

Codage de l'information TP – Mai 2024

Problème

L'objectif est de créer un modèle IA en utilisant un algorithme de machine learning qui pourra déterminer (prédire) si un langage donné est un code ou pas.

On s'intéressera uniquement au langage défini dans l'alphabet binaire composé au plus de 10 mots.

Le grand défi de cet exercice est de trouver une représentation matricielle adéquate pour entraîner le modèle de machine learning.

Étape 1 : Création de données

Dans un premier temps, il faut générer les langages ainsi que les mots les composant. Pour se faire,

- 1) Créez une fonction qui génère aléatoirement des mots binaires de longueur aléatoire entre 1 et 7.
- 2) Créez une fonction qui génère aléatoirement des langages (en utilisant la fonction précédente) dont le nombre de mots le composant est aléatoire en 1 et 10
- 3) Créez des fonctions qui nous donneront des propriétés numériques d'un langage pour en faire des colonnes pour nos données d'entraînement (ici se trouvent votre principal travail)
- 4) Créer alors les données complètes pour notre modèle de machine learning (utiliser l'algorithme de Sardinias et Patterson pour la colonne **classe**). Vous ne pouvez pas utiliser plus de 5000 langages (5000 lignes dans la donnée d'entrainement) .

Étape 2 : Création du modèle

- 1) Créer le modèle en utilisant un modèle de classification que vous connaissez (ML ou DL) en python
- 2) Stocker le modèle
- 3) Créer un script python qui permet de faire une prédiction sur un langage donné (si pas clair voir l'étape 3)

Étape 3 : Utilisation du modèle dans une application web

- 1) Créer une page qui nous permettra de saisir un langage
- 2) Créer une page qui fera office d'interface entre le script python qui fait la prédiction et une page de résultat affichant si le langage donné dans la page de la question précédente est un code ou pas accompagné du résultat si on applique ce langage à l'algorithme de Sardinias et Paterson.