Resultados búsqueda proximal práctica 3

Daniel Ruiz Mayo, Alberto Javier Parramón, Víctor de Juan

UAM -

31 de marzo de 2016 01:43

Apuntes UAM
Doble Grado Mat.Inf.

Código en Github

1. Pruebas del buscador

- '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 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*

[0] Documento compilado el 31 de marzo de 2016 a las 01:43

- '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-

```
linea: termino ESPACIO lista_de_postings

lista_de_postings: posting
| posting ESPACIO lista_de_postings

posting: docId COMA num_posiciones COMA posiciones

posiciones: long
| long COMA posiciones
```

Por ejemplo:

```
libro 1,3,1,2,3 2,2,1,2
```

Significa que el término *libro* aparece en el documento con Id 1, 3 veces, en las posiciones 1,2 y 3. Y en el documento con Id 2, dos veces, en las posiciones 1 y 2.

Para la creación del índice en RAM hemos utilizado las siguientes clases:

- *Posting*: En esta clase guardamos la estructura de un posting, es decir, una lista de posiciones y el id del documento al que pertenecen.
- *Entrada*: En esta clase guardamos la estructura de una entrada, es decir, un término seguido de una lista de postings.
- *Indice*: En esta clase guardamos la estructura de indice, es decir, una lista de entradas.

En RAM se va guardando el índice estructurado de esa manera hasta completar un máximo de 400MB leídos de documentos. En ese momento fusionamos lo que tenemos en RAM con lo que tenemos en disco. Además, al crear el índice en disco, también guardamos dos tablas Hash:

- La primera tabla hash relaciona el id de un documento con el nombre del documento y su módulo.
- La segunda tabla hash relaciona un término del índice con la posición (offset de bytes) en la que se encuentra en el fichero de índice.

2. Búsqueda

Hacemos consultas orientadas a término, es decir leemos secuencialmente todos los postings de todos los docId de cada término de la consulta, y usamos la función de scoring vectorial tf-idf, implementada de la forma vista en clase.

Estos scores los guardamos en un TreeMap con clave el docId, y vamos sumando los tf-idf.

Para la búsqueda literal, procedemos de forma similar.

En primer lugar guardamos los postings del primer documento y a partir de estos vamos reduciendo el número de postings que no cumplan con la función que nos dice si el siguiente termino es el siguiente en posición en el documento respecto del anterior.

De esta manera mantenemos en memoria una lista de docId, posiciones de todos los elementos del primer término y progresivamente eliminamos los que no están de forma consecutiva.

Los docId que siguen estando se van almacenando en un TreeMap como en la anterior búsqueda, y se van sumando los scores.