Travail pratique 2

Synonymes… BD

Pour le TP1, vous aviez créé un système qui faisait la comparaison des contextes de mots. Dans le TP2, vous devrez emmagasiner dans une BD les données qui vous ont permis d'effectuer ces comparaisons.

On veut pouvoir entraîner notre système sur un texte, avec une taille de fenêtre fixe. Stocker les résultats et recommencer. Chaque fois que vous fournissez un texte, la taille du vocabulaire change, donc votre matrice change de dimensions. Vous devez donc vraisemblablement mettre à jour votre vocabulaire avant de calculer les cooccurrences.

Un point TRÈS important, la matrice que vous utilisiez dans le TP1 était assez creuse, c'est-à-dire que la vaste majorité des valeurs dans votre matrice sont 0 (99.9% en fait). Ne conservez donc pas les valeurs nulles, elles prendront trop d'espace mémoire.

Tables

Le nombre de tables et leur structure relève de votre choix, mais n'oubliez pas les contraintes là où elles sont nécessaires.

Type d'application

Vous devez créer une application de type ligne de commande. Remarquez qu'on lance une commande séparée pour chaque corpus individuel, mais que la base de données peut maintenant contenir les données cumulatives de plusieurs textes. Lorsque vous avez entraîné votre modèle sur trois texte, en trois entraînements, ça doit être comme si vous l'aviez entraîné sur un seul gros texte contenant ces trois textes. Notez aussi que l'entraînement et la recherche de « synonymes » se fait en deux exécutions **distinctes**. Vous devez également ajouter la possibilité de régénérer la base de données, également dans une exécution distincte.

Arguments

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

Options pour l'entraînement :

- -e : entraînement
- -t <*taille*> : taille de fenêtre. <*taille*> doit suivre -t, précédé d'un espace.
- --enc <*encodage*> : encodage de fichier. <*encodage*> doit suivre --enc, précédé d'un espace.
- --chemin <*chemin*> : chemin du corpus d'entraînement. <*chemin*> doit suivre --chemin, précédé d'un espace.

Note : les options qui contiennent plus d'une lettre doivent être précédées de 2 tirets afin de respecter la syntaxe GNU/POSIX.

Lorsqu'on fournit l'option -e, on DOIT fournir les options -t, --enc et --chemin, avec leur argument respectif. Le système analysera le corpus et ajoutera les nouveaux mots de vocabulaire et les nouvelles cooccurrences pour cette taille de fenêtre dans la BD, suite à quoi il arrêtera son exécution.

Notez bien que les cooccurrences pour la fenêtre de taille 5, par exemple, et celles pour la taille de fenêtre 7, par exemple, ne sont pas exactement les mêmes…

Exemple d'appel pour l'entraînement :

Y:\Cooccurrences\src>mainBD.py -e -t 5 --enc utf-8 --chemin ..\textes\GerminalUTF8.txt

Options pour la recherche :

- -r : rechercher des synonymes
- -t <*taille*> : taille de fenêtre. <*taille*> doit suivre -t, précédé d'un espace.

Lorsqu'on fournit l'option -r, on fournit l'option -t, avec la taille, évidemment. Le système chargera le vocabulaire et les cooccurrences pour cette taille de fenêtre. Il demandera ensuite à l'utilisateur un mot, le nombre de résultats à afficher et la méthode de calcul, comme au TP1.

Exemple d'appel pour la recherche :

Y: \Cooccurrences\src>mainBD.py -r -t 5

Option pour la régénération de la BD :

- -b : régénérer la BD

Lorsqu'on fournit l'option -b, elle doit apparaître seule. Son effet est que toutes les tables de la base de données sont détruites et ensuite recréées, vides.

Exemple d'appel pour la régénération:

Y: \Cooccurrences\src>mainBD.py -b

Vous DEVEZ vous conformer aux consignes pour la ligne de commande. Je vous conseille de vous occuper de ces consignes avant tout le reste. Des modules existent pour le traitement d'arguments de ligne de commande.

Connecteur BD

Vous devez utiliser sqlite3.

Équipes

Conservez les mêmes équipes qu'au TP1.

À remettre

Les fichiers source Python (incluant ceux pour la création de tables et autre).

Vous devez me rendre votre dépôt accessible via Bitbucket ou GitHib. Donc, S.V.P. donnez les droits de lecture à l'utilisateur identifié par l'adresse ppmonty@cvm.qc.ca

N'oubliez pas de faire le ménage des fichiers inutiles sur votre dépôt et d'inclure, dans le répertoire racine, un README où doivent figurer les noms de tous les membres de votre équipe et toute autre information nécessaire à la mise en place de votre système.