# 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. La taille du corpus d’entraînement pour les cooccurrences aura un effet sur la qualité de vos tests.

En plus de tester des valeurs de configuration différentes, vous devez également innover. Par exemple :

* Optimisation;
* Utilisation des données à d’autres fins;
* Utilisation d’algorithmes d’IA supplémentaires;
* Etc.

Vous devez documenter vos recherches.

#### 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 ? Vous pouvez en reprendre parmi celles qu’on a énoncées en classe que vous ajouterez à des nouvelles pour le volet innovation. De plus, vous devez aussi vous demander pourquoi ce problème est intéressant et expliquer 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 effectuerez vos expériences et comment il a été construit, si applicable.
  + Vous devez aussi présenter les méthodes que vous allez utiliser et pourquoi.
    - Du pseudo-code est de rigueur ici.
* Résultats
  + Vos résultats devront être 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 ? Êtes-vous en mesure de faire une analyse qualitative de vos résultats ? 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 (pas obligé d’être sur deux colonnes). De plus, prenez soin 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.

#### 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 les plus 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.

* -c : clustering
* -t <*taille*> : taille de fenêtre. <*taille*> doit suivre -t, précédé d’un espace
* -n <*nombre*>: nombre de mots à afficher par cluster (à la fin de l’exécution)
* --nc <*nombre*> : nombre de centroïdes, une valeur entière. Avec cette option, vous devez générer <*nombre*> points aléatoirement (qui ne sont donc pas des mots) dans l’espace de recherche, i.e. l’univers de dimension N, où N est le nombre de mots dans votre vocabulaire. Les points ainsi générés aléatoirement sont des centroïdes.

OU BIEN

--mots <"*mot1 mot2 mot3*">: liste de <*nombre*> mots qui serviront de centroïdes initiaux

*NB : Vous devez donner soit --nc ou --mots, mais pas les deux.*

Exemple d’appel : Y:\Cooccurrences\src>main.py -c -t 5 -n 10 --nc 5

#### 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 que 5 centroïdes. Attendez-vous donc à lancer vos tests avant de vous coucher et de récupérer vos résultats le lendemain. Vous feriez bien de rediriger vos résultats dans des fichiers à part.

Exemple d’appel :

Y:\Cooccurrences\src>main.py -c -t 5 -n 10 --nc 5 > resultats\_t5\_n10\_nc5.txt

#### 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) et un README spécifiant les configs qui fonctionnent. N’oubliez pas que les tests se feront à la ligne de commande. Cette remise se fera sur Bitbucket. S.V.P. laissez le tout dans le même dépôt (repository) que lors du TP2.

#### Grille de correction

|  |  |
| --- | --- |
| Code | 10% |
| Exécution | 10% |
| Qualité des résultats pour le clustering de base | 10% |
| Qualité des tests (clustering ET **innovation**) | 20% |
| Tableaux et graphes (résultats) | 10% |
| Analyse des résultats de tests de base et pour l’**innovation** (discussion) | 20% |
| Contenu (résumé, intro, matériel, méthodes) | 10% |
| Qualité du français | 10% |