# 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 de 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 et la qualité du corpus d’entraînement pour les cooccurrences auront aussi un effet sur vos tests.

#### Présentation

Vous devez documenter vos résultats sous forme d'un mini-rapport de laboratoire. Il devra prendre la forme suivante:

* Page titre avec la liste des auteurs
* Résultats
  + Vos résultats devront être sous forme de tableaux et de graphiques.
    - Prenez soin de bien identifier les données que vous présentez.
      * i.e. quelle configuration (taille de fenêtre, nombre de centroïdes, textes, etc.) a mené à de tels résultats
        + Ne vous gênez pas de présenter beaucoup de données.
* Discussion
  + Analysez vos résultats de manière qualitative.
    - Comparez les résultats des différentes configurations.
    - Pourquoi les configurations mènent-elles à des résultats différents?
    - Que pensez-vous de vos expériences?
    - Etc.

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

Cette application à la ligne de commande doit imprimer les résultats une fois les calculs terminés. Voir le fichier resultats.txt. La section à la toute fin est la plus importante. Je vous conseille toutefois d'imprimer une trace concise des états internes (itérations) de votre programme, tel qu'on peut voir au milieu du fichier.

Vous devez afficher -n (voir la section suivante) mots par cluster à la fin de l’exécution de votre programme. Triez-les par ordre de proximité du centroïde auquel le mot appartient.

#### 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 maximal de mots à afficher par cluster (à la fin de l’exécution)
* -k <*nombre*> : nombre de centroïdes, une valeur entière. Avec cette option, vous devez choisir <*nombre*> vecteur(s) de votre matrice aléatoirement. Ces vecteurs sont les coordonnées initiales des centroïdes.

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

#### 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. Vous feriez bien de rediriger vos résultats dans des fichiers à part.

Exemple de redirection :

Y:\Cooccurrences\src>main.py -c -t 5 -k 5 -n 10 > resultats\_t5\_k5\_n10.txt

#### Équipes

Gardez les mêmes équipes que pour le TP2.

#### À remettre

Le mini-rapport, les fichiers source Python et un README spécifiant les configs qui fonctionnent (et vos noms, bien sûr). N’oubliez pas que les tests se feront à la ligne de commande. Cette remise se fera sur Bitbucket ou GitHub. 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 | 20% |
| Tableaux et graphes (résultats) | 20% |
| Analyse des résultats de tests | 20% |
| Qualité du français | 10% |