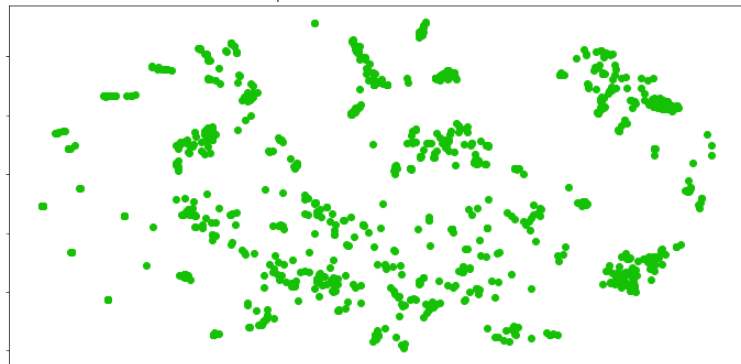
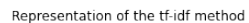
Verb tense



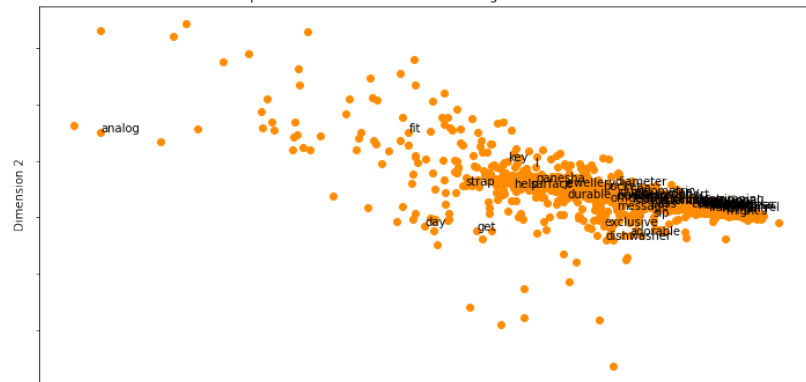
Country-Capital

II ~ Pré-traitement du Texte

Vectorisation :

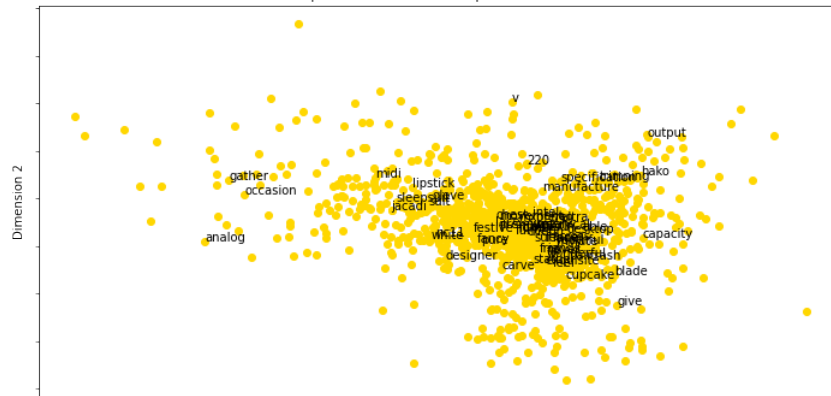


Dimension 1



Dimension 1

Representation of the Skip Gram method



Dimension 1

II ~ Classification du Texte

Aller plus loin, la classification :

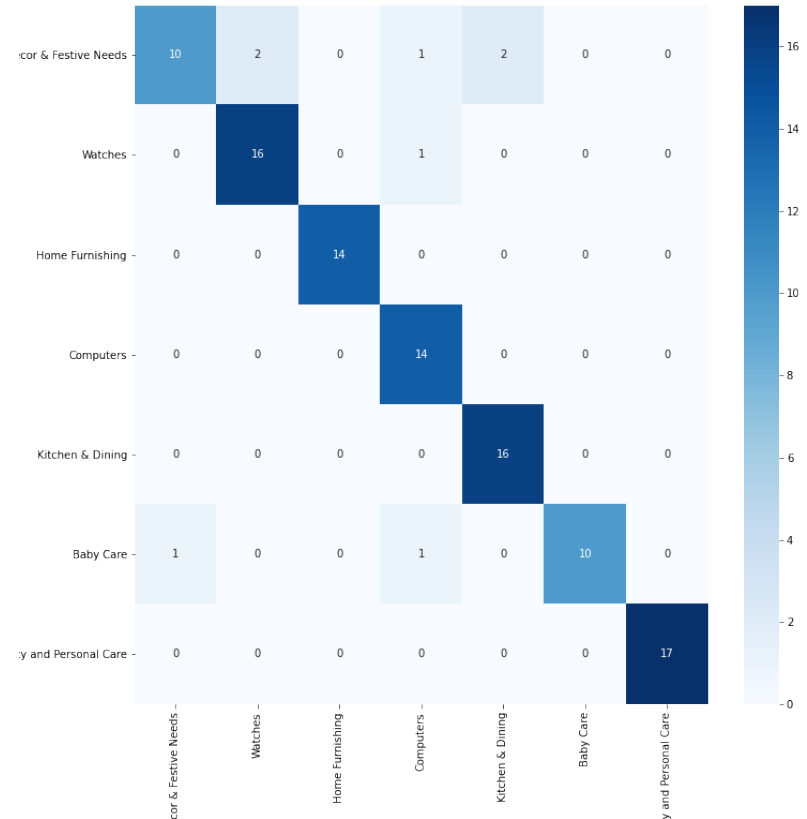
- En entrée : vecteurs
- En sortie : niveaux de catégories
- Séparation en jeux d'entraînement et de test

II ~ Classification du Texte

Classification :

- Naive Bayes Classifier
 - Niveau 1 : 92%
 - Niveau 2 : 74%
 - Niveau 3 : 67%

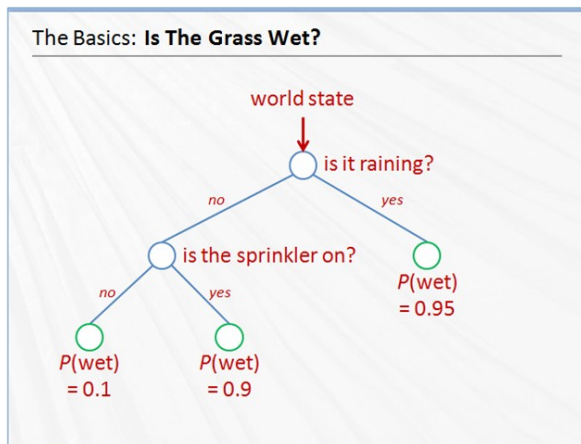
CONFUSION MATRIX - Multinomial Naive Bayes Classifier on the Categories of Level 1



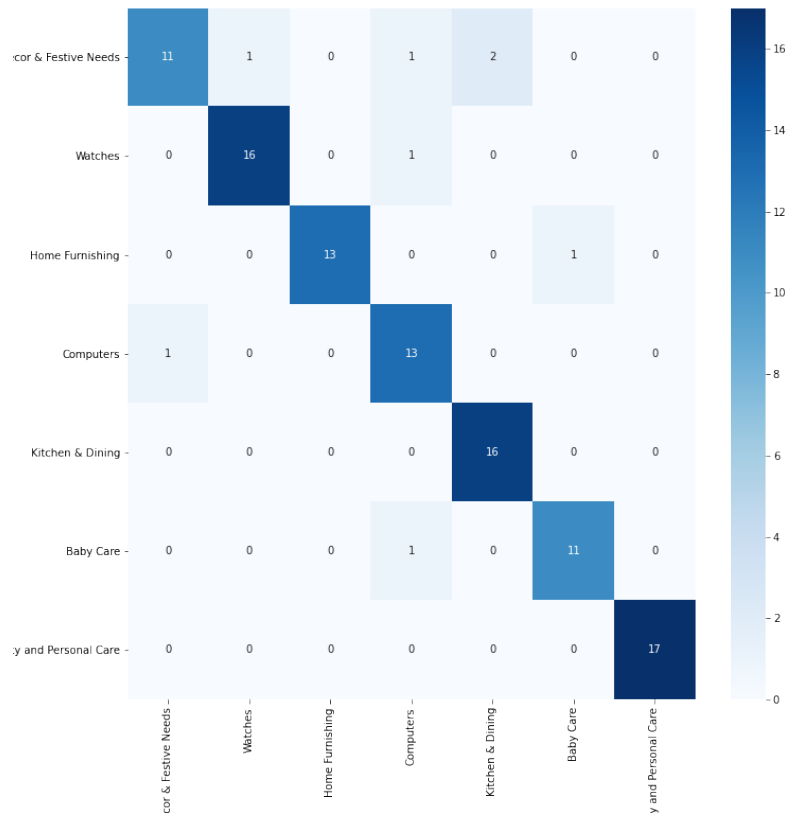
II ~ Classification du Texte

Classification :

- Random Forests
 - Niveau 1 : 92%
 - Niveau 2 : 87%
 - Niveau 3 : 83%



CONFUSION MATRIX - Random Forest Classifier on the Categories of Level 1

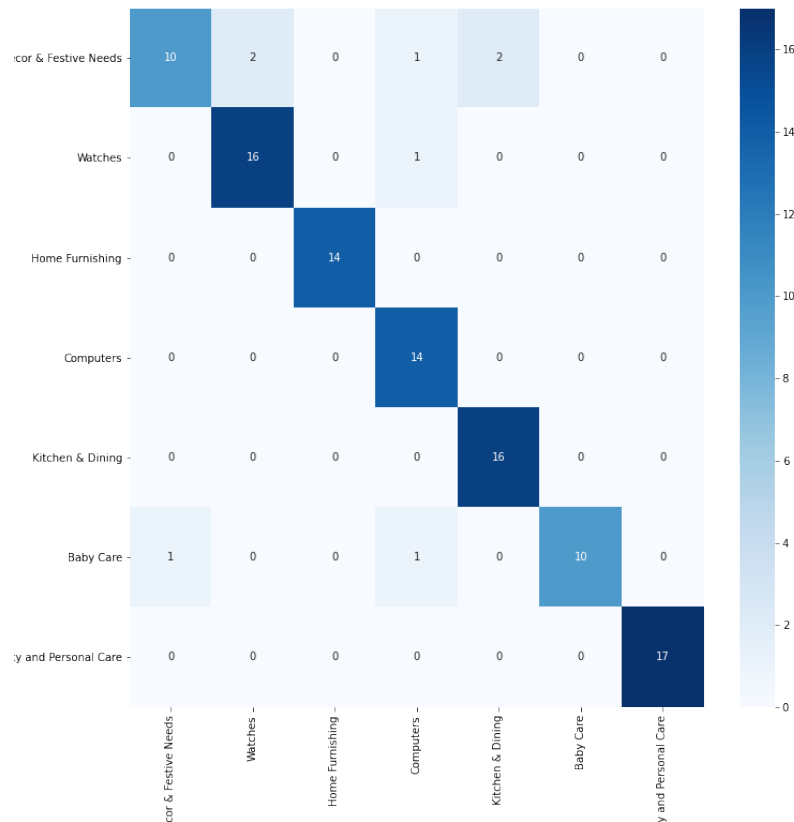


II ~ Classification du Texte

Classification :

- Recurrent Neural Network
 - Niveau 1 : 93%
 - Niveau 2 : 84%
 - Niveau 3 : 80%
- Déployé par moi même
 - Modèle simple, 3 couches
 - Testé sur CBOW et Skip Gram

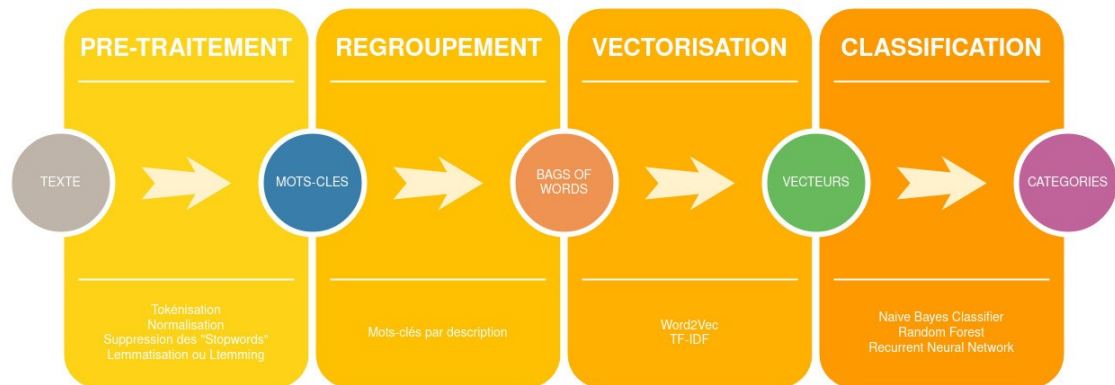
CONFUSION MATRIX - Multinomial Naive Bayes Classifier on the Categories of Level 1



II ~ Classification du Texte

Classification :

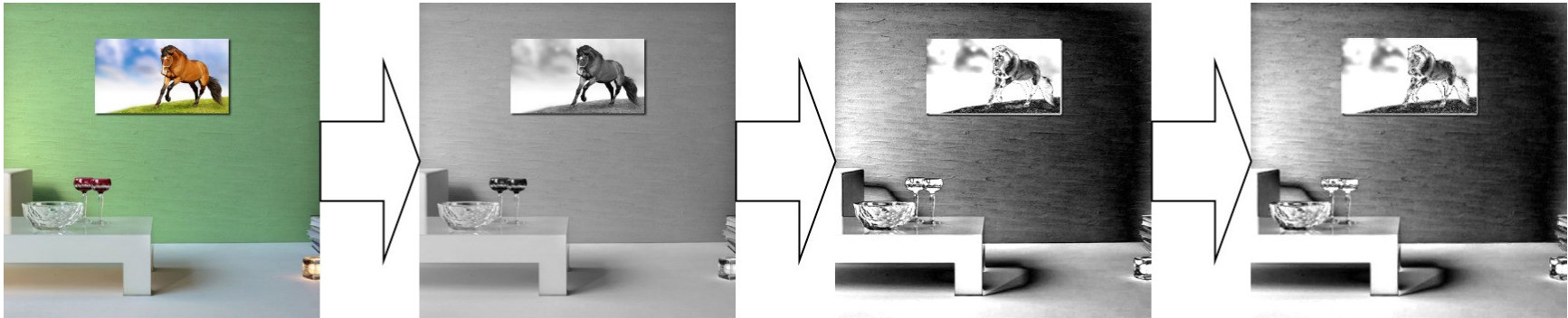
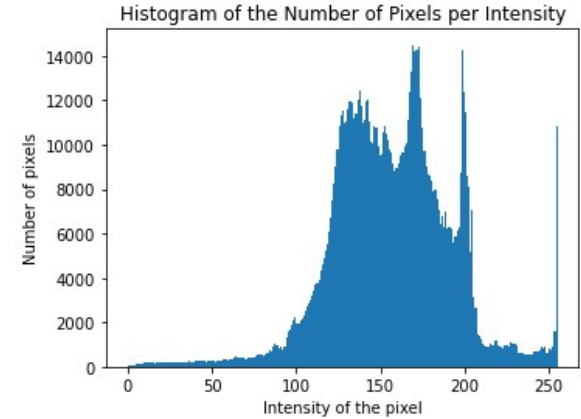
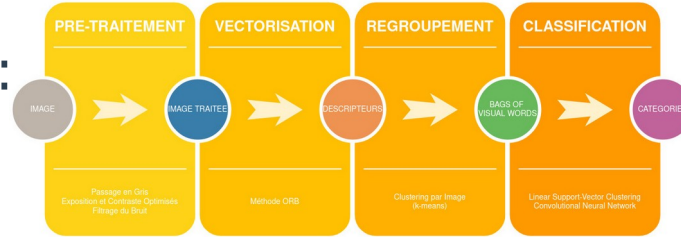
		Accuracy	Time
Level of Category	Model		
Level 1	Multinomial Naive Bayes	92.380952	0.023243
	Random Forest	92.380952	0.520511
	RNN on Continuous Bag of Word	80.000001	13.767050
	RNN on Skip Gram	92.631578	17.020003
Level 2	Multinomial Naive Bayes	74.285714	0.011050
	Random Forest	86.666667	0.568172
	RNN on Continuous Bag of Word	70.526314	29.507907
	RNN on Skip Gram	84.210527	28.094129
Level 3	Multinomial Naive Bayes	66.666667	0.022552
	Random Forest	82.857143	0.608572
	RNN on Continuous Bag of Word	66.315788	41.373782
	RNN on Skip Gram	80.000001	29.876024



III ~ Pré-traitement des Images

Pré-traitement (OpenCV) :

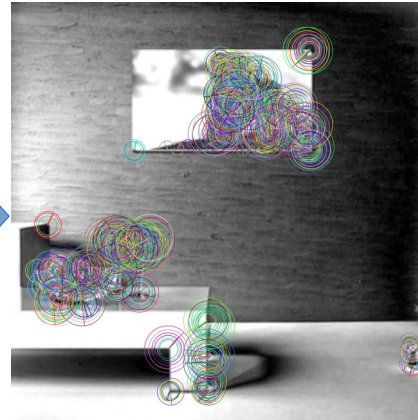
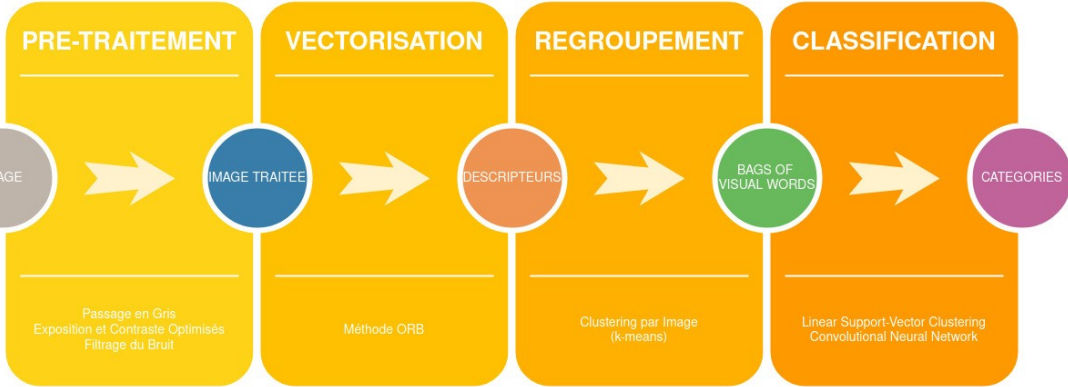
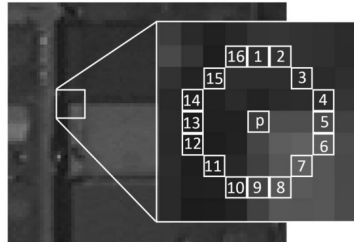
- Passage en Gris
- Exposition et Contraste Optimisés
- Filtrage du Bruit



III ~ Pré-traitement des Images

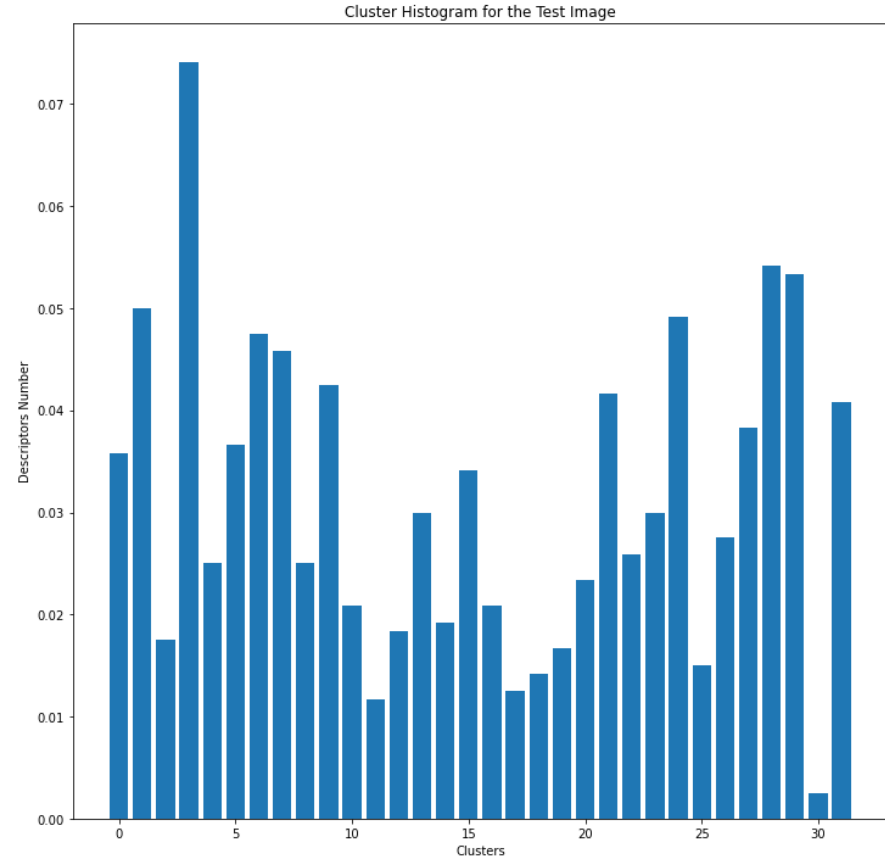
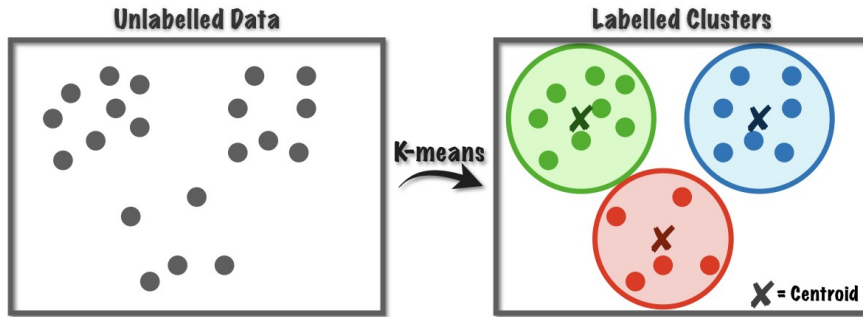
Vectorisation :

Méthode ORB : Oriented FAST and rotated BRIEF



III ~ Pré-traitement des Images

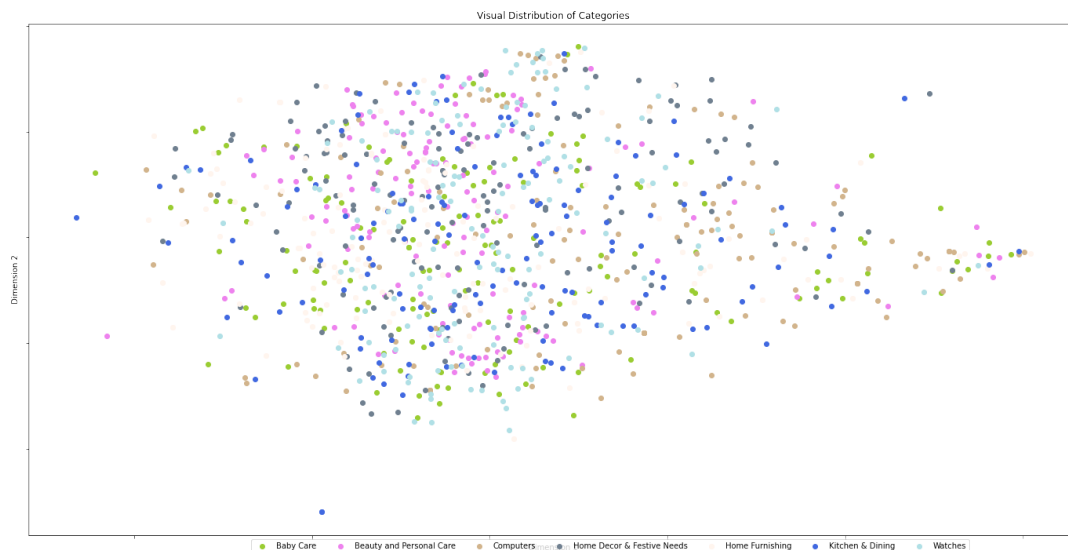
Regroupement :
K-means Clustering



III ~ Pré-traitement des Images

Réduction Dimensionnelle :

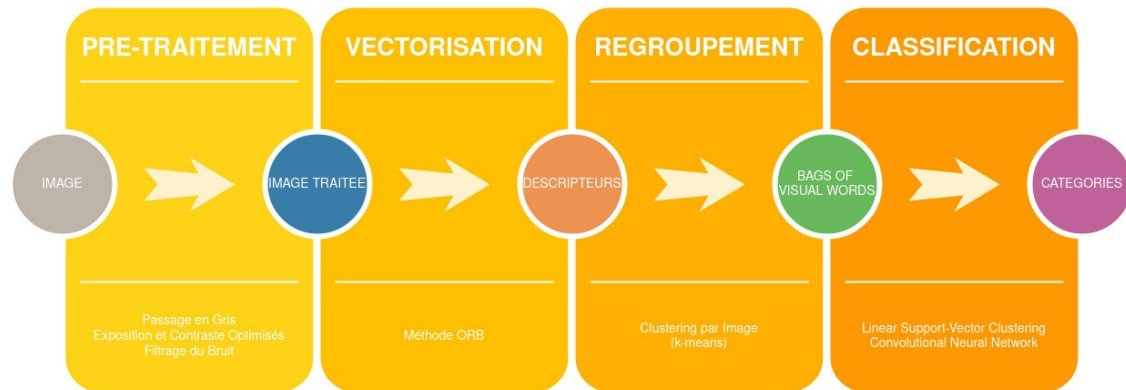
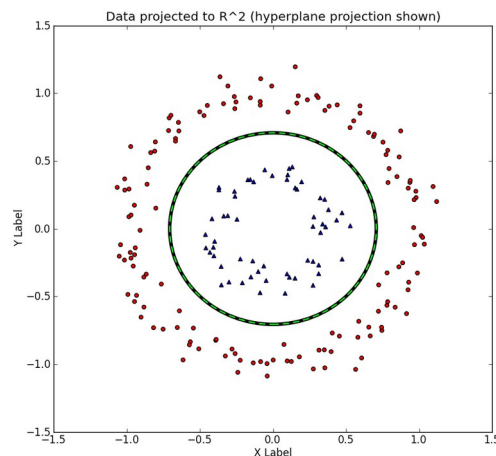
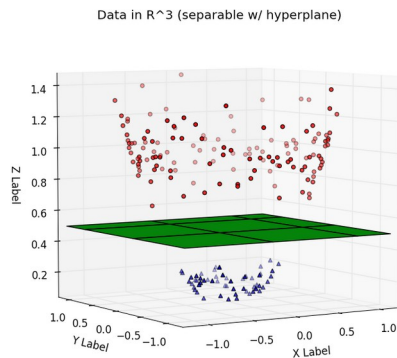
- Décorrélation :
 - Meilleure séparation
 - Réduction du temps de traitement
- PCA :
 - 99% variance expliquée
 - 32 à 28 composantes
- T-SNE :
 - Visualisation



III ~ Classification des Images

Classification:

Linear Support-Vector Clustering (SVC)



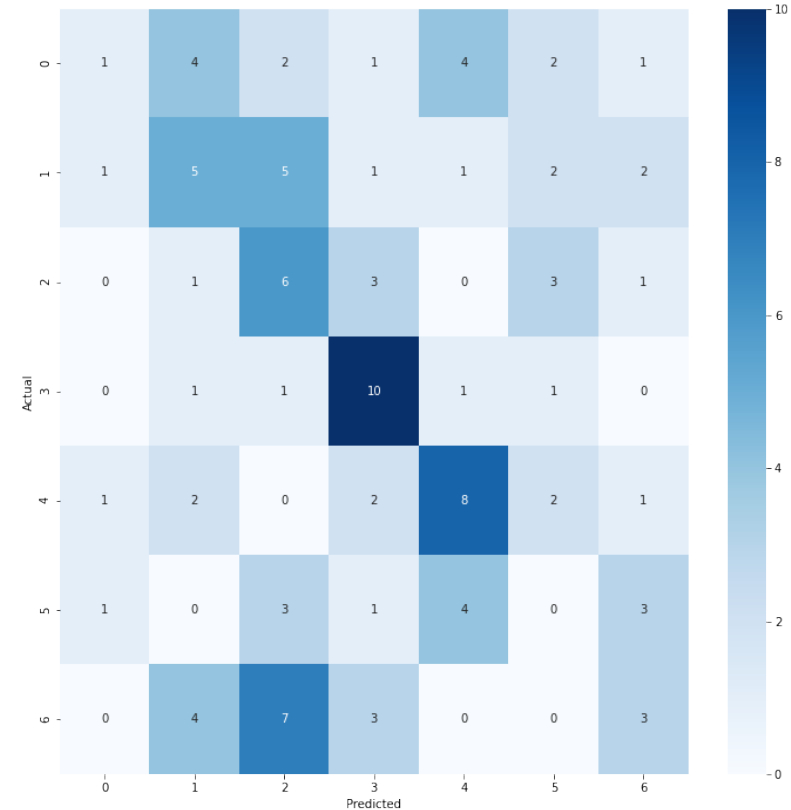
III ~ Classification des Images

Classification:

Linear SVC

- Niveau 1 : 31%
- Niveau 2 : 2%
- Niveau 3 : 5%

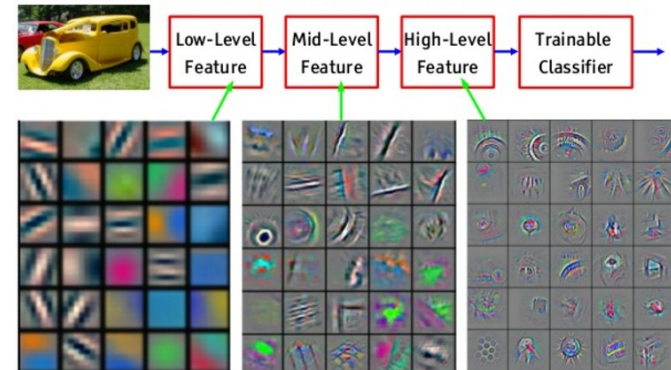
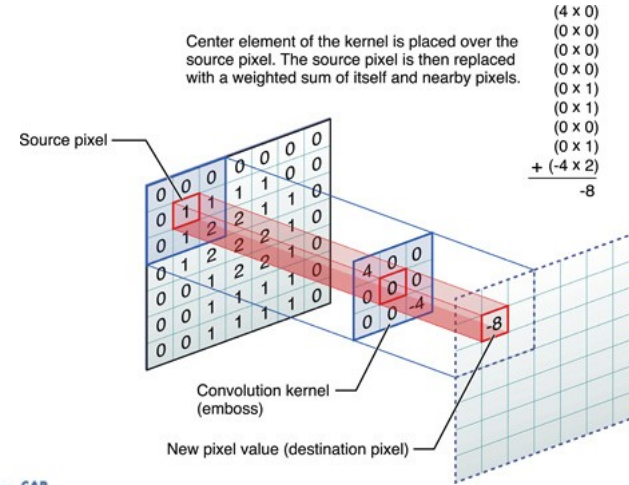
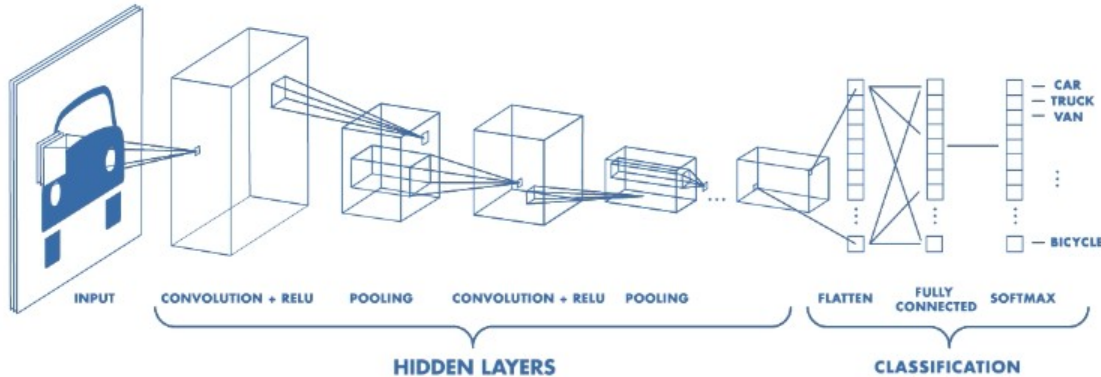
CONFUSION MATRIX - SVC Classifier on the Categories of Level 1



III ~ Classification des Images

Classification:

Convolutional Neural Network (CNN)



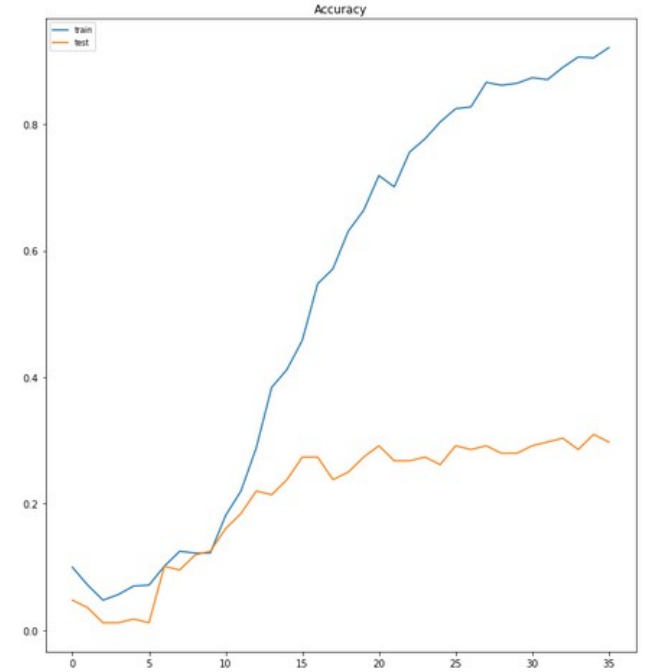
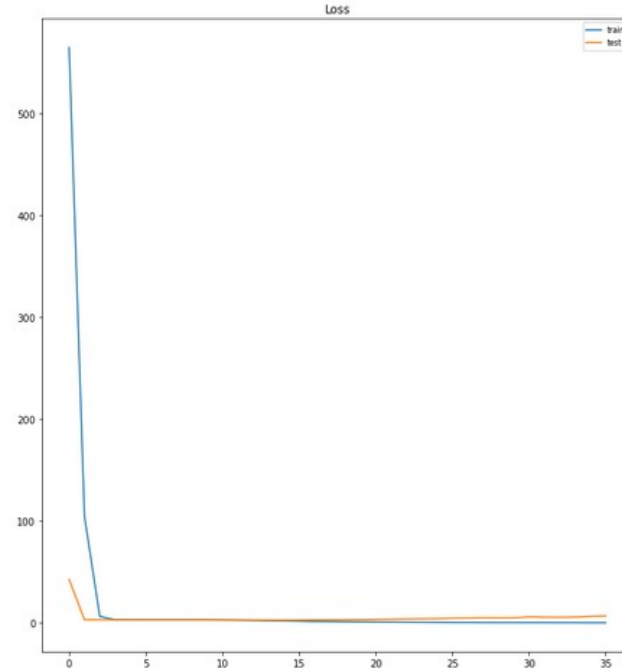
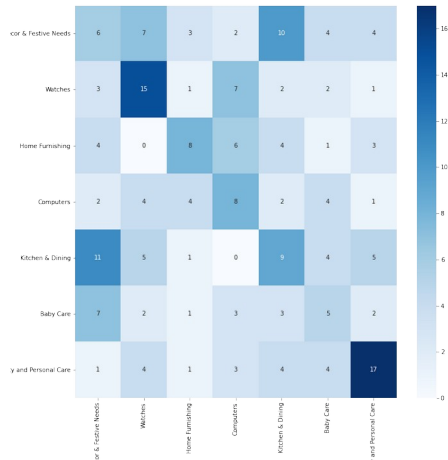
III ~ Classification des Images

Classification:

CNN

- Niveau 1 : 35%
- Niveau 2 : 30%
- Niveau 3 : 39%

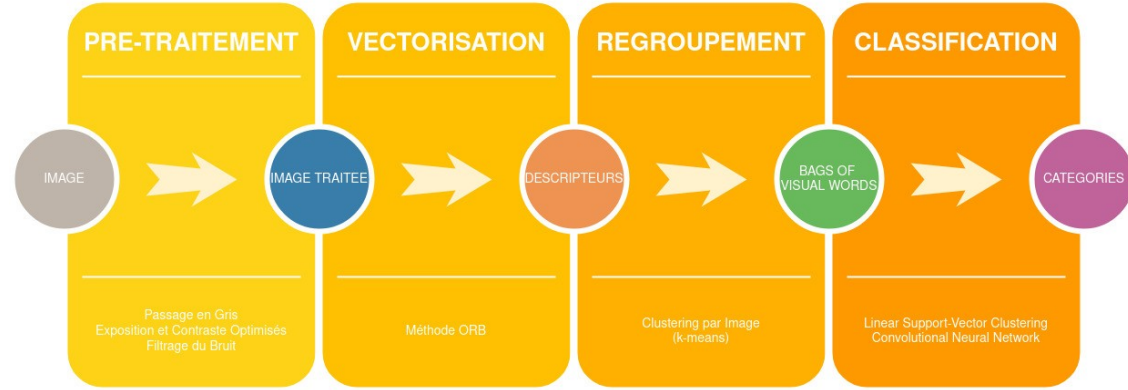
CONFUSION MATRIX - CNN Image Classifier on the Categories of Level 1



III ~ Classification des Images

Classification :

		Accuracy	Time
Level of Category	Model		
Level 1	Linear SVC	31.428571	2.942513
	CNN	34.523809	190.002122
Level 2	Linear SVC	1.904762	4.112777
	CNN	29.761904	207.038256
Level 3	Linear SVC	4.761905	3.729364
	CNN	39.285713	197.103601



Conclusion et Perspectives

Conclusion :

- Traitement de textes
- Traitement d'images
- Classifieur :
 - Texte → possible en l'état
 - Image → non possible en l'état

		Accuracy	Time
Level of Category	Model		
Level 1	Multinomial Naive Bayes	92.380952	0.023243
	Random Forest	92.380952	0.520511
	RNN on Continuous Bag of Word	80.000001	13.767050
	RNN on Skip Gram	92.631578	17.020003
Level 2	Multinomial Naive Bayes	74.285714	0.011050
	Random Forest	86.666667	0.568172
	RNN on Continuous Bag of Word	70.526314	29.507907
	RNN on Skip Gram	84.210527	28.094129
Level 3	Multinomial Naive Bayes	66.666667	0.022552
	Random Forest	82.857143	0.608572
	RNN on Continuous Bag of Word	66.315788	41.373782
	RNN on Skip Gram	80.000001	29.876024

		Accuracy	Time
Level of Category	Model		
Level 1	Linear SVC	31.428571	2.942513
	CNN	34.523809	190.002122
Level 2	Linear SVC	1.904762	4.112777
	CNN	29.761904	207.038256
Level 3	Linear SVC	4.761905	3.729364
	CNN	39.285713	197.103601

Conclusion et Perspectives

Axes de poursuite

- Autres classifieurs
- Optimisation (hyperparamètres, “Transfert Learning”, etc)
- GPU plutôt que CPU
- Plus de données ou “data augmentation”
- Mélange de classifieurs.

**Merci de votre
attention**

Recurrent Neural Network

