

Tut 1: Hướng dẫn sử dụng Scrapper

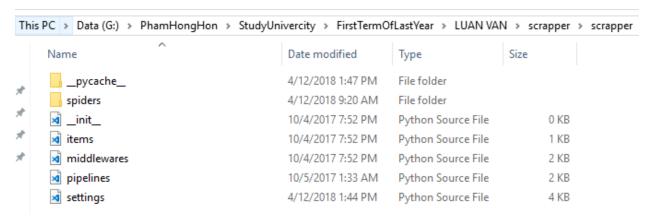
- Tác giả: Mai Đức Trung.
- Nhóm nghiên cứu về AI của Trường Đại học Bách Khoa Thành phố Hồ Chí Minh.

1. Cài đặt môi trường chạy:

- a. MongoDB https://www.mongodb.com/
- b. Python 3 https://www.python.org/
- c. Thu viện Scrapy trong Python 3 https://doc.scrapy.org/en/latest/intro/install.html

2. Hướng dẫn chay:

- B1: Chạy MongoDB (ở dưới là hướng dẫn)
 https://o7planning.org/vi/10265/huong-dan-cai-dat-va-cau-hinh-database-mongodb
- B2: Mở thư mục như trong hình dưới:
 Và mở CMD python tại thư mục này.



B3: chạy lệnh: scrapy crawl mogi
 Trong đó mogi là tên được đặt trong file mogi.py ở trong thư mục spiders.



```
This PC > Data (G:) > PhamHongHon > StudyUnivercity > FirstTermOfLastYear > LUAN VAN > scrapper > scrapper > spiders
                                       Date modified
    Name
                                                        Type
                                                                         Size
    __pycache__
                                      4/12/2018 4:23 PM File folder
                                      10/4/2017 7:52 PM Python Source File
    🖈 _init_
                                                                               1 KB
    a batdongsan
                                      12/7/2017 8:08 AM Python Source File
                                                                               4 KB
    a batdongsanTintuc
                                        11/15/2017 10:46
                                                         Python Source File
                                                                                3 KB
                                        5/6/2018 10:04 AM Python Source File
                                                                               3 KB
   🗖 mogi
     nhandan 🖈
                                        10/8/2017 2:52 AM
                                                         Python Source File
                                                                                2 KB
    4/12/2018 4:22 PM Python Source File
                                                                               31 KB
    d thanhnien
                                        10/4/2017 7:52 PM Python Source File
                                                                               3 KB
    d tuoitre
                                        10/4/2017 7:52 PM Python Source File
                                                                               4 KB
                                        10/5/2017 11:17 AM Python Source File
     vnexpress
                                                                                2 KB
    zing
                                        10/4/2017 7:52 PM Python Source File
                mogi.py
             import scrapy
            from random import randint
            class BatdongsanTintucSpider(scrapy.Spider):
             name = "mogi"
                  BASE_URL = 'https://mogi.vn/'
                count = 51
                 def start_requests(self):
                       url = self.BASE_URL
yield scrapy.Request(url=url, callback=self.parse)
```

- Dữ liệu sẽ được lưu trong MongoDB sau khi chạy.
- Chỉ cần sửa những file như mogi.py thành cái của mình, đổi tên là được

3. Các tài liêu tham khảo:

https://doc.scrapy.org/en/latest/intro/tutorial.html