Sujet : Evaluation des résultats d'un co-clustering

francois.role@parisdescartes.fr

Janvier 2020

1 Contexte

Quand on fait un co-clustering de textes, il est intéressant de ne pas se limiter à évaluer la qualité des clusters de documents.

2 Travail à réaliser

Vous effectuerez le co-clustering de plusieurs corpus de textes (qui vous seront fournis quand les projets démarreront)

Vous calculerez la similarité entre termes pour toutes les paires de termes (vous pourrez pour cela utiliser des *word embeddings*). On considèrera comme un bon choix (TP) le fait d'avoir mis dans le meme cluster deux termes dont la similarité dépasse un seuil α (hyperparamètre du programme que vous ferez varier dans vos expériences). La décision de mettre dans le même cluster deux termes très différents sera notée FP. La décision de mettre dans des clusters différents deux termes très différents sera notée TN. La décision de mettre dans des clusters différents deux termes très similaires sera notée FN. Vous prendrez en compte les $N(N-1)/2$ paires de termes s'il y a N termes dans le corpus. Vous calculerez alors TP + TN / TP + TN + FP + FN.

Pour chaque co-clustering, vous comparerez les valeurs obtenues avec la méthode ci-dessus avec les mesures classiques (ARI, NMI, Clustering Accuracy) de qualité des clusters de documents obtenus avec ce co-clustering.

Vous afficherez également pour chaque cluster les L paires de termes correspondant aux plus mauvaises décisions FP et FN.

Vous utiliserez les algorithmes CoclusMod, CoclustModFuzzy et CoclustInfo du package Coclust. Dans le cas de CoclustModFuzzy, vous utiliserez les assignations strictes renvoyées par le programme mais vous exploiterez aussi les probabilités finales qu'il fournit.

3 Outils

- package Python coclust

4 Rendu

Un notebook Jupyter contenant les affichages et les analyses.

5 Critères d'évaluation

Qualité du code, qualité des expériences et des conclusions (insérées dans des cellules de texte du notebook) que vous pourrez tirer suite à ces expériences.