Data Management and Ethics - Exercises

Axel Oehmichen
ao1011@imperial.ac.uk

Friday 25th June 2021

The task is to download, modify and insert few tweets into a MongoDB database and then run simple mongo queries against the newly created database. We assume that your docker environment is ready; if not, please see the instructions in Week 1.

Twitter:

Create a Twitter account and create a new app to retrieve the consumer key, consumer secret, access token key and access token secret:

- https://apps.twitter.com/

Twitter API dev documentation :

- https://developer.twitter.com/en/docs/api-reference-index

Keep them at hand, we will need them in the notebook.

Spark & MongoDB:

We will be using a different set of containers for this exercise. This exercise is contained within the same environment as the courseworks, so you can use this environment for both the courseworks and the exercises.

If you use git, you can run the command:

git clone https://gitlab.doc.ic.ac.uk/theinis/dockerdm.git

If you do not use git, you can download the files as a zip here:

https://gitlab.doc.ic.ac.uk/theinis/dockerdm

Then open a terminal inside the dockerdm folder and run the command:

docker compose up

NB: The first setup of the different containers takes a bit of time, so please be patient

Once the environment is ready and all the containers are started, your shell should display something like this:

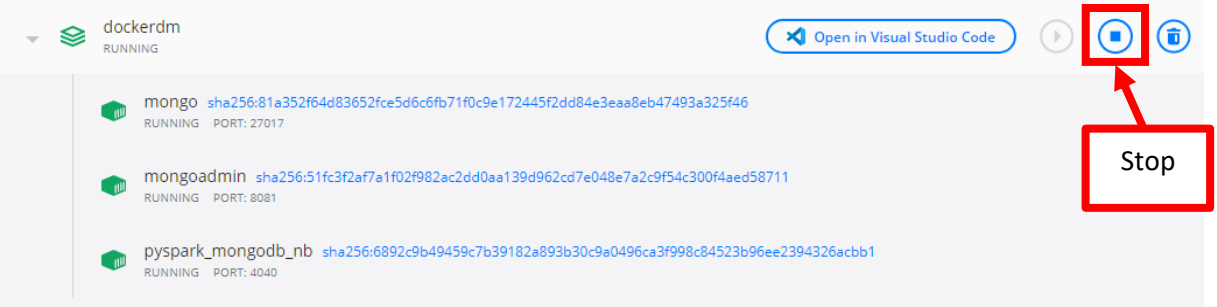
```
[+] Running 4/2
 - Network dockerdm_default      Created
 - Container mongo               Created
 - Container mongoadmin          Created
 - Container pyspark_mongodb_nb  Created
Attaching to mongo, mongoadmin, pyspark_mongodb_nb
```

Once all of this is ready, in an identical fashion to the previous exercise, go to Jupyter to start the exercise:

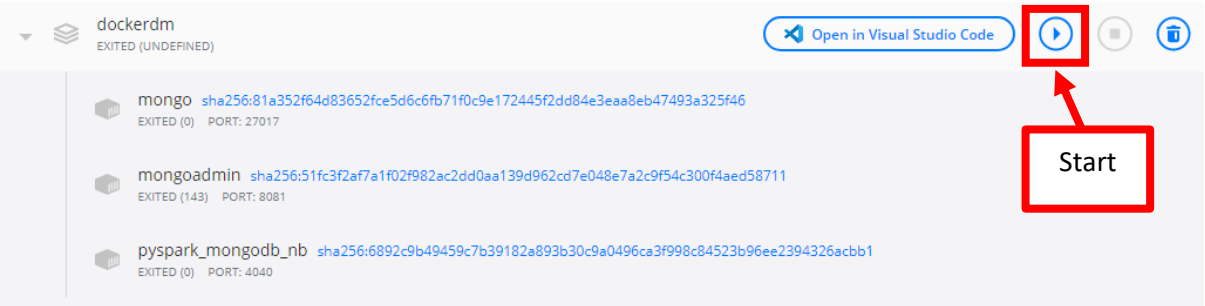
```
pyspark_mongodb_nb |     To access the notebook, open this file in a browser:
pyspark_mongodb_nb |         file:///home/jovyan/.local/share/jupyter/runtime/nbserver-15-open.html
pyspark_mongodb_nb |     Or copy and paste one of these URLs:
pyspark_mongodb_nb |         http://1d0485d0fd9b:8888/?token=4bc0ff234bf89bf68660b8f3f9d5fcf582568c0bc7667a53
pyspark_mongodb_nb |      or http://127.0.0.1:8888/?token=4bc0ff234bf89bf68660b8f3f9d5fcf582568c0bc7667a53
```

Stop and start the environment:

Once the environment is started, you will be able to see it in your docker desktop client. You can also start and stop it from there:



and



The address and token to access Jupyter between sessions stay the same. Thus, you can freely start and stop the containers at will.

Jupyter:

Replace the XXX fields by your own tokens in the notebook.

```
###################################
#        main program             #
###################################

# Twitter key and secret for OAuth
consumer_key = "XXX"
consumer_secret = "YYY"

access_token = "AAA"
access_token_secret = "BBB"

api = twitter.Api(consumer_key=consumer_key,
                  consumer_secret=consumer_secret,
                  access_token_key=access_token,
                  access_token_secret=access_token_secret)
```

Exercises:

Queries:

1) Count the number of tweets and users
2) Print out the name of all the users inserted
3) Find the most retweeted tweet
4) Find the shortest tweet
5) Count all the words used the tweets and find the top 5 most used