**DICCIONARIO DE SAR**

* Recuperación de información: arquitecturas o algoritmos utilizados para encontrar elementos dentro de una gran cantidad (no estructurada) de los mismos, los cuales satisfacen una condición determinada.
* Índices: estructuras creadas para realizar búsquedas rápidas. Se crearon inicialmente para representar conjuntos de categorías. Cuando aparece la informática se empiezan a crear grandes volúmenes de estos de manera automática.
* Consultas: expresiones booleanas de términos que se combinan con operadores AND, OR Y NOT (ejemplo: Brutus AND Caesar AND NOT Calpurina).
* Corpus: conjunto cerrado de textos o datos destinados a la investigación científica. También recopilación de libros o textos de un determinado autor. En definitiva, es un conjunto de textos relacionados entre sí por ser de un mismo autor o tratar un mismo tema.
* Matriz de incidencia: una matriz en la que dado un corpus y todos los textos que hay en él, se numeran todas las palabras distintas que en ellos aparecen, con unos y ceros. Normalmente el número de ceros es mucho mayor que el número de unos.
* Índices invertidos: es una manera más eficiente de guardar las apariencias de una palabra en un texto de un corpus. En ella sólo se guardan las apariciones, es decir, los unos de la matriz de incidencia. Está formado por un diccionario y una lista de ocurrencias.
  + Diccionario: conjunto de palabras distintas que forman los textos del corpus.
  + Lista de ocurrencias (posting list): para cada término del diccionario, se forma una lista con los documentos en los que este aparece (se hace mediante un identificador del texto o documento que lo incluye)
* Full inverted index: en este índice aparece en qué documento está la palabra, cuantas veces sale y en que posiciones sale esas veces. Tiene la siguiente estructura: [documento en el que sale, nº veces que sale [(list) posiciones en las que sale]], [documento 2º en el que sale, … []].
* Índices invertidos: para la creación de estos debemos tener en cuenta que:
  + Diccionario: almacena la palabra y el número de documentos en los que esta aparece.
  + Lista de ocurrencias (posting list): contiene (indexado) el docID de los documentos en los que aparece dicha palabra.
* Pasos en la indexación:
  + Dividir el documento o documentos en secuencias de tokens, es decir, separar los términos, aun con repeticiones de estos, junto con el docID del documento en el que aparecen.
  + Se ordena la tabla resultante del paso anterior por términos en orden alfabético.
  + Se crea el diccionario y la lista de ocurrencias, es decir, se indexan los términos y la frecuencia con la que aparecen en el diccionario (frecuencia = número de apariciones) y en la posting lists se introduce, en el caso de que su frecuencia sea 1 el documento (docID) en el que aparece, si su frecuencia es mayor, entonces se escribe en la posting list los dosID’s en los que aparece la palabra, por orden dentro del docID.
* Consultas de una sola palabra: es la búsqueda mas sencilla, en la que tratamos de encontrar las ocurrencias de esa palabra dentro del diccionario y obteniendo los resultados a través de la posting list del mismo. La búsqueda del diccionario, y por tanto su estructura, sigue eda’s clásicas, como tablas hash o árboles. El diccionario se encuentra en memoria principal (RAM) mientras la posting list lo hace en el almacenamiento secundario.
* Consultas de más de una palabra: tienen dos casos
  + Operación AND: se debe buscar la aparición de las palabras de la consulta con una lista para cada palabra. Después de hace la intersección (operación AND).
  + Operación OR: se buscan las apariciones de las palabras con una lista para cada una. Después, se hace la intersección de estas (operación OR).
  + Con la operación AND se buscan que aparezcan TODAS las palabras de una consulta en un mismo texto y con la operación OR con que aparezca una de las palabras en un texto este ya es válido.