Test technique Data Scientist(e)

Objectif :

Développer un système de classification automatique permettant d'associer une désignation commerciale à une catégorie de référence (FoodEx).

Vous devrez :

1. Explorer et nettoyer les données.
2. Mettre en œuvre différentes approches (règles, NLP, ML/DL).
3. Comparer les résultats de vos modèles (analyse comparative).
4. Fournir un projet reproductible (Docker + GitHub).

Données à télécharger

À chaque requête envoyée à l'API, vous devez passer les paramètres :

email=email_candidat* et *code=identifiant_candidat

- Référentiel FoodEx
  Endpoint :
  https://www.api.chemosenstools.com/api/recrutement/teststechniques/data/download/train
  Output : *referentiel_foodex.xlsx*
  Contenu :
  - Désignations commerciales
  - Catégories de référence
- Jeu de test
  Endpoint :
  https://www.api.chemosenstools.com/api/recrutement/teststechniques/data/download/test
  Output : *test.json*
  Contenu :
  - Désignations commerciales à classer

Travail attendu

1. Prétraitement & nettoyage
   - Normalisation des désignations (minuscule, accents, stopwords, etc.).
   - Extraction de mots-clés pertinents.
   - Application éventuelle d'expression régulières.
2. Classification
   - Étape 1 : Pré-sélection des catégories candidates à partir des mots-clés.

    - Étape 2 : Application d'un ou plusieurs modèles NLP / ML / DL pour affiner la classification.
    - Vous pouvez aussi proposer une approche directe (modèle unique) si vous le justifiez.
3. Analyse comparative
    - Implémentez au moins deux approches différentes (ex. : règles + embeddings, ou TF-IDF + SVM, ou LLM, etc.).
    - Comparez leurs performances sur le jeu de test (test.json).
    - Présentez les résultats de manière synthétique (un DataFrame avec désignation, catégorie prédite par chaque modèle, catégorie attendue si applicable).

Critères d'évaluations

- Qualité du code : propreté, structuration (POO, modules), respect des conventions (PEP8).
- Maîtrise NLP : pertinence du nettoyage, enrichissement des features, choix des modèles.
- Robustesse : gestion des désignations fantaisistes, fautes, synonymes, ambiguïtés.
- Reproductibilité : projet exécutable sans friction (Docker, dépendances claires).
- Clarté de l'analyse : résultats comparés et justifiés.

Livrables

1. Manifest de suivi (manifest.txt)
    - Contiendra vos notes de travail : difficultés rencontrées, intuitions, justification des choix techniques.
    - Renommez le fichier de la manière suivante : *prenom.nom.txt*
    - À uploader via l'endpoint : https://www.api.chemosenstools.com/api/recrutement/teststechniques/data/upload
    - Utilisez la méthode HTTP appropriée.
    - Vous recevrez un mail de confirmation après upload.
2. Projet GitHub
    - Nom du repo : test_technique_chemosens
    - Contenu :
        - Code source complet.
        - Dockerfile et instructions de lancement.
        - Tout élément nécessaire à la reproductibilité (requirements.txt, notebook, etc.).

Durée & Déroulé

- Temps imparti : 4 heures.
- À la fin de l'épreuve :
    - Push sur GitHub.
    - Upload du manifest.txt.