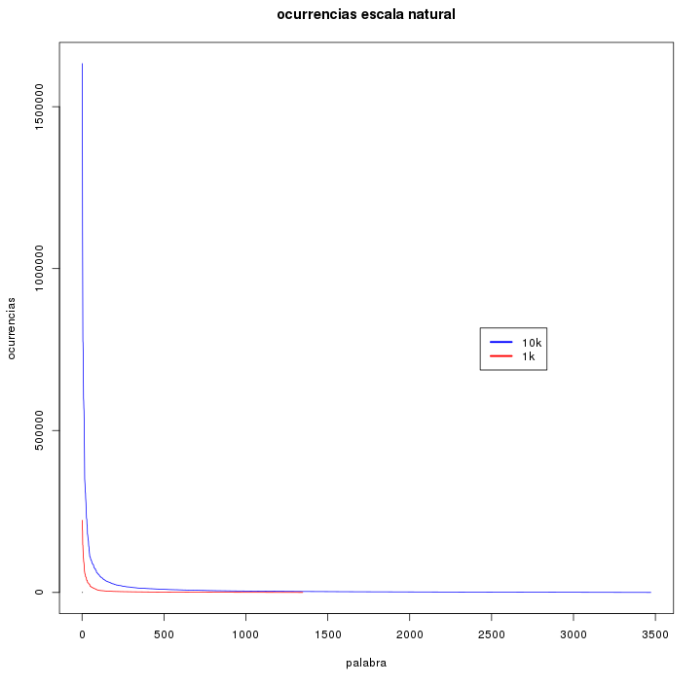
Busqueda y mineria de información

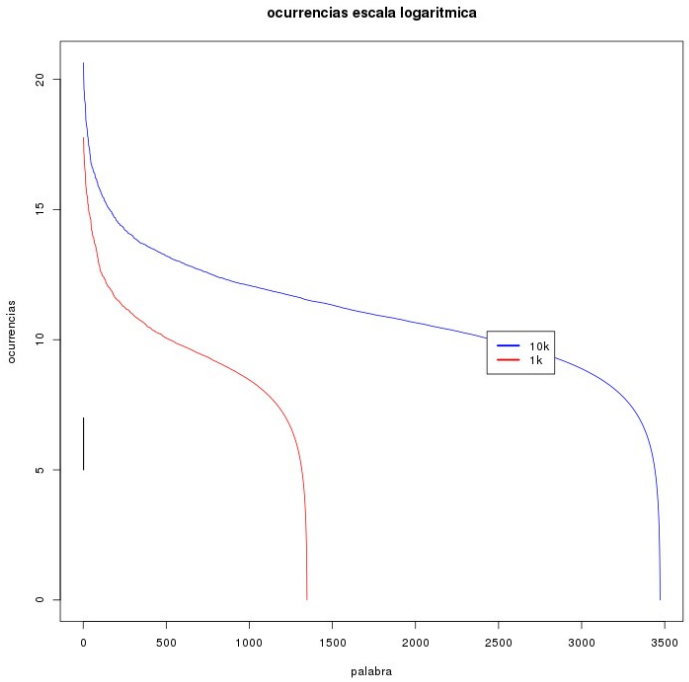
Práctica 1

Grupo 09

Angel Fuente Ortega
Mario Valdemaro García Roque

Frecuencia de los terminos.





En este apartado aporatmos 2 gráficas, una con los resultados en escala natural y otra con los resultados con una escala logaritmica. En amabas la linea roja representa el grupo de 1000 documentos y la azul el de 10000.

Sobre la gáfica en escala natural comentar que obtenemos unas ocurrencias exageradamente altas de algunos terminos al principio, detalle que analizaremos mas tarde. En cuanto a la forma que tienen las dos curvas se ve que tienen un forma parecida a la que encontramos en una distribución exponencial o en una libre de escala, es decir un terminos iniciales muy frecuentes en comparacion con el resto de terminos.

Debido a que en la primera gráfica casi no se puede observar casi nada debido a que casi inmediatamente la grafica se aplana, hemos decidido realizar la segunda gráfica, que esta en una escala logaritmica en base 2. De ella se oberva mejor como los terminos con frecuencias “intermedias” tienen un decrecimiento mas suave, siendo estos probablemente palabras de un lenguaje como el inglés o el castellano.

Si analizamos cuales son los terminos mas frecuentes de, por ejemplo, los 1000 documentos entendemos porque los primeros documentos tienen una frecuencia tan alta.

Term	frec	docs
href	222841	979
td	196531	817
famili	190659	569
div	147196	780
html	18713	999

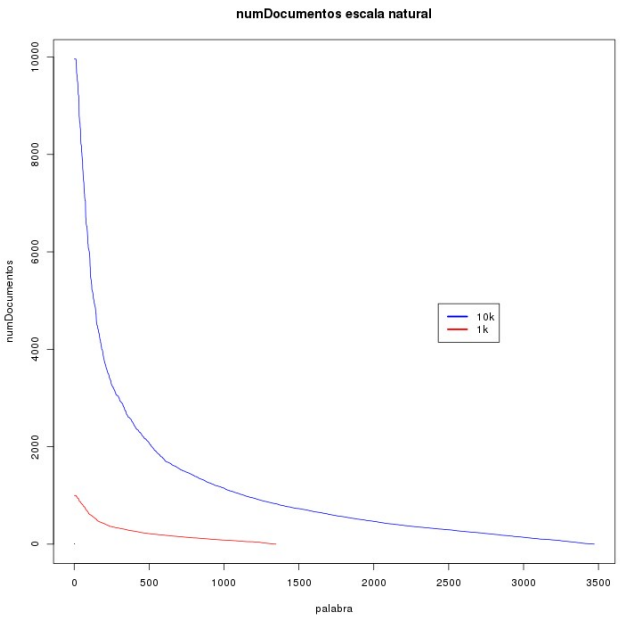
Estos terminos son cercanos a código o directamente pertenecen a html (como por ejemplo div o href). Si nos fijamos no solo tienen unas frecuencias altas sino que algunos salen en casi todos los documento.
¿Pero significa entonces esto que no estamos filtrando el código HTML?
Realmente no, pero tampoco lo estamos filtrando del todo. Es decir nuestro parseador de html borra todo lo que sea estrictamente el marcado pero si echamos un ojo con cualquier editor grafico al código obervamos esto al inicio del documento:

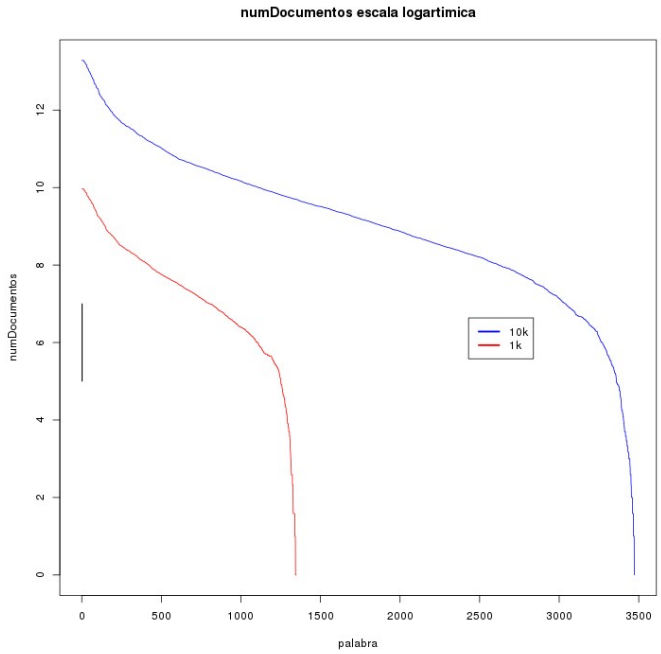
```
HTTP/1.1 200 OK
Content-Type: text/html
Date: Mon, 19 Jan 2009 23:16:05 GMT
Set-Cookie: RemoteCountry=US; domain=.ecplaza.net; path=/
Cache-control: private
Content-Length: 66345
```

Este es un mensaje tipico del protocolo http para indicarnos el estado del servidor. Y, aunque no lo hemos comprobado peronalmente, tiene aspecto de que aparece en cada documento ya que si nos fijamos en las frecuencias que pusimos arriba html sale en 999 documentos.

Por tanto esta parte del documento no es interesante para crear indices ya que es un mensaje generico de establecimiento de conexión en vez del contenido de la pagina. Por tanto nuestro parser no entiende que sea HTML asi que lo deja.

Frecuencia de los documentos:

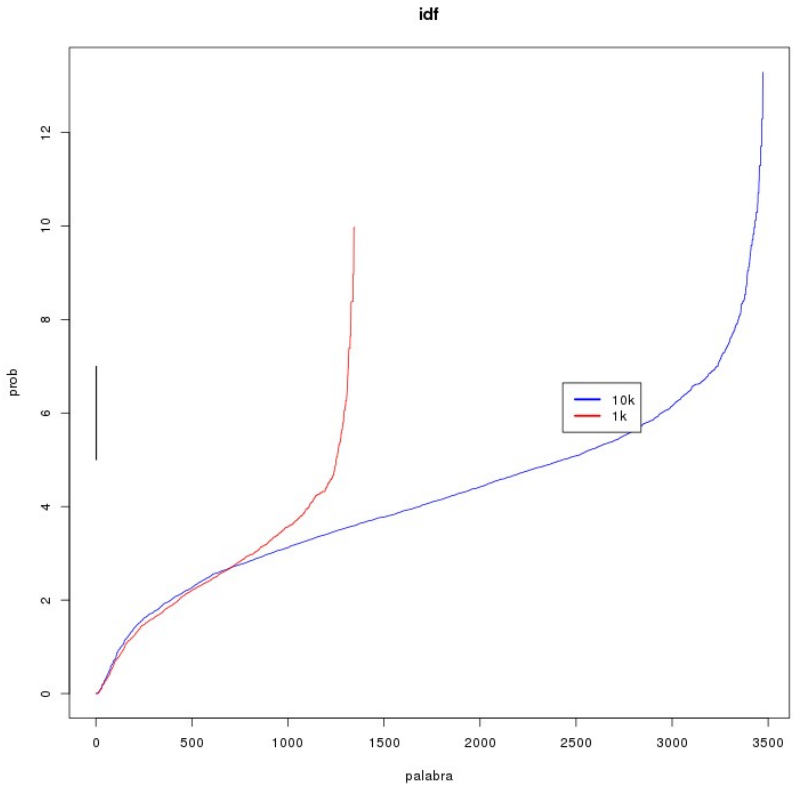




En cuanto a la frecuencia de los terminos en los documentos vuelve a ocurrir lo que ya mencionamos arriba, pero esta vez de una forma mas moderada. Basicamente se ve como los primeros terminos tienen una frecuencia alta debidido a que aparecen en casi todos los documentos.

Tambien volvemos a obervar el comportamiento mencionado en las frecuencias, que el aspecto de las curvas que observamos siguen un patron parecido al de una distribucion exponencial o al de una libre de escala.

IDF:



Estas curvas presentan un aspecto creciente debido a que el documento esta ordenado de mayor ocurrencia en documentos a menor. Por tanto sabiendo que la funcion idf es el logaritmo del número de documentos partido de los documentos en los que se encuentra el termino, si un termino esta presente en muchos documentos este valor se acercara a cero, mientras que si el termino apenas esta presente en algun el valor tendera a maximizarse.