Challenge Data Scientist

Données:

Fichiers parquets, ~70.000 urls. Pour récupérer le dataset zippé: data (3.6Mo)

Énoncé du challenge:

Le but du challenge est de produire un classifieur multi-label capable d'inférer les catégories d'une url (et il n'est pas nécessaire de scrapper les pages web des urls).

Nous vous fournissons pour cela un jeu de données au format parquet comportant les colonnes suivantes:

- Url: url d'une page
- Day: le jour du mois
- Target: la liste des classes associés à l'url ← ce qu'on veut prédire

Échantillon des données:

	url	target	day
0	https://www.cdiscount.com/bricolage/electricit...	[1831, 1751, 1192, 745, 1703]	4
1	https://www.mystalk.net/profile/vitoriafcorrea	[847, 978, 582, 1381, 529]	4
2	https://www.lequipe.fr/Tennis/TennisFicheJoueu...	[20, 1077, 294]	4
3	http://m.jeuxvideo.com/forums/42-32625-6018005...	[381, 935, 1343, 622, 933]	4
4	https://context.reverso.net/traduction/espagno...	[692, 1265, 725, 1264, 1266]	4

Vous pouvez utiliser la méthode, le modèle, le framework qui vous semble le plus pertinent.

Vous devrez produire:

- Le liens vers un repo git de votre code
- le code en Python 3
- Les spécifications de l'environnement permettant de faire tourner votre code, au choix:
  - fichier requirements.txt
  - un Dockerfile ou un container
- un README clair nous indiquant comment utiliser ce que vous nous envoyez

Evaluation:

Vous serez évalué sur votre approche du problème, sur la qualité du code produit et sur la qualité des prédictions de votre modèle.

Puis si vos résultats sont valables, vous serez invité en entretien technique avec nos Data Scientists pour discuter de votre solution.