#!/bin/bash

# Scrapy Tutorial Terminal commands

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

# For Google Docs optimization, use [Tools] > [Preferences...] menu, and disable all autocorrect options.

#rm -rf tutorial

scrapy startproject tutorial

cd tutorial

scrapy genspider tutorial quotes.toscrape.com

scrapy crawl tutorial

touch tutorial/spiders/quotes.py

### start quotes.py

import scrapy

class QuotesSpider(scrapy.Spider):  
 name = "quotes"  
 start\_urls = [  
 'http://quotes.toscrape.com/page/1/',  
 'http://quotes.toscrape.com/page/2/',  
 ]  
  
 def parse(self, response):  
 page = response.url.split("/")[-2]  
 filename = 'quotes-%s.html' % page  
 with open(filename, 'wb') as f:  
 f.write(response.body)

### end quotes.py

scrapy crawl quotes

scrapy shell 'http://quotes.toscrape.com/page/1/'

### enter scrapy shell

response.css('title')

response.css('title::text').extract()

response.css('title').extract()

response.css('title::text').extract\_first()

response.css('title::text')[0].extract()

response.css("div.quote")

quote = response.css("div.quote")[0]

title = quote.css("span.text::text").extract\_first()

author = quote.css("small.author::text").extract\_first()

tags = quote.css("div.tags a.tag::text").extract()

exit

### exit scrapy shell

touch tutorial/spiders/quotesjson.py

### start quotesjson.py

import scrapy  
  
class QuotesjsonSpider(scrapy.Spider):  
 name = "quotesjson"  
 start\_urls = [  
 'http://quotes.toscrape.com/page/1/',  
 'http://quotes.toscrape.com/page/2/',  
 ]  
  
 def parse(self, response):  
 for quote in response.css('div.quote'):  
 yield {  
 'text': quote.css('span.text::text').extract\_first(),  
 'author': quote.css('small.author::text').extract\_first(),  
 'tags': quote.css('div.tags a.tag::text').extract(),  
 }

### end quotesjson.py

#rm quotes.json

scrapy crawl quotesjson -o quotes.json

cat quotes.json