

# Modelos Computacionales

## Tarea 4

**Jose Alberto Benítez Andrades**

**71454586A**

**Modelos Computacionales**

**Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web**

**UNED**

**24/06/2011**

24 de junio de 2011

## ÍNDICE

### Contenido

Actividad 1:.....	3
Actividad 2:.....	8
Actividad 3: Comentarios Finales .....	12

---

24 de junio de 2011**Actividad 1:**

1. Descargar Lucene y crear un pequeño programa en Java que permita indexar y recuperar sobre las 30 páginas del corpus.
2. Hacer una búsqueda con cada una de las 15 preguntas.
3. Comparar los resultados con la tabla de resultados esperados (o ground truth hecha manualmente).

Extensiones optativas:

- a. Probar a indexar las páginas de diferentes maneras (con solo un campo o con varios). Por ejemplo como hay información semántica probar a utilizarla a la hora de indexar utilizando varios campos. Analizar los beneficios/desventajas.
- b. Hacer un módulo de procesamiento de las consultas que, basándose en la gramática identificada, sea capaz de hacer de interfaz entre la pregunta en LN y el motor de búsqueda.

**Resolución:**

La lista de preguntas es la siguiente:

- 1) ¿Qué son las webs con significado sintáctico?
- 2) ¿Qué relación existe entre las estandarizaciones W3C y la web semántica?
- 3) ¿A qué llamamos webs de datos?
- 4) La web semántica ¿necesita una base de conocimiento?
- 5) ¿En qué consiste la estructuración por metadatos?
- 6) ¿Quiénes son los investigadores iniciales de la web semántica?
- 7) ¿Qué evolución ha habido desde la web 1.0?
- 8) ¿Es lo mismo web semántica que web 3.0?
- 9) ¿Qué componentes posee una web semántica?
- 10) ¿Existen herramientas para crear webs semánticas más fácilmente?
- 11) ¿Existen CMS o gestores de contenido orientados a esta nueva tendencia?
- 12) ¿Ayuda el HTML5 a la inserción de la web semántica en la red?
- 13) ¿Es positivo para el posicionamiento SEO en buscadores la creación de webs semánticas?
- 14) ¿En qué posición se encuentran las herramientas OWL, RDF, XML y SPARQL en la web semántica?
- 15) ¿Existen buscadores semánticos?

Basándonos en la lista de webs de la fase 1, la tabla real de las webs que deberían verse en función de las preguntas anteriores, es la siguiente:

24 de junio de 2011

Tablas de valores esperados (las filas representan las 15 preguntas y las columnas las 30 webs)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															

24 de junio de 2011

Tablas de valores obtenidos con el programa (las filas representan las 15 preguntas y las columnas las 30 webs)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															

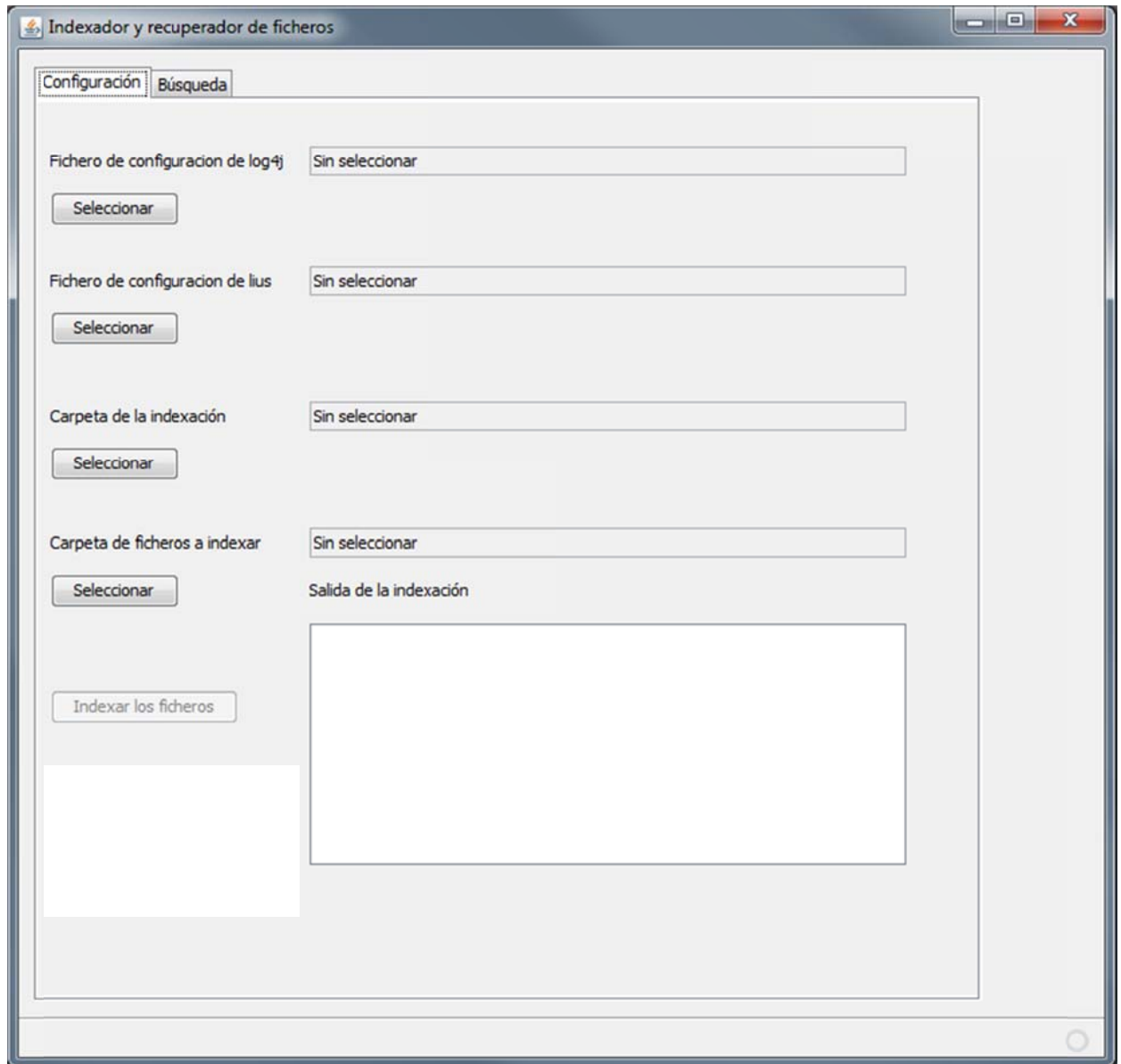
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															

24 de junio de 2011

Observaciones sobre la prueba:

La aplicación que he hecho para poder obtener estos datos, se utiliza así:

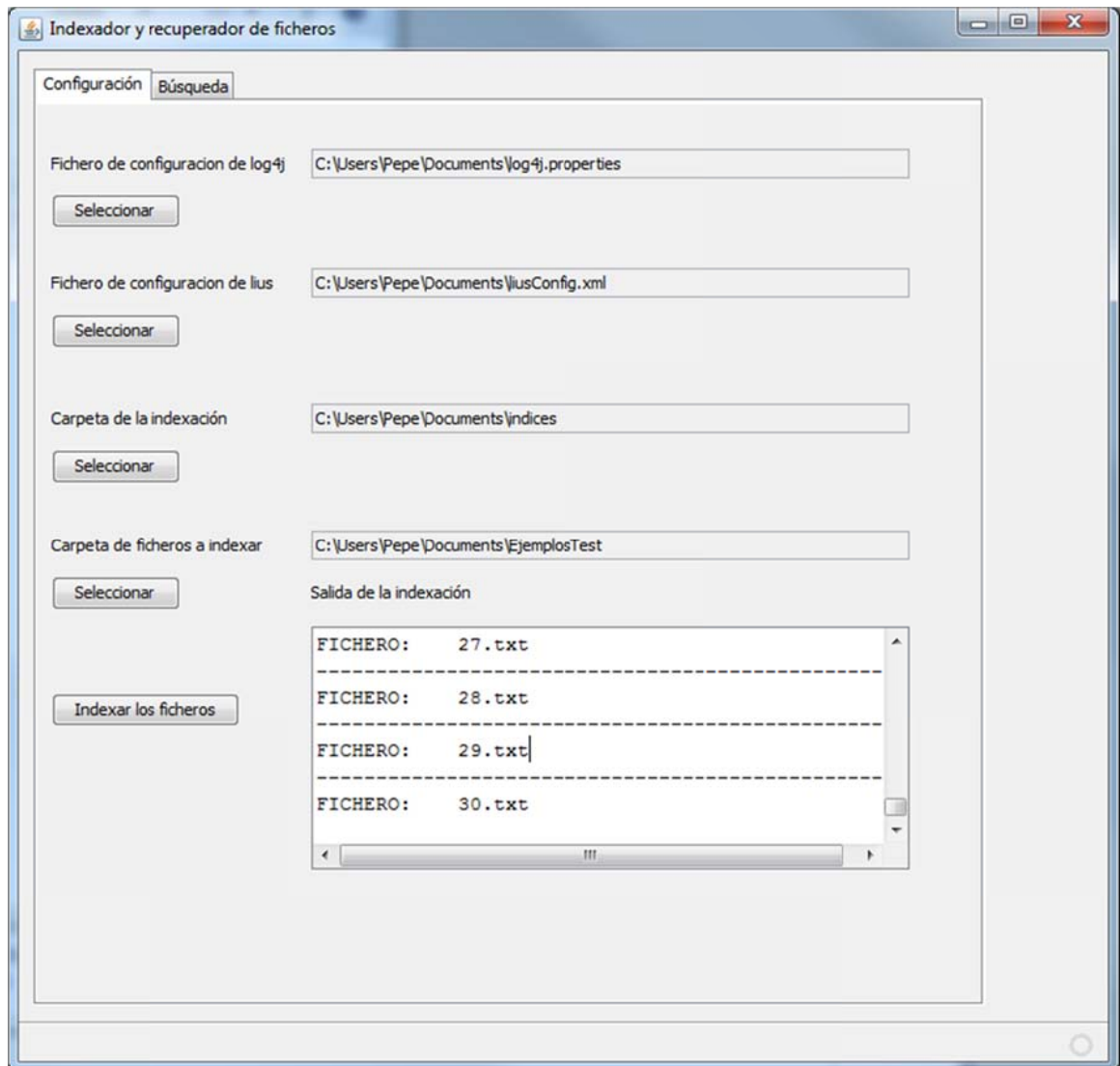
1. En primer lugar, se ejecuta una aplicación de escritorio en modo ventana que contiene la siguiente información:



2. El primer paso es seleccionar los ficheros y las carpetas necesarias para que se pueda indexar y posteriormente recuperar los ficheros que queramos:
  - El fichero log4j.properties en primer lugar.
  - En segundo lugar debemos seleccionar el fichero liusConfig.xml.
  - La carpeta donde queremos que se queden los ficheros indexados con Lucene.
  - Y por último la carpeta que contenga todos los ficheros que queremos indexar.

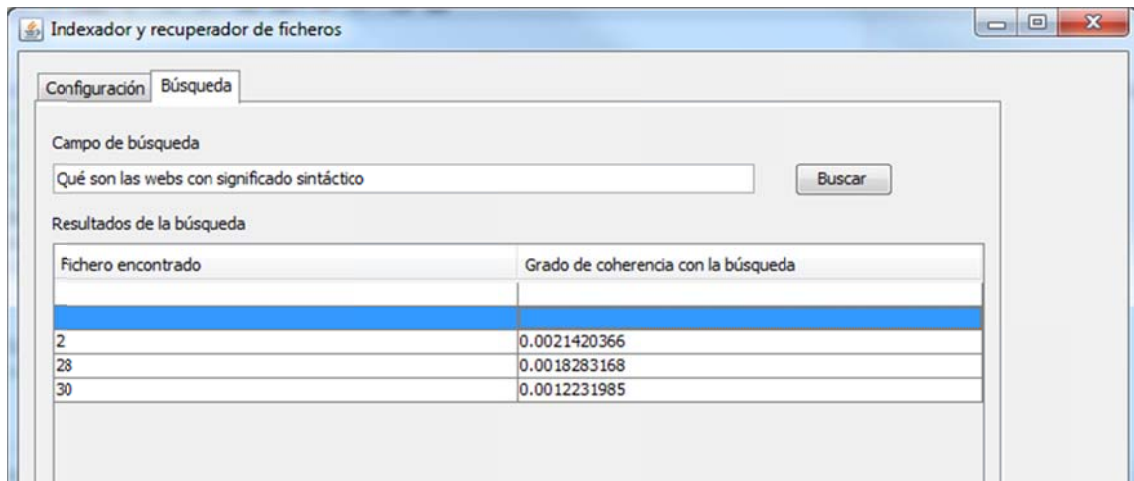
24 de junio de 2011

3. Una vez seleccionado todo, podemos pulsar sobre el botón “Indexar los ficheros” y obtendremos una salida en la que nos indicará uno por uno todos los ficheros que se han indexado.



24 de junio de 2011

4. Una vez indexados los ficheros, podremos ir a la pestaña “búsqueda” y realizar una búsqueda. En este caso he probado a poner la primera pregunta y el resultado es el siguiente:



En este caso particular, la web que debería haber devuelto era la 2 solamente, pero ha devuelto la 2, la 28 y la 30. Cabe destacar, que aunque haya devuelto 3 webs, la web que mayor grado de coherencia tiene, es precisamente la 2.

Lo que he observado con Lucene, es que las búsquedas que realizaba eran tan específicas que no encontraba realmente bien los ficheros que correspondían a esas preguntas. No obstante, sin ser unos resultados excelentes, no han sido malos del todo, ni difieren tanto de los resultados que deberían haberse obtenido.

## Actividad 2:

### Enunciado:

Usar FreeLing, EuroWordNet o MCR para definir (a mano) modificaciones/expansiones de las preguntas y analizar las diferencias con los resultados de la búsqueda con estas variantes y las preguntas originales (sin modificar/expandir) a través del motor de búsqueda de la actividad 1.

1. Variantes morfológicas
2. Variantes semánticas de palabras comunes
3. Variantes ortográficas
4. Expansión con sinónimos u otras palabras relacionadas

Para acceder a través de sus demos en la web, se sugieren:

- <http://nlp.lsi.upc.edu/freeling/demo/demo.php>
- <http://clic.ub.edu/es/eurowordnet-es>
- <http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl>

### Resolución:



---

24 de junio de 2011

El listado de quince preguntas inicial junto con sus variaciones es el siguiente:

1. ¿Qué son las webs con significado sintáctico?
  - a. V1 (morfológica): ¿Qué ser la web con significar sintáctico?
  - b. V2 (semánticas): ¿En qué consisten las webs con significado sintáctico?
  - c. V4 (sinónimos) ¿Qué representan las webs con representación sintáctica?
2. ¿Qué relación existe entre las estandarizaciones W3C y la web semántica?
  - a. V1 (morfológica): ¿Qué relación existir entre las estandarizaciones W3C y la web semántica?
  - b. V2 (semánticas): ¿Qué tienen en común las reglas W3C y la web semántica?
  - c. V4 (sinónimos): ¿Qué relación hay entre las reglas W3C y la web semántica?
3. ¿A qué llamamos webs de datos?
  - a. V1 (morfológica): ¿A qué llamar web de dato?
  - b. V2 (semánticas): ¿Qué son exactamente las webs de datos?
  - c. V4 (sinónimos): ¿A qué nombramos como webs de datos?
4. La web semántica ¿necesita una base de conocimiento?
  - a. V1 (morfológica): La web semántica ¿necesitar un base de conocimiento?
  - b. V2 (semánticas): ¿Realmente es necesaria la existencia de una base de conocimiento en las web semánticas?
  - c. V4 (sinónimos): La web semántica ¿precisa una base de sabiduría?
5. ¿En qué consiste la estructuración por metadatos?
  - a. V1 (morfológica): ¿En qué consistir la estructura por metadato?
  - b. V2 (semánticas): La estructuración por metadatos ¿en qué se fundamenta exactamente?
  - c. V4 (sinónimos): ¿De qué trata la conformación por metadatos?
6. ¿Quiénes son los investigadores iniciales de la web semántica?
  - a. V1 (morfológica): ¿Quién ser el investigador inicio de la web semántica?
  - b. V2 (semánticas): ¿Quiénes serían los primeros en investigar la web semántica?
  - c. V4 (sinónimos): ¿Quiénes son los primeros investigadores de la web semántica?
7. ¿Qué evolución ha habido desde la web 1.0?
  - a. V1 (morfológica): ¿Qué evolución haber desde la web 1.0?
  - b. V2 (semánticas): ¿Existe una evolución desde la web 1.0?
  - c. V4 (sinónimos): ¿Qué evolución existe desde la web 1.0?
8. ¿Es lo mismo web semántica que web 3.0?
  - a. V1 (morfológica): ¿Ser lo mismo web semántica que web 3.0?
  - b. V2 (semánticas): ¿Qué características tienen la web 3.0 y la web semántica entre sí?
  - c. V4 (sinónimos): ¿Es igual web semántica que web 3.0?
9. ¿Qué componentes posee una web semántica?
  - a. V1 (morfológica): ¿Qué componente poseer una web semántica?
  - b. V2 (semánticas): ¿Cuáles son los componentes de una web semántica?
  - c. V4 (sinónimos): ¿Qué componentes tiene una web semántica?
10. ¿Existen herramientas para crear webs semánticas más fácilmente?

---

24 de junio de 2011

- a. V1 (morfológica): ¿Existir herramientas para crear web semántica más fácil?
  - b. V2 (semánticas): Para crear webs semánticas de una manera sencilla, ¿existen herramientas en la actualidad?
  - c. V4 (sinónimos): ¿Hay instrumentos para hacer webs semánticas más sencillamente?
11. ¿Existen CMS o gestores de contenido orientados a esta nueva tendencia?
- a. V1 (morfológica): ¿Existir CMS y gestor de contenido orientado a esta nueva tendencia?
  - b. V2 (semánticas): ¿Los CMS están preparados para la web semántica?
  - c. V4 (sinónimos): ¿Hay CMS o gestores de contenido orientados a esta nueva moda?
12. ¿Ayuda el HTML5 a la inserción de la web semántica en la red?
- a. V1 (morfológica): ¿Ayudar HTML5 a insertar web semántica en red?
  - b. V2 (semánticas): ¿Aporta alguna ventaja HTML5 a la web semántica?
  - c. V4 (sinónimos): ¿Ayuda HTML5 a la introducción de la web semántica en internet?
13. ¿Es positivo para el posicionamiento SEO en buscadores la creación de webs semánticas?
- a. V1 (morfológica): ¿Positivo ser para posicionar SEO en buscador crear webs semánticas?
  - b. V2 (semánticas): ¿Es importante la web semántica en el SEO?
  - c. V4 (sinónimos): ¿Es bueno para el posicionamiento SEO en buscadores la invención de webs semánticas?
14. ¿En qué posición se encuentran las herramientas OWL, RDF, XML y SPARQL en la web semántica?
- a. V1 (morfológica): ¿En qué posición encontrar herramienta OWL, RDF, XML y SPARQL en web semántica?
  - b. V2 (semánticas): OWL, RDF, XML y SPARQL son herramientas web, ¿qué relación tienen con la web semántica?
  - c. V4 (sinónimos): ¿En qué lugar se encuentran los instrumentos OWL, RDF, XML y SPARQL en la web semántica?
15. ¿Existen buscadores semánticos?
- a. V1 (morfológica): ¿Existir buscador semántico?
  - b. V2 (semánticas): En la actualidad ¿se conoce algo sobre los buscadores semánticos?
  - c. V4 (sinónimos): ¿Hay buscadores semánticos?

Después de generar estas variaciones, he podido observar lo siguiente:

- En la aplicación realizada para poder recuperar las webs indexadas, he insertado cada variación de cada pregunta para obtener de nuevo resultados. He observado que por lo general, la variación de tipo V1, mostraba prácticamente los mismos resultados que la pregunta original. Sin embargo, las variaciones V2 y V4, en la mayoría de las ocasiones han logrado un mejor resultado que la primera pregunta propuesta originalmente en este trabajo.

---

24 de junio de 2011

- Estas observaciones demuestran que, la manera de expresar una pregunta es muy importante a la hora de recoger los datos, ya que, puede ser que si no nos expresamos como es debido, no obtengamos los resultados que realmente necesitamos obtener.

### ***Explicación de la aplicación realizada para esta actividad***

La aplicación creada para esta tarea, ha sido realizada con el IDE NetBeans 7.0 en lenguaje JAVA 6 (jdk 1.6.0) con la versión 2.9.4 de LUCENE y con la librería de LIUS LIUS (Lucene Index Update and Search) en su versión 1.0.

Indagando por la red, encontré una web que explicaba con bastante detalle cómo poder utilizar la librería LIUS para la indexación y recuperación de documentos.

Las clases creadas para este proyecto han sido las dos siguientes:

- IndexadorMC.java
- ModCompuAppliApp.java

La primera, **IndexadorMC.java**, contiene la lógica de indexación y recuperación de los ficheros:

- Una de las funciones importantes que están dentro de esta clase es *indexar*. Llamando a esta función, el objeto recoge como parámetro la situación de 4 rutas importantes:
  - ✓ El fichero de configuración para la librería log4j.
  - ✓ El fichero .xml para que funcione la librería LIUS.
  - ✓ La carpeta donde queremos dejar los ficheros que se crean al indexar los documentos.
  - ✓ Y la carpeta donde se encuentran los ficheros que vamos a indexar

Teniendo esta información, en esta función, con ayuda de la librería LIUS, recorre los ficheros de la carpeta que le indicamos y los indexa, creando una serie de ficheros que se leen de forma especial.

- Mediante la función *buscar*, realizamos la búsqueda que queremos sobre los ficheros indexados en la carpeta que hemos indicado con anterioridad.

La clase **ModCompuAppliApp.java** únicamente contiene la parte gráfica de la aplicación, no tiene ningún algoritmo de indexación ni recuperación en su interior.

### Actividad 3: Comentarios Finales

#### Enunciado

Aunque no tiene entidad de tarea, es obligatorio incluir en la última entrega de la práctica algunos comentarios y críticas a la práctica, sobre las dificultades, el interés del enunciado, etc.

#### Resolución

La práctica, a mi parecer, ha sido una buena forma de poder aplicar los conocimientos teóricos que se adquieren a lo largo del curso mediante la realización de los correspondientes resúmenes de cada tema.

En las primeras fases de la práctica, se hace un poco pesada la recopilación y la lectura de tantísima información. Una de las dificultades que tuve en estas fases, fue la de crear distintas categorías y asociarlas a las páginas que ya había buscado. Principalmente porque, al ser webs de un mismo tema concreto, sólo que con algunas diferencias muy particulares, se hace difícil crear 15 preguntas que realmente nos lleven a un número pequeño de las 30 páginas seleccionadas.

Por otra parte, en las últimas fases de la práctica, las dificultades fueron mayoritariamente por no entender muy bien los enunciados por culpa mía. En el momento en el que se solventaron mis dudas, pude realizar las tareas sin ningún problema, con un mayor o menor esfuerzo a la hora de hacer cada actividad, pero en definitiva, fueron buenas tareas para seguir aprendiendo.

La fase 4 y última de la práctica, para mí ha sido la mejor de las cuatro, quizá porque el crear una aplicación para poder probar la eficacia de la indexación y búsqueda de ficheros, me resultaba muy interesante.

Tuve ciertos problemas para familiarizarme bien con Lucene, pero indagando mucho por internet, se encuentra material muy bueno para utilizar. Concretamente, hice uso de unas librerías que había creadas y que ayudaban a hacer la indexación de ficheros y la búsqueda de ficheros de forma más sencilla. Utilicé NetBeans para cargar el proyecto, y para poder indexar las webs, hice 3 pruebas.

1. Comencé intentando indexar los ficheros .htm de las 30 páginas, pero a pesar de tener un parser interno que debía leer estos ficheros bien, en muchas ocasiones fallaba y no realizaba bien la indexación.
2. Después lo que hice fue, volcar la información de todos los artículos en 30 ficheros .txt, y esto funcionó mejor. Aun así me encontré con otro problema
3. El problema eran los acentos. Por ejemplo, la palabra “semántica” me la indexaba por separado “sem” y “ntica”, con lo cual a la hora de realizar la búsqueda, tenía problemas. Así que finalmente esos 30 ficheros .txt, los actualicé quitándole las tildes, y así no tuve ningún problema.

---

24 de junio de 2011

Una vez creada la aplicación, fue algo tedioso el tener que anotar las webs que aparecen como resultado de una búsqueda y realizar la comparación con el valor que tenía que aparecer en la realidad. Son muchos datos y lleva bastante tiempo realizarlo, pero muy interesante.

Por lo general, me ha parecido una práctica que conlleva bastante trabajo y muchas horas de esfuerzo, pero que merecen la pena, ya que, gracias a ella, se pueden comprender mucho mejor todo lo que se estudia en la teoría. En muchas ocasiones, la teoría se leía y se entendía, pero costaba imaginar cómo se aplicaba a la práctica.