网络招聘信息的分析与挖掘

网络招聘信息的分析与挖掘

摘　要：严峻的就业形势，使得越来越多的人期望通过选择职业技能培训或是学历提升来提高自己的综合技能，从而顺利的走上工作岗位。

随着计算机网络的迅速发展，网络招聘信息平台现已成为招聘者进行信息发布，应聘者获取信息的主要渠道，网络招聘信息不仅能够反映用人单位的人才需求,另一方面也反映出了社会和各行业对人才的需求现状。因此，对网络招聘信息进行分析，是具有一定的现实意义。本文采用文本的挖掘方法，建立了文本挖掘模型，对目前招聘现状进行一定的描述分析，分析目前所需的人才类型，还有热门行业，职位，并且预测相关职位的需求情况。

利用 excel 以及 Rstudio 对职业类型进行文本聚类，根据结果，我们进而对新型的职业——数据分析师，进行岗位技能需求分析，得出用人单位重于能力多于专业的结论，并对数据分析师职业进行发展预测；主要行业是 IT 类型，而且职位主要分布在中国东部发达的地区的结论。

本文的特色在于灵活结合了文本挖掘以及时间序列，对新兴的需求职位的进行纵向的分析，较有深度，同时结合分析结果以及专业方向问题，为学校培养人才计划方面提供了可行性的建议。

关键词：网络招聘　招聘信息　文本挖掘　词云 时间时序

目录

前言

严峻的就业形势使得越来越多的人希望通过选择职业技能培训或学历提升来提高自己的综合技能，从而顺利走上工作岗位。

2013 年，高校毕业生人数达到 699 万。此后 3 年间，这一数字不断攀升，到 2016 年将达到新高。

“今年大学生就业形势不会比去年更加严峻。”全国政协委员、新东方教育科技集团的董事长俞敏洪干脆地回答,因为现在许多的“创新、创业”的新公司吸引了许多大学生,俞敏洪委员的判断与今年的政府工作报告中提出的“鼓励以创业带动就业”的政策方向相吻合。

全国政协委员、北京外国语大学原党委书记杨学义的看法也类似，他的信心来自于国家在政策层面对就业问题的重视。这种从国家政策层面上的重视体现在 3 月 5 日财政部与发改委公布的两份最新文件中。财政部将“完善和鼓励高校毕业生到科技型、创新型中小企业和城乡基层就业创业政策，拓宽高校毕业生就业渠道”写入 2016 年国家财政的主要支出政策。国家发改委把做好就业托底工作列为 2016 年经济社会发展的主要任务。国家从政策上全方位支持就业工作。

全国政协委员、华中师范大学党委书记马敏注意到去年我国城镇新增就业 1312 万人，超额完成了全年预期目标。李克强总理在政府工作报告中，将今年城镇新增就业的目标定在了 1000 万人以上。虽然对 2016 年高校毕业生就业情况的预估普遍乐观，但也没有低估眼前的压力。许多重点大学的学生面临的就业压力较小。在这 765 万毕业生中，压力最大的是一些普通院校的学生，那如何破解大学生就业难？在“供给侧改革”思路的引导下，从大学生就业市场“供”与“需”的两端思考，提出“教育供给侧改革”的思路：

在“需”的层面，经济增长无疑是就业岗位增加的必要条件。全国政协委员、南京大学政府管理学院院长张凤阳赞同政府工作报告中提及的“稳增长主要是为了保就业、惠民生，有 6.5%～7%的增速就能够实现比较充分的就业。”

1.研究的方向

1.1研究的背景

根据信息产业部分析，“十五”期间是我国电子信息产业发展的关键时期，预计电子信息产业仍将以高于经济增速两倍左右的速度快速发展，产业前景十分广阔。未来的发展重点是电子信息产品制造业、软件产业和集成电路等产业 ；新兴通信业务如数据通信、多媒体、互联网、电话信息服务、手机短信等业务也将迅速扩展；值得关注的还有文化科技产业，如网络游戏等。目前，信息技术支持人才需求中排除技术故障、设备和顾客服务、硬件和软件安装以及配置更新和系统操作、监视与维修等四类人才最为短缺。此外，电子商务和互动媒体、数据库开发和软件工程方面的需求量也非常大。

1.2 研究的目的和意义

针对大学生就业连年愈加严峻之势，我们将利用 2015 年 11 月 7 日— 2016 年 4 月 5 日的网络招聘信息进行行业、职业分析，并希望可以为大学生就业提供方向和建议。人才需求的市场中，在“需”的层面，经济增长是就业岗位增加的必要条件，国家在经济政策以及中国经济的发展愈加见好；但是在“供”的这个层面，大学生应对行业需求有一定了解进而提升自我，不在竞争中被轻易淘汰。

2. 理论知识

2.1 数据挖掘的简介

数据挖掘又称为资料探勘或者数据采矿。它是数据库知识发现中的一个步骤。数据挖掘通常是指从大量数据中通过算法来搜索隐藏其中信息的过程。数据挖掘一般与计算机科学有关，通过统计、情报检索、专家系统（依靠以前的经验法则）和模式识别等多种方法来实现以上目标。

分析方法包括分类、估计、预测、相关性分组或关联规则、聚类和复杂数据类型挖掘。

（1）分类。先从数据中挑选已经分类好的训练集，在训练集上用数据挖掘分类的技术，建立好分类模型，对于没分类的数据进行分类。

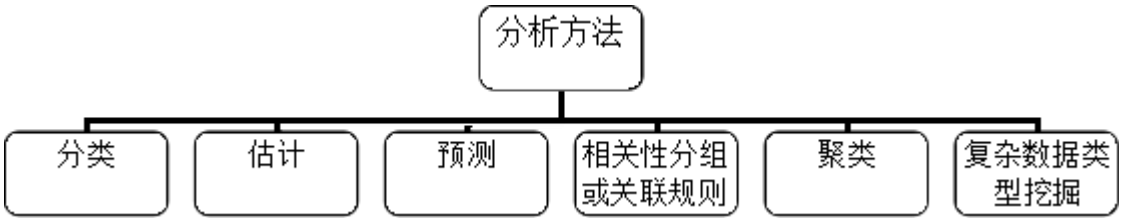
（2）估计。给定输入数据，通过估计，得到未知的连续变量的值，然后根据先设定的阈值，进行分类。

（3）预测。预测一般是通过分类或者估值起作用的，即通过分类或估值得到模型，模型用于未知变量的预言。预言其实没必要分为单独的类。预言目的是对未来未知变量的预测，这种预测是要时间来验证的，也就是说必须经过一段时间后，才能知道预言的准确性是多少。

（4）相关性分组或关联规。相关性分组或关联规则决定什么事情将一起发生。

（5）聚类。聚类是对记录进行分组，把类似的记录在一个聚集里。聚集通常作为数据挖掘的第一步。

（6）复杂数据类型挖掘。描述和可视化是对数据挖掘结果的表示方式。一般只指数据可视化工具，包括报表工具和商业智能分析产品的统称。例如通过 Yonghong Z-Suite 等工具进行数据的展现、分析和钻取，将数据挖掘的分析结果更形象深刻地展现出来。



数据挖掘过程中各步骤如下:

（1）确定业务的对象，清晰定义出业务问题，认清数据挖掘的目的是数据挖掘的重要的一步。

（2）数据准备：

a 数据选择，搜索全部与业务对象有关的内部和外部数据信息，并选择适合数据挖掘应用的数据。

b 数据预处理，研究数据的质量，为下一步的分析作准备，确定要进行的挖掘操作的类型。

c 数据转换，将数据转换成分析模型，这个模型是针对挖掘算法而建立的。建立一个适合于挖掘算法的分析模型是数据挖掘成功的关键。

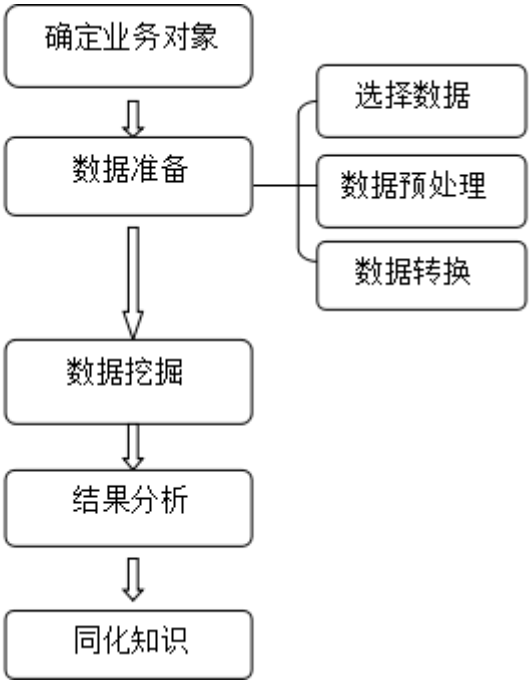
(3) 数据挖掘：对得到的经过了转换的数据进行挖掘。

(4) 结果分析：解释和评估结果，使用的分析方法通常应数据挖掘操作来定，

一般会用可视化技术。

(5)知识的同化：将分析得到的知识汇到业务信息系统的组织结构中。

数据挖掘的步骤图：



数据挖掘的任务主要是关联分析、聚类分析、分类、预测、时序模式和偏差分析等。

⑴关联分析。两个或以上变量的取值间有某种规律性，称为关联。关联分为简单关联、时序关联和因果关联这三种。关联分析目的是找到数据库中隐藏了的关联网。通常用支持度和可信度两个阀值去度量关联规则的相关性，并且不断引入兴趣度、相关性等参数，使所挖掘的规则更加符合需求。

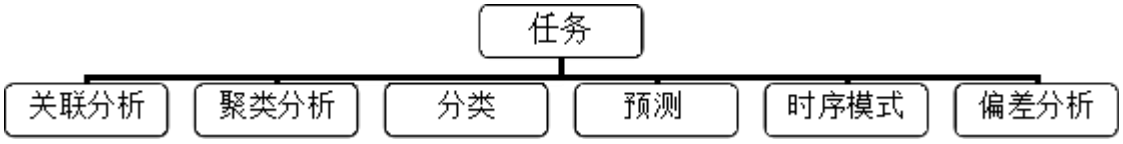
⑵聚类分析。聚类是将数据按相似性来归纳成若干类别，同类中的数据互相相似，不同类中的数据相异。聚类分析能建立宏观的概念，发现数据的分布模式和可能的数据属性间的相互关系。

⑶分类。分类是找出一个类别的概念描述，代表了这类数据的总体信息，并用这种描述构造模型，通常用规则或者决策树模式表示。分类是用训练数据集通过一定算法来求得分类规则。分类可以被用于规则描述和预测。

⑷预测。预测是通过历史数据找到变化规律，建立模型，并由模型对未来数据的种类和特征进行预测。预测关心的是精度及不确定性，一般用预测方差来度量。

⑸时序模式。时序模式是指根据时间序列搜索得到的重复发生概率比较高的模式。与回归相同，它也是用已知的数据来预测未来的值，但是这些数据的分别是变量所处的时间的不同。

⑹偏差分析。在偏差中包括许多有用的知识，数据库中的数据有许多异常的情况，发现数据库中存在的异常情况的数据是非常重要的。偏差检验的基本方法是寻找观察结果和参照之间的差别。



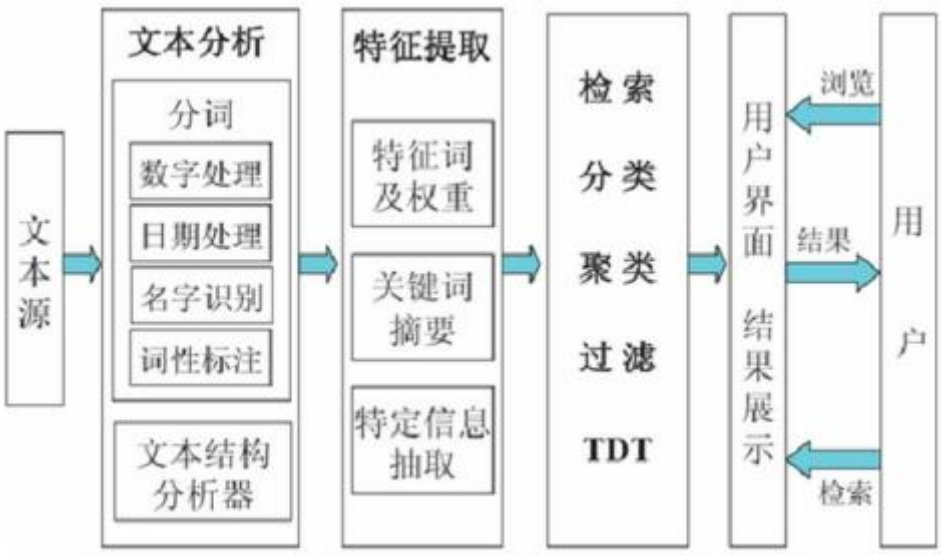
2.2 文本挖掘模型

文本挖掘是信息挖掘的研究分支，用于文本信息的知识发现。文本挖掘用智能算法，如基于案例的推理、神经网络、可能性推理等，并结合文字处理技术，分析大量的非结构化文本源（如电子表格、文档、问题查询、客户电子邮件、网页等），抽取或标记文字间的关系、关键字概念，并按内容对文档进行分类，获得有用的知识和信息。

文本挖掘是一个多学科混杂的领域，涵盖了多种技术，包括信息抽取、数据挖掘技术、机器学习、信息检索、计算语言学、自然语言处理、线性几何、概率理论、统计数据分析甚至还有图论。

文本挖掘的常见方法主要包括词频分析及主题模型、wordcloud 展现、分类评价、文本分类等。分类主要包括无监督分类（系统聚类、string kernals、KMeans），有监督分类（SVM、knn）。

文本挖掘主要过程：文本分析、特征提取、文本分类、文本聚类、模型评价。



K-means 算法是典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。

k 个初始类聚类中心点的选取对聚类结果具有比较大的影响，因为在算法第一步中是随机选取任意 k 个对象作为初始聚类的中心，初始地代表一个簇。算法在每次迭代中对数据集中剩余的每个对象，根据与各个簇中心的距离将每个对象重新赋给最近的簇。当考察完全部数据对象后，一次迭代运算完成后，新的聚类中心被计算出来。如果在一次迭代前后，J 的值没发生变化，说明算法已经收敛。

算法过程：

(1)从 n 个文档随机选取 k 个文档作为质心

(2)对剩余的各个文档测量其到每个质心的距离，并将它归到最近的质心的类

(3）重新计算各个类的质心

(4）迭代 2～3 步至新质心与原质心相等或小于指定阀值，算法结束

说明如下：先从 n 个数据对象任意选择 k 个对象作为初始聚类中心，而对于所剩下其它对象，根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类；再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值）；不断重复这一过程至标准测度函数收敛为止。通常都采用均方差作为标准测度函数. k 个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

要分析文本内容，最常见的分析方法是提取文本中的词语，并统计频率。频率能反映词语在文本中的重要性，通常越重要的词语，在文本中出现的次数就会

越多。词语提取后，还可以做成词云，让词语的频率属性可视化，更加直观清晰。词云就是对文本中出现频率较高的关键词给予视觉上的突出，形成关键词云层或关键词渲染，从而过滤大量的文本信息，使浏览者只需一眼扫过文本就可以知道文本的主旨。

2.3 截面数据和时间序列数据

2.3.1 截面数据

截面数据是指在同一时间截面上反映一个总体的一批个体的同一特征变量的观测值 ，是样本数据中常见类型之一。横截面数据不要求统计对象和范围相同，但是要求统计的时间相同。在分析横截面数据时，应注意两个问题：

一是异方差问题，因为数据是在某一时期对个体或地域的样本采集，不同个体或地域本身存在差异；

二是数据一致性，主要包括变量样本容量是否一致、样本取样时期是否一致、数据统计标准是否一致。

2.3.2 时间序列数据

时间序列数据是指在不同时间点上收集的数据，这类数据反映了某一事物、现象等随时间的变化状态或程度。

时间序列建模基本步骤是：

①用观测、调查、抽样、统计等方法取得被观测系统时间序列动态数据。

②根据动态数据作相关图，进行相关分析，求自相关函数。相关图能显示出变化的趋势和周期，并能发现跳点和拐点。跳点是指和其他数据不一致的观测值。如果跳点是正确的观测值，在建模时应考虑进去，如果是反常现象，则应把跳点调整到期望值。拐点是时间序列从上升趋势突变为下降趋势的点。如果存在拐点，则在建模时一定用不同的模型去分段拟合该时间序列。

③辨识合适的随机模型，进行曲线拟合，即用通用随机模型去拟合时间序列的观测数据。对于短的或简单的时间序列，可用趋势模型和季节模型加上误差来进行拟合。对于平稳时间序列，可用通用 ARMA 模型及其特殊情况的自回归模型、滑动平均模型或组合-ARMA 模型等来进行拟合。当观测值多于 50 个时通常都采

用 ARMA 模型。对于非平稳时间序列则要先把观测到的时间序列进行差分运算，化为平稳时间序列，再用适当的模型去拟合这个差分序列。

3.模型的建立与分析

3.1 数据预处理

3.1.1 异常数据处理

在这里我们获取了 2015 年 11 月 7 日— 2016 年 4 月 5 日的网络招聘信息以及岗位描述数据，在数据的处理时候，有些数据是不符合常理的，我们就要把这些点剔除。

3.1.2 缺失数据处理

在一些数据不完整的时候，我们会采取一定的方法进行处理，例如在 IT 行业的预测中，由于 2015 年 11 月和 2016 年 4 月的数据不完整，如果根据这个数据来分析的话，就会出现很大的误差，所以我们对它的一个处理是剪切掉。在附件一里面的一些岗位名称，还有城市名缺失，同样，我们也是删除掉的。

3.2 人才的需求中的职业类型的现状探索

3.2.1 获取有效信息数据

2015 年 11 月 7 日— 2016 年 4 月 5 日的网络招聘岗位信息文本数据，共 539216 条记录，截取前 7 条数据如下：

Job_Description	PositionId
职位描述 ：基于 Android 平台进行手机软件的设计、开发、需求分析等； 任职要求： 1、熟练掌握 java 技术，熟悉面向对象编程设计，具备扎实的编程基础； 2、精通 Android 开发平台及框架原理；	1

对面向对象开发有深入的理解； 3、具备熟练的技术调研能力并能完成可行性说明； 4、两年以上 Android 端移动互联网开发经验； 5、必须具有熟练的即时通讯开发和消息推送开发经验 6、具备良好的团队合作能力和沟通能力，有较强的自我提升和学习能力 优先录用条件： 1、研究阅读过 Android 系统的源码 2、具有移动支付、网上银行或消息推送、即时通讯等相关软件开发经验 3、熟悉 NDK 编程并有相关经验	
岗位描述：1、前端框架的设计与实现 2、 各业务模块前端代码开发 3.、 平台易用性与用户体验的持续改进 4、 Web 前沿技术研究和新技术调研 岗位要求：1、精通 Web 前端技术，包括 HTML/CSS/Javascript 等 2、精通 JS 对象编程，并能熟练使用 jquery 进行动态网页开发 3、有基于 Ajax 或 Jsonp 的开发经验 4、对 NodeJS / Html5 及其相关技术有一定了解；或者熟悉 HTTP 协议、Apache 模块、cookie 等 Web 技术； 5、技术视野广阔，乐于不断学习新知识与新技术，并能应用到实际工作中 6、个性乐观开朗，逻辑性强，乐于团队合作 7、 Web 前沿技术研究和新技术调研	2
岗位职责： 1、负责日常款项支付，境内外网银转帐汇款，处理日常往来账核对； 2、负责公司及分公司费用报销的核查及支付； 3、负责公司资金日报、周报、月报填报及发送，不定时汇报资金异常情况；4、负责总部及所有分公司的网银管理、银企对账、及资金分配； 5、负责银行的日常对接和沟通，负责各类账户问题处理，如开、销户、外汇结售汇等； 6、完成领导安排的其他工作。 任职要求： 1，出纳工作经验在 5 年左右或以上，最少 3 年最多不超过 15 年；2、工作稳定，不频繁跳槽的 4，有境外银行出纳工作经验优先 6，大专或以上优先 2、电脑操作熟练；熟	3

练使用财务软件、Word、Excel 及运用其中各类函数，数据透视表等； 3、工作细致，责任感强，具有良好的学习能力、独立工作能力和财务分析能力； 4、具备良好的沟通能力和团队精神，客观公正，保守秘密； 5、广州本地户口优先考虑。	
岗位职责： 1、负责日常款项支付，境内外网银转帐汇款，处理日常往来账核对； 2、负责公司及分公司费用报销的核查及支付； 3、负责公司资金日报、周报、月报填报及发送，不定时汇报资金异常情况；4、负责总部及所有分公司的网银管理、银企对账、及资金分配； 5、负责银行的日常对接和沟通，负责各类账户问题处理，如开、销户、外汇结售汇等； 6、完成领导安排的其他工作。 任职要求： 1、出纳工作经验在 5 年左右或以上，最少 3 年最多不超过 15 年；2、工作稳定，不频繁跳槽的；3、有境外银行出纳工作经验优先；4、大专或以上优先；5、电脑操作熟练；熟练使用财务软件、Word、Excel 及运用其中各类函数，数据透视表等； 6、工作细致，责任感强，具有良好的学习能力、独立工作能力和财务分析能力；7、具备良好的沟通能力和团队精神，客观公正，保守秘密；8、广州本地户口优先考虑。	3
岗位职责： 1、无线通信系统性能测试与分析，包括常见标准无线通信系统与私有协议标准系统; 2、射频系统性能预算与系统仿真，系统应用性能分析; 3、射频生产自动测试夹具与测试方案及测试算法开发。 4、板级性能分析与仿真，包括频率规划,链路分析,射频失真分析及系统共存(板级或系统级); 5、射频器件选型，射频电路仿真设计与调试; 任职资格: 1、熟悉射频前端芯片常见测试与校准方法; 2、扎实的射频微波理论和测试经验，熟悉使用常长的射频仪器;具有常见射频电路（如 mixer,pll,lna,vga,pa 滤波器）匹配设计经验，熟练使用射频电路仿真工具; 3、具有射频系统仿真及建模工具（如 ads ptolemy，system vue,awr vss 或 matlab）的使用经验;	4

4、熟悉射频系统性能仿真与预算方法;5、熟悉射频系统生成自动化板级及系统级测试	
岗位职责： 1、无线通信系统性能测试与分析，包括常见标准无线通信系统与私有协议标准系统; 2、射频系统性能预算与系统仿真，系统应用性能分析; 3、射频生产自动测试夹具与测试方案及测试算法开发。 4、板级性能分析与仿真，包括频率规划，链路分析，射频失真分析及系统共存（板级或系统级）; 5、射频器件选型，射频电路仿真设计与调试; 任职资格: 1、熟悉射频前端芯片常见测试与校准方法; 2、扎实的射频微波理论和测试经验，熟悉使用常长的射频仪器;具有常见射频电路（如 mixer,pll,lna,vga,pa 滤波器）匹配设计经验，熟练使用射频电路仿真工具; 3、具有射频系统仿真及建模工具（如 adsptolemy，systemvue,awrvss 或 matlab）的使用经验; 4、熟悉射频系统性能仿真与预算方法; 5、熟悉射频系统生成自动化板级及系统级测试	4
岗位职责： 1. 独立并高质量完成来访客户的接待，向其介绍课程及相关服务，并促成报名； 2. 耐心、详细解答电话咨询的客户,邀约意向客户到中心咨询； 3. 挖掘客户的潜在需求，对课程做出准确推荐； 4. 根据公司制定的销售目标，配合团队完成销售任务； 5. 配合其他部门完成工作。 任职条件： 1. 乐于从事电话销售行业以及客户服务相关工作，具备良好的客户服务意识，良好的自我激励能力； 2. 学习能力强、主动积极、有团队合作精神； 3. 思维敏捷，反应快，普通话标准； 4. 具备良好的理解能力、表达能力和沟通能力，以结果为导向； 5. 能够承受工作压力，对电话销售抱有极高的热忱; 6. 专科及以上学历,有过电话销售经验优先，能力突出者提升为主管	5
工作职责： 1、参与需求的分析和设计，并完成相关技术文档	6

的编写； 2、协助项目开发负责人完成功能原型及说明文档的编写； 3、实现业务模块的开发，按时完成编码任务，对代码质量负责； 4、配合测试部门完成测试环境的搭建，以及对相关业务的培训，并完成测试问题的修改和调试知识技能 任职资格： 1、熟练掌握 PHP、HTML、javascript、CSS； 2、熟悉 JavaScript、Ajax、XML 等相关技术； 3、熟悉 oracle 或 mysql 数据库，有相关数据库编程经验,熟练掌握 SQL 语句，尤其是性能把握； 4、熟悉 Linux 操作系统，掌握常用的 Linux 命令 、shell 脚本； 5、较强的学习能力，工作认真积极，有良好的沟通表达方式，较强的抗压能力 6、具备良好的文档书写习惯与表达能力和沟通能力； 7、责任心强，吃苦耐劳，诚实守信，有较强的责任心及团队合作精神； 8、能独立带领团队完成项目的研发工作。 9、有微信公众平台相关开发经验者优先 素质要求：良好的沟通协调能力及领导力，擅于创新，能够持续改进。	

我们利用上述招聘信息数据，利用数据挖据，建立文本聚类模型，探索出人才的需求中的职业类型，并以此对现今人才市场的人才需求状况展开分析。

3.2.2 对所有行业的宏观分析

A 选择数据

我们选择附件 1 作为需要分析的表格，在如此多的职业中，为了对各职业有一个初步的了解，我们对不同职业做文本聚类，得到初步的职业分类。

B 数据处理

在进行文本聚类时，需要对数据矩阵进行距离的计算，因此为了得到矩阵，首先需要对文本使用分词函数进行分词，将句子分成词语，并且对一些虚词，连词等使用停词函数去除，将维度进行约减，得到文字矩阵后数字化，得到数据矩阵，需要用到的算法有层次算法，划分算法，欧氏距离等其他算法。

C 得到聚类结果（代码在附件）

运行的结果如下：

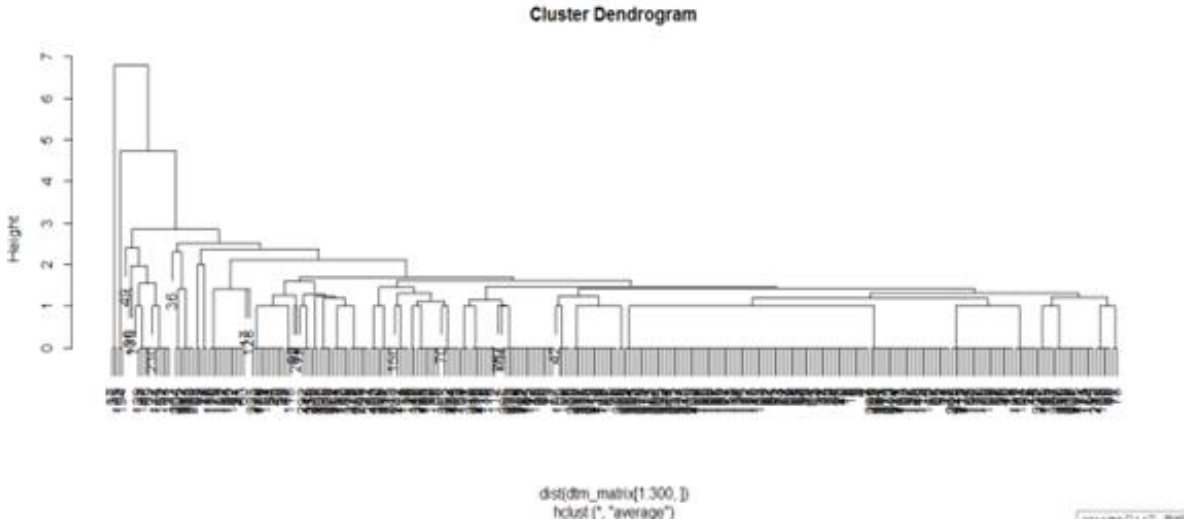


图 1 聚类图

3.2.3 结果及分析

表 1 职业类型表

技术	销售	管理	服务	金融	其他	实习
后端开发	销售	运营	客服	投融资	采购	助理
前端开发	公关	产品经理	法务	风控	供应链	
移动开发	市场/营销	运维	售前咨询	投资		
视觉设计	高端市场职位	人力资源	售后客服	高端金融职位		
测试	高端产品职位	行政	内容编辑	审计税务		
编辑	销售专员	财务	前台	会计		
高端技术职位	销售经理	项目管理	采购专员	投资顾问		
企业软件	市场策划	高端运营职位				
硬件开发	市场营销	高端职能职位	文秘			
dba	品牌公关	运营专员	物流			

交互设计	文案策划	运营经理	系统			
用户研究	电话销售	电商产品经理				
高端设计职位	渠道销售	人事/hr				
产品设计师	销售助理	市场顾问				
java	销售总监	产品助理				
web 前端	商务渠道	培训经理				
平面设计师		客服经理				
测试工程师		项目经理				
it 支持		大客户代表				
数据分析师		客户代表				
游戏策划		媒介经理				
嵌入式		市场总监				
网页产品设计师		总助				
ui 设计师		ceo				
架构师		系统集成				
运维工程师		项目助理				
sqlserver		项目总监				
系统工程师		bd 经理				
系统集成		采购经理				

上表为文本聚类的结果总结，由上表统计可得，职业类型可以分为技术、销售、管理、服务、金融、其他及学习这六个主要类别。我们可以看到技术类的和管理类下的行业分类很多，说明在这两个行业需要的人才类型广泛，可以发展的方向很多，有较多的选择余地，因此在招聘信息中，这些职业的招聘会更加常见，这也符合行业的要求。

3.3 热门招聘城市、行业和岗位的分析

本次源数据为 530544 条数据，利用 excel 对招聘城市、招聘行业、招聘岗位进行频数分析，并对高频出现的城市、招聘行业、以及岗位进行与时间的相关分析，进而了解招聘的热门城市、行业以及岗位。

分析中，对数据进行了预处理，在城市的信息中，无缺失值，在 530444 条信息中得到以下排序，我们只取前九项：

表 2 高频城市 top9 表

City	汇总
北京	197258
成都	15521
广州	46370
杭州	41654
南京	8559
厦门	5901
上海	88408
深圳	64516
武汉	9960
总计	509466

故在招聘的城市中，需求量较大的城市排序为北京、成都、广州、杭州、南京、厦门、上海、深圳、武汉。 我们也看到了大城市对不同职业需求量比起小城市更大，因此在大城市中，求职者获得工作的也对相对小城市的机会多。

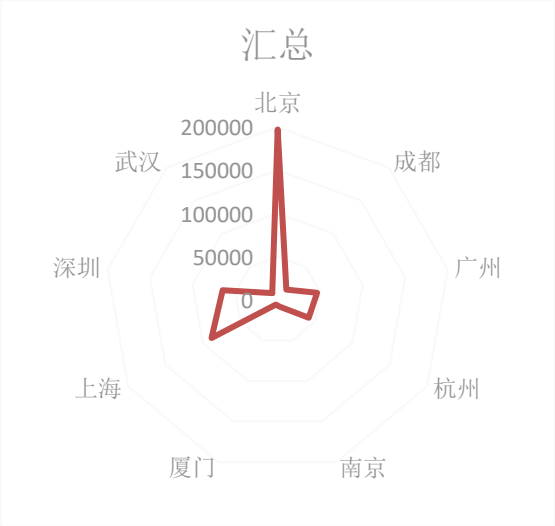


图 2 城市分布雷达图

由上图可得高频城市中，各个城市之间的占比对比，北京对于人才需求的数量相对来数比较大。同时也发现，北上广深依旧是人才需求量大的城市。

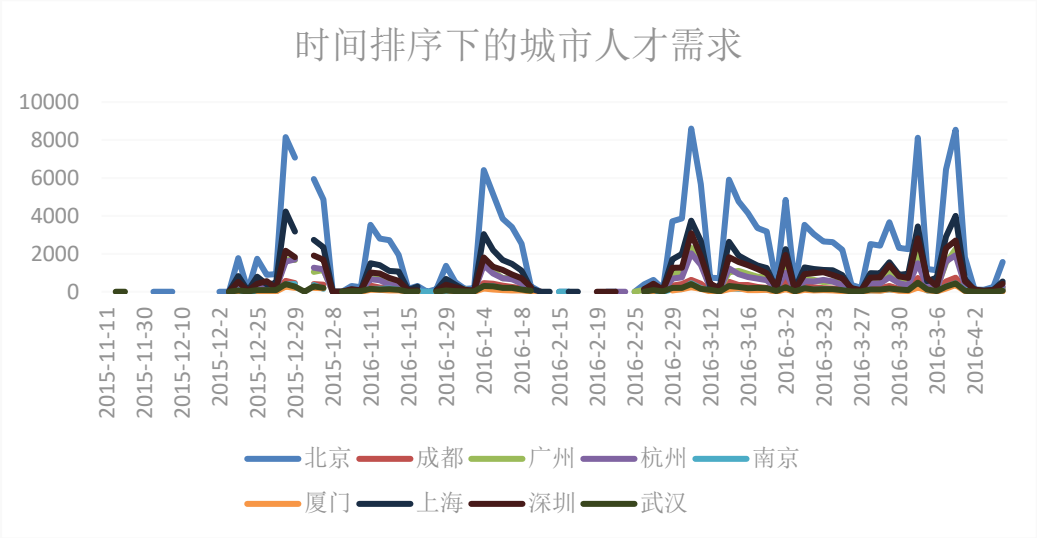


图 3 城市人才需求图

高频出现的招聘城市，只能代表该城市对于人才的需求数量上的大，根据对热门招聘城市的理解--现今需求比期望需求大，故，我们通过 Excel 数据整理如上图可得，在时间的排序下城市对于人才需求的数量的折线图如上，，在 2016 年 2 月以后的城市对于人才需求在时间上相对来说比较频繁，这也反映了一个社会现象，在春节前，较多的职员辞职，因此在春节结束后各公司需要加大在招聘人数，而且这个现象在北上广深更加显著，这也从侧面说明了这几个城市为招聘

城市的热门城市。

在行业的信息中，有效信息为 530431，有 13 个缺失值，在 530444 条信息中得到以下排序，我们只取前九项：

表 3 高频行业 top9 表

IndustryField	汇总
电子商务	29752
金融	39410
移动互联网	75401
移动互联网 • 020	29997
移动互联网 • 电子商务	42227
移动互联网 • 教育	12740
移动互联网 • 金融	25434
移动互联网 • 企业服务	15489
移动互联网 • 数据服务	17444
合计	448949

故在招聘的行业中，需求量较大的招聘行业排序为电子商务、金融、移动互联网、移动互联网 • 020、移动互联网 • 电子商务、移动互联网 • 教育、移动互联网 • 金融、移动互联网 • 企业服务、移动互联网 • 数据服务。

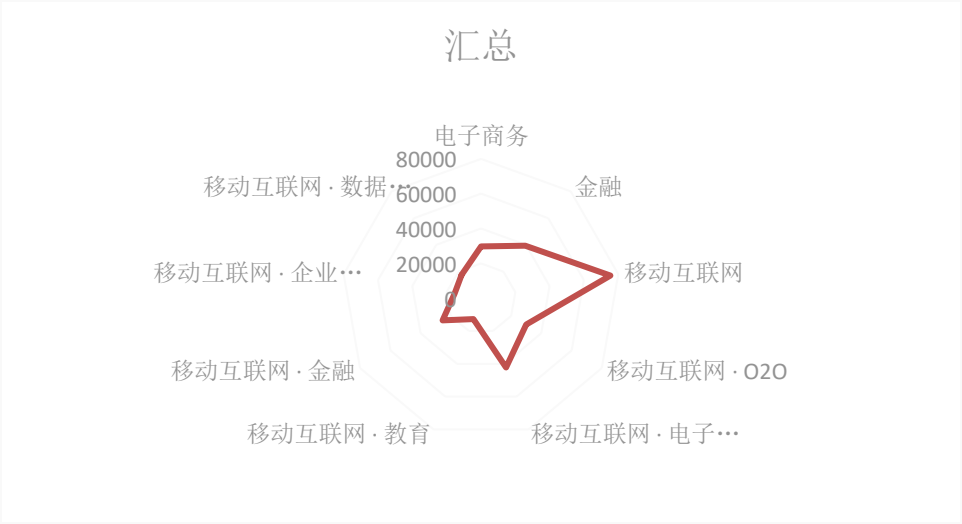


图 4 行业分布雷达图

由上图可得各个招聘行业之间的占比对比，互联网对于人才需求的数量相对来数比较大。

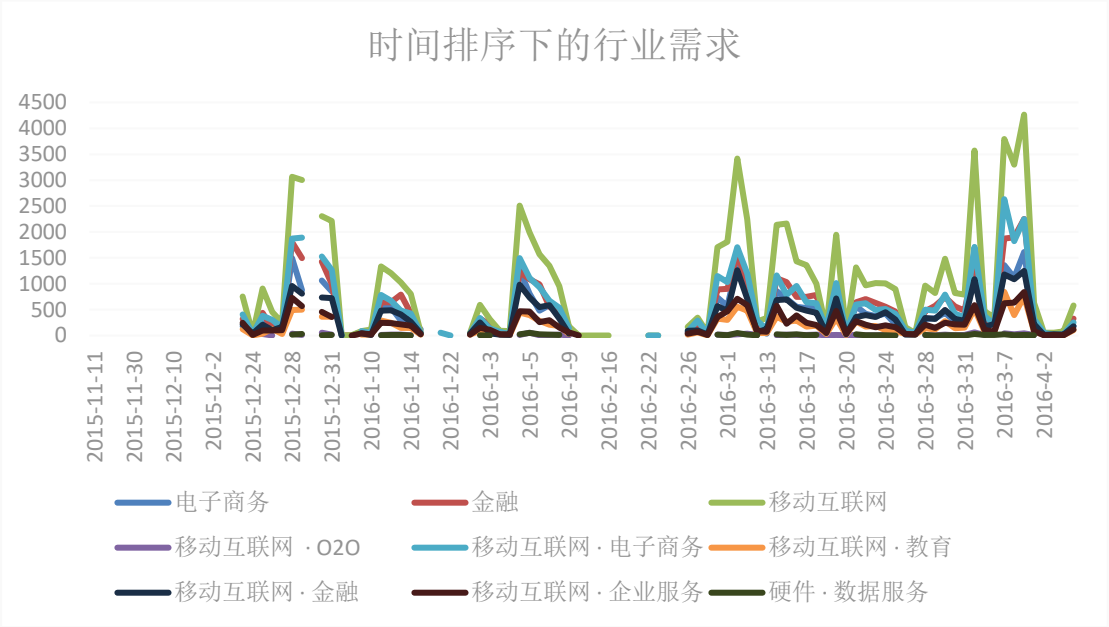


图 5 行业需求折线图

高频出现的招聘行业中，只能代表该行业对于人才的需求数量上的大，根据对热门招聘行业的理解--现今需求比期望需求大，故，我们通过 Excel 数据整理如上图可得，在时间的排序下招聘行业对于人才需求的数量的折线图如上，，在 2016 年 3 月以后的招聘行业对于人才需求在时间上相对来说比较频繁，这也可以理解为，这几个招聘行业的热门招聘行业。

在职业的信息中，有效信息为 530384，有 60 个缺失值，在 530384 条信息中得到以下排序，我们只取前九项：

表 4 高频职位 top9 表

PositionType	汇总
测试	19218
产品经理	30234
后端开发	91999
前端开发	33178

市场/营销	38231
视觉设计	30720
销售	57235
移动开发	31125
运营	53880
总计	499302

故在招聘的岗位中，需求量较大的招聘的岗位排序为测试、产品经理、后端开发、前端开发、市场/营销、视觉设计、销售、移动开发、运营。

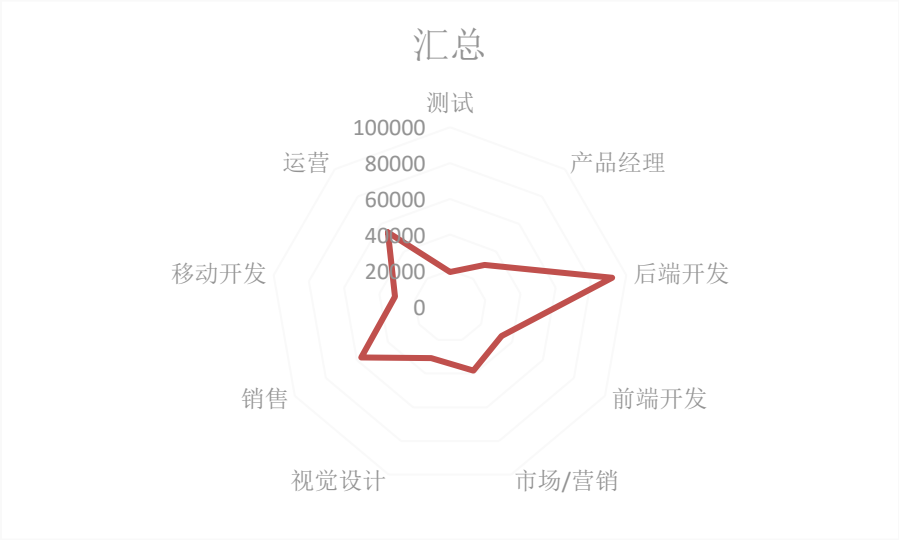


图 6 岗位分布雷达图

由上图可得各个招聘岗位之间的占比对比，后端开发对于人才需求的数量相对来数比较大。

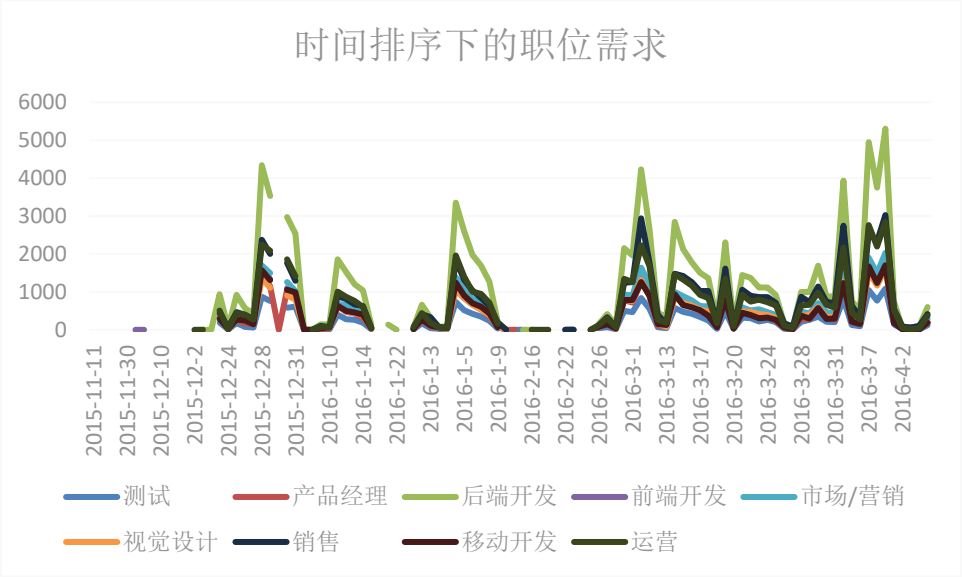


图 7　职位需求折线

故同理，我们通过 Excel 数据整理如上图可得，在时间的排序下招聘岗位对于人才需求的数量的折线图如上，在 2016 年 3 月以后的招聘岗位对于人才需求在时间上相对来说比较频繁，这也可以理解为，这几个招聘岗位的热门招聘岗位。

热门城市、行业以及岗位的走向分析：由上可得，热门的城市有北京、成都、广州、杭州、南京、厦门、上海、深圳、武汉；热门的行业为电子商务、金融、移动互联网、移动互联网 • 020、移动互联网 • 电子商务、移动互联网 • 教育、移动互联网 • 金融、移动互联网 • 企业服务、移动互联网 • 数据服务；而热门的岗位为测试、产品经理、后端开发、前端开发、市场/营销、视觉设计、销售、移动开发、运营。可以看出，在互联网的快速发展下，企业发展快，数量也跟着多，对相关的人才需求自然也变多。

3.4 对新兴职位的分析以及发展预测——数据分析师

3.4.1 对数据分析师职业分析

根据源文件所提供的招聘信息，随着数据分析以及数据挖掘行业的兴起，涌现除了一些新的职位，如数据分析师、数据挖掘工程师等等。我们这里我们将招聘信息对数据分析师进行筛选，共 50 条记录，并进行文本聚类以及词云分析。

截取的部分数据如下：

	Job_Description	PositionId	PositionName
1	岗位名称数据挖掘工程师岗位职责1.负责业务数据收集与整...	277	大数据数据分析师
2	大数据数据分析师 岗位职责 1.负责业务数据收集与整理，...	277	大数据数据分析师
3	岗位职责： 1、面向游戏业务数据，通过数学建模进行分析...	908	游戏数据分析师
4	岗位职责 - 对产品、运营、市场数据进行分析、监测、统计...	920	数据分析师
5	岗位职责 - 对产品、运营、市场数据进行分析、监测、统计...	920	数据分析师
6	岗位职责 1.负责构建数据分析和监控体系，为产品和技术部...	1206	数据分析师
7	岗位职责 1.负责构建数据分析和监控体系，为产品和技术部...	1206	数据分析师
8	工作职责： 1. 负责规划和设计构建 大学生信用模型建设，...	1582	数据分析师
9	工作职责： 1. P2P借贷行业数据分析，报告撰写及相关研究...	1893	数据分析师/研究员
10	工作职责： 1.P2P借贷行业数据分析，报告撰写及相关研究...	1893	数据分析师/研究员
11	1. 参与项目技术方案设计与需求分析，根据方案与需求进行...	1976	数据分析师

对每一个文本向量进行关键词提取，截取部分数据如下：

```
    33.5779     33.4695       29.68     29.0143     25.5956
"数据系统"      "数据" "数据分析"      "相关"      "业务"
    33.5779     33.4695       29.68     29.0143     25.5956
"数据系统"      "数据" "数据分析"      "相关"      "业务"
70.4352 33.4695 20.4765 19.7414 18.8017
  "amp"   "数据"   "业务"   "能力"   "分项"
70.4352 33.4695 20.4765 19.7414 18.8017
  "amp"   "数据"   "业务"   "能力"   "分项"
```

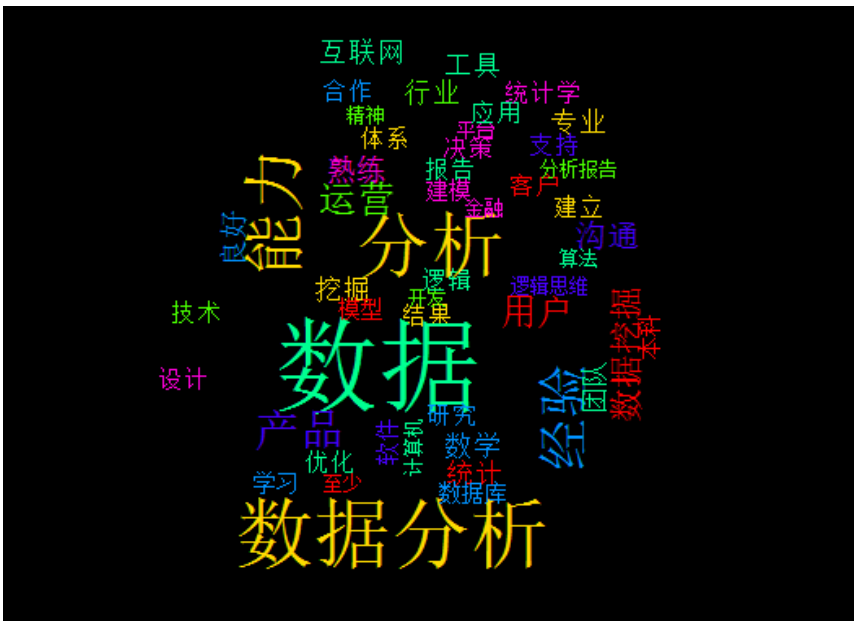
上图可以看出，对每个文本向量进行提取得不出数据分析师这个职业的重要信息，于是我们将对所有数据分析师的招聘岗位信息进行词云分析。

对文本进行分词，并转换为文本矩阵。截取部分数据如下：

```
[4015] "固化"       "分析"       "加入"       "分析"       "系统"       "5."
[4021] "监督"       "数据"       "信息安全"   "数据"       "质量"       "6."
[4027] "建立"       "信用风险"   "评级"       "模型"       "等"         "风控"
[4033] "模型"       "任职"       "需求"       "1."         "本科"       "及"
[4039] "以上学历"   "数据"       "统计"       "或"         "数据挖掘"   "专业"
[4045] "方向"       "信息"       "数学"       "计算机相关" "专业"       "2."
[4051] "熟练"       "运用"       "数据处理"   "及"         "基础"       "数据分析"
[4057] "方法"       "擅长"       "利用"       "模型"       "进行"       "分析"
[4063] "预测"       "3."         "熟练"       "使用"       "EXCEL"      "操作技能"
[4069] "掌握"       "SPSS"       "SAS"        "ACCESS"     "或"         "VBA"
[4075] "优先"       "4."         "为"         "人"         "耐心"       "细致"
[4081] "工作"       "认真"       "学习"       "能力"       "强"         "能够"
[4087] "主动"       "沟通"       "分享"       "5."         "良好"       "的"
[4093] "数据"       "敏感度"     "能"         "从"         "大量"       "数据"
[4099] "提炼"       "核心"       "结果"       "有"         "数据分析"   "挖掘"
```

由于上述信息中有很多我们在分析中不需要的词语，如：“以上学历”、“为”、“有”等等以及数字，在这里我们创建自己的停用词，与我们的信息匹配并去除。

最后创建出出现频率最高的 50 个词，如下：



由图 可以看出，在上述的重要信息中，数据、数据分析、能力、分析、经验等的文本信息字体比较大，可以得出在数据分析师的招聘信息中，招聘单位对该职位的人才对能力、数据分析、经验要求比较多，这些都可归结为能力，能力是可以在时间的累积下不断增加的。对比起能力，我们都知道数据分析师的对口专业是统计学，但是跟能力比起来，专业并不是万能的，招聘单位对能力更加看重。

3.4.2 对数据分析发展预测

主要对数据分析师进行在时间排序下的人才需求量的分布进而进行预测。

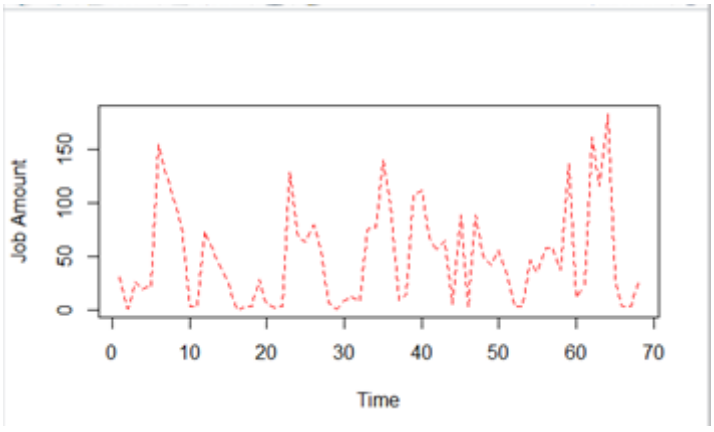


图 8 数据分析师的时间序列图

由上图可得，由于数据在时间的分布下，在同一条直线上下波动，可以得出，

数据分析师的人才需求量与时间的关系不大。在后面我们会直接对数据进行 AcF 图的 Arima 中的 q、p 选取，并对数据进行预测。

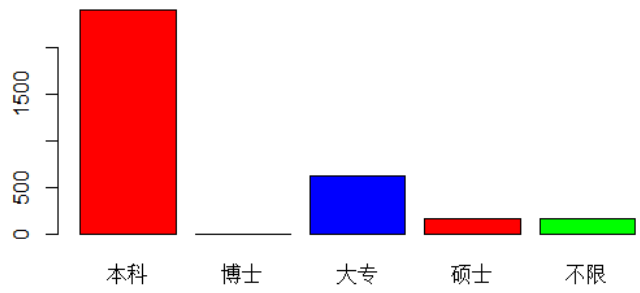


图 9 全部的学历分布图

以上为数据分析师的岗位需求的学历要求，主要是本科为主，大专为辅。这也告诉我们虽然数据分析师看起来很高大上，但是招聘更多的需要还是本科生和大专生，比起事前的学习，入职后的培训更加有效。

表 5 全部地区分布 表

	summary(data$City)
北京	197258
上海	88408
深圳	64516
广州	46370
杭州	41654
成都	15521
武汉	9960
南京	8559
厦门	5901
西安	4950
长沙	4935
苏州	4374
郑州	4203
天津	3683
重庆	3105
合肥	2380
青岛	2184
福州	1984
珠海	1934
济南	1919
东莞	1506
佛山	1303
无锡	1236
大连	1133
宁波	939
沈阳	763
南昌	757

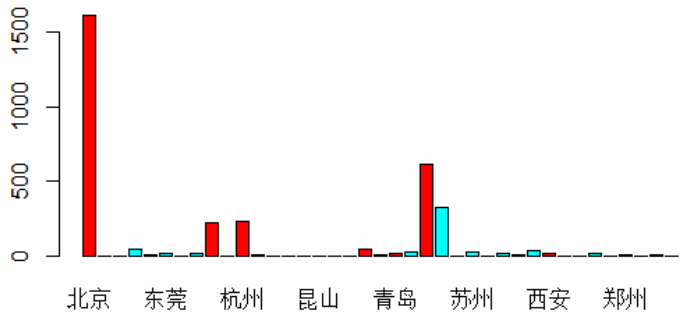


图 10 全部的地区分布图

由图 10 和图 11 可以看出数据分析师的人才需求量在地区的分布，以及数量分布图。热门城市仍旧是北上广深。

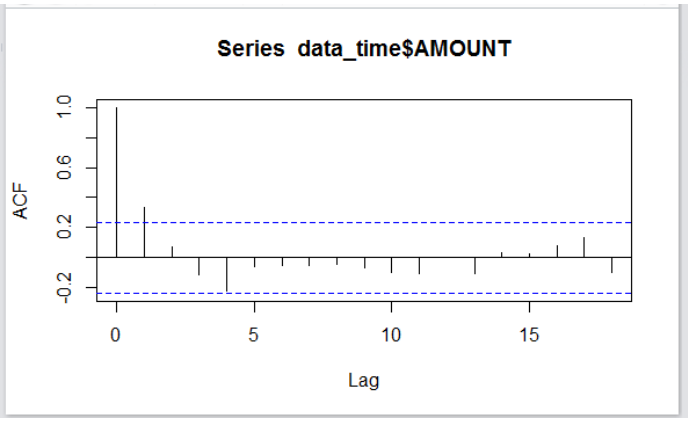


图 11 数据分析师的 ACF 分布图

自相关图显示滞后 1 阶自相关值超过边界值，虽然 4 阶自相关值超出边界，那么很可能属于偶然出现的，而自相关值在其他上都没有超出显著边界。

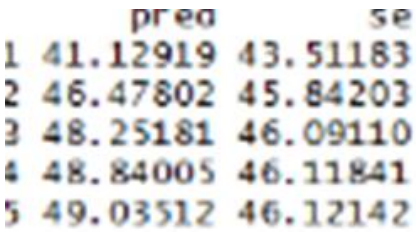


图 12 数据分析师 5 期预测的结果图

预测值如上图所示，一年后的预测为 41.12919，而五年后的数据分析师的人才需求量会在一定程度上向上增长为 49.03512 万，可以得出随着大数据时代的

到来，数据分析的岗位在未来五年内的仍会有一定的需求增长空间。

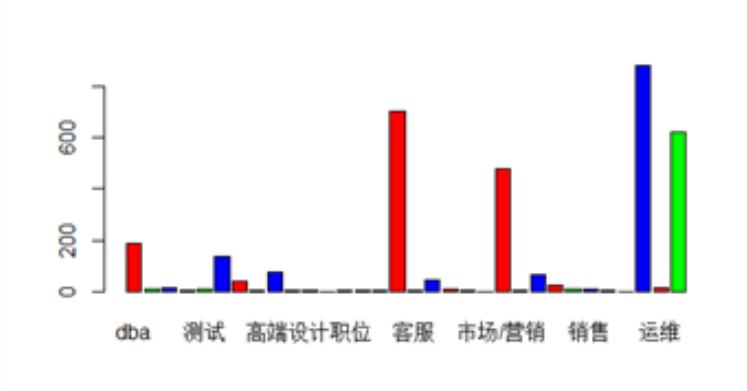


图 13 数据分析师的 POSITIONFIRSTTYPE 的分布

以上为数据分析师的职业类别的分布，主要以运营、客服等服务类的工作为主。

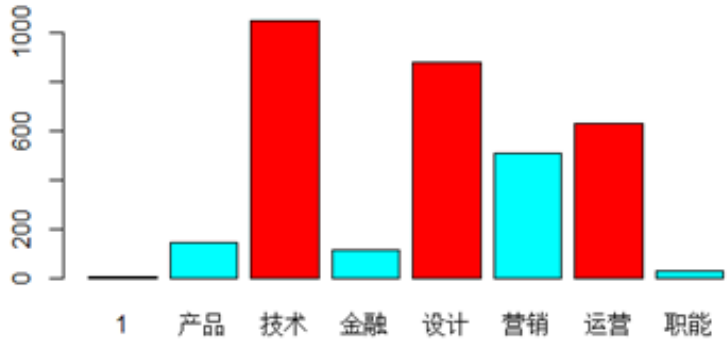


图 14 数据分析师的 TYPE 分布图

而对数据分析师的种类分布中，主要以技术、设计和营运为主。

3.5IT 人才市场的分析

3.5.1 行业分析

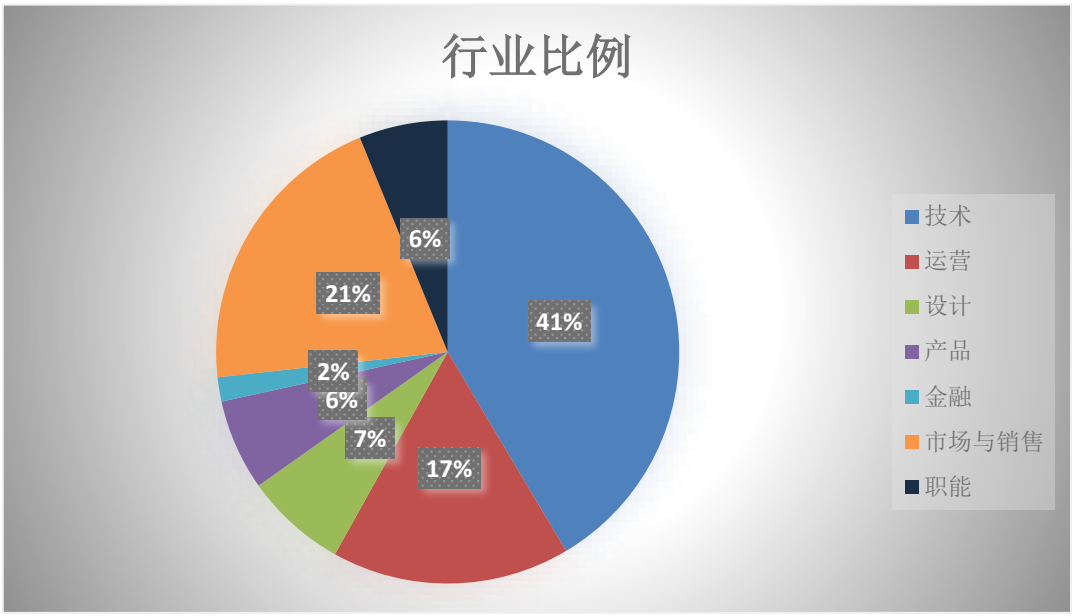


图 15 行业比例饼图

从所给的数据可以分析出 IT 行业在这些数据里面所占的比例为 41%+17%+7%=65%,所有现在的 IT 人才市场的需求很大

3.5.2 行业地域

图 16 IT 行业地域分布图

根据设计行业分类汇总的情况，我们算出了超过 100 个职位的地方主要有以上的地方。主要是分布在一些经济发达的地区，主要是在中国的东部。北上广深，杭州，成都等这些地方的科技比较发达，所以对 IT 行业的人才需求比较多。

3.5.3 学历层次

表 6 学历层次汇总表

博士	92	0.03%
硕士	3024	0.88%
本科	170877	49.49%
大专	125090	36.23%
中专	0	0.00%
高中	2	0.00%
初中	0	0.00%
学历不限	46197	13.38%
总计	345282	100%

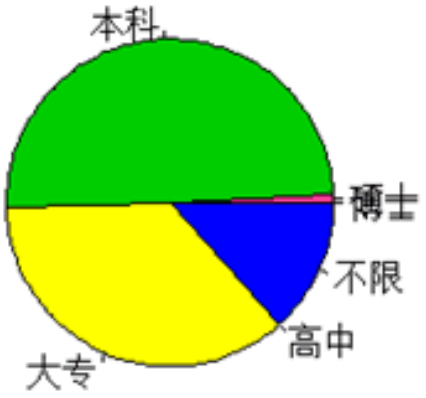


图 17 学历层次饼图

从学历层次的图可以看到，刚刚好友一半的职位是要求是本科的学历的，其次是大专的要求，再有就是也有相当一部分的职位是没有学历要求的。而硕士和博士的要求也占了 10%。

3.5.4 职位

表 7 职位汇总表

PositionType	数量	百分比
移动开发	31125	9.01%
前端开发	33178	9.61%
dba	3907	1.13%
客服	12930	3.74%
交互设计	3144	0.91%
视觉设计	30720	8.90%
运营	53880	15.60%
后端开发	91999	26.64%
运维	14137	4.09%
企业软件	5015	1.45%
编辑	18385	5.32%
高端技术职位	11017	3.19%
测试	19218	5.57%
高端设计职位	1846	0.53%
高端运营职位	2891	0.84%
项目管理	5831	1.69%
硬件开发	4087	1.18%
用户研究	1972	0.57%
总计	345282	100.00%

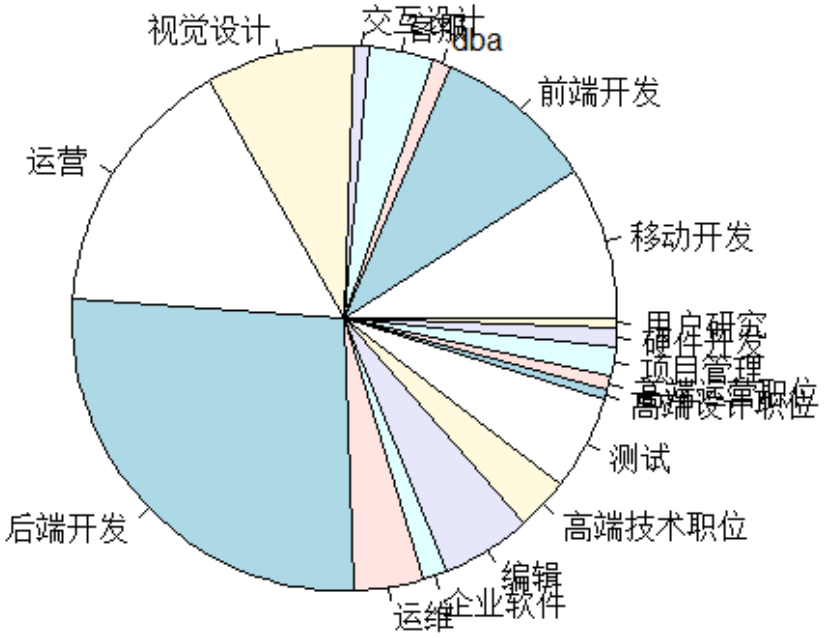


图 18 职位分布并图

从职位这个表格我们可以看出来在目前，后端开发，前端开发，移动开发，运营，视觉设计，测试这几个职位很热门。

3.5.5 未来发展趋势

表 8 预测表

时间	数量
2015 年 11 月	6
2015 年 12 月	60856
2016 年 1 月	69634
2016 年 2 月	11495
2016 年 3 月	197454
2016 年 4 月	5836
统计	345282

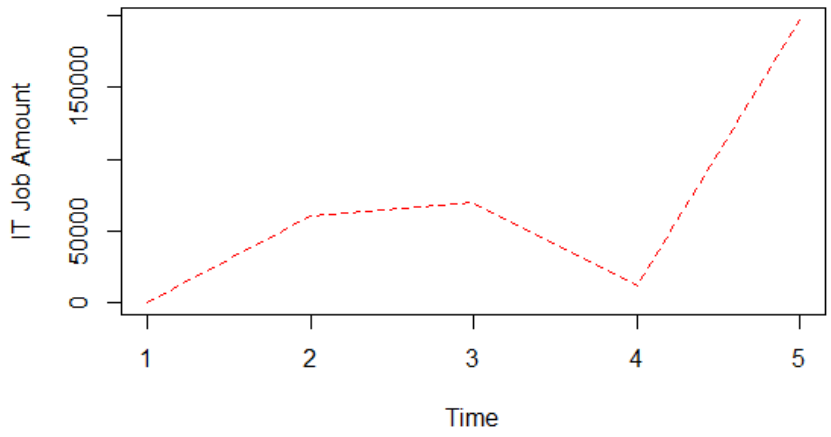


图 19 IT 行业的人才的需求折线图

由于 2015 年 11 月份和 2016 年 4 月的数据不完善，所以我们把它去掉，得到的图可以知道，在 2016 年的 2 月到 3 月是急速上升的，所以我们预测在未来的几个月里面对 IT 行业的人才的需求是急增的。

4 对学校人才培养计划的建议

为了让我们学习的内容更加符合用人单位的要求，我们根据所得到的结果针对我们所学的专业给出了以下的建议：

根据我们小组的情况,有统计和信息与计算科学.我们通过附件 3 的职位描述来刷选出这两个专业的工作职责,岗位职责,任职要求.根据任职要求,我们分析出每个岗位的技术要求 都有要求熟练 excel 这个电子表格软件。现在，无论是信息与计算科学还是统计专业都没有开设这个软件的学习年课程。我身边有好多的同学都不会用 excel，就连很基本的一些功能都不知道，这样的办事效率会大打折扣。而且我们日常生活接触更多的是 excel。我们现在专业学习的软件的技术性都很强,在日常生活上应用的机会很少。我们可以用 excel 处理批量的数据。

针对信息与计算科学这个专业，我们分析出用人单位给出的岗位职责主要是 IT 和数据处理、分析这两个方面。其中我们分析了一下 IT 岗位的要求，除了那些基本之外，相当一部分的单位是要求应聘者熟练 linux 系统，而这个专业是没

有开设这个课程的，所以我们建议开设这门课程。还有，我们在对附加 1 的招聘信息进行分析后，发现前端开发，还有测试这两个职位是比较热门的。但是，这个专业对前端的课程只是在前几年增加了一门。前端开发涉及的东西很多，而我们觉得增加这个方面的课程作为选修。还有测试也是在软件工程工作量里面占了很大的一部分。所以我们建议要开设一门专门的课程对我们进行专业的培训。由于信息与计算科学这个专业还有一个方向可以走，但是在这个方向的要求是至少会一种的数据分析的软件，而我们是没有学习这些的，对于这些同学来说，学习起来会很吃力。所以我们建议开设 R 软件的课程。

统计这个专业走数据分析，还有数据挖掘这两个方向比较多。但是考虑到那些分析需要建模，还有算法的分析的。所以我们建议开设这两门课程作为必修课。

5 参考文献

[1].王立伟.数据挖掘研究现状综述[J].图书与情报,2008,(5):41-43.
[2].安淑芝.数据仓库与数据挖掘[M].北京：清华大学出版社,2005:25-70.
[3].吴婕.浅析数据挖掘软件的发展[J].情报理论与实践,2004,(2):211-214
[4]谌志群,张国煊.文本挖掘研究进展[J].模式识别与人工智能. 2005(01)
[5]周文霞主编.职业生涯管理[M]. 复旦大学出版社, 2004

6 附录

```
library(rJava)
library(Rwordseg)
library(tm)
d=as.character(data3$Job_Description)
words=lapply(d,removeNumbers)
words=gsub(pattern="[a-zA-Z]+","",words)
wordsegment=function(x){
   library(Rwordseg)
   segmentCN(x)
}
words=lapply(words,wordsegment)
corpus=Corpus(VectorSource(words))
meta(corpus,"cluster")=data3$Job_Description
(words.dtm=DocumentTermMatrix(corpus,control=list(wordLenghts=c(1,Inf))))
words.dtm2=removeSparseTerms(words.dtm,sparse=0.9)
dtm_matrix=as.data.frame(inspect(words.dtm2))
dist.dtm<- dissimilarity(words.dtm2, method = "cosine")
job=hclust(dist(dtm_matrix[1:300,]),method="average")
plot(job)


#换表格#
d2=as.data.frame(summary(data$PositionType))
d1=as.character(data$PositionType)
words1=lapply(d1,removeNumbers)
words1=gsub(pattern="[a-zA-Z]+","",words)
wordsegment=function(x){
   library(Rwordseg)
   segmentCN(x)
}
words1=lapply(words1,wordsegment)
corpus1=Corpus(VectorSource(words1))
meta(corpus1,"cluster")=data$PositionType
(words.dtm1=DocumentTermMatrix(corpus1,control=list(wordLenghts=c(1,Inf))))
words.dtm12=removeSparseTerms(words.dtm1,sparse=0.9)
dtm_matrix1=as.data.frame(inspect(words.dtm12))
dist.dtm1<- dissimilarity(words.dtm12, method = "cosine")
job=hclust(dist(dtm_matrix[1:300,]),method="average")
plot(job)
```

```
#简单分析#
barplot(D,col=rainbow(2))
barplot(E,col=rainbow(3))
names(E)=c("本科","博士","初中","大专","高中","硕士","不限","中专")
setwd("E://修改泰迪")
jobdescr=read.csv("E:/修改泰迪/数据分析师.csv",T)
a=as.character(jobdescr$Job_Description)

library(Rcpp)
library(jiebaR)
cutter= worker()
a_segment=cutter[a]
filter_words = c("我","你","它","大家")
filter_segment(a_segment,filter_words)
cutter=worker(bylines = TRUE)
a_segment_line=cutter[a]
cutter$write
cutter["files.path"]

#可以使用 vector_keywords 对一个文本向量提取关键词。
keyworker=worker("keywords",topn=2)#关键词个数
cutter = worker()
vector_keywords(cutter[a],keyworker)

#可以使用 vector_keywords 对一个文本向量提取关键词。
keyworker=worker("keywords",topn=5)#关键词个数
cutter = worker()
for(i in 1:length(a)){
   str1<-a[i]
   print(vector_keywords(cutter[str1],keyworker))
}
mixseg = worker()
mixseg[a]
## 相当于 segment(a,mixseg )
## 或者 mixseg<=a

#词性
tagger=worker("tag")
tagger[a]

#Simhash 与海明距离
simhasher=worker("simhash",topn=2) #关键词的 simhash 距离
```

```
simhasher<=a[1]
distance(a[1],a[2], simhasher)


##快速模式
qseg[a]

library(tm)
a_ovid=Corpus(VectorSource(a_segment_line))
summary(a_ovid)

inspect(a_ovid[1])

library(SnowballC)

a_ovid<-tm_map(a_ovid,removeNumbers)#去除数字
a_ovid<-tm_map(a_ovid,stripWhitespace)#去除多余空格
a_ovid<-tm_map(a_ovid,removePunctuation)#去除标点符号
a_ovid<-tm_map(a_ovid,removeWords, stopwords("english"))#将英文中的停词删掉：例如把
that  at  等英文介词去掉。
a_ovid<-tm_map(a_ovid,PlainTextDocument)#去掉空文件

dtm<-DocumentTermMatrix(a_ovid,control=list(dictionary=as.character(a_segment),removePunc
tuation =TRUE,stopwords=TRUE, wordLengths = c(1, Inf)))
# Punctuation 是否去掉标点符号默认 falseremoveNumbers 是否去掉数字默认 false，
#dictionary 设置要统计的中文词语，如果不设置的话，默认会从所有的语料库里统计
#wordLengths 设置如果词的长度大于 X 时舍去。
dtm2=removeSparseTerms(dtm, sparse=0.9)
df_dtm2<-as.data.frame(inspect(dtm2))#将词频矩阵转换为数据框格式得到

data.frame(inspect(DocumentTermMatrix(a_ovid)))

##文档聚类
d <- dist(dtm2, method = "euclidean")
fit <- hclust(d, method="ward.D")
fit$labels<-jobdescr$PositionId
plot(fit)


library(NLP)
library(tm)
library(rJava)
library(Rwordseg)
library(RColorBrewer)
```

```
library(wordcloud)
library(tmcn)

#创建停止词
mystopwords <- read.table(file = file.choose(), stringsAsFactors = FALSE)
head(mystopwords)
class(mystopwords)
#需要将数据框格式的数据转化为向量格式
mystopwords <- as.vector(mystopwords[,1])
head(mystopwords)
#自定义删除停止词的函数
removewords <- function(target_words,stop_words){
   target_words = target_words[target_words%in%stop_words==FALSE]
   return(target_words)
}
segword <- sapply(X = a_ovid, FUN = removewords, mystopwords)
segword[[1]]
#绘制文字图
word_freq <- getWordFreq(string = unlist(segword))
opar <- par(no.readonly = TRUE)
par(bg = 'black')
#绘制出现频率最高的前 50 个词
wordcloud(words = word_freq$Word, freq = word_freq$Freq, max.words = 50, random.color =
TRUE, colors = rainbow(n = 7))
par(opar)

v = table(unlist(segword)) #计算每个单词的词频
v = sort(v, decreasing = T) #按降序排列
d = data.frame(segword = names(v), freq = v) #将词频矩阵转换为数据框格式
d$segword=as.vector(d$segword) #将单词字段规整为字符串格式
rbind(d[nchar(d$segword)==1,][1:10,],d[nchar(d$segword)==2,][1:20,],d[nchar(d$segword)>=3,][
1:20,])->result_r #提取不同字数的单词中词频最高的 TOP50 单词，作为词云绘制的素材
write.table(result_r,file="E:/修改泰迪/ciyun.csv",sep=",",row.names = F,quote = F)

#问题 3#
data_an=read.csv("E:/T3AN.csv",T)
data_wa=read.csv("E:/T3WA.csv",T)
data_time=read.csv("E:/time.csv",T)
library(tseries)
#时间序列分析接下来的需求情况#
watime=as.data.frame(summary(data_wa$CreateTime))
matplot(data_time,type="l",lty=2,col=2,xlab="Time",ylab="Job Amount")
acf(data_time$AMOUNT)
pacf(data_time$AMOUNT)
```

```
Box.test(data_time$AMOUNT, type="Ljung-Box",lag=6)
m1=arima(data_time$AMOUNT, order = c(1,0,0),method="ML")
summary(m1)
r=m1$residuals
Box.test(r,type="Ljung-Box",lag=6, fitdf=1)
prop.fore = predict(m1, n.ahead =5)
as.data.frame(prop.fore)#predict#
#简单的分析#
#city#
C3=read.csv("E:/CC3.csv",F)
c3=C3$V2
n=as.vector(C3$V1)
names(c3)=n
barplot(c3,col=rainbow(2))
#education#
e3=as.vector(summary(data_an$Education))
names(e3)=c("本科","博士","大专","硕士","不限")
barplot(e3,col=rainbow(3))
#positiontype#
PP3=as.vector(summary(data_an$PositionType))

P3=read.csv("E:/P3.csv",F)
na=as.vector(P3$V1)
names(PP3)=na
barplot(PP3,col=rainbow(3))
plot(PP3,col=rainbow(2))

pf3=as.vector(summary(data_an$PositionFirstType))
names(pf3)=c("1","产品","技术","金融","设计","营销","运营","职能")
barplot(pf3,col=rainbow(2))

#数据挖掘#
wa_time=read.csv("E:/WATIME.csv",T)
matplot(wa_time,type="l",lty=2,col=2,xlab="Time",ylab="Job Amount")
acf(wa_time$AMOUT)
pacf(wa_time$AMOUT)
Box.test(wa_time$AMOUT, type="Ljung-Box",lag=6)
m2=arima(wa_time$AMOUT, order = c(1,0,0),method="ML")
summary(m2)
r1=m2$residuals
Box.test(r1,type="Ljung-Box",lag=6, fitdf=1)
prop.fore = predict(m2, n.ahead =5)
as.data.frame(prop.fore)#predict#
#简单的分析#
```

```
#city#
CC3=as.data.frame(summary(data_wa$City))

CC=read.csv("E:/CC3.csv",F)
barplot(cc,col=rainbow(2))
#education#
E3=as.data.frame(summary(data_wa$Education))
E3=as.vector(summary(data_wa$Education))
names(E3)=c("本科","博士","大专","硕士","不限")
barplot(E3,col=rainbow(3))
#positiontype#
PP3=as.data.frame(summary(data_wa$PositionType))
PP3=as.vector(summary(data_wa$PositionType))

P3=read.csv("E:/ee3.csv",F)
na=as.vector(P3$V1)
names(PP3)=na
barplot(PP3,col=rainbow(3))


pf3=as.vector(summary(data_wa$PositionFirstType))
names(pf3)=c("1","产品","技术","金融","设计","营销","运营","职能")
barplot(pf3,col=rainbow(2))

a=read.csv("E:/1.csv",F)
a=as.vector(a$V1)
names(a)
names(a) <- c("博士","硕士","本科","大专","高中","不限")
pie(a,main = "学历饼图")

a=read.csv("E:/1.csv",F)
b=as.vector(a$V1)
names(b)
names(b)=as.vector(a$V2)
pie(b,main = "职位饼图")

a=read.csv("E:/1.csv",F)
matplot(a$V2,type="l",lty=2,col=2,xlab="Time",ylab="IT Job Amount")
```