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://perso.liris.cnrs.fr/marc.plantevit/ENS/DMTP/BoleyLucchesePauratGartnerKDD2011directLocalPatternSamplingTwoStepRandom.pdf

L'objectif de ce TP est d'implémenter dans le langage de votre choix (Python, Java, C++, etc.) et d'appliquer les algorithmes d'échantillonnage introduits 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. La méthode proposée retourne des motifs (réalisations) à la demande. Toutefois, aucune information sur le motif autre que sa syntaxe n'est donnée (i.e., la fréquence n'est pas communiquée). 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 (attention de ne pas considérer des jeux de données aux caractéristiques particulières[1]):
   - http://fimi.ua.ac.be/data/ (chess, connect, mushroom, etc.)
   - http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php
5. Pour 5 jeux de données différents, 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. On veut mettre en évidence que la probabilité d'un motif d'être tiré est proportionnelle à sa mesure (fréquence ou aire dans notre cas). Mettre en place une expérience pour tester l'échantillonnage. On peut éventuellement comparer les résultats par rapport aux résultats fournis par une approche complète (tous les motifs de minsupp >= 1).
8. Comment se comporte l'algorithme sur des jeux de données contenant au moins une transactions beaucoup plus grande que les autres ? (e.g., Kosarak). Proposer et implémenter une solution.
9. (Bonus) Implémenter l'algorithme 3, et afficher la distribution de 1000 réalisations.
10. (Bonus++) Imaginer un algorithme d'échantillonnage s'appuyant sur une autre mesure.
11. Cette approche est elle adaptée pour échantillonner des motifs fermés ? Justifier et discuter une solution en fonction.

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

Ce travail ne sera pas **évalué.**

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

[1] E.g., les jeux de données où toutes les transactions ont la même taille ou ceux qui ont une transaction beaucoup plus importantes que les autres (2048 vs 5 en moyenne).