Recuperación de Información

Practica 2: Parser de documentos con TIKA

Autores: García Maldonado, Fco Javier
Correo: franelas@correo.ugr.es
DNI: 76654015Y

Autores: Valenzuela Álvarez, Pablo
Correo: pvalenzuela@correo.ugr.es
DNI: 76652136J

1.- Introducción

En esta práctica hemos realizado un trabajo de análisis de documentos básico. Este análisis consistía en procesar varios archivos ubicados en un directorio de los cuales se extraen los metadatos en primer lugar, en segundo lugar la extracción de los enlaces que aparezcan en cada documento y en último lugar se procesan el número de términos en el archivo y la cantidad que hay de cada uno de ellos. Cabe destacar que a estos términos les hemos aplicado un pre-procesado antes de iniciar el conteo, ya que si no realizamos este pre-procesado el documento generado con los términos y la cantidad de veces que aparecen contendría errores al contar como términos independientes aquellas palabras que tengan signos de puntuación antes, después y en mitad de la palabra.

2.- Procesado de documentos con TIKA

Tal y como hemos indicado en el apartado anterior, en esta práctica hemos realizado un procesamiento de archivos con TIKA. En primer lugar, nos hemos centrado en extraer el idioma en el que está escrito el texto. A continuación, una vez extraído el idioma, pasamos a obtener los metadatos del archivo tales como: codificación, nombre del documento y tipo de documento. Seguidamente extraemos los enlaces que contuviese el documento. Por último, lo que hemos realizado ha sido un conteo de las diferentes palabras que hay en cada documento y el número de veces que aparece cada una de ellas, para ello usamos la estructura de datos HashMap.

Una vez que tenemos todas las palabras almacenadas en nuestra ED, necesitamos que estas estén en orden decreciente en función de la frecuencia de aparición de las palabras. Por lo que una palabra que aparezca más que ninguna debe estar la primera. El problema llegado a este punto es que el HashMap no tiene función de ordenación, ni se almacena ordenado, por lo que debemos recurrir a otra ED diferente, esta ED elegida ha sido el ArrayList. Con ella podemos volver a introducir los datos y ordenarlos, usando un comparador, para así poder imprimir en un documento de texto las palabras junto con su frecuencia por orden decreciente.

Cabe destacar que para nuestro proyecto hemos utilizado como librería TIKA. Y que los libros son todos obtenidos del proyecto Gutenberg, los cuales están en diferentes idiomas, desde italiano, inglés y francés hasta islandés o incluso húngaro.

```
Nombre:         55667-h.htm
Tipo:           application/xhtml+xml
Codificación:   UTF-8
Idioma:         hu
------------------------------------------------------------------------
Nombre:         Gutenberg.html
Tipo:           application/xhtml+xml
Codificación:   UTF-8
Idioma:         en
------------------------------------------------------------------------
Nombre:         La femme francaise dans les temps modernes.epub
Tipo:           application/epub+zip
Codificación:   windows-1252
Idioma:         fr
------------------------------------------------------------------------
Nombre:         La vera cuciniera genovese.epub
Tipo:           application/epub+zip
Codificación:   windows-1252
Idioma:         hu
------------------------------------------------------------------------
Nombre:         pg53956.mobi
Tipo:           application/x-mobipocket-ebook
Codificación:   UTF-16LE
Idioma:
------------------------------------------------------------------------
Nombre:         pg55719.mobi
Tipo:           application/x-mobipocket-ebook
Codificación:   windows-1252
Idioma:
------------------------------------------------------------------------
Nombre:         Tika in Action.pdf
Tipo:           application/pdf
Codificación:   IBM866
Idioma:         en
------------------------------------------------------------------------
Nombre:         In the days of queen Mary.txt
Tipo:           text/plain
Codificación:   windows-1252
Idioma:         en
------------------------------------------------------------------------
```

Salida de metadatos

Salida de enlaces

```
Link: Gutenberg_files/load.css
Link: http://www.gutenberg.org/favicon.ico
Link: http://www.gutenberg.org/w/opensearch_desc.php
Link: http://www.gutenberg.org/w/api.php?action=rsd
Link: http://www.gnu.org/copyleft/fdl.html
Link: http://www.gutenberg.org/w/index.php?title=Special:RecentChanges&feed=atom
Link: #navigation
Link: #p-wikisearch
Link: http://www.gutenberg.org/ebooks/
Link: http://www.gutenberg.org/wiki/Category:Bookshelf
Link: http://www.gutenberg.org/catalog/
Link: http://m.gutenberg.org/
Link: http://www.gutenberg.org/wiki/Gutenberg:Contact_Information#Electronic_Mail
Link: http://www.gutenberg.org/wiki/Gutenberg:Terms_of_Use
Link: Gutenberg_files/latest-covers.html
Link: 
Link: http://m.gutenberg.org/
Link: http://m.gutenberg.org/
Link: http://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Needs_Your_Donation
Link: http://www.pgdp.net/
Link: https://librivox.org/
Link: http://www.gutenberg.org/wiki/Gutenberg:Contact_Information#Electronic_Mail
Link: https://www.fcc.gov/ecfs/search/proceedings?q=name:((17-108))
Link: https://law.duke.edu/cspd/publicdomainday
Link: http://www.gutenberg.org/ebooks/
Link: http://www.gutenberg.org/ebooks/search/%3Fsort_order%3Drelease_date
Link: http://www.gutenberg.org/ebooks/search/%3Fsort_order%3Ddownloads
Link: http://www.gutenberg.org/browse/scores/top
Link: http://www.gutenberg.org/wiki/Category:Bookshelf
Link: http://www.gutenberg.org/wiki/Gutenberg:Feeds
Link: http://www.gutenberg.org/catalog/
Link: http://www.gutenberg.org/wiki/Gutenberg:Offline_Catalogs
Link: http://self.gutenberg.org/
Link: http://www.gutenberg.org/wiki/Gutenberg:MobileReader_Devices_How-To
Link: http://www.gutenberg.org/wiki/Category:How-To
Link: http://www.gutenberg.org/wiki/Gutenberg:Contact_Information
Link: http://www.gutenberg.org/wiki/Gutenberg:Readers%27_FAQ#R.26._I.27ve_found_some_obvious_t
Link: http://www.gutenberg.org/wiki/Category:Volunteering
Link: http://www.pgdp.net/
Link: https://librivox.org/
Link: http://www.gutenberg.org/wiki/Gutenberg:Promote_Project_Gutenberg
Link: http://www.gutenberg.org/wiki/Gutenberg:About
Link: http://www.gutenberg.org/wiki/Gutenberg:No Cost or Freedom%3F
```

Salida de conteo

```
the,4640;
and,2050;
of,1995;
to,1457;
a,1113;
in,964;
he,781;
was,776;
his,710;
that,611;
i,513;
with,513;
you,498;
for,471;
it,465;
had,426;
as,407;
at,401;
were,399;
on,388;
their,388;
they,373;
by,341;
be,318;
is,283;
from,258;
this,258;
him,258;
but,254;
said,252;
will,245;
all,244;
not,237;
them,233;
s,226;
sir,213;
which,209;
her,200;
my,195;
your,194;
men,189;
have,185;
we,184;
then,173;
me,160;
```

3.- Verificación de la Ley de Zipf

En este apartado trataremos de verificar que se cumple esta ley. Esta ley viene dada por la ecuación de Booth y Federowicz, y se expresa de la siguiente forma:

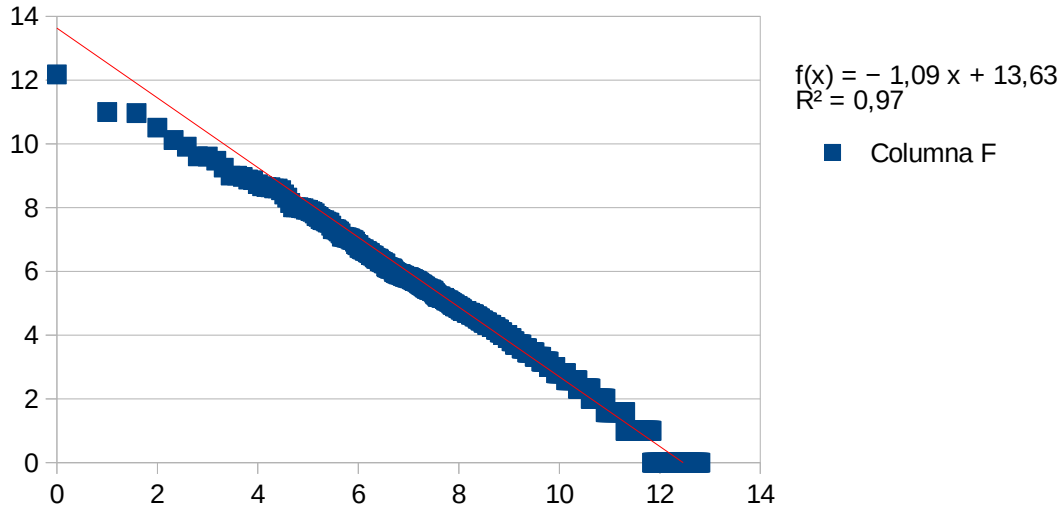
$$F=\frac{k}{R^m}$$

Donde F representa la frecuencia, R la posición en la ordenación que hemos realizado y por último k y m son constantes. Para obtener dichas constantes podemos hacerlo a partir del grafico log-log teniendo en cuenta que:

$$\ln(F)=\ln\frac{k}{R^m}=\ln(k)-m\ln(R)$$

Por tanto, si realizamos sobre el grafico log-log un ajuste lineal, podremos obtener dichas constantes k y m de forma sencilla.

Ejemplo para texto de 7000 palabras en inglés.



Ejemplo para texto de 15000 palabras en francés