[爬虫+数据分析] 分析北京Python开发的现状

程序员共成长

相信各位同学多多少少在拉钩上投过简历，今天突然想了解一下北京Python开发的薪资水平、招聘要求、福利待遇以及公司地理位置。既然要分析那必然是现有数据样本。本文通过爬虫和数据分析为大家展示一下北京Python开发的现状，希望能够在职业规划方面帮助到大家！！！

爬虫

爬虫的第一步自然是从分析请求和网页源代码开始。从网页源代码中我们并不能找到发布的招聘信息。但是在请求中我们看到这样一条POST请求

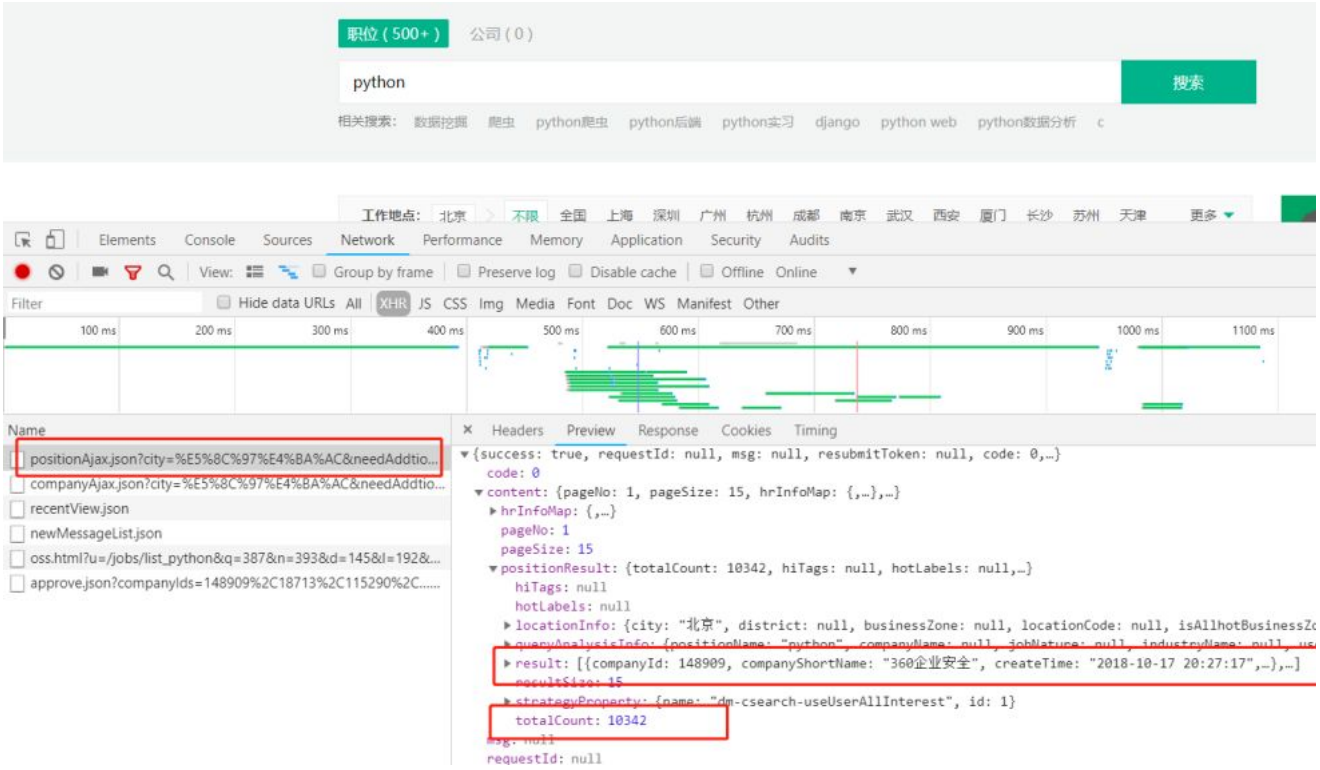
如下图我们可以得知

url：https://www.lagou.com/jobs/positionAjax.json?city=%E5%8C%97%E4%BA%AC&needAddtionalResult=false

请求方式：post

result：为发布的招聘信息

totalCount：为招聘信息的条数



“

通过实践发现除了必须携带headers之外，拉勾网对ip访问频率也是有限制的。一开始会提示 ’访问过于频繁’，继续访问则会将ip拉入黑名单。不过一段时间之后会自动从黑名单中移除。

针对这个策略，我们可以对请求频率进行限制，这个弊端就是影响爬虫效率。

其次我们还可以通过代理ip来进行爬虫。网上可以找到免费的代理ip，但大都不太稳定。付费的价格又不太实惠。

具体就看大家如何选择了

”

通过分析请求我们发现每页返回15条数据，totalCount又告诉了我们该职位信息的总条数。

向上取整就可以获取到总页数。然后将所得数据保存到csv文件中。这样我们就获得了数据分析的数据源！

post请求的Form Data传了三个参数

first ： 是否首页(并没有什么用)

pn：页码

kd：搜索关键字

```
#  获取请求结果

#  kind  搜索关键字
```

```
#  page  页码  默认是1
def  get_json(kind,  page=1,):
        #  post请求参数
        param  =  {
                'first':  'true',
                'pn':  page,
                'kd':  kind
        }
        header  =  {
                'Host':  'www.lagou.com',
                'Referer':  'https://www.lagou.com/jobs/list_python?labelWords=&fromSearch=true&suginput=',
                'User-
Agent':  'Mozilla/5.0  (Macintosh;  Intel  Mac  OS  X  10_12_6)  AppleWebKit/537.36  (KHTML,  like  Gecko)  Chrome/69.0.34
        }
        #  设置代理
        proxies  =  [
                {'http':  '140.143.96.216:80',  'https':  '140.143.96.216:80'},
                {'http':  '119.27.177.169:80',  'https':  '119.27.177.169:80'},
                {'http':  '221.7.255.168:8080',  'https':  '221.7.255.168:8080'}
        ]
        #  请求的url
        url  =  'https://www.lagou.com/jobs/positionAjax.json?
px=default&city=%E5%8C%97%E4%BA%AC&needAddtionalResult=false'
        #  使用代理访问
        #  response  =  requests.post(url,  headers=header,  data=param,  proxies=random.choices(proxies))
        response  =  requests.post(url,  headers=header,  data=param,  proxies=proxies)
        response.encoding  =  'utf-8'
        if  response.status_code  ==  200:
                response  =  response.json()
```

```
            #  请求响应中的positionResult  包括查询总数  以及该页的招聘信息(公司名、地址、薪资、福利待遇等...)
            return  response['content']['positionResult']
        return  None
```

接下来我们只需要每次翻页之后调用 get_json 获得请求的结果 再遍历取出需要的招聘信息即可

```
if  __name__  ==  '__main__':
        #  默认先查询第一页的数据
        kind  =  'python'
        #  请求一次  获取总条数
        position_result  =  get_json(kind=kind)
        #  总条数
        total  =  position_result['totalCount']
        print('{}开发职位，招聘信息总共{}条.....'.format(kind,  total))
        #  每页15条  向上取整  算出总页数
        page_total  =  math.ceil(total/15)
        #  所有查询结果
        search_job_result  =  []
        #for  i  in  range(1,  total  +  1)
        #  为了节约效率  只爬去前100页的数据
        for  i  in  range(1,  100):
            position_result  =  get_json(kind=kind,  page=  i)
            #  每次抓取完成后,暂停一会,防止被服务器拉黑
            time.sleep(15)
            #  当前页的招聘信息
            page_python_job  =  []
            for  j  in  position_result['result']:
                python_job  =  []
                #  公司全名
                python_job.append(j['companyFullName'])
```

```
                # 公司简称
                python_job.append(j['companyShortName'])
                # 公司规模
                python_job.append(j['companySize'])
                # 融资
                python_job.append(j['financeStage'])
                # 所属区域
                python_job.append(j['district'])
                # 职称
                python_job.append(j['positionName'])
                # 要求工作年限
                python_job.append(j['workYear'])
                # 招聘学历
                python_job.append(j['education'])
                # 薪资范围
                python_job.append(j['salary'])
                # 福利待遇
                python_job.append(j['positionAdvantage'])
                page_python_job.append(python_job)
            # 放入所有的列表中
            search_job_result += page_python_job
            print('第{}页数据爬取完毕, 目前职位总数:{}'.format(i, len(search_job_result)))
            # 每次抓取完成后,暂停一会,防止被服务器拉黑
            time.sleep(15)
```

ok! 数据我们已经获取到了，最后一步我们需要将数据保存下来

```
    # 将总数据转化为data frame再输出
        df = pd.DataFrame(data=search_job_result,
```

```
                                    columns=['公司全名', '公司简称', '公司规模', '融资阶段', '区域', '职位名称', '工作经验', '学历要求', '工资', '职位福利'])

        df.to_csv('lagou.csv', index=False, encoding='utf-8_sig')
```

运行main方法直接上结果：

	公司全名	公司简称	公司规模	融资阶段	区域	职位名称	工作经验	学历要求	工资	职位福利
2	北京启明合心科技有限公	合心科技	50-150人	A轮	海淀区	Python开发工程师	1-3年	本科	20k-35k	大数据处理,自由度大,有挑
3	京东商城-技术研发体系	京东商城研发部	2000人以上	上市公司	大兴区	Python	5-10年	本科	30k-50k	行业前沿,部门直招,核心加
4	北京奇艺世纪科技有限公	爱奇艺	2000人以上	上市公司	海淀区	Python高级开发工程师	3-5年	本科	20k-40k	互联网大厂,成长空间,超强
5	乐飞天下信息技术(北京	乐飞天下	50-150人	未融资	朝阳区	Python中级开发工程师	5-10年	大专	17k-34k	福利好,带薪假,员工旅游,调
6	上海海知智能科技有限公	海知智能	50-150人	A轮	海淀区	Python开发工程师	3-5年	本科	20k-40k	创新前沿,技术大牛,成长空
7	北京点石经纬科技有限公	新东方教育行业研究院	150-500人	B轮	海淀区	Python工程师	3-5年	大专	15k-30k	机会多多,K12,平台好
8	上海基分文化传播有限公	趣头条	500-2000人	上市公司	海淀区	python开发工程师	5-10年	本科	20k-40k	五险一金,补充医疗,带薪年
9	掌阅科技股份有限公司	掌阅	500-2000人	上市公司	朝阳区	高级Python研发工程师	3-5年	本科	25k-35k	扁平化管理,福利待遇佳,学
10	北京好巧国际旅行社有限	好巧	150-500人	B轮	朝阳区	Python高级研发工程师	3-5年	本科	20k-35k	六险一金,弹性不打卡,水果
11	北京优虎商务服务有限公	老虎证券	150-500人	C轮	朝阳区	Python工程师	3-5年	本科	18k-35k	6险一金,弹性工作,团队好,
12	贝壳找房（北京）科技有	贝壳	2000人以上	C轮	海淀区	Python开发工程师	1-3年	本科	25k-50k	15薪,13天年假,免费三餐,
13	北京智分科技有限公司	NewsJet	50-150人	不需要融资	海淀区	高级Python开发工程师	3-5年	本科	25k-45k	Geek范,清华系,福利待遇好
14	京东金融	京东金融	2000人以上	上市公司	大兴区	高级Python开发工程师	5-10年	本科	20k-40k	班车,餐补
15	北京旷视科技有限公司	Face++	500-2000人	C轮	海淀区	高级Python开发工程师	3-5年	本科	25k-50k	大牛多
16	北京汉迪移动互联网科技	iHandy	150-500人	不需要融资	海淀区	高级Python工程师	3-5年	本科	25k-50k	硅谷氛围,高手如云,零食水
17	北京启明合心科技有限公	合心科技	50-150人	A轮	海淀区	中级python开发工程师	不限	本科	20k-35k	领导好,不打卡,午晚餐,年
18	北京点石经纬科技有限公	新东方教育行业研究院	150-500人	B轮	海淀区	高级python工程师	5-10年	大专	20k-30k	机会多多,平台好,领导好
19	北京步鼎万舟科技有限公	微车	150-500人	C轮	海淀区	Python高级开发工程师	3-5年	本科	20k-30k	一日三餐、带薪年假、弹性
20	北京数美时代科技有限公	数美	150-500人	B轮	朝阳区	python研发工程师	1年以下	硕士	15k-30k	飞速发展,技术氛围浓,大牛
21	北京首都在线科技股份有	CDS	150-500人	上市公司	海淀区	python研发工程师	5-10年	本科	15k-30k	薪酬福利好,云计算公司
22	中电信用服务有限公司	中电信用	50-150人	未融资	海淀区	Python工程师	5-10年	本科	15k-25k	五险一金,带薪年假,餐补车
23	北京博派通达科技有限公	博派通达	50-150人	A轮	朝阳区	Python开发工程师	3-5年	不限	15k-30k	不限量零食,福利多多,年终
24	上海谦问万答吧云计算科	学霸君	2000人以上	C轮	朝阳区	Python研发工程师	3-5年	本科	20k-35k	地铁周边
25	柔持（北京）科技有限公	柔持英语	150-500人	B轮	朝阳区	Python后端开发工程师	1-3年	本科	20k-30k	交通方便,弹性时间,不加班
26	北京旷视科技有限公司	Face++	500-2000人	C轮	海淀区	Python 高级开发工程师	3-5年	本科	20k-40k	人工智能,黑科技,大牛多
27	北京好巧国际旅行社有限	好巧	150-500人	B轮	朝阳区	Linux C++/Python中高级研	5-10年	本科	25k-35k	弹性工作,六险一金,百万医
28	北京智者天下科技有限公	知乎	500-2000人	D轮及以上	海淀区	资深Python开发	5-10年	本科	30k-45k	三餐,健身房,nac
29	达而观信息科技（上海）	达观数据	50-150人	B轮	海淀区	Python高级开发工程师	5-10年	本科	18k-36k	团队技术强,发展空间大,人
30	北京汇滴科技有限公司	汇滴科技	15-50人	不需要融资	朝阳区	中高级Python开发工程师	3-5年	本科	20k-30k	五险一金,全勤奖,餐补,年
31	北京闪银奇异科技有限公	闪银奇异	500-2000人	D轮及以上	朝阳区	Python工程师	3-5年	本科	20k-40k	七险一金,核心项目,高新企
32	北京步鼎万舟科技有限公	微车	150-500人	C轮	海淀区	Python开发工程师	3-5年	本科	10k-18k	一日三餐,六险一金,14薪,
33	柔持（北京）科技有限公	柔持英语	150-500人	B轮	朝阳区	Python开发工程师	3-5年	本科	20k-40k	五险一金,pytho
34	北京国双科技有限公司	Gridsum 国双	500-2000人	上市公司	海淀区	资深Python研发工程师	3-5年	本科	20k-40k	上市公司,六险一金,发挥空
35	北京一览科技有限公司	一览科技	150-500人	B轮	朝阳区	python开发工程师	3-5年	本科	20k-40k	福利优厚
36	北京果壳互动科技传媒有	果壳网	150-500人	C轮	朝阳区	高级Python开发工程师	3-5年	本科	20k-40k	八险一金,免费三餐,不加班,
37	北京好巧国际旅行社有限	好巧	150-500人	B轮	朝阳区	Python研发工程师	3-5年	大专	18k-30k	弹性工作,百万医疗,10天年
38	北京智者天下科技有限公	知乎	500-2000人	D轮及以上	海淀区	Python开发工程师	3-5年	本科	15k-30k	薪资优厚、氛围宽松、每日
39	北京汉迪移动互联网科技	iHandy	150-500人	不需要融资	海淀区	Python开发工程师	1-3年	本科	15k-30k	硅谷氛围,高手如云,零食水
40	人人贷商务顾问（北京）	人人贷	2000人以上	A轮	海淀区	Python开发高级工程师	3-5年	本科	30k-60k	五险一金,节假日礼金,竞争
41	北京右划网络科技有限公	右划	50-150人	A轮	朝阳区	高级python开发工程师	3-5年	本科	30k-60k	短视频,大牛云集,工程师文
42	北京旷视科技有限公司	Face++	500-2000人	C轮	海淀区	Python工程师	1-3年	本科	20k-40k	技术驱动,极客氛围,大牛团
43	北京天广汇通科技有限公	天广汇通	50-150人	不需要融资	海淀区	python开发工程师	3-5年	本科	15k-25k	软件工程师
44	车好多旧机动车经纪(北	瓜子二手车直卖网	2000人以上	C轮	海淀区	Python开发工程师	3-5年	本科	20k-40k	最高19薪 独角兽公司 学习
45	炫一下（北京）科技有限	一下科技	500-2000人	D轮及以上	朝阳区	Python研发工程师	3-5年	本科	35k-55k	硕士最好,不加班,18薪,有其
46	遨游酒店信息技术(深圳	遨游酒店信息技术(深圳	500-2000人	不需要融资	朝阳区	高级Python研发工程师	3-5年	本科	15k-30k	弹性上下班,高速发展,高速

数据分析

通过分析cvs文件，为了方便我们统计，我们需要对数据进行清洗

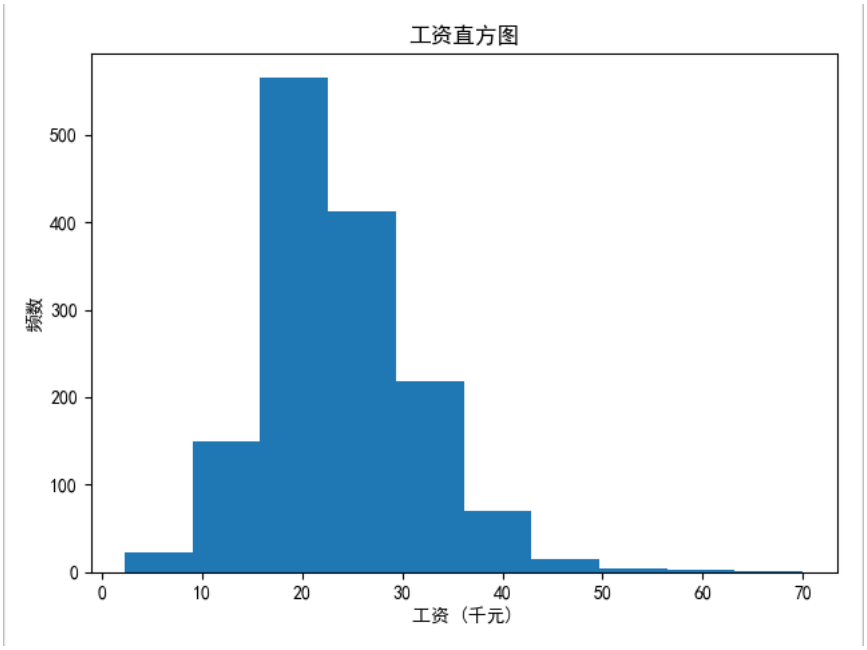
比如剔除实习岗位的招聘、工作年限无要求或者应届生的当做 0年处理、薪资范围需要计算出一个大概的值、学历无要求的当成大专

```
#  读取数据

df  =  pd.read_csv('lagou.csv',  encoding='utf-8')

#  数据清洗,剔除实习岗位

df.drop(df[df['职位名称'].str.contains('实习')].index,  inplace=True)

#  print(df.describe())

#  由于CSV文件内的数据是字符串形式,先用正则表达式将字符串转化为列表,再取区间的均值

pattern  =  '\d+'

df['work_year']  =  df['工作经验'].str.findall(pattern)
```

```
#  数据处理后的工作年限
avg_work_year  =  []
#  工作年限
for  i  in  df['work_year']:
      #  如果工作经验为'不限'或'应届毕业生',那么匹配值为空,工作年限为0
      if  len(i)  ==  0:
            avg_work_year.append(0)
      #  如果匹配值为一个数值,那么返回该数值
      elif  len(i)  ==  1:
            avg_work_year.append(int(''.join(i)))
      #  如果匹配值为一个区间,那么取平均值
      else:
            num_list  =  [int(j)  for  j  in  i]
            avg_year  =  sum(num_list)/2
            avg_work_year.append(avg_year)
df['工作经验']  =  avg_work_year
#  将字符串转化为列表,再取区间的前25%，比较贴近现实
df['salary']  =  df['工资'].str.findall(pattern)
#  月薪
avg_salary  =  []
for  k  in  df['salary']:
      int_list  =  [int(n)  for  n  in  k]
      avg_wage  =  int_list[0]+(int_list[1]-int_list[0])/4
      avg_salary.append(avg_wage)
df['月工资']  =  avg_salary
#  将学历不限的职位要求认定为最低学历:大专\
df['学历要求']  =  df['学历要求'].replace('不限','大专')
```

数据通过简单的清洗之后，下面开始我们的统计

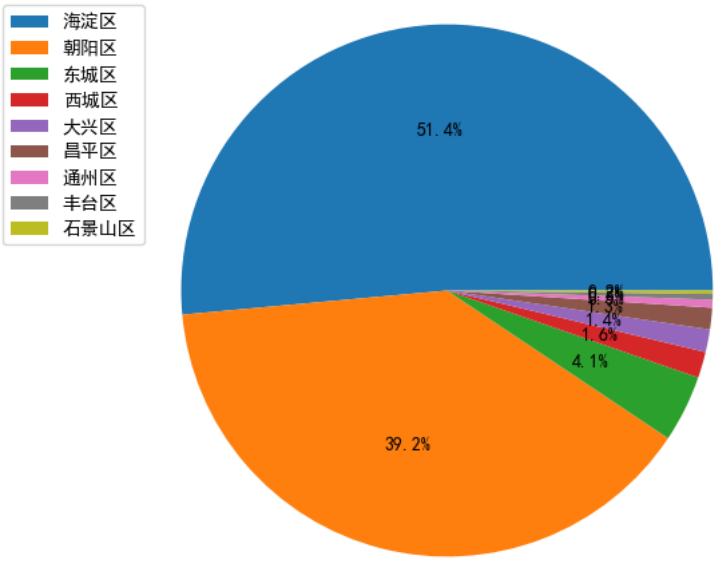
```
#  绘制频率直方图并保存
plt.hist(df['月工资'])
plt.xlabel('工资  (千元)')
plt.ylabel('频数')
plt.title("工资直方图")
plt.savefig('薪资.jpg')
plt.show()
```



结论：北京市Python开发的薪资大部分处于15~25k之间

```
#  绘制饼图并保存
count  =  df['区域'].value_counts()
plt.pie(count,  labels  =  count.keys(),labeldistance=1.4,autopct='%2.1f%%')
plt.axis('equal')     #  使饼图为正圆形
plt.legend(loc='upper  left',  bbox_to_anchor=(-0.1,  1))
```

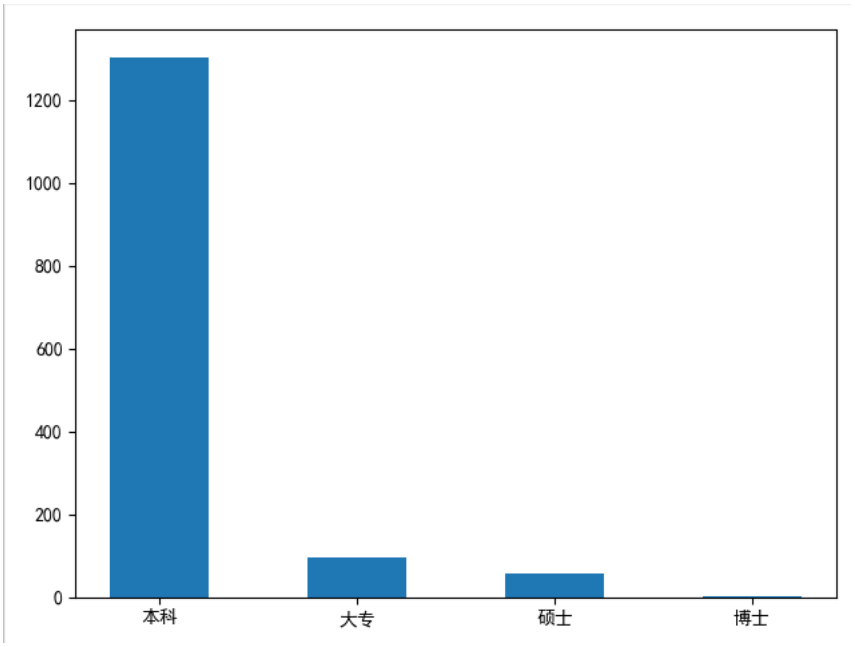
```
plt.savefig('pie_chart.jpg')
plt.show()
```



结论：Python开发的公司最多的是海淀区、其次是朝阳区。准备去北京工作的小伙伴大概知道去哪租房了吧

```
# {'本科': 1304, '大专': 94, '硕士': 57, '博士': 1}
dict = {}
for i in df['学历要求']:
    if i not in dict.keys():
        dict[i] = 0
    else:
        dict[i] += 1
index = list(dict.keys())
print(index)
num = []
for i in index:
    num.append(dict[i])
print(num)
plt.bar(left=index, height=num, width=0.5)
```

```
plt.show()
```



结论：在Python招聘中，大部分公司要求是本科学历以上。但是学历只是个敲门砖，如果努力提升自己的技术，这些都不是事儿

```
#  绘制词云,将职位福利中的字符串汇总
text  =  ''
for  line  in  df['职位福利']:
      text  +=  line
#  使用jieba模块将字符串分割为单词列表
cut_text  =  ' '.join(jieba.cut(text))
#color_mask  =  imread('cloud.jpg')    #设置背景图
cloud  =  WordCloud(
        background_color  =  'white',
        #  对中文操作必须指明字体
        font_path='yahei.ttf',
        #mask  =  color_mask,
        max_words  =  1000,
        max_font_size  =  100
```

```
        ).generate(cut_text)

#  保存词云图片

cloud.to_file('word_cloud.jpg')

plt.imshow(cloud)

plt.axis('off')

plt.show()
```

结论：弹性工作是大部分公司的福利，其次五险一金少数公司也会提供六险一金。团队氛围、扁平化管理也是很重要的一方面。

至此，此次分析到此结束。有需要的同学也可以查一下其他岗位或者地区的招聘信息哦~

希望能够帮助大家定位自己的发展和职业规划。

- End -

文章转载自公众号

程序员共成长

程序员共成长