Resultados búsqueda proximal práctica 3

Daniel Ruiz Mayo, Alberto Javier Parramón, Víctor de Juan

UAM -

5 de abril de 2016 20:00

Apuntes UAM
Doble Grado Mat.Inf.

Código en Github

1. Ejecución de la práctica

Para probar las distintas funcionalidades implementadas tenemos que tener como directorio trabajo el directorio de la práctica. En este caso $bmi1415_p3_12$.

Si se desea crear el índice de nuevo. Se debe ejecutar la clase $Indexador.java$. Cogerá automáticamente el indice de clueweb-1K situado en el directorio $colecciones/clueweb-1K$.

2. Pruebas del buscador

Para probar el buscador se debe ejecutar la clase $ProximalSearcher.java$. Se le pedirá al usuario que introduzca por teclado la búsqueda que desea realizar.

Si se desea más información como el $score$ de cada documento o más detalles del modo en el que esta se realiza se puede consultar el código del método $search$ y descomentar algunas de las últimas líneas de dicho método, en concreto:

Para ver el score de cada documento.

```
//System.out.println("Score documento " + this.getDocName(docId) + ": " + score);
```

Para ir viendo la evolución del algoritmo de búsqueda proximal:

```
//System.out.println("(" + a + "," + b + ")");
```

Son comentarios que se han ido utilizando a modo de debug y que hemos preferido dejarlos con el objetivo de que sea fácil de ver el funcionamiento del algoritmo implementado; el cual es el mismo que se nos expone en las transparencias de la asignatura.

Aquí mostramos algunos resultados del fichero queries.txt:

- 'Obama family tree'
  - Mejor documento: clueweb09-en0010-79-2218.html
  - Extracto de la búsqueda:
    *l down our category **tree** and find the topic that fits your interest best. Then start creating your game. Remember to assign a clear and descriptive title to your trivia quiz. People use our search function in a logical way. For example, if you are writing a quiz about "african elephants", entitle your quiz "African Elephants" or "The African Elephant Quiz". But if your quiz is about "african elephant poaching", your title should reflect the content of your quiz as faithfully as possible. Hence, "African Elephant Poaching" or "Poaching of African Elephants" would be appropriate titles. Try to capitalize the first word of each noun, as in the example above. It is important that all trivia quizzes obey a specific length rule. Quizzes can have any maximum number of questions (heck, our Trivia Marathon is endless) but should have a minimum of 12 questions each or more in order to provide a relative challenge for the player. Of course, always make sure to triple check your sources of information and verify the accuracy of the quizzes you develop. After all, your reputation is at stake*

[0] Documento compilado el 5 de abril de 2016 a las 20:00

here. The Masters of Trivia Community will reward or penalize depending on how good or bad a job you do. Finally, needless to say, we don't tolerate any profanity or discriminatory language on our site. No need to waste your time trying; offensive quizzes will be swiftly removed by the site administrators or the Community. Games Played Game Category Points Barack ***Obama Family Tree*** *Trivia Politic*

- 'french lick resort and casino'
  - Mejor documento: clueweb09-en0006-85-33176.html
  - Extracto de la búsqueda:
    549 Luxury ***Resort****s:* ***French Lick*** *Springs Indiana In Midwest area hotels spas casinos conferences centers suites romantic family vacations getaways packages weddings meetings planning Phone: 812-936-9300 Fax: 812-936-2100 8670 West State Road 56* ***French Lick****, Indiana 47432 1-888-936-9360 12/19/08 Conde Nast Traveler Gold List 12/15/08 Renowned Performers to Headline at* ***French Lick*** *11/3/08- Condé Nast Traveler Readers Pick West Baden Springs Hotel for Top 100 List 10/31/08- More than 500,000 Reasons to Visit* ***French Lick Resort*** *for the Holidays 10/9/08-* ***French Lick*** *Frightens* ***and*** *Delights with Halloween Activities 10/3/08- Autumn Magic Festival 8/7/08-* ***Resort*** *Hosts Special Guest from Chicago Boys & Girls Club 7/21/08-* ***French Lick*** *on Schedule to Open Pete Dye Course 7/8/08- Entertainment Performers to Headline at* ***French Lick Resort*** *6/25/08-* ***French Lick*** *Announces World Class Driving Festival 5/19/08-* ***French Lick*** *Announces New Activities Program 5/12/08- Tickets now on sale for Styx at* ***French Lick*** *5/7/08- Bass Tournament Nets $13,658 for Children 4/28/08-* ***French Lick Resort Casino*** *wins 2008 Gold Tee*
- 'getting organized'
  - Mejor documento: clueweb09-en0002-44-10207.html
  - Extracto de la búsqueda:
    Month Why Organize ***Getting Organized*** *Tools of the Trade*
- 'toilet'
  - Mejor documento: clueweb09-en0001-14-26411.html
  - Extracto de la búsqueda:
    ntent-Length: 66288 ***toilet*** *- wax seal* ***toilet*** *-*
- 'mitchell college'
  - Mejor documento: clueweb09-en0006-23-17720.html
  - Extracto de la búsqueda:
    72.0 2.Colorado ***College****.... 24 14-10 1712 71.3 3.Chapman University.. 7 6-1 487 69.6 4.Meredith* ***College****.... 24 13-11 1638 68.2 5.University of Dallas 25 11-14 1679 67.2 6.Finlandia University 18 6-12 1176 65.3 7.Nebraska Wesleyan... 26 10-16 1693 65.1 8.****Mitchell College****.... 24 9-*

3. Evaluación

Para ver los valores de $p@5$ y $p@10$ se debe ejecutar la clase $TestSearcher.java$. Se obtienen los siguientes resultados para el fichero queries.txt:

- Para $p@5$

```
Query: 1 0.2
Query: 2 0.0
Query: 3 0.2
Query: 4 0.0
Query: 5 0.0
```

promedio: 0.08

- Para $p@10$

```
Query: 1 0.5
Query: 2 0.0
Query: 3 0.1
Query: 4 0.0
Query: 5 0.0
```

promedio: 0.12

4. Ejercicio 2

El algoritmo de PageRank lo hemos implementado de forma matricial, utilizando la librería ejml (*Efficient Java Matrix Library*)

La implementación matricial del algoritmo la estudiamos en la asignatura *Modelización* del grado en Matemáticas y hemos utilizado los apuntes ([Valero,2015]) de la asignatura para la implementacíon del algoritmo.

5. Ejercicio 3

Para este crawler hemos usado como web de referencia wikipedia, y hemos tomado 100 paginas, generando un fichero llamado graph.txt con las urls salientes de estas 100 páginas. Si se quiere ejecutar, requiere un proceso manual, debido a que no se ha conseguido automatizar la creación de un zip en el que se incluyeran todos los html de la carpeta WEBCRAWLER sin que apareciera esta carpeta dentro del docs.zip. Se requiere una primera ejecución para descargar todos los documentos html y una vez realizada comprimir todos los archivos de la carpeta WEBCRAWLER en una zip llamado docs.zip

Como podemos ver en pagerank.txt las páginas con mayor score son

Por otra parte los resultados a la consulta Minería de datos en el indice son los que se ven en la imagen.

Estas páginas no nos sorprenden, pues corresponden a las páginas de wikipedia de Minería de datos, Tratamiento de base de datos,R (lenguaje de programación), SPSS y RapidMiner. Como detalle decir que si buscamos minería de datos en Google la primera opción es la misma que la nuestra, la pagina de mineria de datos de wikipedia.

https://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos	5.659309564233164E-4
https://es.wikipedia.org/wiki/Tratamiento_de_base_de_datos	5.659309564233164E-4
https://es.wikipedia.org/wiki/Inteligencia_artificial	5.659309564233164E-4
https://es.wikipedia.org/wiki/Variable_dependiente	5.659309564233164E-4
https://es.wikipedia.org/wiki/Histograma	5.659309564233164E-4
https://es.wikipedia.org/wiki/Mapa_autoorganizado	5.659309564233164E-4
https://es.wikipedia.org/wiki/M%C3%A1quina_oracle	5.659309564233164E-4
https://es.wikipedia.org/wiki/Ciencia	5.659309564233164E-4
https://es.wikipedia.org/wiki/Gen%C3%A9tica	5.659309564233164E-4
https://es.wikipedia.org/wiki/Ingenier%C3%ADa_el%C3%A9ctrica	5.659309564233164E-4

Tabla 1: *Top 10 de páginas con mayor puntuación*

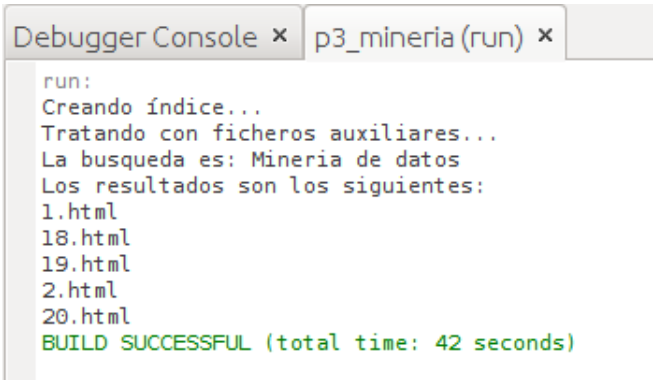

Figura 1: *Resultados para la consulta "Minería de datos"*