PYTHON
SCRAPY 爬虫项目

编译环境：

PyCharm 2019.2.4 (Community Edition)

Build #PC-192.7142.42, built on November 1, 2019

Runtime version: 11.0.4+10-b304.77 amd64

VM: OpenJDK 64-Bit Server VM by JetBrains s.r.o

Windows 10 10.0

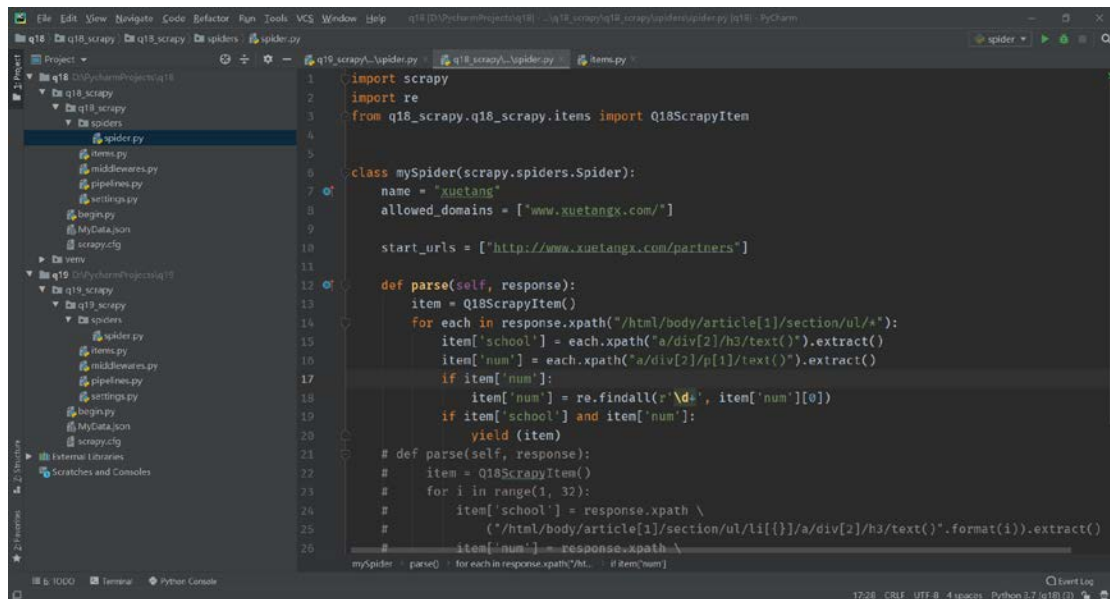GC: ParNew, ConcurrentMarkSweep

Memory: 1963M

Cores: 8

Registry:

Non-Bundled Plugins:

python 版本：3.7（Anaconda3）

scrapy 版本：1.8

1、

Spider.py

```python
import scrapy
import re
from q18_scrapy.q18_scrapy.items import Q18ScrapyItem


class mySpider(scrapy.spiders.Spider):
    name = "xuetang"
    allowed_domains = ["www.xuetangx.com/"]

    start_urls = ["http://www.xuetangx.com/partners"]

    def parse(self, response):
        item = Q18ScrapyItem()
        for each in response.xpath("/html/body/article[1]/section/ul/*"):
            item['school'] = each.xpath("a/div[2]/h3/text()").extract()
            item['num'] = each.xpath("a/div[2]/p[1]/text()").extract()
            if item['num']:
                item['num'] = re.findall(r'\d+', item['num'][0])
            if item['school'] and item['num']:
                yield (item)
```

Items.py
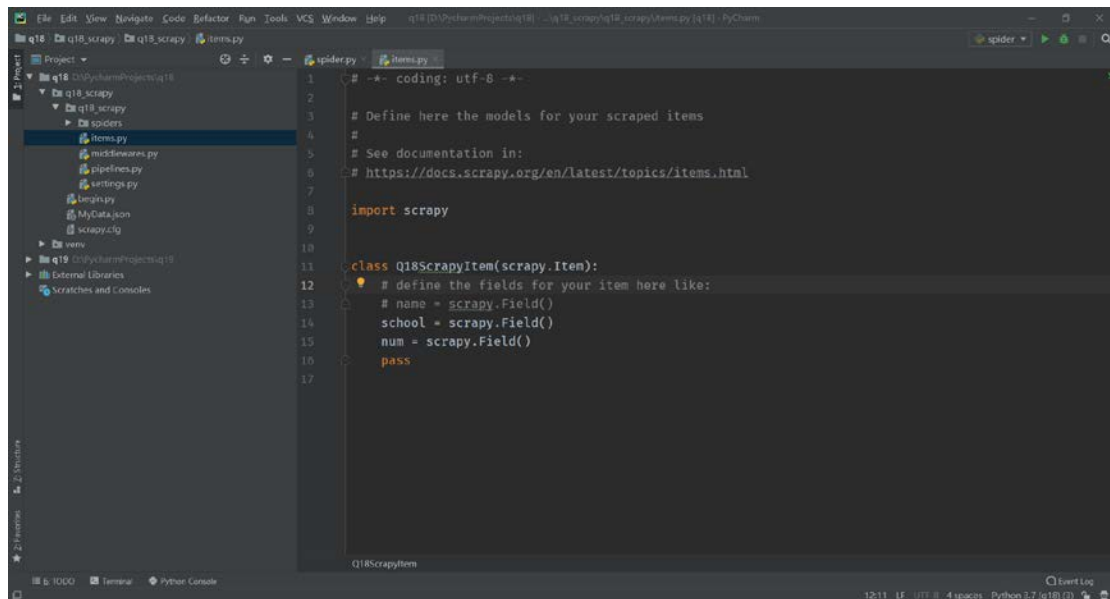
school 为学校名称，num 为课程数量

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

# Define here the models for your scraped items
#
# See documentation in:
# https://docs.scrapy.org/en/latest/topics/items.html

import scrapy


class Q18ScrapyItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    school = scrapy.Field()
    num = scrapy.Field()
    pass
```

运行：

结果：

{"school": ["清华大学"], "num": ["152"]}
{"school": ["台湾清华大学"], "num": ["1"]}
{"school": ["台湾交通大学"], "num": ["1"]}
{"school": ["斯坦福大学"], "num": ["1"]}
{"school": ["中央社会主义学院"], "num": ["0"]}
{"school": ["西安交通大学"], "num": ["24"]}
{"school": ["中南财经政法大学"], "num": ["7"]}
{"school": ["复旦大学"], "num": ["1"]}
{"school": ["北京理工大学"], "num": ["5"]}
{"school": ["中国科学技术大学"], "num": ["0"]}
{"school": ["吉林大学"], "num": ["5"]}
{"school": ["中国农业大学"], "num": ["4"]}

{"school": ["华南理工大学"], "num": ["18"]}
{"school": ["哈尔滨工业大学"], "num": ["14"]}
{"school": ["山东大学"], "num": ["5"]}
{"school": ["苏州大学"], "num": ["1"]}
{"school": ["东南大学"], "num": ["4"]}
{"school": ["台湾云林科技大学"], "num": ["0"]}
{"school": ["武汉大学"], "num": ["2"]}
{"school": ["华中科技大学"], "num": ["1"]}
{"school": ["同济大学"], "num": ["5"]}
{"school": ["南开大学"], "num": ["1"]}
{"school": ["北京工业大学"], "num": ["2"]}
{"school": ["南京理工大学"], "num": ["2"]}
{"school": ["西北工业大学"], "num": ["12"]}
{"school": ["台北艺术大学"], "num": ["0"]}
{"school": ["国防科技大学"], "num": ["11"]}
{"school": ["台湾静宜大学"], "num": ["1"]}
{"school": ["中南大学"], "num": ["11"]}
{"school": ["南京大学"], "num": ["6"]}
{"school": ["北京邮电大学"], "num": ["1"]}
{"school": ["武汉理工大学"], "num": ["4"]}
{"school": ["长安大学"], "num": ["1"]}
{"school": ["ACCA"], "num": ["8"]}
{"school": ["中国创业学院"], "num": ["9"]}
{"school": ["华东师范大学"], "num": ["1"]}
{"school": ["河南消防总队"], "num": ["1"]}
{"school": ["阿德莱德大学"], "num": ["7"]}
{"school": ["Microsoft"], "num": ["7"]}
{"school": ["陆军边海防学院"], "num": ["1"]}
{"school": ["麻省理工学院"], "num": ["65"]}
{"school": ["宁波大学"], "num": ["2"]}
{"school": ["南昌大学"], "num": ["2"]}
{"school": ["台湾宜兰大学"], "num": ["0"]}
{"school": ["莱斯大学"], "num": ["31"]}
{"school": ["北京大学"], "num": ["28"]}
{"school": ["台湾师范大学"], "num": ["0"]}
{"school": ["教育部在线教育研究中心"], "num": ["1"]}
{"school": ["科学松鼠会"], "num": ["1"]}
{"school": ["加州大学伯克利分校"], "num": ["32"]}
{"school": ["暨南大学"], "num": ["7"]}
{"school": ["瓦赫宁根大学"], "num": ["4"]}
{"school": ["中国人民解放军陆军工程大学"], "num": ["3"]}
{"school": ["韦尔斯利大学"], "num": ["5"]}
{"school": ["北京体育大学"], "num": ["2"]}
{"school": ["重庆大学"], "num": ["2"]}

{"school": ["中国石油大学（北京）"], "num": ["0"]}
{"school": ["华东理工大学"], "num": ["1"]}
{"school": ["河北师范大学"], "num": ["2"]}
{"school": ["哈尔滨工程大学"], "num": ["11"]}
{"school": ["浙江大学"], "num": ["1"]}
{"school": ["江苏大学"], "num": ["1"]}
{"school": ["中国人民大学"], "num": ["0"]}
{"school": ["西南交通大学"], "num": ["2"]}
{"school": ["台湾中原大学"], "num": ["0"]}
{"school": ["大叶大学 "], "num": ["3"]}
{"school": ["昆士兰大学"], "num": ["12"]}
{"school": ["台湾空中大学"], "num": ["0"]}
{"school": ["台湾慈济大学"], "num": ["1"]}
{"school": ["红河学院"], "num": ["0"]}
{"school": ["黑龙江大学"], "num": ["3"]}
{"school": ["湖北大学"], "num": ["17"]}
{"school": ["集美大学"], "num": ["0"]}
{"school": ["河北工业大学"], "num": ["6"]}
{"school": ["北京交通大学"], "num": ["2"]}
{"school": ["北京语言大学"], "num": ["3"]}
{"school": ["北京师范大学"], "num": ["10"]}
{"school": ["昆明理工大学"], "num": ["6"]}
{"school": ["宁波城市职业技术学院"], "num": ["7"]}
{"school": ["温州大学"], "num": ["0"]}
{"school": ["浙江金融职业学院"], "num": ["4"]}
{"school": ["浙江纺织服装职业技术学院"], "num": ["1"]}
{"school": ["云南农业大学"], "num": ["0"]}
{"school": ["石河子大学"], "num": ["0"]}
{"school": ["兰州大学"], "num": ["2"]}
{"school": ["东北大学"], "num": ["14"]}
{"school": ["浙江农林大学"], "num": ["0"]}
{"school": ["北京服装学院"], "num": ["2"]}
{"school": ["福建农林大学"], "num": ["10"]}
{"school": ["防灾科技学院"], "num": ["0"]}
{"school": ["江苏大学"], "num": ["1"]}
{"school": ["中央民族大学"], "num": ["0"]}
{"school": ["宁波大红鹰学院"], "num": ["1"]}
{"school": ["陕西工业职业技术学院"], "num": ["0"]}
{"school": ["国际关系学院"], "num": ["0"]}
{"school": ["西安工程大学"], "num": ["0"]}
{"school": ["邢台学院"], "num": ["0"]}
{"school": ["燕山大学"], "num": ["1"]}
{"school": ["昆明医科大学"], "num": ["1"]}
{"school": ["辽宁对外经贸学院"], "num": ["5"]}

{"school": ["青海大学"], "num": ["1"]}
{"school": ["云南民族大学"], "num": ["0"]}
{"school": ["云南大学"], "num": ["2"]}
{"school": ["湖州师范学院"], "num": ["1"]}
{"school": ["大连海事大学"], "num": ["3"]}
{"school": ["河南科技大学"], "num": ["0"]}
{"school": ["洛阳理工学院"], "num": ["1"]}
{"school": ["青海师范大学"], "num": ["0"]}
{"school": ["北京林业大学"], "num": ["3"]}
{"school": ["首都师范大学"], "num": ["1"]}
{"school": ["重庆电子工程职业学院"], "num": ["0"]}
{"school": ["中国传媒大学"], "num": ["4"]}
{"school": ["贵州理工学院"], "num": ["1"]}
{"school": ["贵州中医药大学"], "num": ["0"]}
{"school": ["广州中医药大学"], "num": ["1"]}
{"school": ["深圳职业技术学院"], "num": ["0"]}
{"school": ["咸阳师范学院"], "num": ["0"]}
{"school": ["郑州大学"], "num": ["1"]}
{"school": ["电子科技大学"], "num": ["0"]}
{"school": ["台湾静宜大学"], "num": ["1"]}
{"school": [" 宝鸡文理学院"], "num": ["0"]}
{"school": ["北京工商大学"], "num": ["0"]}
{"school": ["首都医科大学"], "num": ["1"]}
{"school": ["长沙民政职业技术学院 "], "num": ["0"]}
{"school": ["中国石油大学（北京）"], "num": ["0"]}
{"school": ["大连工业大学"], "num": ["2"]}
{"school": ["大连理工大学"], "num": ["6"]}
{"school": ["南方医科大学"], "num": ["0"]}
{"school": ["哈尔滨工程大学"], "num": ["11"]}
{"school": ["黑龙江大学"], "num": ["3"]}
{"school": ["辽宁工程技术大学"], "num": ["6"]}
{"school": ["华北电力大学"], "num": ["0"]}
{"school": ["东北大学"], "num": ["14"]}
{"school": ["汕头大学"], "num": ["1"]}
{"school": ["西南大学"], "num": ["2"]}
{"school": ["芜湖职业技术学院"], "num": ["0"]}
{"school": ["武汉科技大学"], "num": ["0"]}
{"school": ["西安建筑科技大学"], "num": ["0"]}
{"school": ["延安大学"], "num": ["0"]}
{"school": ["燕京理工学院"], "num": ["1"]}
{"school": ["杨凌职业技术学院"], "num": ["3"]}

2、

Spider.py
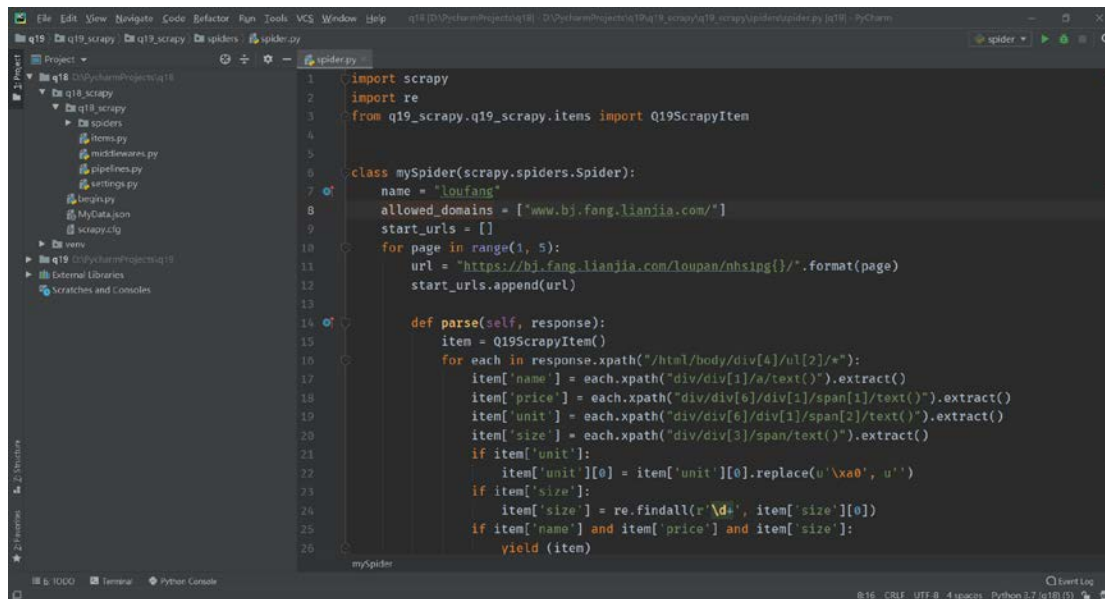
```python
import scrapy
import re
from q19_scrapy.q19_scrapy.items import Q19ScrapyItem


class mySpider(scrapy.spiders.Spider):
    name = "loufang"
    allowed_domains = ["www.bj.fang.lianjia.com/"]
    start_urls = []
    for page in range(1, 5):
        url = "https://bj.fang.lianjia.com/loupan/nhs1pg{}/".format(page)
        start_urls.append(url)

    def parse(self, response):
        item = Q19ScrapyItem()
        for each in response.xpath("/html/body/div[4]/ul[2]/*"):
            item['name'] = each.xpath("div/div[1]/a/text()").extract()
            item['price'] = each.xpath("div/div[6]/div[1]/span[1]/text()").extract()
            item['unit'] = each.xpath("div/div[6]/div[1]/span[2]/text()").extract()
            item['size'] = each.xpath("div/div[3]/span/text()").extract()
            if item['unit']:
                item['unit'][0] = item['unit'][0].replace(u'\xa0', u'')
            if item['size']:
                item['size'] = re.findall(r'\d+', item['size'][0])
            if item['name'] and item['price'] and item['size']:
                yield (item)
```
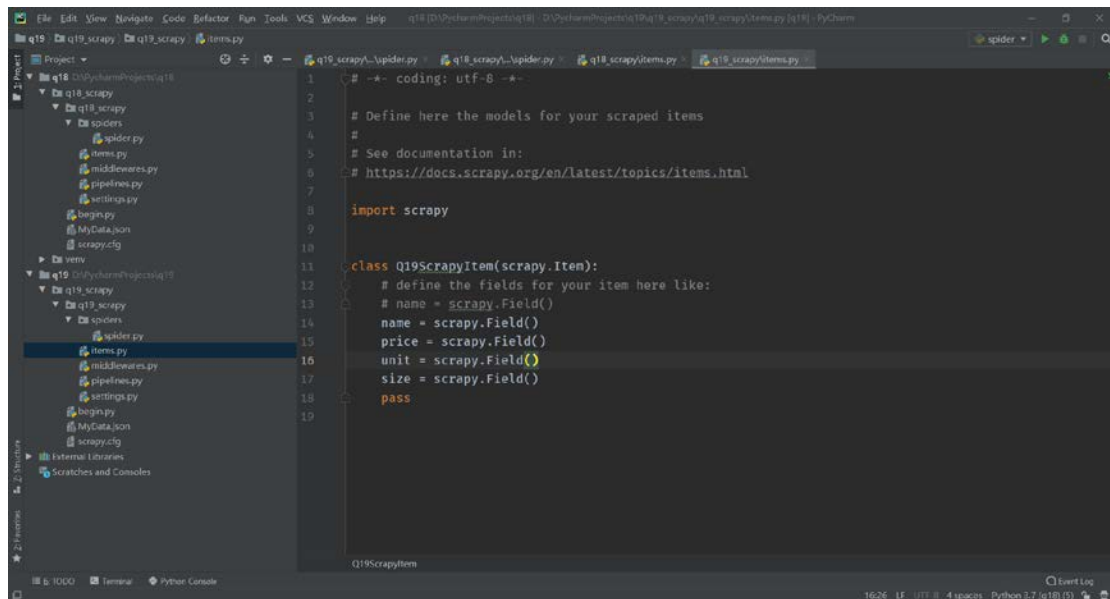
Items.py

name 为楼盘名称，price 为价格，unit 为单位，size 为平米数

```
# -*- coding: utf-8 -*-

# Define here the models for your scraped items
#
# See documentation in:
# https://docs.scrapy.org/en/latest/topics/items.html

import scrapy


class Q19ScrapyItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    name = scrapy.Field()
    price = scrapy.Field()
    unit = scrapy.Field()
    size = scrapy.Field()
    pass
```
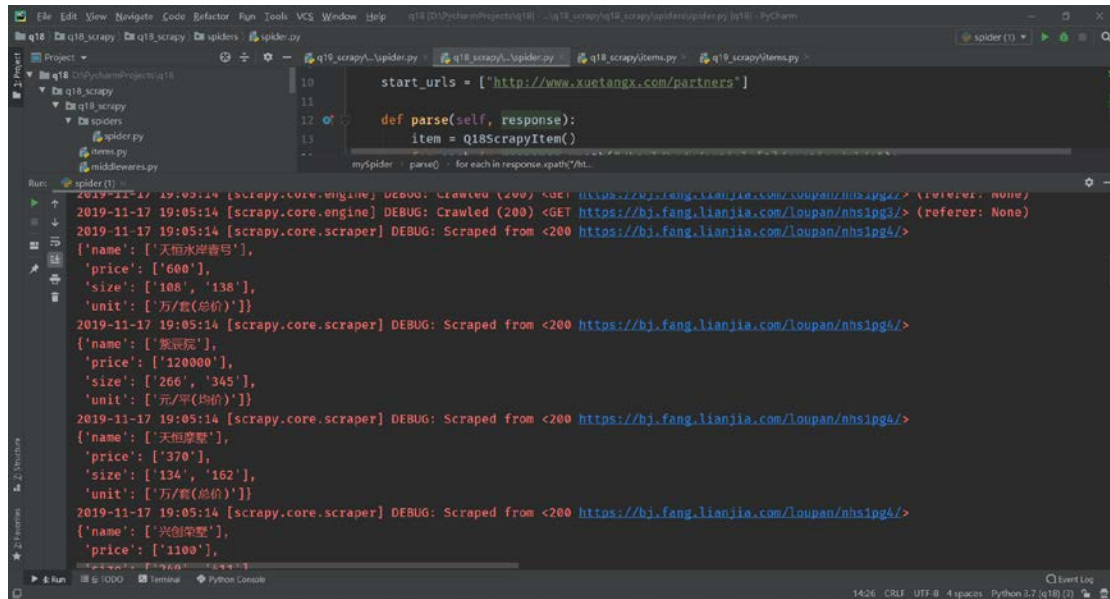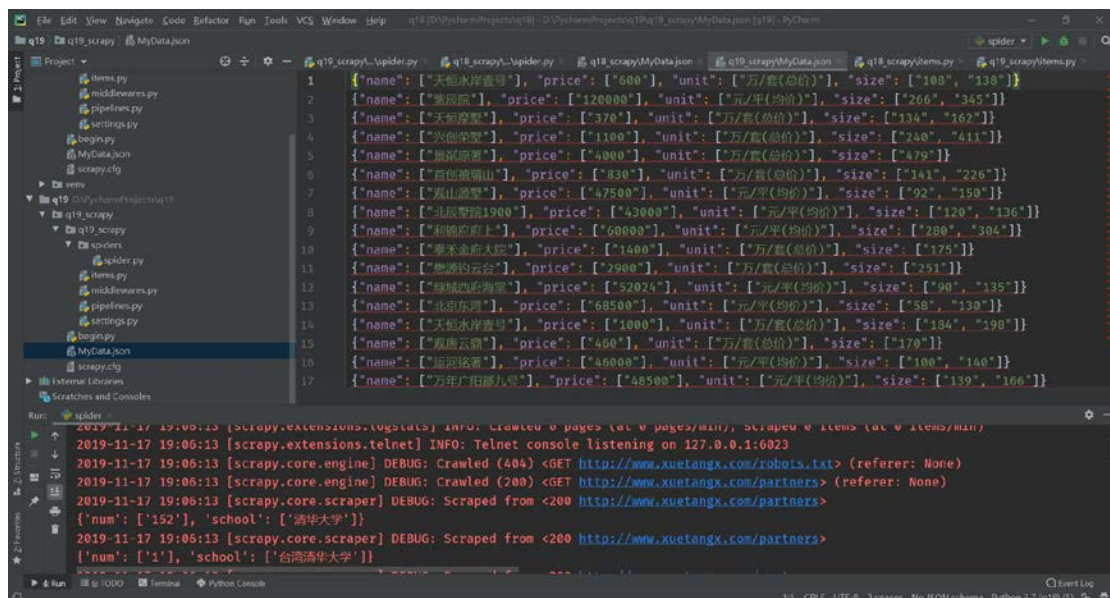
运行：

结果：

{"name": ["天恒水岸壹号"], "price": ["600"], "unit": ["万/套(总价)"], "size": ["108", "138"]}

{"name": ["紫辰院"], "price": ["120000"], "unit": ["元/平(均价)"], "size": ["266", "345"]}

{"name": ["天恒摩墅"], "price": ["370"], "unit": ["万/套(总价)"], "size": ["134", "162"]}

{"name": ["兴创荣墅"], "price": ["1100"], "unit": ["万/套(总价)"], "size": ["240", "411"]}

{"name": ["景郡原著"], "price": ["4000"], "unit": ["万/套(总价)"], "size": ["479"]}

{"name": ["首创禧瑞山"], "price": ["830"], "unit": ["万/套(总价)"], "size": ["141", "226"]}

{"name": ["观山源墅"], "price": ["47500"], "unit": ["元/平(均价)"], "size": ["92", "150"]}

{"name": ["北辰墅院1900"], "price": ["43000"], "unit": ["元/平(均价)"], "size": ["120", "136"]}

{"name": ["利锦府上"], "price": ["60000"], "unit": ["元/平(均价)"], "size": ["280", "304"]}

{"name": ["泰禾金府大院"], "price": ["1400"], "unit": ["万/套(总价)"], "size": ["175"]}

{"name": ["懋源钓云台"], "price": ["2900"], "unit": ["万/套(总价)"], "size": ["251"]}

{"name": ["绿城西府海棠"], "price": ["52024"], "unit": ["元/平(均价)"], "size": ["90", "135"]}

{"name": ["北京东湾"], "price": ["68500"], "unit": ["元/平(均价)"], "size": ["58", "130"]}
{"name": ["天恒水岸壹号"], "price": ["1000"], "unit": ["万/套(总价)"], "size": ["184", "198"]}
{"name": ["观唐云鼎"], "price": ["460"], "unit": ["万/套(总价)"], "size": ["170"]}
{"name": ["运河铭著"], "price": ["46000"], "unit": ["元/平(均价)"], "size": ["100", "140"]}
{"name": ["万年广阳郡九号"], "price": ["48500"], "unit": ["元/平(均价)"], "size": ["139", "166"]}
{"name": ["首开璞瑅公馆"], "price": ["106000"], "unit": ["元/平(均价)"], "size": ["236"]}
{"name": ["华远裘马四季"], "price": ["60000"], "unit": ["元/平(均价)"], "size": ["156", "191"]}
{"name": ["V7九间堂"], "price": ["1700"], "unit": ["万/套(总价)"], "size": ["220", "420"]}
{"name": ["北京城建国誉府"], "price": ["48000"], "unit": ["元/平(均价)"], "size": ["143", "297"]}
{"name": ["御汤山熙园"], "price": ["1800"], "unit": ["万/套(总价)"], "size": ["300", "470"]}
{"name": ["华远和墅"], "price": ["1580"], "unit": ["万/套(总价)"], "size": ["295"]}
{"name": ["天资华府"], "price": ["42000"], "unit": ["元/平(均价)"], "size": ["93", "276"]}
{"name": ["檀香府"], "price": ["55000"], "unit": ["元/平(均价)"], "size": ["208", "320"]}
{"name": ["观山源墅"], "price": ["47500"], "unit": ["元/平(均价)"], "size": ["290", "437"]}
{"name": ["中国铁建花语金郡"], "price": ["71000"], "unit": ["元/平(均价)"], "size": ["150"]}
{"name": ["中国铁建万科翡翠长安"], "price": ["57000"], "unit": ["元/平(均价)"], "size": ["166", "204"]}
{"name": ["电建金地华宸"], "price": ["64000"], "unit": ["元/平(均价)"], "size": ["180", "247"]}
{"name": ["北辰璟院1900"], "price": ["43000"], "unit": ["元/平(均价)"], "size": ["251", "282"]}
{"name": ["首创天阅西山"], "price": ["1500"], "unit": ["万/套(总价)"], "size": ["180", "211"]}
{"name": ["西山燕庐"], "price": ["65000"], "unit": ["元/平(均价)"], "size": ["123", "172"]}
{"name": ["北科建泰禾丽春湖院子"], "price": ["1020"], "unit": ["万/套(总价)"], "size": ["468", "800"]}
{"name": ["绿地海珀云翡"], "price": ["67000"], "unit": ["元/平(均价)"], "size": ["102", "178"]}
{"name": ["都丽华府"], "price": ["300"], "unit": ["万/套(总价)"], "size": ["90", "172"]}
{"name": ["中粮京西祥云"], "price": ["58000"], "unit": ["元/平(均价)"], "size": ["115", "140"]}
{"name": ["燕西华府"], "price": ["49800"], "unit": ["元/平(均价)"], "size": ["60", "125"]}