Échantillonnage 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 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). Écrire 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):
   - https://bitbucket.org/anesbendimerad/sigibbssamplingcode/src/master/ItemsetDatasets/
   - http://fimi.ua.ac.be/data/
   - http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php
5. Mettre en place une étude empirique (sur au moins 4 jeux de données réels + éventuellement 2 synthétiques) afin de vérifier la qualité de l'échantillonnage, notamment en permettant de répondre aux questions suivantes :
   - Est-ce que les motifs sont bien tirés proportionnellement à leur mesure (e.g., plus il est fréquent, plus sa probabilité d'être tirer est grande, idem pour l'aire).
   - Est-ce la méthode d'échantillonnage permet d'obtenir une bonne diversité des motifs tirés ?
   - …
6. Comment se comporte l'algorithme sur des jeux de données contenant au moins une transaction beaucoup plus grande que les autres ? (e.g., Kosarak). Proposer et implémenter une solution.
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 (notebook), java, C++, etc.

Ce travail sera **évalué** et peut s'effectuer en **[bi|tri]nôme.**

Modalité de rendu : Un rapport plus le code. Idéalement un notebook (et son pdf), à minima du code documenté avant le 23/11/2019 (07h59) en déposant le rapport et le code (dans une archive) dans la colonne Tomuss dédiée. **Ne déposer qu'une seule fois par groupe.**
Pour le nom des fichier, faites clairement apparaitre vos noms. Pour le code, si trop volumineux, vous pouvez déposer un fichier contenant un lien vers un dépôt du code.

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