Echantillonnage direct de l'espace des motifs

L'extraction de l'ensemble complet des motifs vérifiant une contrainte (e.g., fréquence, aire, etc.) est un problème NP-difficile. Par conséquent, nous n'avons aucune garantie sur les temps d'exécution d'une approche exhaustive même si cette dernière exploite pleinement différentes propriétés d'élagage de l'espace de recherche. Pour pallier à ce problème et permettre de présenter « instantanément » des motifs pertinents à l'analyste, de nombreux travaux visant à échantillonner directement l'espace des motifs ont été développés. Ce TP s'intéresse à l'un d'eux :

Boley, M., Lucchese, C., Paurat, D., & Gärtner, T. (2011, August). *Direct local pattern sampling by efficient two-step random procedures.* In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 582-590). ACM.

Disponible à l'adresse suivante : https://cremilleux.users.greyc.fr/pub/papers/BoleyLucchesePauratGartnerKDD2011directLocalPatternSamplingTwoStepRandom.pdf

L'objectif de ce TP est d'implémenter (et appliquer) les algorithmes d'échantillonnage introduit dans cet article, notamment l'échantillonnage de motifs par rapport à la **fréquence** et à l'**aire** :

1. Implémenter l'algorithme d'échantillonnage des motifs fréquents.
2. Implémenter l'algorithme d'échantillonnage basé sur l'aire.
3. Ecrire une fonction qui étant données k réalisations, retourne les valeurs réelles de la fréquence et/ou l'aire en une seule passe sur les données.
4. Tester avec des données réelles : http://fimi.ua.ac.be/data/ (chess, connect, mushroom, etc.)
5. Pour différents 4 jeux de données, afficher la distribution de 1000 réalisations. Attention, l'approche s'appuie sur un tirage avec remise, il est donc possible d'avoir des doublons qu'il faudra veiller à supprimer.
6. Mettre en place une expérience pour évaluer la diversité de k tirages.
7. (Bonus) Implémenter l'algorithme 3, et afficher la distribution de 1000 réalisations.
8. (Bonus++) Imaginer un algorithme d'échantillonnage s'appuyant sur une autre mesure.

Langages de programmation possibles : python, java, C++, etc.

Ce travail sera **évalué** et peut s'effectuer en **binôme.**

Modalité de rendu : 2 versions de travail: une en fin de séance (idéalement un notebook, à minima du code documenté), puis une version révisée avant la prochaine séance de TP (i.e., 27/03/2018, 23:59 GMT) par mail à marc.plantevit@univ-lyon1.fr.

Pour aller plus loin :

Mario Boley, Sandy Moens, Thomas Gärtner: Linear space direct pattern sampling using coupling from the past. KDD 2012: 69-77
http://win.ua.ac.be/~adrem/bibrem/pubs/boley12cftp.pdf