1. Set up your mongodb

1. If you are using linux or mac, download and install mongodb and open mongodb in your terminal. Download mongodb.

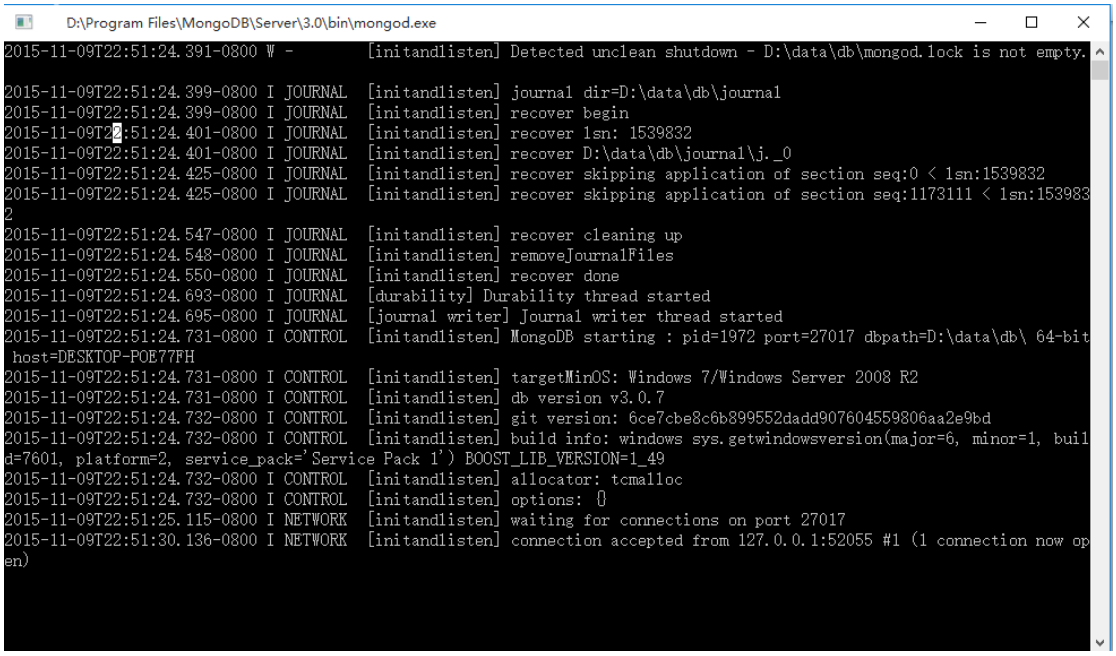
   https://www.mongodb.org/downloads#production

2. If you are using windows, open mongod at first

D:\Program Files\MongoDB\Server\3.0\bin

Open mongod, which means you open mongo database service and wait users to connect to mongo.

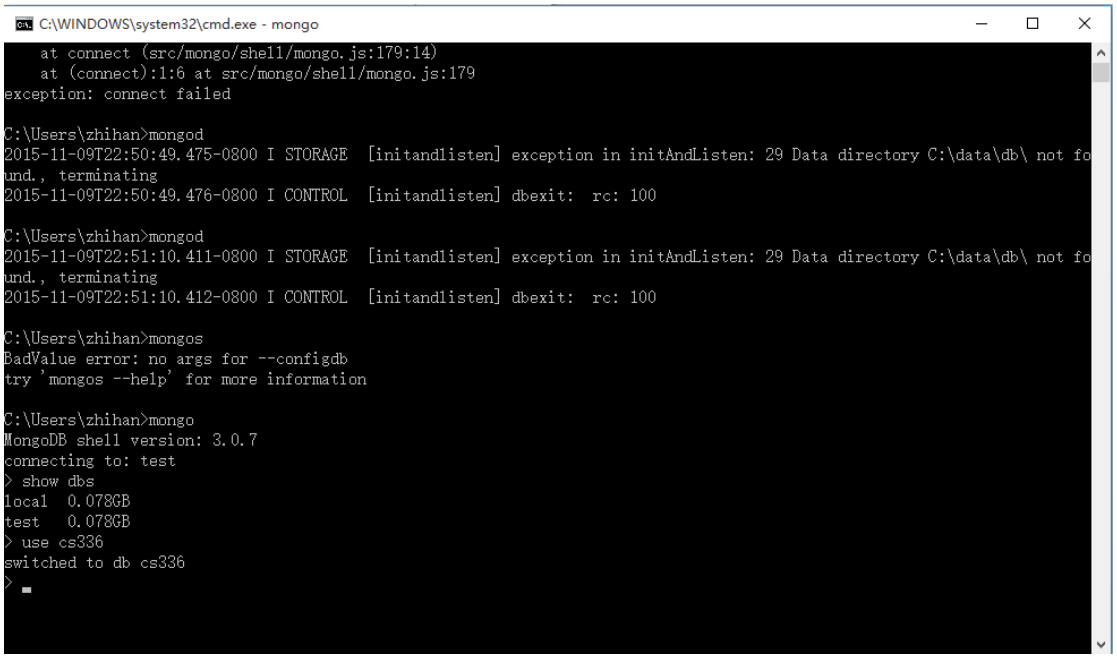
If you get a problem “mongodb data/db does not exist ”, you can create a folder “c:/data/db” (or “d:/data/db”, it depends on which disk drive the mongo is installed in)or mongod.exe --dbpath "your data path" to reset your data path.



3. Set mongo bin file in your computer environment path

   Find your mongo bin folder then add that after your environment path(right click on my computer then attribute and then advanced setting and then environment variable).

4. Then open cmd, in cmd use command “mongo”, now you are connected to mongodb.



5. Show all database instances in the database

   show dbs

6. Create a new database instance

   use cs336

   Mind that now you are under cs336 database instance.

7. Under the database instance, create two collections

   db.createCollection("unlabel_review")

   db.createCollection("unlabel_review_after_splitting")

   After that, if you use “show collections”, you can find three new collections have been created.

   You can check those two collections are all empty by using command find() or count(), find will return all contents in the collection and count will return how many documents in the collection(the number of documents).

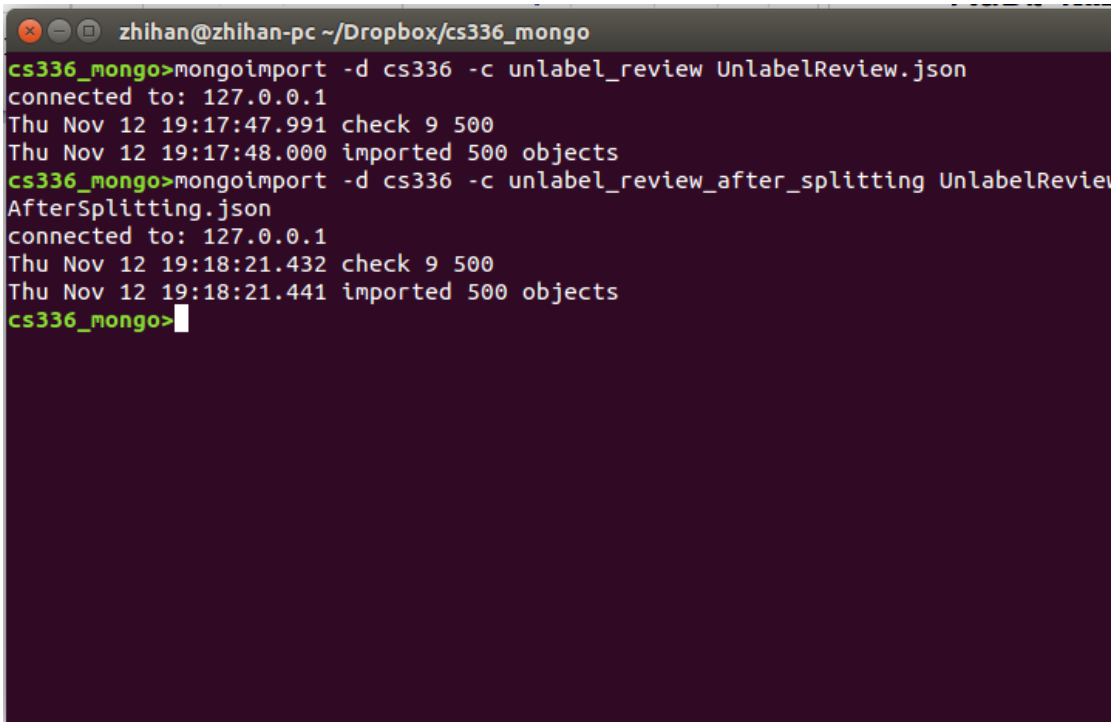
   using db.unlabel_review.count(), db.unlabel_review_after_splitting.count(), or replace “count” with “find”. The results are as follows:

8. Import json data in your database collections, use two json files to import the data in your database.

   Open a new cmd console and then use mongoimport command.

   mongoimport -d cs336 -c unlabel_review UnlabelReview.json

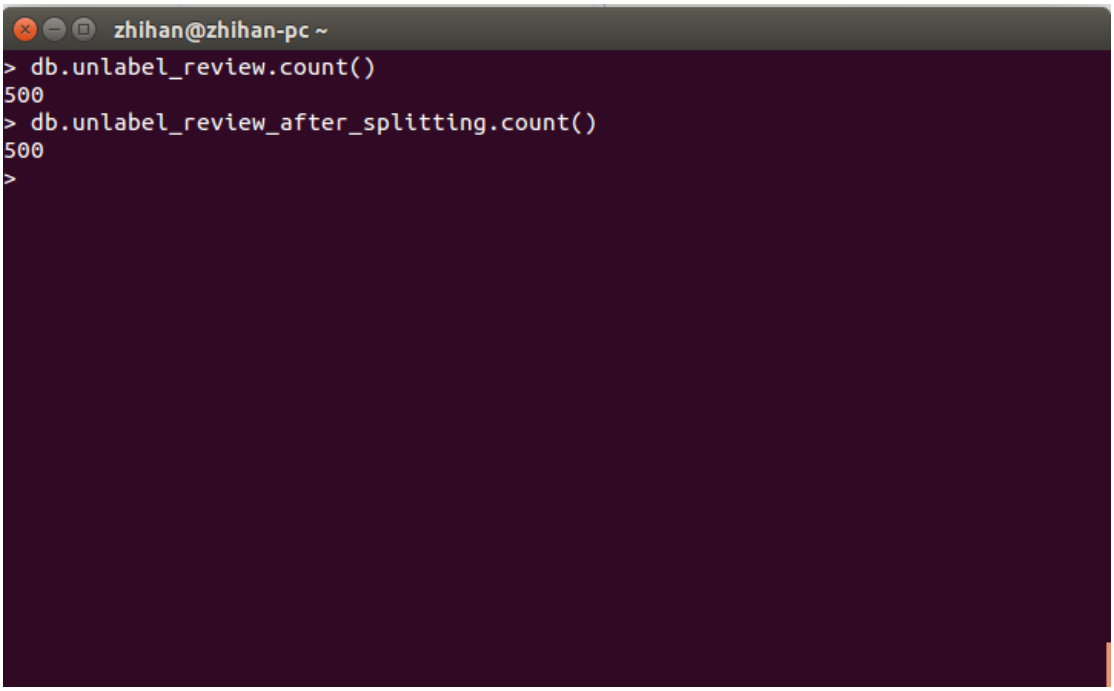
   mongoimport -d cs336 -c unlabel_review_after_splitting UnlabelReviewAfterSplitting.json



Then check data again.

db.unlabel_review.count()

db.unlabel_review_after_splitting.count()



We have import 500 documents in each unlabeled collection.

9. If want to check what it contain in each collection, you can use find() function.

10. Connect mongo from other programming language. Mongodb support js as default programming language. If you want to connect it from java, refer to following link.

    https://docs.mongodb.org/getting-started/java/

    other programming languages can be found in the following link.

    https://docs.mongodb.org/ecosystem/drivers/

unlabel_review is raw data. unlabel_review_after_splitting is processed data, containing words in reviews and its corresponding frequency(count).

Its format is as follows:


```
{
"_id" : ObjectId("5644077d0b539b0990979d13"),
"review" : [
{
"count" : 1,
"word" : "everyone"
},
{
"count" : 1,
"word" : "version"
},{
"count" : 2,
"word" : "entertaining"
},
{
"count" : 3,
"word" : "classic"
```

}],

"id" : "2381_9"

}

As you see, each review has been split to words and its frequency. Because the number of words may be different in each review, it is not a good way to store such a data structure in classic relational database system.

review: customers' reviews about a movie

id: the review's id

You do not need to concern about "_id", it's a special id automatically generated by mongodb for each document (record).