SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Projet 2 – projet *online*. Classification par indice de centralité.

BM Bui-Xuan

ATTENTION CHOIX DE PROJET : Ce projet peut être fusionné avec le projet final *cloud*. Chaque groupe est libre de choisir entre :

— Choix A : rendre séparément ce projet *offline* et le projet final *cloud* ;

— Choix B : ne rendre qu'un seul projet parmi ces deux, dans ce cas, le projet en question doit obligatoirement être le projet final *cloud*.

Cependant, ce choix doit être explicitement rappelé à la première page de chaque rapport concerné par ce choix.

Index des documents textuels : Dans ce projet, tout document textuel D est considéré comme une liste $l(D)$ de mots de l'alphabet latin `[A-Za-z]` séparés par des caractères pris dans la liste `[^A-Za-z]` [1]. Alors, l'index de D est l'ensemble

$index(D) = \{(m,k)$: le mot m apparaît k fois dans la liste $l(D)\} \cup \{(m,0)$: le mot m n'apparaît pas dans la liste $l(D)\}$.

Distance de Jaccard : Soit deux documents textuels D_1 et D_2, nous définissons la distance entre D_1 et D_2 comme :

$$d(D_1,D_2) = \frac{\sum_{(m,k_1)\in index(D_1)\wedge(m,k_2)\in index(D_2)} \max(k_1,k_2) - \min(k_1,k_2)}{\sum_{(m,k_1)\in index(D_1)\wedge(m,k_2)\in index(D_2)} \max(k_1,k_2)}.$$

Graphe géométrique : Un graphe géométrique dans un espace métrique de fonction de distance d est défini par un ensemble de points dans l'espace métrique appelés sommets, et un seuil sur la distance entre les points : il existe une arête entre deux sommets si et seulement si la distance d dans l'espace métrique entre les deux sommets est inférieure à ce seuil.

Indice de centralité de Jaccard : Etant donné un ensemble de documents textuels et un seuil appelé seuil de similarité θ, nous définissons le graphe de Jaccard de ces documents comme le graphe géométrique de seuil θ et de fonction distance celle de Jaccard dans le modèle de graphe géométrique défini par ces documents textuels. Alors, l'indice de centralité de Jaccard de ces documents textuels est une fonction associant tout document à un réel dans $[0,1]$ exprimant une certaine notion de centralité de chaque document. Pendant le cours, nous avons vu au moins 3 indices de centralité différents : (séance 4) indice dit de *closeness centrality* ; (séance 4) indice dit de *betweenness centrality* ; (séance 5) indice spectral de la théorie des chaîne de Markov, dit de *pagerank*.

L'objectif de ce projet est double :

— d'abord, il s'agit de calculer la matrice de distance de Jaccard entre toute paire de documents textuels d'une cinquantaine de livres ;

— puis, proposer un tri de ces livres par ordre d'indice de centralité décroissant. Chaque groupe est libre de décider l'indice de centralité utilisé dans le projet. Cependant, il est important de bien le décrire dans le rapport. On veillera en particulier à expliciter la définition, le calcul, ainsi que le résultat de cet indice sur la base de ces cinquantaine de livres.

1. Ceci est plus spécifique que la situation du projet 1 *offline*, où un document textuel est la plupart du temps considéré comme : une liste de chaînes de caractères séparés par le caractère '`\n`' ; autrement dit, des lignes de chaînes de caractères séparées par le caractère '`\n`'.

1 L'énoncé du projet 2 – projet *online*

ATTENTION : Il est obligatoire d'indiquer à la première page du rapport le "CHOIX A" dans le rendu de projet.

Il s'agit de construire une base de données d'une bibliothèque personnelle de livres, stockés sous forme de documents textuels. La taille minimum de chaque bibliothèque est 50 livres. La taille minimum de chaque livre est 10000 mots. L'objectif principal de ce projet peut être vu comme un ordre de lecture de ces livres en commençant par ceux jugés les plus "centraux" et terminant par ceux les moins "centraux" dans le graphe géométrique de l'espace de Jaccard défini par ces livres. Le critère de centralité est laissé libre à chaque groupe d'interpréter. Cependant, il est obligatoire d'utiliser au moins un critère parmi les trois indices vu en cours. Plus précisément, il faut obligatoirement fournir un premier classement de ces livres par l'un des trois indices de *closeness*, *betweenness*, ou *pagerank*. Puis, chaque groupe est libre de fournir (ou pas) un deuxième classement par l'indice de son choix.

Un effort particulier doit être mis sur la phase de test et celle de la rédaction du rapport. Il y aura ainsi deux notations distinctes pour le rendu de ce projet : une note pour la réalisation ($\approx 20\%$ de la note totale de l'UE) ; et une note pour le rapport ($\approx 10\%$ de la note totale de l'UE).

On veillera à expliciter les points suivants, pour chaque fonction implémentée, e.g. celle calculant la distance de Jaccard et celle calculant l'indice de centralité :

— définition du problème et la structure de données utilisée.

— analyse et présentation théorique des algorithmes connus dans la littérature.

— argumentation concise appuyant toute appréciation, amélioration, ou critique à propos de ces algorithmes existants dans la littératures.

— partie test : méthode d'obtention des *testbeds*. En particulier, il est important de citer la provenance des fichiers sources, e.g. pour les documents textuels.

— test de performance : mieux vaut privilégier les courbes, diagrammes bâton (moyenne + écart type) et diagrammes de fréquence, plutôt qu'exhiber les colonnes de chiffre sans fins...

— une discussion sur les résultats de test de performance est toujours la très bienvenue.

— conclusion et perspectives sur ce problème de classement par similarité d'une bibliothèque de documents textuels (a.k.a. le cas *online* des moteurs de recherche de documents textuels).

Il est prudent d'avoir entre 5 et 10 pages pour un rapport avec un contenu moyen. La limitation formelle pour ce rapport est 12 pages. Il convient de bien respecter cette limitation : les pages 13+ ne seront pas lues !

Contraintes :

— A réaliser en binôme ou en individuel.

— Archiver la totalité du rendu en un seul fichier compressé contenant de la documentation (rapport $\approx$ 5-10 pages), un binaire et un README expliquant comment exécuter le binaire, le code source commenté, un Makefile (ou Apache Ant si Java), un répertoire contenant les instances de test, et tout ce dont on juge utile à la lecture du projet sans toute fois dépasser la dizaine de Méga-octet.

— Envoyer ce fichier à `buixuan@lip6.fr`, 3 emails maximum par groupe. L'utilisation des hébergeurs en ligne (drive et compagnies) est proscrite. La nomination de préférence est daar-CHOIX-A-projet-online-NOM.piki, où piki peut être un élément de $\{tgz, zip, rar, 7z, etc\}$.

— Deadline : 02 Décembre 2019, 13h45, cachet de serveur de messagerie faisant foi. Pénalité de retard : malus de $2^{h/24}$ points pour h heures de retard.