Projet 6

Classification des Biens de Consommation

Eva Bookjans

Etude de Faisabilité d'un Moteur de Classification

place de marché

Photo

Description

"Rockmantra Water Fire Ceramic Mug (5.5 l) Price: Rs.199 Give a thrilling yet fresh start to your day. An exclusive creation by Rockmantra,… "

- Baby Care
- Beauty & Personal Care
- Computers
- Home Decor & Festive Needs
- Home Furnishing
- **Kitchen & Dining**
- Watches
- …

Automatisation

- **Passage à l'échelle**
- **Expérience utilisateur fluide**
  - **Vendeurs** : faciliter la mise en ligne de nouveaux articles
  - **Acheteurs** : faciliter la recherche de produits

Le Jeu des Données

7 Catégories principals

- Baby Care
- Beauty & Personal Care
- Computers
- Home Decor & Festive Needs
- Home Furnishing
- Kitchen & Dining
- Watches

- 150 entrées par catégorie
- **1050 données en totale**

- **Petit** jeu de données
- **Equilibré** entre catégories
- **Complète** (photo + description)

➢ Suffisant pour une **première étude de faisabilité**

Outils et Méthodes NLP

Natural Language Processing

- Tokenisation
- Lemmatisation
- Vectorisation
- Réduction de Dimension
- Clustering
- Classification

La **Description** du Produit

Traitement de Texte

Traitement de Texte

- **Tokenisation**
- **Lemmatisation**
- **Vectorisation**
- **Réduction de Dimension**
- Clustering
- Classification

WordCloud – technique de visualisation des données textuelles

➢ la taille du mot indique sa fréquence

Simple Traitement de Texte (inclus dans le module):

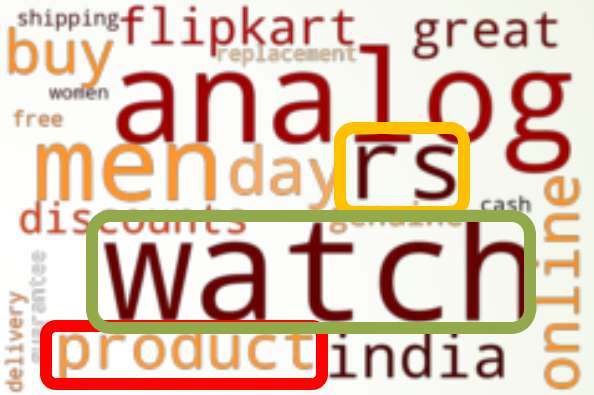
- **Tokenisation -** le texte est coupé en mots (simples ou/et en paires) —> 'tokens'
- **Lemmatisation -** traitement des pluriels (oui/non)
- **Vectorisation -** la fréquence du 'token' dans le texte
- **Réduction de Dimension**
  - Élimination des 'stopwords' (= mots/tokens sans signifiance)
  - Maximum nombre de mots/tokens à afficher

La **Description** du Produit avec WordCloud

Home Furnishing

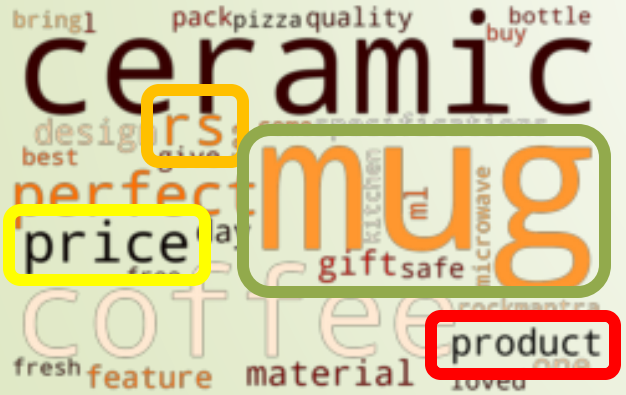
Baby Care

Watches



Home Decor & Festive Needs

Kitchen & Dining



Beauty and Personal Care

Computers

- Stopwords métier
- Mots-clés

Traitement de Texte

- **Tokenisation**
- **Lemmatisation**
- Vectorisation
- Réduction de Dimension
- Clustering
- Classification

Bibliothèques NLP

NLTK +

WordNet (base de données lexicale)

- faite pour la recherche / enseignement

SpaCy

- plus robuste

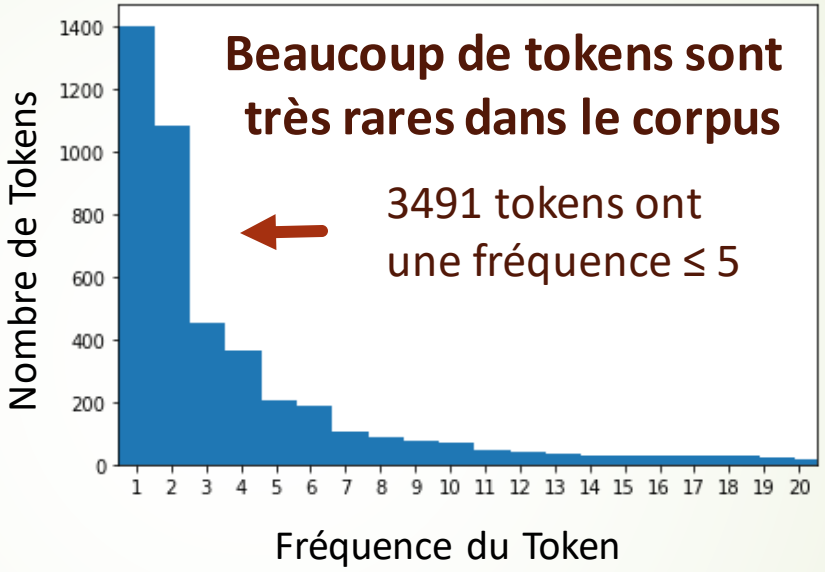
Bases de données lexicales :

- 'stopwords' standards
- fonction sémantique des mots

Traitement de Texte

- **Tokenisation**
- **Lemmatisation**
- Vectorisation
- Réduction de Dimension
- Clustering
- Classification

Les Tokens



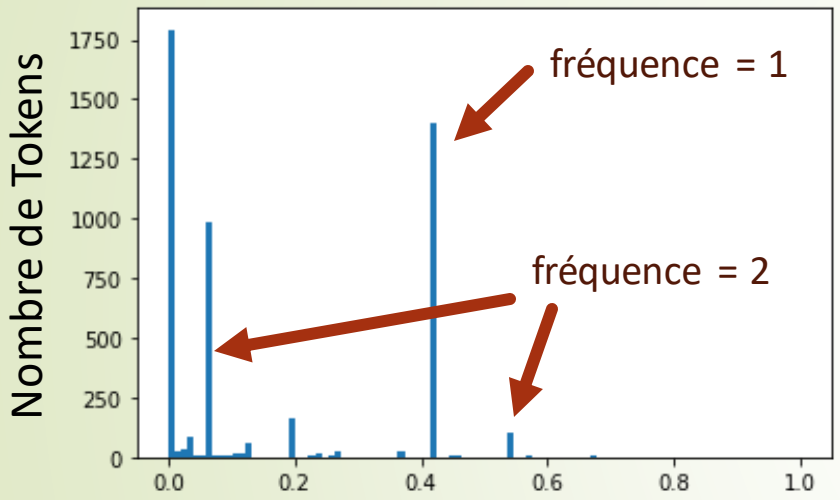
Fréquence moyenne des token : 12
Tokens par document en moyenne : 53
Tokens uniques : 4799

Les 22 tokens le plus fréquents

TOKEN	Count
rs	911
free	618
cm	594
buy	583
products	577
delivery	567
cash	564
genuine	564
replacement	559
day	549
price	541
flipkart	481
guarantee	473
com	473
mug	406
online	396
shipping	381
color	343
features	337
watch	336
pack	328
baby	321

Les Tokens

p-value du token
vis-à-vis les catégories (Chi2)



7 Catégories

Fréq.	min. p-Value
1	0.423190
2	0.061969
3	0.006232
4	0.000522
5	0.000039

Stopwords du corpus

Les fréquences de document le plus élevées

TOKEN	p-Value	Fréq.	Doc. Fréq.
rs	6.0e-02	911	911
free	6.5e-13	618	595
buy	1.3e-13	583	578
products	1.3e-14	577	569
delivery	1.4e-15	567	566
cash	1.8e-15	564	564
genuine	1.8e-15	564	564
price	2.1e-38	541	525
day	6.3e-30	549	512
replacement	1.2e-61	559	489
guarantee	7.8e-43	473	471
flipkart	3.3e-68	481	392
online	4.4e-44	396	389
com	1.1e-68	473	385
shipping	3.0e-54	381	381
specifications	3.1e-18	321	309
general	2.0e-20	288	284
box	1.1e-09	297	251
features	8.1e-12	337	241
type	1.5e-10	318	237
color	1.8e-05	343	221
sales	3.7e-27	261	218

Règles de Sélection:

- p-Value ≤ 0.01
- Max. DF = 2/7 (≤ 300 documents)
- Min. DF = 0.03/7 (≥ 4.5 documents)

Traitement de Texte

- **Tokenisation**
- **Lemmatisation**
- Vectorisation
- Réduction de Dimension
- Clustering
- Classification

```
Buy Nagar Handloom Floral Double Quilts & Comforters
Multicolor at Rs. 1350 at Flipkart.com. Only Genuine
Products. Free Shipping. Cash On Delivery!
```

- Suppression des signes de ponctuation et des chiffres
- Tokenisation (NLTK et Wordnet / SpaCy)
- Transformation en minuscule

```
buy nagar handloom floral double quilts comforters multicolor
at rs at flipkart com only genuine products free shipping
cash on delivery
```

- Lemmatisation (NLTK et Wordnet / SpaCy)
- Suppression des 'Stopwords'
  - TLTK et Wordnet / SpaCy
  - Stopwords du corpus (e.g. 'buy', 'rs', 'flipkart', 'com', 'product', 'free', 'shipping', 'cash', 'delivery', …)

```
nagar handloom floral double quilts comforters multicolor
```

Traitement de Texte

- **Tokenisation**
- **Lemmatisation**
- Vectorisation
- Réduction de Dimension
- Clustering
- Classification

```
Key Features of Lula Baby Girl's Dark Blue Bodysuit Fabric: cotton spandex Brand Color: ROYAL BLUE,Lula Baby Girl's Dark Blue Bodysuit Price: Rs. 330 Lula babywear is designed to caress the baby like rose petals. Softest and safest cotton wear made for the comfort of the babies.
```

- Suppression des signes de ponctuation et des chiffres
- Tokenisation (NLTK et Wordnet / SpaCy)
- Transformation en minuscule

```
key features of lula baby girl s dark blue bodysuit fabric cotton spandex brand color royal blue lula baby girl s dark blue bodysuit price rs lula babywear is designed to caress the baby like rose petals softest and safest cotton wear made for the comfort of the babies
```

- Lemmatisation (NLTK et Wordnet / SpaCy)
- Suppression des 'Stopwords'
  - TLTK et Wordnet / SpaCy
  - Customisés (e.g. 'features', 'rs', 'flipkart', 'com', 'price', ...)

```
key lula baby girl dark blue bodysuit fabric cotton spandex royal blue lula baby girl dark blue bodysuit lula babywear caress baby like rise petal softest safe cotton wear make comfort baby
```

Traitement de Texte

- Tokenisation
- Lemmatisation
- **Vectorisation**
- **Réduction de Dimension**
- Clustering
- Classification

Les Tokens

4799 Tokens (après tokenisation et lemmatisation)

Vectorisation + Règles de Sélection

- CountVectorizer(max_df = 2/7, min_df = 0.03/7)
- SelectorFpr(chi2, alpha = 0.01)
- TfidfTransformer()

814 Tokens

Réduction de Dimensions

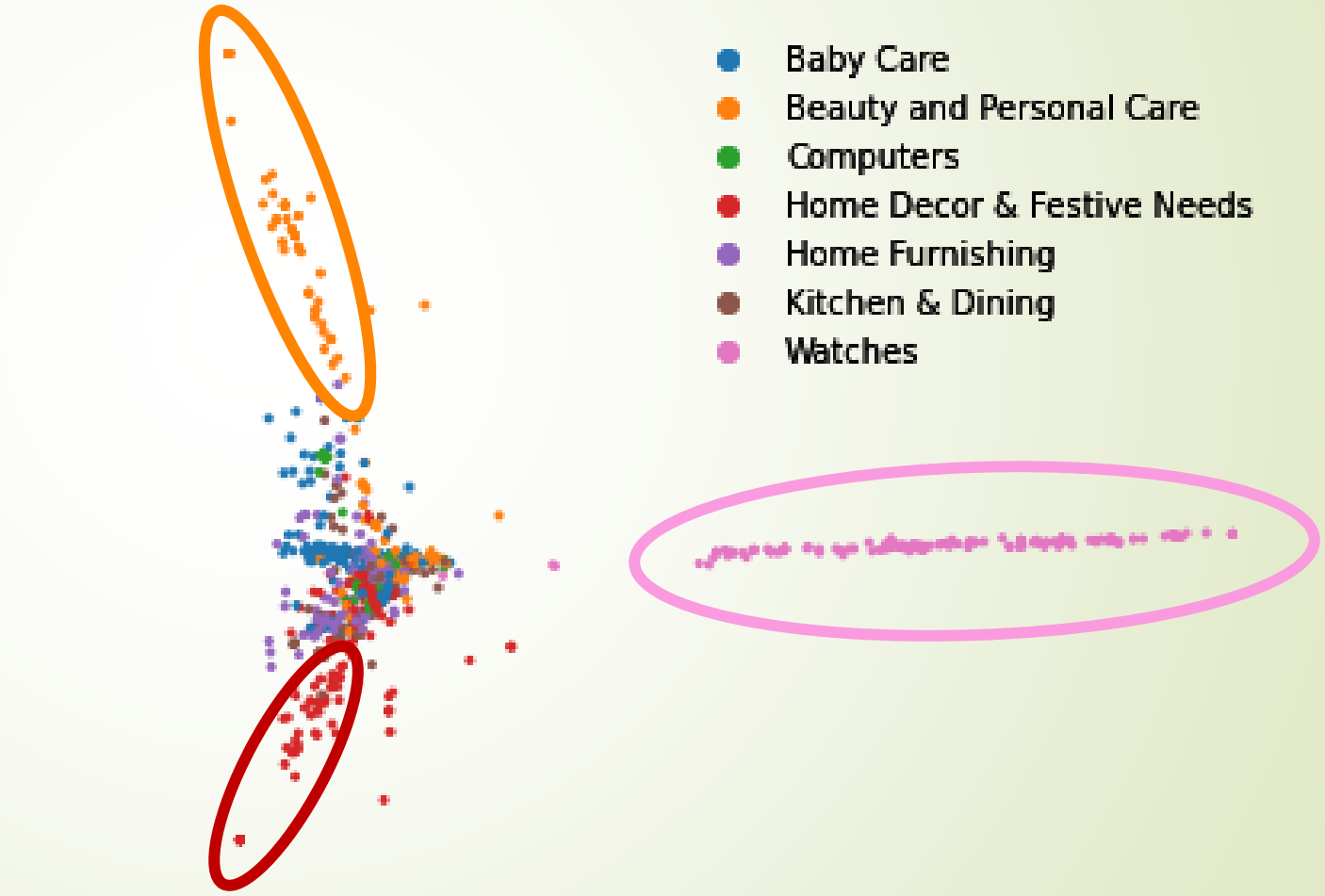
- PCA(n_components = 0.99)

440 Features

Traitement de Texte

- Tokenisation
- Lemmatisation
- Vectorisation
- **Réduction de Dimension**
- Clustering
- Classification

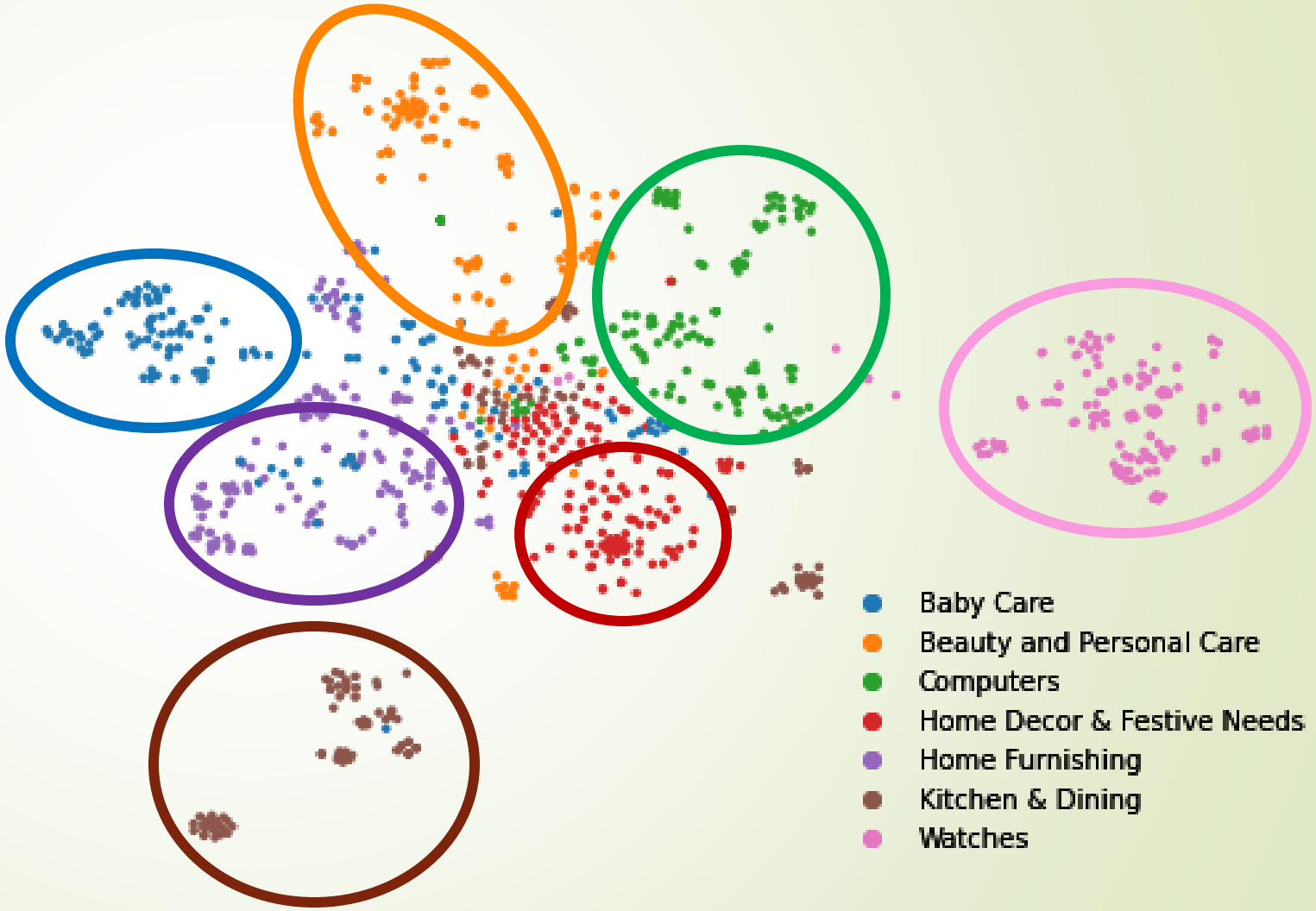
Projection en 2D – PCA



Traitement de Texte

- Tokenisation
- Lemmatisation
- Vectorisation
- **Réduction de Dimension**
- Clustering
- Classification

Projection en 2D – t-SNE

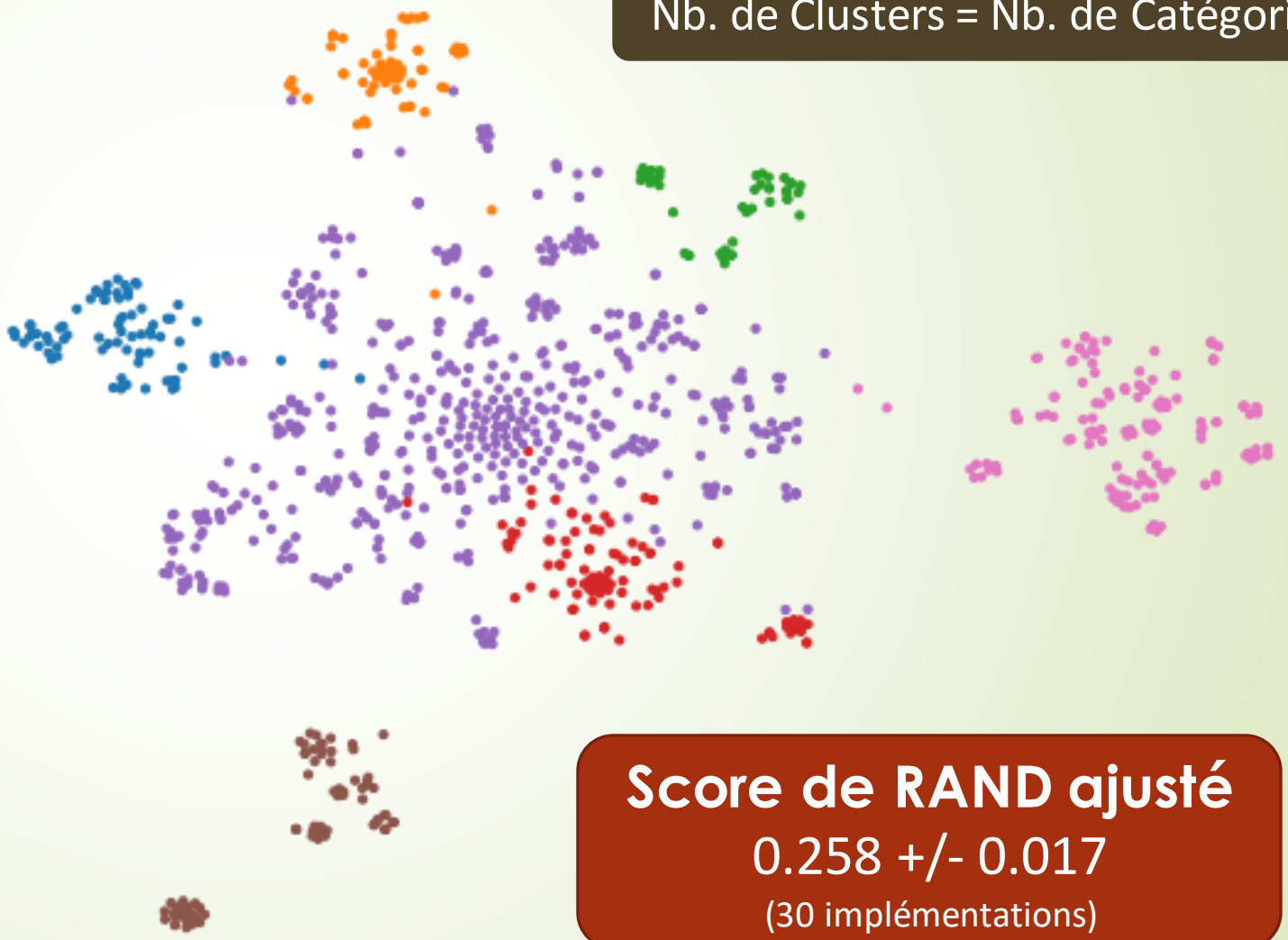


Traitement de Texte

- Tokenisation
- Lemmatisation
- Vectorisation
- Réduction de Dimension
- **Clustering**
- Classification

KMeans Clustering

Nb. de Clusters = Nb. de Catégories

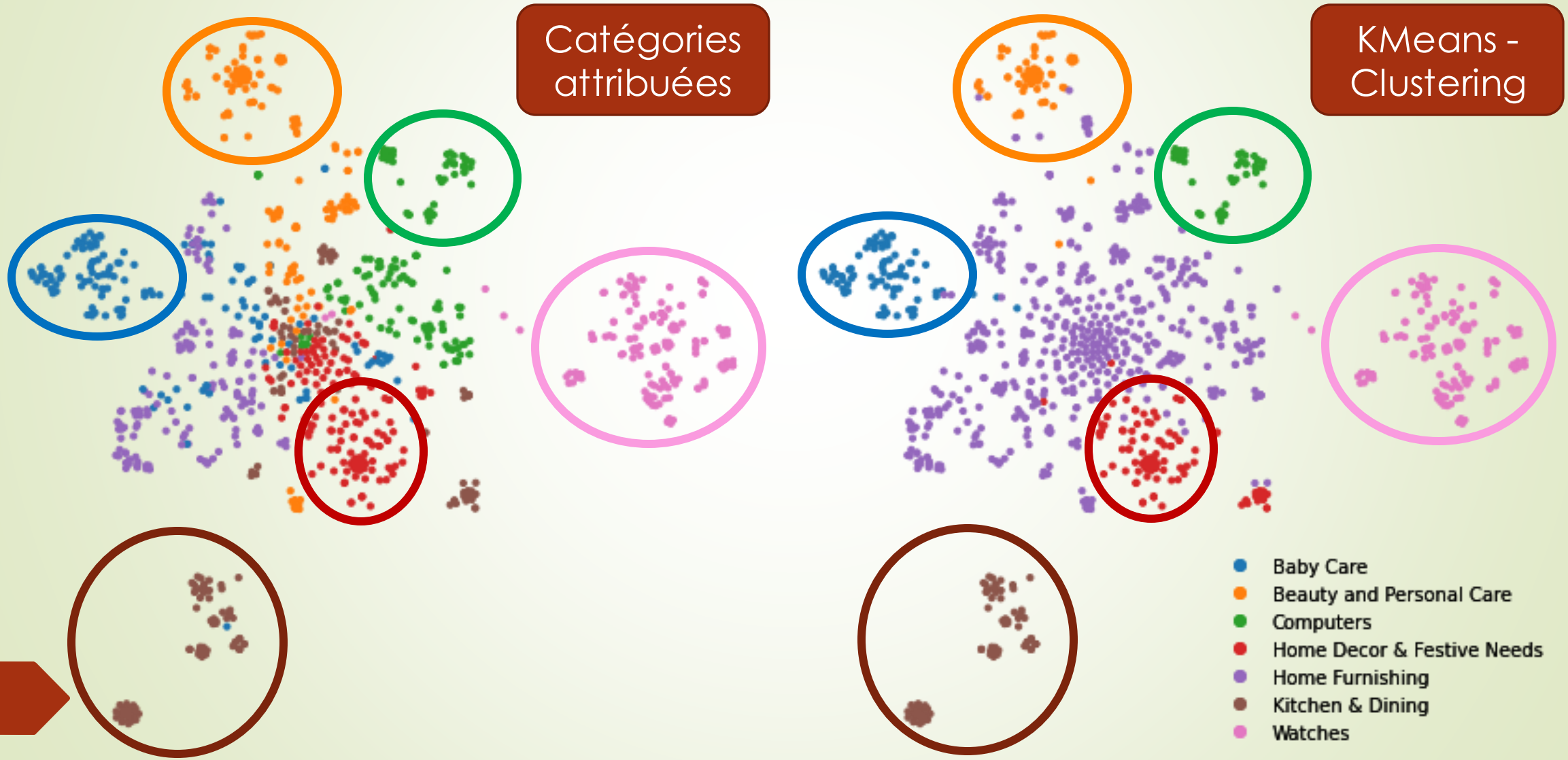


Score de RAND ajusté

0.258 +/- 0.017

(30 implémentations)

KMeans Clustering – Evaluation

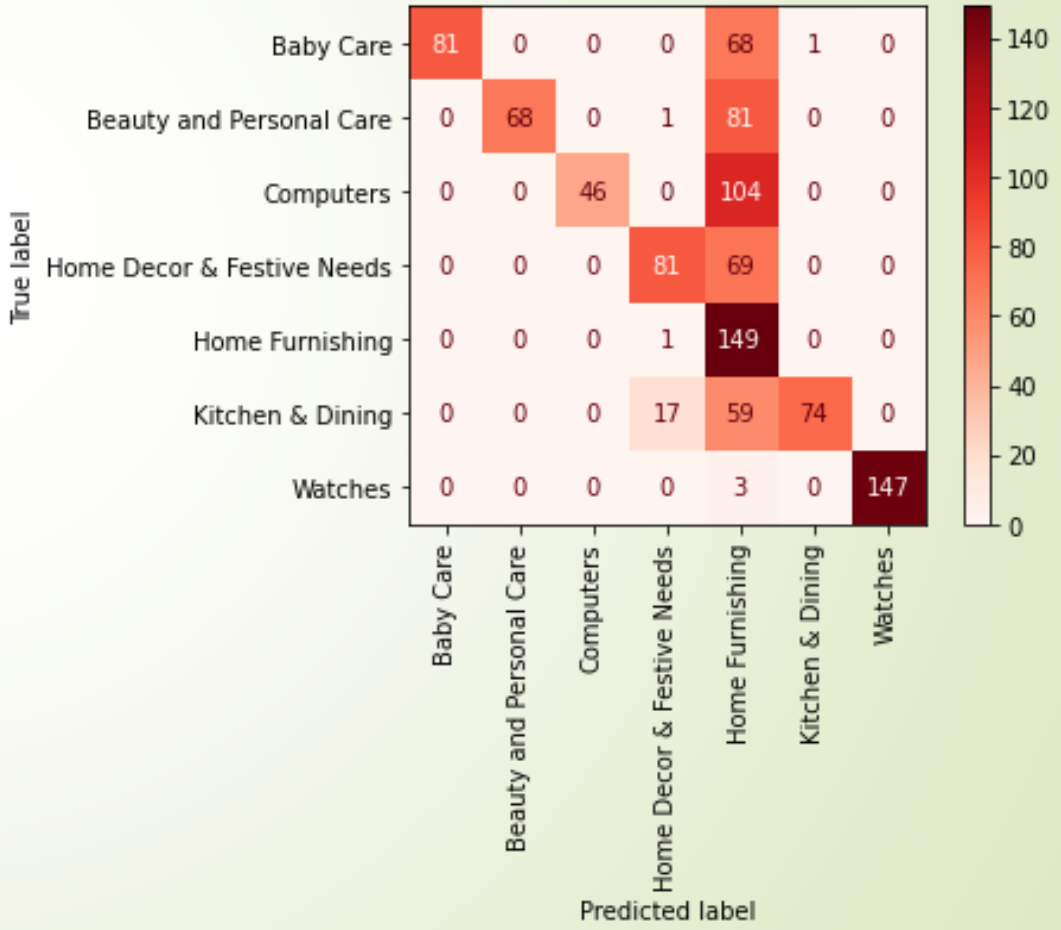


KMeans Clustering – Evaluation

Rapport de Classification

	precision	recall	f1-score	support
Baby Care	1.00	0.54	0.70	150
Beauty and Personal Care	1.00	0.45	0.62	150
Computers	1.00	0.31	0.47	150
Home Decor & Festive Needs	0.81	0.54	0.65	150
Home Furnishing	0.28	0.99	0.44	150
Kitchen & Dining	0.99	0.49	0.66	150
Watches	1.00	0.98	0.99	150
accuracy			0.62	1050
macro avg	0.87	0.62	0.65	1050
weighted avg	0.87	0.62	0.65	1050

Matrice de Confusion



Accuracy : 0.62

Traitement de Texte

- Tokenisation
- Lemmatisation
- Vectorisation
- Réduction de Dimension
- **Clustering**
- Classification

Clustering autres modèles

Tokenisation et Lemmatisation

- **NLTK et Wordnet, Spacy**

+

Vectorisation + Règles de Sélection

Version1

- CountVectorizer(max_df = 2/7, min_df = 0.03/7)
- SelectorFpr(chi2, alpha = 0.01)
- TfidfTransformer()
- PCA(n_components = 0.99)

Version2

- CountVectorizer(max_df = 2/7, min_df = 0.03/7, **stop_words** = [...])
- SelectorFpr(chi2, alpha = 0.01)
- TfidfTransformer()
- **TruncatedSVD**(n_components = 450)

+

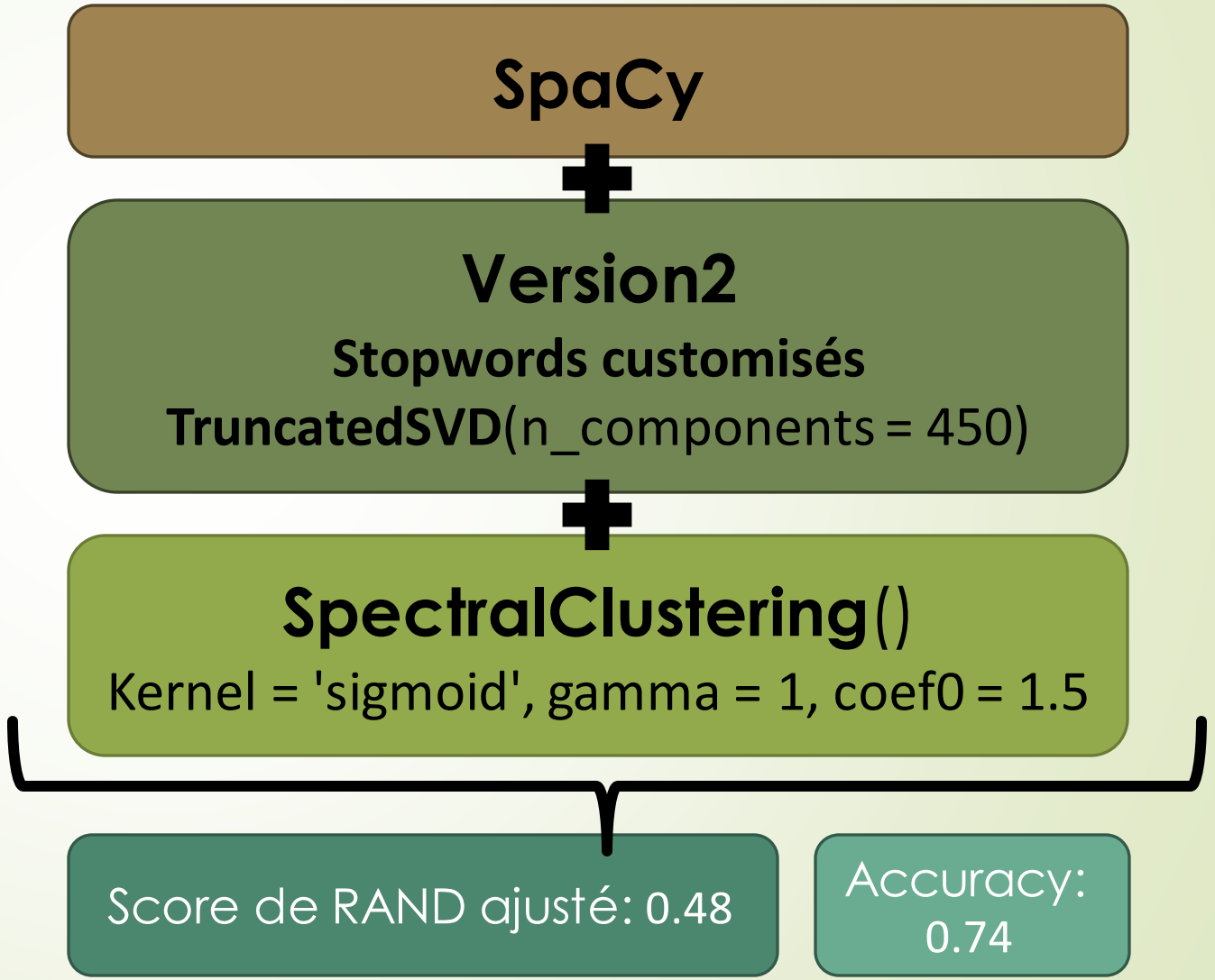
Algorithme de Clustering

- **KMeans, GaussianMixture, AgglomerativeClustering, SpectralClustering**

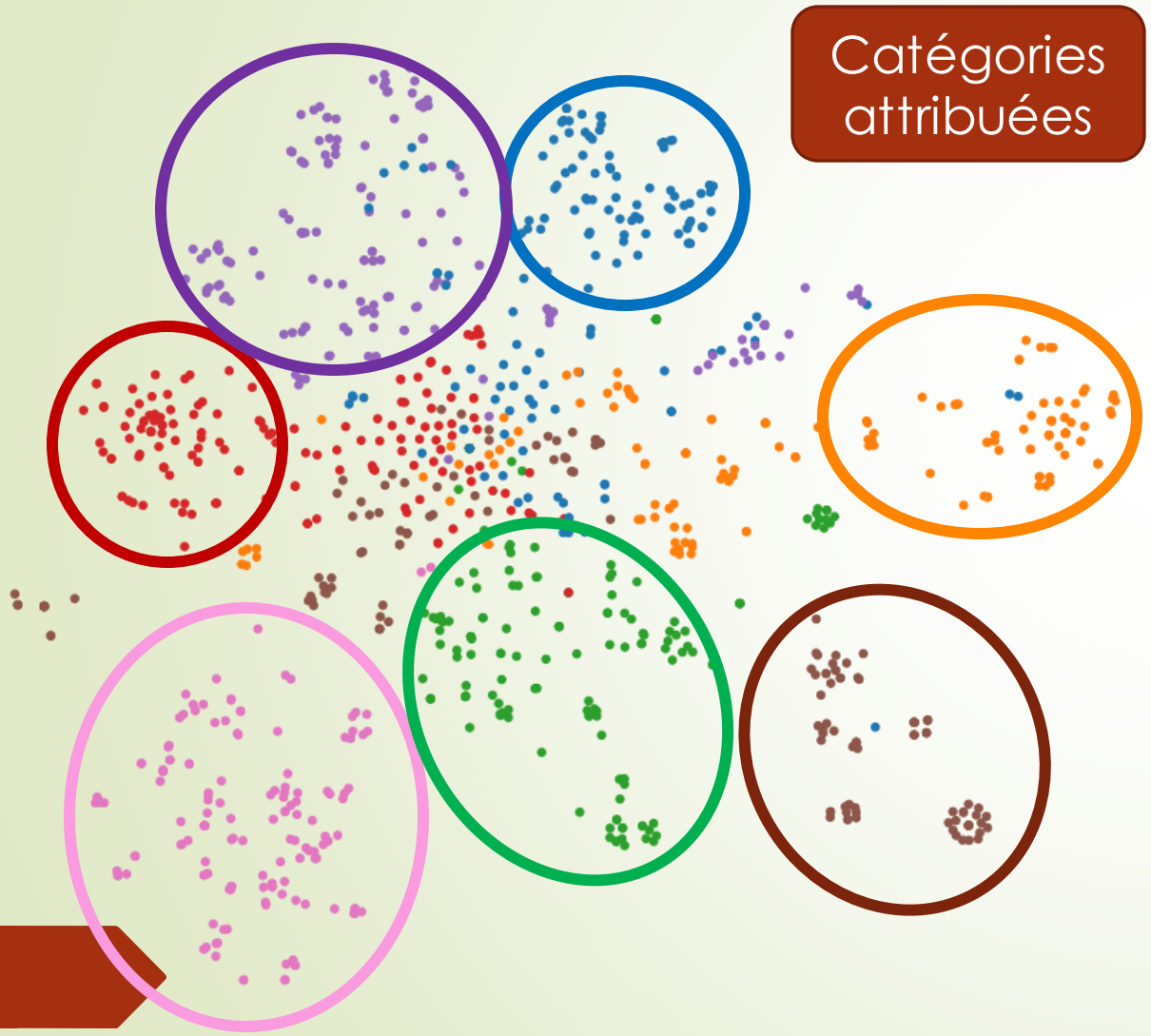
Traitement de Texte

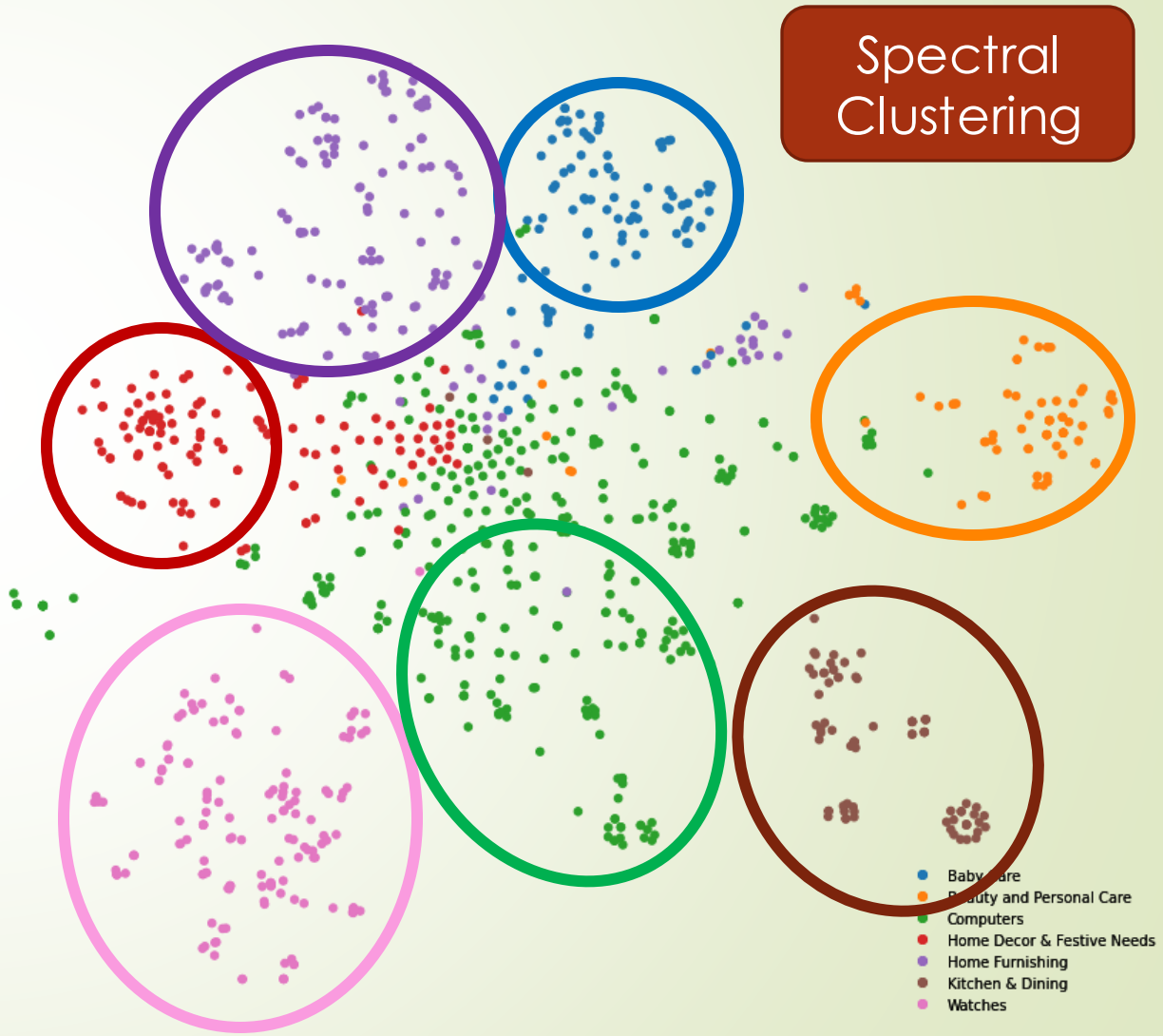
- Tokenisation
- Lemmatisation
- Vectorisation
- Réduction de Dimension
- **Clustering**
- Classification

SpectralClustering()



SpectralClustering – Evaluation



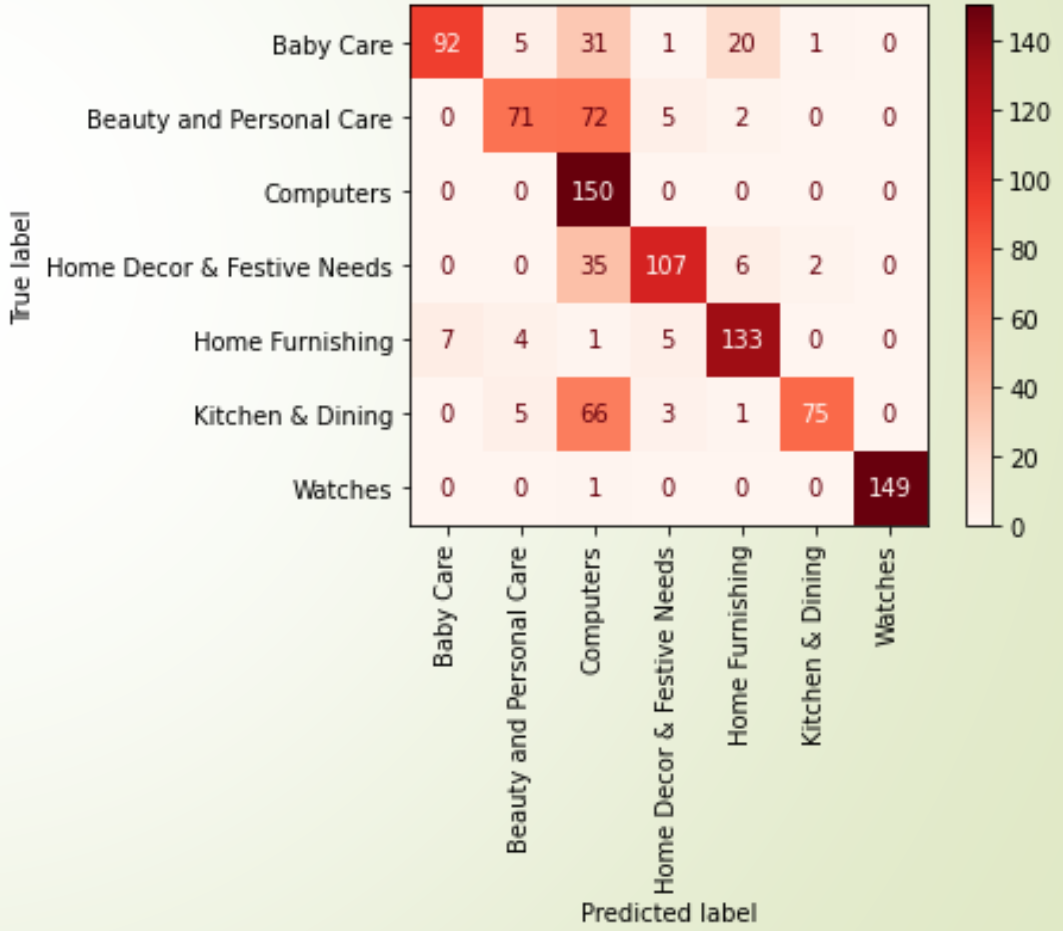


SpectralClustering – Evaluation

Rapport de Classification

	precision	recall	f1-score	support
Baby Care	0.93	0.61	0.74	150
Beauty and Personal Care	0.84	0.47	0.60	150
Computers	0.42	1.00	0.59	150
Home Decor & Festive Needs	0.88	0.71	0.79	150
Home Furnishing	0.82	0.89	0.85	150
Kitchen & Dining	0.96	0.50	0.66	150
Watches	1.00	0.99	1.00	150
accuracy			0.74	1050
macro avg	0.84	0.74	0.75	1050
weighted avg	0.84	0.74	0.75	1050

Matrice de Confusion



Accuracy : 0.74

Traitement de Texte

- Tokenisation
- Lemmatisation
- Vectorisation
- Réduction de Dimension
- Clustering
- **Classification**

Classification

Modèles Testés

- KNeighborsClassifier()
- ComplementNB()
- MultinomialNB()

Accuracy: 0.90-0.92

un moteur de classification basé sur la description du produit doit être **faisable**

Résumé de Traitement de Texte

Un Moteur de Classification est faisable

- Un simple Clustering montre déjà les différentes catégories avec une accuracy > 0.6
- Des premiers classifications donnent une accuracy de 0.90-0.92

Axes d'Amélioration

- Stopwords customisés : best, good, perfect, …
- Meilleur dictionnaire pour la lemmatisation : quilt(s), conforter(s), ...
- Inclure des Mots composés (n grams) : e.g. hair spray, …

Outils et Méthodes CV

Computer Vision

- Keypoints
- Descripteurs
- Vectorisation
- Réduction de Dimension
- Clustering
- Classification

Les **Images** du Produit

Computer Vision

Traitement d'Image

- **Keypoints**
- **Descripteurs**
- Vectorisation
- Réduction de Dimension
- Clustering
- Classification

Bibliothèque OpenCV

SIFT / ORB / SURF

- algorithmes pour la détection et la description des caractéristiques/features (locales) dans les images
- outils de computer vision (e.g. détection d'objets)

Manipulation d'images

- filtres, transformations, couleurs, affichage, ...

Computer Vison

- détection des features, objects, ...

Les **Keypoints** et **Descripteurs SIFT**

Image originale → En Gris → Egalisation → Keypoints Descripteurs

"Mots virtuels"

1216 points clés décrits par un **descripteur SIFT**

(= vecteur de longueur 128)

SIFT – Scale-invariant Feature Transform

Traitement d'Image

- Keypoints
- Descripteurs
- **Vectorisation**
- **Réduction de Dimension**
- Clustering
- Classification

Création de Features

989655 Descripteurs SIFT (longueur = 128, max. features = 1000)

MiniBatchKMeans Clustering
- nb. clusters = sqrt(nb. Descripteurs)

995 Descripteur Clusters ("Sacs de Mots Virtuels")

Vectorisation des Descripteurs d'Image
- affectation aux clusters
- nombre de descripteurs par cluster (histogramme)

995 Features

Réduction de Dimensions
- PCA(n_components = 0.99)

440 Features

Traitement d'Image

- Keypoints
- Descripteurs
- Vectorisation
- **Réduction de Dimension**
- Clustering
- Classification

Projection en 2D – PCA

Traitement d'Image

- Keypoints
- Descripteurs
- Vectorisation
- **Réduction de Dimension**
- Clustering
- Classification

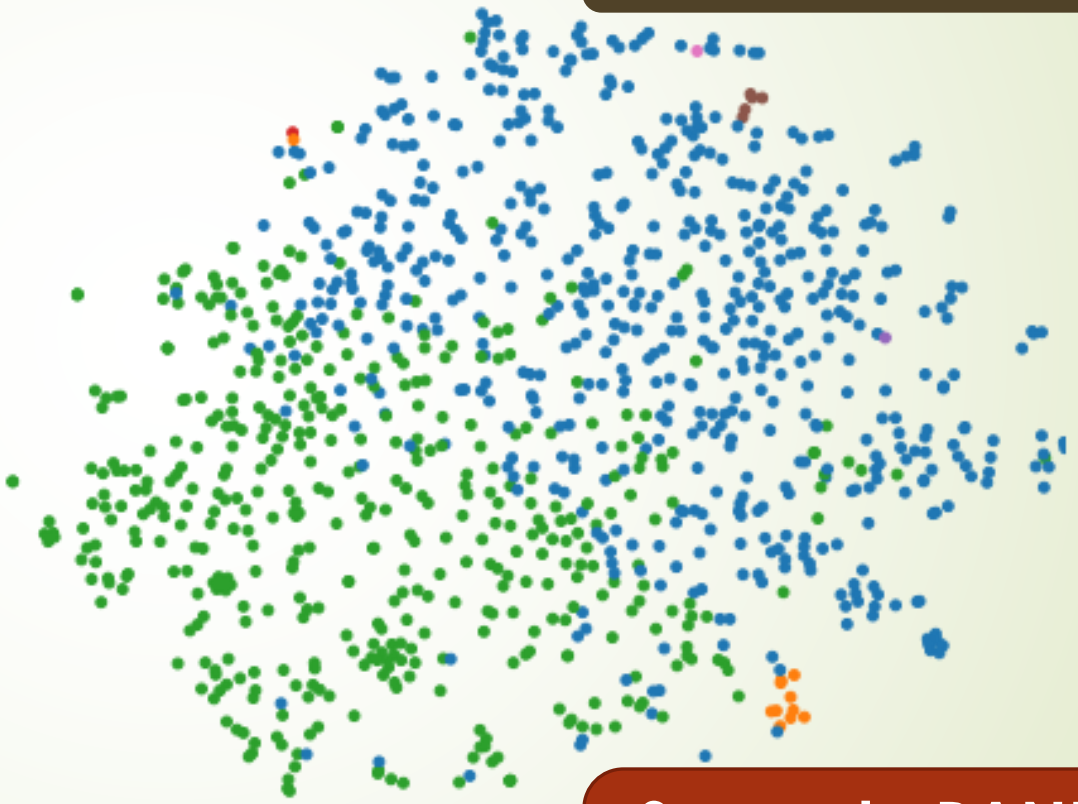
Projection en 2D – t-SNE

Traitement d'Image

- Keypoints
- Descripteurs
- Vectorisation
- Réduction de Dimension
- **Clustering**
- Classification

KMeans Clustering

Nb. de Clusters = Nb. de Catégories



Score de RAND ajusté
0.027 +/- 0.006
(10 implémentations)

Traitement d'Image

- Keypoints
- Descripteurs
- Vectorisation
- Réduction de Dimension
- **Clustering**
- Classification

SpectralClustering

Nb. de Clusters = Nb. de Catégories

Score de RAND ajusté
0.045

Traitement d'Image

- Keypoints
- Descripteurs
- Vectorisation
- Réduction de Dimension
- Clustering
- **Classification**

Classification

Modèles Testés

- KNeighborsClassifier() — **Accuracy**: 0.43
- SVC() — **Accuracy**: 0.54

un moteur de classification basé sur les descripteurs SIFT de l'image du produit **beaucoup plus difficile**

Examples des Images


Computers


Home Furnishing


Home Decor & Festive Needs


Watches


Beauty and Personal Care


Watches


Home Furnishing


Beauty and Personal Care


Home Decor & Festive Needs


Home Furnishing


Baby Care


Watches


Computers


Home Furnishing

Computers


Computers


Kitchen & Dining


Home Furnishing

Résumé de Traitement d'Image

Les Descripteurs SIFT ne sont pas adaptés

- Pas de clusters évidents
- Pas de regroupement en fonction des catégories
- Des premiers classifications donnent une accuracy faible: 0.43-0.54

Un Moteur de Classification n'est pas évident

- Tester autre algorithmes de reconnaissance d'image
  - CNN Transfer Learning

Etude de Faisabilité
d'un Moteur de Classification

Résumé

Un moteur de classification basée sur

- Description --> **faisable**
  - Dictionnaire customisé / adapté
- Image --> **pas évident !**
  - CNN Transfer learning (?)

Faisable

- Combiner Features
- Méthodes Ensemblistes