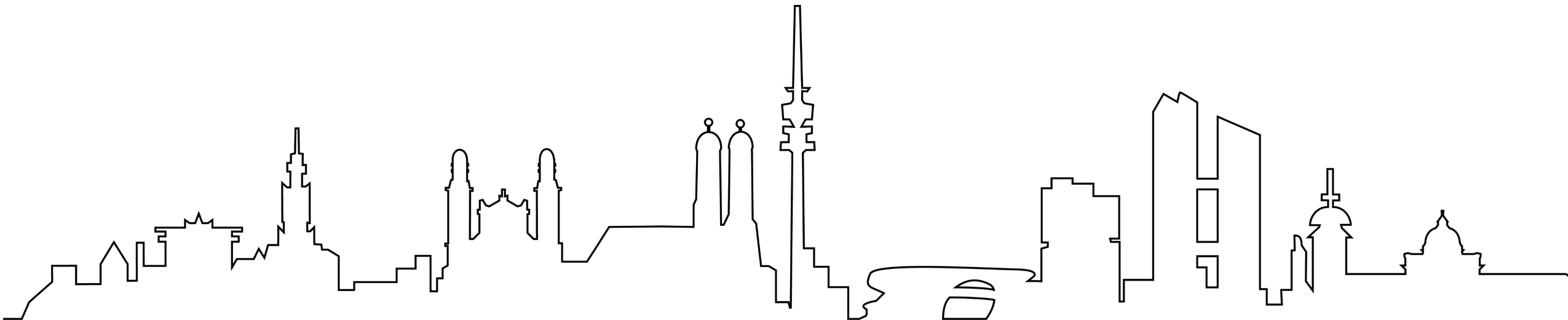
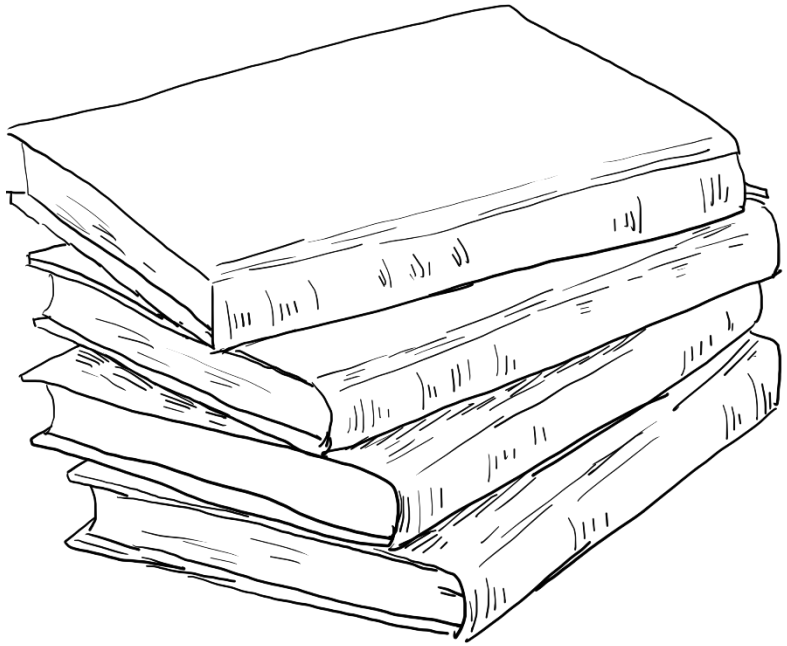
自 如 租 房

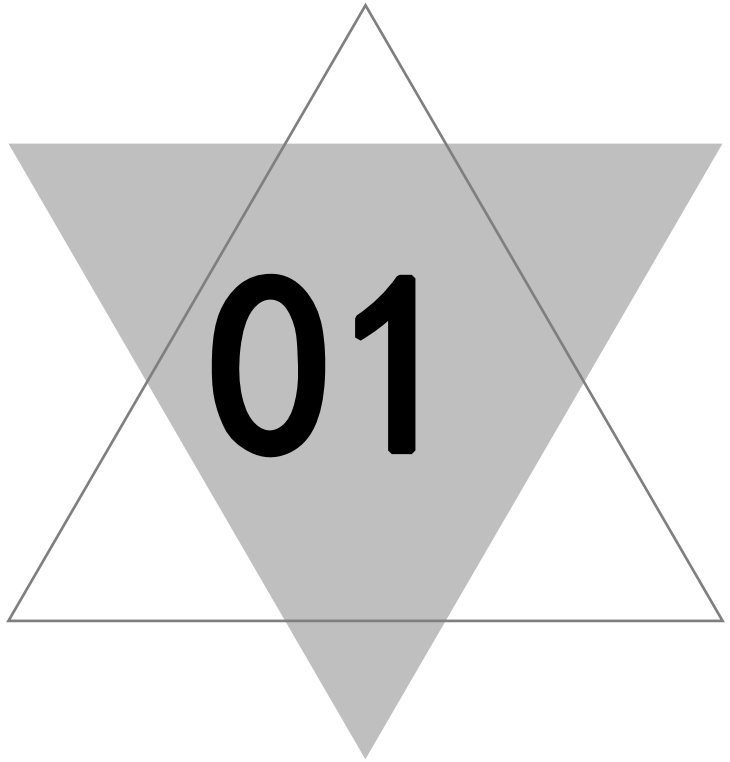
—— 上海市租房分析

第五组：姜灼洁 卢益康 欧盈池 杨澜 周剑波





目　录



背　景

01 背景

租房需求

—— 实习和工作中存在租房需求

—— 租房信息过载

中国最大020青年居住社区

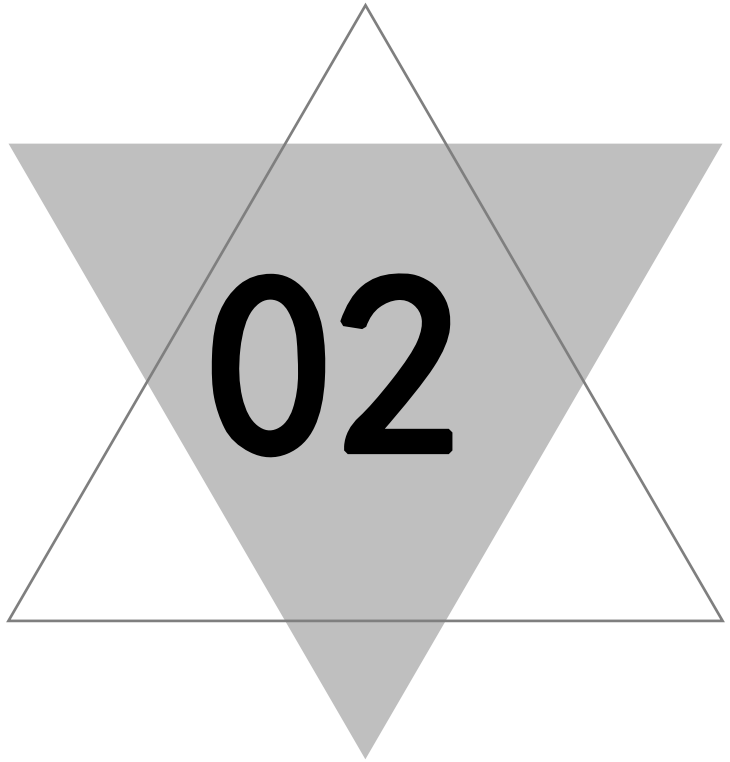
—— 省去传统租房模式冗余环节

—— 众多年轻人的租房选择

新的知识

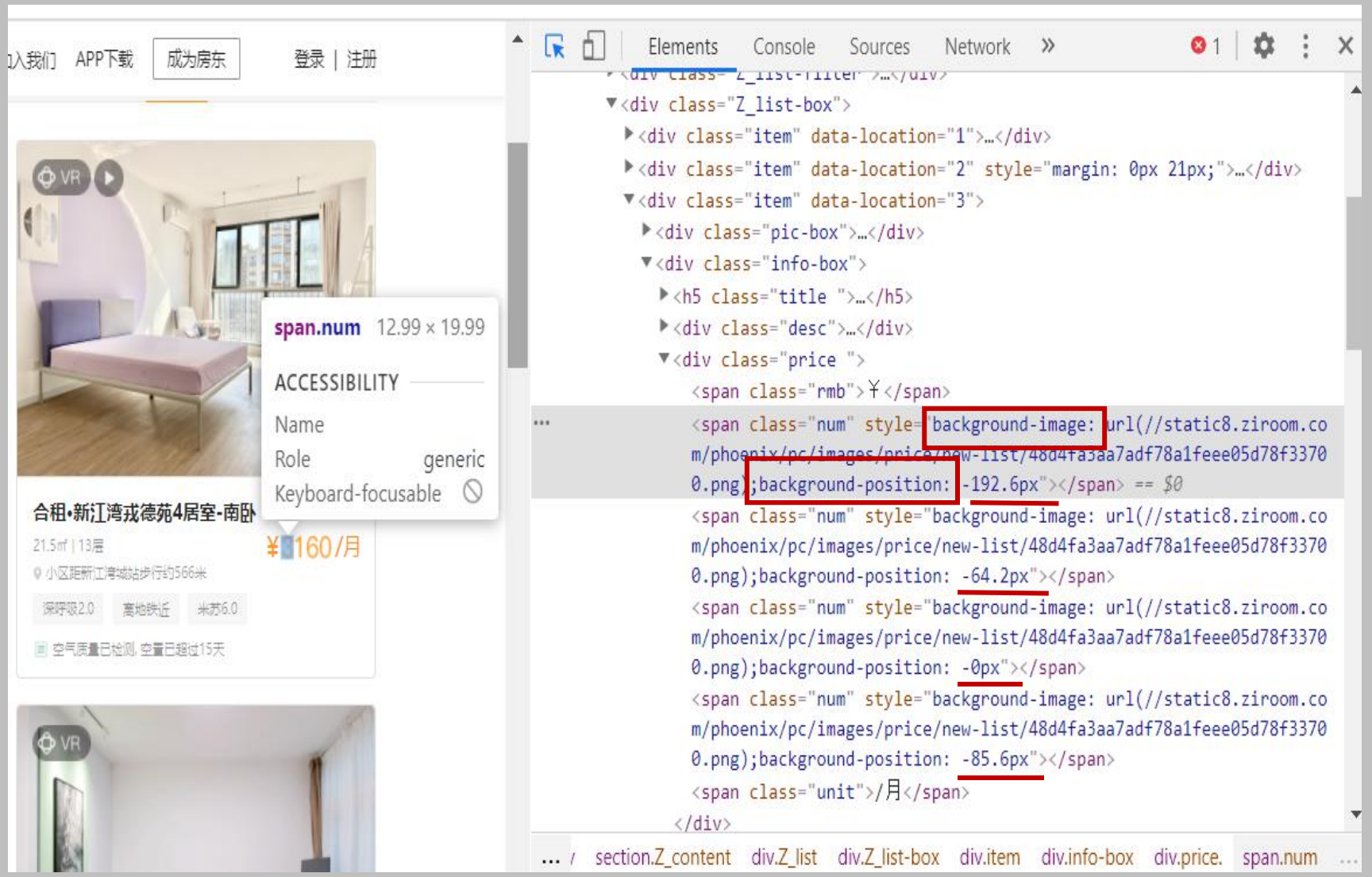
—— 雪碧图反爬

—— k-means聚类分析



随机样本爬取

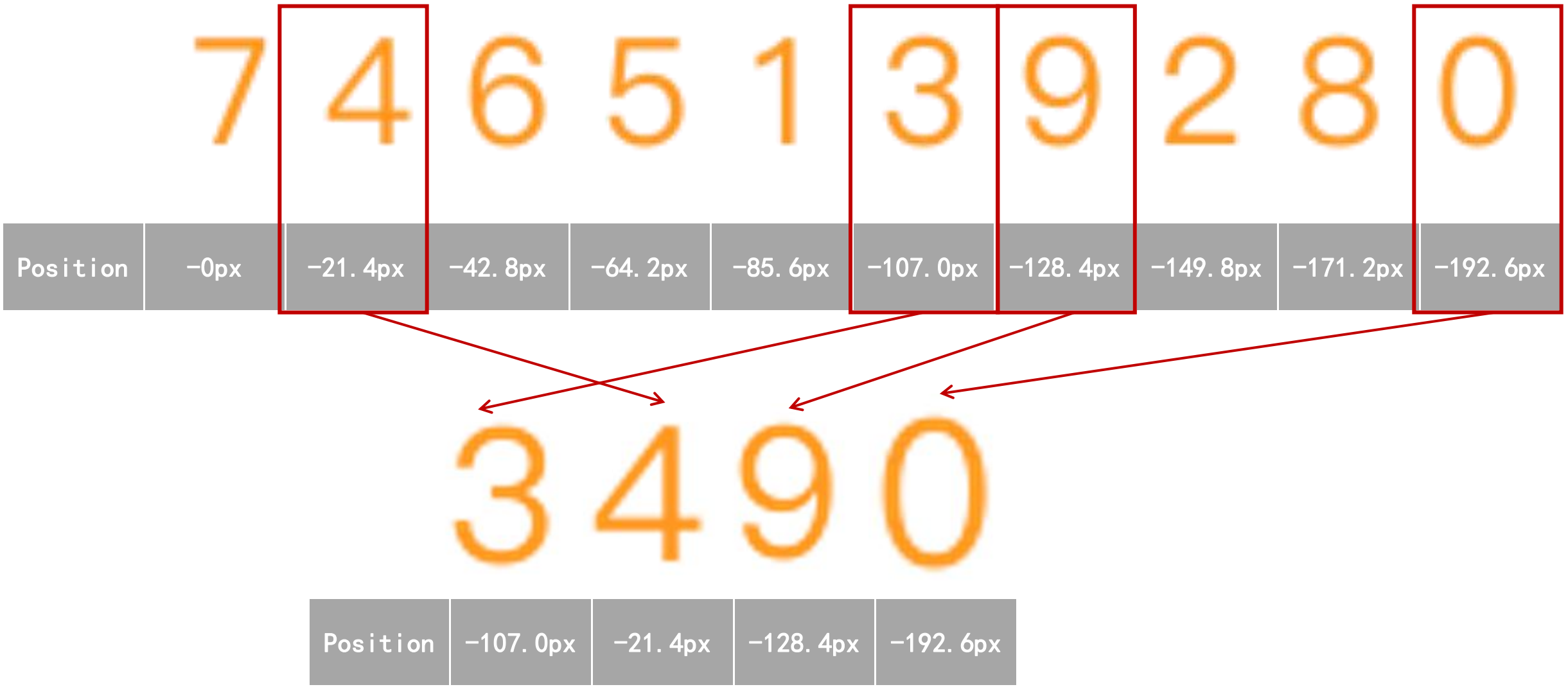
自如的反爬机制



雪碧图

把网页中一些背景图片整合到一张图片文件中，再利用“background-image”，“background-position”的组合进行背景定位，background-position可以用数字精确的定位出背景图片的位置。

自如的反爬机制


7 4 6 5 1 3 9 2 8 0
Position -0px -21.4px -42.8px -64.2px -85.6px -107.0px -128.4px -149.8px -171.2px -192.6px
3 4 9 0
Position -107.0px -21.4px -128.4px -192.6px

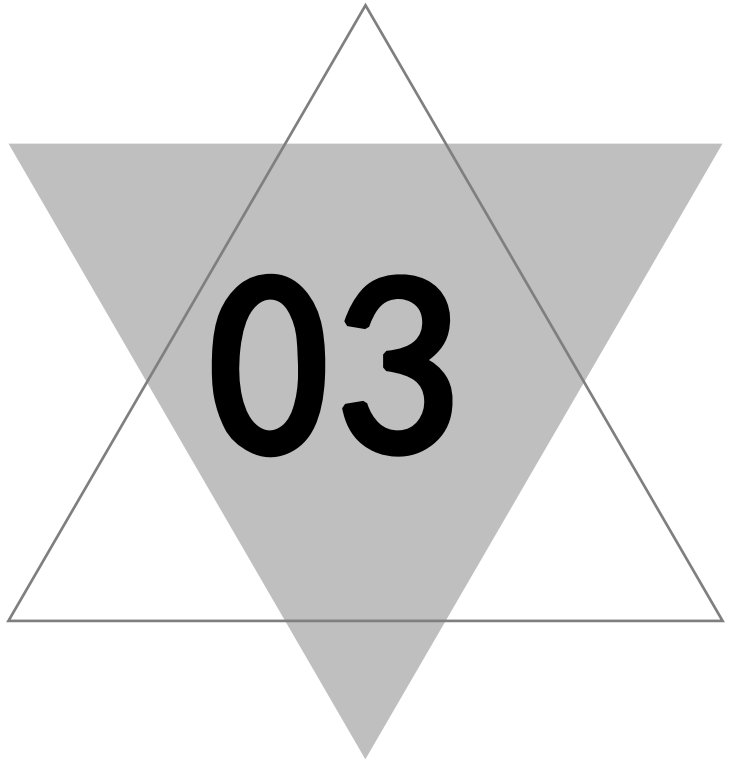
自如的反爬机制

解决反爬

1. 抓取图片，用图片识别包pytesseract，识别出图片中的每一个数字

2. 建立一个数字元素和图片位置对应的字典

3. 根据网页中抓取到的位置信息，找出相应的数字，表示价格



数据处理与可视化

03 数据处理与可视化

01 数据预处理

- 去掉重复值
- 数据分割
- 提取数值
- 0值转换为空值
- 租金统一以月为单位

02 K-means聚类

- 中位数填充缺失值
- 标准化
- 设置哑变量
- 占比分析

03 数据处理与可视化

数据预处理前

	0
0	3560 合租·地杰国际城F欧香四季4居室-南卧 18㎡ 14/18层 小区距御桥站步行约...
1	9460 整租·华升公寓2室1厅-南 96.41㎡ 20/34层 小区距江浦路站步行约150米
2	3290 合租·博爱家园4居室-南卧 17.5㎡ 6/6层 小区距高科西路站步行约234米
3	3030 合租·紫叶花园东园4居室-南卧 17.5㎡ 5/6层 小区距北蔡站步行约216米
4	3990 合租·仁和苑4居室-南卧 17.41㎡ 7/12层 小区距江湾体育场站步行约243米

数据预处理后

	0	1	2	3	4	5	area	distance	floor	floors	price
0	3560	合租·地杰国际城F欧香四季4居室-南卧	18㎡		14/18层	小区距御桥站步行约191米	18	191	14	18	3560
1	9460	整租·华升公寓2室1厅-南	96.41㎡		20/34层	小区距江浦路站步行约150米	96	150	20	34	9460
2	3290	合租·博爱家园4居室-南卧	17.5㎡		6/6层	小区距高科西路站步行约234米	17	234	6	6	3290
3	3030	合租·紫叶花园东园4居室-南卧	17.5㎡		5/6层	小区距北蔡站步行约216米	17	216	5	6	3030
4	3990	合租·仁和苑4居室-南卧	17.41㎡		7/12层	小区距江湾体育场站步行约243米	17	243	7	12	3990

03 数据处理与可视化

01 数据预处理

- 去掉重复值
- 数据分割
- 提取数值
- 0值转换为空值
- 租金统一以月为单位

02 K-means聚类

- 中位数填充缺失值
- 标准化
- 设置哑变量
- 占比分析

03 数据处理与可视化

k-means聚类简介

- 非监督学习的算法
- 以空间中k个点为中心进行聚类，对最靠近他们的对象归类，通过迭代的方法，逐次更新各聚类中心的值，直到得到最好的聚类结果

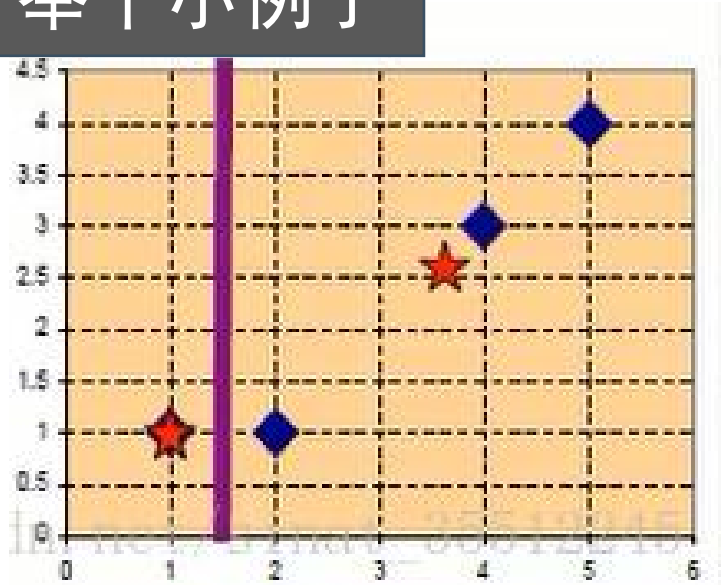
聚类过程

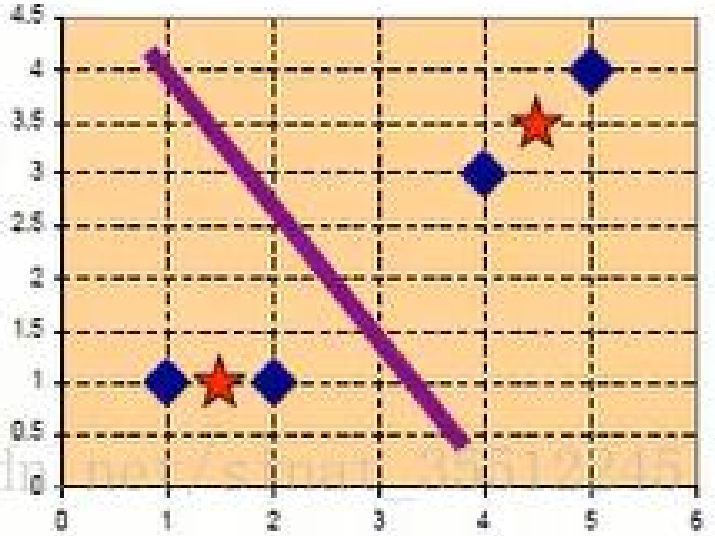
- 适当选择c个类的初始中心
- 在第k次迭代中，对任意一个样本，求其到每个类中心的距离，将该样本归到距离最短的那个中心所在的类
- 利用均值等方法更新该类的中心值
- 对于所有的聚类中心，如果利用的迭代法更新后，每个中心值均保持不变，则迭代结束；否则继续迭代

03 数据处理与可视化

理解K-means聚类——举个小例子

药物名词	药物重量	药物PH值
A	1	1
B	2	1
C	4	3
D	5	4





c1=(1,1) c2=(2,1)

D0= [0.00, 1.00, 3.61, 5.00
1.00, 0.00, 2.83, 4.24]

G0= [1, 0, 0, 0
0, 1, 1, 1]

c1=(1,1) c2=(13/3,8/3)

D1= [0.00, 1.00, 3.61, 5.00
3.14, 2.36, 0.47, 1.89]

G1= [1, 1, 0, 0
0, 0, 1, 1]

c1=(1.5,1) c2=(4.5,3.5)

D2= [0.50, 0.50, 3.20, 4.61
4.30, 3.54, 0.71, 0.71]

G2= [1, 1, 0, 0
0, 0, 1, 1]

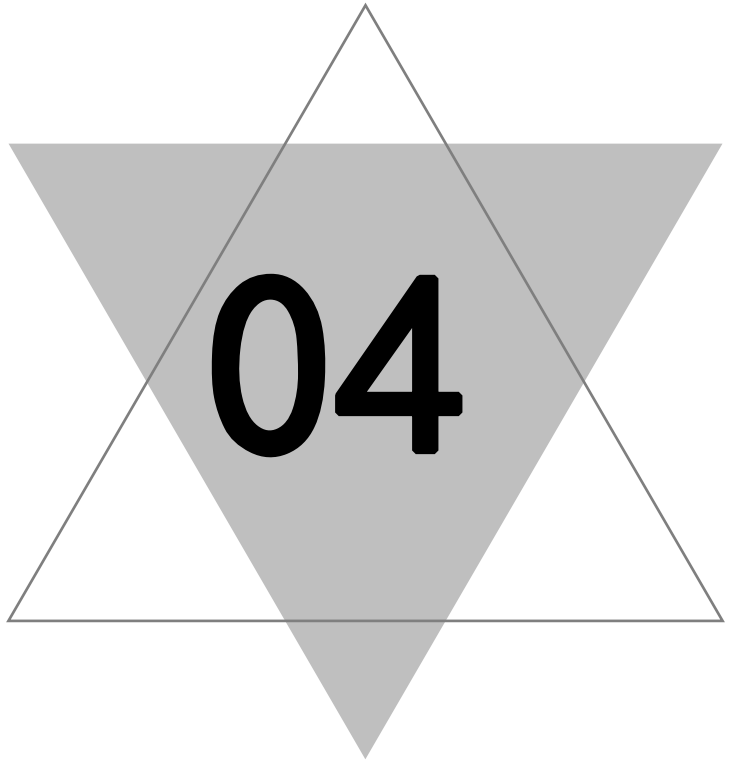
03 数据处理与可视化

上海各区房源占比

	district	交通不便普通	交通一般普通	交通便利普通	高端房源	豪宅
浦东新区	364	91	65	116	91	1
普陀区	163	18	16	39	87	3
闵行区	138	75	39	12	10	2
宝山区	103	60	36	5	0	2
嘉定区	85	53	29	1	0	2
长宁区	85	0	5	45	34	1
徐汇区	69	1	2	17	48	1
松江区	62	43	15	2	2	0
杨浦区	61	6	10	19	25	1
静安区	59	3	4	28	22	2
虹口区	47	8	6	15	18	0
黄浦区	38	0	0	5	32	1
青浦区	16	16	0	0	0	0

结果分析

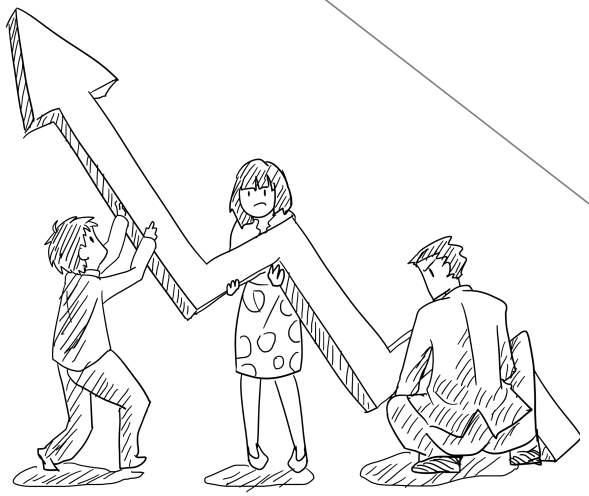
- 交通便利的普通房源
- 交通不便的普通房源
- 交通一般的普通房源
- 豪宅与高端房源



全样本爬取+推荐算法

那些踩过的坑

- 反爬设计
- 动态加载
- 页面跳转
- 广告
- 元素交互性
- 网页404
- 翻页
- HTTP拒绝访问



总结

爬虫最大的敌人

Sufe-TEL-1x