Extraction des listes d'ingrédients depuis les fiches techniques de produits alimentaires

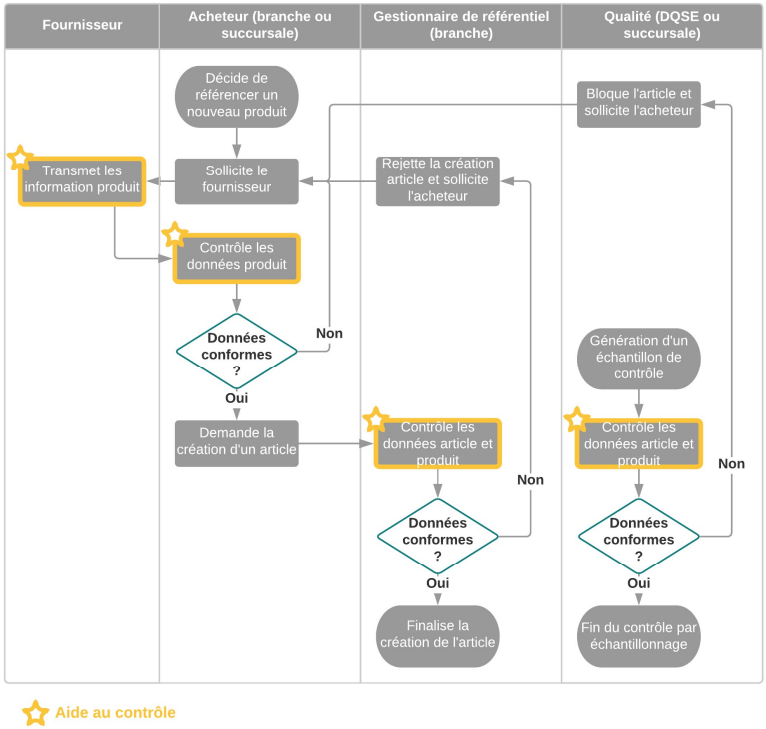
DSSP 14 – Pierre MASSE

Juin 2020

SOMMAIRE

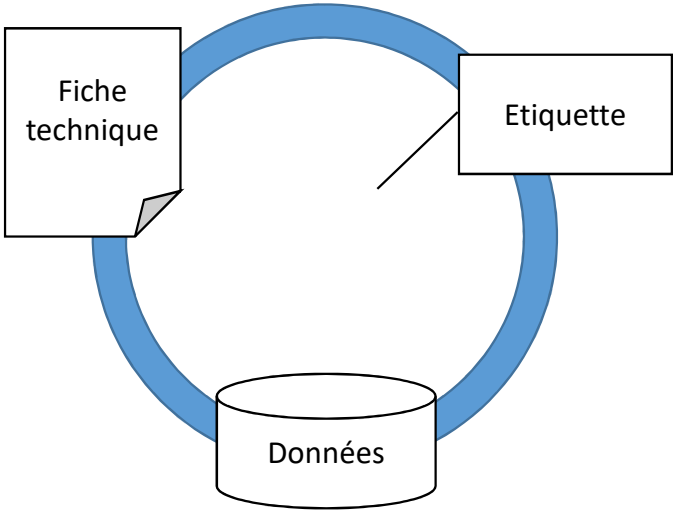
- Exploiter le contenu de documents pour la qualité des données produit, pourquoi ?
- Traitement du langage ou traitement de l’image ?
- Le fonctionnement du modèle
- Meilleurs paramètres et résultats obtenus
- Les apports de la formation qui ont rendu ce projet possible

La qualité de l'information produit passe par l'exploitation du contenu de documents



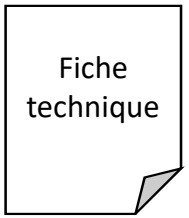
De multiples étapes de contrôle ont lieu dans le processus de gestion de l'information produit au sein du Groupe Pomona.

Pour l'essentiel, ces contrôles visent à s'assurer de la cohérence entre les données qui ont été transmises par les fabricants des produits, et les documents qu'ils ont transmis par ailleurs.



Ces contrôles consistent uniquement à prendre connaissance du contenu des documents (fiches techniques et étiquettes), et à évaluer la cohérence avec les données, avec peu d'interprétation.

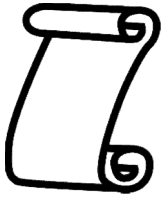
Plusieurs méthodes d'extraction de connaissance depuis les documents peuvent être mises en en œuvre



Traitement de l'image

- Multiples textes
- Avec leurs localisations relatives

✓ Nombreuses fonctionnalités accessibles
✗ Forte complexité à la mise en œuvre



Traitement du langage

- Long texte unique
- mais possibilité de découper selon des règles de gestion

Classification de chaque morceau de texte candidat (ex : allégation nutritionnelle, dénomination réglementaire, …)

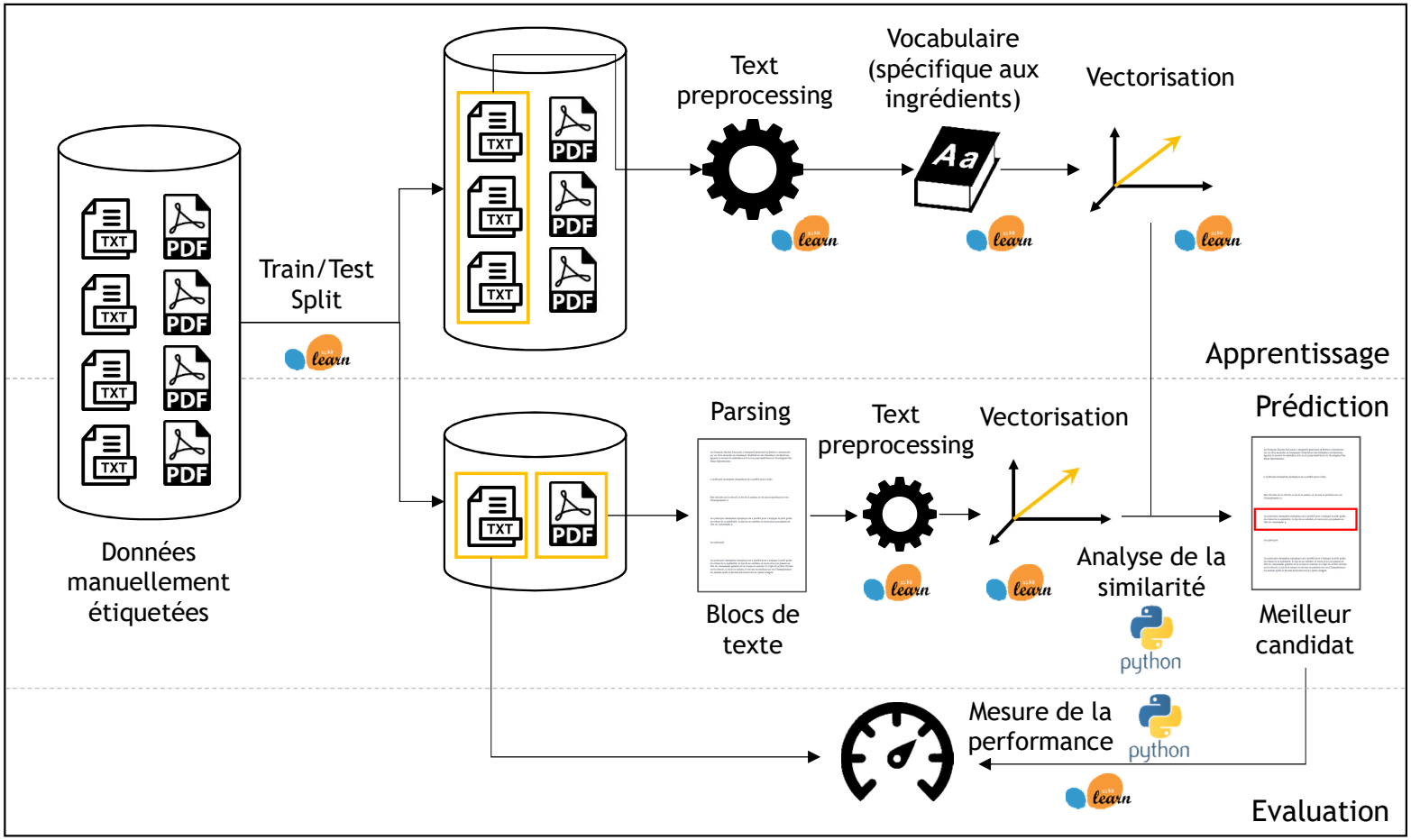
✓ Gestion de plusieurs types d'information simultanément
✗ Nécessite un grand nombre de données étiquetées actuellement non disponibles

Requêtage parmi les textes candidats en fonction d'un vocabulaire

✓ Mise en œuvre simplifiée sur certaines données (listes d'ingrédients)
✗ N'adresse qu'un seul cas d'usage

Essentiellement pour des raisons de faisabilité technique, il a été décidé d'offrir la possibilité d'extraire **les listes d'ingrédients** via des fonctionnalités de **traitement du langage**, en utilisant des techniques de **requêtage de l'information.**

L'utilisation de diverses briques logicielles et bibliothèques de Machine Learning permet d'extraire les listes d'ingrédients des fiches techniques



Le modèle a été entraîné et ajusté sur des données étiquetées manuellement.

L'utilisation de fonctionnalités d'Optical Caracter Recognition n'a pas été nécessaire, les outils de pdf mining étant largement suffisants.

La performance du modèle a été mesurée en utilisant la **distance d'édition de Levenshtein**

Les principaux paramètres sont le **preprocessing, le mode de vectorisation des textes** (blocs de texte et vocabulaire cible) et le **calcul de similarité**.

Le modèle a pu être optimisé via des grid searches, et les résultats obtenus sont très prometteurs

Paramètres retenus

Preprocessing

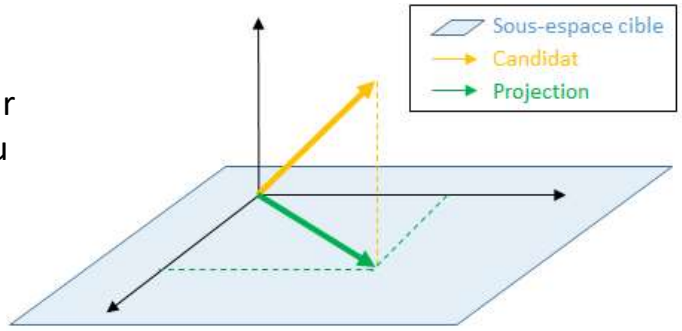
- Retrait des accents
- Mise en minuscules
- Retrait des stopwords

Vectorisation

- Bag Of Word, sans calcul d'embedding
- Représentation binaire (présence / absence du mot)
- Sans calcul de l'inverse document frequency
- Avec les monogrammes, brigrammes et trigrammes

Calcul de similarité

- Rapport des normes du vecteur dans l'espace de départ et de sa projection sur le sous ensemble généré par les mots du vocabulaire des ingrédients
- Norme dans l'espace initial : L4
- Norme dans le sous-espace de projection : L3



Résultats

Similarité de Levenshtein	
Cross-validation sur train set	Evaluation sur test set
63.31% ± 3.83%	**67.18%**

Accuracy	
Cross-validation sur train set	Evaluation sur test set
24.25% ± 3.07%	**27.00%**

La mise en place de ce modèle n'aurait pu être possible sans les compétences acquises lors de la formation

1 **Extraction de features depuis des textes**

Ce qui a rendu possible la construction de ce modèle est la compréhension des techniques utilisées dans le traitement du langage en Machine Learning.

La représentation « Bag of Words » qui permet ensuite de passer dans un univers vectoriel des textes, puis d'appliquer d'autres techniques plus élaborées (embeddings, calculs de normes et de similarité, …) a été centrale tout au long de ce projet.

2 **Preprocessing en traitement du langage**

Un des facteurs qui a eu le plus d'impact sur la performance du modèle est l'identification et le retrait des stopwords.

Les techniques de base (gestion des accents, de la casse, …) et la connaissance des outils associés a clairement permis d'améliorer les résultats.

3 **Evaluation et amélioration de la performance**

Les principes de base (train/test split, …) ont permis de mettre en œuvre une méthodologie robuste d'évaluation de la performance.

Les techniques d'ajustement des hyperparamètres (grid search), et la connaissance des outils mis à disposition dans la bibliothèque scikit-learn ont été des facteurs déterminants dans les résultats obtenus.

Je tire une grande satisfaction des résultats obtenus dans le cadre de ce projet. La conjoncture n'est pas favorable pour le lancement de projets novateurs, mais ce modèle pourrait clairement être mis en production pour le Groupe Pomona.

Questions ?

Merci pour votre attention