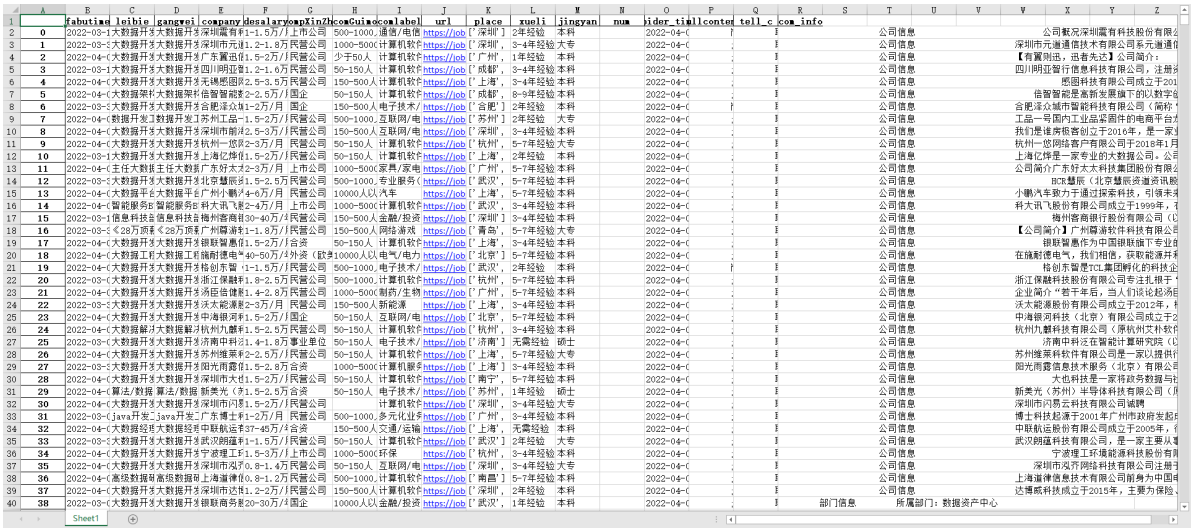
数据预处理

关于数据预处理的步骤解释说明

首先一开始数据是很乱的，所以我们需要一步步进行数据处理，确保数据的整洁



	A	fabutime	leibie	gangwei	company	desalary	wenZhicon	Guimo	com_label	url	place	xueli	jingyan	num	jider_tit	llconten	tell_c	com_info
1																		
2	0	2022-03-1	大数据开发	大数据开发	深圳震有	1.5-1.5万	上市公司	500-1000	通信/电信	https://job	['深圳']	2年经验	本科		2022-04-(			公司信息
3	1	2022-03-1	大数据开发	大数据开发	深圳市元	2.1-2.8万	民营公司	1000-5000	计算机软件	https://job	['深圳',	3-4年经验	大专		2022-04-(			公司信息
4	2	2022-04-(	大数据开发	大数据开发	广东翼迅	1.5-2万/月	民营公司	少于50人	计算机软件	https://job	['广州',	1年经验	本科		2022-04-(			公司信息
5	3	2022-03-1	大数据开发	大数据开发	四川明亚	2.1-2.6万	民营公司	50-150人	计算机软件	https://job	['成都',	3-4年经验	本科		2022-04-(			公司信息
6	4	2022-04-(	大数据开发	大数据开发	无锡思强	2.5-3.5万	民营公司	150-500人	计算机软件	https://job	['上海',	3-4年经验	本科		2022-04-(			公司信息
7	5	2022-04-(	大数据开发	大数据架构	信智智能	2-2.5万/月	国企	50-150人	计算机软件	https://job	['成都',	8-9年经验	本科		2022-04-(			公司信息
8	6	2022-03-1	大数据开发	大数据开发	合肥泽众	1-2万/月	国企	150-500人	电子技术/	https://job	['合肥']	2年经验	本科		2022-04-(			公司信息
9	7	2022-04-(	数据开发工	数据开发工	苏州工品	1.5-2万/月	民营公司	500-1000	互联网/电	https://job	['苏州']	2年经验	大专		2022-04-(			公司信息
10	8	2022-04-(	大数据开发	大数据开发	深圳市前	2.5-3万/月	民营公司	150-500人	互联网/电	https://job	['深圳',	3-4年经验	大专		2022-04-(			公司信息
11	9	2022-04-(	大数据开发	大数据开发	杭州一悠	2-3万/月	民营公司	500-1000	计算机软件	https://job	['杭州']	5-7年经验	大专		2022-04-(			公司信息
12	10	2022-03-1	大数据开发	大数据开发	上海亿怖	1.5-2万/月	民营公司	50-150人	计算机软件	https://job	['上海',	2年经验	本科		2022-04-(			公司信息
13	11	2022-04-(	主任大数据	主任大数据	广东好太	2-3万/月	上市公司	1000-5000	家具/家电	https://job	['广州',	5-7年经验	本科		2022-04-(			公司信息
14	12	2022-03-1	大数据开发	大数据开发	北京慧眼	1.5-2.5万	民营公司	500-1000	专业服务(	https://job	['武汉',	5-7年经验	本科		2022-04-(			公司信息
15	13	2022-04-(	大数据平台	大数据平台	广州小鹏	4-6万/月	民营公司	10000人以	汽车	https://job	['上海',	5-7年经验	本科		2022-04-(			公司信息
16	14	2022-04-(	智能服务E	智能服务E	科大讯飞	2-4万/月	上市公司	1000-5000	计算机软件	https://job	['武汉',	3-4年经验	本科		2022-04-(			公司信息
17	15	2022-03-1	信息科技主	信息科技主	梅州客商	30-40万/4	民营公司	150-500人	金融/投资	https://job	['深圳']	3-4年经验	本科		2022-04-(			公司信息
18	16	2022-03-3	≤28万顶薪	≤28万顶薪	广州尊源	1-1.8万/月	民营公司	150-500人	网络游戏	https://job	['青岛',	5-7年经验	大专		2022-04-(			公司信息
19	17	2022-04-(	大数据开发	大数据开发	银联智惠	1.5-2万/月	合资	50-150人	计算机软件	https://job	['上海',	3-4年经验	本科		2022-04-(			公司信息
20	18	2022-04-(	大数据工程	施耐德电气	40-50万/4	外资（欧美		10000人以	电气/电力	https://job	['北京']	5-7年经验	本科		2022-04-(			公司信息
21	19	2022-04-(	大数据开发	大数据开发	格创东智	1-1.5万/月	民营公司	500-1000	电子技术/	https://job	['武汉',	2年经验	本科		2022-04-(			公司信息
22	20	2022-03-1	大数据开发	大数据开发	浙江保融	1.8-2.5万	民营公司	500-1000	计算机软件	https://job	['杭州',	5-7年经验	本科		2022-04-(			公司信息
23	21	2022-04-(	大数据开发	大数据开发	汤臣倍健	1.4-2.8万	民营公司	1000-5000	制药/生物	https://job	['广州',	5-7年经验	本科		2022-04-(			公司信息
24	22	2022-03-1	大数据开发	大数据开发	沃太能源	2-3万/月	民营公司	150-500人	新能源	https://job	['上海',	3-4年经验	本科		2022-04-(			公司信息
25	23	2022-04-(	大数据开发	大数据开发	中海银河	1.5-2万/月	国企	50-150人	互联网/电	https://job	['北京',	5-7年经验	本科		2022-04-(			公司信息
26	24	2022-04-(	大数据解决	大数据解决	杭州九麟	1.5-2.5万	民营公司	50-150人	计算机软件	https://job	['杭州',	3-4年经验	本科		2022-04-(			公司信息
27	25	2022-03-1	大数据开发	大数据开发	济南中科	1.4-1.8万	事业单位	50-150人	电子技术/	https://job	['济南']	无需经验	硕士		2022-04-(			公司信息
28	26	2022-04-(	大数据开发	大数据开发	苏州维莱	2-2.5万/月	民营公司	50-150人	计算机软件	https://job	['上海',	5-7年经验	大专		2022-04-(			公司信息
29	27	2022-03-1	大数据开发	大数据开发	阳光雨露	1.5-2.8万	合资	1000-5000	计算机服务	https://job	['上海']	3-4年经验	本科		2022-04-(			公司信息
30	28	2022-04-(	大数据开发	大数据开发	深圳市大	1.5-2万/月	民营公司	50-150人	计算机软件	https://job	['南宁',	5-7年经验	本科		2022-04-(			公司信息
31	29	2022-04-(	算法/数据	算法/数据	新美光（	1.5-2.5万	合资	50-150人	电子技术/	https://job	['苏州',	1年经验	硕士		2022-04-(			公司信息
32	30	2022-04-(	大数据开发	大数据开发	深圳市闪	1.5-2万/月	民营公司		计算机软件	https://job	['深圳',	3-4年经验	大专		2022-04-(			公司信息
33	31	2022-03-(	java开发工	java开发工	广东博士	1-2万/月	民营公司	500-1000	多元化业务	https://job	['广州',	3-4年经验	本科		2022-04-(			公司信息
34	32	2022-04-(	大数据经理	大数据经理	中联航运	37-45万/4	合资	150-500人	交通/运输	https://job	['上海',	无需经验	本科		2022-04-(			公司信息
35	33	2022-03-1	大数据开发	大数据开发	武汉朗蓝	1-1.5万/月	民营公司	50-150人	计算机软件	https://job	['武汉']	2年经验	大专		2022-04-(			公司信息
36	34	2022-04-(	大数据开发	大数据开发	宁波理工	1.5-3万/月	上市公司	1000-5000	环保	https://job	['杭州',	3-4年经验	本科		2022-04-(			公司信息
37	35	2022-04-(	大数据开发	大数据开发	深圳市泓	0.8-1.4万	民营公司	50-150人	互联网/电	https://job	['深圳',	3-4年经验	大专		2022-04-(			公司信息
38	36	2022-04-(	高级数据研	高级数据研	上海道傅	0.8-1.2万	民营公司	500-1000	计算机软件	https://job	['南昌']	5-7年经验	本科		2022-04-(			公司信息
39	37	2022-04-(	大数据开发	大数据开发	深圳市达	1.2-2万/月	民营公司	150-500人	计算机软件	https://job	['深圳',	2年经验	本科		2022-04-(			公司信息
40	38	2022-03-(	大数据开发	大数据开发	银联商务	20-30万/4	国企	10000人以	金融/投资	https://job	['武汉',	1年经验	本科		2022-04-(	部门信息	所属部门：数据资产中心	

Sheet1

我们一步步来做，先从简单的开始

首先我们需要对城市这一列进行数据清洗，把一些多余的字符串，以及城市所在的区删去，这些数据都是无效数据

```
def chuli_city(x):
    x = str(x)
    x = x.split(',')
    x = x[0]
    x = x.replace("'",'').replace("[",'').replace("]",'')
    return x

df1['place'] = df1['place'].apply(chuli_city)
```

这一步就是删除无效数据，

接着把无用的行删除，像num就是空列不需要，所以也是删除

```
df1 = df1.drop(['num','Unnamed: 0'],axis=1)
```

然后对后面三列，进行清洗，因为有很多空白处，这些都是我们需要删除的，并且存在很多缺失值，所以也是要把缺失值全部删除的，并且判断是否为中文，如果不是中文也是全部删除

```python
def chuli_fullcontent(x):
    def str_count(str):
        count_en = count_dg = count_sp = count_zh = count_pu = 0
        for s in str:
            if s in string.ascii_letters:
                count_en += 1
            elif s.isdigit():
                count_dg += 1
            elif s.isspace():
                count_sp += 1
            elif s.isalpha():
                count_zh += 1
            else:
                count_pu += 1
        return count_zh
    x = str(x)
    x1 = x.replace('_x000D_\n',_).replace(' ',_).strip(' ;')
    x1 = x1.replace(' ',_)
    count = str_count(x1)
    if count >= 10:
        return x1
    else:
        return np.NAN
```

这个代码就是对数据进行一个清洗工作，做好上面的步骤之后

然后我们再回到城市分类的问题，我们把城市划分为一线城市，新一线城市，二线城市，其他城市

```python
def city_type(x):
    if '北京' in x or '深圳' in x or '广州' in x or '上海' in x:
        return '一线城市'
    if '天津' in x or '成都' in x or '南京' in x or '西安' in x or '重庆' in x or '长沙' in x or '杭州' in x or '武汉' in x or '苏州' in x or '合肥' in x or '沈阳
        return '新一线城市'
    if '合肥' in x or '福州' in x or '泉州' in x or '厦门' in x or '兰州' in x or '贵阳' in x or '珠海' in x or '惠州' in x or '中山' in x or '南宁' in x or '石家
        return '二线城市'
    if '太原' in x or '昆明' in x or '嘉兴' in x or '金华' in x or '绍兴' in x or '台州' in x or '温州' in x:
        return '二线城市'
    else:
        return '其他城市'
```

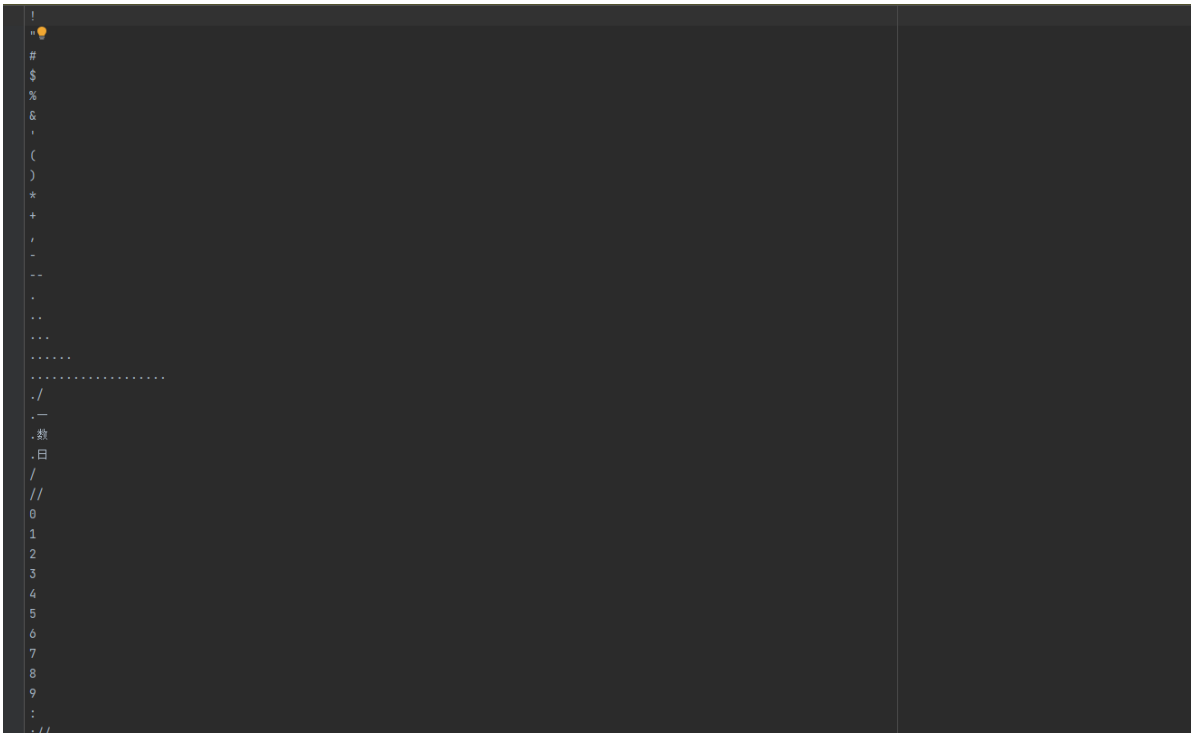
把上面的步骤全部整理好之后我们把整理好的数据，保存为一个新的表格

原始数据为2128行

整理好的数据为1802行，从这里可以看出还是删了挺多无效数据的

关于分词这边

首先，关于数据，会有很多无意义的词，所以我们需要准备一个停用词



停用词的作用就是帮助我们把无效的字符全部删除

删除好无效词之后，

我们再进行筛选，2个词以上的我们才保留下来，不然一个词太多了，这样也没什么意义

```
# 分词
word_num = jieba.lcut(content_series.str.cat(sep='。'), cut_all=False)

# 条件筛选
word_num_selected = [i for i in word_num if i not in stop_words and len(i) >= 2]
return word_num_selected
```

然后再筛选出前100个词生成词云

```
# 绘制词云图
text1 = get_cut_words(content_series=df['fullcontent'])
stylecloud.gen_stylecloud(text=','.join(text1), max_words=100,
```

最后这个便是结果图了

关于聚类的说明步骤

首先就是对原来的内容进行分词处理

```
# 对原文本分词
def cut_words():
    # 获取当前文件路径
    df = pd.read_excel('招聘数据.xlsx').loc[:,['fullcontent']]
    text1 = df.astype('str').values
    content = ''
    for t in text1:
        text = jieba.cut(t[0], cut_all=False)
        for i in text:
            content += i
            content += " "
        content += "\n"
    return content
```

把分词处理好之后，生成一个分词的文本后面要用到

这些是对应的分好词的文本

接着再把这些词进行一个tf-idf的计算

```
# 读取预料 一行预料为一个文档
for line in open('数据.txt', 'r',encoding='utf-8').readlines():
    corpus.append(line.strip())
# 将文本中的词语转换为词频矩阵 矩阵元素a[i][j] 表示j词在i类文本下的词频
vectorizer = CountVectorizer()

# 该类会统计每个词语的tf-idf权值
transformer = TfidfTransformer()

# 第一个fit_transform是计算tf-idf 第二个fit_transform是将文本转为词频矩阵
tfidf = transformer.fit_transform(vectorizer.fit_transform(corpus))
# 获取词袋模型中的所有词语
word = vectorizer.get_feature_names()

# 将tf-idf矩阵抽取出来 元素w[i][j]表示j词在i类文本中的tf-idf权重
weight = tfidf.toarray()

# 打印特征向量文本内容
print('Features length: ' + str(len(word)))
```

去查看每个词tf-idf的权重，计算好之后用k-means进行聚类

```
from sklearn.cluster import KMeans

clf = KMeans(n_clusters=4)
print(clf)
pre = clf.fit_predict(weight)
df = pd.read_excel('招聘数据.xlsx')
result = pd.concat((df, pd.DataFrame(pre)), axis=1)
result.rename({0: '聚类结果'}, axis=1, inplace=True)
result.to_excel('招聘数据-聚类.xlsx')
print(pre)
#
# 中心点
print(clf.cluster_centers_)
print(clf.inertia_)
```

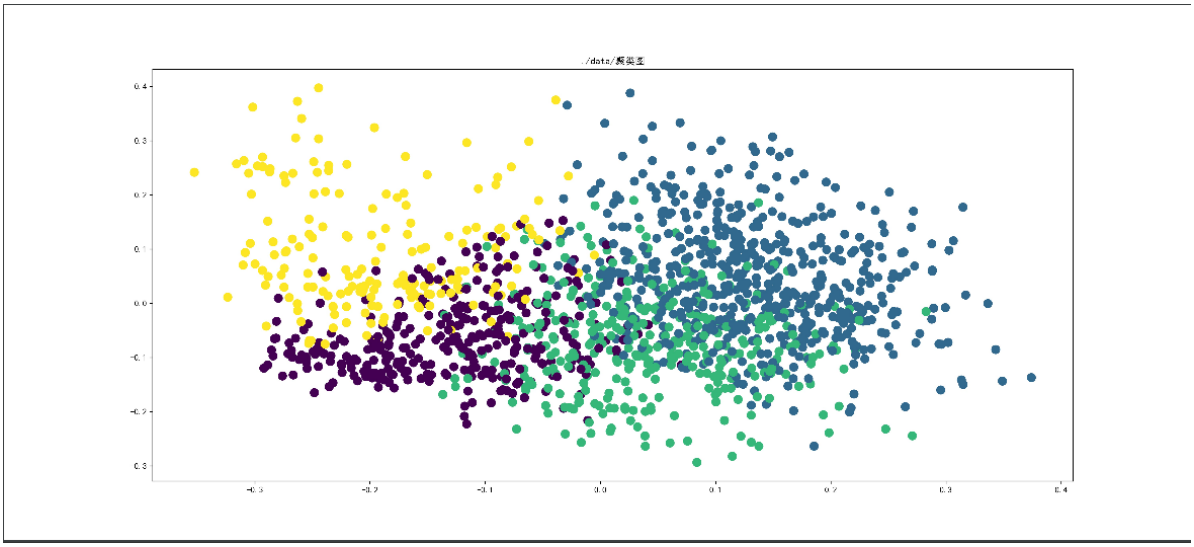
聚好的结果，我们把它保存起来并且生成一个新的文档

接着我们再进行一个pca降维的处理，为了后面的图把每一类进行区分出来

```
pca = PCA(n_components=4)  # 输出两维
newData = pca.fit_transform(weight)  # 载入N维
print(newData)

x = [n[0] for n in newData]
y = [n[1] for n in newData]
plt.rcParams['font.sans-serif'] = ['SimHei']  # 支持中文
plt.rcParams['axes.unicode_minus'] = False
plt.figure(figsize=(20,9),dpi=300)
plt.scatter(x, y, c=pre, s=100)
# plt.legend()
plt.title("./data/聚类图")
plt.savefig('./data/聚类图.jpg')
plt.show()
```

最后降维好的图形如下



基本可以看出每一类差不多都聚在一起，不过因为数据量的原因，所以还是比较密集的