Problema JAVA
(I/O.- Adaptación de datos a requerimientos de base de datos)

Supongamos que tenemos una base de datos con referencias bibliográficas, y recibimos un fichero con datos extraídos de scholar.google.es. Será preciso adaptar estos datos a nuestras necesidades concretas para incorporarlos. En este fichero (***DB.txt***) tenemos una línea por cada dato bibliográfico, que estará precedido por un número de referencia y un carácter que indica el tipo de dato (t=título, y=año, p=publicación, r=referencias, a=autor, l=link). La información correspondiente a referencias(r) y links(l) puede no aparecer para un determinado artículo, y la correspondiente a autores(a) puede aparecer varias veces, de modo que en cada ocasión se referirá a un autor y llevará siempre asociada la filiación del mismo entre comillas. El número de referencia y el carácter que indica el tipo de información van juntos y seguidos por un espacio en blanco. No hay un orden predeterminado para las lineas, de modo que la información referente a los distintos artículos puede intercalarse sin restricciones.

Ejemplo:

```
101t Polymerase chain reaction can detect bacterial DNA in aseptically loose total hip arthroplasties
101y 2004
101p journals.lww.com
101r 8 455 30 532 31 234 21 333 13 19
101a MT. Clarke "University of Toronto, utoronto.ca, Canada"
101a CP. Roberts "Rheinische Friedrich Wilhelms Universitat Bonn, uni-bonn.de, Germany"
101a PTH. Lee "University of Massachusetts Amherst, umass.edu, USA"
101a J. Gray "University of California Irvine, uci.edu, USA"
101l http://journals.lww.com/corr/Abstract/2004/10000/Polymerase_Chain_Reaction_Can_Detect_Bacterial_DNA.23.aspx
```

La tarea consiste en recoger toda la información que se nos aporta y generar los listados (en otros ficheros de salida) correspondientes a cada una de las tablas que serán necesarias para actualizar posteriormente la base de datos.

Las tablas a generar son:

- ***Autores.txt***: dos columnas separadas por ",". La primera un número de orden dado por el programa y la segunda un autor sacado de ***DB1.txt***. Obviamente los autores se repiten numerosas veces en ***DB.txt*** pero no deben aparecer repeticiones en ***Autores.txt***.
- ***Filiaciones.txt***: dos columnas separadas por ",". La primera un número de orden dado por el programa y la segunda una institución sacada de las lineas de autor. Obviamente las filiaciones se repiten numerosas veces en ***DB.txt*** pero no deben aparecer repeticiones en ***Filiaciones.txt***.
- ***Autor-filiacion.txt***: dos columnas separadas por "," La primera un número de orden asociado a un autor conforme a ***Autores.txt***, y la segunda con el número de orden asociado a Filiación que le corresponde codificado según ***Filiaciones.txt***.
- ***Publicaciones.txt***: dos columnas separadas por ",". La primera un número de orden dado por el programa y la segunda una publicación sacada de las lineas de publicación. Obviamente las filiaciones se repiten numerosas veces en ***DB.txt*** pero no deben aparecer repeticiones en ***Publicaciones.txt***.
- ***Info-articulo.txt***: 5 columnas separadas por ",". La primera el número de referencia, la segunda el título, la tercera el año, la cuarta la publicación y la quinta columna la dirección web, que en caso de no estar disponible se consignará "null"
- ***Articulo-autores.txt***: dos columnas separadas por ","; la primera el número de referencia asociado a al artículo conforme a ***Info-artículo.txt*** y la segunda un autor conforme a ***Autores.txt***. Debe tenerse en cuenta que un articulo puede tener varios autores, y en tal caso se generará una nueva línea en este fichero por cada autor.
- ***Articulo-referencias.txt***: dos columnas separadas por ","; la primera el número de referencia asociado a un artículo conforme a ***Info-artículo.txt*** y la segunda una referencia (de la linea de entrada correspondiente al código "r"). Debe tenerse en cuenta que un articulo tendrá normalmente varias referencias, y en tal caso se generará una nueva línea en este fichero por cada referencia.

A tener en cuenta:

El fichero ***DB.txt*** se ha generado extrayendo automáticamente la información de Google Scholar y puede presentar las faltas de homogeneidad propias de dicha aplicación web (p. ej. Un mismo autor con variantes en su denominación que le harán aparecer como más de uno). Al examinar los resultados esto puede inducir en algún caso a pensar que el programa no ha funcionado correctamente, por lo que conviene "sospechar" también de los datos de origen. En todo caso lo que se pretende es que la salida sea consistente con la entrada[1].

1 Tomamos el camino sencillo por tratarse de un ejercicio, pero podría ganarse en "robustez" haciendo cosas como considerar que los separadores pueden no ser únicos o seguir un patrón determinado (p.ej. es igual <J. Gray> que <J. Gray>), no permitir que la capitalización de caracteres influya (p.ej. es igual <J. Gray> que <j. gray>), admitir líneas vacías, etc.