Classifier automatiquement des biens de consommation

03/02/2022 - Parcours Data Scientist

Sébastien Bourgeois

Sommaire

1. Problématique & dataset
2. Prétraitements
3. Résultats du clustering
4. Approche supervisée
5. Conclusion

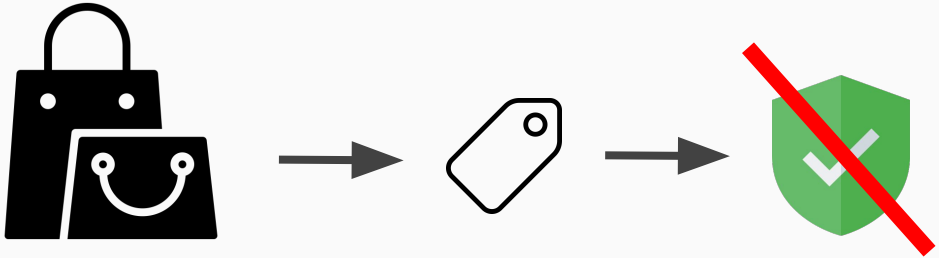
1. Problématique & dataset

Problématique

Les icônes de cette slide ont été conçues par Freepik, TravisAvery, Dreamstale et "I Wayan Wika" du site Flaticon.com

1. Problématique & dataset

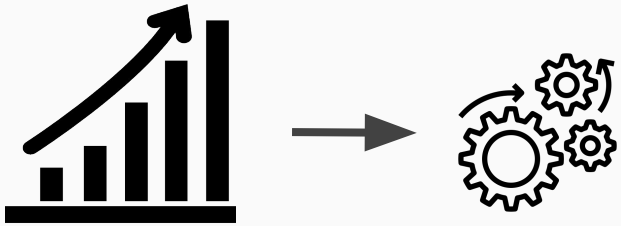
Problématique



Les icônes de cette slide ont été conçues par berkahicon, Pixel perfect et "I Wayan Wika" du site Flaticon.com

1. Problématique & dataset

Problématique



Les icônes de cette slide ont été conçues par Freepik et Becris du site Flaticon.com

1. Problématique & dataset

Dataset

```
<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
```

1. Problématique & dataset

Dataset

```
<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
```

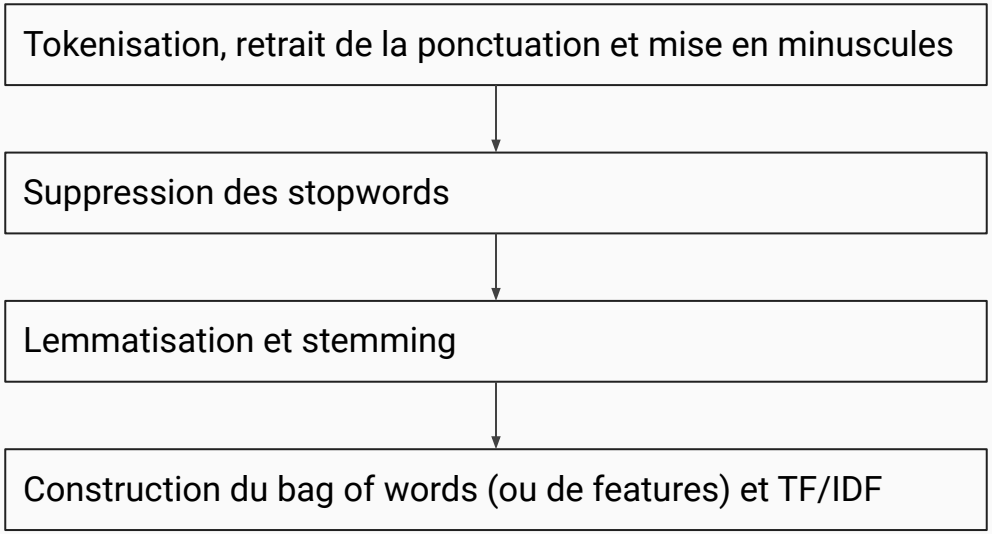
1. Problématique & dataset

Dataset

```
<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
```

2. Prétraitements

Descriptions



2. Prétraitements

Descriptions

Flipkart.com: Buy Denver RO,Black Code Gift Set Combo Set online only for Rs. 355 from Flipkart.com. Only Genuine Products. 30 Day Replacement Guarantee. Free Shipping. Cash On Delivery!

tokenisation

[flipkart, com, buy, denver, ro, black, code, gift, set, combo, set, online, only, for, rs, from, flipkart, com, only, genuine, products, day, replacement, guarantee, free, shipping, cash, on, delivery]

2. Prétraitements

Descriptions

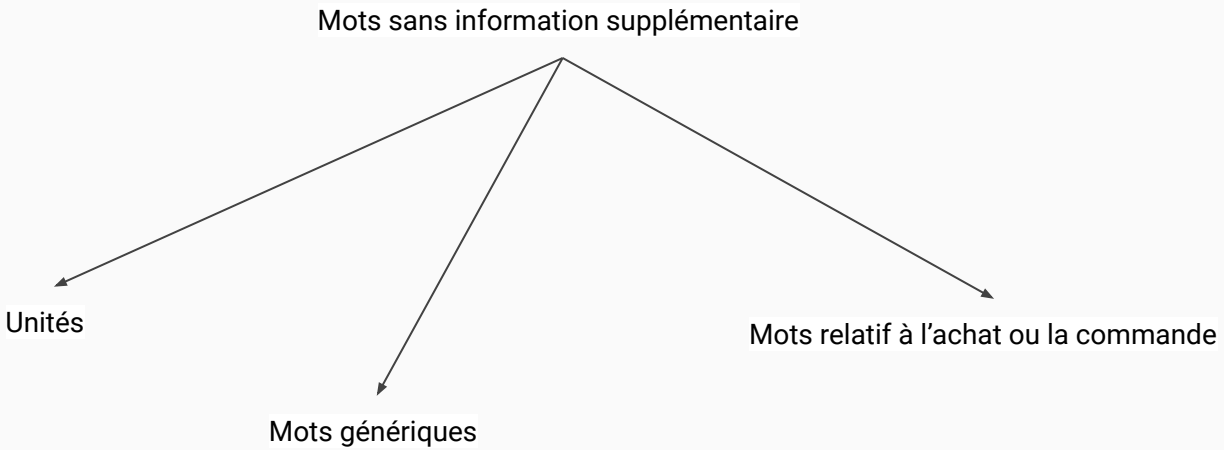
[flipkart, com, buy, denver, ro, black, code, gift, set, combo, set, online, only, for, rs, from, flipkart, com, only, genuine, products, day, replacement, guarantee, free, shipping, cash, on, delivery]

retrait des stopwords classiques

[flipkart, com, buy, denver, ro, black, code, gift, set, combo, set, online, ~~only, for~~, rs, ~~from~~, flipkart, com, ~~only~~, genuine, products, day, replacement, guarantee, free, shipping, cash, ~~on~~, delivery]

2. Prétraitements

Descriptions



2. Prétraitements

Descriptions

[flipkart, com, buy, denver, ro, black, code, gift, set, combo, set, online, ~~only, for~~, rs, ~~from~~, flipkart, com, ~~only~~, genuine, products, day, replacement, guarantee, free, shipping, cash, ~~on~~, delivery]

retrait des mots sans info

[~~flipkart, com, buy~~, denver, ro, black, code, gift, ~~set~~, combo, ~~set, online~~, ~~only, for~~, ~~rs, from, flipkart, com~~, ~~only~~, ~~genuine, products, day, replacement, guarantee, free, shipping, cash, on, delivery~~]

[denver, ro, black, code, gift, combo]

2. Prétraitements

Descriptions

[flipkart, com, buy, denver, ro, black, code, gift, set, combo, set, online, ~~only, for~~, rs, ~~from~~, flipkart, com, ~~only~~, genuine, products, day, replacement, guarantee, free, shipping, cash, ~~on~~, delivery]

retrait des mots sans info

[~~flipkart, com, buy~~, denver, ro, black, code, gift, ~~set~~, combo, ~~set, online~~, ~~only, for~~, ~~rs, from, flipkart, com~~, ~~only~~, ~~genuine, products, day, replacement, guarantee, free, shipping, cash, on, delivery~~]

[denver, ro, black, code, gift, combo]

2. Prétraitements

Descriptions

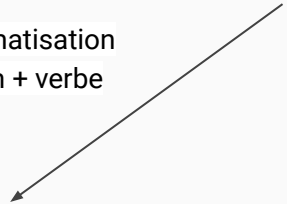
[denver, ro, black, code, gift, combo]

2. Prétraitements

Descriptions

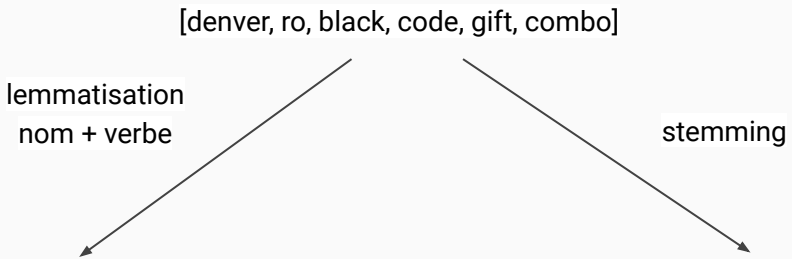
[denver, ro, black, code, gift, combo]

lemmatisation
nom + verbe



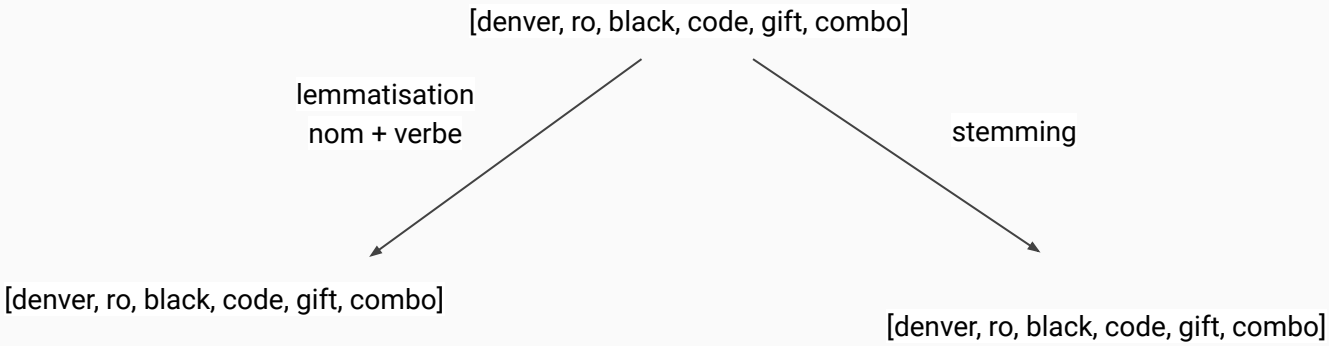
2. Prétraitements

Descriptions



2. Prétraitements

Descriptions



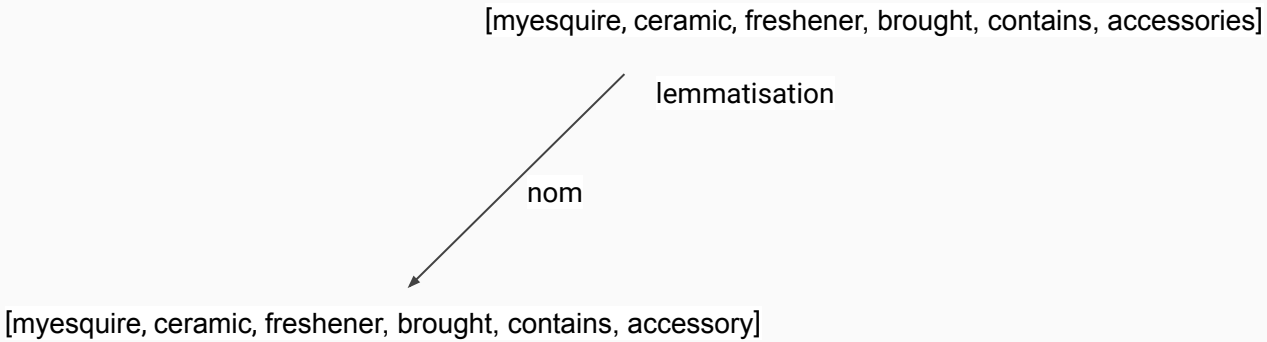
2. Prétraitements

Descriptions

[myesquire, ceramic, freshener, brought, contains, accessories]

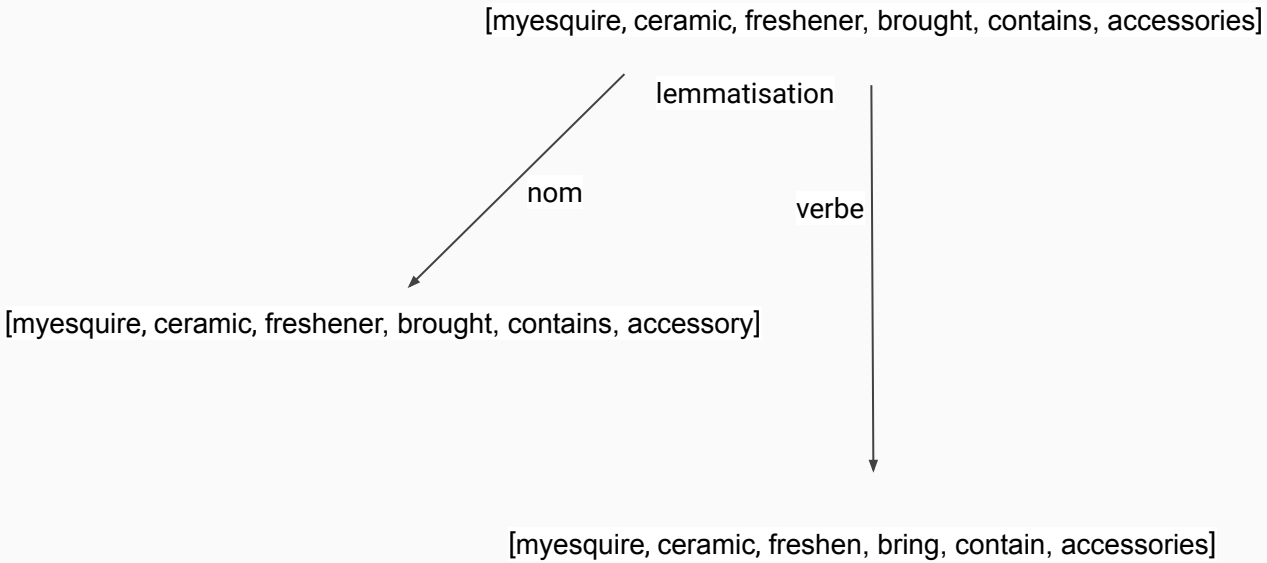
2. Prétraitements

Descriptions



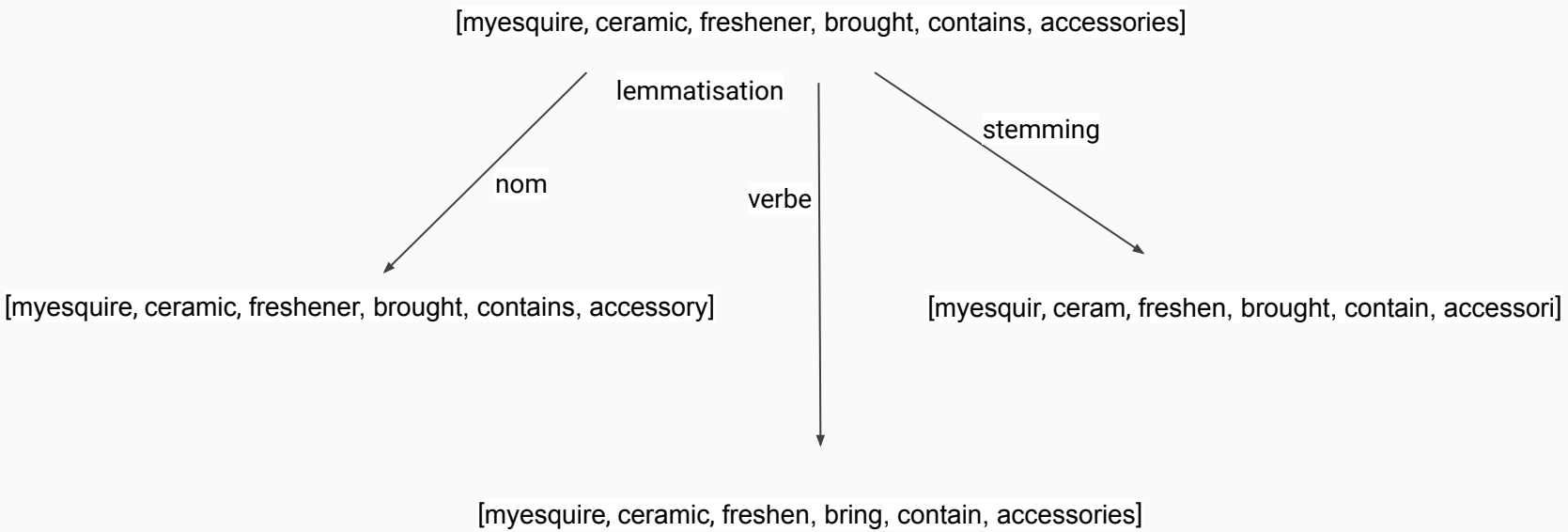
2. Prétraitements

Descriptions



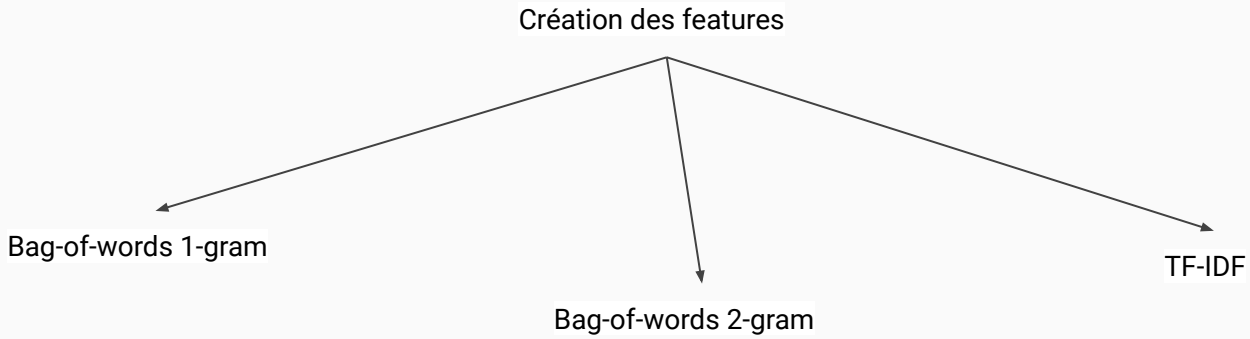
2. Prétraitements

Descriptions



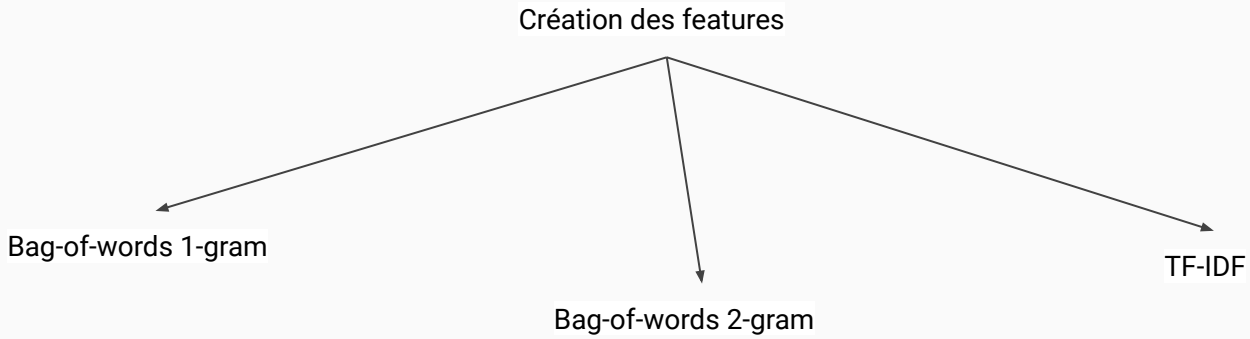
2. Prétraitements

Descriptions



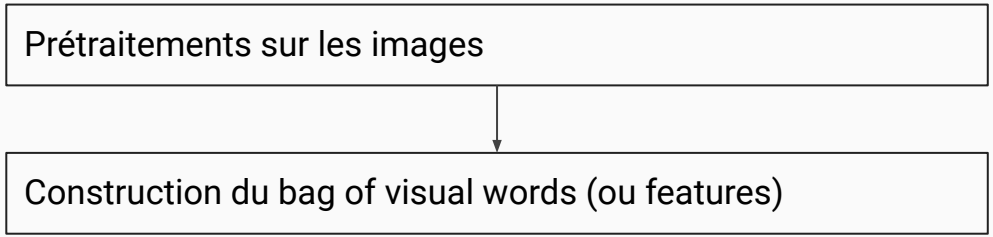
2. Prétraitements

Descriptions



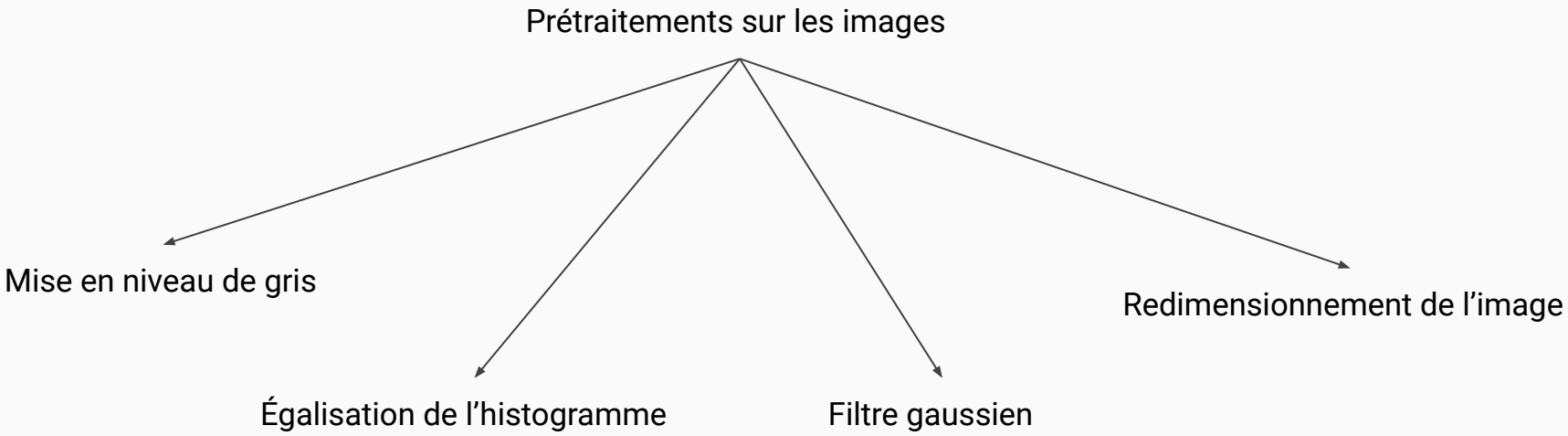
2. Prétraitements

Images



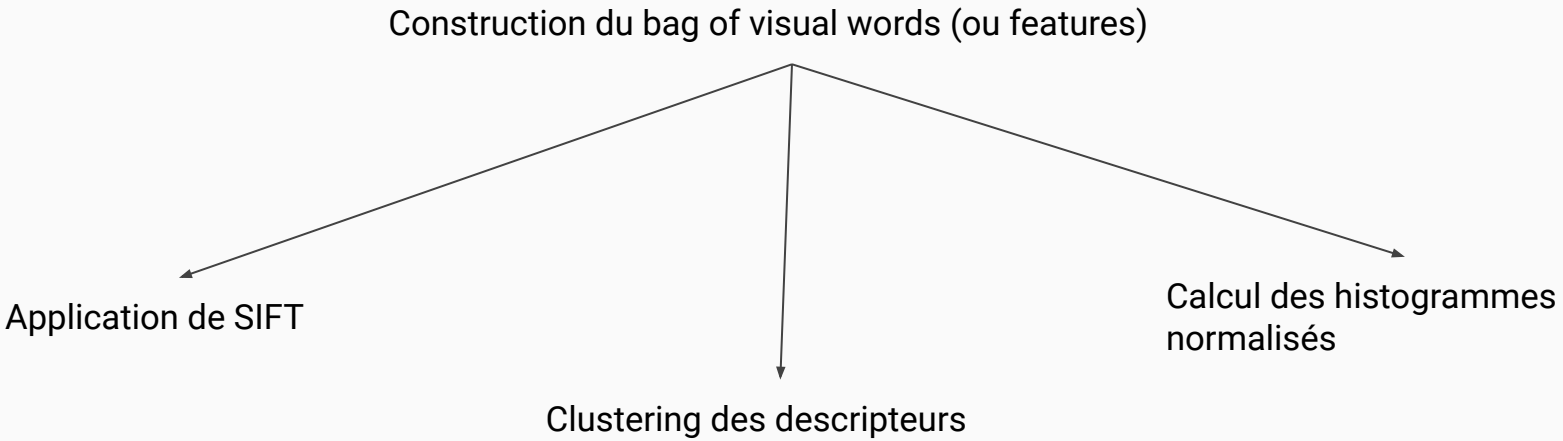
2. Prétraitements

Images



2. Prétraitements

Images



3. Résultat du clustering

Descriptions

LDA -> bag-of-words 1-gram & lemmatisation

```
Catégorie 0:
dohar dark cell hp dv pavilion bag oil bluetooth battery
Catégorie 1:
watch analog men great discount women dial strap water bowl
Catégorie 2:
baby cotton color box sales number fabric wall design print
Catégorie 3:
combo oil best kadhai face soap cream kit beauty care
Catégorie 4:
showpiece best towel paper bath bottle green quilt light brass
Catégorie 5:
mug design ceramic gift make perfect coffee eyelet material love
Catégorie 6:
laptop skin warranty shape print pad usb mouse multicolor quality
```

3. Résultat du clustering

Descriptions

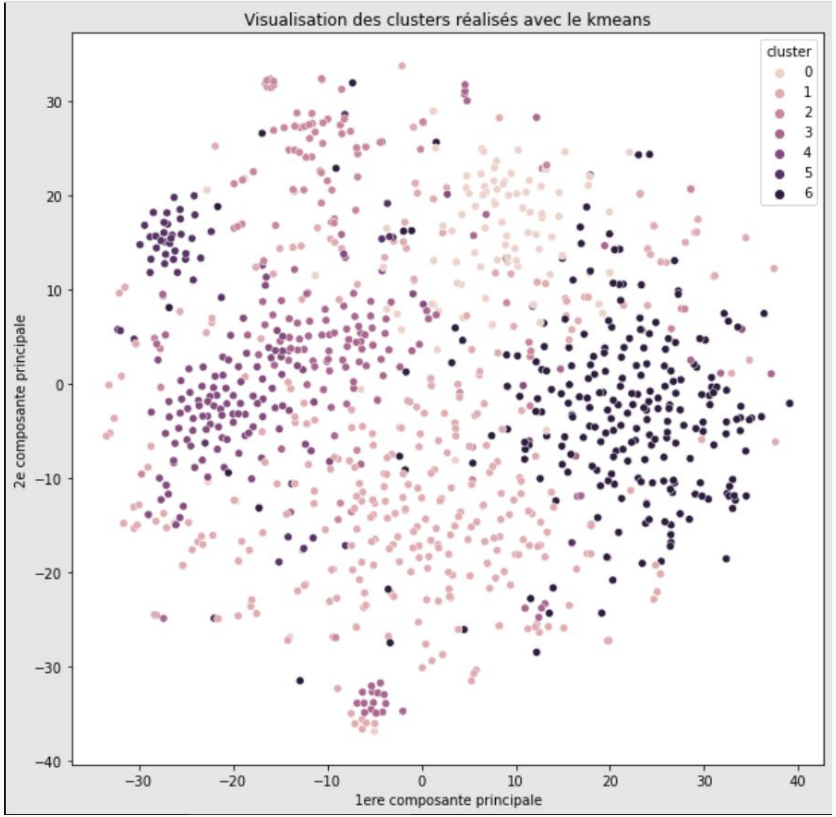
LDA -> bag-of-words 1-gram & lemmatisation

Mettre visualisation

3. Résultat du clustering

Images

KMeans -> 7 clusters



4. Approche supervisée

Descriptions

Modèle	Précision
Méthode naïve	12%
Régression logistique	89%

Résultat sur le jeu de test : 94% de produits bien catégorisés

4. Approche supervisée

Images

Modèle	Précision
Méthode naïve	14,61%
k-NN	25,2%
Régression logistique	46,75%
SVM	46,76%

Résultat sur le jeu de test : 51% de produits bien catégorisés

4. Approche supervisée

Mixte : descriptions & images

Réduction dimension : 4,8K variables -> 163 CP ~70% variance expliquée

Modèle	Précision
Méthode naïve	15%
k-NN	70%
Régression logistique	79%
SVM	73%

Résultat sur le jeu de test : 82% de produits bien catégorisés

5. Conclusion

Descriptions

Images

Approche mixte

Faisabilité

Les icônes de cette slide ont été conçues par Freepik, Pixel Buddha du site Flaticon.com

Questions/Réponses

Fin.