南京财经大学 数据挖掘报告

2019 —— 2020 第 二 学期

课程名称：数据挖掘

任课教师：韩伟

学生姓名：方泽强、奚宇星

班　　级：软件 1601

学　　号：2120162016、2120163203

报告题目：基于文本挖掘分析谷歌招聘信息

内容摘要：本文使用 Python 编程语言结合 Echarts 可视化工具，对谷歌一年内发布的招俜信息进行自然语言处理与文本挖掘，最后进行聚类分析，结合数据可视化结果给出针对谷歌的求职意见

关 键 词：词频；词云；聚类

基于文本挖掘分析谷歌招聘信息

我们先是在网上搜罗有趣的数据集，看到有人爬取了 2018 年谷歌在全球招聘的信息，里面包含了职位，类别，职责，最低条件和优先资格，这引起了笔者的注意。因为笔者正处周围同学都在找工作的氛围中（南财即将就业的大四学子），许多同学为了科技大厂不断的刷题，改简历去追求与自己较为对口的岗位，这让我产生一个想法：能不能用数据分析的方法提高自己进入某个企业的概率；故我们找到这个数据集，虽然不是最新的数据，但不失为一个模拟锻炼的机会。

一、 前期准备

1.数据集

共有 1250 像这样的数据，由于是英文，无需像中文那样进行复杂的分词工作，这就给我们文本挖掘带来很大的便利，不过要提前注意在词频统计阶段要剔除对数据分析无用的单词词，比如人称主语，停用词，数字。

job_skills.csv (1.79 MB) — 7 of 7 columns — Views

	Company Company name. Because the data set contain only Google jobs, the company field will always have value "Google"	Title The title of the job	Category Job Category	Location Location of the job	Responsibilities Responsibilities of the job	Minimum Qualifica Minimum qualifications for the job	Preferred Qualificat Preferred qualifications for the job
	Google 98% YouTube 2%	Business Intern 2... 3% MBA Intern, Sum... 3% Other (792) 94%	Sales & Account ... 13% Marketing & Co... 13% Other (21) 73%	Mountain View, ... 15% Sunnyvale, CA, ... 12% Other (90) 72%	Responsibilities a... 3% Google interns ar... 3% Other (800) 95%	Currently enrolle... 3% Must be pursuing... 3% Other (806) 95%	Previous internshi... 3% Interest in the tec... 3% Other (820) 95%
1	Google	Google Cloud Program Manager	Program Management	Singapore	Shape, shepherd, ship, and show technical programs designed to support the work of Cloud Customer Engineers and Solutions Architects. Measure and report on key metrics tied to those programs to identi...	BA/BS degree or equivalent practical experience. 3 years of experience in program and/or project management in cloud computing, enterprise software and/or marketing technologies.	Experience in the business technology market as a program manager in SaaS, cloud computing, and/or emerging technologies. Significant cross-functional experience across engineering, sales, and marketi...
2	Google	Supplier Development Engineer (SDE), Cable/Connector	Manufacturing & Supply Chain	Shanghai, China	Drive cross-functional activities in the supply chain for overall Technical Operational readiness in all NPI phases leading into mass production. Collaborate with suppliers and Engineering teams in	BS degree in an Engineering discipline or equivalent practical experience. 7 years of experience in Cable/Connector Design or Manufacturing in an NPI role. Experience working with Interconnect	BSEE, BSME or BSIE degree. Experience of using Statistics tools for Data analysis, e.g. distribution histogram/pareto chart, process control chart, Design of Experiment (DOE), Correlation Analysis, et...

图 1 数据集预览

2.实验工具

Python

针对这次数据挖掘的文本处理，使用 Python 编程语言处理会较为简洁优雅，其高拓展性也为数据挖掘后进行可视化提供了极大的方便性；Python 的扩展包 nltk 就是一个为自然语言处理而设计的工具，而另一个机器学习扩展包 sklearn 可以让我们有能力进行聚类分析，降低机器学习实现的门槛。

Echarts

可视化工具 Echarts 能给我们的文本分析带来极大助力。这是一个使用 JavaScript 实现的开源可视化库，可以流畅的运行在 PC 上，兼容当前绝大部分浏览器 Chrome，Firefox，Safari 等，底层依赖矢量图形库 ZRender，可以为使用者提供直观，交互丰富，可高度个性化定制的数据可视化图表。

数据预处理

1.去除标点符号

在数据集中的最低资格属性中内容都是以几句话的形式出现，故在步长一定的限定下提取单个单词，会收集到各种标点符号，这对我们的文本分析显然是种阻碍，如果不去标点符号，那么高频词中就会出现大量的标点符号。如下代码，我们需要先调用 python 的一个自然语言处理扩展包 nltk，使用正则匹配，为其规定如下一个范式，包含了所有在英文语句中可能出现的标点符号。然后在循环语句中去匹配相应的标点符号，将新的词干集合覆盖原先的初始数据。

```
pattern = re.compile(r'\s|,|\.|!|\?|:|;|\\|‘|’|\/')

for d in datas:

    t=[lemmatizer.lemmatize(x.lower()) for x in re.sub(pattern,' ',
```

```
d["Minimum Qualifications"]).split()]
    words_each.append(t)
    words+=t
```

2.去除停用词

此外在数据集中还会收集到各种停用词，如 The,We,There,Are...如果不去除停用词，那么高频词中就会出现大量停用词，影响文本分析的过程和数据可视化的最终视觉效果。故这里依然采用 nltk 扩展包，其自带 stopwords 英文停用词语料库，在最后一步将其停用词删去，并再次排序，以防词频顺序再被打乱。

```
from nltk.corpus import stopwords
sr = stopwords.words('english')
freq = nltk.FreqDist(words)
freq_sorted=[w for w in sorted(freq.items(),reverse=True,key=lambda
x:x[1]) if w[0] not in sr and len(w[0])>1]
```

3.工作地点标准化

我们统计出国家后发现有些相同国家的表示方法不同，比如美国 United States 和 USA 其实表示的都是美国，而 Dubai - United Arab Emirates 表示的是沙特阿拉伯，通过如下代码，我们使得如迪拜 Dubai - United Arab Emirates 变换为阿联酋 United Arab Emirates；United States\USA 统一变换成 United States；香港台湾等地区统一划归为中国。

```
country_counts['United States']+=country_counts.pop('USA')
country_counts['United Arab Emirates']=country_counts.pop('Dubai - United
Arab Emirates')
```

```
country_counts['China']+=country_counts.pop("Taiwan")+country_counts.pop("
Hong Kong")
```

再者,我们需要对统计后的词频进行排序，从而达到关注较高频率词的目的，因此采用自带的 sort 函数，通过 lambda 函数实现键值对按照其值排序。

```
country_counts_sorted=sorted(country_counts.items(),key=lambda
x:x[1],reverse=True)
```

文本挖掘

1. 工作地点

我们先对数据集里的工作地点进行词频统计，如图所示，只要国家名在工作地点的键中出现，则 country_counts 对应的国家的数目就加一，以此达到工作地点词频统计的效果。

```
for d in datas:
    country=d['Location'].split(",")[-1].strip()
    if country in country_counts.keys():
        country_counts[country]+=1
    else:
        country_counts[country]=1
```

再键入以下代码进行画图，将 country_counts 的键值对输入地图图表的对象中，最后渲染导出 html 文件，而绘出地理位置热点图可以在浏览器如 Chrome 中预览，并进行相应的数据分析了。

```
from pyecharts import Map
attr = country_counts.keys()
```

```
value = country_counts.values()

map0 = Map("工作地点", width=1200, height=600)

map0.add("工作地点", attr, value, maptype="world", is_visualmap=True,

visual_text_color='#000')

map0.render(path="工作地点.html")
```

我们词频统计的结果如图 2 所示，最终结果是一个 Python 中的字典类，一个国家对应自己的计数值，国家名称是字符串类型，数值是整型。

图 2 location 词频统计结果

通过图表数据可知，最明显的一种提高成功进入 google 机会就是申请就业机会比例较高的地区（国家）。没有任何悬念，google 公司当年在美国有 640 个职位发布远大于世界上其它地区。

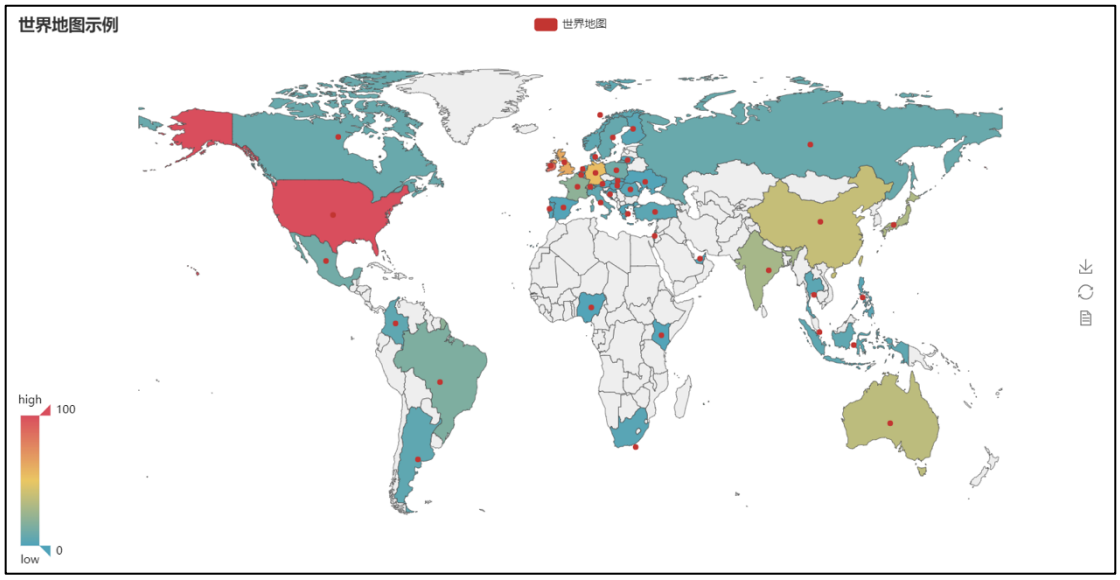

图 3 职位热点图（全部）

为了找出职位较多的区域，利用 Echarts 高交互性的特点，我们在图表中调节使用左下角的热力柱，找出热点区域，调整到 41 时发现地图上剩下 6 个国家，

美国、爱尔兰、英国、德国、新加坡和中国。

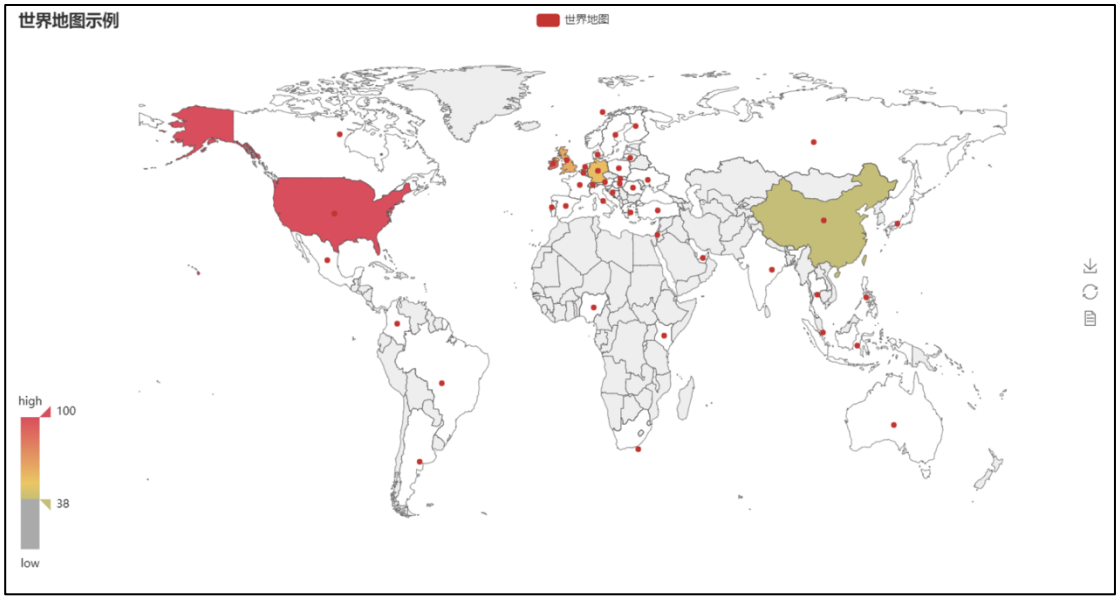

图 4 职位热点图（前 6）

按职位数量排序：美国（640）、爱尔兰（87）、中国（77）、英国（62）、德国（54）、新加坡（41）；其实我们还能发现，还有另一种提高进入 Google 公司的方法是去追求那些非科技中心但在 Google 就业机会中占很大比例的地区，比如爱尔兰，新加坡等不似硅谷那样高度集聚科技企业的地方。

2.工作类别

除了了解职位发布的位置，了解哪些团队获得职位发布的比例最大也是很有帮助的。这里工作类别的词频统计方法和上文提及的类似，如下也是运用循环结构，在工作类别字典类中匹配关键字技术。

```
for d in datas:
    category=d['Category'].strip()
    if category in category_counts.keys():
        category_counts[category]+=1
```

```
    else:
        category_counts[category]=1
```

不同的地方在于画图部分，我们选用的是 echarts 里的柱状图包 Bar，为了方便显示我们将横纵坐标旋转约 45 度，以便实现能容纳长文本。

```
bar = Bar("工作类别",width=1200,height=600)
attr=[x[0] for x in category_counts_sorted]+['']
values=[0]+[x[1] for x in category_counts_sorted]+[0]
bar.add("工作类别", attr, values, xaxis_interval=0, xaxis_rotate=15,
yaxis_rotate=45)
bar.render("rental.html")
```

如图 5 所示，我们发现销售&客户管理以及市场营销&沟通是一个这个公司招聘岗位里的大头，约占所有职位的 27%。但是，职位招聘最少的岗位是（基于职位发布数）：数据中心与网络、技术写作、IT 与数据管理、开发者关系、网络工程，偏重于技术开发，这里就能看出公司经费开销的比重。

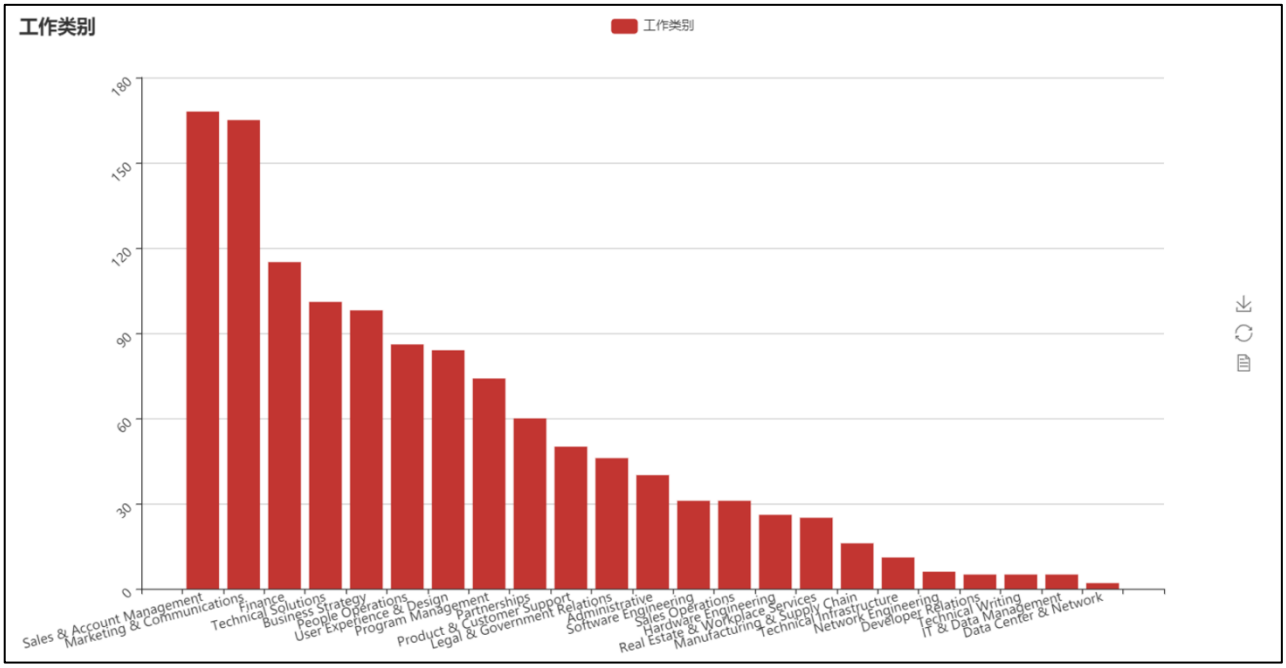

图 5 工作类别词频柱图

如果从该柱图中点向右看去，可以发现仍有较大比例的团队在工作范围上更具技术性。尽管 Google 是一家技术公司，但他们在招聘方面的重点是面对客户的机会和业务运营。从图表上看，针对的投递公司面向客户的业务岗可以一定程度上增加了获得 Google 面试的机会，然后增大被录用的机会，这将使得你在工作中有较快晋升并在以后获得你想进入的技术岗位。

3.岗位职责

除了根据工作位置和工作类别需求确定机会之外，该如何增加进入 Google 这种科技巨头公司的机会呢？其实我们在求职的时候都会听到别人说尽量将自己的简历与职位描述写得匹配。故接下来我们将对岗位职责进行文本挖掘，这可以帮助我们发现数据集中更重要的内容（而不仅仅是职位类别），使你能够改善简历以增加接到面试的几率。在 1,250 个职位发布的“岗位职责”即"Responsibility"部分上运行文本分析后，如图可以看到最常用的术语。有两个地方需要注意，首先，这些术语是词干，用于对相似的词进行分组（英文还是相对复杂的）。其次，散点图横坐标的数字是该术语被看到的总次数（总数）；而纵坐标数字是该术语所对应的职位数（职位数）。

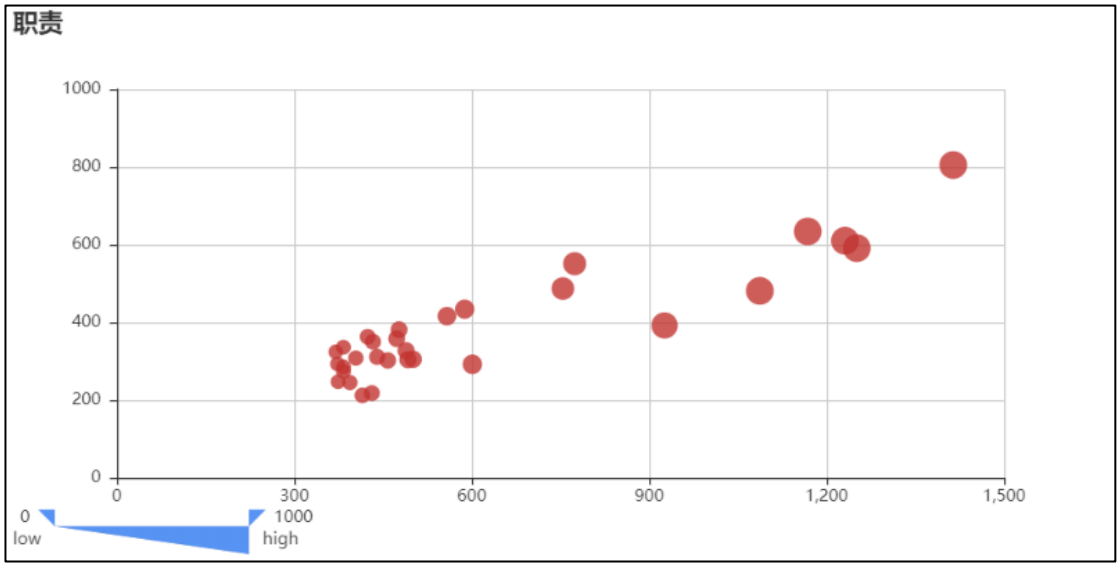

图 6 岗位职责词频散点分布状况

约 50%或更多的职位发布中出现了协作 team，开发 development，管理 management 和产品 product。其他值得注意的术语包括：支持 support，市场 market，过程 process，设计 design，项目 project，驱动 drive，解决方案 solution，辨识能力 identify 等。在这些领域中拥有经验非常重要；但最终目的是要让你的简历能够引起 HR 关注，只要将排名前几的术语囊括进来，无论职位发布的信息如何变化（在合理的范围内）极有可能提高你命中 Google 职位发布描述信息相匹配的几率，进而获得网申面试机会。

4.最低要求

除了作为职位申请人所应该匹配的职责岗位之外，了解其职业描述所对应的最低要求也很重要，这涉及到申请的底线，在简历中你也要仔细思量这些语句；在完成相应的词频统计后，我们编码实现词云分析。我们发现最低要求也非常的看重申请人的学历 degree，不仅如此，他还非常重视申请人的任职年限 year 等较为基础的信息；此外我们可以关注到业务分析能力 BA，表达流利度 fluently，岗位相关性 related 等第二梯队的高频词也可作为简历投递时参考的对象。

图 7 最低条件词云

虽然实际情况中最低要求往往会取决于取决于团队 team,地理位置 location 和工作级别 level。但经过分析后，数据集中大部分工作说明中都显示我们需要注意一些重要内容；在大约 80%的职位中，具有适当的实践经验 experience 和学位 degree 是最低要求。其实对于任何面临求职的人来说，这都不是什么新鲜的内容。而还有其他术语是较为新颖的，可以用来使得简历与众不同。

从分析来看，起码 Google HR 会寻找在书面和对话中都较为高效的人，流利的外语也会让你与众不同。对编程 programming，市场营销 marketing，工程学 engineering 的有一定理解以及在技术方面的经验可以使得你大大超过大部分职位中的最低要求。

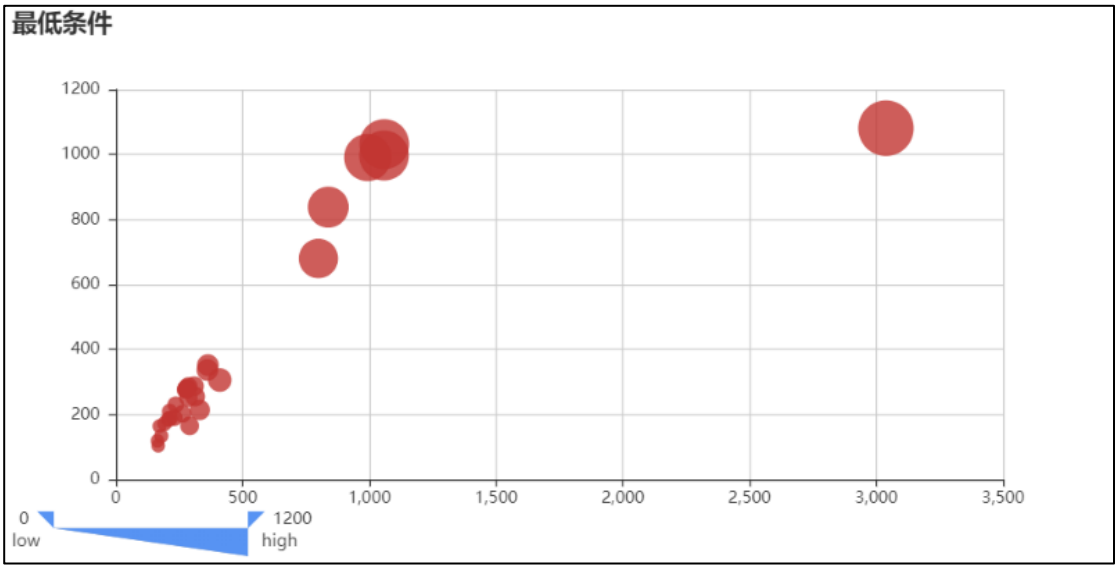

图 8 最低要求词频散点分布状况

5.优先资格

我分析的最后一部分是“优先资格”，沿着上文最低条件的分析思路，我们进行词频统计，得到高频词键值对通过 pyecharts 调用词云工具实现；通过对词云结果的分析，我们发现优先资格依然重视 degree,只不过这个条件会要求到硕士学位；而且优先资格会更加重视项目经历，同样的也更加看重管理能

力和其岗位领域专业技术技能的匹配度。

图 9 最优条件词云

您会注意到，其中许多术语与最低资格相似。那么，为什么有两个不同的部分？不难发现，雇主们是想在这里看到你所展现的独特能力与技能。他们希望你有项目经历 experience。但有趣的是，对于许多职位，他们都希望看到你是知识渊博的 knowledgeable，知道如何处理人际关系 relationships，在不同团队 teams 和环境 environments 中工作得很好；而且作为求职者你也要知道如何使用数据 data，甚至是如何使用 Microsoft Excel。其中很多都是和技术人员十分对口的，但要注意不要因为简单而没有在简历中提及这些关键词而错失面试的机会。

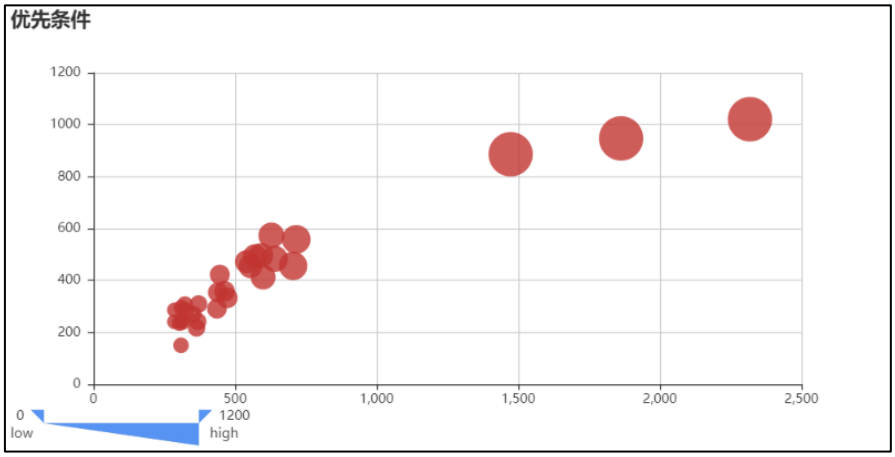

图 10 优先资格词频散点分布状况

四、聚类分析

我们希望挖掘工作类别之间的关系，用人工理解工作类别的方式，对职责中的高频词汇，取有效值进行人工分类，各取五个，先粗略的分为市场营销，管理决策，技术类；先对各个类别进行词频统计，跳过无关项。

```
rt={'管理决策':['manage','strategy','solution','management','lead'],
    '技术':['develop','technical','cloud','client','build'],
    '市场营销':['product','business','customer','sale','marketing']}
for i in rt.keys():
    tmp[i]=sum(map(lambda s:getCounts(s,freq),rt[i]))/len(t)
tmp['n']=category
```

如下代码对每个类别进行类似打分的操作，让三个维度的数值维持在 0~10 之间，方便平衡三个尺度在一个合理的范围内进行聚类分析。

```
   for i in scores:
       cs[i['n']][0]+=1
       cs[i['n']][1]+=i['管理决策']
       cs[i['n']][2]+=i['技术']
       cs[i['n']][3]+=i['市场营销']
```

最后调用 Python 的 sklearn 包直接实现聚类，先设置类簇为 3，再设置随机种子为 9，使得每一次相同条件的聚类结果相同，不过目前暂时不需要调整 kmeans 的参数来优化模型，kmeans 的聚类结果足够我们找出一些规律。

```
    from sklearn.cluster import KMeans
    for k,v in cs.items():
        kmeans=KMeans(n_clusters=3, random_state=9)
        res=kmeans.fit_predict(datasets)
```

如图 11 所示，我们发现偏重技术与市场营销的岗位如 Technical Solution 技术解决方案，销售运营 Sales Operations，销售与会计管理 Sales&Account Management，这些岗位联系较大，申请人在投递对应岗位的时候可以参考这些联系较为紧密的招聘信息，对自己的简历和面试做好相应准备，比如第二志愿可以选择相似的职位，而不是仅仅凭借岗位的名字判断。

第二个类簇主要是那些对市场营销和技术要求中等的岗位，产品与客户支持 Product & Customer Support，用户体验与设计 User Experience & Design，营销与交流 Marketing & Communications 等，在投递该类型岗位的时候就可以一次投递相似的多个岗位，因为从聚类的结果来看这几个岗位在高频词上具有相关性。

第三个类簇主要是那些单纯的技术岗位，如软件工程 Software Engineering，技术设施维护 Technical Infrastructure，网络工程 Network Engineering，硬件工程 Hardware Engineering，在投递这些岗位的时候应该侧重自身的一些相关的技术特长，可以减少描述一些市场营销与决策管理方面的能力，为你的其它匹配技能与项目经历腾出空间。

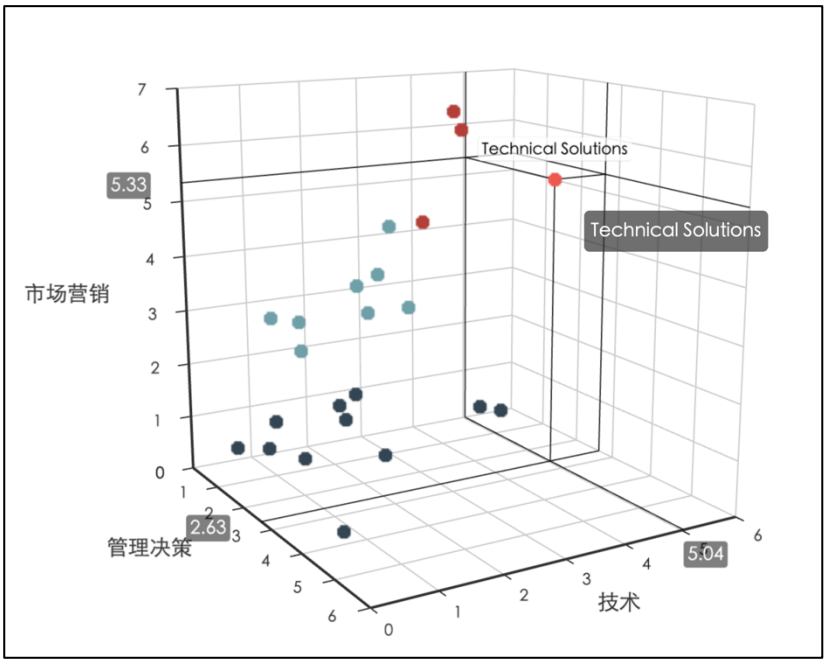

图 11 三个维度聚类结果

五、结语

总之，从分析的结果来看，Google 比较在乎求职者是否拥有学位（可能更喜欢硕士学位），也关注您是否具备最低要求和技能。但即使这样，职位发布的具体情况也可能与我们分析后得到的观点不一样。可能有一些方法可以增加在正确的职位上获得面试的机会。以下是此次分析的一些总结，为可能要申请 Google 的人提供一些较为粗糙的建议。

首先，目标尽量锁定美国。并寻找提供大量岗位但不会被视为科技中心的工作地理位置。其次，针对面向业务和运营的团队。他们在职位发布中占了很大的份额，并且可能有利于你的晋升。避免投递技术导向型的职位除非你在某一方面出类拔萃，尤其是在您的职业生涯较早的阶段。最后，简化个人履历内容。无论职位还是团队，他们都有他们要找的那些特定能力，如最低要求中需要有高效的沟通能力（书面和口语），流利的外语，编程，营销，工程思维，技术等方面的经验，比对这些高频词求职者可以确保自己命中要害而不是赘述无关内容。