

# MongoDB 实习

薛飞跃 1700017831

## 一 . 环境配置

1. 下载安装包并解压, 将文件夹移动至指定位置, 并将其可执行文件添加到PATH中

```
wget https://fastdl.mongodb.org/linux/mongodb-linux-x86_64-3.6.3.tgz
tar -zxvf mongodb-linux-x86_64-3.6.3.tgz
sudo mv mongodb-linux-x86_64-3.6.3/ /usr/local/mongodb
sudo chmod -R 777 !$
export PATH=/usr/local/mongodb/bin:$PATH
```

2. 创建数据库目录, 并启动MongoDB服务。注意: 如果数据库目录不是/data/db, 可以通过 --dbpath 来指定。

```
sudo mkdir -p /data/db
sudo chmod -R 777 !$
cd /usr/local/mongodb/bin/
./mongod
```

此时新开一个窗口

3. 连接MongoDB, 进入Shell模式

```
cd /usr/local/mongodb/bin/
./mongo
```

显示如下信息表示连接成功:

```
MongoDB shell version v3.6.3
connecting to: mongodb://127.0.0.1:27017
MongoDB server version: 3.6.3
welcome to the MongoDB shell.
...
...
>
```

## 二 . python 安装 pymongo

使用 pip 下载即可

## 三 . 导入数据

连接到数据库后, 选择相应的集合, 然后读取 json 并插入数据库

```

import json
import os
import pymongo
from pymongo import MongoClient
client = MongoClient('localhost',27017)

db = client.test

covid = db.covid
for file in os.listdir('CORD-19-research'):
    with open(os.path.join('CORD-19-research',file),'r') as f:
        dic=json.load(f)
        covid.insert_one(dic)

```

#### 四．查询任务

1. json 中使用“.”号来获取 json 中嵌入对象的值,如下查询 metadata 中 authors 中有三项的文档，在 find\_one() 函数中，第二项代表返回的项，设置“metadata.authors”为 1，代表只返回这一项。

```

result=covid.find_one({"metadata.authors" : {"$size": 3}},{"metadata.authors":1})

In [19]: runfile('/home/xuefeiyue/文档/mongodb.py', wdir='/home/xuefeiyue/文档')
{'_id': ObjectId('5ed4ef6b29b3e6464e764ccd'), u'metadata': {u'authors': [{u'last': u'Lauro', u'suffix': u'',
u'affiliation': {u'laboratory': u'', u'institution': u'University of Sussex', u'location': {u'country':
u'UK', u'settlement': u'Falmer, Brighton', u'postCode': u'BN1 9QH'}}}, u'middle': [u'Di'], u'email': u'',
u'first': u'Francesco'}, {u'last': u'Kiss', u'suffix': u'', u'affiliation': {u'laboratory': u'',
u'institution': u'University of Sussex', u'location': {u'country': u'UK', u'settlement': u'Falmer, Brighton',
u'postCode': u'BN1 9QH'}}}, u'middle': [u'Z'], u'email': u'', u'first': u'Istv\xe1n'}, {u'last': u'Miller',
u'suffix': u'', u'affiliation': {u'laboratory': u'', u'institution': u'La Trobe University', u'location':
{u'country': u'Australia', u'settlement': u'Bundoora'}}}, u'middle': [u'C'], u'email': u'', u'first':
u'Joel']}]}
```

2. 使用 count()函数查询有三位作者的文档的总数。结果中有 warning, 于是尝试使用最新的 count\_documents()函数代替

```

result=covid.find({"metadata.authors" : {"$size":3}}).count()

In [49]: runfile('/home/xuefeiyue/文档/mongodb.py', wdir='/home/xuefeiyue/文档')
1
/home/xuefeiyue/文档/mongodb.py:24: DeprecationWarning: count is deprecated. Use Collection.count_documents
instead.
    result=covid.find({"metadata.authors" : {"$size": 3}}).count()

result=covid.count_documents({"metadata.authors" : {"$size": 3}})
```

3. 尝试查询作者数少于 3 的文章总数，发现报错：

```
result=covid.find({"metadata.authors" : {"$size": {"$lt",3}}},{"metadata.authors":1}).count()
```

**InvalidDocument: cannot encode object: set([3, '\$lt']), of type: <type 'set'>**

经过查询手册，发现不能\$size 不能匹配一定范围内的元素，非要查询，必须使用一个字段来保存元素的数量。

You cannot use \$size to find a range of sizes (for example: arrays with more than 1 element). If you need to query for a range, create an extra size field that you increment when you add elements.

4. 查询某一个作者参与的文档数：

```
result=covid.count_documents({"metadata.authors" :{
    "first": "Jiajie",
    "middle": [],
    "last": "Qian",
    "suffix": "",
    "affiliation": {},
    "email": ""
}})
print(result)
```

这里 metadata.authors 是一个列表，里面有多个作者，本来应该使用\$elemMatch 来进行匹配，但尝试直接匹配发现，版本支持直接查询。

```
In [53]: runfile('/home/xuefeiyue/文档/mongodb.py', wdir='/home/xuefeiyue/文档')
2
```

## 五．心得与总结

1. MongoDB 是由数据库-文档-集合-字段来组织的。
2. MongoDB 的键区分大小写。
3. MongoDB 单个查询速度非常快。
4. pymongo 在实际中有一些细节需要注意，应该经常查阅手册。