TP3 : Croisement de variables pour l'analyse et présentation de quelques visualisations

Yoann Pitarch

1 Objectifs

Au cours des précédents TPs, vous avez découvert les sources d'informations à votre disposition, appris à décrire la structure des données obtenues après un export pour qu'elles soient *comprises* par TETRALOGIE et avez enfin appréhendé les filtres négatifs, *i.e.,* le fichier des mots vides, et les dictionnaires de synonymes.

L'objectif de ce TP est de vous familiariser avec les analyses issues du croisement de variables. Typiquement, il s'agira par exemple d'étudier le comportement des auteurs par rapport à une autre *dimension* (ou attribut) telle que les mots du résumé ou les journaux. Au delà d'un aspect purement technique inhérent à la prise en main du logiciel TETRALOGIE, il s'agira aussi d'exercer votre capacité d'analyse pour *faire ressortir quelque chose de ces croisements des variables*. Enfin, vous visualiserez et interprèterez le résultat d'une analyse de différentes manières (matrice 2D, carte et graphe).

La section suivante donne les consignes de ce TP. Vous trouverez en annexe (dernière section de ce document) des indications et des captures d'écran pour vous aider dans la réalisation de ce TP.

IMPORTANT : le rendu prendra la forme suivante. Comme pour le précédent TP, vous créerez un dossier selon le format convenu et déposerez les fichiers à l'intérieur de celui-ci.

2 Instructions

Préambule. Les questions doivent être réalisées dans l'ordre. Il est donc nécessaire de lire le sujet ;-)

1. Sur la base PubMed, recherchez "HIV vaccine" et filtrez les résultats pour ne garder que les publications de moins de 10 ans, qui concernent l'homme et dont l'abstract est disponible. Combien de résultats obtenez vous ?
2. Exportez les résultats et ouvrez ce jeu de données dans TETRALOGIE.
3. Effectuez un premier comptage de tous les champs.
4. Appliquez, pour les champs qui vous semblent pertinents, un filtre négatif pour les mots vides et recherchez les synonymes.

5. Une fois cela fait, trouvez pour les champs Titre et Abstract, les 1000 mots les plus fréquents et appliquez cette liste comme un filtre positif.
6. Vous êtes désormais prêts à entamer le croisement de variables. A ce propos, quelles paires de dimensions vous semblent pertinentes pour les croisements ? Par exemple, discuter la pertinence de la paire $(Titre, Abstract)$ par rapport à la paire $(Abstract, Auteur)$.
7. Effectuez les croisements qui vous semblent intéressants.
8. Pour chaque croisement, quelle est la répercussion sur votre répertoire de travail ?
9. Visualiser la matrice 2D de ces croisements et effectuez un tri par bloc (absolu). Interprétez le résultat.
10. Visualiser la dimension Titre sur une carte. Discutez la méthode et les résultats.
11. Visualiser la relation *co-auteur* sur un graphe. Discutez la méthode et les résultats.

3 Annexes techniques

3.1 Filtres négatifs

La mise en place d'un filtre négatif permet de ne pas considérer certaines valeurs d'un attribut lors d'une analyse. Typiquement, cette technique sera appliquée pour supprimer des mots à la sémantique faible tels que les pronoms, déterminants, etc... Un filtre négatif n'est rien d'autre qu'un fichier texte contenant une chaîne de caractères par ligne. Ces chaînes seront exclues de l'analyse. Le nom donné à ce fichier est libre mais son extension doit **obligatoirement** être : Filtre. Pour mettre en place ce filtre, le nom du fichier associé doit être renseigné dans la case *Liste de mots* (présente lors du calcul de fréquences absolues entre autres) précédé du signe − pour signifier qu'il s'agit d'un filtre négatif. Notez que l'extension ne doit pas apparaître. Ainsi, si le nom est *motsVides.Filtre* seul *motsVides* devra être inscrit.

Enfin, ce filtre est générique dans le sens où il ne dépend pas d'une dimension et est seulement limité par la langue de la collection de documents. Vous avez tout intérêt à le construire petit à petit et à l'utiliser fréquemment dans les analyses.

3.2 Filtres positifs

La mise en place d'un filtre positif permet de ne considérer que certaines valeurs d'un attribut seulement lors d'une analyse. Typiquement, cette technique sera appliquée pour ne considérer que les mots les plus fréquents (voir sous-section suivante pour une méthodologie d'extraction de tels termes). Un filtre positif n'est rien d'autre qu'un fichier texte contenant une chaîne de caractères par ligne. Ces chaînes seules seront considérées dans l'analyse. Le nom donné à ce fichier est libre mais son extension doit **obligatoirement** être : Filtre. Pour mettre en place ce filtre, le nom du fichier associé doit être renseigné dans la case *Liste de mots* (présente lors du calcul de fréquences absolues entre autres).

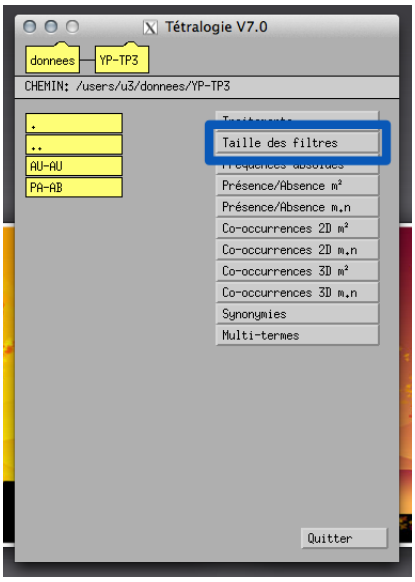

FIGURE 1 – Conserver les occurrences fréquentes

Notez que l'extension ne doit pas apparaître. Ainsi, si le nom est *filtrePos.Filtre* seul *filtrePos* devra être inscrit.

Enfin, ce filtre est beaucoup moins générique dans le sens où il est très lié à la dimension considérée.

3.3 Conservez les occurrences les plus fréquentes

Grâce au menu *Taille des filtres*, voir Figure 1, il est possible de générer automatiquement un filtre positif à partir des termes les plus fréquents. Une fois dans ce menu, il suffit de cliquer sur le ou les attributs dont on veut créer un filtre positif.

La Figure 2 présente un exemple d'interface. Les deux premières colonnes représentent la dimension considérée (version longue et courte). La troisième colonne est cliquable et indique si cette dimension doit être considérée. Le cas échéant, les colonnes 5, 6, 7 et 8 se voient instanciées et doivent être interprétées comme sur la Figure 2. Pour valider la création de ce filtre, il suffit de cliquer sur la case correspondante et de cliquer ensuite sur *EXECUTER*. Autant de fichiers que de cases cliquées seront crées et potentiellement utilisables comme filtres positifs. Si l'on considère la ligne encadrée dans la Figure 2, le nom du fichier créé pour le filtre 1000 sera *0R-14.Filtre*, *i.e.*, le nom court de l'attribut suivit de la fréquence minimale considérée.

3.4 Dictionnaire des synonymes

Un fichier de synonymes peut être créé pour considérer deux termes syntaxiquement différents comme sémantiquement identiques. Pour réaliser cela, un fichier de synonymes devra avoir le format suivant. Chaque ligne sera composé

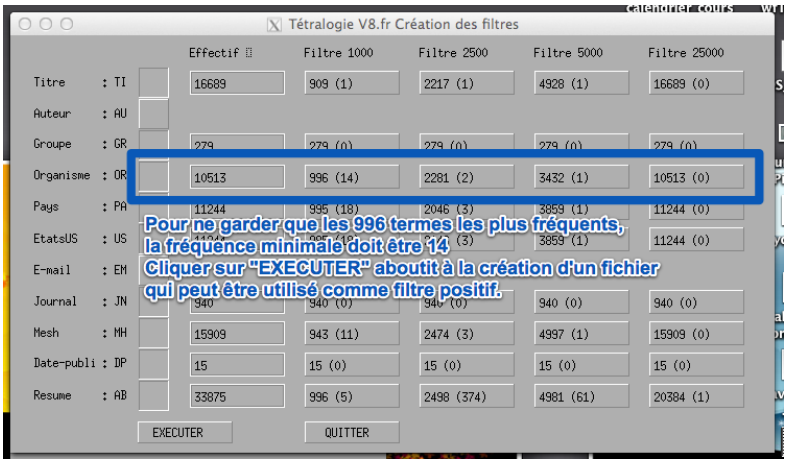

FIGURE 2 – Interface de création de filtre positifs basés sur les occurrences fréquentes

de deux chaînes de caractères séparées par une tabulation. La chaîne de gauche sera la chaîne à remplacer. La chaîne de droite représente le synonyme qui remplacera la chaîne de gauche. Similairement aux filtres, le nom de ce fichier est libre. Son extension est elle contrainte est doit être *Syn*.

Pour mettre en place ce dictionnaire de synonymes, le nom du fichier associé doit être renseigné dans la case *Synonymes* (présente lors du calcul de fréquences absolues entre autres). Notez que l'extension ne doit pas apparaître. Ainsi, si le nom est *AU.Syn* seul *AU* devra être inscrit.

L'utilisation de ces dictionnaires peut être pertinente pour répondre à certains besoins tels que la gestion des hiérarchies, l'uniformisation des noms des structures de recherche, etc...

Grâce au menu de synonymage (Figure 3), vous créez automatiquement à partir des "champ.ind" des synonymes par champ obtenu après l'opération de comptage simple. Il vous faudra relancer de nouveaux comptage en tenant compte des fichiers de synonymes. Il est préférable de faire vérifier le synonymage par un expert du domaine.

3.5 Effectuez un croisement

Le menu *Co-occurrences 2D m.n* (Figure 4) permet d'accèder à l'interface de calcul des co-occurrences (Figure 5). Le principe est simple, il suffit de cliquer sur la case correspondant à l'intersection des attributs que l'on souhaite croiser puis de cliquer ensuite sur *EXECUTER*. Dans la mesure où les répercussions de cette action font l'objet d'une question du présent TP, elles ne sont pas détaillées.

Notez que les filtres et les dictionnaires de synonymes peuvent également être spécifiés ici.

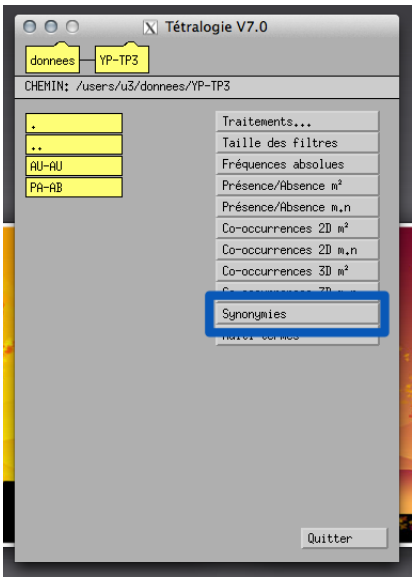

FIGURE 3 – Synonymes

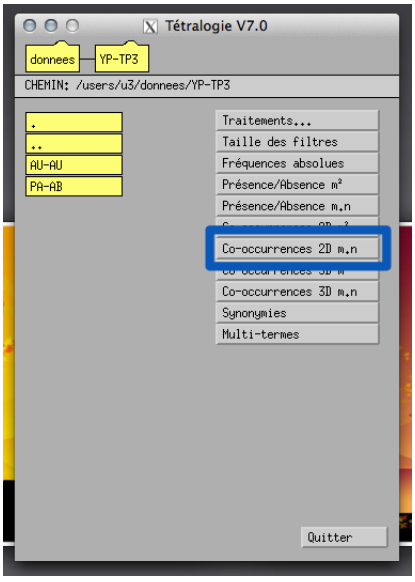

FIGURE 4 – Accéder au menu de calcul des co-occurrences

FIGURE 5 – Interface de calcul de co-occurrences

3.6 Visualisation : quelques techniques

Une fois les croisements souhaités effectués, il est possible de visualiser le résultat de ces croisements. Quelques exemples sont détaillés ci-dessous (Figure 6).

3.6.1 Visualisation sous forme de matrice 2D

La première technique de visualisation est la matrice 2D (Figure 7 et 8). Typiquement pour chaque intersection, le nombre de documents dans lesquels les deux termes apparaissent est inscrit. Un clic droit sur la matrice engendre l'ouverture d'un menu contextuel. Celui permet de manipuler la matrice. Parmi les différents outils proposés, deux sont particulièrement utiles (Figure 7) : le tri (absolu ou relatif) par bloc permet de réordonner la matrice pour obtenir des zones denses et faire émerger des *clusters* et la fonction *Zoom* qui, comme son nom ne l'indique pas, permet de *Dezoomer* la matrice et d'avoir une vision globale de celle-ci (Figure 8). Cette fonction est particulièrement utile après avoir effectué un tri par blocs pour vérifier la pertinence de ce tri.

3.6.2 Visualisation sous forme de carte

Lorsque la première dimension du croisement contient une information spatiale, celle-ci peut être visualisée sur une carte (Figure 9). Ici aussi, l'utilisation de cette fonction fait l'objet d'une question dans ce TP et n'est donc pas plus détaillée.

3.6.3 Visualisation sous forme de graphe

Enfin, il est possible de visualiser les co-occurrences à l'aide d'un graphe (Figure 10). Cela s'avère particulièrement utile lorsque l'on veut croiser une dimension avec elle-même (auteur-auteur par exemple).

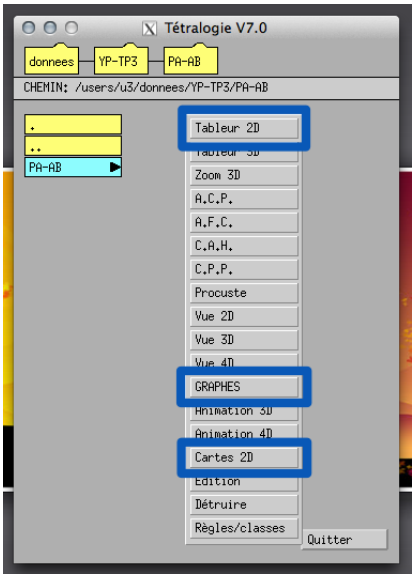

FIGURE 6 – Quelques outils de visualisation pratiques

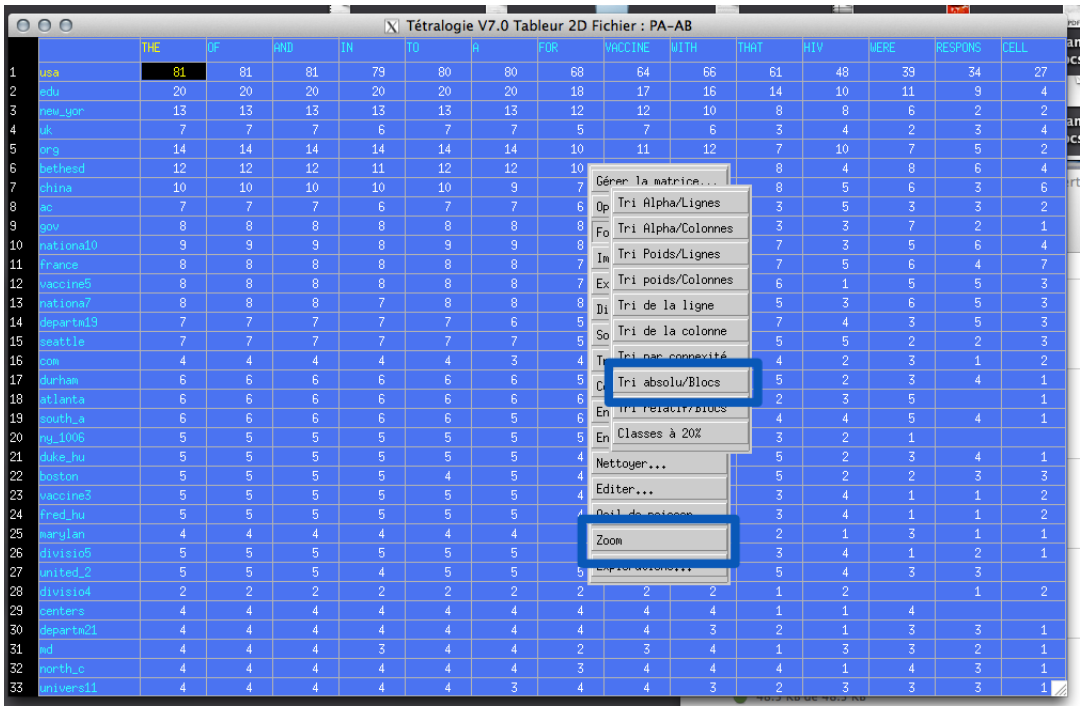

FIGURE 7 – Fonctions intéressantes dans le menu contextuel associé à la visualisation de type matrice 2D

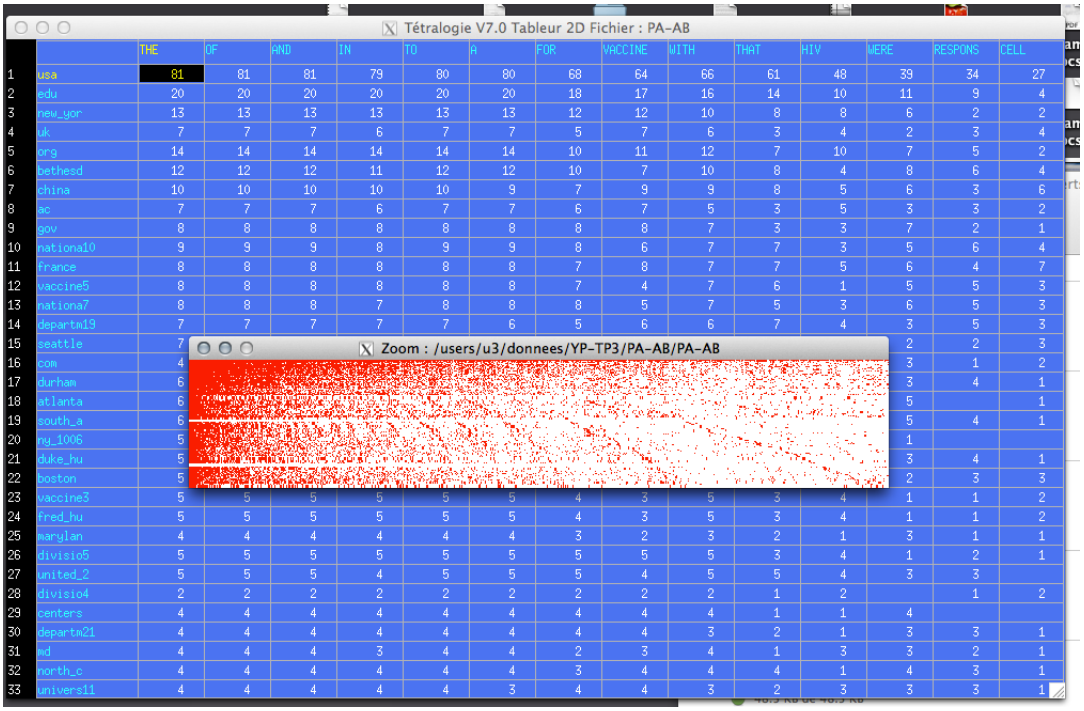

FIGURE 8 – Illustratin de la fonction Zoom

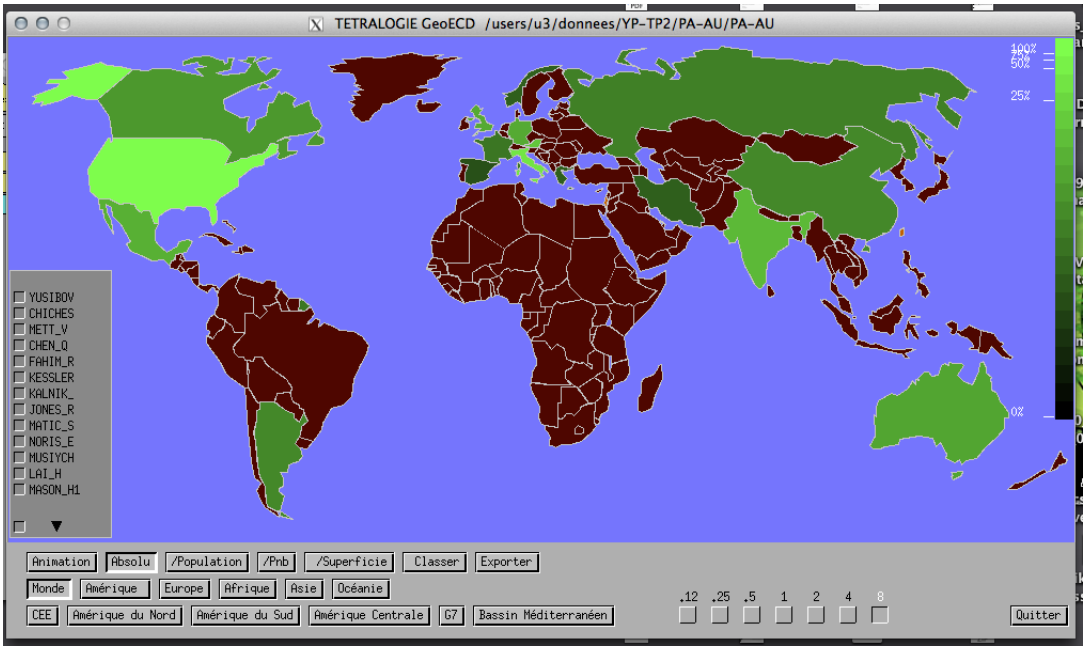

FIGURE 9 – Visualisation cartographique de l'information

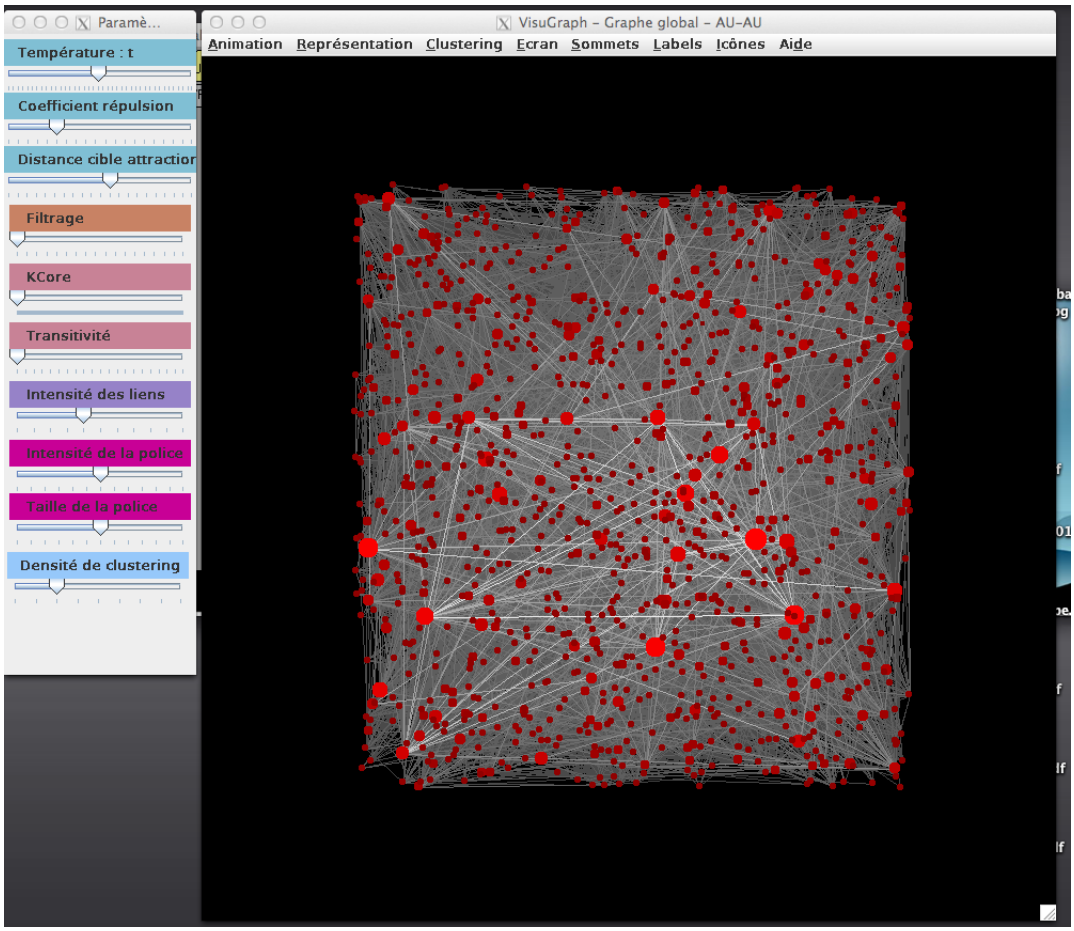

FIGURE 10 – Visualisation de la relation co-auteur