# Travail pratique 3

## Clustering… et article scientifique

Vous avez bâti un ensemble de données impressionnant. Vous pouvez maintenant l’utiliser pour faire des expériences. La méthode que je vous demande d’explorer est le partitionnement des données à l’aide de centroïdes (Centroid-based data clustering).

Vous devrez tester plusieurs configurations différentes, les principales étant la taille de fenêtre (de cooccurrences) et le nombre de centroïdes, mais la taille du corpus d’entraînement pour les cooccurrences aura un effet sur votre capacité à effectuer des tests.

#### Article scientifique

Vous devrez écrire un article scientifique qui relatera vos expériences. Il devra prendre la forme suivante :

* Liste d’auteurs
* Résumé (abstract)
* Introduction
  + Quelles questions vous êtes-vous posées? Quelles sont vos hypothèses? Pourquoi ce problème est-il intéressant? Expliquez pourquoi vous pensez que vos expériences permettront de calmer votre curiosité.
    - Cette section présente habituellement la revue de la littérature scientifique sur laquelle vous basez votre raisonnement. Je vous épargne cette exigence.
* Matériel et méthodes
  + Ici, vous devez spécifier l’ensemble du matériel (données) à l’aide duquel vous effectuerai vos expériences et comment il a été construit, si applicable.
  + Vous devez aussi présenter la méthode que vous allez utiliser et pourquoi.
    - Du pseudo-code est de rigueur ici.
* Résultats
  + Vos résultats sous forme de tableaux et/ou de graphiques.
    - Prenez soin de bien identifier les données que vous présentez.
* Discussion
  + Suite à vos expériences, votre but a-t-il été atteint? Croyez-vous que votre hypothèse a été vérifiée ou infirmée? Pourquoi? Est-ce que vous pourriez faire les choses autrement dans le futur? Comment?

Le pdf d’un article scientifique vous a été fourni comme exemple. De plus, prenez soin de la qualité de votre français. Des points y seront attribués. Vous devez utiliser Word.

#### Code

Vous devez fournir tout ce qui est nécessaire afin de pouvoir reproduire vos expériences. Laissez les données que vous avez utilisées sur votre BD et arrangez-vous pour laisser une connexion valide dans votre code. Les vecteurs de cooccurrences doivent donc DÉJÀ être calculés avant la remise, i.e. le lancement de votre code ne doit pas engendrer un réentraînement des cooccurrences.

#### Type d’application

Ligne de commande, voir le fichier resultats.txt. La ligne de commande et la section à la toute fin sont les parties importantes. Dans le milieu, il y a la trace de l’exécution du programme. À chaque étape de son exécution, je vous suggère de garder une trace de l’état interne de votre programme. Elle peut prendre la forme de votre choix, mais soyez concis.

Vous devez afficher -n (voir la section suivante) mots par cluster à la fin de l’exécution de votre programme. Prenez donc soin de regrouper les mots par cluster.

#### Arguments

Vous devez gérer des options. Elles peuvent être fournies dans n’importe quel ordre.

* -t <*taille*> : taille de fenêtre. <*taille*> doit suivre -t, précédé d’un espace
* -nc <*nombre*> : nombre de centroïdes, une valeur entière.  
    
  OU BIEN  
    
  -m <"*mot1 mot2 mot3*">: liste de mots qui serviront de centroïdes
* -n <*nombre*>: nombre de mots à afficher par cluster (à la fin de l’exécution)

#### Tests

Ne vous gênez pas de demander beaucoup de votre ordinateur. C’est-à-dire que vous pouvez faire des tests avec 10, 20, 50, 100 centroïdes, ou plus encore. Vous verrez la nécessité d’afficher des états internes lorsque vos tests nécessiteront plusieurs heures (!) à effectuer. Le fichier resultats.txt montre la trace d’une exécution typique. Notez le temps d’exécution à la toute fin. Cette configuration ne demandait qu’une taille de fenêtre de 5 et 5 centroïdes. De plus, seules les cooccurrences du corpus « Le Germinal » servaient à faire le partitionnement. Attendez-vous donc à lancer vos tests avant de vous coucher et de récupérer vos résultats le lendemain. Vous feriez bien de sauvegarder vos résultats dans des fichiers à part.

#### Équipes

Gardez les mêmes équipes que pour le TP2.

#### À remettre

L’article scientifique, les fichiers source Python, les scripts Oracle (création de table et autre), la stopliste et un README spécifiant les configs qui fonctionnent. N’oubliez pas que les tests se feront à la ligne de commande, donc le path pour la stopliste doit fonctionner dans ce contexte.

Cette remise se fera sur Bitbucket. S.V.P. laissez le tout dans le même dépôt (repository) que lors du TP2.

#### Quand

Le 13 décembre, avant minuit.