**Tạo 1 project scrapy:**

Scrapy startproject nameproject

**Tạo 1 spider cơ bản tên basic, urls là linkweb:**

Scrapy genspider basic linkweb

**Liệt kê các cấu hình khác:**

Scrapy genspider –l

**Chạy spider tên basic:**

Scrapy crawl basic

**Tùy chọn cũng chạy 1 spider, cho kết quả tương tự trên:**

scrapy parse --spider=basic link-web

**Thêm items LazadaItem trong module lazada.items(kiểu đường dẫn):**

From lazada.items import LazadaItem

**Lưu dữ liệu ra file:**

Scrapy crawl basic –o items.json (items.jl, items.csv, items.xml)

**Lấy link từ file:**

( file urls.txt có nội dung:

http://web:9312/properties/property\_000000.html

http://web:9312/properties/property\_000001.html

http://web:9312/properties/property\_000002.html  
)

urls = [i.strip() for i in open('urls.txt').readlines()]

**Copy nội dung file này sang file khác:**

cp properties/spiders/basic.py properties/spiders/manual.py

**Cài apperyio để kết nối với csdl được tạo trên appery.io, khi crawl sẽ đẩy dữ liệu lên:**

Pip install scrapyapperyio

**Thêm vào file settings.py như sau:**

ITEM\_PIPELINES = {'scrapyapperyio.ApperyIoPipeline': 300}

APPERYIO\_DB\_ID = '<<Your API KEY here>>' (lấy trong mục setting trên appery.io)

APPERYIO\_USERNAME = 'root' (tạo trong mục Users)

APPERYIO\_PASSWORD = 'pass' (cũng tạo trong mục Users cùng pass)

APPERYIO\_COLLECTION\_NAME = ‘lazada’ (tên Collection tạo trên )

**Cho spider chạy đến 1 số lần thì dừng: (chạy nó sẽ tự động đẩy lên, có thể lưu ra file như trước)**

Scrapy crawl mediate –s CLOSESPIDER\_ITEMCOUNT=50

**Truyền tham số bên ngoài hàm:**

<https://github.com/scrapy/scrapy/issues/1138>

def parse\_page1(self, response):

item = MyItem(main\_url=response.url)

request = scrapy.Request("http://www.example.com/some\_page.html",

callback=self.parse\_page2)

request.meta['item'] = item

return request

def parse\_page2(self, response):

item = response.meta['item']

item['other\_url'] = response.url

return item