Classification d'exercices d'algorithmique

Challenge Tech Data Science

Contexte

Le site *Codeforces* rassemble un catalogue d'exercices d'algorithmique. Chaque exercice est classifié dans différentes catégories qui caractérisent les notions mises en jeu par l'exercice. Quelques exemples de catégories sont les suivantes: *"math", "strings", "sortings", "combinatorics", "greedy", "graphs"*, etc… Le dataset xCodeEval réunit ces différents exercices ainsi que les solutions validées des participants.

Pour ce challenge, nous ne travaillerons pas sur le dataset original mais sur un sous-ensemble de ce dataset disponible à ce lien qui est composé de 4982 exemples correspondant à des exercices tous différents.

Problème

Proposez un algorithme permettant de prédire les **tags** associés à un exercice d'algorithmique. Il n'est pas interdit d'utiliser d'autres features que la description du problème (**prob_desc_description**), comme par exemple la solution écrite en Python (**source_code**), ou tout autre feature qui vous paraîtra intéressant.

Certains tags sont beaucoup plus difficiles à prédire et moins bien annotés. On pourra concentrer ses efforts sur les 8 tags suivants: *['math', 'graphs', 'strings', 'number theory', 'trees', 'geometry', 'games', 'probabilities'].*

Objectif

L'objectif est de proposer une approche pertinente que nous pourrons challenger ensemble lors d'un entretien physique. Il n'y a pas une bonne réponse mais de nombreuses possibilités d'arriver au résultat. En plus de la pertinence des approches utilisées, la qualité du code sera également évaluée.

Livrables

- Choisir une ou plusieurs métriques permettant d'évaluer la qualité de l'algorithme. A noter que nous sommes intéressés par savoir quels tags sont les mieux prédits par l'algorithme.
- Développer un module Python exécutable à l'aide d'une CLI permettant d'effectuer une prédiction et une évaluation sur un dataset de test. La prédiction pour un exemple doit pouvoir être effectuée en un temps raisonnable (moins de dix secondes). Si vous avez choisi une approche par apprentissage, le code d'entraînement du modèle fait aussi partie des livrables.

- Préparer quelques slides en vue de la restitution orale du challenge, pour présenter l'approche choisie, les métriques et les pistes d'amélioration.

Remarques

- Si besoin pour l'une des solutions mises en œuvre, Google Colab met à disposition un GPU gratuitement (mais il n'est pas obligatoire d'utiliser un GPU).