Caracterización

Para este ejercicio se han cogido 11 páginas de la web http://elpixeblogdepedja.com/. Una vez se han elegido se han guardado en formato texto. Para componer el archivo arff primero se han quitado las comillas para que no de problemas a la hora de leerlo y se ha copiado todo el texto en una sola línea.

Una vez se ha importado el archivo en Weka se han separado los nombres de las páginas de sus contenidos y se han quitado las palabras stop con un diccionario obteniendo el resultado de *textosCompletos.xlsx.*

Entre los resultados se puede ver que hay entradas como los links de los menus que se repiten en todos los documentos que tienen un peso bajo ya que no se repiten mucho dentro de los mismos como /category/cine>.

También se han encontrado palabras que dentro del mismo documento que tienen un gran peso ya que se han repetido mucho ya que son el tema principal del articulo como por ejemplo Funk, Bob o Colossal que corresponden con títulos de algunos temas tratados.

Como la web es un blog tiene un menú lateral con los últimos post subidos por lo que se puede ver que estas palabras se repiten en todos los documentos con un peso bajo ya que solo están una vez.