Projet_06 : Classifiez automatiquement des biens de consommation

Jérôme LE GAL
Etudiant OpenClassRooms – parcours Data Scientist
Le 27/10/2024

Contexte et objectifs

- Mission pour “place de marché”
- Faisabilité de classification automatique
- Classification supervisée
- Test d’une API

Description du jeu de données

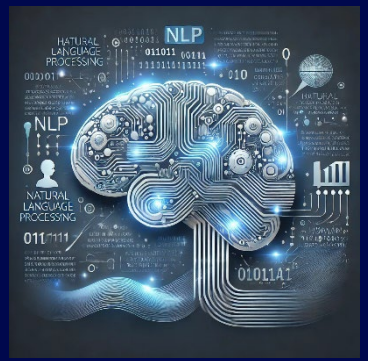
- Fichier csv :
  - Id
  - Nom de l'article
  - Description
  - …
- Images

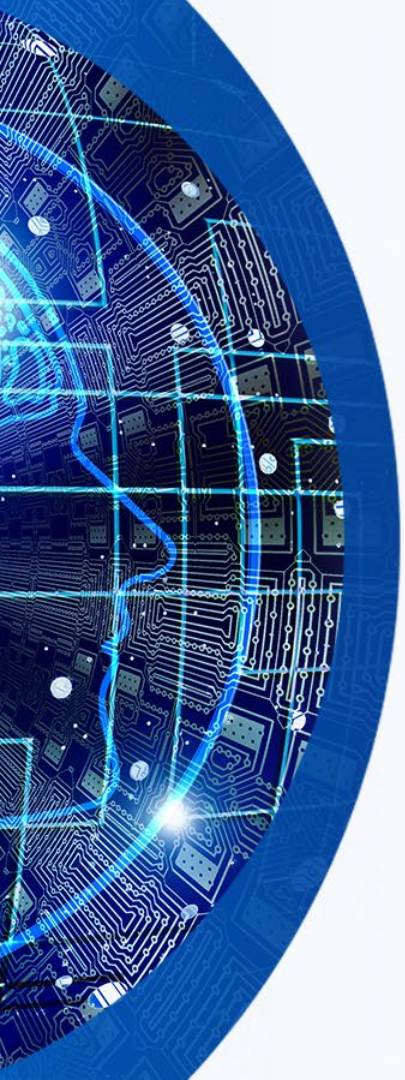
Quantité : 1050

Etude de faisabilité : Natural Language Process

NLP :

- Preprocessing
- Approches basiques
- Approches avancées (Deep Learning)





Faisabilité NLP : Preprocessing

➔ Fonction preprocessing

- Corpus : 47651 mots
- Vocabulaire : 3940 mots

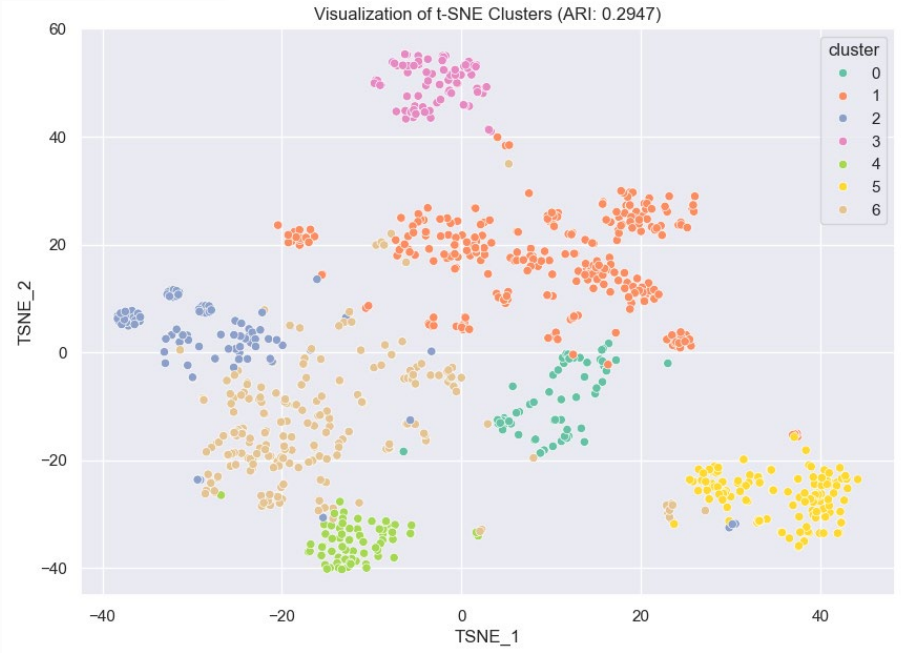
Faisabilité NLP : Approches basiques

- Bag of Words (fréquence de chaque mots) :
  - Score ARI : 0,423

- TF-IDF (évaluation de l'importance de chaque mots) :
  - Score ARI : 0,426

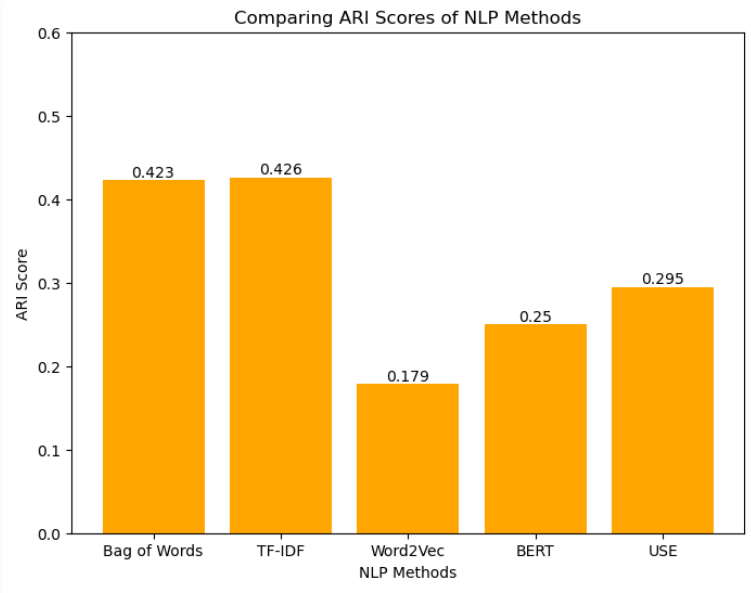
Faisabilité NLP : Approches avancées

- Word2Vec : Embeddings dans un Espace vectoriel
  - ➢ Score ARI : 0,179
- BERT : Transformer qui comprend le contexte bidirectionnel des phrases
  - ➢ Score ARI : 0,25
- USE : capte le sens des phrases complètes
  - ➢ Score ARI : 0,295



Faisabilité NLP : Comparaison des performances

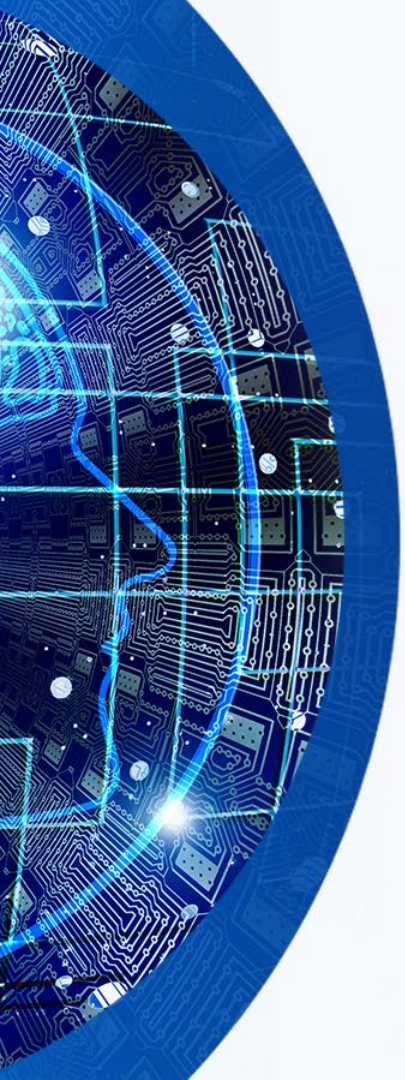
- **TF-IDF obtient le meilleur score ARI (0,426),** surpassant les modèles avancés.
- **Raison probable** : Le jeu de données est trop modeste pour que les modèles avancés montrent tout leur potentiel.
- **Perspectives** : Plus de données et un fine-tuning des modèles avancés pourraient inverser cette tendance.



Etude de faisabilité : Computer Vision

Computer Vision :

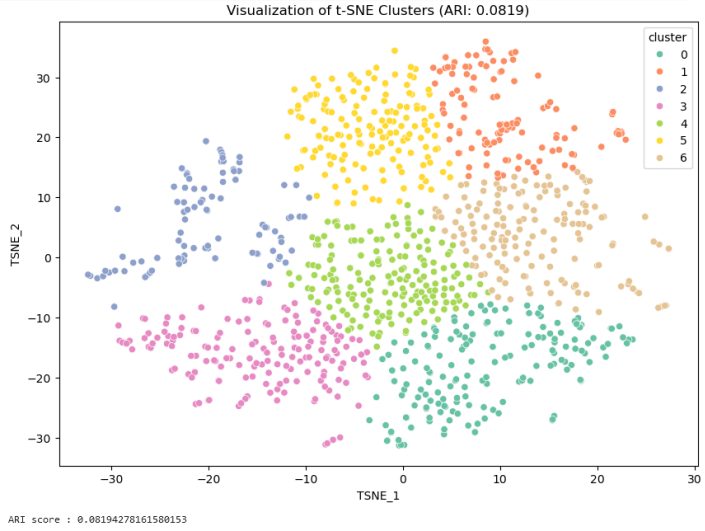
- Approche basique :
  - SIFT
- Approche avancée :
  - CNN

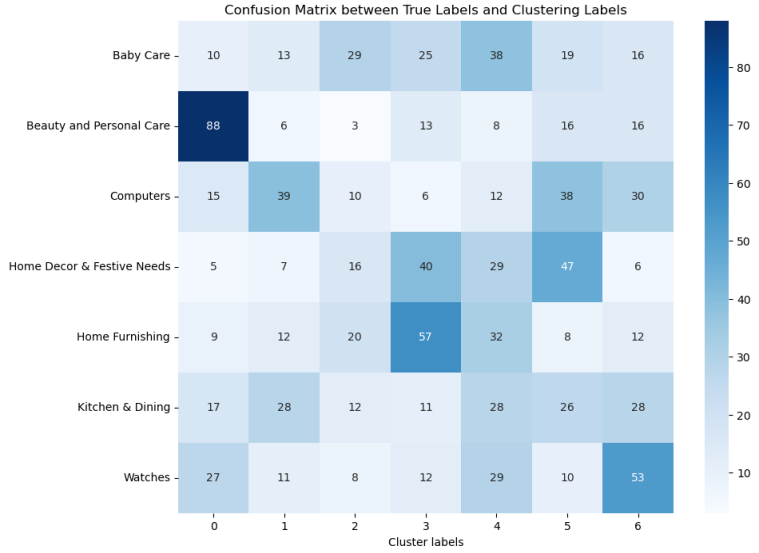


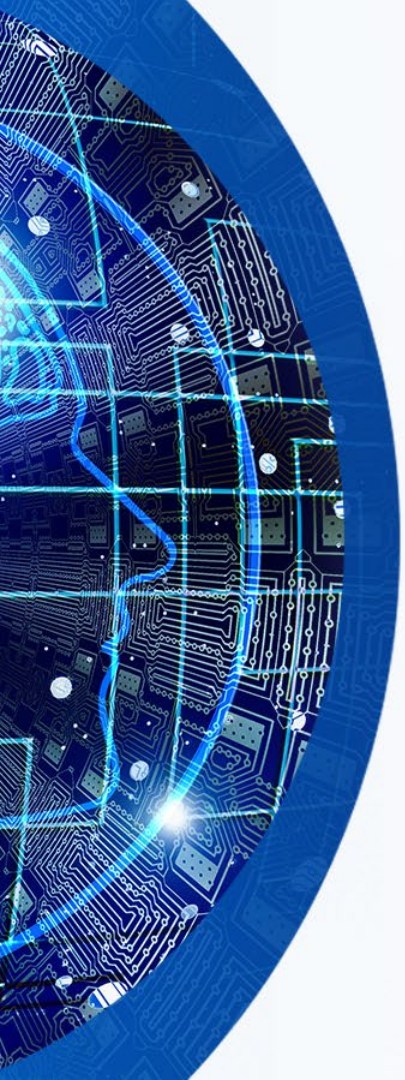
Faisabilité Computer Vision : Approche basique - SIFT

1. Preprocessing
2. Extraction des descripteurs – Shape : (514634, 128)
3. Création des Bags of Visual Words - Kmeans avec 717 clusters de descripteurs
4. Création des histogrammes - Shape : (1050, 717)
5. Réduction de dimension (ACP 99 % → 456 dimensions)
6. Classification Kmeans
7. Réduction de dimension T-SNE à 2D
8. Affichage et score ARI

Faisabilité Computer Vision : Approche basique - SIFT



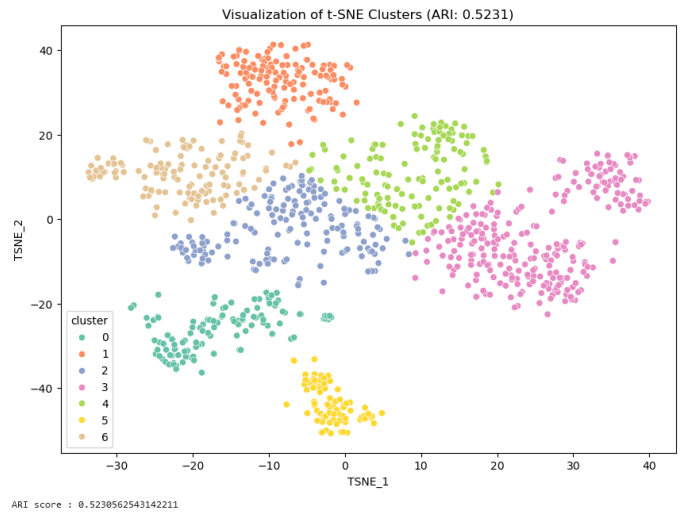


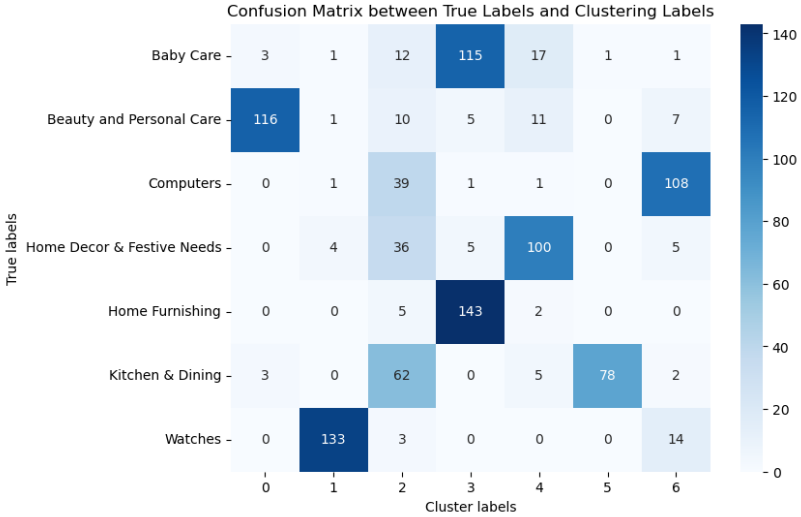


Faisabilité Computer Vision : Approche avancée - CNN

1. Preprocessing
2. Extraction des features (vgg-16 pré-entrainé)
3. Réduction de dimension (ACP 99% → 741 dimensions)
4. Classification Kmeans
5. Réduction de dimension T-SNE à 2D
6. Affichage et score ARI

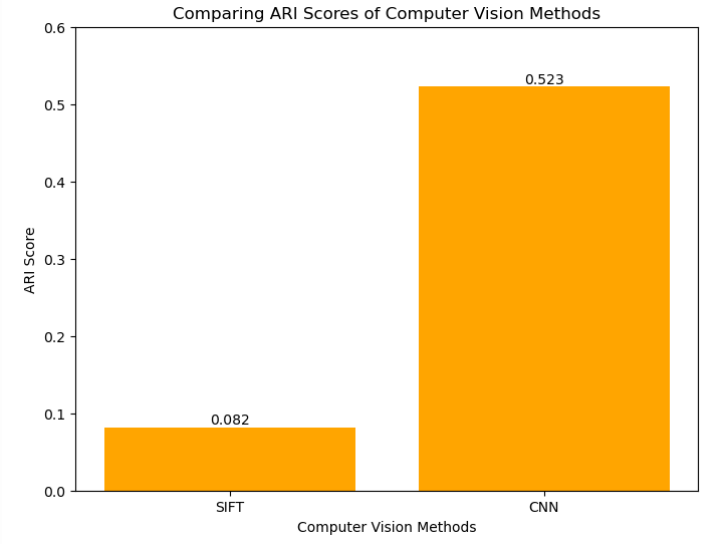
Faisabilité Computer Vision : Approche avancée - CNN



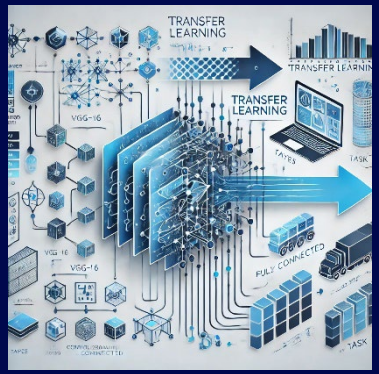


Faisabilité Computer Vision : Comparaison des méthodes

- CNN obtient le meilleur score
- Score honorable sans optimisations

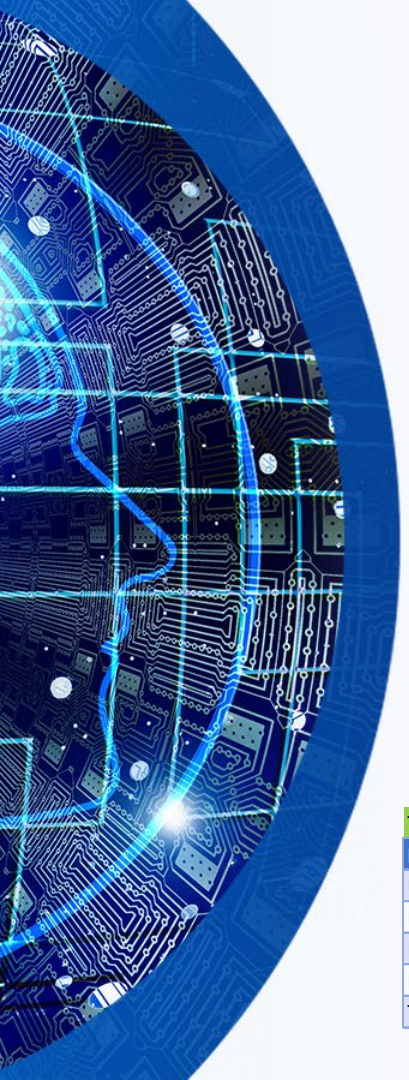


Classification supervisée



Transfer Learning

- Modèle vgg-16 :
  - Recherche d'hyperparamètres optimaux
- Preprocessing
- Entrainements :
  - Dataset de base
  - Dataset + Data Augmentation



Classification supervisée : Transfer Learning - CNN

A - **SANS** Data Augmentation

B - **AVEC** Data Augmentation

1. Preprocessing
2. Split des données (Train, Test, Validation)
3. Tuner Keras
4. Entrainement du modèle
5. Évaluation des performances

Tuner in 90 trials		
Parameters	**Basic dataset**	**Dataset + DA**
Units in dense layer	384	416
Dropout rate	0,2	0,5
Learning rate	0,001	0,001
Best epochs	14	4
Trainning duration	37min with GPU	5h with CPU

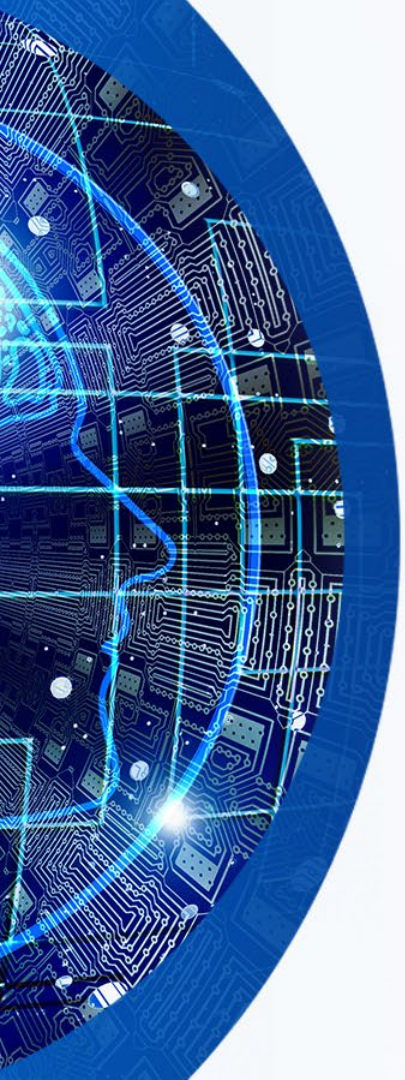
Method	Train Loss	Train Accuracy	-	Test Loss	Test Accuracy
Basic Dataset	0.189	0.959	-	0.811	0.829
With Data Augmentation	0.966	0.794	-	1.36	0.776

Collecte de données par API

Test d’une API - EDAMAM

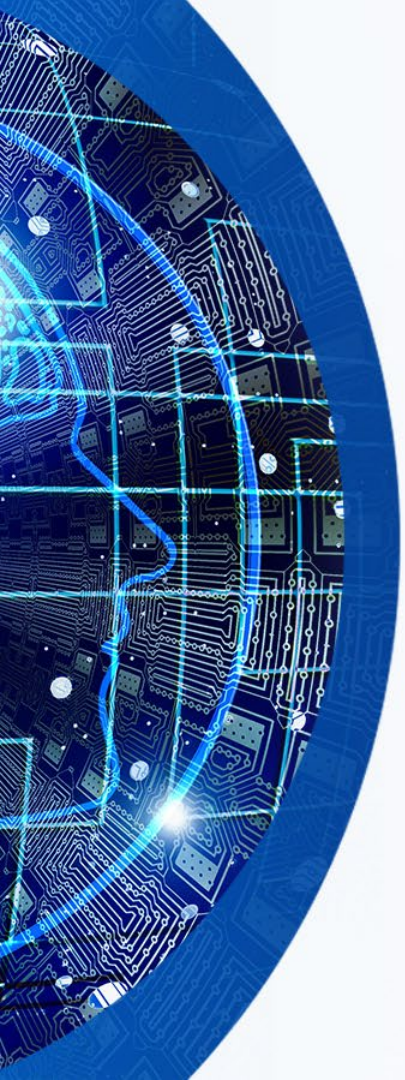
- Critères RGPD
- Clé API
- Requête
- Mise en forme résultats

EDAMAM



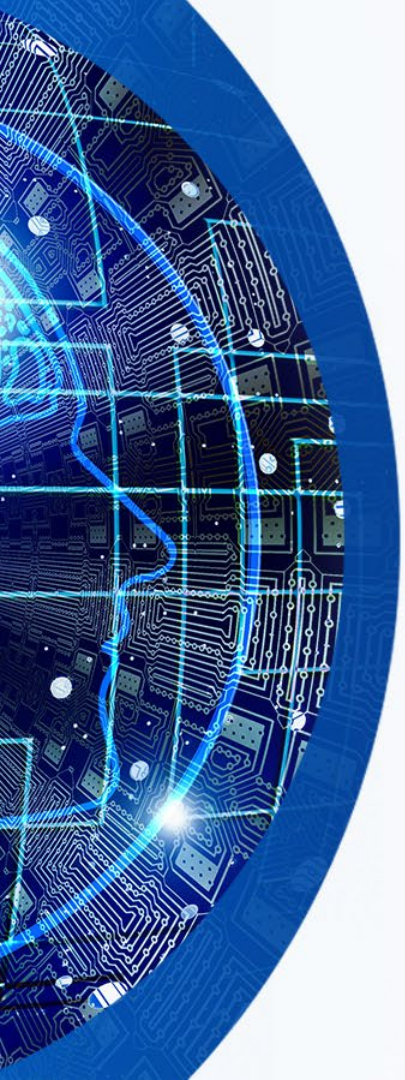
Collecte de données par API : Critères RGPD

1. Licéité, loyauté, transparence



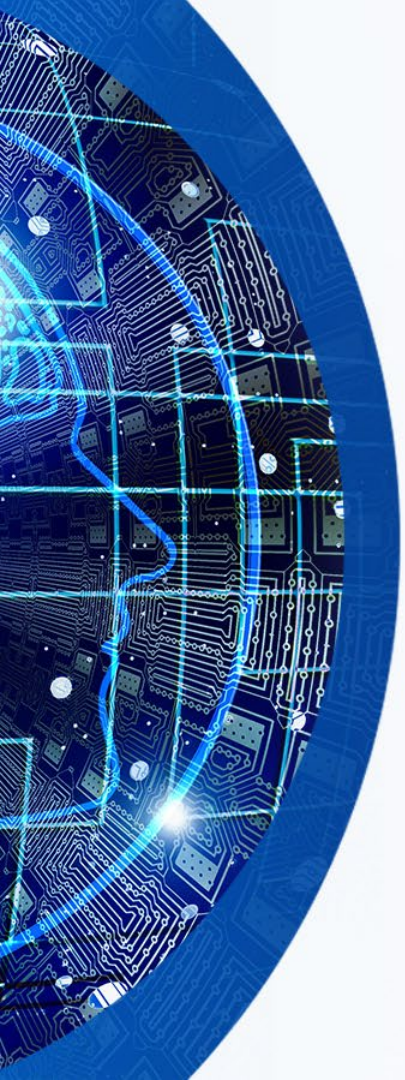
Collecte de données par API : Critères RGPD

1. Licéité, loyauté, transparence
2. Limitation des finalités



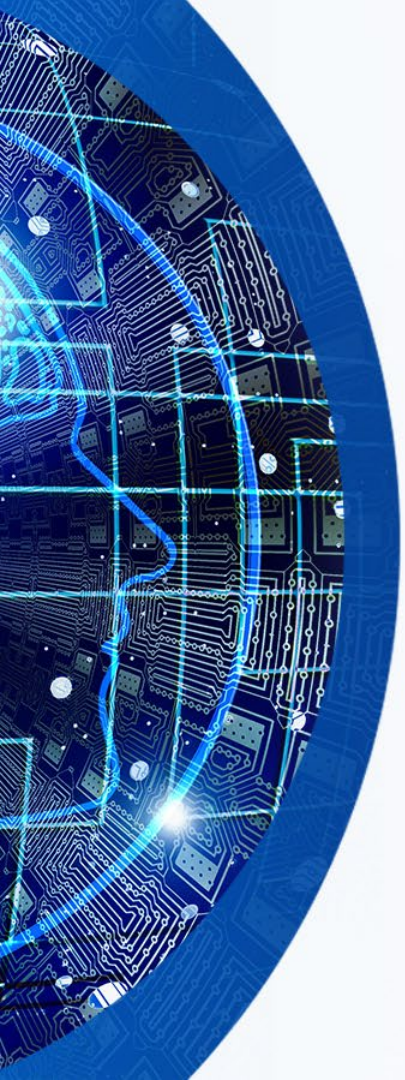
Collecte de données par API : Critères RGPD

1. Licéité, loyauté, transparence
2. Limitation des finalités
3. Minimisation des données



Collecte de données par API : Critères RGPD

1. Licéité, loyauté, transparence
2. Limitation des finalités
3. Minimisation des données
4. Exactitude des données



Collecte de données par API : Critères RGPD

1. Licéité, loyauté, transparence
2. Limitation des finalités
3. Minimisation des données
4. Exactitude des données
5. Limitation de la conservation

Collecte de données par API :

- Clé API
- Requête :
  - app_id=********
  - app_key=**********
  - ingr=champagne
  - nutrition-type=logging

Request URL

```
https://api.edamam.com/api/food-database/v2/parser?app_id=a586f207&app_key=b28ac44e9a58f42fdbacc41603e1f053&ingr=champagne&nutrition-type=logging
```

- Réception Fichier JSON
- Mise en forme des données

➔ Export d'un fichier CSV

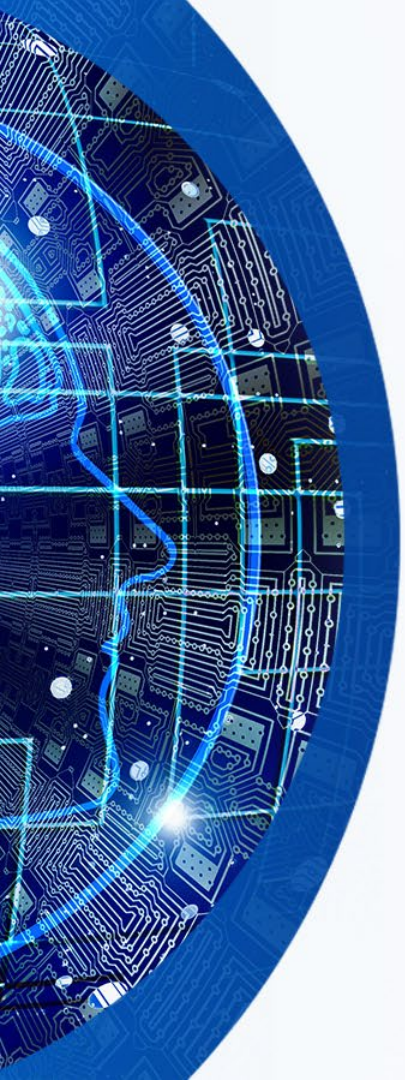
➔ Fonction

Conclusion

- **Contexte** : Projet de classification automatique des biens de consommation à partir de données textuelles et visuelles.
- **Réalisations principales** : Exploration d'approches avancées en NLP (Word2Vec, BERT, USE). Mise en place de modèles de vision par ordinateur (SIFT, CNN).
- **Résultats significatifs** : NLP (BoW – T-SNE) atteint un ARI de 0.426, et la classification d'images par CNN obtient une précision de 77,5 %.
- **Insight majeur** : L'intégration du NLP et de la vision par ordinateur offre des perspectives prometteuses pour la classification automatique.
- **Perspectives** : Des améliorations sont possibles pour optimiser la précision et l'efficacité du système.

Perspectives

- **Augmentation du jeu de données** : Ajouter plus de données textuelles et d'images pour améliorer la généralisation des modèles grâce à plus de variabilité.
- **Fine-tuning et ressources** : Utiliser des ressources plus puissantes (GPU/TPU) afin d'entrainer les couches du modèle sur nos données pour affiner les poids et ainsi réduire les erreurs de classification.
- **Combinaison NLP et Vision par ordinateur** : Développer un modèle multimodal combinant les informations textuelles (description des produits) et visuelles (images) pour une classification plus précise.



Questions ?