Stage de Master 2 (5 mois)

Lieu : Laboratoire LIRIS (Lyon, France)

Responsables du stage (Emails) :

- Hamida Seba (hamida.seba@univ-lyon1.fr)

Thématiques : Graphes, Apprentissage automatique sur les graphes, software héritage

Type de stage : Fin d'études bac +5, Master 2

Durée : 5 mois

Période souhaitée : à partir de février 2024

Rémunération : gratification de stage

Intitulé : Exploration du graphe de softawre heritage

Sujet :

L'exploration de logiciels (software mining) est un domaine de l'ingénierie logicielle empirique qui vise à étudier des ensembles de données de logiciels existants afin de découvrir des modèles et des connaissances qui peuvent aider à améliorer les logiciels futurs [1]. A partir des connaissances obtenues de l'exploration de logiciels, les chercheurs peuvent construire des modèles à l'aide de statistiques et de techniques d'apprentissage automatique, qui peuvent ensuite être interrogés pour améliorer la qualité des logiciels, découvrir des bogues [2], ou même obtenir une vue architecturale de haut niveau de la façon dont les composants logiciels interagissent ensemble entre eux [3]. Software Heritage [4] est une archive de logiciels contenant la plus grande collection publique de fichiers de code source ainsi que l'historique de leur développement, sous la forme d'un immense graphe de centaines de milliards d'arêtes [5]. Dans ce projet, nous nous intéressons à appliquer des méthodes d'apprentissage automatique pour des fins de prédiction de l'évolution de ce graphe et de la production de logiciels de manière générale [5]. Durant ces dernières années, plusieurs méthodes permettant l'apprentissage sur les graphes ont vu le jour [6]. Cependant, un défi persiste sur le passage à l'échelle de ces solutions, en particulier celles basées sur la notion de convolution: Graph Nonvolutional Neural networks (GCN) [7].

L'objectif de ce projet est d'utiliser ces techniques d'apprentissage pour l'exploration du graphe de software héritage, d'étudier leurs limites sur d'aussi grands graphes et de proposer des solutions pour y remédier.

Les tâches à réaliser sont comme suit :

- Prendre en main le graphe de software héritage.
- Etude des GNNs
- Etudier les possibilités de mining réalisées par apprentissage automatique sur ce graphe.
- Applique ces méthodes sur Software Heritage
- Proposer des solutions adaptées à ce graphe

Profil du/de la stagiaire : Compétences avancées (niveau M2) en informatique (d'apprentissage machine et de théorie des graphes fortement souhaitées).

- **Merci d'adresser votre candidature avec un CV ainsi insi que vos notes de l'année universitaire en cours et de l'année dernière à** hamida.seba@univ-lyon1.fr

Références

[1] Nachiappan Nagappan, Thomas Ball, and Andreas Zeller. "Mining metrics to predict component failures". In: Proceedings of the 28th international conference on Software engineering. ACM. 2006, pp. 452–461.

[2] Chadd C Williams and Jeffrey K Hollingsworth. "Automatic mining of source code repositories to improve bug finding techniques". In: IEEE Transactions on Software Engineering 31.6 (2005), pp. 466–480. [3] Hofer, D., Jäger, M., Mohamed, A., & Küng, J. (2020, November).

[3] Ahmed E Hassan. "The road ahead for mining software repositories". In: Frontiers of Software Maintenance, 2008. FoSM 2008. IEEE. 2008, pp. 48–57.

[4] https://www.softwareheritage.org

[5] Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli: The Software Heritage Graph Dataset: Large-scale Analysis of Public Software Development History. MSR 2020: 1-5

[6] William L. Hamilton. Graph Representation Learning. E-bbok Springer

[7] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (2017). https://openreview.net/forum?id=SJU4ayYgl