OPENCLASSROOMS CentraleSupélec

Classifiez automatiquement des biens de consommation

Formation Data Scientist - Janvier / Novembre 2022

Auteur: Eric TREGOAT

Mentor: Benjamin TARDY

Evaluatrice: Julide YILMAZ

Ordre du jour de la soutenance

❑ Présentation

- Rappel de la problématique et présentation du jeu de données
- Explication des prétraitements et des résultats du clustering
- Conclusion sur la faisabilité du moteur de classification et recommandations pour sa création éventuelle

❑ Discussion

❑ Débriefing

Problématique et démarche

❑ Problématique

- **Objectif**: examiner la faisabilité de classifier de manière non supervisée la liste des produits d'un site de e-commerce à partir de leur description sous forme de texte ou d'image

- **Données de base** : 1 fichier de données relatifs aux produits du site et les fichiers image des produits

❑ Démarche

1. Prise de connaissance des données, arborescence produit
2. Fonctions support d'évaluation et visualisation du clustering
3. Classification par natural language processing (NLP) à partir des descriptions textuelles: preprocessing, Bag of Words, sujets clés, word embedding, word/sentence embedding
4. Classification par computer vision (CV) à partir des images des produits: SIFT et CNN
5. Classification combinée NLP et CV
6. Conclusion

Prise de connaissance des données, arborescence produit

- **Fichiers de données**
  - → Echantillon de 1050 produits décrits par 15 caractéristiques
  - → Nom du produit
  - → Arborescence produit
  - → Lien vers le fichier image du produit
  - → Description textuelle du produit
- **Arborescence produit**
  - → Profondeur max: 6 niveaux
  - → Tous les produits ont au moins 2 niveaux
  - → Nombre de classes vues selon la profondeur d'arborescence:
    7 » 62 » 243 » 460 » 596 » 633
- **Classification vs arborescence**
  - → Classification au niveau 1 pour la faisabilité
  - → La classification des autres niveaux peut s'effectuer en cascade
  - → Evaluation vs arborescence nécessiterait d'augmenter la taille de l'échantillon

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1050 entries, 0 to 1049
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   uniq_id                1050 non-null   object
 1   crawl_timestamp        1050 non-null   object
 2   product_url            1050 non-null   object
 3   product_name           1050 non-null   object
 4   product_category_tree  1050 non-null   object
 5   pid                    1050 non-null   object
 6   retail_price           1049 non-null   float64
 7   discounted_price       1049 non-null   float64
 8   image                  1050 non-null   object
 9   is_FK_Advantage_product 1050 non-null  bool
 10  description            1050 non-null   object
 11  product_rating         1050 non-null   object
 12  overall_rating         1050 non-null   object
 13  brand                  712 non-null    object
 14  product_specifications 1049 non-null   object
dtypes: bool(1), float64(2), object(12)
memory usage: 116.0+ KB
```

Fonctions support d’évaluation et visualisation du clustering

A. Fonction de réduction de dimensionnalité: t-SNE ou PCA

B. Fonction de clustering: k-Means sur les features réduites dimensionnellement

C. Fonction d'évaluation du clustering: ARI

D. Matrice de confusion et ordination des labels prédits: ordination en fonction de la matrice de confusion

E. Fonction de prédiction des labels de cluster (s’appuie sur A, B et C)

F. Fonction de visualisation des clusters: représentation de la population avec t-SNE et coloré selon les catégories de produit

G. Fonction de recherche sur grille: implémente la recherche multi-paramètres

H. Fonction d'affichage graphique du résultat de la recherche sur grille: ARI en fonction de chaque paramètre de la recherche

Classification par NLP - Bilan

- Preprocessing du texte:
  - → Transformation unicode UTF-8
  - → Découpage en phrases et mots
  - → Filtrage de la ponctuation (exceptions possibles)
  - → Transformation en minuscule (exceptions possibles)
  - → Filtrage des stop-words
  - → Filtrage des mots contenant des chiffres
  - → Filtrage optionnel des mots courts
  - → Lemmatisation optionnelle (anglais seulement)
  - → Filtrage optionnel des mots hors dictionnaire
  - → Stemming optionnel

Approche	ARI	Clustering	Recherche
TfidfVectorizer	0.72	TfidfVectorizer - Clustering	TfidfVectorizer - search_graph
LDA	0.59	LDA - Clustering	LDA - search_graph
NMF	0.57	NMF - Clustering	NMF - search_graph
CountVectorizer	0.55	CountVectorizer - Clustering	CountVectorizer - search_graph
USE	0.48	USE - Clustering	USE - search_graph
Word2Vec	0.38	Word2Vec - Clustering	Word2Vec - search_graph
BERT	0.32	BERT - Clustering	BERT - search_graph

Classification par NLP – TF-IDF (1/2)

Classification vs catégories:

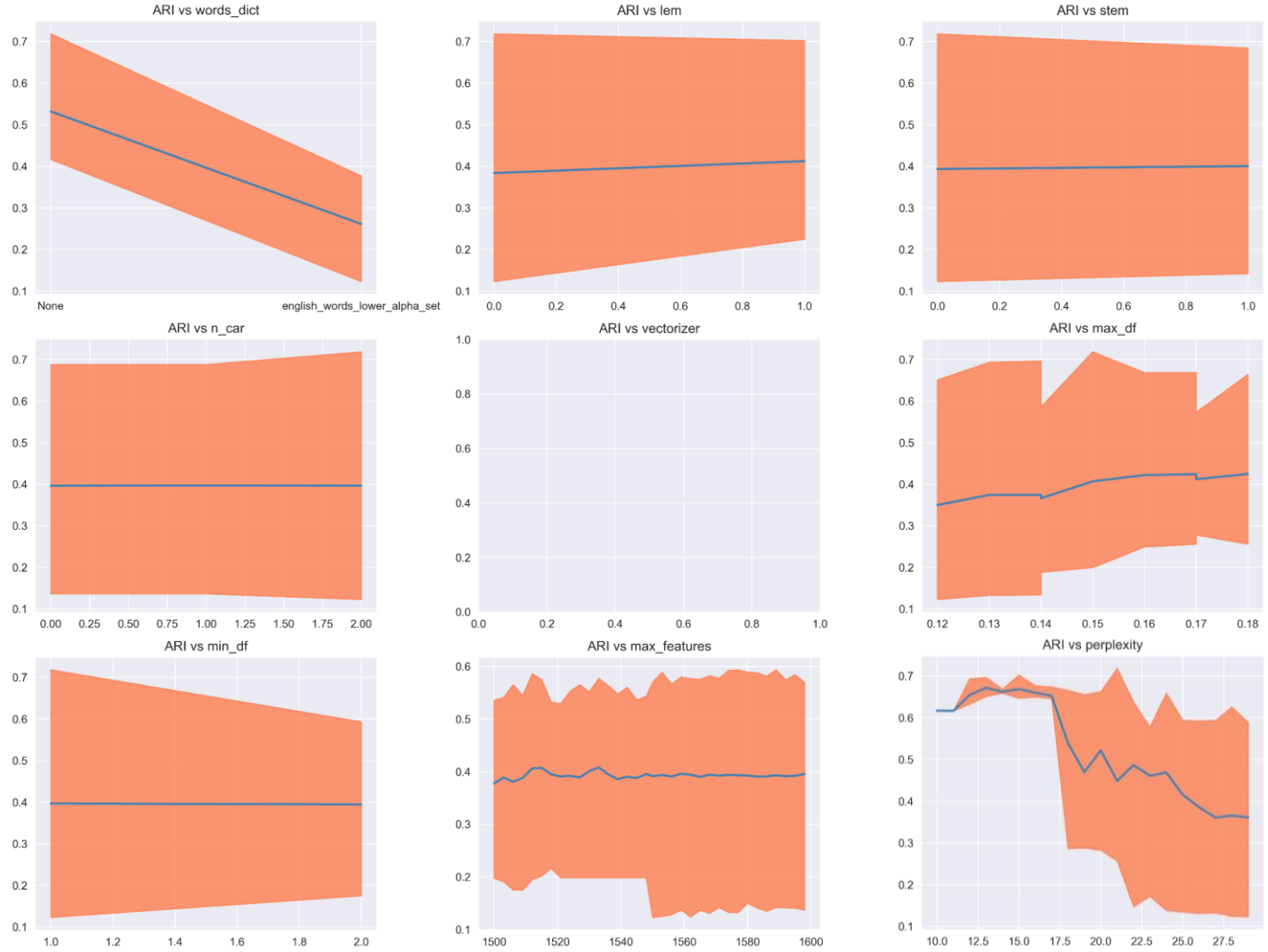
- → ARI = 0,72
- → Correspondance : 83%
- → Catégorie « Watches » la mieux identifiée
- → Catégories « Baby care » et « Kitchen & Dining » les moins bien identifiées

	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
Baby Care	0.733	0.020	0.000	0.113	0.107	0.027	0.000
Beauty and Personal Care	0.000	0.887	0.013	0.080	0.000	0.020	0.000
Computers	0.000	0.000	0.887	0.013	0.007	0.093	0.000
Home Decor & Festive Needs	0.000	0.013	0.087	0.820	0.033	0.033	0.013
Home Furnishing	0.140	0.000	0.000	0.007	0.853	0.000	0.000
Kitchen & Dining	0.007	0.007	0.060	0.160	0.033	0.733	0.000
Watches	0.000	0.000	0.000	0.013	0.000	0.000	0.987

Classification par NLP – TF-IDF (2/2)

Meilleurs paramètres:

- → Sans dictionnaire
- → Sans lemming (faible influence)
- → Sans stemming (faible influence)
- → Filtrage des mots de 2 lettres ou moins
- → Filtrage des mots présents dans plus de 15% des documents
- → Pas de filtrage d'occurrence minimale des mots
- → Pas de seuil de filtrage du nombre maximum de mots (max features)
- → Perplexité optimum pour la réduction de dimensionnalité avec t-SNE: 21



OPENCLASSROOMS
CentraleSupélec

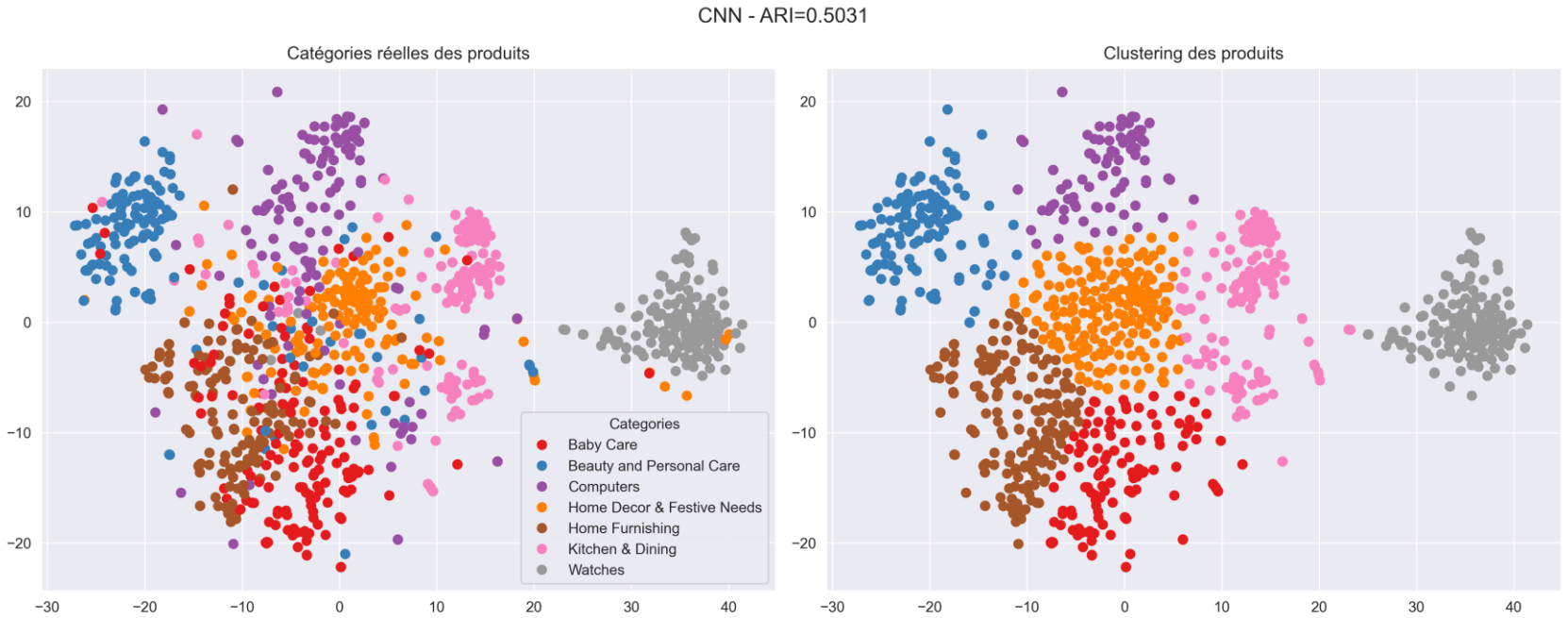
Classification par CV

Approche	Modele	ARI	Clustering	Recherche
CNN	EfficientNetB7	0.50	CNN - Clustering	CNN - search_graph
SIFT		0.05	SIFT - Clustering	SIFT - search_graph

- Processing des images avec SIFT:
  - → Transformation en nuances de gris
  - → Egalisation + débruitage (gaussien + median)
  - → Descripteurs SIFT de chaque image avec OpenCV
  - → Bag of visual words (BoVW) par clustering k-Means
    - → Clustering entre 10*n_categories et sqrt(n_descripteurs)
    - → BoVW = centroïdes
    - → Recherche de maximisation de l'ARI
  - → Vectorisation des images avec le BoVW
  - → Recherche sur grille

- Processing des images avec CNN:
  - → Utilisation de modèles pré-entrainés sur ImageNet pour extraire les features juste avant la couche fully connected
  - → Mise des images à la dimension requise par le modèle
  - → Preprocessing propre à chaque modèle de CNN
  - → Prédiction pour extraire les features + aplatissement
  - → Recherche sur grille pour maximiser l'ARI

Classification par CV – CNN



Classification vs catégories:

- → Meilleur modèle: EfficientNetB7, ARI = 0,5
- → Autres testés: Xception (0,49)
  InceptionResNetV2 (0,37)
  VGG16 (0,35)
  VGG19 (0,34)
- → Correspondance : 72%
- → Catégorie « Watches » la mieux identifiée
- → Catégories « Baby care » et « Computers » les moins bien identifiées

	Baby Care	Beauty and Personal Care	Computers	Home Decor & Festive Needs	Home Furnishing	Kitchen & Dining	Watches
Baby Care	0.560	0.027	0.000	0.133	0.247	0.027	0.007
Beauty and Personal Care	0.027	0.780	0.007	0.087	0.047	0.053	0.000
Computers	0.087	0.027	0.613	0.167	0.047	0.060	0.000
Home Decor & Festive Needs	0.047	0.053	0.000	0.687	0.067	0.113	0.033
Home Furnishing	0.053	0.013	0.020	0.067	0.840	0.007	0.000
Kitchen & Dining	0.053	0.067	0.027	0.120	0.007	0.727	0.000
Watches	0.000	0.000	0.000	0.013	0.007	0.013	0.967

Classification combinée NLP et CV

A. Récupération des features des meilleurs modèles

B. Concaténation des features NLP et CV

→ Redimensionnement (ou pas) des vecteurs pour leur donner la même taille

→ Affectation d'un poids relatif: NLP (weight) vs CV (1-weight)

→ Normalisation (ou pas du vecteur résultant)

Non probant
ARI < 0,5
pour weight < 1

C. Combinaison des probabilités de label des modèles

→ Modèles GMM NLP (ARI=0,65) et CV (ARI=0,49)
pour prédire les probabilités d'appartenance aux clusters

→ Combinaison des probabilités:

→ **Max**

→ Max (Σ prob par cluster)

→ Max (Π prob par cluster)

Probant (ARI=0,6),
mais n'améliore pas le
meilleur clustering

Conclusion

A. La classification non supervisée de la liste des produits à partir de leur description sous forme de texte ou d'image est faisable avec des indices de similarité respectifs de 0,72 et 0,5.

   Nota: les dissimilarités peuvent provenir soit de l'algorithme de clustering, en fonction de la qualité des descriptions et des images, soit des erreurs de catégorisation des opérateurs

B. A partir des descriptifs textuels, l'algorithme de classification le plus performant est le plus simple (Bag of Words).

   Nota: classification par mots clés

C. A partir des images, la classification la plus performante est basée sur les réseaux de neurones convolutifs pré-entrainés, très complexes de conception et entrainement, mais très simples à utiliser pour ce type d'application.

   Nota: le choix du CNN pré-entrainé s'effectue sur la base de sa performance sur un jeu de données similaire ou généraliste (ex: ImageNet).

Echanges avec l’évaluateur

- **Discussion**
- **Débriefing**

OPENCLASSROOMS CentraleSupélec

Contact: Eric TREGOAT

eric.tregoat@gmail.com

06 49 99 79 59

https://www.linkedin.com/in/erictregoat/