

Instituto Tecnológico de Costa Rica

# Documentación TP2

Recuperación de Información Textual

Daniel Cortés Sáenz, Isaac Ramírez Solano  
24/06/2013

## Contenido

Introducción .....	2
Ambiente de Desarrollo .....	2
Instrucciones para ejecutar los programas .....	2
Archivo Invertido .....	2
Consultas .....	2
Page Rank .....	3
Corridas .....	3
Archivo Invertido .....	3
Consultas .....	3
Page Rank .....	6
Comentarios Finales .....	6

## Introducción

La segunda tarea programada consiste en la implementación de un archivo invertido y la realización de búsquedas usando ese archivo invertido. Para ello, se debe implementar la estructura del archivo invertido que consta del diccionario, documentos y archivo de postings.

Además, se debe implementar un algoritmo para calcular el Page Rank de la colección. Este algoritmo recibe como parámetros la diferencia esperada y la cantidad de iteraciones. Al final muestra el Page Rank de los documentos de forma descendente.

## Ambiente de Desarrollo

Para esta segunda tarea programada, se usó Perl (5.16.3) y Java (1.7.25). El sistema operativo usado fue Windows 8. Se decidió usar Perl porque es un lenguaje de scripting bastante rápido. Java se encarga de procesar los documentos que genera Perl ya que las matrices y arreglos son más fáciles de manejar que Perl.

Además, se utilizó Github como manejador de versiones. El repositorio de la tarea está disponible públicamente en [este](#) enlace.

## Instrucciones para ejecutar los programas

### Archivo Invertido

Las instrucciones para obtener el archivo invertido son las siguientes:

1. Se deben procesar los documentos con perl ejecutando el siguiente comando:  
perl parser.pl analizar  
Esto tendrá como salida el archivo de Documentos y un archivo Vocabulario que será utilizado por java.
2. Seguidamente en la carpeta Java se puede encontrar un .bat que se ejecuta dando doble click. Este archivo va a crear el archivo Vocabulario y Postings final del archivo invertido.

Todos los archivos deben estar y estarán creados en la raíz de D:/.

### Consultas

Para las consultas se utilizara en modelo vectorial, estas consultas serán procesadas por perl con el siguiente comando:

1. perl parser.pl consulta "TEXTO DE LA BÚSQUEDA"

El comando anterior tendrá como salida 2 documentos:

Un archivo.txt que contiene el escalafón con los archivos y sus similitudes.

Un archivo.html que se abre automáticamente y muestra el escalafón de similitud.

## Page Rank

Para correr el Page Rank, se debe ingresar alguno de los siguientes comandos:

1. perl parser.pl pr
2. perl parser.pl 0.0003 500

El primer comando asume que la diferencia esperada y la cantidad iteraciones serán 0.0001 y 100. En el segundo, se especifica que la diferencia esperada será 0.0003 y 500.

Es importante aclarar que este programa debe ser ejecutado **después** de haber creado el archivo invertido.

## Corridas

### Archivo Invertido

### Consultas

Las consultas que se ejecutaran para probar el programa serán las pruebas de la guía de documentación.

La primera consulta será “permisos archivos configuración”:



The screenshot shows a web browser window with the address bar displaying 'file:///C:/Users/Sirlsaac/Documents/GitHub/rit2/permisos\_archivos\_configuracion.html'. The page title is 'Resultados búsqueda “permisos archivos configuracion”'. Below the title, it says 'Consulta hecha a las: 21:25:48, Dom 23 Jun, 2013'. The main content area displays three search results, each with a table showing document details and a preview of the file content.

Pos.	ID Documento	Similitud	Ruta
1.	2	50.2583866792817	D:/HTML/cmdline-attr.html

Vista preliminar del archivo:  
Manipulación de los atributos de los archivos Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 3. Introducción a la línea de comandosSiguiente3.2. Manipulación de

Pos.	ID Documento	Similitud	Ruta
2.	64	36.4872789438887	D:/HTML/x2788.html

Vista preliminar del archivo:  
Características de los niveles de seguridad Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 7. MSEC -- Mandrake SEcurity tools (Herramientas de seguridad de Mandr

Pos.	ID Documento	Similitud	Ruta
3.	76	32.250546931745	D:/HTML/x432.html

Vista preliminar del archivo:  
Nociones básicas sobre los archivos Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 2. Conceptos básicos de UnixSiguiente2.2. Nociones básicas sobre los archivos

La consulta fue exitosa y se obtuvieron todos los documentos que se esperaban.

La segunda consulta fue con el texto “archivos y permisos”.



Resultados búsqueda “archivos y permisos”

Consulta hecha a las: 21:28:47, Dom 23 Jun, 2013”

Pos.	ID Documento	Similitud	Ruta
1.	2	50.2583866792817	D:/HTML/cndline-attr.html

Vista preliminar del archivo:

Manipulación de los atributos de los archivos Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 3. Introducción a la línea de comandosSiguiente3.2. Manipulación de

Pos.	ID Documento	Similitud	Ruta
2.	76	32.250546931745	D:/HTML/x432.html

Vista preliminar del archivo:

Nociones básicas sobre los archivos Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 2. Conceptos básicos de UnixSiguiente2.2. Nociones básicas sobre los archivos

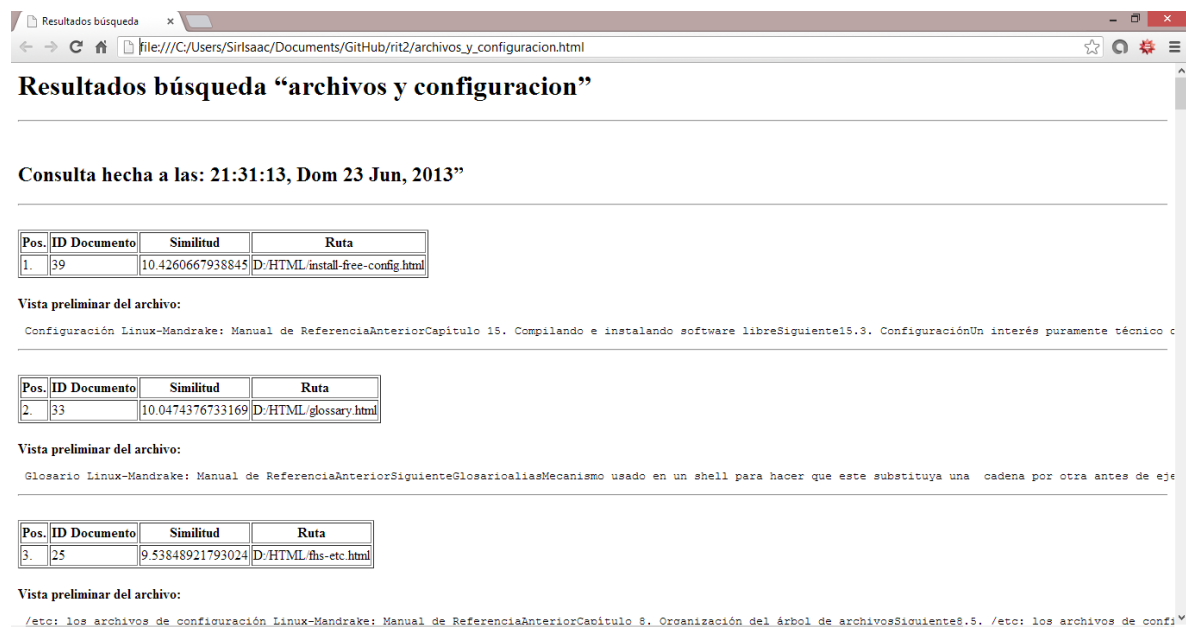
Pos.	ID Documento	Similitud	Ruta
3.	64	30.5575508792275	D:/HTML/x2788.html

Vista preliminar del archivo:

Características de los niveles de seguridad Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 7. MSEC -- Mandrake SEcurity tools (Herramientas de seguridad de Mandr

La consulta fue exitosa y se obtuvieron todos los documentos que se esperaban.

La tercera consulta se realizó con el texto “archivos y configuración”.



Resultados búsqueda “archivos y configuracion”

Consulta hecha a las: 21:31:13, Dom 23 Jun, 2013”

Pos.	ID Documento	Similitud	Ruta
1.	39	10.4260667938845	D:/HTML/install-free-config.html

Vista preliminar del archivo:

Configuración Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 15. Compilando e instalando software libreSiguiente15.3. ConfiguraciónUn interés puramente técnico c

Pos.	ID Documento	Similitud	Ruta
2.	33	10.0474376733169	D:/HTML/glossary.html

Vista preliminar del archivo:

Glosario Linux-Mandrake: Manual de ReferenciaAnteriorSiguienteGlosarioaliasMecanismo usado en un shell para hacer que este substituya una cadena por otra antes de eje

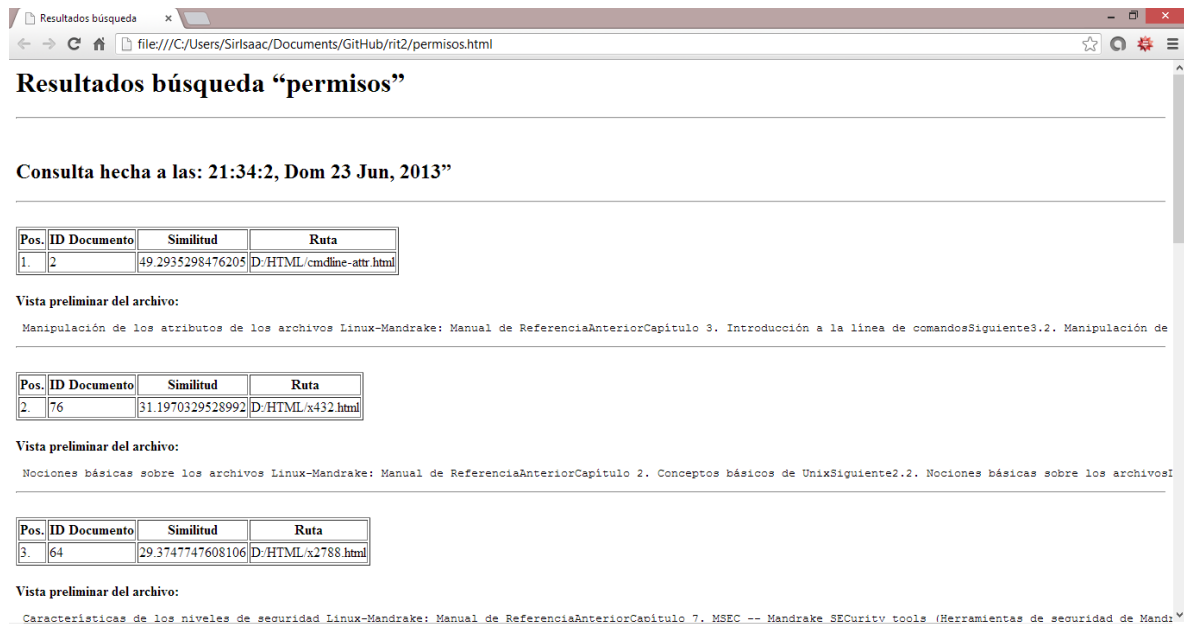
Pos.	ID Documento	Similitud	Ruta
3.	25	9.53848921793024	D:/HTML/this-etc.html

Vista preliminar del archivo:

/etc: los archivos de configuración Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 8. Organización del árbol de archivosSiguiente8.5. /etc: los archivos de confi

La consulta fue exitosa y se obtuvieron todos los documentos que se esperaban.

Para la cuarta consulta se procesó el texto “permisos”.



Resultados búsqueda “permisos”

Consulta hecha a las: 21:34:2, Dom 23 Jun, 2013”

Pos.	ID Documento	Similitud	Ruta
1.	2	49.2935298476205	D:/HTML/cmdline-attr.html

Vista preliminar del archivo:

Manipulación de los atributos de los archivos Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 3. Introducción a la línea de comandosSiguiente3.2. Manipulación de

Pos.	ID Documento	Similitud	Ruta
2.	76	31.1970329528992	D:/HTML/x432.html

Vista preliminar del archivo:

Nociones básicas sobre los archivos Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 2. Conceptos básicos de UnixSiguiente2.2. Nociones básicas sobre los archivos

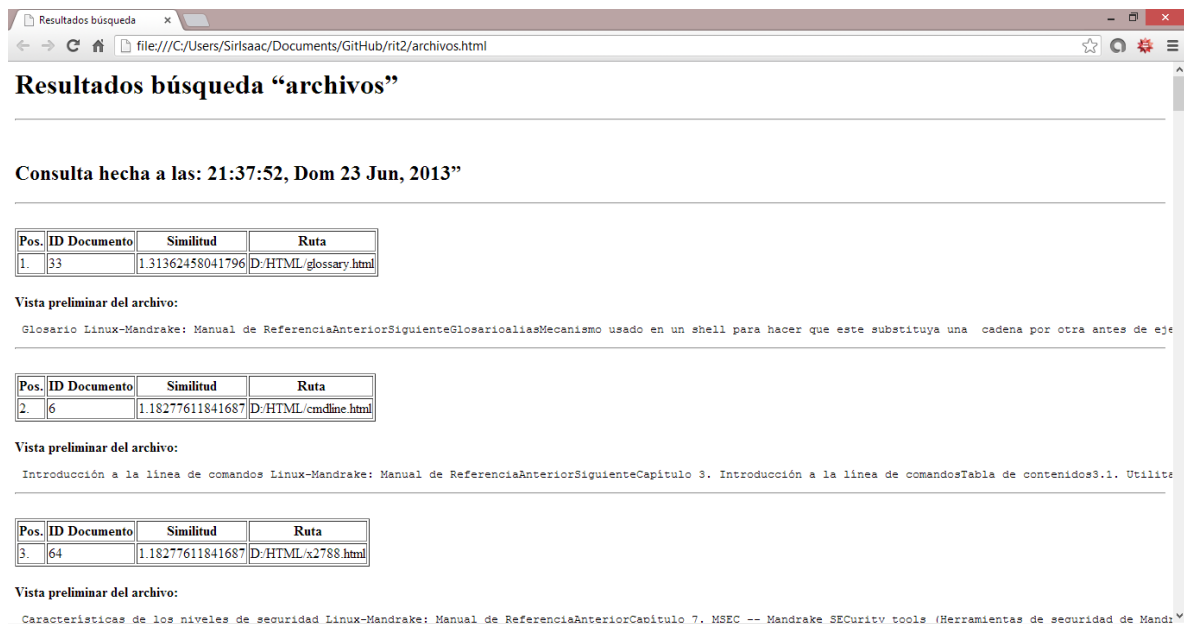
Pos.	ID Documento	Similitud	Ruta
3.	64	29.3747747608106	D:/HTML/x2788.html

Vista preliminar del archivo:

Características de los niveles de seguridad Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 7. MSEC -- Mandrake SECurity tools (Herramientas de seguridad de Mandr

La consulta fue exitosa, sin embargo no se obtuvo el archivo especificado en la guía pues ese archivo lo que contiene es la palabra “permiso” y no “permisos”.

La quinta consulta se realizó con el texto “archivos”.



Resultados búsqueda “archivos”

Consulta hecha a las: 21:37:52, Dom 23 Jun, 2013”

Pos.	ID Documento	Similitud	Ruta
1.	33	1.31362458041796	D:/HTML/glossary.html

Vista preliminar del archivo:

Glosario Linux-Mandrake: Manual de ReferenciaAnteriorSiguienteGlosarioaliasMecanismo usado en un shell para hacer que este substituya una cadena por otra antes de eje

Pos.	ID Documento	Similitud	Ruta
2.	6	1.18277611841687	D:/HTML/cmdline.html

Vista preliminar del archivo:

Introducción a la línea de comandos Linux-Mandrake: Manual de ReferenciaAnteriorSiguienteCapítulo 3. Introducción a la línea de comandosTabla de contenidos3.1. Utilite

Pos.	ID Documento	Similitud	Ruta
3.	64	1.18277611841687	D:/HTML/x2788.html

Vista preliminar del archivo:

Características de los niveles de seguridad Linux-Mandrake: Manual de ReferenciaAnteriorCapítulo 7. MSEC -- Mandrake SECurity tools (Herramientas de seguridad de Mandr

La consulta fue exitosa y se obtuvieron todos los documentos que se esperaban.

## Page Rank

Se adjunta el archivo ResultadoPageRank.txt dentro del directorio “Documentación” de la tarea que tiene el Page Rank de los documentos con 1 iteración.

## Comentarios Finales

La tarea se terminó completamente. Se pudo implementar tanto el archivo invertido como las consultas y el Page Rank. Dentro de los problemas encontrados se pueden mencionar los siguientes:

1. No se sabía cómo extraer los URL de las etiquetas de <a> ni cómo formatear el HTML a que solo se procese el texto. Para esto, se usó el HTML Parser que trae Perl.
2. Los arreglos multidimensionales son difíciles de usar el Perl. Para este caso, se decidió usar Java para que procese él estas partes y Perl se encargue de todo lo demás.

En general a todos los problemas encontrados se les encontró una solución. La única limitación es que el Page Rank requiere el archivo invertido para funcionar.