目录

基于文本聚类的招聘信息中技能要求提取与量化

——以实习僧数据分析实习为例

一、研究背景与目的

网上对爬取招聘网站并对爬取的数据进行分析的技术博客多如牛毛，但对爬取的数据进行分析仅集中在分析薪资与地域、学历要求、工作年限、行业、公司规模等十分容易量化因素的关系，从职位描述中提取对应聘者的技能要求等少之又少，但技能因素是求职者评估自己是否能胜任一个岗位的重要因素，与其求职的准备、选择息息相关。

本文通过爬取实习僧网站“数据分析”一职的实习信息，对“职位描述”的文本进行预处理、分句，使用文本聚类的方式提取每条实习信息中其中的描述专业技能的句子，并对其描述的专业技能进行量化，从而探究专业技能对薪资的影响。本文所述的方法还可用于提取其他岗位、其他要求等，为大学生提供最直接、最真实的岗位信息，从而使他们对感兴趣的职业有所了解，对他们的学习方向提供建议，使其和能更明确地为求职作准备。

二、实习招聘信息数据的获取与说明

本文选择实习僧网站中的招聘信息进行数据的抓取。目前国内市场上的招聘平台虽多，垂直于实习领域的却只有“实习僧”一个代表性产品。实习僧网站作为近几年大学生找实习的热门平台，各大公司在上面发布的实习信息更多更全。在本次抓取中，一共抓取了实习僧上所有职位名称包含“数据分析”的实习信息351条，数据的主体为文本形式的数据。数据抓取的方式为使用python的request库获取具体实习信息的网页源代码，通过 re 模块使用正则表达式匹配出需要的信息。爬取的数据简介如下表 1 所示：

表 1 数据简介

	变量	含义	数据类型	取值范围	备注
职位因素	City	实习地点	定性变量（共 8 个水平）	北京、上海、广州、深圳、杭州、武汉、成都以及其他	以其他为准
	Education	学历要求	定性变量（共 4 个水平）	不限、专科、本科、硕士	以不限为准
	Day_per_week	周实习天数	定量变量（单位：天）	2—6	
	time_span	实习时长	定量变量（单位：月）	2—21	
	salary	实习工资	定量变量	9-425	根据薪资上下限计算得到平均工资
	content	职位描述	文本型	无	
	benefit	职位福利	文本型	无	
公司因素	Comp_industry	所属行业	定性变量（共 9 个水平）	计算机、互联网、金融、电子、电子商务、企业服务、广告、文化传媒以及其他	
	Comp_size	公司规模	定性变量（共 6 个水平）	少于 15 人、15-50 人、50-150 人、150-500 人、500-2000 人、2000 人以上	后续合并成 3 个水平：小型企业、中型企业、大型企业

三、文本聚类提取技能要求

本部分通过对招聘信息中“职位描述”的文本进行预处理、分句，利用 kmeans、GMM、NMF 三种文本聚类方法提取出其中与专业技能有关的句子，为后面量化专业技能作准备。

（一）职位描述文本的预处理

由于职位描述是掺杂着标点，特殊符号，及对文本含义无意义的语助词和语气词的完整中文语句，不能直接被计算机理解，在做分析前需进行文本预处理。

文本预处理主要分为分句，分词，删除停用词，删除低频词，文本向量化处理。

1.分句

由于招聘信息中的“职位描述”是大多按序号列出对应聘者的多条要求，技能要求一般包含在其中的某一句或某几句，因此首先要对每条“职位描述”的文本进行分句，分割的符号为句号、分号、冒号、换行符等。

表 2 分句示例

原文	分句
职位描述:数据分析工程师/实习生岗位职责：1、股票、期货程序化交易数据分析，包括各类高频交易数据的管理、维护、清洗等；2、构建、训练机器学习模型；3、研究、学习各类金融数据以及获取途径任职要求:1、物理、数学、电子、计算机等相关专业在读研究生；2、熟悉 Python，了解 Python 基本语法、数据结构、性能特征，熟悉动态语言的基本性质；3、熟悉机器学习、深度学习，熟悉 C#、SQL/MySQL 数据库者优先；4、具有较强的沟通能力	职位描述
	数据分析工程师/实习生岗位职责
	1、股票、期货程序化交易数据分析，包括各类高频交易数据的管理、维护、清洗等
	2、构建、训练机器学习模型
	3、研究、学习各类金融数据以及获取途径任职要求
	1、物理、数学、电子、计算机等相关专业在读研究生
	2、熟悉 Python，了解 Python 基本语法、数据结构、性能特征，熟悉动态语言的基本性质
	3、熟悉机器学习、深度学习，熟悉 C#、SQL/MySQL 数据库者优先
	4、具有较强的沟通能力

2.分词

文本分词是指将文章或语句中的词语按照一定标准进行划分的过程。相对于英语文本而言，汉语由于文本之间没有天然的分隔，处理其有一定的难度。将较长的语句或文章转化成较短的单词或词组，这一过程即中文分词。本研究中，采用基于统计的分词方法，通过隐马尔可夫（HMM）模型的 Viterbi 算法得到分词结果，具体分词过程是通过 Python 中 jieba 分词包实现。另外，由于数据分析领域存在不少专有词汇，如果只用 jieba 包默认的词典进行分析，则会无法识别这些专有词汇，因此在 jieba 包添加了自定义词典。

除本人对数据分析的了解而添加的词汇外，大部分词汇是通过统计 bigram 词频从而发现被误分的词组而添加的。

添加的部分词汇如表 3 所示：

表 3 添加的部分自定义词汇示例

技能型词汇	专业、年级词汇	通用技能、品质词汇
机器学习	本科以上学历	注重细节
深度学习	相关专业	合作精神
数据运营	在读研究生	逻辑清晰
数据挖掘	暑期实习	团队协作
统计分析		
数学建模		
文本挖掘		
自然语言处理		
R 语言		
办公软件		

3.去除停用词和特殊字符

去除停用词指过滤文本中的特殊字符和对文本含义无意义的词语。例如“的”，“啊”一类的语气语助词，对文本情感倾向判定无意义，却在文本向量表示时由于占据较大比重而对后续分析造成干扰，降低情感分类的准确性。同时，根据分词文本主题不同，停词表需要进行针对性地修改来提高准确性。因此，研究中用到的停词表在《哈工大停用词表》的基础上，根据帖子文本特点进行了修改。

4.去除超高频词与低频词

去除停用词后先做词频统计，发现词频极高的词，如“数据分析”、“职位描述”、“工作职责”、“负责”“工作”等不能体现具体岗位要求的词，因此删除前 10 个超高频词。

由于存在大量无意义的低频词（本文定义出现的频率仅为 1 次的为低频词）可能会降低分类精度，因此对去除停用词后的文本再删除低频词。

（二）文本预处理效果

文本预处理后的文本如表 4 所示，可以看到，每一句职位描述都有大致能看出其明确的类别，日常工作任务描述通常包含“整理”“录入”“搜集”这些动词；用人单位对应聘者专业的要求通常会指定具体专业和年级，如“大三”、“大四”、“研一”、“研二”、“统计学”、“数学”等；专业技能的描述则会指定应聘者需要掌握什么软件，如“excel”、“sql”等；通用技能、品质描述一般是要求应聘者“具有良好职业道德”、“细心”、“认真”等；实习时间描述一般是要求应聘者能保证实习“三个月”、“六个月”等，每周到岗“三天”、“四天”等。

由此可以预见，之后的文本聚类将会取得良好效果。

表 4 分词分句示例

序号	预处理后的文本	描述类别
1	产品库 日常 内容 维护 编辑 录入 整理 撰写 发布	任务描述
2	参与 产品库 优化 问题 整理 反馈	任务描述
3	协助 对接 部门 录入 需求	任务描述
4	大三 大四 学生 理工科 含 专业 专业 专业 考虑	专业、学历描述
5	熟练 使用 各类 办公 设计 软件	专业技能描述
6	较强 逻辑思维 归纳 总结 较强	通用技能、品质描述
7	具有 良好 职业道德 踏实 认真 注重细节	通用技能、品质描述
8	协助 数据运营 中心 进行 资料 搜集 整理 资料 审核	任务描述
9	协助 数据分析师 公司 数据库 内 完成 数据 清洗 配置 规则 监控 辅助	任务描述
10	保证 半年 以上 内 每周 至少 天到 岗 时间	实习时间描述
11	诚实 成熟 稳重 善于 交流	通用技能、品质描述
12	良好 沟通 协调 团队协作 精神	通用技能、品质描述
13	相关专业 统计学 数学 信息工程 计算机 本科	专业、学历描述
14	熟练 使用 msoffice 办公软件 excel powerpoint	专业技能描述
15	基于 公司 大数据 平台 海量 用户 运用 数据 挖掘 理论 方法 准确 快速 处理	任务描述

（三）文本聚类

因为计算机并不认识中文，因此需要将中文词转特征向量，本研究中文本向量化采用 tf-idf，用稀疏方式储存词-文档矩阵。矩阵维度为 t*n，t 代表句子个数，n 代表词语个数。 本文预处理后的词汇有 1793 个，句子 2817 条，提取 1000 个 tf-idf 特征，得到 2817*1000 的文档词频矩阵。下面将用三种聚类方法对“职位描述”中的句子进行聚类，根据聚类结果的解释性选择聚类数目。

1.Kmeans 聚类

Kmeans 是一种基于相似度的聚类方法。在聚类之前，需要用户显式地定义一个相似度函数。聚类算法根据相似度的计算结果将相似的文本分在同一个组。在这种聚类模式下，每个文本只能属于一个组,这种聚类方法也叫“硬聚类”。K-Means 方法是 MacQueen1967 年提出的，原理是给定一个数据集合 X 和一个整数 K（K<n），K-Means 方法将 X 分成 K 个聚类并使得在每个聚类中所有值与该聚类中心距离的总和最小。

经过多次尝试不同的聚类个数，发现把聚类个数定为6类时，能取得较好的聚类效果，即各个类别的文本能表达清晰明确的共同含义。图1中六个词云图展示了kmeans聚类的每个聚类中心的关键词，关键词大小与kmeans输出的权重大小有关。可以看到，前最上方的两张词云图都是描述日常工作任务，权重大的词有“整理”、“研究”、“相关”、“用户”、“数据处理”，“完成”、“项目”、“整理”、“收集”。中间左图中，权重大的词有“接收”、“暑期实习”、“每周”、“四天”等，可以看出这个类别的句子是跟实习时间有关。

中间右图中出现了很多数据分析常用的软件，如“python”、“excel”、“sql”、“spss”、“sas”，说明这个类别的句子是描述专业技能的，从中可以看出，office软件仍然是最为基础的要求，同时也要求应聘者能熟练掌握sql语言、使用数据库，当python、R软件等编程软件兴起时，像sas、spss等传统的统计分析软件仍然占据半壁江山，另外还有些数据分析实习要求掌握大数据相关的软件如hadoop、hive等。

最下方左图中权重大词有“逻辑思维”、“沟通”、“责任心”、“团队精神”、“细心”等，说明这些品质是数据分析岗最为看重的。最下方右图则是对应聘者的学历、专业的要求描述，要求最多的专业是“统计学”、“数学”。

2.GMM聚类

GMM 聚类是一种基于模型的聚类方法，它并不要求每个文本只属于一个组，而是给出一个文本属于不同组的概率。这种聚类方法也叫“软聚类”。这类方法通常假设数据满足一定的概率分布，聚类的过程就是要尽力找到数据与模型之间的拟合点。GMM假设数据服从高斯混合分布(Gaussian Mixture Distribution)，GMM中的k个组件对应于k个族，所以GMM聚类的过程实际上是以似然函数作为评分函数，求使得似然函数最大化的k组 ω_i μ_i Σ_i 参数。

由于文本聚类的稀疏性，且本文所使用句子都有明显的特征，因此GMM聚类后给出的每个句子的概率都十分接近0或1，相当于kmeans的效果。经过多次尝试不同的聚类个数，发现把聚类个数定为8类时，能取得较好的聚类效果，各个类别的关键词能体现明确的文本摘要。

表5展示了每个聚类中心的关键词。从类别的样本分布可以看出，除类别3有1000多条样本外，其他类别样本分布均匀。说明职位描述中，篇幅最大的是日常工作任务描述。

图 1 kmeans 聚类词云图

从表中关键词可以看出，第 1、6 类是专业技能描述，第 2、 3、 8 类是任务描述，第 4 类是实习时间描述，第 5 类是专业、学历描述，第 7 类是通用技能、品质描述。另外，尽管第 1、6 类是专业技能描述，但却略有不同，第 1 类出现的软件为 python、java、hive、hadoop，还出现了算法、数据挖掘、机器学习等词，说明此类职位描述的对编程、对分布式、算法要求更高些，而第 6 类则只是要求应聘者会 office 办公软件，以及传统的 sas、spss 统计分析软件，也要求 python。这说明数据分析岗仍可以往下细分为普通统计分析和偏向算法工程师的数据分析。

表 5 GMM 聚类关键词

类别	1	2	3	4	5	6	7	8
样本数	212	179	1163	202	220	209	301	331
关键词	熟悉 了解 python sql 一定 工具 数据挖掘 熟练 掌握 一种 算法 掌握 语言 经验 方法 hive 数据库 java hadoop 常用 机器学习	提供 支持 运营 产品 决策 日常 报告 部门 协助 报表 公司 提供 数据 相关 团队 需求 包括 优化 指标 建议 用户	相关 完成 经验 项目 协助 公司 报告 整理 研究 参与 行业 部门 维护 系统 学习 管理 撰写 产品 问题 平台	实习 以上 每周 至少 接受 长期 时间 暑期实习 保证 一周 实习期 能够 天及 转正 经验 工作日 全职 期间 考虑 四天	数学 以上 学历 相关 专业 本科 计算机 统计学 专业 硕士 研究生 金融 经济学 在读 全日制 大学 本科	使用 熟练 excel sql 软件 ppt 办公软件 python 工具 熟悉 office 操作 spss 熟练掌握 常用 掌握 精通 sas 运用	沟通 良好 具备 学习 责任心 逻辑思维 较强 团队协作 敏感 团队 协调 逻辑 合作精神 表达能力 优秀 敏感度 抗压 善于精神	进行 需求 用户 提出 整理 协助 产品 报告 模型 相关 项目 理解 挖掘 行为 优化 收集 建议 业务部门 信息 研究

3.NMF 聚类

NMF 是一种非线性降维方法，降维后的矩阵相当于对原文档词频矩阵进行了

特征提取，过滤噪声特征项，因此提取的特征更能反映样本的局部特征，聚类效果更好。用 NMF 进行文档聚类的原理为，给定一个文档语料库，首先构造一个 $n \times p$的文档-词频矩阵 X，其中 n 代表单词个数，p 代表文档个数。使用 NMF 分解矩阵 $X_{n\times p}$，聚类个数即为 k，得到分解矩阵 $U_{n\times k}$ 和 $V_{k\times p}^{T}$。

$$X_{n\times p} \approx U_{n\times k} V_{k\times p}^{T}$$

对 $U_{n\times k}$、$V_{p\times k}$ 归一化后，$V_{p\times k}$ 中的元素 v_{ij} 表示第 i 篇文档属于第 j 个类别的概率，因此，如果第 i 篇文档属于类别 m，则 v_{im} 在 $V_{p\times k}$ 中将取最大值，同时 $V_{p\times k}$ 第 i 行剩下的元素的值将会很小。

$$m = \operatorname{argmax}_j \{v_{ij}\}$$

NMF 聚类的结果如表 6 所示，可以看到 NMF 仅聚成 5 个类别，即可使每个类别的文本有清晰明确的文本摘要。

表 6 NMF 聚类关键词

类别	1	2	3	4	5
样本数	275	1355	389	480	318
关键词	以上学历	协助	excel	沟通	实习
	专业	相关	熟练	良好	每周
	本科	数据挖掘	sql	团队	以上
	数学	进行	python	具备	暑期
	统计学	运营	使用	学习	至少
	相关	报告	软件	责任心	接受
	计算机	需求	熟悉	逻辑思维	时间
	统计	用户	ppt	精神	长期
	在读	整理	熟练掌握	具有	实习期
	研究生	项目	工具	逻辑	转正
	专业本科	产品	sas	合作	保证
	全日制	完成	office	表达能力	一周
	硕士	报表	办公	较强	全职
	大三	建模	spss	抗压	实习生
	金融	支持	数据库	敏感	

（四）聚类方法的比较

1.聚类效果

从以上展示的聚类效果来看，NMF 的聚类类别最少，仅 5 类就可以使得每个类别具有明确清晰的，符合预先设想的文本摘要。

从聚类算法运行速度来看，同样都是从 python 中 sklearn 调用的函数，NMF 最快,仅需要 0.46s；kmeans 与之相差无几，需 0.56s， GMM 最慢，需要 16.78s 才能运行完毕。

2.兰德指数

兰德指数是一种基于聚类相似度的评价指标，它通过观察一对样本点 xi,yi 在两种聚类方法中是否被分在同一个类别来判断两种聚类方法的相似性。从表 7 可以看出，GMM、kmeans 与 NMF 的聚类效果相似度低，兰德指数少于 0.3；kmeans 和 GMM 的聚类效果较相似，两者的兰德指数为 0.43，说明两者中有 43%对样本点是分在了同一个类别或不在同一个类别。这可能是因为 GMM 是 k-Means 方法的概率变种，其基本算法框架和 k-Means 类似，都是通过多次迭代逐步改进聚类结果的质量。

表 7 三种聚类方法的两两兰德指数

兰德指数	Kmeans	GMM	NMF
Kmeans			
GMM	0.4310		
NMF	0.2480	0.2862	

四、文本聚类量化技能要求

（一）专业技能关键词与薪资的关系

从职位描述的文本中提取专业技能关键词，并对需求频率最高的前 10 个技能进行统计计算，得出每一个技能对应的平均薪酬水平，如图 2 所示，点的大小代表该技能需求量的多少。

在前 10 项技能中，excel 需求最大，但平均薪资最低，仅为 144 元，因为 excel 是数据分析工作最应该掌握的工具；Hadoop， Spark 这两者需求少，但平均薪酬水平最高，超过 200 元，并且相对其他技能来说有比较大的差异，因为 Hadoop， Spark 都是应用于分布式数据处理；其他软件对应得平均薪资在 160-200 之间。因此专业技能对薪资有明显影响。

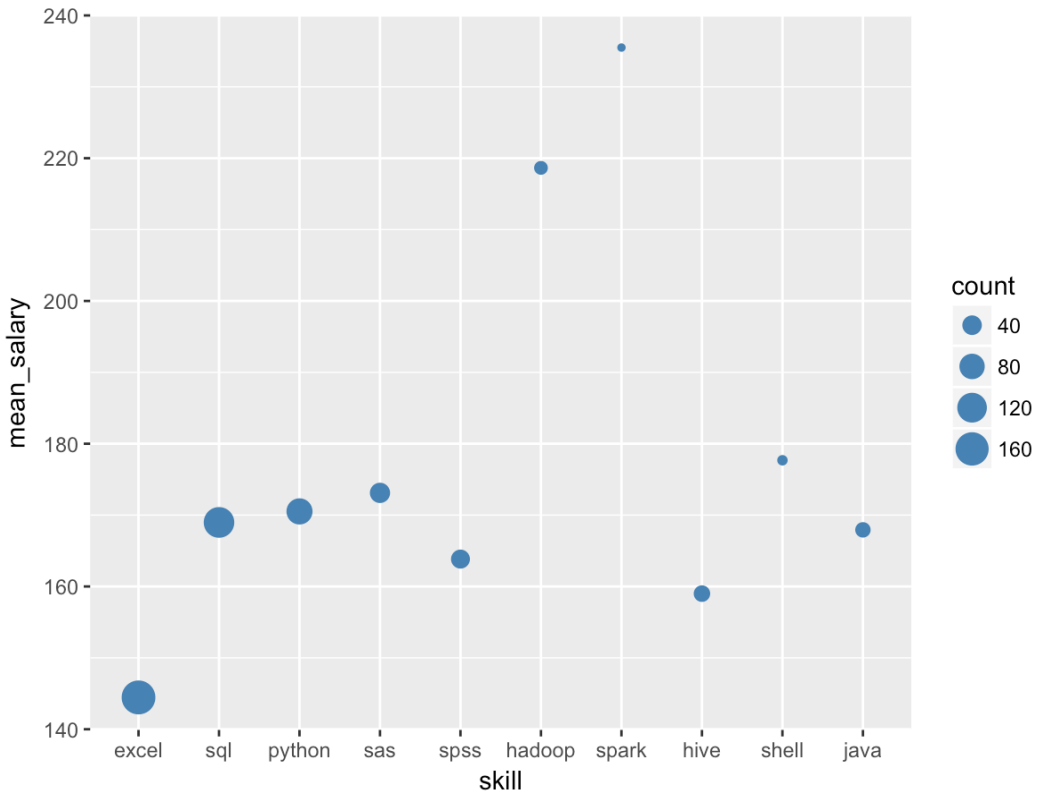

图 2 需求排名前 10 的技能与平均薪资的散点图

（二）文本聚类量化技能要求

通过 GMM 聚类可以看出，专业技能描述也有高低之分，从前面的分析也可看出，要求应聘者掌握 hadoop、spark 等大数据分析相关技能的实习工资更高些，但仅通过从文本中提取技能关键词来衡量技能与薪资的关系，一来需要预先知道有哪些重要技能，二来提取的技能太多会使得技能因素分散在每个技能变量上因此每个技能变量包含的信息较少，使得这种方法更为繁琐，缺乏普适性，且不利于分析技能与薪资的关系。

因此可以将每条样本的职位描述中专业技能描述的句子挑出来再进行聚类，使用聚类方法为句子所述的技能进行评分，这样无需一个个提取技能关键词，且把句子中的关键词综合考量。考虑到以上三种聚类算法的相似度并不高，因此将以上几种聚类方法挑出的描述专业技能的句子取并集，构成小型专业技能描述句子语料库，对该语料库再次进行聚类，使用以上三种方法的聚类效果表 8 所示：

从关键词一栏可以看到，三种聚类方法均能把描述专业技能的句子聚成三种相似的类别。第一种仅要求应聘者掌握 msoffice 软件和 SQL 查询语言，第二种除了要求掌握 msoffice 和 SQL 查询语言以外，还要求掌握其他统计分析软件，如 sas、spss、python 等，而第 3 种则还要求应聘者会应用与大数据、计算机有

关的软件，如 hive、hadoop、Java 等。

表 8 描述专业技能的句子聚类情况

聚类方法	类别	样本数	关键词
Kmeans	1	90	办公软件 ppt office 软件 操作 word 精通 熟练掌握 数据处理 制作 报告 较强 尤其 sql 功底 运用 英语 powerpoint 能够 撰写
	2	36	工具 spss 常用 sas 方法 了解 sql 掌握 统计 熟悉 ppt tableau 数据挖掘 python 语言 熟练掌握 matlab 任一 模型 统计分析
	3	142	熟悉 sql python 熟练掌握 数据库 语言 掌握 了解 软件 相关 经验 一种 java hive 至少 工具 sas 进行 运用 基础
GMM	1	65	ppt 办公软件 office 熟练掌握 制作 报告 熟悉 word 函数 功底 良好 较强 msoffice 文字 操作 能够 透视 sql 一定 数据处理
	2	50	软件 office 精通 统计 操作 运用 数据处理 sql 公式 办公软件 spss ppt 熟悉 办公 英语 相关 统计分析 薪资 基本 word
	3	153	sql 熟悉 python 工具 了解 熟练掌握 语言 掌握 sas 数据库 常用 spss 方法 经验 一种 至少 hive 进行 java 相关
NMF	1	59	Excel ppt 熟练 办公软件 使用 office 熟练掌握 word sql 功底 较强 报告 数据处理 制作 良好 熟悉 文字 具备 常规 敏感性
	2	54	Excel 软件 熟练 使用 office sql 精通 统计 ppt spss python 数据处理 熟悉 运用 统计分析 操作 相关 Ps 办公软件 sas
	3	155	sql excel python 熟练 熟悉 使用 熟练掌握 工具 sas 数据库 spss hive java 语言 了解 Hadoop 数据挖掘 掌握 经验 常用

将每种方法聚出的这三类分别打分为 1 分、2 分、3 分，该职位的技能要求分数取三种方法打分的平均值，从而量化职位的技能要求。

图 3 散点图显示了技能分数与平均工资的关系，可以看出大部分实习工资集中在 100-200 之间，而当技能分数超过 2.5 时，有一些实习的工资能超过 300，从 loess 拟合的回归曲线可以看出轻微渐升的趋势，说明技能要求越高，公司愿意支付的工资越多。

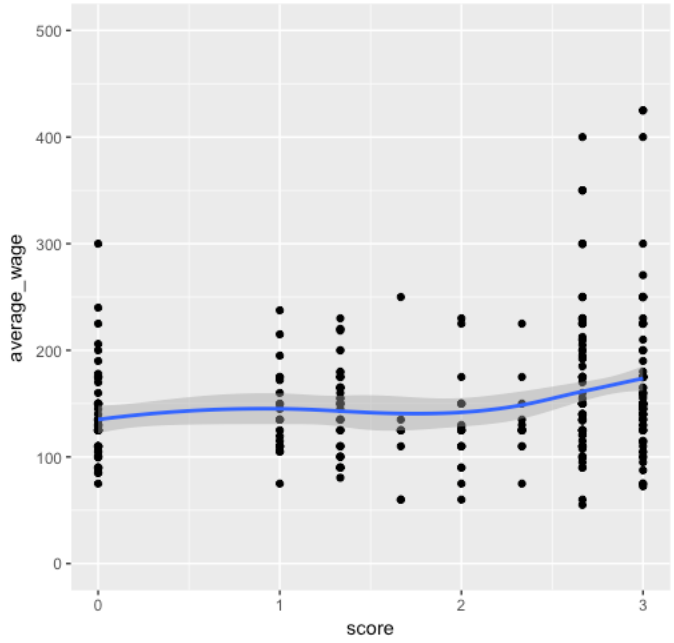

图 3 技能分数与实习工资的散点图

五、技能与薪资的回归分析

实习工资的高低还跟很多因素有关，如地域、行业等，因此接下来把这些因素考虑进去，以实习工资为因变量进行回归分析，重点观察技能分数对实习工资的影响。

从表 9 可以看出，技能分数对实习工资有显著影响，实习分数每多一分，即多掌握一门常用统计软件甚至多掌握一门大数据分析相关软件，则平均实习工资涨约 7 元。因为仅仅是实习而不是正式员工，不同的实习，日实习工资几乎只在 100-200 内浮动，因此技能对工资上涨影响不太大。

其他方面，从实习时间上看，要求一周实习天数越多，说明公司越需人数，愿意开出的实习工资越高；从学历要求上看，要求学历是本科生的实习工资比不限专业的工资低 12.59 元；从专业要求上看，要求专业是计算机的实习工资比专业要求为其他的实习工资高 16.44，计算机专业出身的学生仍是就业市场中的热点需求；从实习地点上看，北上广深杭的实习工资比其他城市多 20 元以上，其中杭州的实习比其他城市的实习高 37 元，而成都、武汉则比其他城市少 5 元以上；从公司行业上看，互联网、计算机行业的公司更为大方些，开出的实习工资更高；从公司规模上看，中型企业比小型企业开出的实习工资少 10 元，而大型企业则比小型企业多 3 元，工资条件仍是大公司吸引就业者的优势。

在该回归方程中，F 检验显著，R 方仅为 0.6，说明自变量对实习工资的波

动仅解释了 60%，另外实习工资还跟具体公司规定，市场行情有关。

表 9 实习工资回归系数表

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	63.23	20.21	3.13	0.0019 **
Skill_score	6.96	2.77	2.52	0.0124 *
day_per_week	12.49	3.61	3.46	0.0006 ***
time_span	0.05	0.79	0.06	0.9497
education 专科	0.13	7.72	0.02	0.9866
education 本科	-12.59	10.98	-1.15	0.2524 *
education 硕士	1.16	10.26	0.11	0.9103
subject_统计	-10.96	7.09	-1.54	0.1234
subject_计算机	16.44	7.49	2.20	0.0288 *
subject_数学	-1.74	7.98	-0.22	0.8275
city_北京	20.11	9.49	2.12	0.0349 *
city_上海	26.20	9.98	2.63	0.0090 **
city_杭州	37.36	19.34	1.93	0.0542 .
city_深圳	25.41	16.54	1.54	0.1255
city_广州	11.16	12.08	0.92	0.3562
city_成都	-6.77	19.09	-0.35	0.7232
city_武汉	-8.37	20.10	-0.42	0.6775
industry_互联网	10.18	9.74	1.04	0.2971 *
industry_计算机	10.98	8.83	1.24	0.2146 *
industry_金融	-16.10	10.54	-1.53	0.1277
industry_电子商务	1.93	25.12	0.08	0.9387
industry_企业服务	-12.12	13.10	-0.93	0.3556
industry_广告	-26.34	15.33	-1.72	0.0866 .
industry_文化传媒	-34.82	16.54	-2.11	0.0360 *
industry_电子	-5.08	19.87	-0.26	0.7985
industry_通信	-29.19	19.92	-1.47	0.1439
comp_size 中型企业	-10.59	9.31	-1.14	0.2563
comp_size 大型企业	3.42	6.66	0.51	0.6084
F-statistic: 3.269 on 25 and 319 DF				
p-value: 6.059e-07				
R_square = 0.61				
Adjusted_R_square = 0.59				

六、结论

本文通过爬取实习僧网站“数据分析”一职的实习信息，对“职位描述”的

文本进行预处理、分句，使用文本聚类的方式提取其中的描述专业技能的句子，并对这些句子再一次进行聚类，区分不同层次的技能要求，并对职位的技能要求进行打分，从而实现岗位信息中技能要求的量化，使得技能与薪酬的关系能更深入地分析。通过以上分析，可以得出以下三个结论：

第一，数据分析师需求频率排在前列的技能有：SQL，Excel， SAS，SPSS，Python， Hadoop 和 MySQL 等，其中 SQL 和 Excel 简直可以说是必备技能；

第二，海量数据、分布式处理框架是走向高薪的正确方向；

第三，SQL 语言和传统的 SAS，SPSS 两大数据分析软件，能够让你在保证中等收入的条件下，能够适应更多企业的要求，也就意味着更多的工作机会。

本文仅以实习僧网站的数据分析实习岗为例，阐述如何通过文本聚类的方法提取并量化职位描述中的专业技能要求，因此数据量比较小，代表性不够好，另外结果适合于实习方面的数据分析岗而不是正式工作。另外本次分析主要针对工具型的技能进行了分析。但实际上数据分析师所需要具备的素质远不止这些，还需要有扎实的数学、统计学基础，良好的数据敏感度，开拓但严谨的思维等。