Laboratorio 3: Recuperación de Información

Objetivo: Pre-procesar la colección CACM utilizando los aprendido en los laboratorios 1 y 2

Materiales:

Instalar Python y NLTK, terminar laboratorios 1 y 2

Descripción:

1. Descargar la colección CACM

   http://ir.dcs.gla.ac.uk/resources/test_collections/cacm/

2. Encontrarás los siguientes archivos:

a) **README** - Short details on the files
b) **cacm.all** - Text of documents
c) **cite.info** - Key to citation info
d) **common_words** - Stop words used by smart
e) **qrels.text** - List of relevance judgements
f) **query.text** - Original text of the query

3. Realiza el mismo pre-procesamiento de la práctica 1 y 2 a los documentos y consultas de la colección. Para facilitar las operaciones, construyan un archivo para consultas y otro para documentos con el siguiente formato:

   Número consulta/documento | texto

   Por ejemplo, para las dos consultas tendrían:

   1 | título | texto

   2 | the relationship of blood and cerebrospinal fluid oxygen concentrations or partial pressures| a method of interest is polarography.

   De esta manera en un archivo tendrán todas las consultas y en otro todos los documentos facilitando su pre procesamiento. Pre-procesen los archivos generados y escriban los documentos y consultas pre-procesados a archivos de salida.

   4.Escribe tu reporte en el formato establecido y colócalo en Teams, asegúrate de ilustrar las salidas obtenidas. Coloca también como parte de tu entrega el código generado.