Querying Documents in MongoDB

By the end of this activity, you will be able to:

1. Find documents in MongoDB with specific field values.
2. Filter the results returned by MongoDB queries.
3. Count documents in a MongoDB collection and returned by queries.

Step 1. **Start MongoDB server and MongoDB shell**. Open a terminal window by clicking on the square black box on the top left of the screen.



Next, change to the *mongodb* directory, and start the server:



1

2

cd Downloads/big-data-3/mongodb

./mongodb/bin/mongod --dbpath db

The arguments *--dbpath db* specify that the directory *db* should be used for the MongoDB directory for datafiles. After starting the MongoDB server, you will see the following lines indicating that the server is running:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/vUHvPVT-EeaqTxIkdCEfsw_e7e71e936f7860924a71d716af7ba1db_server-start.png?expiry=1595289600000&hmac=xWJceXaZwZQy18MdUE8qqe_RAfXUnKJtgv1KUYSWYf4

Next, let's run the MongoDB shell so that we can query the server. Open a new terminal shell window, change to the *mongodb* directory, and start the shell:



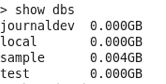
1

2

cd Downloads/big-data-3/mongodb

./mongodb/bin/mongo

Step 2. **Show Databases and Collections**. Run the *show dbs* command to see the databases:



The database named *sample* has been created and loaded with Twitter JSON data. Let's switch to that database by running the *use* command:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/XbHn51UAEeaX4QpLJOK7gQ_667612c1f60e7b8e88d160836d94f814_use-sample.png?expiry=1595289600000&hmac=XIwCj638KzsZ083x869v3OgckSgV0AMyQwVMaZkpE8A

We can see the collections in the *sample* database by running *show collections*:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/R6vI_lUAEealpAoth2FRAw_5725c1daf122e4a7aa2765c28e66f505_show-collections.png?expiry=1595289600000&hmac=WKkR-O53Jf0RddiNssuKn6RdVQdR3zDJqbILrPYZgNs

The Twitter data is stored in the *users* collection. We can run *db.users.count()* to count the number of documents:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/8ETyM1UAEeaztg6Pg6w09w_6f1d10814edde8375f05bcad18204c1f_users-count.png?expiry=1595289600000&hmac=PzG7xYEqsH1HJXtyJd4tAyEbRocTfcFAzbj6TdZ6rrs

Step 3. **Look at document and find distinct values**. We can examine the contents of one of the documents by running *db.users.findOne():*

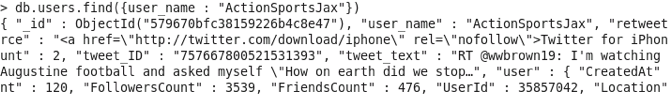


The document has several fields, e.g., *user\_name, retweet\_count, tweet\_ID, etc.*, and nested fields under *user,*e.g., *CreatedAt, UserId, Location*, etc.

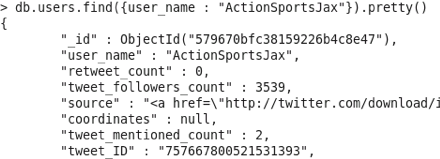
We can find the distinct values for a specific field by using the *distinct()* command. For example, let's find the distinct values for *user\_name*:



Step 4. **Search for specific field value**. We can search for fields with a specific value using the *find()* command. For example, let's search for *user\_name* with the value *ActionSportsJax*:



By appending *.pretty()* to the end of the find command, the results will be formatted:



Step 5. **Filter fields returned by query**.We can specify a second argument to the *find()* command to only show specific field(s) in the result. Let's repeat the previous search, but only show the *tweet\_ID* field:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/dhyyI1UIEeaubA6-qtnryw_89f1bb7b3938d3272130af19729933ed_find-filter.png?expiry=1595289600000&hmac=_xSgTDcX520QVR3kuoJcFRw7b-SJm6BmT-d_HbhzG28

The *\_id* field is primary key for every document, and we can remove it from the results with the following filter:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/lHwdR1UIEeaFpQoKLtdtHw_5ce2bd8165fe95919d41e67de270ad89_find-filter-no-id.png?expiry=1595289600000&hmac=la2J75OGSnRkgdJfoiyx7_6Cc24Z0gzRXN24qLUIqRc

Step 6. **Perform regular expression search**. MongoDB also supports searching documents with regular expressions. If we search for the value *FIFA* in the *tweet\_text* field, there are no results:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/ULQ1FlUKEeaFpQoKLtdtHw_77f0ff1da6ce5073f56e5b1e9c45195b_tweet-text-fifa.png?expiry=1595289600000&hmac=eCBUl5zZdpmtKgpU5c-gIzyPNAfPzGXL4ft1lEwVQ68

However, if we search using a regular expression, there are many results:

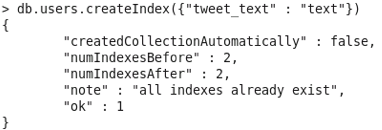


The difference between the queries is that the first searched for where the *tweet\_text* field value was exactly equal to *FIFA*, and the second searched for where the field value contained *FIFA*.

We can append *.count()* to the command to count the number of results:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/g5kUTlUKEeaqTxIkdCEfsw_83b90f0a247664cb2dcb4fffcf5393ec_tweet-text-re-count.png?expiry=1595289600000&hmac=TZCL_OWiTOD_RlVlW4iuqL81qNApAc6BEOhhe9jQHB0

Step 7. **Search using text index**. A text index can be created to speed up searches and allows advanced searches with *$text*. Let's create the index using *createIndex():*



The argument *tweet\_text* specifies the field on which to create the index.

Next, we can use the *$text*operator to search the collection. We can perform the previous query to find the documents containing *FIFA:*

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/HjV6dl2vEeaWHg4NuBHTLw_37e181106d07cd7e1d067de31f4c1433_text-search-fifa.png?expiry=1595289600000&hmac=xFyFSWuxsMXNWpBh_vD77xCLSADwvYZqXOrgBPOUHC4

We can also search for documents not containing a specific value. For example, let's search for documents containing *FIFA,*but not *Texas:*

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/KzVscl2vEeazDA6HAv4GcQ_62133c32f77c77606c0e5199008ec8e2_text-search-no-texas.png?expiry=1595289600000&hmac=1RMZBtXuiJAj0f3yX0r8pXdbsPXvZVrYFvzCnmhJEdo

Step 8. **Search using operators**. MongoDB can also search for field values matching a specific criteria. For example, we can find where the *tweet\_mentioned\_count* is greater than six:



The *$gt* operator search for values greater than a specific value. We can use the *$where* command to compare between fields in the same document. For example, the following searches for *tweet\_mentioned\_count* is greater than *tweet\_followers\_count*:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/lyxIy12-EeaJrxIK7EGdww_8512cd11bd19da57f3533f43b5711e2a_where.png?expiry=1595289600000&hmac=qNpEvm7VWj9AbGpVRyxZSsqgqHu2a1o1TY0vStRYfzg

Note that the field names for *$where* are required to be prefixed with *this,*which represent the document.

We can combine multiple searches by using *$and*. For example, let's search for *tweet\_text* containing *FIFA* and *tweet\_mentioned\_count* greater than four:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/iM3ZDl2-EeaJrxIK7EGdww_5359918a5d09a31b0696abdf0ae0db6e_find-and.png?expiry=1595289600000&hmac=LKppcU9Iqd3OMRAGNxTZOmGSit9S1i2E9yvnobfkb94

When you are done querying MongoDB, run *exit* in the MongoDB shell, and *Control-C*in the terminal window running the server.