CORPEX : Analyse exploratoire d'un corpus biomédical à l'aide de la classification croisée

Amine Ferdjaoui, Amira Tlati, Séverine Affeldt, Mohamed Nadif

CORPEX : Analyse exploratoire d'un corpus biomédical à l'aide de la classification croisée

Amine Ferdjaoui[*,**], Amira Tlati[*]
Séverine Affeldt[*], Mohamed Nadif[*]

[*] Centre Borelli UMR 9010, Université Paris Cité, 75006, France.
[**] SogetiLabs, 147 Quai du Président Roosevelt, 92130, Issy-les-Moulineaux.
<prénom.nom>@u-paris.fr

Résumé. Nous proposons une interface d'aide à l'analyse de corpus via la visualisation interactive de *coclusters* afin d'accompagner l'exploration des thématiques pour un ensemble de textes. Les saisies de l'utilisateur permettent la création ou le chargement d'un corpus de documents, son nettoyage et l'étude interactive et simultanée des termes et des documents. Cet article détaille les fonctionnalités en lien avec la génération dynamique de corpus, notamment dans un cadre biomédical, et également le chargement de matrices documents-termes pour des corpus déjà pré-traités. L'analyse du corpus par la classification croisée (*co-clustering*) et la visualisation conjointe des termes et des documents, suivant le co-partitionnement obtenu sur les deux ensembles, sont des outils efficaces pour une compréhension rapide des sujets abordés dans un corpus. La sauvegarde automatique des résultats permet de relancer facilement différentes analyses par un *co-clustering* approprié et d'obtenir des vues croisées des thématiques à différents niveaux de granularité.

1 Introduction

L'information est aujourd'hui disponible en abondance sous forme textuelle, dans de nombreux domaines. Le Traitement Automatique des Langues (TAL) permet l'automatisation à grande échelle de tâches telles que l'annotation ou la classification. Les méthodes existantes, aisément accessibles via de nombreuses librairies de programmation en `R` ou `Python`, permettent d'analyser et de valoriser de larges corpus comportant par exemple des articles de presse, des compte-rendus d'entretiens ou des commentaires de consommateurs.

Dans le domaine biomédical, de très nombreux articles sont aujourd'hui disponibles en ligne et leur exploitation peut permettre d'identifier des relations d'intérêt pour une éventuelle meilleure prise en charge des patients. Toutefois, on produit de nos jours bien plus d'articles biomédicaux qu'on ne peut en lire. A titre d'exemple, la plateforme PubMed comprend plus de 34 millions de citations pour la littérature biomédicale provenant de MEDLINE, de revues de sciences de la vie et de livres en ligne. Recouper l'ensemble des documents mis à disposition nécessite l'emploi d'approches de TAL avancées. Avec `CORPEX` (*CORPus EXploration*), nous mettons à la disposition de la communauté des chercheurs, mais également des praticiens, une

interface ergonomique et légère pour l'exploration de corpus, notamment dans un contexte biomédical.

L'utilisateur peut soit générer un corpus sur la base de mots-clefs directement à partir de PubMed, soit charger dans l'interface un corpus pré-existant. CORPEX permet également de nettoyer les textes d'un corpus et de les analyser via une approche de classification croisée (*co-clustering*) performante. L'interface proposée permet donc, de façon interactive, d'identifier les sous-thématiques d'un corpus, de relancer des analyses en adaptant la granularité à partir du co-partitionnement, et de consulter en ligne les articles pertinents du corpus en fonction des termes les plus représentatifs pour les sous-thématiques découvertes.

Dans les sections suivantes, nous présentons dans un premier temps les motivations et l'état de l'art pour le *co-clustering* dans un contexte de données textuelles. Puis, nous détaillons les éléments de l'approche de co-partitionnement sur laquelle se base CORPEX. Nous présentons ensuite les différentes fonctionnalités de notre interface en lien avec l'analyse et la visualisation, avant de décrire un cas d'application dans le domaine biomédical.

2 Le *co-clustering* pour les données textuelles

2.1 Motivations et travaux connexes

La classification croisée ou simultanée sur un ensemble de caractéristiques et un ensemble d'objets conduit à une réorganisation de la matrice de données en une structure de blocs homogènes, appelé *coclusters* [1]. Il s'agit en fait d'une extension du *clustering unilatéral* qui prend cependant en compte simultanément les deux ensembles (Govaert et Nadif, 2008, 2013).

Diverses méthodes de *co-clustering* ont été appliquées dans différents domaines, notamment en bioinformatique (Cho et Dhillon, 2008 ; Hanczar et Nadif, 2012) pour regrouper les gènes et les conditions expérimentales, dans le filtrage collaboratif (Hofmann et Puzicha, 1999 ; Deodhar et Ghosh, 2010) pour regrouper les utilisateurs et les éléments, et dans l'exploration de texte (Govaert et Nadif, 2018 ; Salah et Nadif, 2019) pour regrouper des termes et des documents. Grâce à sa capacité à mettre en relation les lignes et les colonnes d'une matrice donnée, le *co-clustering* donne généralement de meilleurs résultats que le clustering unilatéral. Il s'avère même particulièrement efficace pour les données massives de grande dimension et éparses. En outre, le *co-clustering* effectue une réduction de dimensionnalité adaptative implicite qui permet l'utilisation d'algorithmes efficaces et évolutifs pour les données textuelles éparses à haute dimension. Cet aspect est crucial dans le domaine de l'exploration de textes, car la croissance exponentielle des documents en ligne a créé un besoin urgent de méthodes efficaces de traitement et d'interprétation des matrices de documents-termes éparses à haute dimension, c'est-à-dire des matrices où les documents sont représentés dans l'espace des termes, et vice versa. Plus important encore, le *co-clustering* de textes peut identifier les termes les plus discriminants qui caractérisent les sujets abordés dans les classes de documents.

Les approches classiques de *co-clustering* n'intègrent généralement pas d'informations supplémentaires comme un score complémentaire externe qui quantifie les relations sémantiques entre les termes, ou les similarités dans le contenu des documents. Les informations

1. Soit une matrice $\mathbf{X} = (x_{ij}), i \in I, j \in J$, un co-cluster est une sous-matrice $I_k \times J_\ell (I_k \subseteq I, J_\ell \subseteq J)$.

secondaires sur l'espace latent des documents *et* sur l'espace latent des termes peuvent cependant améliorer le *co-clustering* des données document-mot. Inspiré par le succès récent des modèles neuronaux d'intégration des termes, dans (Ailem et al., 2017 ; Febrissy et al., 2022) les auteurs ont proposé un modèle basé sur la NMF (Non-negative Matrix Factorisation) exploitant conjointement les matrices documents-termes et termes-contextes. Récemment, une extension du *co-clustering* basé sur la NMTF (Non-negative Matrix Tri-Factorization), à savoir le `WC-NMTF` (Word Co-Occurence regularized NMTF) (Salah et al., 2018), une technique qui tient compte des relations sémantiques entre les termes, a été appliquée avec succès sur divers ensembles de données textuelles. En plus d'être de grande dimension et éparses, les classes recherchées de documents ou de termes peuvent également être fortement déséquilibrées, et les méthodes de *co-clustering* qui se concentrent sur ce type de données doivent en tenir compte. L'algorithme `DCC` (Directional Coclustering with a Conscience) (Salah et Nadif, 2019), s'est avéré particulièrement adapté à la résolution de ce problème. Il exploite le fait que les données textuelles sont de nature *directionnelle*, ce qui signifie que seules les directions des vecteurs de documents/termes sont pertinentes, et non leur magnitude (Mardia et Jupp, 2009). Il s'appuie ainsi sur le modèle de mélange de von Mises-Fisher (vMF) et introduit un mécanisme de conscience pour éviter les classes vides ou fortement déséquilibrées (Salah et Nadif, 2017a). Cependant, et contrairement à `WC-NMTF`, `DCC` n'utilise aucune régularisation.

2.2 Le *co-clustering* avec conscience régularisé

Dans ce travail, nous exploitons la nature directionnelle des données textuelles via l'algorithme de `RBDCo` (Regularized Bi-Directional Co-clustering)(Affeldt et al., 2021). L'aspect bidirectionnel de cette approche réside dans l'utilisation d'informations issues des représentations des documents et des termes. L'intérêt du modèle `RBDCo` est qu'il propose un cadre général basé sur une formulation matricielle du *co-clustering* s'appuyant initialement sur les modèles de mélange de von Mises-Fisher (vMF) (Banerjee et al., 2005 ; Salah et Nadif, 2017b). Un résultat significatif de la formulation de `RBDCo` est un cadre très riche et flexible pour le *co-clustering* de textes qui permet une régularisation multiplicative simple, à la fois sur les relations sémantiques *mot-mot* et les similarités de contenu *document-document*. Contrairement aux méthodes existantes, qui reposent généralement sur une incorporation *additive* des similarités, nous proposons une régularisation *multiplicative* et *bidirectionnelle* qui encapsule mieux la structure sous-jacente des données textuelles.

Plus précisément, `RBDCo` propose une formulation matricielle de `DCC` (Salah et Nadif, 2017a) qui pénalise chaque classe ligne/colonne en rendant les paramètres de concentration inversement proportionnels à la cardinalité. `DCC` entrelace les regroupements en g classes de n documents/lignes et de d termes/colonnes à chaque itération. Considérons les matrices de classification binaire $\mathbf{Z} \in \{0,1\}^{n\times g}$ and $\mathbf{W} \in \{0,1\}^{d\times g}$ (Fig. 1), où les tailles des classes de $\mathbf{Z}$ et $\mathbf{W}$ sont sur la diagonale de $\mathbf{D_z} = \mathbf{Z}^\top\mathbf{Z}$ et $\mathbf{D_w} = \mathbf{W}^\top\mathbf{W}$. Etant donnée une matrice documents-termes $\mathbf{X}$, les formules de mise à jour de l'algorithme `DCC` peuvent être réécrites sous forme de matrice : $\mathbf{Z} = \textbf{Binarize}(\mathbf{X}\widetilde{\mathbf{W}}\mathbf{D}_\mathbf{z}^{-0.5})$ et $\mathbf{W} = \textbf{Binarize}(\mathbf{X}^\top\widetilde{\mathbf{Z}}\mathbf{D}_\mathbf{w}^{-0.5})$, où $\widetilde{\mathbf{Z}} = \mathbf{Z}\mathbf{D}_\mathbf{z}^{-0.5}$, $\widetilde{\mathbf{W}} = \mathbf{W}\mathbf{D}_\mathbf{w}^{-0.5}$ et $\textbf{Binarize}(\mathbf{B})$, signifie $\forall i, \mathbf{b}_{ik} = \text{argmax}_{k'}\, \mathbf{b}_{ik'}$. Nous obtenons les partitions $\mathbf{Z}$ et $\mathbf{W}$ en alternant ces deux règles jusqu'à ce qu'un point fixe soit atteint. Ainsi, le clustering de documents $\mathbf{Z}$ est dérivé comme une projection pondérée de la matrice de données $\mathbf{X}$ sur le sous-espace couvert par la partition des termes $\mathbf{W}$. De même, la partition

de termes est dérivée comme une projection pondérée de la matrice de données $\mathbf{X}$ sur le sous-espace couvert par la partition des documents $\mathbf{Z}$.

RBDCo intègre deux matrices de données régularisées, $\mathbf{M_z}$ et $\mathbf{M_w}$ avec des valeurs prises dans $\{\mathbf{X}, \mathbf{S}_r\mathbf{X}, \mathbf{X}\mathbf{S}_c, \mathbf{S}_r\mathbf{X}\mathbf{S}_c\}$ aux formules précédentes. Pour la matrice documents-termes, $\mathbf{S}_r$ contient les similarités entre les documents et $\mathbf{S}_c$, les corrélations sémantiques entre les termes. La tâche de regroupement des données pour RBDCo est effectuée en calculant itérativement $\mathbf{Z}$ et $\mathbf{W}$ sur la base de l'interaction entre les deux règles de mise à jour dérivées de la maximisation du critère objectif J. L'Equation.(1) et l'Algorithme (1) rendent compte de cette alternance. En particulier, RBDCo exploite la dualité des documents et des termes, et renforce leur regroupement conjoint avec des régularisations multiplicatives doubles utilisant à la fois $\mathbf{S}_c$ et $\mathbf{S}_r$.

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{W} \end{bmatrix} \leftarrow \begin{bmatrix} 0 & \mathbf{M_z} \\ \mathbf{M_w}^\top & 0 \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{Z}}\mathbf{D}_\mathbf{w}^{-0.5} \\ \widetilde{\mathbf{W}}\mathbf{D}_\mathbf{z}^{-0.5} \end{bmatrix} = \begin{bmatrix} \mathbf{M_z}\widetilde{\mathbf{W}}\mathbf{D}_\mathbf{z}^{-0.5} \\ \mathbf{M_w}^\top\widetilde{\mathbf{Z}}\mathbf{D}_\mathbf{w}^{-0.5} \end{bmatrix}. \quad (1)$$

Algorithm 1 RBDCo. Regularized Bi-Directional Co-Clustering

Input : $\mathbf{X}$ ($\mathbf{x}_i \in \mathbb{S}^{d-1}$), g, $\mathbf{S}_r$, $\mathbf{S}_c$
Output : partitions $\mathbf{Z}$ and $\mathbf{W}$
Initialization : random initialization of $\mathbf{Z}$ and $\mathbf{W}$
repeat
  1. Assignment of documents (1)
    • $\mathbf{Z} \leftarrow \mathbf{M_z}\widetilde{\mathbf{W}}\mathbf{D}_\mathbf{z}^{-0.5}$
    • **Binarize** $\mathbf{Z}$: $\forall i \quad z_i = \arg\max_{k'} z_{ik'}$
  2. Assignment of words (1)
    • $\mathbf{W} \leftarrow \mathbf{M_w}^\top\widetilde{\mathbf{Z}}\mathbf{D}_\mathbf{w}^{-0.5}$
    • **Binarize** $\mathbf{W}$: $\forall j \quad w_j = \arg\max_{k'} w_{jk'}$
until convergence of $J \equiv \frac{1}{2}Tr\big(\widetilde{\mathbf{Z}}^\top(\mathbf{M_z}+\mathbf{M_w})\widetilde{\mathbf{W}}\big)$

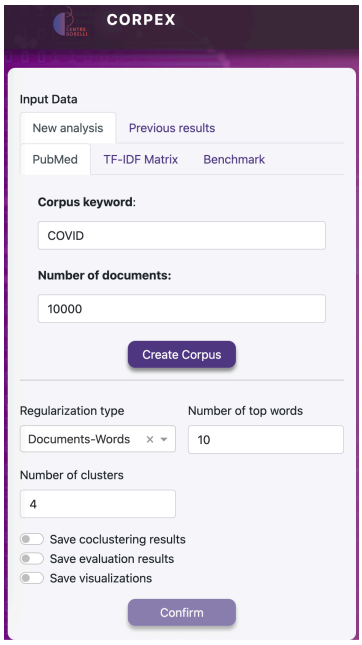
$\mathbf{S}_c$ and $\mathbf{S}_r$ sont construites à partir de la matrice documents-termes originale $\mathbf{X} \in \mathbb{R}^{n\times d}$. Pour $\mathbf{S}_c$, nous utilisons une transformation non-lineaire de la co-occurrence des termes, la Pointwise Mutual Information (PMI), qui est définie comme $\log\left(p(w_i, w_j)/p(w_i)p(w_j)\right)$ pour deux termes w_i and w_j. Avec $\mathbf{C} = \mathbf{X}^\top\mathbf{X}$, la PMI est

$$\mathrm{PMI_C}(w_i, w_j) = \log\frac{c_{ij} \times c_{..}}{c_{j.}c_{.j}}, \quad (2)$$

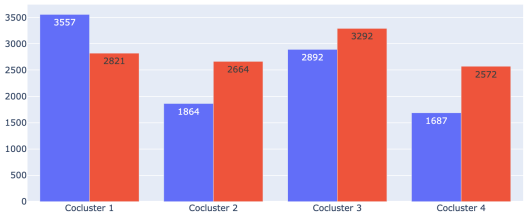
où $c_{..} = \sum_{ij} c_{ij}$, $c_{i.} = \sum_j c_{ij}$ et $c_{.j} = \sum_i c_{ij}$. Une approximation généralement admise consiste à remplacer toutes les valeurs négatives de PMI par 0, donnant une PMI positive (PPMI). $\mathrm{PPMI_c}$ est notre matrice de régularisation de termes $\mathbf{S}_c$. Nous pouvons aussi définir une $\mathrm{PMI_r}(d_i, d_j)$ entre les documents d_i et d_j et considérer la $\mathrm{PPMI_R}$ comme étant $\mathbf{S}_r$.

3 Interface d'analyse et de visualisation

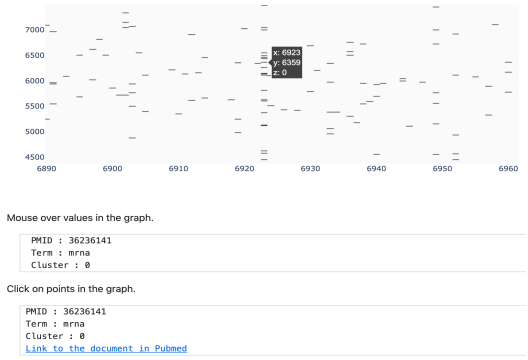
Le système d'exploration CORPEX est instrumenté par un *panneau latéral* qui centralise la définition du corpus et le lancement de l'analyse par le *co-clustering*. L'outil propose ensuite, via le *panneau central*, différentes visualisations des termes et articles pertinents permettant de faciliter l'exploration du corpus, les thématiques découvertes et les termes les plus discriminants. Pour cette partie, nous prendrons l'exemple de la thématique du COVID.



(a) Panneau latéral



(b) Distribution des documents (bleu) et des termes (rouge)



(c) Agrandissement matrice documents×termes et selection

FIG. 1 – Vue globale de l'interface.

3.1 Panneau latéral : Définition et analyse du corpus

L'interface CORPEX permet la création d'un corpus de documents directement via PubMed (Fig. 1 (a) ; onglet **[New Analysis > PubMed]**). L'utilisateur saisit un mot-clef pour télécharger un ensemble de documents (eg. COVID). Le corpus est *nettoyé* (eg. suppression des ponctuations, des *stopwords*) et converti en matrice documents-termes pondérée (TF-IDF ; *Term Frequency-Inverse Document Frequency*). Le corpus peut être sauvegardé au format texte ou matriciel (objet Python). L'analyse de *co-clustering* peut ainsi être relancée ultérieurement sur un même corpus (via l'onglet **[New Analysis > TF-IDF Matrix]**).

L'interface permet également le chargement de corpus *benchmarks* typiquement utilisés dans le domaine du clustering de textes (eg. CLASSIC3&4, PUBMED5&10 ; onglet **[New Analysis > Benchmark]**). Elle permet alors le calcul de métriques d'évaluation sur la base des labels des classes de documents. Les métriques sont l'ARI (Adjusted Rand Index) et la

NMI (Normalized Mutual Information). Cette étape pourrait également servir à comparer deux co-partitionnements. Il est possible de recharger des analyses et visualisations obtenues avec CORPEX sans avoir à relancer un co-clustering(onglet **[Previous results]**).

⇝ **Application au COVID-19** Après avoir saisi le mot-clef COVID (*Corpus keyword*) ainsi que le nombre de documents (*Number of documents*), nous pouvons télécharger un corpus de $10,000$ documents *nettoyés* par un simple clic sur le bouton *Create Corpus* (Fig. 1 (a)). Nous définissons ensuite le type de régularisation pour l'algorithme RBDCo (*Regularization type*) et le nombre de co-clusters (*Number of clusters*). Les boutons basculant situés en bas du panneau latéral permettent d'activer la sauvegarde en local du co-partitionnement (*coclustering results*), des métriques d'évaluation (*evaluation results*) et des visualisations de co-clusters et mots représentatifs des thématiques découvertes (*visualizations*). Le bouton *Confirm* lance l'analyse par RBDCo.

3.2 Panneau central : Exploration de résultats

L'onglet **Coclustering overview** permet d'accéder à la répartition des termes et documents par cocluster, et à la matrice *interactive* documents-termes réorganisée selon les co-clusters. Le co-partitionnement obtenu répartit les termes et les documents en coclusters cohérents, autour de certaines thématiques. Un histogramme (Fig. 1 (b)) rend compte des nombres de termes (en rouge) et de documents (en bleu) par cocluster. La matrice interactive documents-termes réorganisée permet de se déplacer dans les différents co-clusters et d'agrandir certaines sous-matrices. L'utilisateur peut ainsi accéder, par un simple clic, à un terme et un document d'intérêt qu'il souhaite étudier (Fig. 1 (c)). En effet, un lien *url* est fourni afin d'accéder à l'article en ligne via la plateforme PubMed (*Link to the document in PubMed*).

⇝ **Application au COVID-19** L'histogramme de distributions des mots et documents montre le co-partitionnement des données dans le cadre de notre corpus *COVID* de $10,000$ documents. Après un agrandissement sur une partie du 3^e cocluster dans la matrice interactive (Fig. 1 (c)), nous accédons par exemple au terme mrna (*Messenger RNA*), connu pour son rôle dans les vaccins COVID-19. Une url redirige vers un article accessible via PubMed (PMID : 36236141).

L'onglet **Topics overview** donne les termes les plus représentatifs par cocluster sous forme d'histogrammes (Fig. 2) et de nuages de mots (Fig. 3). Les histogrammes indiquent précisément la fréquence de ces termes pour chaque cocluster, par un simple glissement de la souris. Le champs *Number of top words* du panneau latéral permet de définir le nombre de mots représentatifs à considérer pour l'exploration des thématiques.

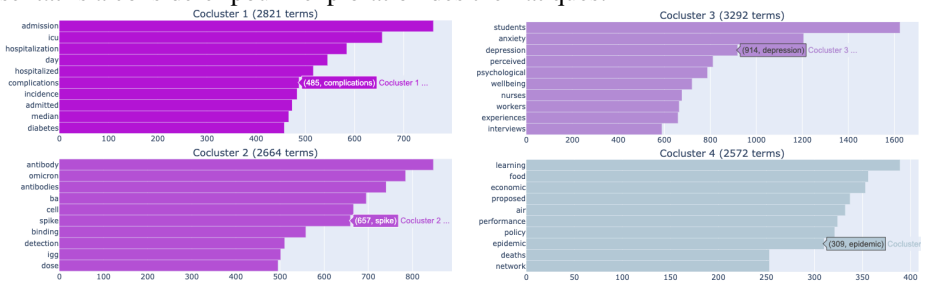


FIG. 2 – Termes représentatifs pour les co-clusters 1 à 4

⇝ **Application au COVID-19** Nous identifions ici quatre thématiques (Fig. 2). Le cluster 1 concentre un vocabulaire lié aux hospitalisations, plus particulièrement dans les services de soins intensifs (*icu*, *admission*) ou pour des patients présentant des comorbidités (*diabetes*). Il est intéressant de noter que *cancer*, visible dans le nuage de mots (Fig. 3), n'apparaît pas dans l'histogramme. En effet, les nuages s'appuient sur les fréquence brutes, tandis que les histogrammes utilisent les valeurs pondérées. Ainsi, *cancer* est très fréquent dans le corpus, mais n'est pas fortement discriminant pour le cluster 1. Le cluster 2 se réfère aux aspects biologiques de la maladie. On retrouve notamment les appellations de variants (*omicron*, *ba* pour BA.4 ou BA.5) ou de protéine (*spike*, *igg*). Le cluster 3 est quant à lui en lien avec les effets psychologiques dus à la pandémie de COVID-19 (*anxiety*, *depression*) et le type de population touchée (*students*, *nurses*, *workers*). Enfin, le cluster 4 se rapporte aux politiques de santé publique et au suivi épidémiologique pendant la pandémie (*policy*, *epidemic*).

FIG. 3 – Nuages des termes pour les coclusters 1 à 4 (gauche à droite)

4 Conclusion

Dans ce travail, nous proposons un outil convivial et efficace permettant d'explorer un corpus biomedical. Il s'appuie sur des avancées récentes en terme de co-clustering. Notons que cet outil pourrait d'une part être utilisé pour d'autres applications dont les données à explorer sont de grande dimension éparses ou non et d'autre part intégrer d'autres nouvelles régularisations.

Références

Affeldt, S., L. Labiod, et M. Nadif (2021). Regularized bi-directional co-clustering. *Statistics and Computing 31*(3), 1–17.

Ailem, M., A. Salah, et M. Nadif (2017). Non-negative matrix factorization meets word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1081–1084.

Banerjee, A., I. S. Dhillon, J. Ghosh, et S. Sra (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res. 6*, 1345–1382.

Cho, H. et I. S. Dhillon (2008). Coclustering of human cancer microarrays using minimum sum-squared residue coclustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 5*(3), 385–400.

Deodhar, M. et J. Ghosh (2010). Scoal : A framework for simultaneous co-clustering and learning from complex data. *TKDD 4*(3), 1–31.

Febrissy, M., A. Salah, M. Ailem, et M. Nadif (2022). Improving nmf clustering by leveraging contextual relationships among words. *Neurocomputing 495*, 105–117.

Govaert, G. et M. Nadif (2008). Block clustering with bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis 52*(6), 3233–3245.

Govaert, G. et M. Nadif (2013). *Co-clustering : models, algorithms and applications*. New York : John Wiley & Sons.

Govaert, G. et M. Nadif (2018). Mutual information, phi-squared and model-based co-clustering for contingency tables. *Advances in Data Analysis and Classification 12*(3), 455–488.

Hanczar, B. et M. Nadif (2012). Ensemble methods for biclustering tasks. *Pattern Recognition 45*(11), 3938–3949.

Hofmann, T. et J. Puzicha (1999). Latent class models for collaborative filtering. In *IJCAI*, Volume 99, Stockholm, Sweden, pp. 688–693. Morgan Kaufmann.

Mardia, K. V. et P. E. Jupp (2009). *Directional statistics*, Volume 494. New York, NY, USA : John Wiley & Sons.

Salah, A., M. Ailem, et M. Nadif (2018). Word co-occurrence regularized non-negative matrix tri-factorization for text data co-clustering. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3992–3999.

Salah, A. et M. Nadif (2017a). Model-based von mises-fisher co-clustering with a conscience. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 246–254. SIAM.

Salah, A. et M. Nadif (2017b). Social regularized von mises–fisher mixture model for item recommendation. *Data Mining and Knowledge Discovery 31*(5), 1218–1241.

Salah, A. et M. Nadif (2019). Directional co-clustering. *Adv. Data Analysis and Classification 13*(3), 591–620.

Summary

We propose an interface that supports corpus analysis via interactive visualizations of *co-clusters* to explore the topics for a set of texts. The user can create or load a corpus of documents, clean them and study simultaneously the terms and the documents. This article details the functionalities related to the dynamic generation of corpora, especially in a biomedical context, and also the loading of document-term matrices for already pre-processed corpora. The analysis of the corpus by cross-classification (*co-clustering*) and the joint visualization of the terms and documents according to the co-partitioning, are effective tools for a quick understanding of the topics in a corpus. The automatic saving of the results allows to easily relaunch different *co-clustering* analyses and obtain crossed views of the topics at different levels of granularity.