Classez automatiquement des biens de consommation

PLAN

1/ Présentation de la problématique
2/ Présentation du jeu de données
3/ Exploration et préparation du jeu de données
4/ Traitement automatique du langage naturel (texte)
5/ Vision par ordinateur (image)
6/ Conclusion
7/ Suite du projet

Data scientist chez **“Place de marché”**

Lancer un marketplace e-commerce

Objectifs :

- Etudier faisabilité d’un **moteur de classification** des articles, à partir de la description et de l’image

- Fichier CSV renseignant sur les articles :
  Prix, nom, lien internet, description, catégorie et sous catégories etc.

- Fichier de photos correspondant aux articles

- Valeurs manquantes des colonnes cibles
- Dupliqués colonnes *nom de produit* et *description*
- Extraire la catégorie principale et sous catégories :
  Garder la catégorie principale (7 catégories)
- Encodage de la catégorie principale
- Equilibre des catégories

- Concaténer le nom produit et description

- Créer des fonctions de prétraitement:
  - *Antonymes*
  - *Tokenization*
  - *Alphanumériques*
  - *Stop words (anglais)*
  - *Mise en minuscule*
  - *Lemmatisation*

- Créer des fonctions de prétraitement pour les différentes méthodes (deep learning)

- Créer des fonctions de visualisation

- Comptage de mots

```
Nombre de mots : 5037

15 mots les plus utilisés : ['r', 'product', 'for', 'only', 'cm', 'free', 'buy', 'replacement', 'delivery', 'genuine', 'shippin
g', 'cash', 'price', 'day', 'mug']

15 mots les moins utilisés : ['clinic', 'apron', 'uniform', 'vastu', 'aura', 'negativity', 'disintegrates', 'disintegrating',
'main', 'rotate', 'clockwise', 'amplifies', 'intention', 'played', 'deeper']

Compte de mots uniques : 1123
```

- **Méthodes**

  Bag of words : comptage/fréquence de chaque mot
  Vocabulaire de 4817 mots

  TF-IDF : comptage du mot dans document * nb documents contenant le mot
  Vocabulaire de 4817 mots

- **Méthodes**

Word2Vec : créer un espace vectoriel des mots avec un forme de similarité
Vocabulaire de 4734 mots

Doc2Vec : créer un espace vectoriel des mots avec un forme de similarité selon le contexte des documents
Vocabulaire de 5037 mots

- **Méthodes**

USE : réseau de neurones entrainé sur des millions de textes, renvoi vecteurs de taille 512, traite texte brut

BERT : lit les documents dans les deux sens et masque des mots pour les prédire selon le contexte
réseau de neurones renvoie vecteurs de taille 768
méthode innovante et populaire

Mesure de faisabilité

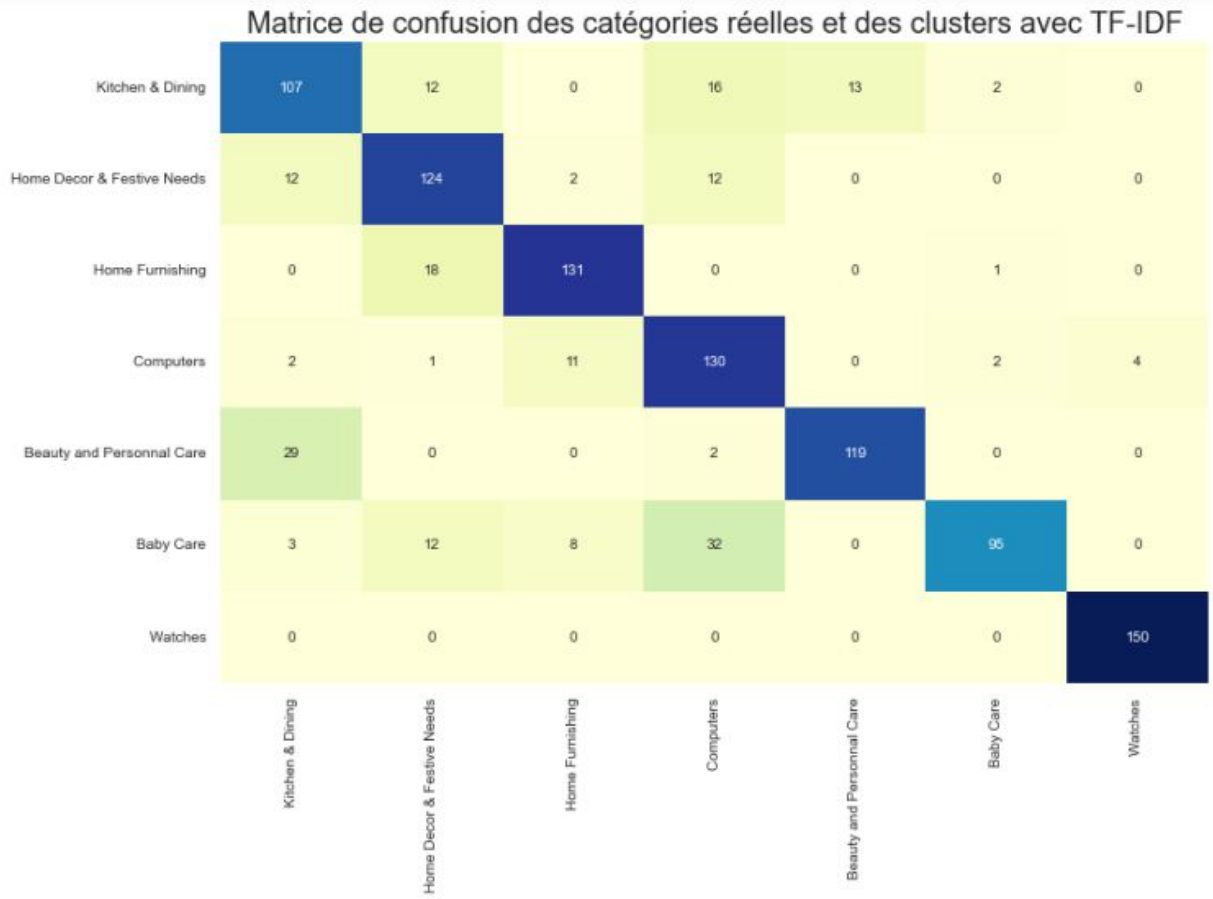
- Visualisation TSNE (2 dimensions)
- Clustering KMeans (7 clusters)
- Calcul score ARI

Bag of words : 0.4622
TF-IDF : 0.6326
Word2Vec : 0.4329
Doc2Vec : 0.454
BERT : 0.3933
USE : 0.4822


Représentation des produits par catégories réelles
Catégorie
0
1
2
3
4
5
6
40
20
0
−20
−40
−60
−40
−20
0
20
40
60


Représentation des produits par clusters
Clusters
0
1
2
3
4
5
6
40
20
0
−20
−40
−60
−40
−20
0
20
40
60

ARI : 0.6326


Matrice de confusion des catégories réelles et des clusters avec TF-IDF
Kitchen & Dining
Home Decor & Festive Needs
Home Furnishing
Computers
Beauty and Personnal Care
Baby Care
Watches
107 12 0 16 13 2 0
12 124 2 12 0 0 0
0 18 131 0 0 1 0
2 1 11 130 0 2 4
29 0 0 2 119 0 0
3 12 8 32 0 95 0
0 0 0 0 0 0 150
Kitchen & Dining
Home Decor & Festive Needs
Home Furnishing
Computers
Beauty and Personnal Care
Baby Care
Watches

- Récupérer les images pour les associer à leur catégorie

- Méthodes
    SIFT
    CNN (VGG16)

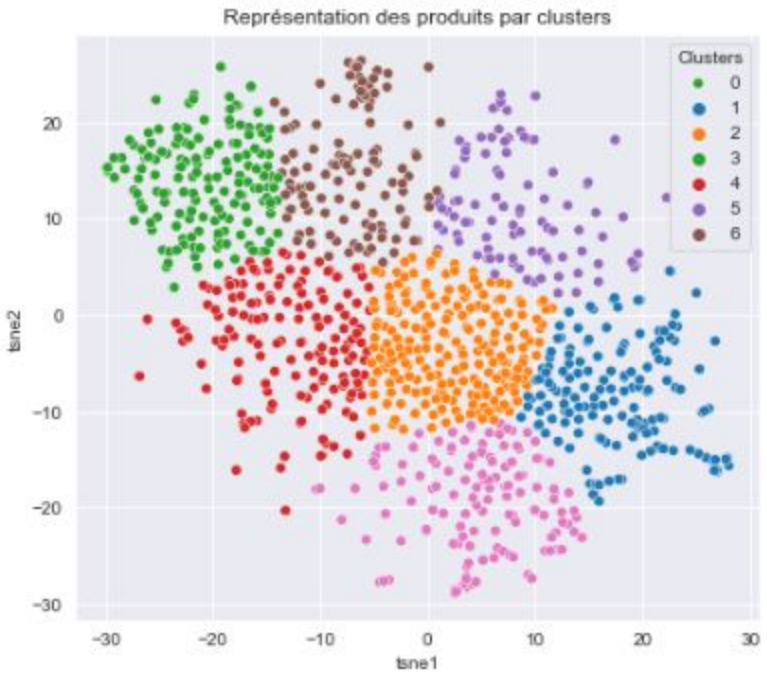
- **SIFT**

  Charger l'image
  Passer en niveaux de gris
  Redimensionner l'image (20%)
  Egaliser l'histogramme (ajuster le contraste)
  Identifier et extraire les descripteurs (277 304 descripteurs)

  Bag of Visual Words
  Créer des clusters de descripteurs (256 clusters)
  Construire un histogramme (vecteur) à partir des clusters de descripteurs

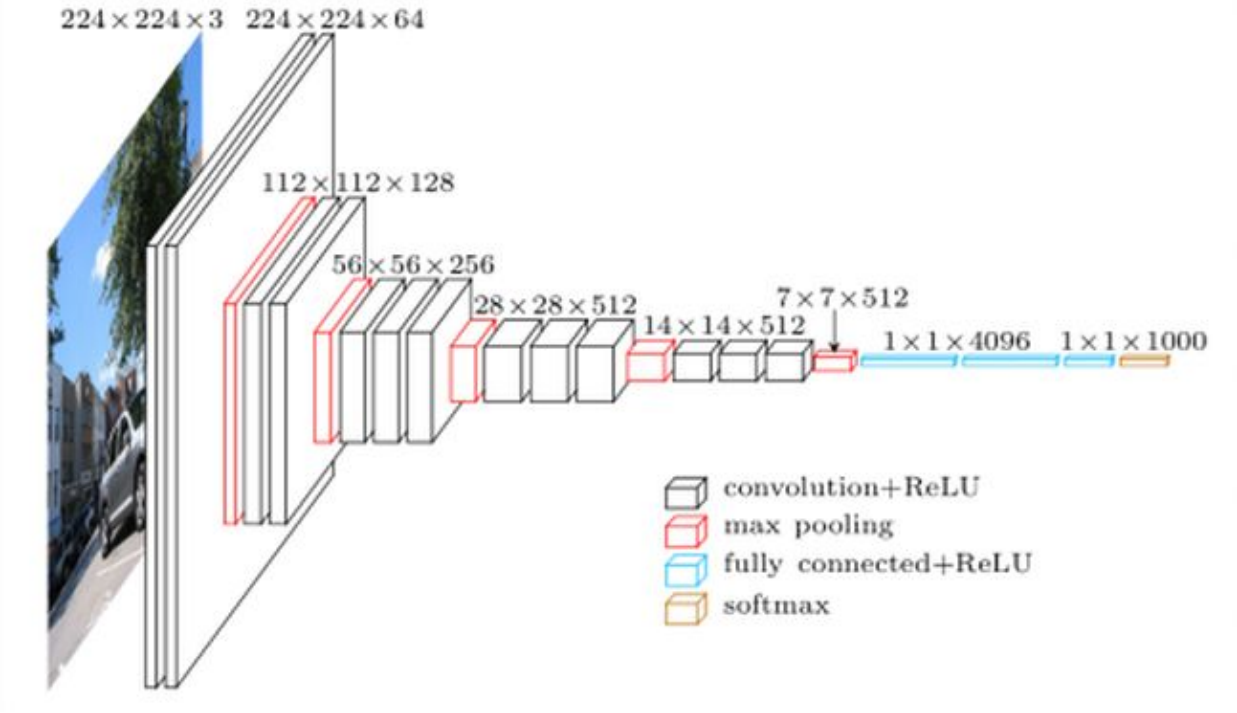
  Estimation de la faisabilité
  Réduction de dimensions (PCA)
  Visualisation TSNE (2 dimensions)
  Clustering KMeans
  Calcul score ARI


Représentation des produits par catégories réelles
tsne2
tsne1
Catégorie
0
1
2
3
4
5
6


Représentation des produits par clusters
tsne2
tsne1
Clusters
0
1
2
3
4
5
6

ARI : 0.0014

- CNN (VGG16)


224 × 224 × 3
224 × 224 × 64
112 × 112 × 128
56 × 56 × 256
28 × 28 × 512
14 × 14 × 512
7 × 7 × 512
1 × 1 × 4096
1 × 1 × 1000
convolution+ReLU
max pooling
fully connected+ReLU
softmax

- **CNN (VGG16)**

  Charger le modèle pré-entraîné
  Supprimer la couche de classification (de sortie)
  Pas procéder à un entraînement

  Prédire chaque image à partir du modèle pour récupérer les features

  Estimation de la faisabilité
  Visualisation TSNE (2 dimensions)
  Clustering KMeans
  Calcul score ARI


Représentation des produits par catégories réelles
Catégorie
0
1
2
3
4
5
6
80
60
40
20
0
−20
−40
−60
−40
−20
0
20
40
60


Représentation des produits par clusters
Clusters
0
1
2
3
4
5
6
80
60
40
20
0
−20
−40
−60
−40
−20
0
20
40
60

ARI : 0.4237

Le moteur de classification semble faisable avec la description des produits

Bonus :

- Entrainement d'un classifieur (GradientBoostingClassifier)
- Score de précision de 0.76

- Présentation des résultats et en discuter
- Trouver le meilleur classifieur ainsi que ses meilleurs paramètres
- Proposer une mise en production d'une API