**Compte rendu projet de RI**

Vous trouverez dans cette archive le code que nous avons écrit pour les trois premiers TME de RITAL.

Nous avons testé notre code sur l’extrait Camcshort pour le TME 1 et 2 puis nous sommes passés sur les données complètes de cisi et de cacm pour tester l’ensemble des trois TMEs.

Organisation du code :

Dans TME1\_2\_Projet.py : Code des TME 1 et 2 en ce qui concerne le projet

Dans TME3.py : Code du projet pour le TME 3 et tests sur les données cisi et cacm. Les résultats sont présents en dessous de la ligne exécutant le test en commentaire.

Dans exercice1.py : Code des exercices 1 des TME. Cependant, nous avons eu un problème.

Dans le fichier Test.py, il y a le même code que le fichier TME1\_2\_Proje.py mais il nous servait à tester sur le fichier Camcshort.

Les .txt d’index sont liés à l’exercice 1 du premier TME.

Les autres fichiers sont les fichiers que vous nous avez fournis.

Difficultés :

* Il nous semble que les modèles de RI renvoient des scores donc nous avons dû fixer arbitrairement une limite de documents renvoyés pour simuler le fait que le modèle ne renvoie que les documents ayant les meilleurs scores
* Nous avons remarqué qu’il y avait des query sans documents pertinents comme le 105 dans cisi. Cela nous a posé des problèmes pour certaines mesures comme le rappel, la précision moyenne.
* On a aussi eu beaucoup de mal à déterminer si nos résultats attendus semblaient corrects vu la taille des corpus. C’est pour cela qu’on avait commencé sur l’extrait de cacmShort.
* Nous n’avons pas réussi à faire tourner le modèle de langue sur autre chose que cacmshort car cela prenait trop de temps.
* Nous n’avons pas fait les bonus.

Résultats :

Les résultats du TME 3 se trouvent à la fin du fichier python du TME3. Nous avons analysé cisi puis cacm.