Demande
Stage Master 2 recherche 2025

12 novembre 2024

1. Titre : Utilisation des chemins de l'arbre syntaxique (path contexts) dans Code2Vec

2. Encadrant(e)(s) : Cyril Fonlupt, Denis Robilliard

3. Durée : 5 à 6 mois

4. Description du sujet : Lorsqu'on applique les grands modèles de langage (LLMs) à l'analyse et la génération de code informatique, plusieurs approches sont proposées, notamment soit en partant directement du texte du programme comme par exemple CodeBert, soit en se référant à l'arbre syntaxique abstrait (AST - abstract syntax tree) issu de la grammaire du langage considéré, comme Code2Vec. Les travaux de la thèse de O. Belmoudden menés dans l'équipe montrent que l'utilisation par Code2Vec de chemins (path contexts) dans l'AST permet d'obtenir des propriétés sémantiques plus riches, moins liées à la simple syntaxe que CodeBert. Toutefois le codage des path contexts de Code2Vec est améliorable comme montré par Sun et al. (voir [1]).
Dans ce stage on s'intéressera d'abord à répliquer les travaux de [1], puis on testera des pistes d'amélioration, comme la sélection des path contexts à retenir selon leur taille ou selon la présence d'éléments syntaxiques particuliers.

5. Contexte et objectifs de la demande : La demande s'inscrit en continuité avec la thèse de Oumaïma Belmoudden sur les modèles code2vec. Le projet augmentera l'expertise de l'équipe sur les nombreuses modèles LLM dédiés à la programmation automatique.

[1] Improvements to code2vec: Generating path vectors using RNN, Xuekai Sun, Chunling Liu, Weiyu Dong, Tieming Liu, Computers & Security Volume 132, September 2023