TP : Visualisation, Clustering, Co-clustering

- Ce TP composé de deux parties est à réaliser par binome ou seul. Il sera noté.
- Remise d'un repport sous forme d'un Notebook à 17h30 même non fini, toute ressemblance entre TPs sera sanctionnée.

1 Partie 1

Dans ce projet on s'appuiera d'abord sur le travail intitulé "Scénario" qui permet de mettre en pratique quelques méthodes de visualisation telles que ACP, AFC, AFCM, MDS et de classification telles que kmeans, CAH, NMF. Les données, l'objectif et ces méthodes sont décrites dans le document disponible dans `http://wikistat.fr/pdf/st-scenar-explo-spam.pdf`. Il s'agit de 58 variables qui sont observées sur 4601 messages dont 1813 pourriels (spams). Les données sont disponibles dans le répertoire `http://wikistat.fr/data`. La variable binaire **Spam** est présente à titre illustratif mais qui sera utile dans la visualisation et la comparaison des méthodes.

1. Exécuter toutes les étapes décrites dans le document, bien assimiler les objectifs.
2. Faire une synthèse de tous ces résultats.
3. Proposer une autre normalisation appropriée et mesurer son impact sur les résultats.
4. Sachant que la matrice des données est sparse et que le co-clustering serait très approprié.
   (a) Utiliser le package `Blockcluster`[1] et `blockmodels`[2] avec le modèle approprié. Justifier votre réponse.
   (b) Visualiser les classes de messages.
   (c) Compléter votre analyse en termes de visualisation et de clustering en utilisant d'autres méthodes disponibles dans le package `Coclust` [3]. Commenter vos résultats.
   (d) *Question facultative (Bonus)*: Même question en utilisant d'autres méthodes disponibles dans le package python `scikit-learn`[4].
   (e) *Question facultative (Bonus)*: Investiguer le package `biclust` et particulièrement la méthode `BCQuest` sur la matrice codée en catégories. Commenter vos résultats.
   (f) Faire une analyse synthétique de vos résultats.

2 Partie 2

Sur le dossier `DATA` sont diponibles plusieurs matrices documents-termes.

1. Faire un descriptif de ces tables de données en termes de taille, sparsité et d'autres indicateurs pertinents à la fois sur l'ensemble des documents et l'ensemble des termes.
2. Choisir 4 tables et visualiser sur un plan l'ensemble des documents et l'ensemble des termes avec au moins deux méthodes, que peut-on dire ?
3. Choisir 3 méthodes de clustering, porter les classes obtenues sur les plans et comparer les résultats obtenus en terme de clustering sur l'ensemble des documents.
4. Choisir 3 méthodes de co-clustering appropriées, comparer les résultats obtenus en terme de clustering sur l'ensemble des documents.
5. Faire une conclusion.

[1] `https://https://cran.r-project.org/web/packages/blockcluster/vignettes/blockcluster_tutorial.pdf`
[2] https://cran.r-project.org/web/packages/blockmodels/blockmodels.pdf
[3] `https://pypi.python.org/pypi/coclust`
[4] `http://scikit-learn.org/stable/`