IIC2413 — Bases de Datos

MongoDB

Durante esta semana trabajaremos con MongoDB, una base de datos de documentos JSON. Tu primera labor será familiarizarte un poco con mongoDB. Busca el tutorial en

`http://www.tutorialspoint.com/mongodb/mongodb_create_database.htm`

y aprende a:

- Crear una base de datos con un nombre en particular
- Insertar documentos en una base de datos
- Usar el comando `find()` para buscar documentos
- Usar el comando `pretty()` para mostrar documentos
- Crear índices

1. Wikidata

Baja y descomprime el archivo `"wiki_500_json.zip"`. Tienes en tus manos un pequeño documento con 500 documentos JSON. El documento tiene este formato:

```
[
{JSON 1},
{JSON 2},
    .
    .
    .
{JSON 500}
]
```

Los documentos JSON que tienes representan información de Wikidata, la base de datos detrás de wikipedia. Busca más información acá:

> https://www.mediawiki.org/wiki/Wikibase/DataModel/JSON

La estructura de alto nivel de estos documentos es la siguiente:

```
{
  "id": "Q33",
  "type": "item",
  "labels": {},
  "descriptions": {},
  "aliases": {}
  "claims": {},
  "sitelinks": {},
}
```

Este documento representa entidades de wikidata. El campo "id" es el identificador de cada entidad. Por ejemplo, Q298 corresponde a "Chile", Q33 corresponde a "Finlandia" y Q60 corresponde a "Nueva York". Para ver como se despliega la información de Chile, ingresa a

https://www.wikidata.org/wiki/Q298

Y para ver el archivo JSON,

https://www.wikidata.org/wiki/Special:EntityData/Q298.json

Puedes copiar y pegar acá para que aparezca más legible:

http://jsonprettyprint.com

Importando

Tu primer deber es importarlo a tu base de datos. Para eso usa, desde el terminal, el siguiente comando (asumiendo que el archivo está en tu carpeta de trabajo actual, que la base de datos es test y la collection a la que van estos documentos es entidades):

```
mongoimport --db test --collection entidades --drop --file wiki_500.json --jsonArray
```

Listo! Documentos importados. Conéctate a mongo y ejecuta lo siguiente (deberás ver documentos json):

```
use test
db.entidades.find();
```

2. Consultas

La consulta básica se realiza mediante la función `find()`. Opcionalmente puedes ejecutar find con dos parámetros que también son documentos JSON. El primer parámetro corresponde a filtros (selección) y el segundo a proyección. Por ahora, mira la estructura de uno de estos documentos. Puedes escribir lo siguiente, para hacer que aparezcan los documentos de una forma un poco más legible:

```
use test
db.entidades.find().pretty();
```

Filtros

El primer parámetro son los filtros, o selección. La selección por igualdad es simple: la siguiente consulta busca el documento con id igual a "Q33":

```
db.entidades.find({id: "Q33"}, {}).pretty();
```

Puedes ver el tutorial para ver qué otros tipos de filtros se pueden usar.

Proyección

El segundo parámetro de `find()` corresponde a los campos a proyectar. Para ver las entidades que tenemos en nuestra base de datos, escribe (no es necesario usar comillas para los strings, el sistema las asume):

```
db.entidades.find({}, {id: 1});
```

El primer parámetro es vacío por que no especificamos selecciones, y el segundo es un documento que contiene `{id:1}`. Esto significa que quieres proyectar el documento JSON, mostrando solo el campo `id`. En general puedes listar cualquier cantidad de parámetros a mostrar `{key1:1,key2:1,...}` o, si quieres dejar afuera algunos parámetros, cambiando los 1s por 0s. Por ejemplo, ejecuta

```
db.entidades.find({}, {labels: 0, aliases: 0, claims: 0, sitelinks: 0}).pretty();
```

Para dejar fuera la parte de labels, aliases, claims y sitelinks.
Exploremos un poco tu base de datos!

Wikidata: entidades

Veamos lo que tiene nuestra base de datos. La parte bajo "labels"(etiquetas, en inglés) de los documentos contienen las etiquetas en diferentes idiomas. Ejecuta

```
db.entidades.find({}, {"id": 1, "labels.en": 1}).pretty();
```

para buscar los ids de los documentos junto con sus etiquetas en inglés (aquí `"labels.en"` se refiere a el objeto con llave "en", que está dentro del objeto con llave "labels"). Ubica el label en inglés de Finlandia (Q33).

1. Escribe una consulta que, dada una entidad, retorne su label o etiqueta correspondiente al idioma inglés, junto a su descripción. Prueba tu consulta para Q298 (Chile) y P36 (Capital).

Una vez que tenemos la etiqueta, podemos agregar su descripción. El formato para la descripción es similar a las etiquetas:

```
"descriptions": {
   .
   .
   .
   "en": {
     "language": "en",
     "value": "country in South America"
   }
   .
   .
   .
 }
```

Escribe las siguientes consultas. ¡Busca en la documentación si lo necesitas!

2. Escribe una consulta que te de todas las entidades junto a sus descripciones, ordenadas alfabéticamente según su id.

3. Utiliza `limit` en el resultado de tus consultas para obtener porciones de los documentos más manejables.

(Bonus) Wikidata: conexiones

Si miras la página de wikidata de chile, verás que aparece una cajita que dice "capital", y al lado aparece Santiago.
¿Qué está pasando? Wikidata almacena propiedades sobre las entidades. En el caso anterior, la propiedad es "capital", y relaciona a "Chile" con "Santiago". En realidad, el id de "Santiago" es Q2887. Entonces, Wikidata relaciona a Q298 con Q2887 mediante la etiqueta P36.
La estructura de esta relación está en el mismo documento de Chile (Q298), buscalo con:

```
db.entidades.find({id: "Q298"}, {id: 1, "labels.en": 1, "claims.P36": 1}).pretty();
```

Además de la información de la etiqueta de la arista (P36), el valor que nos importa es donde sale "numeric-id". Verás que aparece solo un número, pero si anteponemos una Q nos queda Q2887, que es justo la entidad que corresponde a "Santiago".
Para las siguiente consultas es más fácil usar python y otro lenguaje de programación. ¿Puedes hacerlo en mongoDB?

3. Escribe una consulta que, dada una entidad, retorne (i) su label o etiqueta en inglés, (ii) su descripción en inglés, e (iii) para cada "claim" de esa entidad, retorne la etiqueta (en inglés) de la propiedad de ese claim y la etiqueta (en inglés) de la entidad que está siendo relacionada por ese "claim". Puede que algunas etiquetas no existan en tu pedazo de la base de datos, y en ese caso habría que poner algo así como "etiqueta no encontrada".

3. (Bonus) Índices

Acá hay un pedazo mucho más grande de Wikidata. Bájalo, e impórtalo tal como antes.

`https://drive.google.com/file/d/0B32bx9rxQY_CVWJrVElVcFZnbTA/view`

Ahora intenta buscar nuevamente a Chile o a Finlandia. No están, pero ¿ves cuánto se demora? Lo que pasa es que la base de datos no está indexada en el campo id.

1. Agrega un índice al campo id de cada documento
2. Busca nuevamente a Chile o Finlandia. ¿Notas la diferencia?
3. Busca alguna entidad que si esté en la base de datos