Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Une image contenant texte, capture d’écran, logiciel

Description générée automatiquement

Le code possède la fonction train\_and\_classify\_emails qui entraîne un modèle de machine learning pour classifier des emails comme contenant ou non des événements, en utilisant le français comme langue de traitement (via Spacy). Voici une description des fonctionnalités principales de ce code :

1. **Chargement du modèle Spacy** : Le modèle linguistique français fr\_dep\_news\_trf est chargé via Spacy pour le prétraitement des emails.
2. **Vérification des répertoires d'entrainement** : Le code vérifie l'existence des dossiers contenant les emails avec événements (path\_avec\_eve) et sans événements (path\_sans\_eve). S'ils n'existent pas, ils sont créés.
3. **Chargement ou entraînement du modèle** :
   * Si un fichier modèle (par défaut "trained\_model.joblib") existe déjà, le modèle est chargé via joblib.load.
   * Sinon, les emails des dossiers sont chargés, étiquetés (1 pour les emails avec événements, 0 pour ceux sans), et prétraités avec la fonction preprocess\_email.
   * Un pipeline est créé avec un TfidfVectorizer (pour transformer les emails en vecteurs de caractéristiques) et un classificateur SVC (Support Vector Machine) avec un noyau linéaire. Ce modèle est entraîné sur les emails prétraités et les étiquettes.
   * Le modèle entraîné est sauvegardé avec joblib.dump.
4. **Prédiction et évaluation** :
   * Les emails du dossier de test (path\_teste) sont chargés et prétraités. Pour chaque email, le modèle prédit la probabilité qu'il contienne un événement.
   * Si la prédiction est incorrecte, l'utilisateur peut manuellement corriger la classification via une entrée utilisateur (input) dans la console.
5. **Affichage des résultats** :
   * Après la classification, le programme affiche les emails classifiés comme contenant ou non des événements, avec les probabilités associées.

**Fonctionnalités supplémentaires :**

* Le modèle sauvegardé peut être réutilisé sans nécessiter de réentraînement, ce qui optimise le temps d'exécution pour les futures classifications.
* La correction manuelle permet d'améliorer la précision du modèle en ajustant les erreurs de classification faites par l'IA.