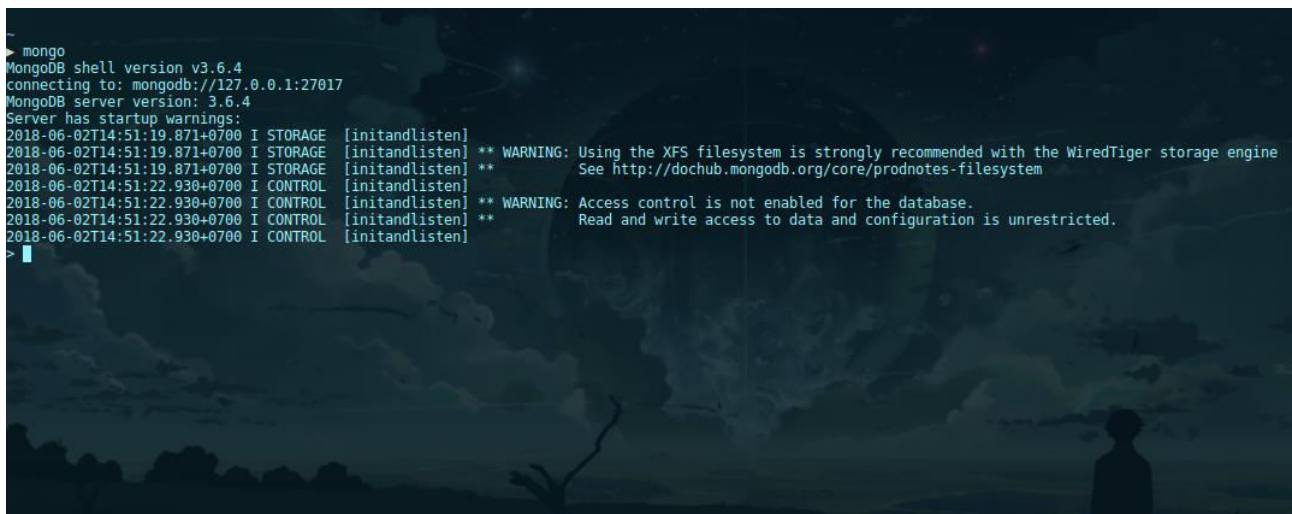


Tut 3: Hướng dẫn xuất dữ liệu từ Mongodb và đánh label cho các dữ liệu

- Sau khi các bạn crawl dữ liệu từ trang web tin tức mà bạn chọn, dữ liệu được lưu ở trong mongodb nên cần phải được xuất ra file để tiến hành đọc và phân loại bài báo.
- Nếu bạn đã crawl thành công và export ra được file JSON rồi thì nên xem code phân loại ở bên dưới.
- Trước tiên cần kiểm tra xem dữ liệu đã được lưu trong máy hay chưa.
- ✓ Các bạn mở terminal/cmd rồi chạy lệnh sau:

mongo

- ✓ Lệnh sẽ mở mongo shell



```
mongo
MongoDB shell version v3.6.4
connecting to: mongodb://127.0.0.1:27017
MongoDB server version: 3.6.4
Server has startup warnings:
2018-06-02T14:51:19.871+0700 I STORAGE [initandlisten]
2018-06-02T14:51:19.871+0700 I STORAGE [initandlisten] ** WARNING: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine
2018-06-02T14:51:19.871+0700 I STORAGE [initandlisten] See http://dochub.mongodb.org/core/prodnotes-filesystem
2018-06-02T14:51:22.930+0700 I CONTROL [initandlisten]
2018-06-02T14:51:22.930+0700 I CONTROL [initandlisten] ** WARNING: Access control is not enabled for the database.
2018-06-02T14:51:22.930+0700 I CONTROL [initandlisten] Read and write access to data and configuration is unrestricted.
2018-06-02T14:51:22.930+0700 I CONTROL [initandlisten]
```

- ✓ Tiếp tục nhập lệnh sau:

use locatfamily
db.articles.count()

- ✓ Nếu thành công, nó sẽ xuất ra 1 số lớn hơn 0, số đó chính là số bài báo đã crawl được.

```

> mongo
MongoDB shell version v3.6.4
connecting to: mongodb://127.0.0.1:27017
MongoDB server version: 3.6.4
Server has startup warnings:
2018-06-02T14:51:19.871+0700 I STORAGE [initandlisten]
2018-06-02T14:51:19.871+0700 I STORAGE [initandlisten]
2018-06-02T14:51:19.871+0700 I STORAGE [initandlisten]
2018-06-02T14:51:22.930+0700 I CONTROL [initandlisten]
2018-06-02T14:51:22.930+0700 I CONTROL [initandlisten]
2018-06-02T14:51:22.930+0700 I CONTROL [initandlisten]
2018-06-02T14:51:22.930+0700 I CONTROL [initandlisten]
> use locatefamily
switched to db locatefamily
> db.articles.count()
10579
>

```

- ✓ Sau đó thoát ra ngoài terminal chính bằng lệnh:

exit

- Rồi chạy lệnh sau để export data từ MongoDB:

mongoexport --db locatefamily --collection articles --out articles.json --jsonArray

- Lệnh này sẽ xuất ra 1 file tên là articles.json tại đường dẫn hiện tại của terminal. Nó chứa toàn bộ dữ liệu của các bài báo mà chúng ta đã crawl về.

```

Documents/python/scrapper
> mongoexport --db locatefamily --collection articles --out articles.json --jsonArray
2018-06-02T17:05:17.482+0700 connected to: localhost
2018-06-02T17:05:18.467+0700 [.....] locatefamily.articles 0/10579 (0.0%)
2018-06-02T17:05:19.467+0700 [#####] locatefamily.articles 8000/10579 (75.6%)
2018-06-02T17:05:19.915+0700 [#####] locatefamily.articles 10579/10579 (100.0%)
2018-06-02T17:05:19.915+0700 exported 10579 records

```

- Sau khi đã có dữ liệu JS, chúng ta cần phải phân loại dữ liệu trong đó ra thành từng category và đánh label cho từng bài báo dựa vào thuộc tính "type". Để làm điều này thì có nhiều cách, sau đây mình chỉ hướng dẫn cách đọc file JSON rồi ghi vào file .txt, mỗi file .txt tương ứng với một bài báo và các file này sẽ được phân loại theo "type" và lưu vào folder tương ứng.
- Ví dụ ta có 2 bài báo gồm có các trường như sau:

```

{
  _id: ...,
  title: abc,
  content: xyz,
  type: a
}

```

```
{  
  _id: ...,  
  title: a,  
  content: x,  
  type: b  
}
```

- Mình sẽ lưu bài báo thứ nhất có type là a vào thư mục là "a" và đặt tên file là 0.txt (số 0 ở đây là số thứ tự của bài báo trong folder, tên này giúp cho việc đọc file từ sklearn dễ dàng hơn). Còn bài báo thứ 2 có type là "b" thì cũng sẽ lưu ở trong thư mục "b" với tên là 0.txt.
- Các bạn có thể dựa vào ý tưởng trên hoặc có thể sử dụng cách khác, mình cung cấp sẵn cho các bạn hàm để đọc dữ liệu từ file json, việc xử lý như thế nào thì phụ thuộc vào các bạn.
- Lưu ý là khi chuyển bài báo từ dạng JSON sang text thì cần phải loại bỏ những syntax có trong file json như các dấu { },: và bỏ luôn các key như _id, title, content... Phần cần giữ lại là value của các trường title, summary, content (tùy bạn chọn sao cho hợp lý). Ví dụ trong bài báo thứ nhất ở trên thì file a/0.txt có nội dung như sau:

Abc
xyz

Code:

```
from __future__ import print_function  
import json  
  
# Remember to check the path to articles.json relative to this file before executing  
with open('articles.json') as json_data:  
    # Load JSON  
    articles = json.load(json_data)  
    print(len(articles), "Articles loaded succesfully")  
    # Loop through every article in the json file  
    for article in articles:  
        # Your code lies here :)  
        pass
```