爬虫案例6(随机代理池+随机usragent+时间间隔)

https://www.cnblogs.com/cnkai/p/7401343.html

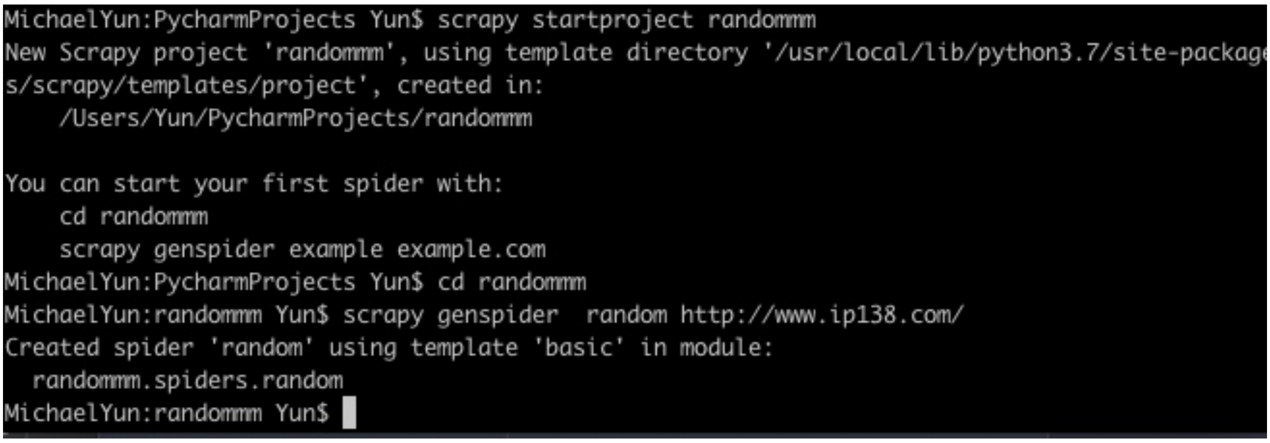
https://www.cnblogs.com/cnkai/p/7401526.html

https://blog.csdn.net/mouday/article/details/81512748 随机时间间隔

先固定间隔 在setting中 设置
`DOWNLOAD_DELAY=3`

————————

1.创建项目



```
MichaelYun:PycharmProjects Yun$ scrapy startproject randommm
New Scrapy project 'randommm', using template directory '/usr/local/lib/python3.7/site-package
s/scrapy/templates/project', created in:
    /Users/Yun/PycharmProjects/randommm

You can start your first spider with:
    cd randommm
    scrapy genspider example example.com
MichaelYun:PycharmProjects Yun$ cd randommm
MichaelYun:randommm Yun$ scrapy genspider  random http://www.ip138.com/
Created spider 'random' using template 'basic' in module:
  randommm.spiders.random
MichaelYun:randommm Yun$ 
```

2.在setting.py 中设置user_agent 和 代理ip池

注意:以后的思路可以自己爬，读取json或者自己爬的json文件

```
MY_USER_AGENT = [
    "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; AcooBrowser; .NET CLR 1.1.4322; .NET CLR 2.0.50727)",
    "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; Acoo Browser; SLCC1; .NET CLR 2.0.50727; Media Center PC 5.0; .NET CLR 3.0.04506)",
    "Mozilla/4.0 (compatible; MSIE 7.0; AOL 9.5; AOLBuild 4337.35; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)",
    "Mozilla/5.0 (Windows; U; MSIE 9.0; Windows NT 9.0; en-US)",
    "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Win64; x64; Trident/5.0; .NET CLR 3.5.30729; .NET CLR 3.0.30729; .NET CLR 2.0.50727; Media Center PC 6.0)",
    "Mozilla/5.0 (compatible; MSIE 8.0; Windows NT 6.0; Trident/4.0; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; .NET CLR 1.0.3705; .NET CLR 1.1.4322)",
    "Mozilla/4.0 (compatible; MSIE 7.0b; Windows NT 5.2; .NET CLR 1.1.4322; .NET CLR 2.0.50727; InfoPath.2; .NET CLR 3.0.04506.30)",
    "Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN) AppleWebKit/523.15 (KHTML, like Gecko,
Safari/419.3) Arora/0.3 (Change: 287 c9dfb30)",
    "Mozilla/5.0 (X11; U; Linux; en-US) AppleWebKit/527+ (KHTML, like Gecko, Safari/419.3) Arora/0.6",
    "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.2pre) Gecko/20070215 K-Ninja/2.1.1",
    "Mozilla/5.0 (Windows; U; Windows NT 5.1; zh-CN; rv:1.9) Gecko/20080705 Firefox/3.0 Kapiko/3.0",
    "Mozilla/5.0 (X11; Linux i686; U;) Gecko/20070322 Kazehakase/0.4.5",
    "Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.9.0.8) Gecko Fedora/1.9.0.8-1.fc10 Kazehakase/0.5.6",
    "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/535.11 (KHTML, like Gecko) Chrome/17.0.963.56 Safari/535.11",
    "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_7_3) AppleWebKit/535.20 (KHTML, like Gecko) Chrome/19.0.1036.7 Safari/535.20",
    "Opera/9.80 (Macintosh; Intel Mac OS X 10.6.8; U; fr) Presto/2.9.168 Version/11.52",
    "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/536.11 (KHTML, like Gecko) Chrome/20.0.1132.11 TaoBrowser/2.0 Safari/536.11",
    "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.71 Safari/537.1 LBBROWSER",
    "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; WOW64; Trident/5.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0; .NET4.0C; .NET4.0E; LBBROWSER)",
    "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; QQDownload 732; .NET4.0C; .NET4.0E; LBBROWSER)",
    "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/535.11 (KHTML, like Gecko) Chrome/17.0.963.84 Safari/535.11 LBBROWSER",
    "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.1; WOW64; Trident/5.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0; .NET4.0C; .NET4.0E)",
    "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; WOW64; Trident/5.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0; .NET4.0C; .NET4.0E; QQBrowser/7.0.3698.400)",
    "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; QQDownload 732; .NET4.0C; .NET4.0E)",
    "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0; SV1; QQDownload 732; .NET4.0C; .NET4.0E; 360SE)",
    "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; QQDownload 732; .NET4.0C; .NET4.0E)",
    "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.1; WOW64; Trident/5.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0; .NET4.0C; .NET4.0E)",
    "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.89 Safari/537.1",
    "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.89 Safari/537.1",
    "Mozilla/5.0 (iPad; U; CPU OS 4_2_1 like Mac OS X; zh-cn) AppleWebKit/533.17.9 (KHTML, like Gecko) Version/5.0.2 Mobile/8C148 Safari/6533.18.5",
    "Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:2.0b13pre) Gecko/20110307 Firefox/4.0b13pre",
    "Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:16.0) Gecko/20100101 Firefox/16.0",
    "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.11 (KHTML, like Gecko) Chrome/23.0.1271.64 Safari/537.11",
    "Mozilla/5.0 (X11; U; Linux x86_64; zh-CN; rv:1.9.2.10) Gecko/20100922 Ubuntu/10.10 (maverick) Firefox/3.6.10",
    "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36",
    ]


PROXIES = [
    'http://218.2.80.59:8080',
    'http://60.208.44.228:80',
    'http://218.202.219.82:80',
    'http://139.180.220.37:2333'
]
```

3.编写自定义中间件，随机

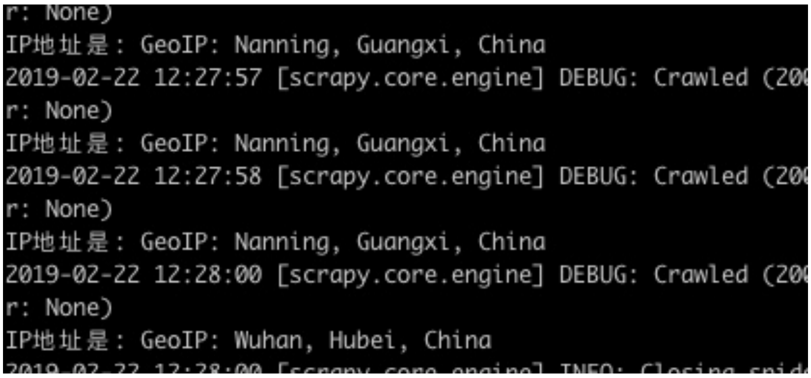
```

from scrapy.downloadermiddlewares.useragent import UserAgentMiddleware
import random


#设置随机代理
class MyUserAgentMiddleware(UserAgentMiddleware):


    def __init__(self, user_agent):
        self.user_agent = user_agent

    @classmethod
    def from_crawler(cls, crawler):
        return cls(
            user_agent=crawler.settings.get('MY_USER_AGENT')
        )

    def process_request(self, request, spider):
        agent = random.choice(self.user_agent)
        request.headers['User-Agent'] = agent


#设置随机ip
class ProxyMiddleware(object):

    def __init__(self, ip):
        self.ip = ip

    @classmethod
    def from_crawler(cls, crawler):
        return cls(ip=crawler.settings.get('PROXIES'))

    def process_request(self, request, spider):
        ip = random.choice(self.ip)
        request.meta['proxy'] = ip
```

4.添加配置文件中

```
DOWNLOAD_DELAY=3

DOWNLOADER_MIDDLEWARES = {
    'scrapy.downloadermiddleware.useragent.UserAgentMiddleware': None,
    'randommm.middlewares.MyUserAgentMiddleware': 400,
    'randommm.middlewares.ProxyMiddleware': 401,
}
```

5.逻辑代码参考

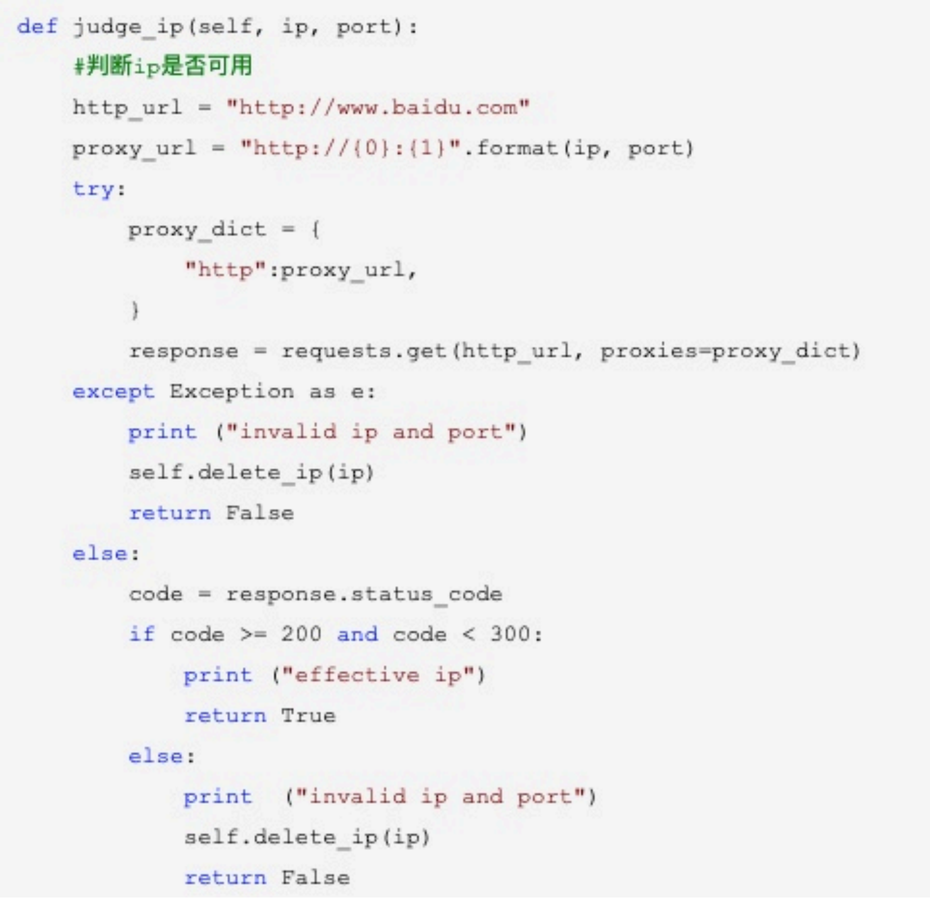
```python
# -*- coding: utf-8 -*-
import scrapy



class RandomSpider(scrapy.Spider):
    name = 'random'
    allowed_domains = []

    def start_requests(self):

        url = 'https://www.ip.cn/'

        for i in range(4):
            yield scrapy.Request(url=url, callback=self.parse, dont_filter=True)

    def parse(self,response):
        print('IP地址是:',response.css('#result > div > p:nth-child(3)::text').extract_first())

```



```
r: None)
IP地址是: GeoIP: Nanning, Guangxi, China
2019-02-22 12:27:57 [scrapy.core.engine] DEBUG: Crawled (20
r: None)
IP地址是: GeoIP: Nanning, Guangxi, China
2019-02-22 12:27:58 [scrapy.core.engine] DEBUG: Crawled (20
r: None)
IP地址是: GeoIP: Nanning, Guangxi, China
2019-02-22 12:28:00 [scrapy.core.engine] DEBUG: Crawled (20
r: None)
IP地址是: GeoIP: Wuhan, Hubei, China
2019-02-22 12:28:00 [scrapy.core.engine] INFO: Closing spide
```

可以看到有随机切换IP

注意参考:



```
def judge_ip(self, ip, port):
    #判断ip是否可用
    http_url = "http://www.baidu.com"
    proxy_url = "http://{0}:{1}".format(ip, port)
    try:
        proxy_dict = {
            "http":proxy_url,
        }
        response = requests.get(http_url, proxies=proxy_dict)
    except Exception as e:
        print ("invalid ip and port")
        self.delete_ip(ip)
        return False
    else:
        code = response.status_code
        if code >= 200 and code < 300:
            print ("effective ip")
            return True
        else:
            print  ("invalid ip and port")
            self.delete_ip(ip)
            return False
```