Taller de Indexación de Texto

Objetivo

Realizar consultas en la colección de tweets de la ciudad de Cali y en las páginas web de Wikipedia utilizando Solr, y añadir información en índice.

Procedimiento

1. Acceder a Solr

El programa Solr ya se encuentra instalado. Para acceder al servidor web de Solr ingrese en su buscador web de preferencia la siguiente URL: **http://cursobigdata.cloudapp.net:8983/solr/** Al hacerlo podrá visualizar la interfaz de Solr tal como se muestra en la siguiente figura.

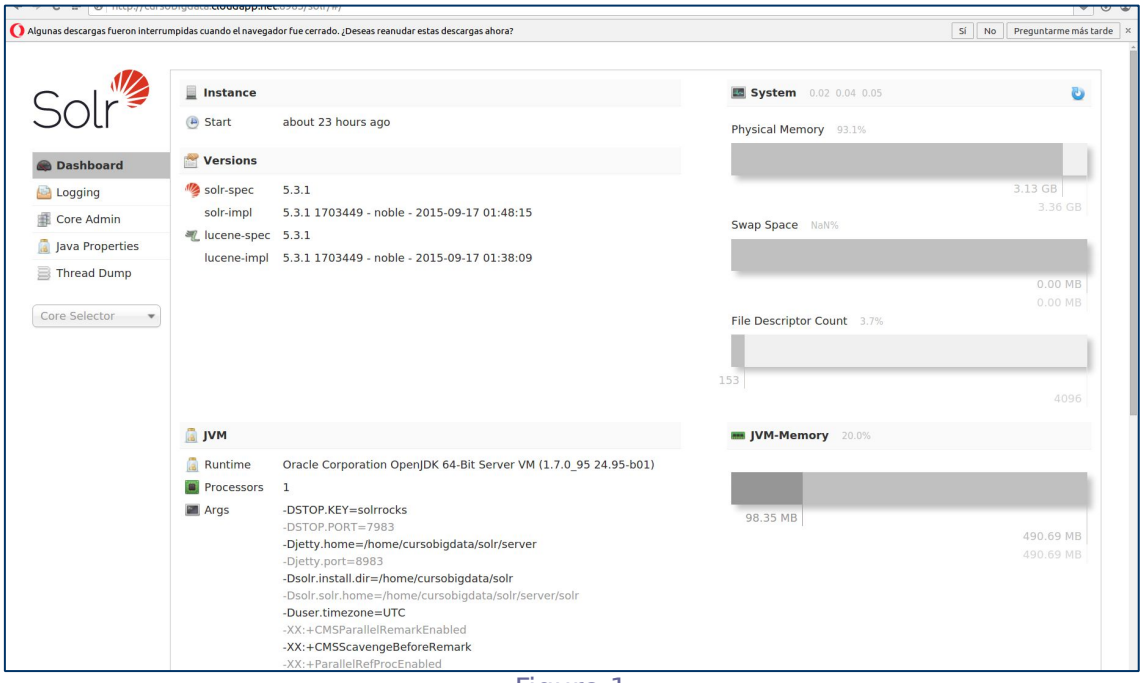

Figura 1

2. Ver las colecciones existentes

Para visualizar la lista de las colecciones existentes de clic en **Core Selector**, ubicado en el panel del costado izquierdo de la pantalla (recuadro de color rojo, ***Figura 1***). Como se observa en la ***Figura 2*** deberán aparecer listadas la colección de **Tweets** y de **Wikipedia**.

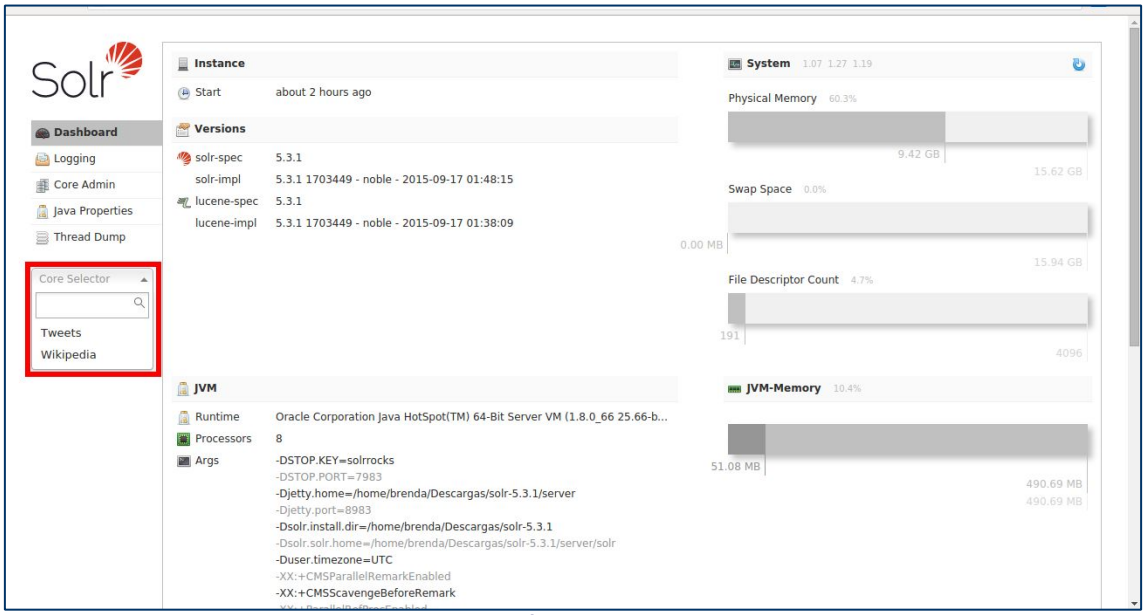

Figura 2

3. Realizar consultas en la colección de Tweets

Acceda a la colección de **Tweets** haciendo clic sobre esta una vez desplegadas las colecciones existentes (***Figura 2***). Al abrir esta colección se puede observar la cantidad de documentos que contiene.

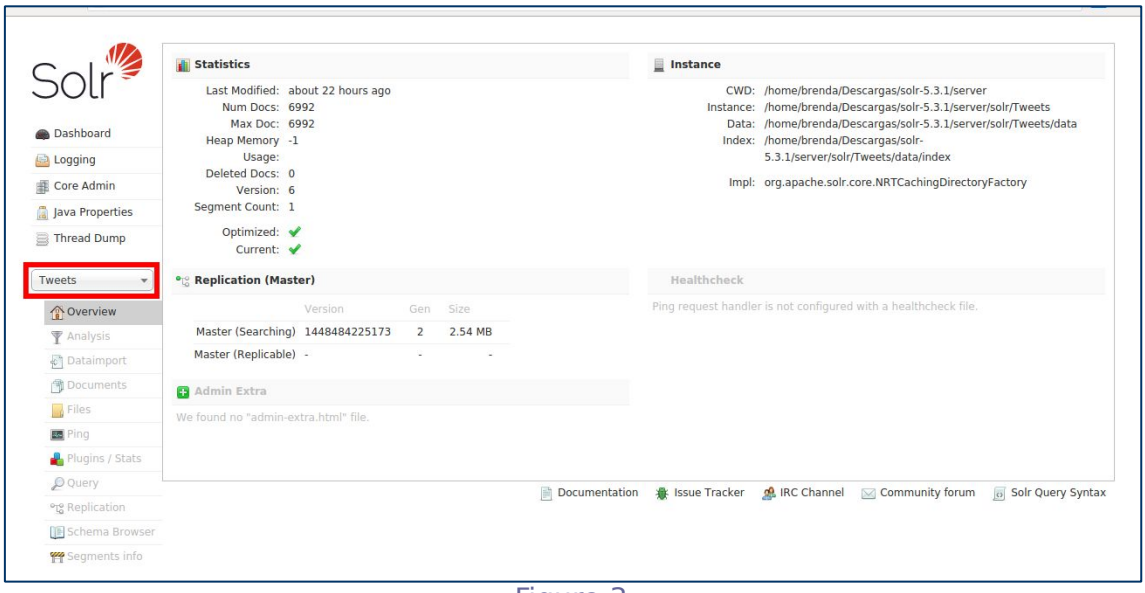

Figura 3

Para realizar consultas en los tweets existentes de clic en **Query**, ubicado en el panel del costado izquierdo de la ventana (***Figura 4***, recuadro de color rojo).

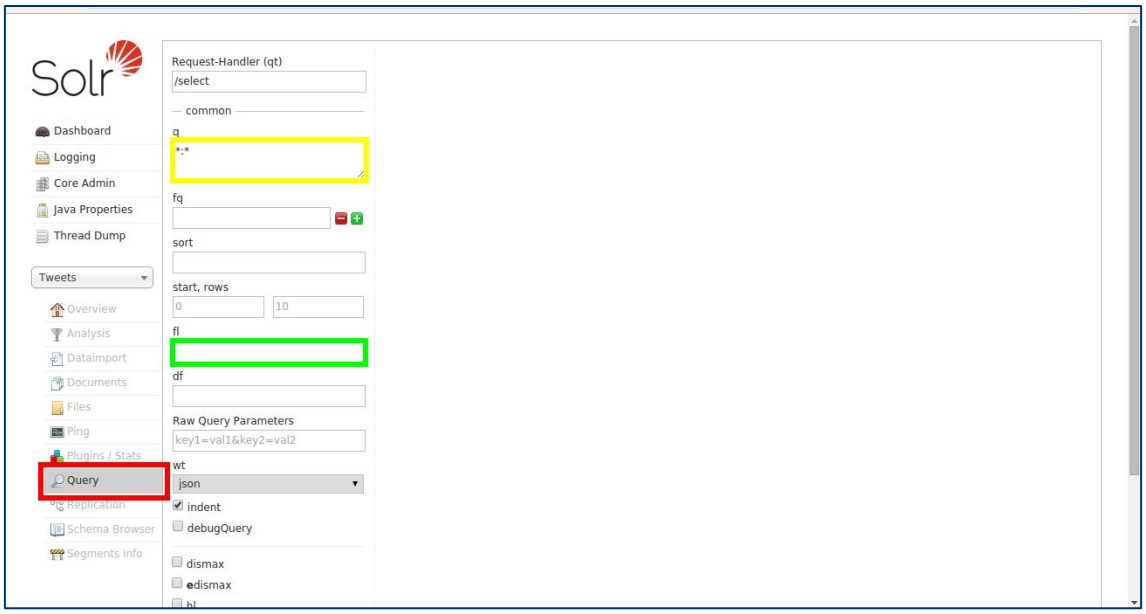

Figura 4

Ingrese la palabra, o la serie de palabras separadas por coma (**,**) que desea buscar reemplazando los caracteres existentes en el recuadro de color amarillo (***Figura 4***).

Si desea puede ingresar los campos que quiere visualizar separados por coma (**,**) en el recuadro señalado de color verde en la ***Figura 4***. Algunos de los campos existentes son: **userName**, **country**, **place**, **text** y **date**. De no ingresar un campo o una serie de campos específicos para visualizar, al realizar la consulta se mostrarán todos los campos existentes en el documento.

Desplácese hacia abajo y de clic en **Execute Query**, como se observa en el recuadro de color azul de la siguiente figura.

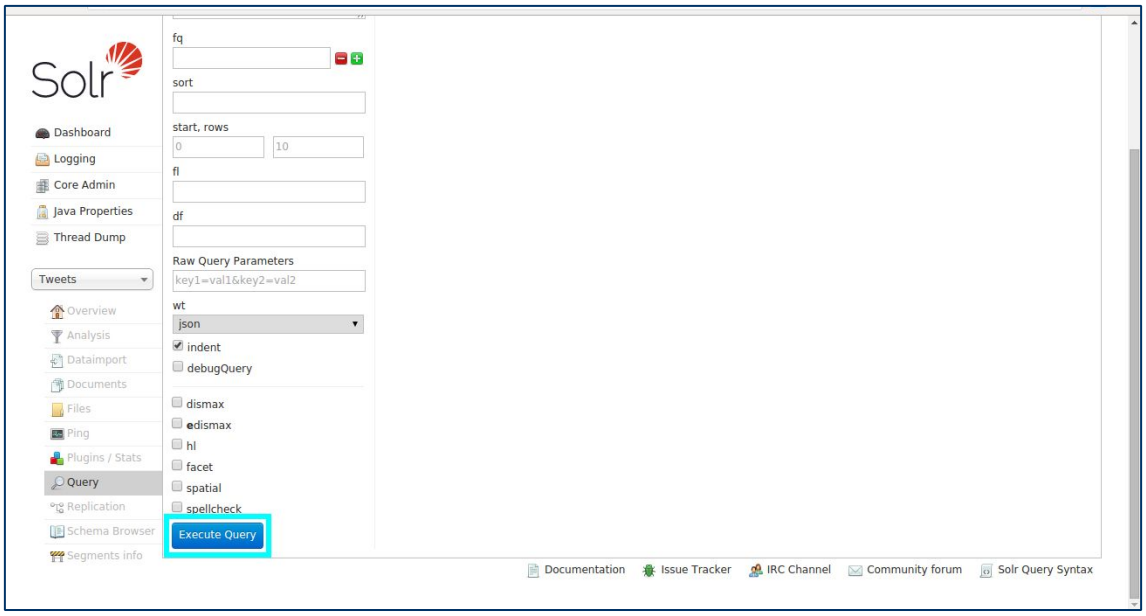

Figura 5

Una vez de clic en **Execute Query** podrá observar al costado derecho de la pantalla el resultado de su búsqueda.

4. Añadir un documento a la colección de Tweets

Para añadir un documento, de clic en **Documents** ubicado en el panel del costado izquierdo de la pantalla (***Figura 6***, recuadro de color rojo). Posteriormente escriba en el recuadro señalado de color amarillo lo siguiente:

`{"id": "`**`CÉDULA`**`", "country": "`**`PAÍS`**`", "place": "`**`LUGAR DE NACIMIENTO`**`", "date": "`**`FECHA DE NACIMIENTO`**`", "userName": "`**`NOMBRES Y APELLIDOS`**`", "text": "`**`TEXTO`**`"}`

Reemplace **CÉDULA** por su número de cédula, **PAÍS** por su país de origen, **LUGAR DE NACIMIENTO** por su lugar de nacimiento, **FECHA DE NACIMIENTO** por su fecha de nacimiento (AAAA-MM-DD), **NOMBRES Y APELLIDOS** por sus respectivos nombres y apellidos, y **TEXTO** por el texto que desee compartir. En todos los casos, conserve las comillas existentes (**“ ”**).

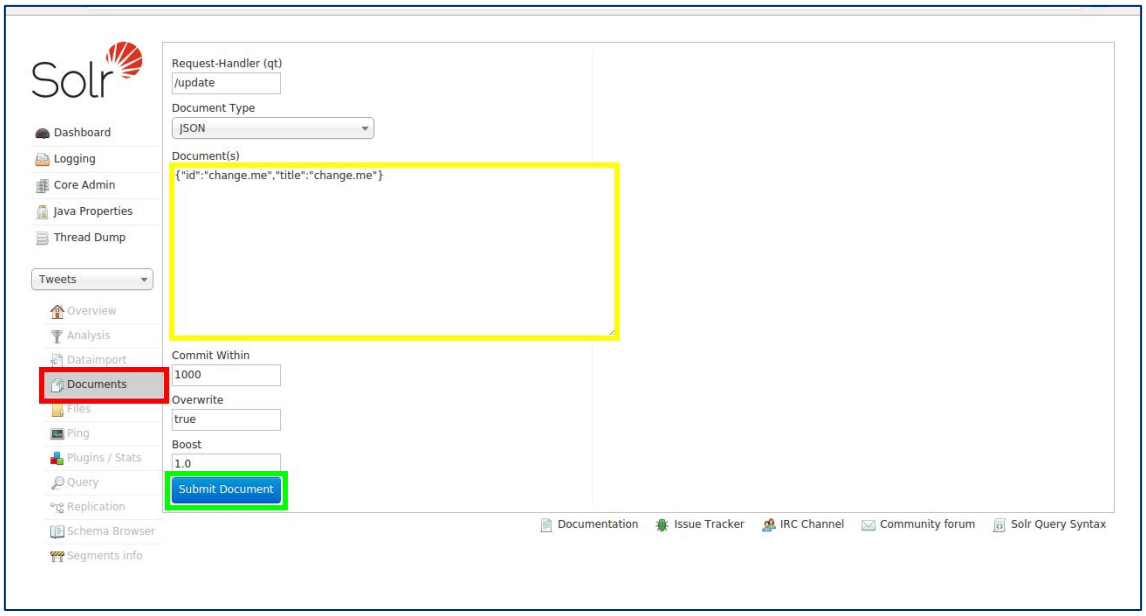

Figura 6

Posteriormente de clic en **Submit Document** (recuadro de señalado de color verde en la ***Figura 6***) para enviar el documento. Si no se encuentra ningún error en el documento deberá aparecer al costado inferior derecho de la pantalla **Status: success.**

5. Buscar su documento en la colección

Verifique la existencia de su documento en la colección haciendo una query con su **número de identificación** siguiendo las instrucciones del ítem número tres del presente taller (***Realizar consultas en la colección de Tweets***).

En el recuadro de color amarillo señalado en la ***Figura 4*** ingrese su número de cédula y posteriormente de clic en **Execute Query**, recuadro de color azul señalado en la ***Figura 5***.

Una vez ejecute la consulta al costado derecho de la pantalla deberá aparecer el documento que previamente ingreso a la colección **Tweets** con todos los campos existentes, a menos que especifique aquellos campos que desea visualizar en el recuadro de color verde en la ***Figura 4***.

6. Actualice un documento

Para actualizar un documento se debe conocer el **id** (número único identificador) de dicho documento. Siga los pasos del ítem número cuatro del presente taller (***Añadir un documento a la colección de***

Tweets) para insertar un documento a la colección e ingrese en el recuadro señalado de color amarillo (***Figura 6***) lo siguiente:

`{"id": "CÉDULA", "country": "PAÍS", "place": "LUGAR DE EXPEDICIÓN", "date": "FECHA DE EXPEDICIÓN", "userName": "NOMBRES Y APELLIDOS", "text": "TEXTO"}`

Reemplace **CÉDULA** por su número de cédula (identificador único del documento), **PAÍS** por su país de origen, **LUGAR DE EXPEDICIÓN** por el lugar de expedición de su cédula, **FECHA DE EXPEDICIÓN** por la fecha de expedición de su cédula (AAAA-MM-DD), **NOMBRES Y APELLIDOS** por sus respectivos nombres y apellidos, y **TEXTO** por el texto que desee compartir. En todos los casos, conserve las comillas existentes (**“ ”**).

Ya que había insertado previamente un documento con ese número identificador, el documento a insertar lo va a reemplazar. Una vez de clic en **Submit Document** (recuadro de señalado de color verde en la ***Figura 6***) podrá comprobar los cambios realizando nuevamente una consulta con su número de identificación (cédula) siguiendo los pasos del ítem número cinco (**Buscar su documento en la colección**).

7. Abrir una terminal

Para abrir una terminal, siga las instrucciones de los ítems número 1 y 2 de la **Guía de conexión**.

8. Eliminar su documento

Para eliminar el documento que insertó a la colección de **Tweets** escriba en la terminal el comando que aparece a continuación luego del símbolo **$** y presione la tecla ENTER. Reemplace **ID_DOCUMENT** por su número de cédula.

```
~$ /home/cursobigdata/solr/bin/post -c Tweets -d
"<delete><id>ID_DOCUMENT</id></delete>"
```

Bloque de código 1

Una vez presione la tecla ENTER si realiza una consulta con su número de identificación no encontrará ningún documento que coincida con la búsqueda ya que ha sido eliminado.

9. Insertar las páginas web descargadas en el Taller de Crawling a la colección Wikipedia

Para indexar las páginas web descargadas de Wikipedia a la colección **Wikipedia** de Solr escriba en la terminal el comando que aparece a continuación luego del símbolo **$** y presione la tecla ENTER. Reemplace **CODE** por su **CÓDIGO**.

```
~$ /home/cursobigdata/solr/bin/post -c Wikipedia
CODE/Taller1-02/*
```

Bloque de código 2

La indexación de los 1000 archivos demora no más de 15 segundos y observará el transcurso del proceso. Una vez culminado dicho proceso podrá efectuar las operaciones vistas previamentes con la colección de **Tweets**.

10. Cierre la terminal y desconéctese del servidor

Para desconectarse del servidor, siga las instrucciones de los ítems número 5 y 6 de la **Guía de conexión**.

11. Realizar consultas en la colección de Wikipedia

Acceda a la colección de **Wikipedia** haciendo clic sobre esta una vez desplegadas las colecciones existentes (***Figura 2***). Al abrir esta colección se puede observar la cantidad de documentos que contiene.

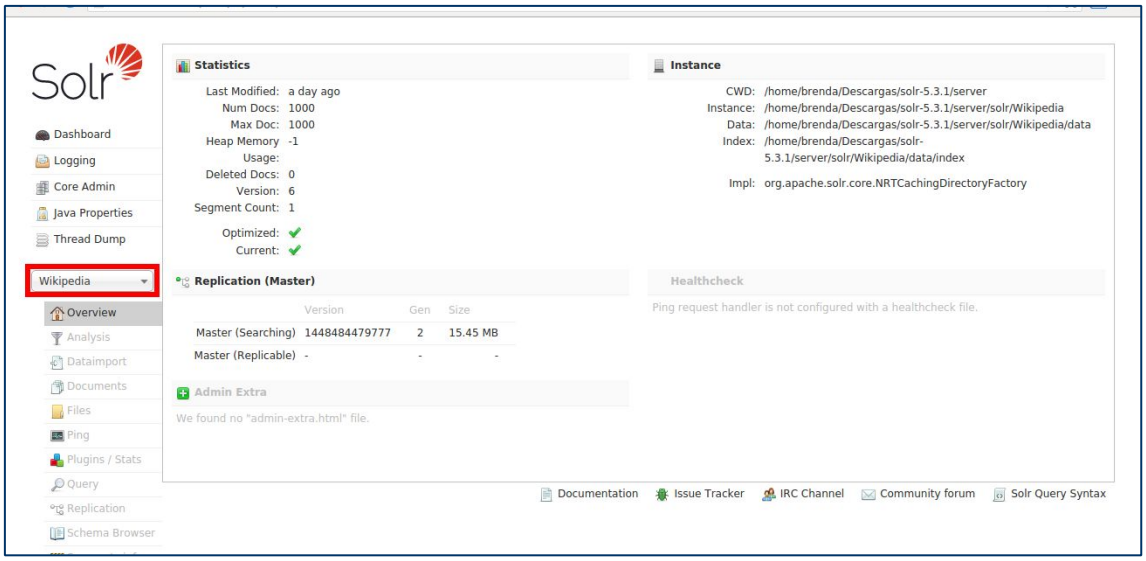

Figura 7

Para realizar consultas en las páginas de Wikipedia indexadas de clic en **Query** (***Figura 8***, recuadro de color rojo).

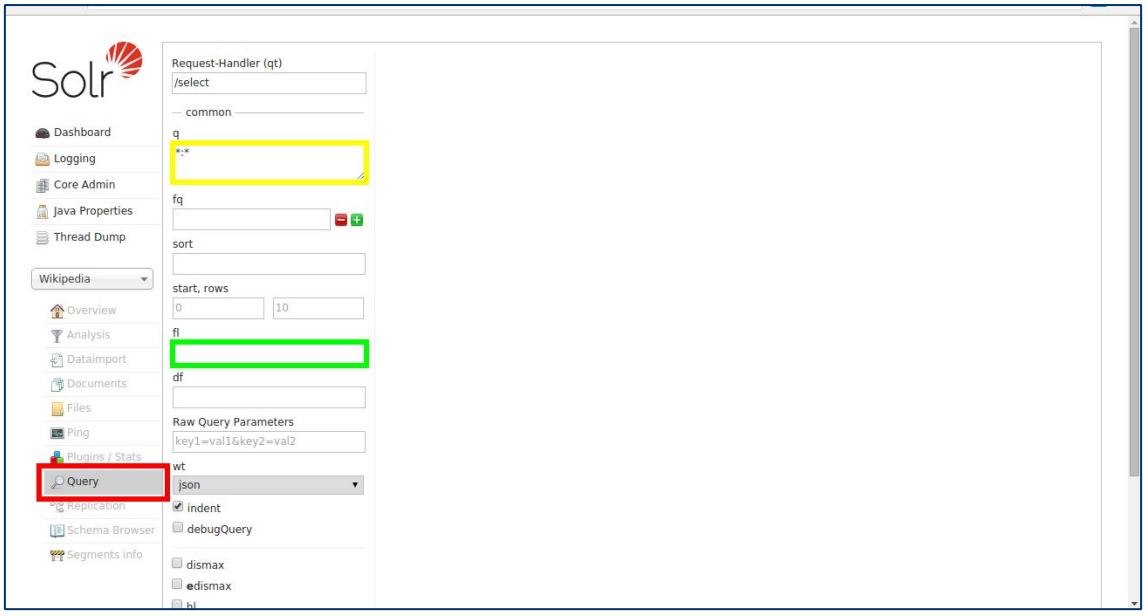

Figura 8

Ingrese la palabra, o la serie de palabras separadas por coma (**,**) que desea buscar reemplazando los caracteres existentes en el recuadro de color amarillo (***Figura 8***).

Si desea puede ingresar los campos que quiere visualizar separados por coma (**,**) en el recuadro señalado de color verde en la ***Figura 8***. Algunos de los campos relevantes son: **id** (contiene la ruta del archivo indexado por si desea acceder a una de las páginas web) y **title** (contiene el título de la página web). De no ingresar un campo o una

serie de campos específicos para visualizar, al realizar la consulta se mostrarán todos los campos existentes en el documento.

Desplácese hacia abajo y de clic en **Execute Query**, como se observa en el recuadro de color azul de la siguiente figura.

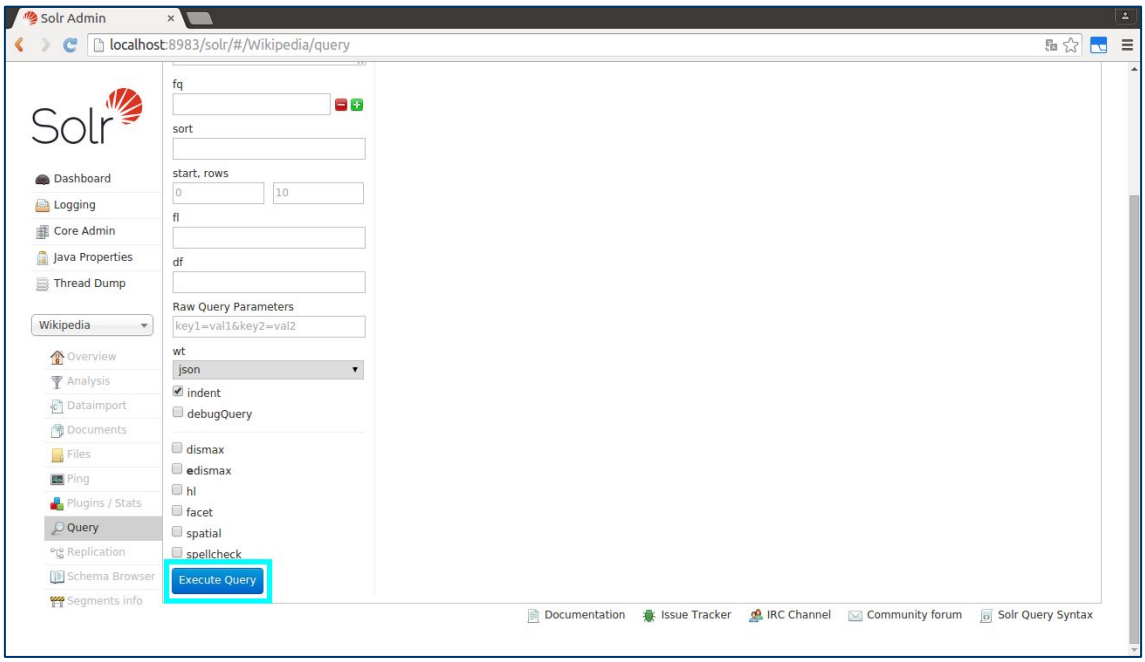

Figura 9

Una vez de clic en **Execute Query** podrá observar al costado derecho de la pantalla el resultado de su búsqueda. Ya que la búsqueda se efectúa en toda la página web puede que exista coincidencia de las palabras ingresadas con el contenido de ésta y no necesariamente con el correspondiente título.

Fin de la Guia