

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	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	0	substitue	jobin	gangeng	company	desulargoxpin	zico	Guia	co	label	url	place	xueli	jingyan	num	hider	tiail	conten	tel	c	com	info			
2	0	2022-03-(大数据开发工程师)	深圳	有研1-1.5万/月	上市公司	500-1000	通信/电信	https://job			深圳	2年经验	本科												
3	1	2022-03-(大数据开发工程师)	深圳市元	1.2-1.8万	民营企业	1000-5000	计算机软	https://job			深圳	3-4年经验	大专												
4	2	2022-04-(大数据开发工程师)	广东	1.5-2万/月	民营企业	少于50人	计算机软	https://job			广州	1年经验	本科												
5	3	2022-03-(大数据开发工程师)	四川	1.2-1.6万	民营企业	50-150人	计算机软	https://job			成都	3-4年经验	本科												
6	4	2022-04-(大数据开发工程师)	无锡	2.5-3.5万	民营企业	150-500人	计算机软	https://job			上海	3-4年经验	本科												
7	5	2022-04-(大数据开发工程师)	信智智能	2-2.5万/月	国企	50-150人	计算机软	https://job			成都	6-9年经验	本科												
8	6	2022-03-(大数据开发工程师)	合肥	2-2万/月	国企	150-500人	电子技术	https://job			合肥	2年经验	本科												
9	7	2022-04-(大数据开发工程师)	苏州	1.5-2万/月	民营企业	500-1000	互联网/电	https://job			苏州	2年经验	大专												
10	8	2022-04-(大数据开发工程师)	深圳	2-3万/月	民营企业	150-500人	互联网/电	https://job			深圳	3-4年经验	本科												
11	9	2022-04-(大数据开发工程师)	杭州	2-3万/月	民营企业	50-150人	计算机软	https://job			杭州	5-7年经验	大专												
12	10	2022-03-(大数据开发工程师)	上海	2-3万/月	民营企业	50-150人	计算机软	https://job			上海	2年经验	本科												
13	11	2022-04-(主任/数据主任)	大数据	2-3万/月	上市公司	1000-5000	数据/深	https://job			广州	5-7年经验	本科												
14	12	2022-03-(大数据开发工程师)	北京	2-3万/月	民营企业	500-1000	专业服务	https://job			武汉	5-7年经验	本科												
15	13	2022-04-(大数据开发工程师)	广州	4-6万/月	民营企业	10000人以上	汽车	https://job			上海	5-7年经验	本科												
16	14	2022-04-(智能服务/智能服务)	科大讯飞	2-4万/月	上市公司	1000-5000	计算机软	https://job			武汉	3-4年经验	本科												
17	15	2022-03-(信息科技/信息科技)	台州	30-40万/月	民营企业	150-500人	金融/融资	https://job			深圳	3-4年经验	本科												
18	16	2022-03-(C2B/C2B)	南京	1-1.8万/月	民营企业	150-500人	网络/游戏	https://job			南京	5-7年经验	本科												
19	17	2022-04-(大数据开发工程师)	银联	1.5-2万/月	国企	50-150人	计算机软	https://job			上海	3-4年经验	本科												
20	18	2022-04-(大数据开发工程师)	施耐德	40-50万/月	外资	(数据)10000人以上	电气/电力	https://job			北京	5-7年经验	本科												
21	19	2022-04-(大数据开发工程师)	信创东	1-1.5万/月	民营企业	500-1000	电子技术	https://job			武汉	2年经验	本科												
22	20	2022-03-(大数据开发工程师)	信创东	1.5-2万/月	民营企业	500-1000	计算机软	https://job			杭州	5-7年经验	本科												
23	21	2022-04-(大数据开发工程师)	汤臣倍健	1.4-2.8万	民营企业	1000-5000	制药/生物	https://job			广州	3-4年经验	本科												
24	22	2022-03-(大数据开发工程师)	沃太	2-3万/月	民营企业	150-500人	新能源	https://job			上海	3-4年经验	本科												
25	23	2022-04-(大数据开发工程师)	中海	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												
26	24	2022-04-(大数据开发工程师)	杭州	1.5-2万/月	民营企业	50-150人	计算机软	https://job			杭州	3-4年经验	本科												
27	25	2022-03-(大数据开发工程师)	济南	4-1.8万	事业单位	50-150人	电子技术	https://job			济南	无经验	硕士												
28	26	2022-04-(大数据开发工程师)	苏州	2-2.5万/月	民营企业	50-150人	计算机软	https://job			上海	5-7年经验	大专												
29	27	2022-03-(大数据开发工程师)	阳光	1.5-2.8万	合资	1000-5000	计算机软	https://job			上海	3-4年经验	本科												
30	28	2022-04-(大数据开发工程师)	深圳	1.5-2万/月	民营企业	50-150人	计算机软	https://job			南宁	5-7年经验	本科												
31	29	2022-04-(算法/数据/数据)	新奥	1.5-2.5万	合资	50-150人	电子技术	https://job			苏州	1年经验	硕士												
32	30	2022-04-(大数据开发工程师)	深圳	1.5-2万/月	民营企业	500-1000	多元化业务	https://job			深圳	3-4年经验	大专												
33	31	2022-03-(java开发/java开发)	广东	1-2万/月	民营企业	500-1000	多元化业务	https://job			广州	3-4年经验	本科												
34	32	2022-04-(大数据开发工程师)	中	37-40万/月	合资	150-500人	交通/运输	https://job			上海	无经验	本科												
35	33	2022-03-(大数据开发工程师)	武汉	1-1.5万/月	民营企业	50-150人	计算机软	https://job			武汉	2年经验	大专												
36	34	2022-04-(大数据开发工程师)	宁波	1.5-3万/月	上市公司	1000-5000	环保	https://job			杭州	3-4年经验	本科												
37	35	2022-04-(大数据开发工程师)	深圳	7.0-8-1.4万	民营企业	50-150人	互联网/电	https://job			深圳	3-4年经验	大专												
38	36	2022-04-(高级数据/高级数据)	上海	2-3万/月	民营企业	500-1000	计算机软	https://job			南昌	5-7年经验	本科												
39	37	2022-04-(大数据开发工程师)	深圳	1.2-2万/月	民营企业	150-500人	计算机软	https://job			深圳	2年经验	本科												
40	38	2022-03-(大数据开发工程师)	银联	30-30万/月	国企	10000人以上	金融/融资	https://job			武汉	1年经验	本科												

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

```
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次清洗工作，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

```
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' in x:
        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 '石家庄' in x:
        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

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
2096	2094	2022-04-C	大数据产业大数据产业空间规划	11.1-1.8万/	民营企业	50-150人	互联网/电	https://job/	【重庆】	5-7年经验	本科			2022-04-C						公司信息		空间规划	【重庆】	科技股份有限公司	20	
2097	2095	2022-03-C	医疗大数据医疗大数据	深圳市岩1-2万/月	民营企业	少于50人	医疗设备/	https://job/	【深圳】	3-4年经验	本科			2022-04-C						公司信息		深圳市岩向科技	【yashu】	是全球领先		
2098	2096	2022-04-C	大数据产业大数据产业字节跳动	2-4万/月	民营企业	10000人	以互联网/电	https://job/	【北京】	3-4年经验	本科			2022-04-C						公司信息		字节跳动成立于2012年3月，公司使命：				
2099	2097	2022-04-C	大数据产业大数据产业成都博泰	1.8-2千/月	民营企业	150-500人	计算机软	https://job/	【武汉】	在校生/应	本科			2022-04-C						公司信息		武汉新加坡科技有限公司	是厦门	是领先		
2100	2098	2022-03-C	大数据产业大数据产业成都博泰	1.8-2千/月	民营企业	10000人	以计算机软	https://job/	【深圳】	3-4年经验	大专			2022-04-C						公司信息		博泰科技	【深文】	上市公司，股票代码		
2101	2099	2022-04-C	大数据产业大数据产业广州博泰	0.8-1.1万/	民营企业	50-150人	计算机软	https://job/	【北京】	2年经验	大专			2022-04-C						公司信息		广州博泰企业				
2102	2100	2022-03-C	大数据产业大数据产业广州博泰	1.2-2万/月	民营企业	少于50人	专业服务	https://job/	【广州】	2年经验	本科			2022-04-C						公司信息		广州博泰企业				
2103	2101	2022-04-C	大数据产业大数据产业深圳博泰	1.8-2.5万/	上市公司	1000-5000	计算机软	https://job/	【深圳】	3-4年经验	大专			2022-04-C						部门信息		所属部门：服务与外包事业部		【凌佳科技集团简介】		
2104	2102	2022-03-C	大数据产业大数据产业济南博泰	0.4-1万/	民营企业	50-150人	计算机软	https://job/	【济南】	1年经验	大专			2022-04-C						公司信息		所属部门：服务与外包事业部		【凌佳科技集团简介】		
2105	2103	2022-04-C	大数据产业大数据产业民航电信	1.2-2.4万/	国企	150-500人	通信/电	https://job/	【北京】	在校生/应	硕士			2022-04-C						公司信息		公司简介：民航电信开发有限责任公司		凌佳科		
2106	2104	2022-04-C	大数据产业大数据产业深圳博泰	1.8-2万/月	上市公司	1000-5000	计算机软	https://job/	【成都】	3-4年经验	大专			2022-04-C						部门信息		所属部门：服务与外包事业部		【凌佳科技集团简介】		
2107	2105	2022-04-C	CDU大数据CDU大数据	上海博泰2.8-3万/月	国企	150-500人	计算机软	https://job/	【深圳】	10年以上	本科			2022-04-C						公司信息		公司介绍上海博泰信息技术有限公司		凌佳科		
2108	2106	2022-04-C	大数据产业大数据产业上海博泰	0.8-1.1万/	国企	500-1000	通信/电	https://job/	【昆明】	3-4年经验	本科			2022-04-C						公司信息		大康移动通信设备有限公司	【以	凌佳科		
2109	2107	2022-03-C	大数据产业大数据产业广州博泰	1.5-30万/	国企	150-500人	专业服务	https://job/	【广州】	2年经验	本科			2022-04-C						公司信息		公司介绍广州人【集团】下属价值中心，		凌佳科		
2110	2108	2022-04-C	大数据产业大数据产业广州博泰	1.1-1.5万/	民营企业	少于50人	计算机软	https://job/	【天津】	在校生/应	大专			2022-04-C						公司信息		公司简介：凌佳科技集团		凌佳科		
2111	2109	2022-04-C	大数据产业大数据产业深圳博泰	1.2-2.5万/	上市公司	1000-5000	计算机软	https://job/	【上海】	3-4年经验	大专			2022-04-C						部门信息		所属部门：服务与外包事业部		【凌佳科技集团简介】		
2112	2110	2022-04-C	证券大数据证券大数据南京立	1.5-2万/月	民营企业	少于50人	金融/经	https://job/	【南京】	无经验	硕士			2022-04-C						公司信息		公司简介：凌佳科技集团		凌佳科		
2113	2111	2022-04-C	大数据产业大数据产业重庆博泰	1.2-1.5万/	民营企业	500-1000	计算机软	https://job/	【长沙】	3-4年经验	本科			2022-04-C						公司信息		重庆足下科技有限公司	是凌佳科	凌佳科		
2114	2112	2022-03-C	大数据产业大数据产业北京博泰	1.8-2千/	民营企业	50-150人	计算机软	https://job/	【武汉】	3-4年经验	本科			2022-04-C						公司信息		北京凌佳科技主要从事应用软件的开		凌佳科		
2115	2113	2022-04-C	大数据产业大数据产业山东博泰	0.8-1.2万/	民营企业	500-1000	计算机软	https://job/	【青岛】	在校生/应	本科			2022-04-C						公司信息		山东丰丰信息技术有限公司	是凌佳科	凌佳科		
2116	2114	2022-04-C	大数据产业大数据产业南京博泰	1.2-2.5万/	国企	少于50人	电气/电	https://job/	【南京】	1年经验	大专			2022-04-C						公司信息		南京丰丰信息技术有限公司	是凌佳科	凌佳科		
2117	2115	2022-04-C	大数据产业大数据产业深圳博泰	1.2-2万/	上市公司	1000-5000	计算机软	https://job/	【深圳】	3-4年经验	本科			2022-04-C						部门信息		所属部门：服务与外包事业部		【凌佳科技集团简介】		
2118	2116	2022-04-C	大数据产业大数据产业成都博泰	4.8-6千/	非盈利组织	10000人	以教育/培	https://job/	【成都】	2年经验	硕士			2022-04-C						公司信息		成都博泰酒店管理学院	【成都博泰酒店	凌佳科		
2119	2117	2022-04-C	大数据产业大数据产业上海博泰	10-15万/	国企	150-500人	计算机软	https://job/	【上海】	在校生/应	本科			2022-04-07 11:11:12												
2120	2118	2022-03-C	大数据产业大数据产业厦门博泰	1.4-2.5万/	民营企业	1000-5000	计算机软	https://job/	【宁波】	2年经验	本科			2022-04-C						公司信息		厦门博泰丰源软件技术有限公司	是凌佳科	凌佳科		
2121	2119	2022-04-C	大数据产业大数据产业深圳博泰	1.1-1.5万/	民营企业	少于50人	计算机软	https://job/	【杭州】	3-4年经验	本科			2022-04-C						公司信息		云歌信息技术有限公司	【江苏】	凌佳科		
2122	2120	2022-04-C	大数据产业大数据产业深圳博泰	1.8-3万/	民营企业	150-500人	互联网/电	https://job/	【西安】	5-7年经验	本科			2022-04-C						公司信息		深圳市云歌信息技术有限公司	【南京云歌	凌佳科		
2123	2121	2022-04-C	大数据产业大数据产业常州博泰	0.8-1千/	民营企业	50-150人	贸易/进	https://job/	【常州】	1年经验	大专			2022-04-C						公司信息		常州博泰进出口有限公司	是凌佳科	凌佳科		
2124	2122	2022-04-C	大数据产业大数据产业深圳博泰	0.8-1.2万/	外企	500-1000	服装/纺织	https://job/	【广州】	在校生/应	本科			2022-04-C						公司信息		中国太平洋行	【集团】	是凌佳科		
2125	2123	2022-04-C	大数据产业大数据产业深圳博泰	0.8-1.2万/	国企	500-1000	保险	https://job/	【上海】	5-7年经验	本科			2022-04-C						公司信息		中国平安保险	【集团】	是凌佳科		
2126	2124	2022-04-C	大数据产业大数据产业深圳博泰	0.8-1千/	民营企业	150-500人	计算机软	https://job/	【深圳】	无经验	大专			2022-04-C						公司信息		经营范围包括一般经营项目是：技术服		凌佳科		
2127	2125	2022-04-C	大数据产业大数据产业北京博泰	1.2-2万/	民营企业	10000人	以汽车	https://job/	【深圳】	3-4年经验	本科			2022-04-C						部门信息		所属部门：信息中心		凌佳科		
2128	2126	2022-03-C	大数据产业大数据产业深圳博泰	1.1-1.5万/	民营企业	500-1000	互联网/电	https://job/	【深圳】	2年经验	本科			2022-04-C						公司信息		深圳市一凌网络股份有限公司		凌佳科		
2129	2127	2022-04-C	大数据产业大数据产业南京博泰	0.8-1.2万/	民营企业	150-500人	计算机软	https://job/	【南京】	在校生/应	本科			2022-04-C						公司信息		南京森博科技股份有限公司	是凌佳科	凌佳科		
2130	2128	2022-04-C	大数据产业大数据产业深圳博泰	1.8-3万/	国企	10000人	以保险	https://job/	【深圳】	5-7年经验	本科			2022-04-C						部门信息		所属部门：金融银行科技中心		凌佳科		
2131																										
2132																										
2133																										
2134																										
2135																										

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


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



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

```
岗位职责 负责 数 平台 开发 架构 优化 项目 运营 支撑 职 资 验 计算机 相关 专业 学 年 数 运维 开发 工作 验 熟练 linux 开发 环境 熟练掌握 JAVA Shell Python 语言 熟 143
负责 公司 数分析 数 实时 线 处理 业务 开发 优化 工作 参 数 产品 架构设计 案 讨 负责 踪 解决 产品 研发 团队 客户 遇 产品 问题 负责 产品 系统 持续 进 优化 岗位职责 熟练 jav
岗位职责 NBSP 负责 数 数分析 务 日 数 探索 数 统计 数分析 工作 通 专项 数分析 效 发现问题 提 改善 建议 推动 解决 NBSP 负责 数 支撑 服务 接 业务 研发 需求 参 数 支撑 类 工作 持续 完
数 开发 工程师 岗位职责 基 Spark Hadoop 数 平台 ETL 建设 开发 维护 优化 协助 业务 数 进行 分析 建模 业务部门 数 化 运营 提供 支持 业务 需求 进行 数 产品 规划 设计 开
岗位 描述 根 业务 需求 进行 数 仓库 数 平台 环境 搭建 架构设计 程序开发 负责 线 数 数 采集 清洗 加载 数 层面 发现 支持 解决 业务 系统 面 问题 负责 分布式 批量 计算 分布式 存 计算 数 库
岗位职责 搭建 基 Hadoop Spark Flink 相关 技术 数 平台 产品 利 数 相关 技术 数 进行 分析 处理 负责 数 相关 项目 技术支持 服务 工作 负责 数 平台 性 参数 调整 优化 职 责
岗位职责 负责 数 处理 分析 平台 服务 框架 设计 开发 工作 协助 相关 业务 需求 分解 服务 接口 制 负责 业务 模型 抽象 数 模型 设计 开发 具备 数 技术 研究 力 进 数 技术 发展 职
NBSP 负责 统 数 数 采集 规范 建设 NBSP 负责 线 数 数 仓库 建设 NBSP 负责 数 平台 数 基础 建设 提高 数 服务 质量 NBSP 负责 提供 统 标准化 数 流 业务 支撑 业务 团队 敏捷 开发 适
工作 职责 NBSP 参 数 平台 建设 维护 NBSP 参 设计 数 数 仓库 模型 构建 分层 体系 元 数 理 核心 应 开发 NBSP 参 数 服务 开发 部门 提供 数 服务 岗位 求 NBSP 科 学 计算机 相关 专业
职位 描述 参 产品 系统 架构设计 数 库 设计 开发 计划 完成 模块 开发 元 测试 工作 完成 数 库 模块 设计 开发 参 产品 进行 代码 优化 功 维护 性 维护 需求 变 协助 系统 部 进行 系统 维
工作 职责 负责 构建 公司 实时 数 流 实时 数 分析 挖掘 系统 完成 面 业务 目标 数 实 时 流 模型 义 应 开发 负责 数 实 时 流 应 技术 实现 案 设计 实现 负责 数 采集 数 清洗 数 处理 数 库
岗位职责 科 计算机 相关 专业 少 年 数 架构设计 验 丰富 数 仓库 数 挖掘 机器 学习 项目 验 具备 丰富 型 分布式 数 库 设计 开发 部署 验 包括 主流 结构化 数 据 MySQL Postgresql Or
岗位职责 负责 数 平台 开发 建设 数 仓库 包括 数 接入 存储 应 数 类型 通 流 数 非 结构化 数 结构化 数 负责 数 库 监控 优化 数 计算 性 职 求 科 学 计算机 相关 专业 年 开发
岗位职责 负责 动 驾驶 数 平台 案 设计 构建 数 闭环 足 算法 开发 验证 仿真 测试 开发 数 采集 分析 视 化 标注 平台 工具 树 加速 算法 功 开发 负责 解决 攻克 数 系统 平台 技术 难题 岗位 求
参 项目 需求 分析 撰写 数 工作 计划 详细 设计 实施 NBSP 根 项目 核心 架构 负责 搭建 数 数 框架 框架 编写 系统 开发 环境 完成 系统 数 框架 核心 代码 实现 NBSP 负责 数 采集 清洗 整合
岗位职责 负责 行 数 风 控系统 相关 平台 日 运维 负责 行 风险 数 集市 风险 画 体系 建设 负责 行 相关 风险 外部 数 需求 沟通 接 测试 线 参 海量 数 处理 高 性 分布式 计算 架构 设计
岗位职责 研发 高 易 公司 统 数 平台 负责 公司 数 平台 相关 架构 开发 工作 指标 设计 数 建模 数 架构 架构 优化 数 质量 理 性 优化 数 处理 职 资 验 科 学 年 数 开 发 验
岗位职责 负责 广 告 业务 数 分析 平台 服务端 架构设计 性 优化 负责 广 告 业务 基 数 技术 开源 框架 选型 开发 负责 广 告 业务 数 挖掘 建模 相关 核心 算法 代码 实现 负责 广 告 业务 数 平台 业
岗位职责 负责 数 业务 平台 架构设计 搭建 基 Hadoop Spark TimescaleDB ETL 设计 开发 海量 数 进行 调 优 理解 业务 场景 产品 逻辑 分析 需求 规划 数 间 关联 关系 职 条件 计
岗位职责 NBSP 负责 数 平台 数 仓库 建设 ETL 开发 数 分析 NBSP 负责 数 实 时 计算 平台 业务 开发 NBSP 负责 数 平台 建设 维护 岗位 求 学 科 计算机 相关 专业 年 数 处理 研 发 验
岗位职责 参 数 类 底层 组件 平台 系统 开发 优化 参 数 相关 层 数 产品 应 开发 推广 公司 类 业务 开发 团队 提供 技术支持 跨 业界 技术 动态 推动 公司 数 相关 技术 持续 进 步
NBSP 数 平台 数 应 需求 开发 工作 NBSP 负责 理 优化 公司 核心 产品 算法 系统 断 迭代 优化 算法 效果 提升 核心 业务 指标 NBSP 负责 项目 进度 推 进 踪 模型 设计 数 清洗 数 转换 模型 开
参 数 架构 规划设计 参 业务 建模 数 化 运营 平台 搭建 运维 业务流程 数 结合 提 建议 解决 案 完成 基 数 平台 业务 项目 开发 实施 维护 工作 负责 解决 数 平台 建设 程 中 技术 问
岗位职责 负责 数 仓库 数 分析 数 挖掘 设计 建模 研 发 工作 负责 数 仓库 数 挖掘 设计 建模 研 发 工作 负责 数 仓库 数 挖掘 设计 建模 研 发 工作 负责 数 仓库 数 挖掘 设计 建模 研 发 工作
```

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_)
```

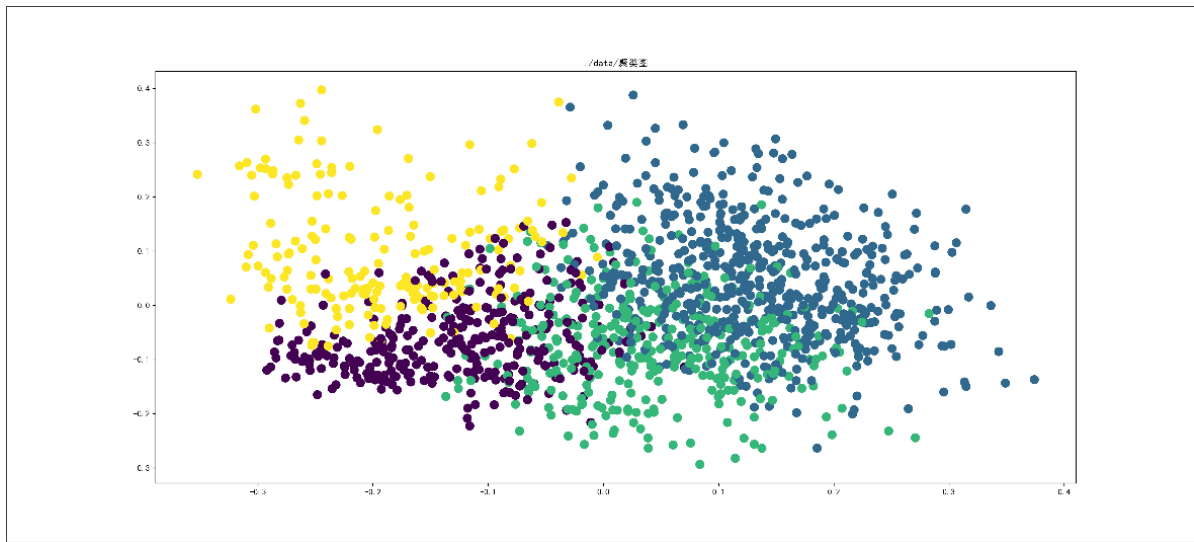
聚好的结果，我们把它保存起来并且生成一个新的文档

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