**USTHB, département d’Informatique 2018/2019**

**Série d’exercices de TP**

**Module : RI**

Créez un répertoire appelé « TPRI », créez dans ce répertoire N documents de type texte. Chaque document i est nommé « Di » exemple le document 1 est D1, le document 2 est D2 …etc. Mettez dans chaque document le texte que vous voulez. Ce répertoire sera notre collection de travail pour cette série d’exercices.

Réalisez les exercices suivants avec le langage Python.

**Exercice 1**

Ecrire un programme python qui permet de (en supprimant les mots vides et les ponctuations) :

1. Créer et afficher l’index des fréquences pour chaque document.
2. Créer et afficher le Fichier Inversé des fréquences de la collection.

**Exercice 2**

En utilisant la structure du Fichier Inversé obtenue dans l’exercice 1, programmer deux fonctions d’accès de base. La première reçoit un numéro de document, et retourne la liste de mots avec leurs fréquences. La deuxième reçoit un mot, et retourne la liste de documents et les fréquences d’occurrences du mot dans chaque document.

**Exercice 3**

En utilisant la structure du Fichier Inversé obtenue dans l’exercice 1, créer et afficher le Fichier Inversé de la collection par la pondération TF\*IDF. Utilisez la formule : poids(ti, dj)=(freq(ti,dj)/Max(freq(t, dj))\*Log((N/ni) +1)

**Exercice 4**

En utilisant la structure du Fichier Inversé obtenue dans l’exercice 3 par pondération TF\*IDF, programmer deux fonctions d’accès de base. La première reçoit un numéro de document, et retourne la liste de mots avec leurs poids. La deuxième reçoit un mot, et retourne la liste de documents et les poids du mot dans chaque document.