CC52D: Recuperación de Información
Tarea 1

Profesores: Ricardo Baeza Yates y Carlos Castillo

Plazo de entrega: 28 de Mayo

Escriba un programa que:

- Extraiga el vocabulario de un texto, normalizando todo a minúsculas y eliminando palabras vacías (en inglés).
- Usando lo anterior, obtener a partir del vector asociado a un documento (decida usted si quiere darle más peso a algunas partes de cada documento) y poder calcular la distancia coseno entre dos documentos.
- Utilizando la colección de Fibrosis Cística en sunsite.dcc.uchile.cl/irbook/ obtenga:
  - Los 10 documentos más parecidos al tercer documento en la colección.
  - Tomando las consultas como documentos, obtenga las respuestas de su sistema y calcule el gráfico de precisión-recuperación para las últimas tres consultas y el gráfico de precisión-recuperación normalizado del promedio de estas tres consultas, en base a las respuestas ya evaluadas de acuerdo a su relevancia.

Se sugiere usar la colección en formato XML aunque también es válido usar el formato en texto plano.