How to create "test" DB dump

Download the original Amazon reviews dataset from https://snap.stanford.edu/data/web-Amazon.html and load it in a MongoDB.

Otherwise clone it from an existing MongoDB instance (depending on YourUniversity policies a VPN and a SSH tunneling might be necessary):

```
$ sudo openfortivpn vpn.your_university.tld:443 --username ******** --password ********

$ ssh -fN -L 127117:localhost:21017 student**@descartes.departement.your_university.tld

$ sudo mongod    --config   /etc/mongodb.conf

$ mongodump      --host     127.0.0.1 --port     27117            \
                 --username ********  --password ********         \
                 --db       test      --archive           -j 8    \
  | mongorestore --host     127.0.0.1 --port     27017    -j 8
```

Then (inside the MongoDB shell) switch to "test" DB and since the dataset is really big, reduce its size by:

- dropping unneeded collections (like *restaurants*)
- cutting off from the "meta" and "reviews" collections the unneeded fields (*brand*, *price*, *related*, *sales rank*, *title*, but also *helpful*, *review text*, *review time*, *summary*, *unix review time*)
- removing the documents without a "description" field or with an empty one
- populating an array with the "asin" fields of all the documents in the "meta" collection
- removing from the "reviews" collection documents about items with an "asin" field not present in the above mentioned array

```
$ mongo
> use test
> db.restaurants.drop()
> db.meta.update({},
                 { $unset: { brand:     1,
                             price:     1,
                             related:   1,
                             salesRank: 1,
                             title:     1,
                           },
                  },
                 { multi:true })
```

Execution time: 10.224 sec
Updated documents: 106 474

```
> db.reviews.update({},
                    { $unset: { helpful:        1,
                                reviewText:     1,
                                reviewTime:     1,
                                summary:        1,
                                unixReviewTime: 1,
                              },
                     },
                    { multi:true })
```

Execution time: 19 min 21 sec
Updated documents: 23 831 908

```
> db.meta.deleteMany({description: {$exists: false}})
> db.meta.deleteMany({description: ''})
```

Execution time: 0.664 sec
Updated documents: 26 264

```
> var items = db.meta.find({},
                           {_id:0, asin:1}).map(
                                   function(d) {return d.asin})
> db.reviews.deleteMany({asin: {$not: {$in: items}}})
```

Execution time: 10 min 2 sec
Dropped documents: 21 867 827

```
> db.meta.aggregate([{
        $addFields: {
                categories: {
                        $reduce: {
                                input: "$categories",
                                initialValue: [],
                                in: { $concatArrays: [
                                        "$$value",
                                        { $cond: {
                                                if: {$isArray: "$$this"},
                                                then: "$$this",
                                                else: []
                                                  }
                                         }
                                      ]
                                     }
                                  }
                             }
                     }
        },
        {$out: "meta"},
])
```

Execution time: 2.529 sec

Finally create an index on the "reviewerID" field in the "reviews" collection and compact "meta" and "reviews" collections:

```
> db.reviews.createIndex({ reviewerID: 1 })
> db.runCommand ( { compact: "meta",    force: true } )
> db.runCommand ( { compact: "reviews", force: true } )
```

Execution time: 19.434 sec

Then restart MongoDB and create a dump of the shrinked "test" db:

```
$ mongodump --host 127.0.0.1 --port    27017                                \
            --db   test      --archive=shrinked_test_db_at_descartes.mongodump.gz    \
            --gzip            -j       8

$ ls -sh    shrinked_test_db_at_descartes.mongodump.gz

  86M shrinked_test_db_at_descartes.mongodump.gz
```