python 编写的抓京东商品的爬虫

闲着没事尝试抓一下京东的数据，需要使用到的库有：BeautifulSoup，urllib2，在 Python2 下测试通过 from creepy import Crawler from BeautifulSou...

闲着没事尝试抓一下京东的数据，需要使用到的库有：BeautifulSoup，urllib2，在 Python2 下测试通过

```
from creepy import Crawler
from BeautifulSoup import BeautifulSoup
import urllib2
import json

class MyCrawler(Crawler):
    def process_document(self, doc):
        if doc.status == 200:
            print '[%d] %s' % (doc.status, doc.url)
            try:
                soup = BeautifulSoup(doc.text.decode('gb18030').encode('utf-8'))
            except Exception as e:
                print e
                soup = BeautifulSoup(doc.text)
            print soup.find(id="product-intro").div.h1.text
            url_id=urllib2.unquote(doc.url).decode('utf8').split('/')[-1].split('.')[0]
            f = urllib2.urlopen('http://p.3.cn/prices/get?skuid=J_'+url_id,timeout=5)
            price=json.loads(f.read())
            f.close()
            print price[0]['p']
        else:
            pass

crawler = MyCrawler()
crawler.set_follow_mode(Crawler.F_SAME_HOST)
crawler.set_concurrency_level(16)
crawler.add_url_filter('\.(jpg|jpeg|gif|png|js|css|swf)$')
crawler.crawl('http://item.jd.com/982040.html')
```

部分运行结果：

```
[200] http://item.jd.com/519836.html
三星 HM1200 原装蓝牙耳机 黑色
118.00
[200] http://item.jd.com/603133.html
```

捷波朗 EASYVOICE+ 易音 蓝牙耳机 黑色
-1
[200] http://item.jd.com/1030552473.html
【年终热卖】嘉源手机 N699 2.8 寸双屏双卡双待双电 2000 毫安翻盖手机 黑
268.00