**Colecciones de bases de datos a utilizar en el trabajo**

En el trabajo puedes utilizar una colección de test del sitio web “Latent Semantic Indexing” (http://www.cs.utk.edu/~lsi/). Por ejemplo la colección (CISI) que consta de información referente a 5.609 términos en 1460 documentos y viene dada mediante un fichero de consultas textual, una lista de términos y una matriz (dispersa) de términos por documento en formato Harwell-Boeing. En el poliformat también se ha dejado la versión en formato \*.mat de la matriz CISI.

Conversión de formato de las matrices de la colección (http://web.eecs.utk.edu/research/lsi/)

Con el fin de poder trabajar en el entorno MATLAB con todas las colecciones, el primer paso debe ser convertir las matrices CISI, CRAN, MED de términos por documentos en formato Harwell-Boeing a una matriz en formato disperso de MATLAB. Para ello, en primer lugar, se debe modificar la cabecera del fichero de la matriz para que incorporase información relativa a la estructura de la misma (número de líneas para punteros, datos …). A continuación se ha realizado la lectura y conversión mediante el programa “hb\_to\_msm” disponible en la siguiente dirección: <http://people.sc.fsu.edu/~burkardt/m_src/hb_to_msm/hb_to_msm.html>. Una vez modificada y cargada la matriz, se puede proceder a almacenarla en un fichero .mat para su posterior lectura.

Nota: Para conversión de formatos puedes utilizar los ficheros de la carpeta del poliformat/  [Curso 2020/21](https://poliformat.upv.es/portal/site/DOC_34573_2020/tool/6ae4ced0-9f60-4fa8-9466-8297411f6b0b?panel=Main" \o "Ir a esta carpeta) /[Trabajos](https://poliformat.upv.es/portal/site/DOC_34573_2020/tool/6ae4ced0-9f60-4fa8-9466-8297411f6b0b?panel=Main) /[Técnica LSI- Sistemas de Recuperación de Información](https://poliformat.upv.es/portal/site/DOC_34573_2020/tool/6ae4ced0-9f60-4fa8-9466-8297411f6b0b?panel=Main) /HB-conversion