Contexte du Test Technique

Nous souhaitons développer un modèle capable de prédire la note associée à un avis en se basant uniquement sur le texte de cet avis. Les avis fournis sont des avis déposés par des clients sur des hôtels. L'objectif est de vérifier si la note attribuée par le client correspond bien au contenu de son avis. Cette prédiction nous permettra d'identifier des incohérences potentielles et d'améliorer la fiabilité des évaluations des hôtels.

Structure du projet

Utilisez les fichiers `train.csv` et `test.csv` présent sous l'URL suivante:
https://drive.google.com/file/d/1TrnRqGvOoif7kLgenEwIyf0JlQlheMCx/view?usp=sharing

- `train.csv` contient les colonnes suivantes : `establishment_id`, `review_id`, `review_text`, `global_rate`.
- `test.csv` contient les colonnes suivantes : `establishment_id`, `review_id`, `review_text`.

Entraînez un modèle sur les données d'entraînement.

Utilisez le modèle entraîné pour prédire les notes (`global_rate`) des avis dans le fichier `test.csv`.

Livrable attendu

Projet Python avec une interface en ligne de commande qui permet :

1. De générer un modèle.
2. D'utiliser un modèle pour faire des prédictions sur un dataset de test.

Le fichier CSV de prédictions.

Conclusion

N'hésitez pas à utiliser toutes les bibliothèques externes nécessaires à l'implémentation de votre solution. Nous voulons finalement pouvoir exécuter votre projet localement via (quelques) lignes de commande. Vous pouvez fournir et partager des ressources supplémentaires qui vous ont aidé à résoudre le problème. Vous pouvez également partager vos réflexions sur les améliorations potentielles futures.

Nous évaluerons vos habitudes de codage, vos connaissances sur les étapes habituelles de ML et le processus de réflexion suivi tout au long de l'exercice. Notez que l'objectif n'est PAS la performance du modèle (vous n'aurez pas assez de données), ni la scalabilité, ni la complexité temporelle (pas besoin de cache, de base de données rapide, de cloud computing, de GPU…).

- **Nous ne voulons pas que vous passiez plus de 3 heures sur la tâche, mais si vous le faites, veuillez nous en informer lors du debrief.**
- **Nous préférerions une solution testable (même avec un modèle basique) plutôt qu'une solution optimisée mais non testable.**