Classifiez automatiquement des biens de consommation

Soutenance P6 OpenClassrooms - Orkun Selçuk - 2023

Classifiez automatiquement des biens de consommation

SOMMAIRE

1. Projet
2. Nettoyage
3. Bag of words
4. WORD2VEC
5. BERT
6. UNIVERSAL SENTENCE ENCODER
7. ORB
8. VGG16

Projet

Vous êtes Data Scientist au sein de l’entreprise "Place de marché”, qui souhaite lancer une marketplace e-commerce. Sur la place de marché, des vendeurs proposent des articles à des acheteurs en postant une photo et une description.

Mission

- Ma mission est de réaliser une première étude de faisabilité d'un moteur de classification d'articles, basé sur une image et une description, pour l'automatisation de l'attribution de la catégorie de l'article.
- Je dois analyser le jeu de données en réalisant un prétraitement des descriptions des produits et des images, une réduction de dimension, puis un clustering

Dataset

Il y a 1050 lignes et 15 colonnes

	uniq_id	crawl_timestamp	product_url	product_name	product_category_tree	pid
0	55b85ea15a1536d46b7190ad6fff8ce7	2016-04-30 03:22:56 +0000	http://www.flipkart.com/elegance-polyester-mul...	Elegance Polyester Multicolor Abstract Eyelet ...	["Home Furnishing >> Curtains & Accessories >>...	CRNEG7BKMFFYHQ8Z
1	7b72c92c2f6c40268628ec5f14c6d590	2016-04-30 03:22:56 +0000	http://www.flipkart.com/sathiyas-cotton-bath-t...	Sathiyas Cotton Bath Towel	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEGFZHGBXPHZUH
2	64d5d4a258243731dc7bbb1eef49ad74	2016-04-30 03:22:56 +0000	http://www.flipkart.com/eurospa-cotton-terry-f...	Eurospa Cotton Terry Face Towel Set	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEG6SHXTDB2A2Y
3	d4684dcdc759dd9cdf41504698d737d8	2016-06-20 08:49:52 +0000	http://www.flipkart.com/santosh-royal-fashion-...	SANTOSH ROYAL FASHION Cotton Printed King size...	["Home Furnishing >> Bed Linen >> Bedsheets >>...	BDSEJT9UQWHDUBH4
4	6325b6870c54cd47be6ebfbffa620ec7	2016-06-20 08:49:52 +0000	http://www.flipkart.com/jaipur-print-cotton-fl...	Jaipur Print Cotton Floral King sized Double B...	["Home Furnishing >> Bed Linen >> Bedsheets >>...	BDSEJTHNGWVGWWQU
...	...	...	...	...	...	...
1045	958f54f4c46b53c8a0a9b8167d9140bc	2015-12-01 10:15:43 +0000	http://www.flipkart.com/oren-empower-extra-lar...	Oren Empower Extra Large Self Adhesive Sticker	["Baby Care >> Baby & Kids Gifts >> Stickers >...	STIE88ZGTX65GH4V

Nettoyage

Je remplace les NaN de product_spécifications par ‘Rien’

Je remplace les NaN de brand par ‘Autre’

Je remplace les NaN de retail_price et discounted_price par la moyenne

J’enleve les colonnes uniq_id, crawl_timestamp, product_url, pid, product_rating, overall_rating

```
1 df.isna().sum()
```

```
uniq_id                    0
crawl_timestamp            0
product_url                0
product_name               0
product_category_tree      0
pid                        0
retail_price               1
discounted_price           1
image                      0
is_FK_Advantage_product    0
description                0
product_rating             0
overall_rating             0
brand                    338
product_specifications     1
dtype: int64
```

Nettoyage

Vu qu'il y a beaucoup de sous catégorie, je vais garder que les catégories principaux pour faciliter le clustering

```
df_drop2['product_category_tree'].value_counts()
```

```
["Home Furnishing >> Bed Linen >> Blankets, Quilts & Dohars"]                                                    56
["Kitchen & Dining >> Coffee Mugs >> Prithish Coffee Mugs"]                                                      26
["Watches >> Wrist Watches >> Maxima Wrist Watches"]                                                             23
["Kitchen & Dining >> Coffee Mugs >> Rockmantra Coffee Mugs"]                                                    22
["Watches >> Wrist Watches >> Sonata Wrist Watches"]                                                             19
                                                                                                                 ..
["Home Decor & Festive Needs >> Showpieces >> Adaa Showpieces"]                                                   1
["Beauty and Personal Care >> Body and Skin Care >> Face Care >> Sunscreen >> Richfeel Sunscreen"]                1
["Home Decor & Festive Needs >> Showpieces >> Unique Design Showpieces"]                                          1
["Kitchen & Dining >> Kitchen Tools >> Kitchen Implements >> Pizza Cutters >> Step4deal Pizza Cutters"]          1
["Baby Care >> Infant Wear >> Baby Boys' Clothes >> Shirts >> Beebay Shirts >> Beebay Baby Boy's Checkered Casual Shirt"]    1
Name: product_category_tree, Length: 642, dtype: int64
```

Nettoyage

product_category_tree
["Home Furnishing >> Curtains & Accessories >>...
["Baby Care >> Baby Bath & Skin >> Baby Bath T...
["Baby Care >> Baby Bath & Skin >> Baby Bath T...
["Home Furnishing >> Bed Linen >> Bedsheets >>...
["Home Furnishing >> Bed Linen >> Bedsheets >>...

```
df_drop2['categorie1'] = df_drop2['product_category_tree'].str.split().str[0]
df_drop2['categorie2'] = df_drop2['product_category_tree'].str.split().str[1]
```

```
df_drop2['categorie'] = df_drop2['categorie1'] + ' ' + df_drop2['categorie2']
df_drop2 = df_drop2.drop(['categorie1', 'categorie2'], axis = 1)
```

categorie
["Home Furnishing
["Baby Care
["Baby Care
["Home Furnishing
["Home Furnishing
...

Nettoyage

```
1 df_drop2['categorie'].value_counts()
```

```
Home Decor & Festive Needs    150
Watches                       150
Home Furnishing               150
Kitchen & Dining              150
Computers                     150
Beauty and Personal Care      150
Baby Care                     150
Name: categorie, dtype: int64
```

Nettoyage de la colonne description

- J'enlève la ponctuation
- Je remplace les majuscules par des minuscules
- Je vais lemmatiser les mots

```
df_drop['sentence_bow'] = df_drop['description'].apply(lambda x : transform_bow_fct(x))
df_drop['sentence_bow_lem'] = df_drop['description'].apply(lambda x : transform_bow_lem_fct(x))
df_drop['sentence_dl'] = df_drop['description'].apply(lambda x : transform_dl_fct(x))
```

```
df_drop['sentence_stem'] = df_drop['description'].apply(lambda x : stem_dataset(x))
```

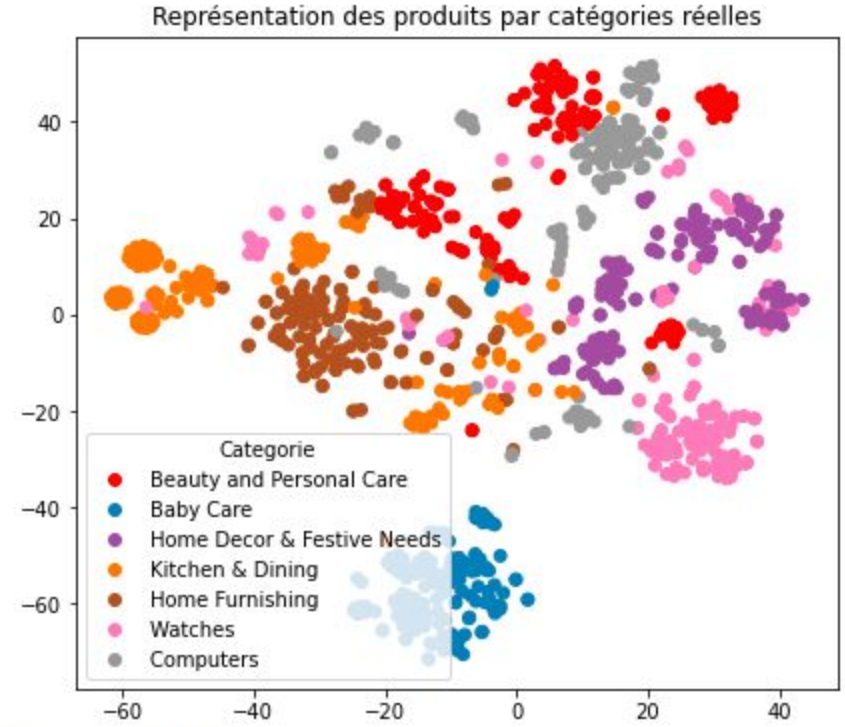
Traitement de texte


```
{'key': 2925,
 'feature': 2237,
 'elegance': 2047,
 'polyester': 3827,
 'multicolor': 3378,
 'abstract': 755,
 'eyelet': 2186,
 'door': 1934,
 'curtain': 1736,
 'floral': 2318,
 '213': 234,
 'height': 2621,
 'pack': 3621,
 'price': 3900,
```

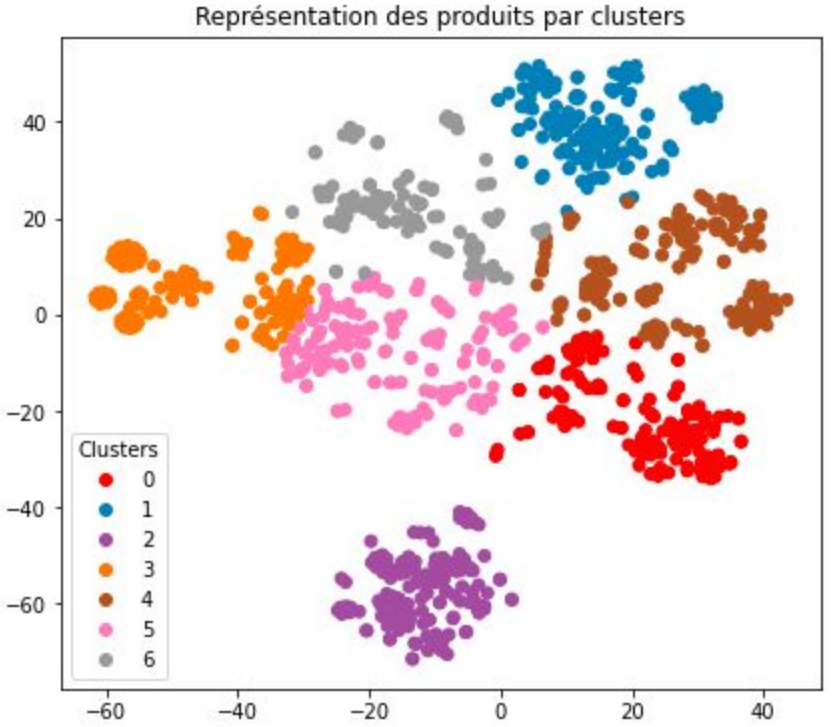

CountVectorizer
Compte le nombre de fois qu'un mot
apparaît dans un document
ARI : 0.32

TfidfVectorizer
Compte le nombre de fois où le mot apparaît mais aussi son importance dans tout le corpus. En pénalisant les mots qui
reviennent le plus souvent.
ARI : 0.44

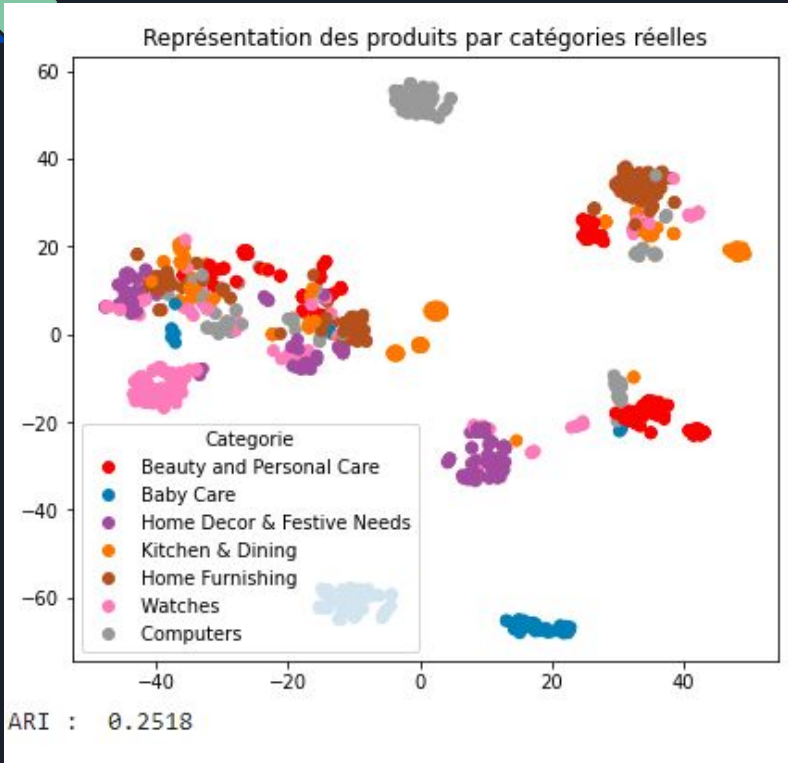
TFIDF

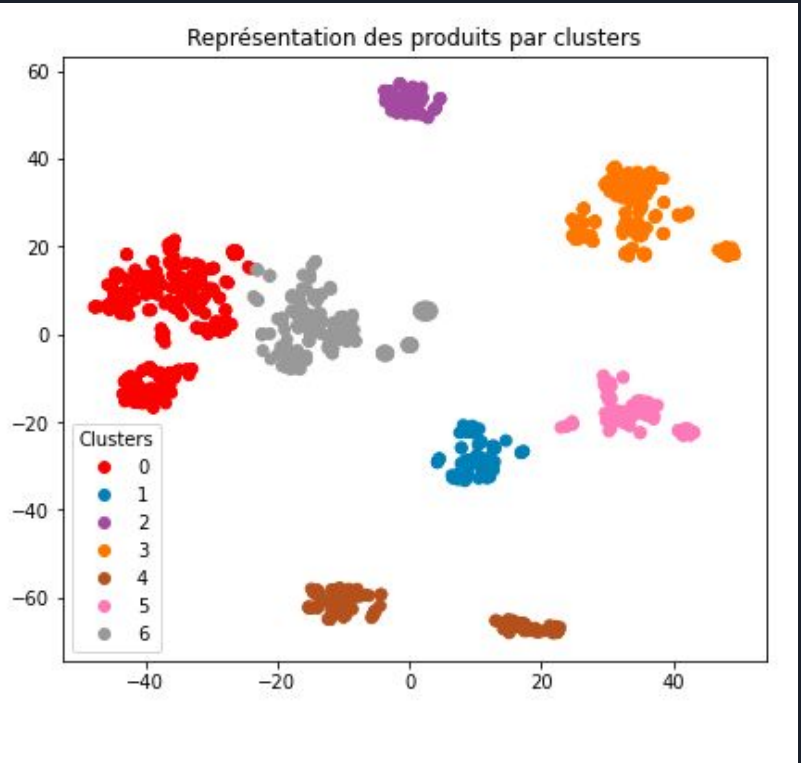


ARI : 0.4449



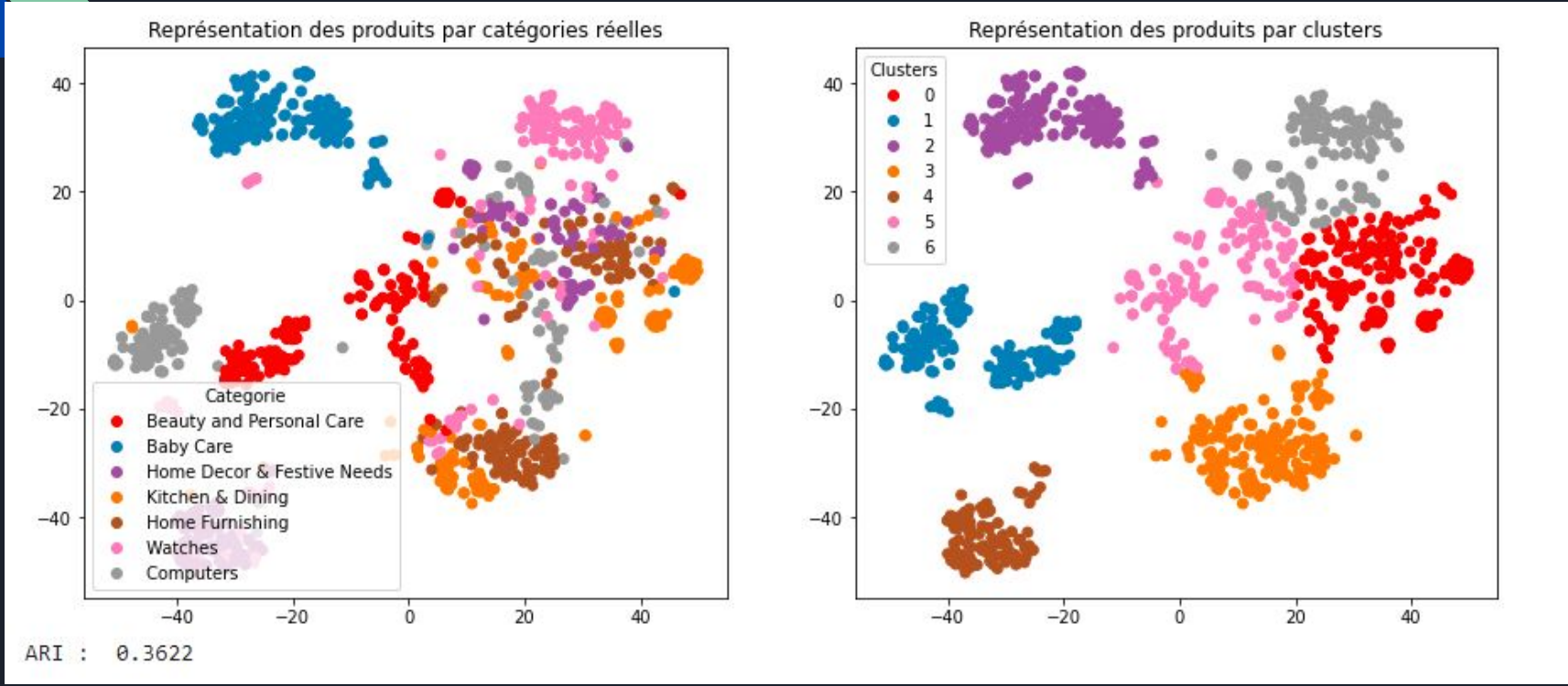
WORD2VEC





Similarité entre les mots en utilisant l'embedding (vecteur de mots)

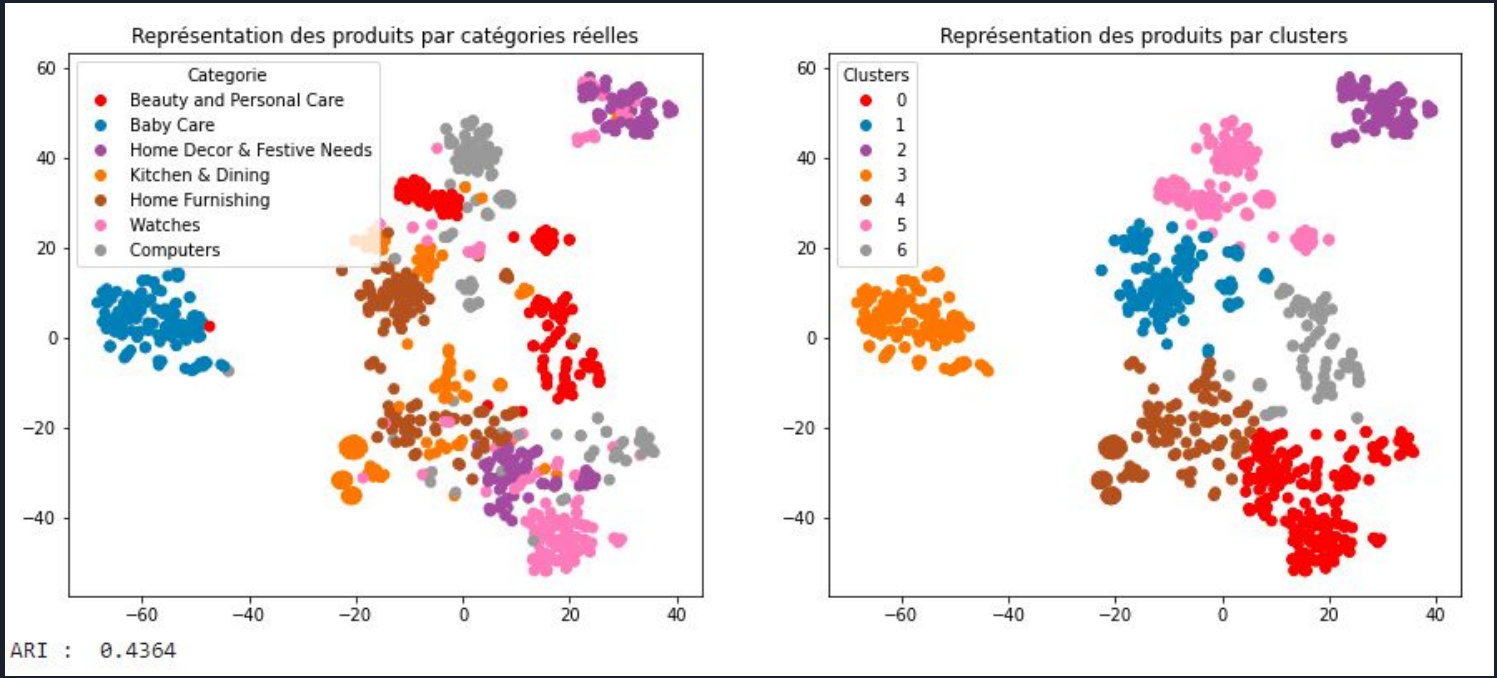
BERT



Bidirectional Encoder Representations from Transformers

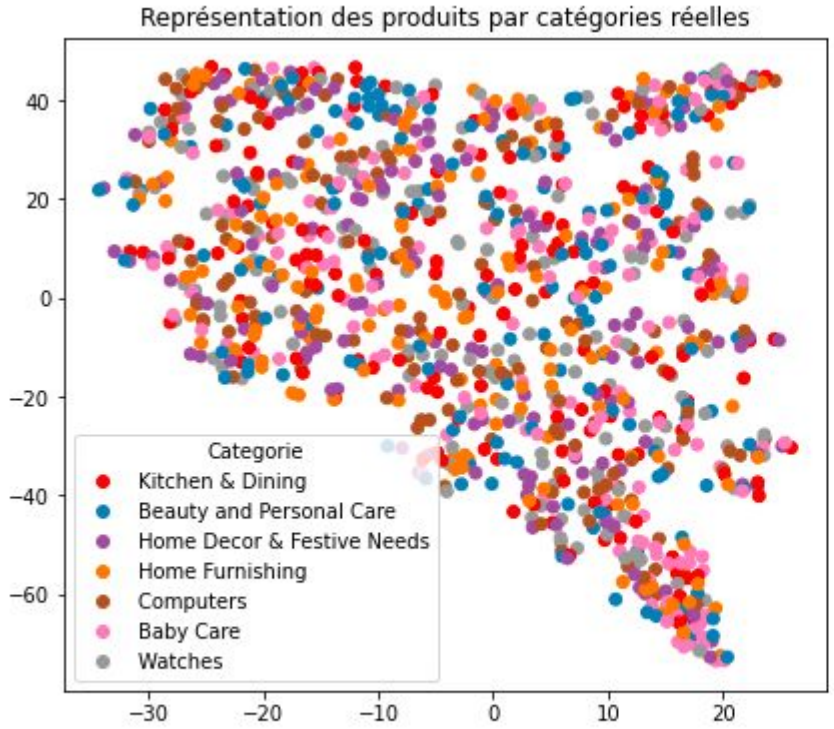
Transformers : mécanisme d'attention, apprends les relations contextuelles entre les mots dans un texte

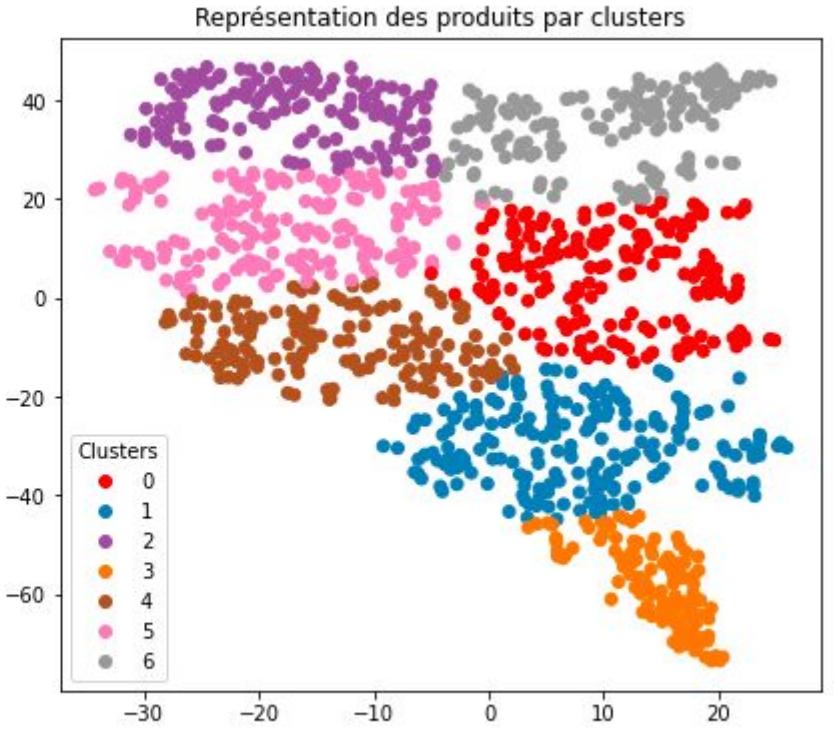
UNIVERSAL SENTENCE ENCODER



Code le texte en vecteur qui peuvent être utilisé pour faire de la classification de texte

ORB

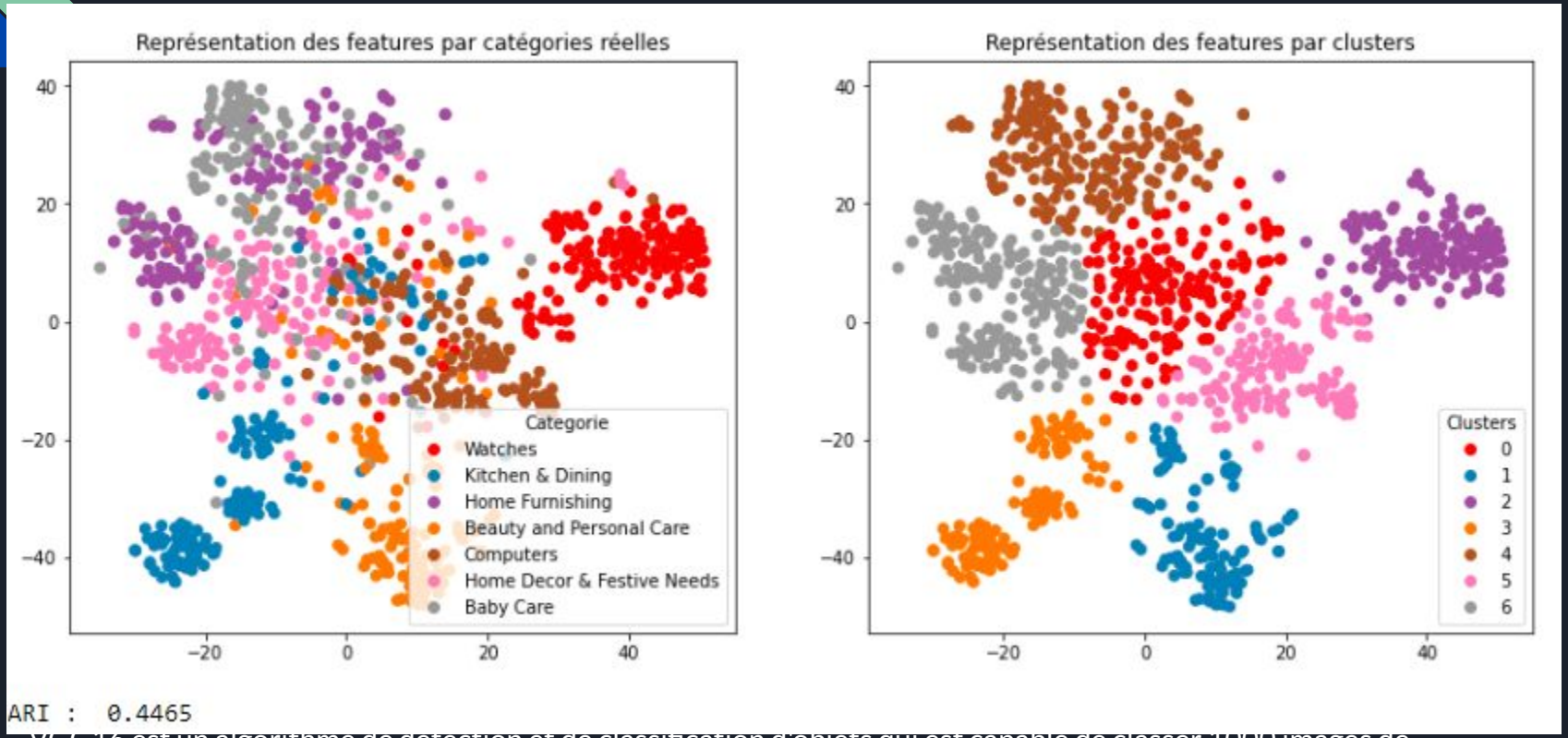




ARI : 0.0005

ORB utilise une version modifiée du détecteur de points clés FAST et du descripteur BRIEF.

VGG16



ARI : 0.4465

VGG16 est un algorithme de détection et de classification d'objets qui est capable de classer 1000 images de 1000 catégories différentes avec une précision de 92,7%. Le VGG utilise des champs réceptifs très petits et il y a également moins de paramètres.

Nom du modèle	ARI Score
TFIDF	0.44
WORD2VEC	0.25
BERT	0.36
Universal Sentence Encoder	0.43
ORB	0.0005
VGG16	0.4465

	product_name	categorie	cat_labels	image
1	Sathiyas Cotton Bath Towel	Baby Care	Home Furnishing	7b72c92c2f6c40268628ec5f14c6d590.jpg

Wallmantra Medium Vinyl Sticker Sticker Baby Care Home Furnishing c3edc504d1b4f0ba6224fa53a43a7ad6.jpg

Conclusion

- Réorganiser les catégories
- Créer notre propre modèle
- L’étude de faisabilité est positif

Axe d’amélioration

- Faire en sorte que le modèle reconnaisse mieux le “Home Furnishing”

Merci