Test technique DGFIP (e-contact)

Classifieur de messages

L'objectif de ce test est de présenter un problème typique de classification de textes qui pourrait se rencontrer à la DGFIP.

Le dataset proposé comporte 2 tables :
- une table data_train comportant un ensemble de messages labellisés en deux classes : 1 et 8. Attention, certains messages relèvent simultanément des 2 classes (étiquetés 1:8).
- une table data_test ne contenant qu'une liste de messages.

Le principe est simple : il s'agit de réaliser un classifieur, c'est-à-dire un algorithme permettant de déterminer la classe d'un texte. Ce classifieur pourra être entraîné sur le data_train et ainsi fournir une liste de prédictions de labels des messages du data_test. L'enjeu principal est de comparer des outils classiques de nlp (tf-idf par exemple) avec des techniques plus récentes (embedding, llm).
Nous attacherons une attention particulière au workflow, au preprocessing, aux métriques utilisées ainsi qu'à l'évaluation finale de la pertinence des résultats obtenus.

Pour ce qui est du rendu, nous recommandons un notebook écrit en Python accompagné de commentaires, ainsi qu'un fichier csv contenant les prédictions des labels de la table data_test.

Contact : hermann.woehrel@dgfip.finances.gouv.fr