Classifiez automatiquement des biens de consommation

Ilaria Mereu

Soutenance Projet 6

Parcours Data Scientist

Mission

Étudier la faisabilité d'un moteur de classification des articles en vente en différentes catégories

Entreprise qui souhaite lancer une marketplace e-commerce.

Objectifs:

Sur la base d'une image et une description, notre objectif est automatiser l'attribution de la catégorie de l'article.

Demarche indiquée:

- **prétraitement des descriptions des produits et des images,**
- **une réduction de dimension,**
- **puis un clustering.**

Sommaire

1. Description de la base des données d'origine
2. Modèles explorés pour la partie texte et résultat de la recherche
    - Bag of words (CountVectorizer, Td-idf)
    - Word embedding (Word2Vec, BERT base uncased, BERT hub Tensorflow, USE)
3. Modèles explorés pour la partie images et résultat de la recherche
    1. SIFT
    2. Reseaux de neurones convolutifs (Mobilenet et convolutif)
4. Conclusions et perspectives

I - Description du jeu de données d'origine

Description du jeu de données

1) 1 fichier contenant 1050 lignes et 15 colonnes

2) Un dossier de 1051 images

À chaque ligne du fichier correspond un article en vente et ses caractéristiques, en particulier:

- Une classification de catégorie de l'objet exécutée manuellement ayant moins trois niveaux, le dernier étant le nom de l'article. Exemple:
  "Baby Care >> Baby Bath & Skin >> Baby Bath Towels >> Sathiyas Baby Bath Towels >> Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Y…"

- Une description:
  "Specifications of Sathiyas Cotton Bath Towel (3 Bath Towel, Red, Yellow, Blue) Bath Towel Features Machine Washable Yes Material Cotton Design Self Design General Brand Sathiyas Type Bath Towel GSM 500 Model Name Sathiyas cotton bath towel Ideal For Men, Women, Boys, Girls Model ID asvtwl322 Color Red, Yellow, Blue Size Mediam Dimensions Length 30 inch Width 60 inch In the Box Number of Contents in Sales Package 3 Sales Package 3 Bath Towel"

- Le nom du fichier de l'image contenu dans le dossier

Catégories

1) Les catégories de premier niveau (N1) sont 7, exactement 150 articles par catégorie : 'Baby Care ', 'Beauty and Personal Care ', 'Computers ', 'Home Decor & Festive Needs ', 'Home Furnishing ', 'Kitchen & Dining ', 'Watches '

2) Les catégories de deuxième niveau (N2) sont 61: 'Baby & Kids Gifts ', 'Baby Bath & Skin ', 'Baby Bedding ', 'Baby Grooming ', 'Bar & Glassware ', 'Bath Linen ', 'Bath and Spa', 'Beauty Accessories ', 'Bed Linen ', 'Body and Skin Care ', 'Candles & Fragrances ', 'Clocks ', 'Coffee Mugs', 'Combos and Kits ', 'Computer Components ', 'Computer Peripherals ', 'Consumables & Disposables ', 'Containers & Bottles ', 'Cookware ', 'Curtains & Accessories ', 'Cushions, Pillows & Covers ', 'Decorative Lighting & Lamps ', 'Diapering & Potty Training ', 'Dinnerware & Crockery ', 'Eye Care ', 'Feeding & Nursing ', 'Floor Coverings ', 'Flowers, Plants & Vases ', 'Fragrances', 'Furniture & Furnishings ', 'Garden & Leisure ', 'Hair Care ', 'Health Care ', 'Housekeeping & Laundry ', 'Infant Wear ', 'JMD Home Furnishing ', 'Kitchen & Dining Linen', 'Kitchen Tools ', 'Kripa's Home Furnishing ', 'Laptop Accessories ', 'Laptops ', 'Lighting ', 'Living ', 'Living Room Furnishing ', 'Makeup ', 'Men's Grooming ', 'Network Components ', 'Pressure Cookers & Pans ', 'Religion & Devotion ', 'Showpiece ', 'Showpieces ', 'Software ', 'Storage ', 'Strollers & Activity Gear ', 'Table Decor & Handicrafts ', 'Tablet Accessories ', 'Tableware & Cutlery ', 'Tidy Home Furnishing ', 'Wall Decor & Clocks ', 'Women's Hygiene ', 'Wrist Watches '.

Pouvons-nous trouver une segmentation des articles proche des catégories N1 à partir de leur description par le biais d'une analyse NLP?

Modèles explorés pour la partie texte

Protocole

1) Prétraitement

    1) "bag-of-words", comptage simple de mots (CountVectorizer)

    2) "bag-of-words", Tf-idf (TfidfVectorizer)

    3) une approche de type word/sentence embedding classique avec Word2Vec (ou Glove ou FastText) ;

2) Réduction de dimensionalité: t-SNE

3) clustering: k-means

Réseaux de neurones:

- BERT
- USE

Modèles explorés (1 - bag of words)

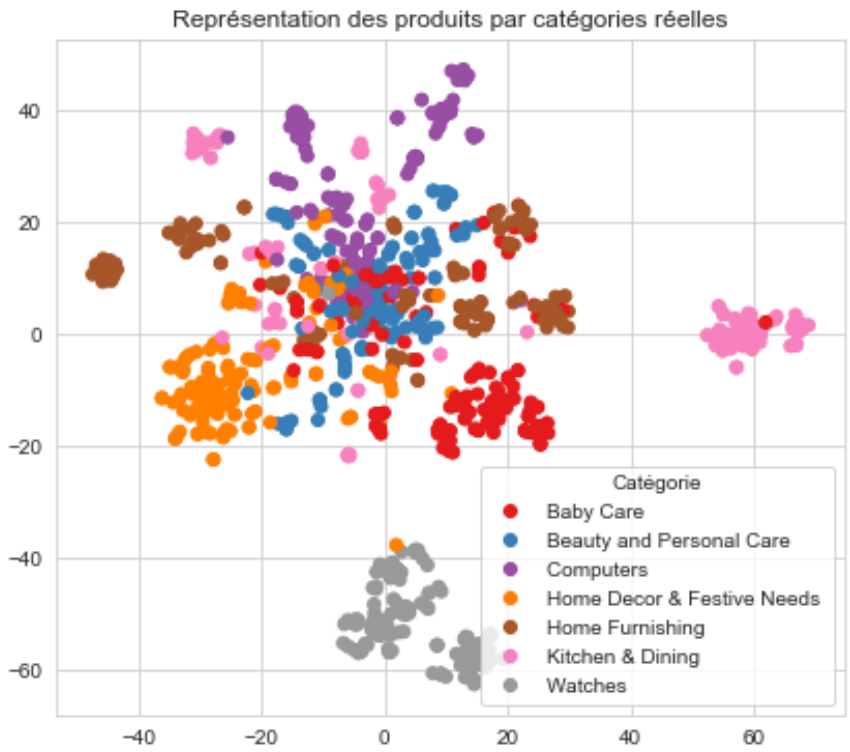
Méthode de représentation	Méthode BOW	Corpus	Méthode extraction	
Bag of words	CountVectorizer	Description	PCA + t-SNE + k-means	Modèle 1
		Nom du produit		Modèle 2
		Description en combinaison avec le nom du produit		Modèle 3
	Td-idf	Description		Modèle 4
		Nom du produit		Modèle 5
		Description en combinaison avec le nom du produit		Modèle 6

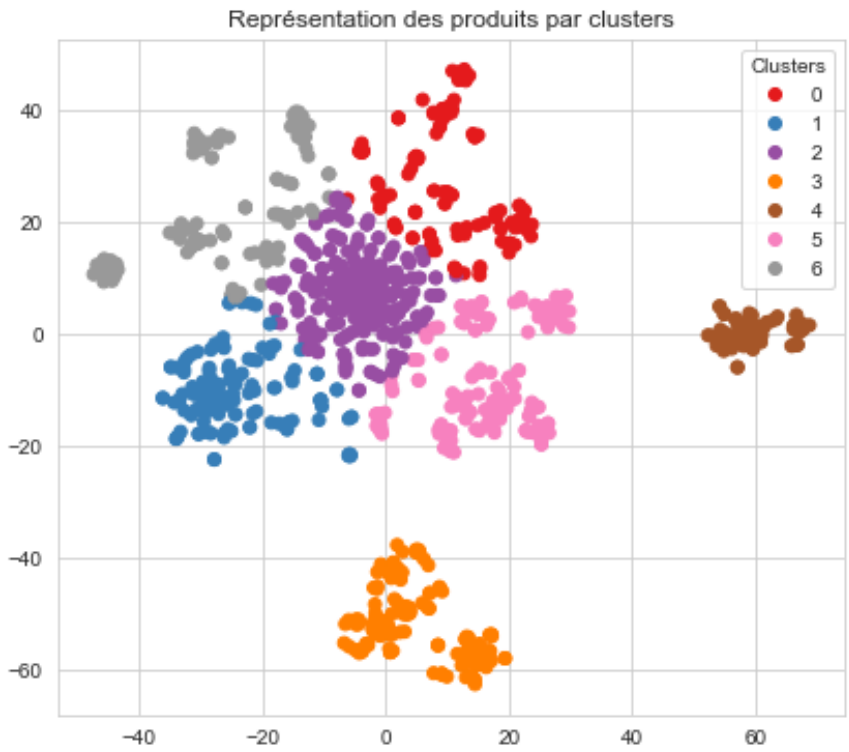
Prétraitement

Réduction de dimensions

Clustering

Modèles explorés (1 - bag of words) - exemple

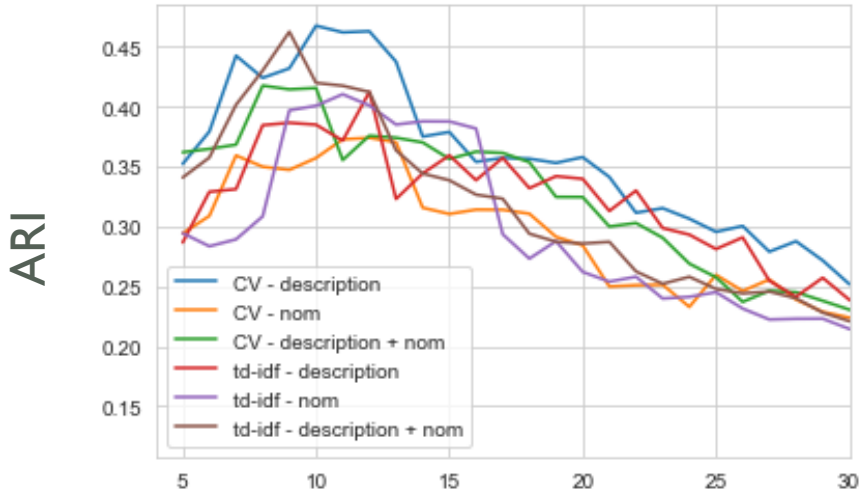




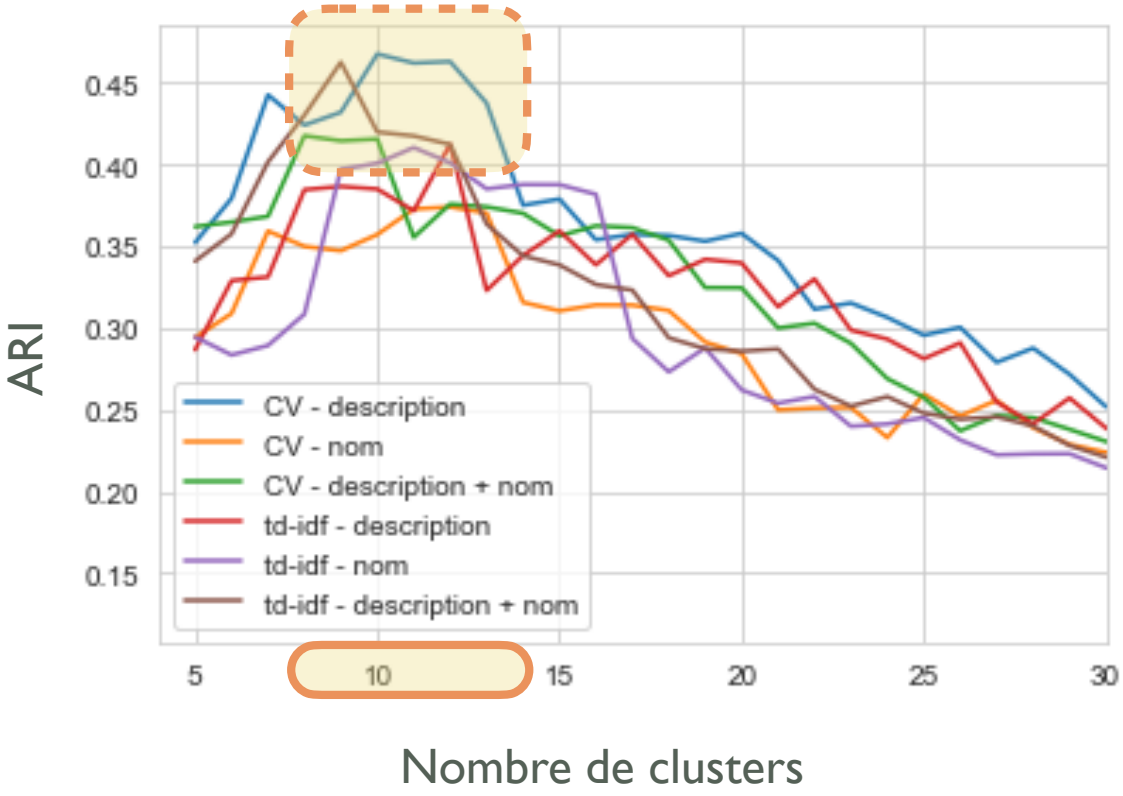
CountVectorizer; n_clusters: 7; ARI : 0.3684; time : 27.0 s

Modèles explorés (1 - bag of words)

Méthode de représentation	Méthode BOW	Corpus	Méthode extraction	
Bag of words	CountVectorizer	Description	**PCA + t-SNE + k-means**	Modèle 1
		Nom du produit		Modèle 2
		Description en combinaison avec le nom du produit		Modèle 3
	Td-idf	Description		Modèle 4
		Nom du produit		Modèle 5
		Description en combinaison avec le nom du produit		Modèle 6



Modèles explorés (1 - bag of words)



Le nombre de catégories N1 (7) ne semble pas aligné aux meilleures performances des modèles (le choix des categories n'est pas univoque)

Modèles explorés (2 - word embedding) - ARI pour 7 clusters

	Méthode de représentation			
Corpus	Word2Vec	BERT base uncased	BERT hub Tensorflow	**USE**
Description	0.2084	0.3265	0.3156	**0.4309**
Nom du produit	0.5236	0.6483	0.6433	**0.6543**
Description en combinaison avec le nom du produit	0.4376	0.3904	0.3909	**0.6062**

Modèles explorés pour la partie image

Protocole SIFT

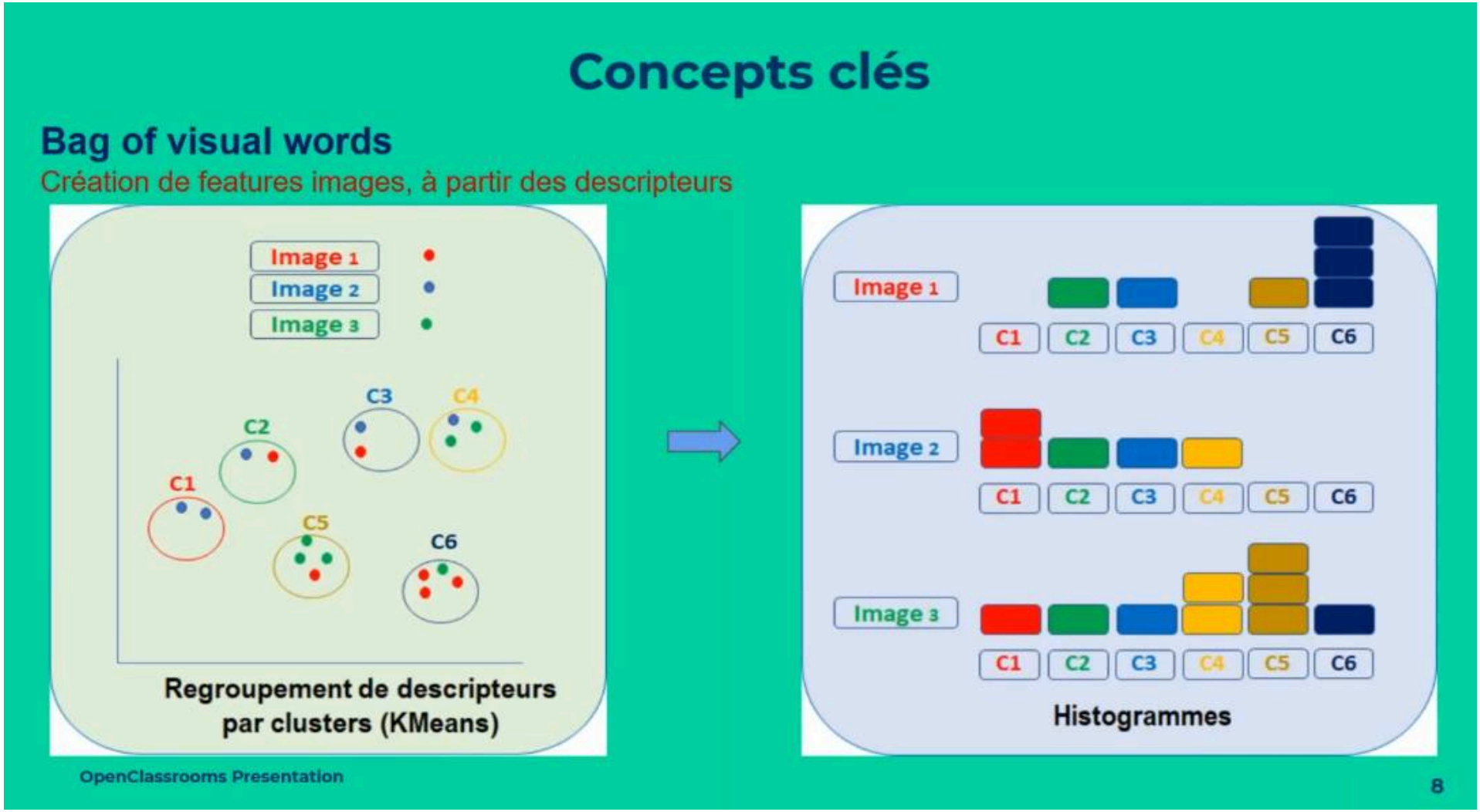
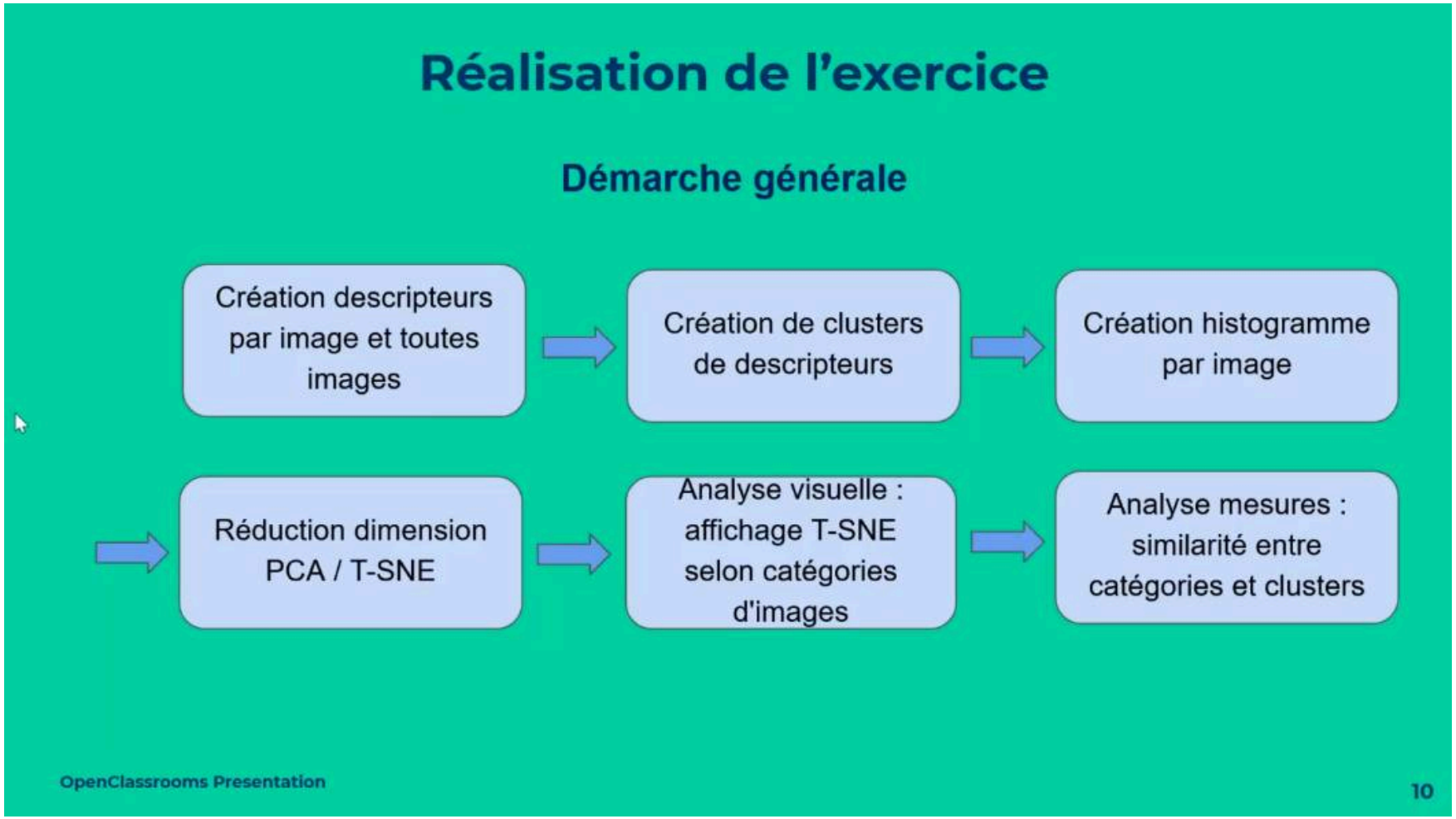


Image réalisée par M. Olivier Chotin

Protocole SIFT



Résultat ARI: 0.049

Image réalisée par M. Olivier Chotin

Protocole CNN

L'approche des réseaux de neurones convolutifs (CNN) est significativement plus puissant de la méthode SIFT. L'extraction des features est elle même automatisée et optimisée par rétropropagation.

1) Prétraitement des images

2) Construction du CNN - la réduction de dimensionalité et le clustering en 7 catégories sont intégrés dans l'architecture du CNN - respectivement par les couches de pooling et le nombre de neurones de la dernière couche.

3) Choix d'un CNN pré-entraîné.

4) Transfer Learning: le CNN est modifié afin d'être entrainé sur une partie de nos données

5) Entraînement

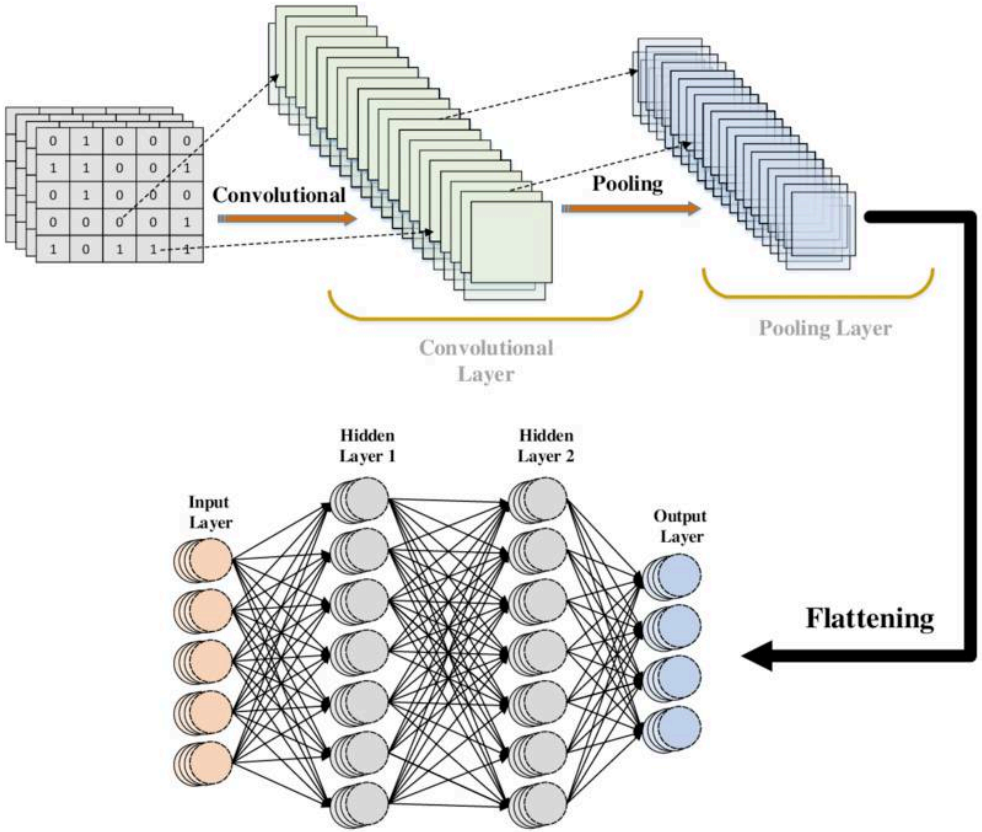
6) Résultats

Protocole CNN

L'approche des réseaux de neurones convolutifs (CNN) est significativement plus puissant de la méthode SIFT. L'extraction des features est elle même automatisée et optimisée par rétropropagation.

1) Prétraitement des images

2) Construction du CNN - la réduction de dimensionalité et le clustering en 7 catégories sont intégrés dans l'architecture du CNN - respectivement par les couches de pooling et le nombre de neurones de la dernière couche.



Protocole CNN - résultats

ARI pour le CNN Mobilenet: 0.6099

ARI pour le CNN VGG16: 0.5959

Conclusions

Meilleures performances

Pour la partie texte:

	ARI
Corpus	**Méthode: USE**
Description	**0.4309**
Nom du produit	**0.6543**
Description en combinaison avec le nom du produit	**0.6062**

Pour la partie image:

	ARI	
CNN Mobilenet	**0.6099**	

Performance améliorable

Perspectives

Compte tenu des catégories données:

- l'autonomie de ce système n'est pas complète
- progrès par rapport à l'affectation purement manuelle

Point d'amélioration:

Ambiguïté de certaines catégories:

- 'Baby Care ', 'Beauty and Personal Care ',
- 'Home Decor & Festive Needs ', 'Home Furnishing ', 'Kitchen & Dining '

  Plusieurs catégories assignées automatiquement?

Points d'amélioration - tenir compte des ambiguïtés

Certains produits peuvent appartenir `plusieurs catégories sans être limité à une seule. Exemple:

- 'Baby Care ', 'Beauty and Personal Care ',
- 'Home Decor & Festive Needs ', 'Home Furnishing ', 'Kitchen & Dining'

1. Proposition: Les descriptions semblent apporter plus de bruit que de clarté. Utilisation du nom du produit.
2. Affectation insuffisante dans ces catégories:
    1. Sol. 1
        1. plusieurs catégories assignées automatiquement?
        2. le moteur assigne plusieurs parcours qui peuvent tenir compte des ambiguïtés
        3. Si les catégories sont incompatible renvoi à la vérification humaine;.
    2. Sol 2. : si le parcours doit être unique:
        1. le moteur assigne plusieurs étiquettes qui peuvent tenir compte des ambiguïtés
        2. parcours assigné sur la base de la catégorie la plus probable.
        3. si les catégories sont incompatible renvoi à la vérification humaine;.

Merci

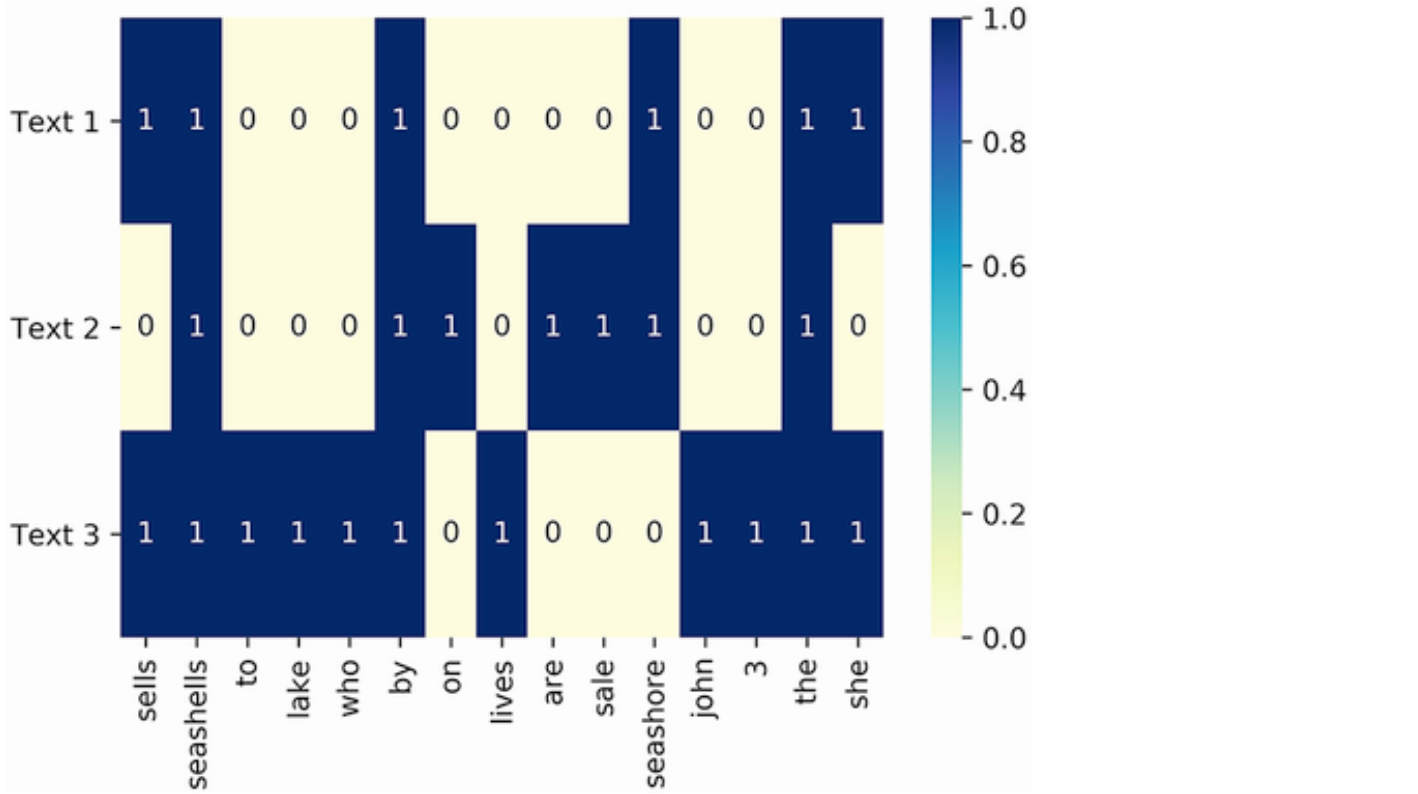
LDA - td-idf - description - modèle 4

```
Topic 0: baby mug girl fabric coffee dress cotton sleeve
neck boy


Topic 1: pizza help cranberry cutter bluetooth agromech
drive slice overwhelms wheel


Topic 2: com flipkart free shipping cash genuine
delivery buy product watch


Topic 3: smart 08 1042 pioneer quits dandruff infused
richness wash intenso


Topic 4: showpiece replacement guarantee day usb genuine
cash shipping delivery product


Topic 5: hair conditioner quilt color display white mat
dw100243 yarn boy


Topic 6: rockmantra mug ceramic dishwasher creation
making permanent porcelain thrilling crafting
```

- Baby Care
- Beauty and Personal Care
- Home Decor & Festive Needs
- Computer
- Kitchen & Dining
- Watches

Bag of words

```
1 text1 = 'She sells seashells by the seashore.'
2 text2 = '"Seashells! The seashells are on sale! By the seashore.
3 text3 = 'She sells 3 seashells to John, who lives by the lake.'
```



[Data Science Bookcamp](#)
[Leonard Apeltsin](#)

Word/sentence embedding

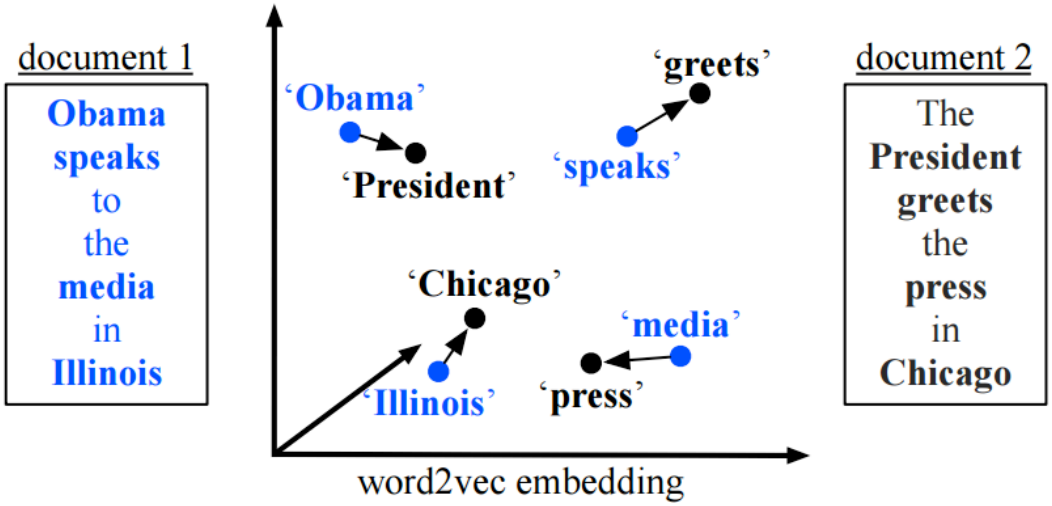


Figure 1. An illustration of the *word mover's distance*. All non-stop words (***bold***) of both documents are embedded into a *word2vec* space. The distance between the two documents is the minimum cumulative distance that all words in document 1 need to travel to exactly match document 2. (Best viewed in color.)

"From Word Embeddings To Document Distances" published in EMNLP'14.