基于 LDA 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 LDA 主题模型提取每条实习信息中以专业技能为主题的句子，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、LDA 主题模型提取技能要求

本部分通过对招聘信息中“职位描述”的文本进行文本预处理，包括分句、词，删除停用词，删除低频词，然后用生成 tf-idf 文档矩阵，再利用 LDA 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）文本预处理效果

文本预处理后的文本如表 2 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 2 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

（二）建立 LDA 主题模型

文本预处理后，一共有 2834 个句子，2013 个词，根据主题关键词结果的明确性、可解释性来选择主题数目。

1. 主题数为 6 时各个主题的关键词能形成明确的文本摘要

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，能取得较好的效果，即各个主题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

每个主题的关键词，关键词形状大小与权重大小有关。

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，可以看出这个主题是跟实习时间有关。

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”，说明这个主题是描述专业技能的，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. 主题数为 10 时技能主题分化

当 LDA 提取的主题数设为 10 时，如图 2 所示，出现了两个专业技能主题，却略有不同，一个主题中出现的软件为 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，而另一个主题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、大数据方面的数据分析。

图 1 LDA 主题模型 6 个主题关键词（主题数=6）

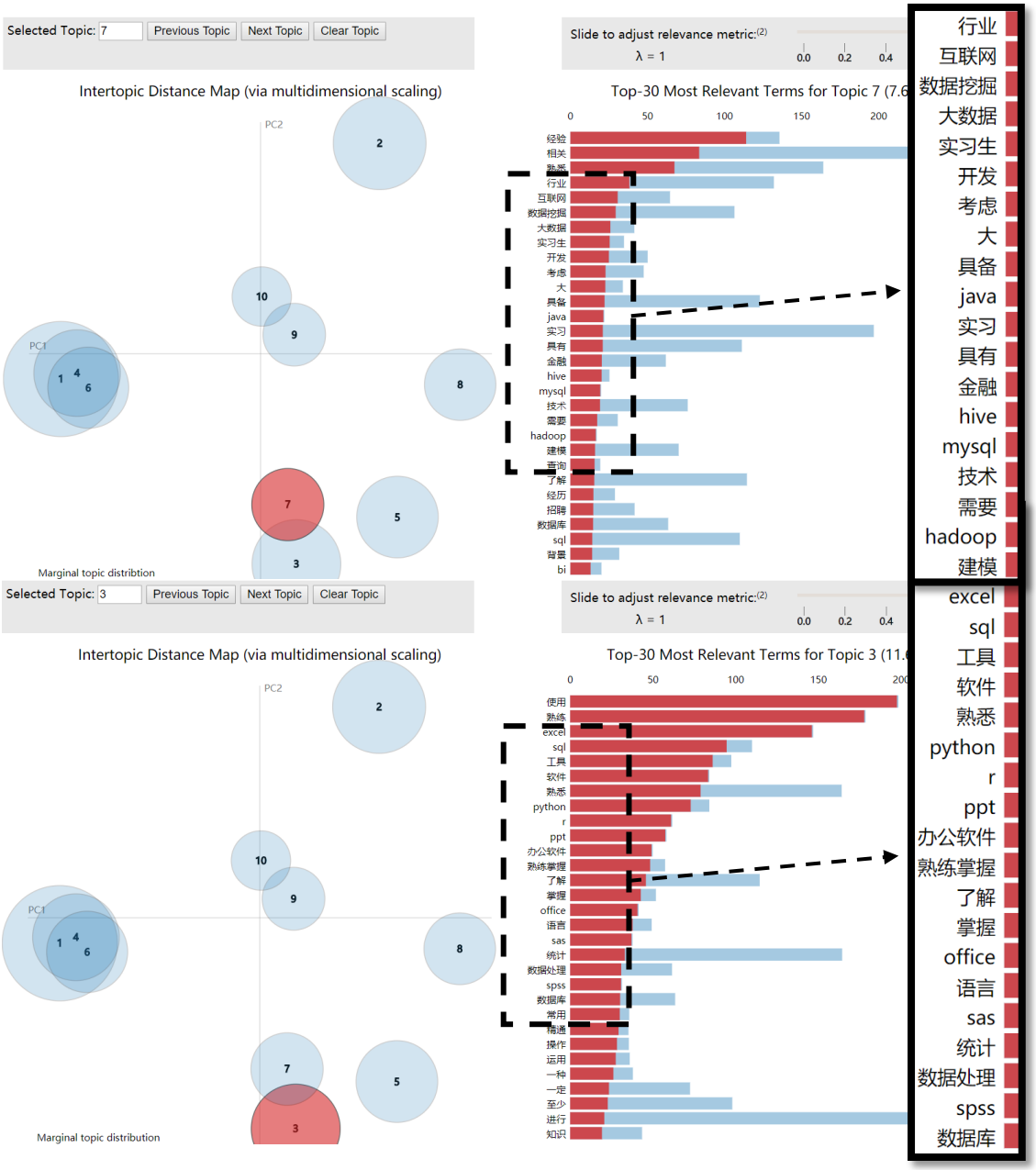

图 2 LDA 提取的两个技能主题（主题数=10）

四、LDA 主题模型量化技能要求

（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，选出需求频率最高的前 10 个技能，挑选出包含这些技能的招聘信息，根据包含某一个技能的所有招聘信息的工资计算出该技能对应的平均工资，如图 3 所示，横轴为技能关键词，纵轴为平均工资，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。

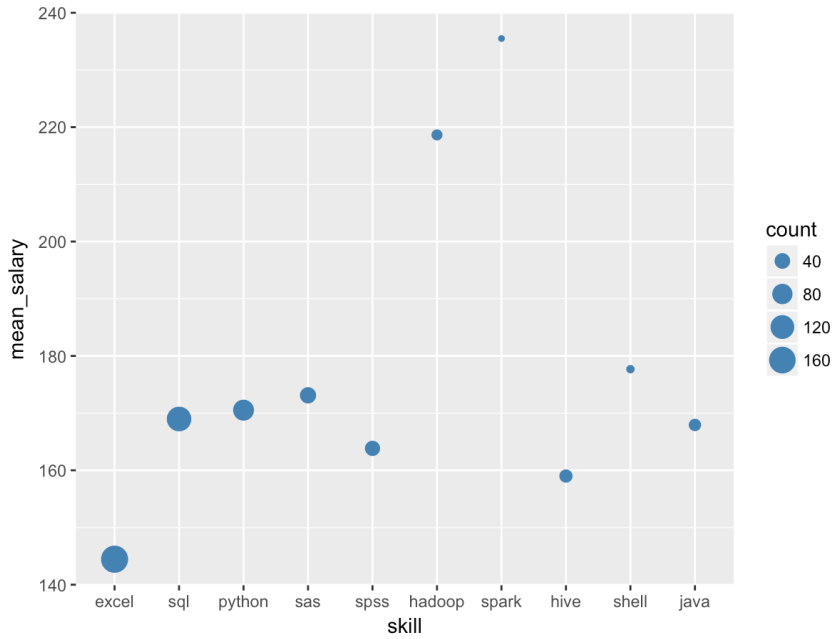

图 3 需求排名前 10 的技能与平均薪资的散点图

（二）LDA 量化技能要求

通过 LDA 的提取 10 个主题的图中可以看出，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，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提取细分的技能主题，根据句子所倾向技能主题的高低为句子所述的技能要求进行评分，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。因此将提取了 6 个主题的 LDA 模型的第 4 个主题（技能主题）取值大于 0.5 的句子取出来，构成小型专业技能描述句子语料库，一共 268 句，587 个词，对该语料库再次建立 LDA 主题模型，效果如图 4 所示：

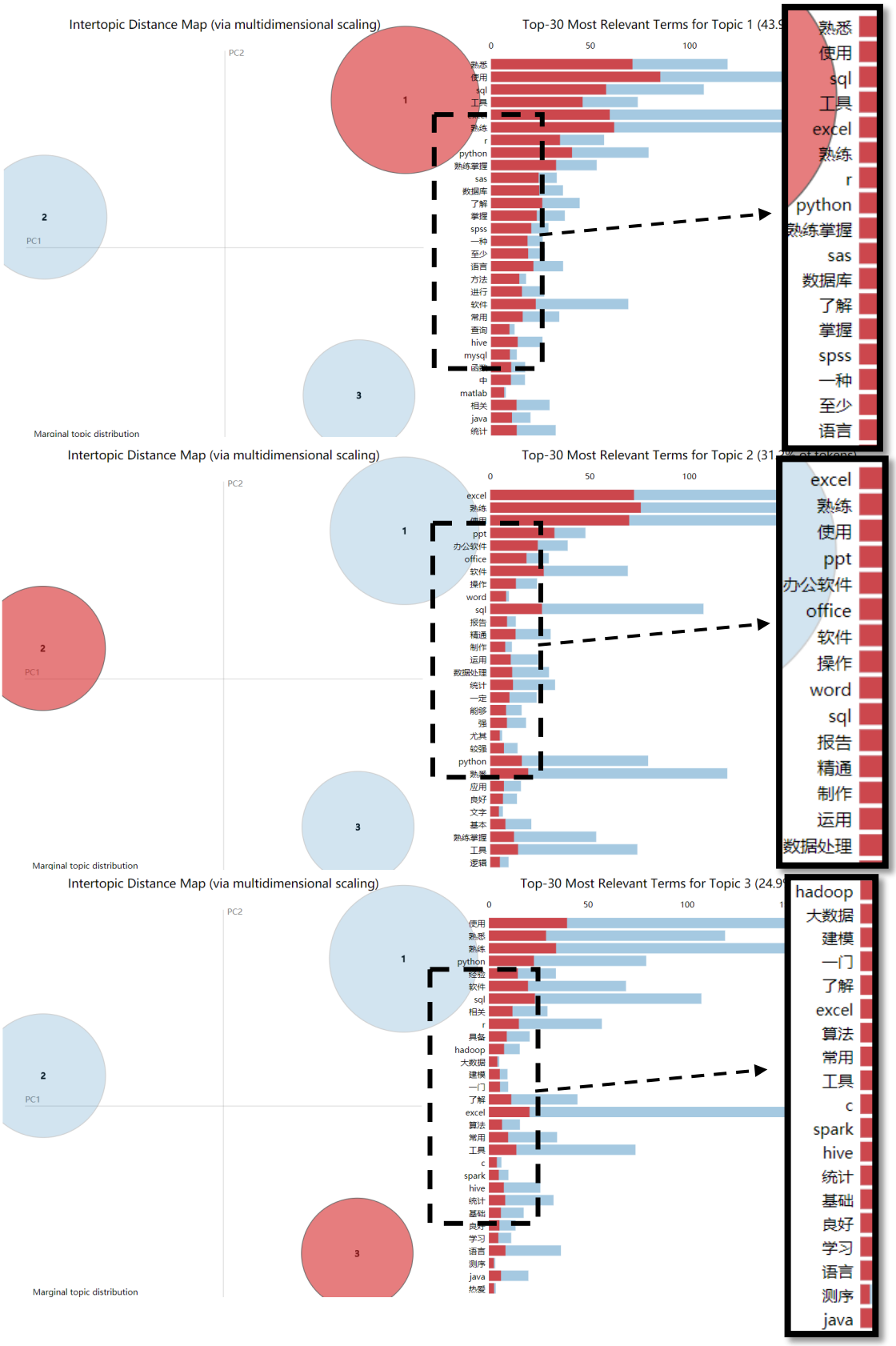

图 4 LDA 提取的 3 个技能主题

当 LDA 提取 3 个主题时，每个主题的关键词正好能表达技能主题的高低。第 2 个主题的句子仅要求应聘者掌握 msoffice 软件和 SQL 查询语言，第 1 个主题除了要求掌握 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 个主题则还要求应聘者会应用与大数据、分布式有关的软件，如 hive、hadoop、spark、Java 等。

按技能高低给技能主题打分 1 分、2 分、3 分，该职位的技能要求分数为其在三个主题上的概率乘以三个主题的分值再求和，从而量化职位的技能要求。

图 5 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。

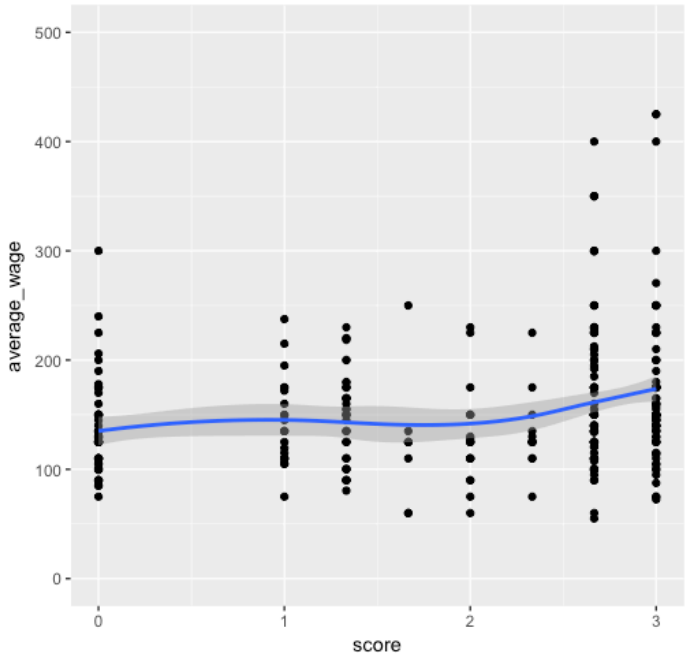

图 5 技能分数与实习工资的散点图

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 3 回归系数表可以看出，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。

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 LDA 提取其中包含技能主题的句子，并对这些句子再次提取技能主题，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 LDA 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。

表 3 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				