GUIONES DE PRÁCTICAS – MÓDULO 3

Práctica 4

Contexto

Para implementar búsquedas eficientes en documentos se suelen utilizar índices invertidos. De esta forma a partir de un conjunto de documentos se obtiene una lista de palabras y se asocia con cada palabra los documentos en los que aparece. Este tipo de índice se dice que es a nivel de registro, existe otra variante en la que además del documento se asocia con la palabra la posición que ésta ocupa dentro del documento, es el índice invertido a nivel de palabra.

Suponemos que tenemos los siguientes documentos con sus datos:
Doc1: IMF, Financial Economics Crisis
Doc2: IMF, Financial Crisis
Doc3: Harry Economics
Doc4: Financial Harry Potter Film
Doc5: Harry Potter Crisis

Su índice invertido a nivel de registro:
IMF → Doc1, Doc2
Financial → Doc1, Doc2, Doc4
Economics → Doc1, Doc3
Crisis → Doc1, Doc2, Doc5
Harry → Doc3 Doc4, Doc5
Potter → Doc4, Doc5
Film → Doc4

Su índice invertido a nivel de palabra sería el siguiente
IMF → Doc1:1, Doc2:1
Financial → Doc1:6, Doc2:6 Doc4:1
Economics → Doc1:16, Doc3:7
Crisis → Doc1:26, Doc2:16, Doc5:14
Harry → Doc3:1 Doc4:11, Doc5:1
Potter → Doc4:17, Doc5:7
Film → Doc4:24

Ejercicio

Considerando los ficheros *texto1.txt, texto2.txt* y *texto3.txt* vamos a obtener su índice invertido a nivel de registro pero considerando solo una serie de palabras clave que nos interesan. Las palabras claves están en el fichero *claves.txt.* Hay que tener en cuenta que la palabra puede aparecer tanto en mayúscula como en minúscula y se deben eliminar los signos de puntuación. Como nombre de archivo que se muestre solo el nombre sin la ruta y que no aparezca repetido el nombre de un archivo en una misma palabra, si ésta aparece más de una vez en él. En los ficheros se han eliminado las tildes para evitar problemas.

Entrega

Se debe entregar un fichero pdf en el que se especifique el enunciado de cada ejercicio, la solución en spark (con algún comentario explicativo) y captura de pantalla con la ejecución y los resultados obtenidos en la Shell.