数据分析求职指南——猎聘网'数据分析'职位解析

通过 python 编写爬虫，爬取了猎聘网关键词“数据分析”的全国范围内 1 月内发布的企业职位。

本次分析源数据的职位发布日期：2017 年 9 月 19 日-2017 年 10 月 19 日

爬虫介绍：爬取猎聘网职位信息

爬虫及数据分析源代码：Ruiww/LiePinAnalysis

目录

数据分析求职指南——猎聘网'数据分析'职位解析
数据清洗
    薪酬数据
    工作地数据
    发布时间
    异常及空值处理
    用于分析的数据
数据分析
    分析思路
    招聘需求分布
        职位集中于计算机、金融领域，互联网·电商行业职位数占比达 1/3，保险、银行行业每家公司平均的职位需求最多
        职位集中于一线城市，北京、上海占据总量的近 60%
        需求聚集在京津冀、长三角、珠三角与川渝四个地区，与当前互联网行业的分布基本一致
        中等规模公司需求量最大，1-49 人初创公司的平均招聘需求最旺盛
        未细化的数据分析职位占 2/3，其后依次为大数据、数据挖掘与数据运营
    薪酬与福利
        数据分析薪酬在前四年稳步增长，4-6 年出现瓶颈期，此后出现较大分化
        福利水平较高，社保、假期与发展空间相关词出现最频繁
    技能需求
        数据分析：SQL、EXCEL、统计学
        数据运营：EXCEL、分析报告、SQL
        数据挖掘：PYTHON、统计学、R
        大数据：HADOOP、SPARK、JAVA

数据清洗

观察原始数据（共 27633 条）情况，并使用 MySQL 对爬取的原始数据进行清洗。

数据清洗 Query 代码：datacleaning.sql

数据清洗结果：用于分析的数据

原始数据字段如下所示：

- JobTitle VARCHAR(255) ：职位名称
- company VARCHAR(255)：公司名称
- salary VARCHAR(255)：薪酬
- position VARCHAR(255)：工作地
- PubTime VARCHAR(255)：发布时间
- qualification VARCHAR(255)：职位要求，包含学历、工作经验、语言、年龄
- tag_list VARCHAR(255)：职位标签，'tag1,tag2,...'
- description TEXT：职位描述，详细介绍职位工作内容，任职要求等信息
- industry VARCHAR(255)：行业
- industry_detail VARCHAR(255)：猎聘更加细化的行业区分
- companySize VARCHAR(255)：公司规模，人员数
- is_end INT(255)：职位是否已结束，未结束为 0，结束为 1

薪酬数据

salary 格式如'4-5 万 72 小时反馈'、'面议 24 小时反馈'，包含职位反馈时间信息

- 新建字段 min_salary INT(255)、max_salary INT(255)、average_salary FLOAT
- 从 salary 字段中提取 min_salary、max_salary，并有 average_salary = avg(min_salary,max_salary)
- 薪酬为'面议'时，min_salary = max_salary = average_salary = 0

工作地数据

position 格式为'国家' 、 '城市' 、 '城市-区县'，如：'新加坡'、'北京'、'广州-天河区'，将'-'前后两部分分开

- 新建字段 position1 VARCHAR(255)、position2 VARCHAR(255)
- 若 position 包含'-'，前一部分为 position1，后一部分为 position2
- 若 position 不包含'-'，则 position1 = position，position2 = null

发布时间

PubTime 格式为'XXXX 年 XX 月 XX 日'，是字符串形式

- 新建字段 pdate date
- 提取 pubtime 中年月日，转化成 pdate = '%Y-%m-%d'

异常及空值处理

- 因猎聘网职位详情页源码格式问题，若职位未给出公司规模数据，则爬取得到的公司规模字段实际为公司地址，令出现问题的行 comAddress = companySize，并删除 companySize 字段内容
- industry_detail 字段为空时，令 industry_detail = industry
- comAddress 字段为空时，令 comAddress = position
- companySize、tag_list、description 字段为空，填充'null'

用于分析的数据

根据分析目的筛选用于分析的数据，获得 2404 条

- 因猎聘网搜索较为模糊，结果中包含'会计助理'、'项目运营'等无关职位，对 JobTitle 进行筛选：包含'数据分析'、'大数据'、'数据运营'、'数据挖掘'等
- 只分析未结束职位：is_end = 0
- 去重

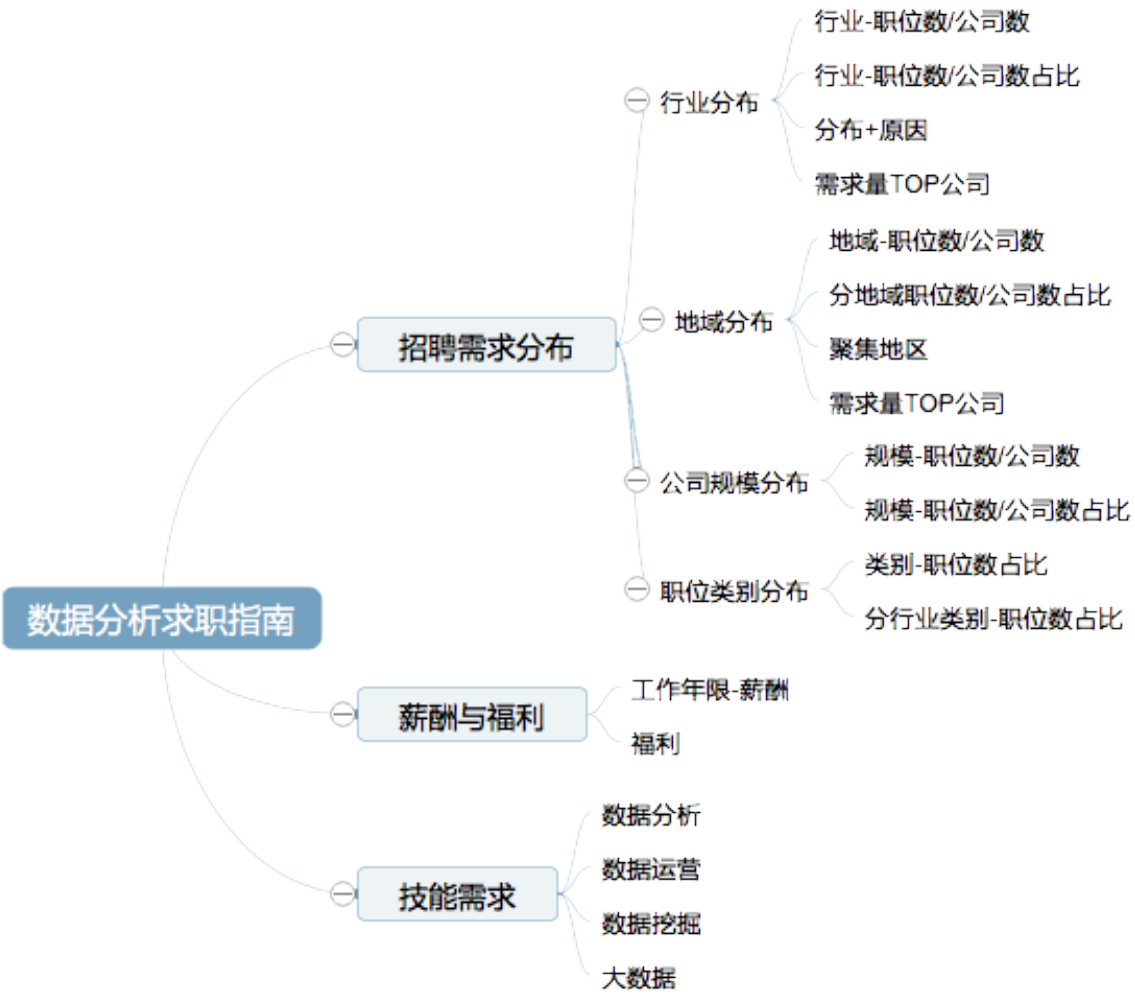
数据分析

数据分析是数学与计算机科学相结合的产物，指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。越来越多的企业将选择拥有项目数据分析经验的专业人士为他们的项目做出科学、合理的分析，以便正确决策项目。

通过偶然的机会，我接触到数据分析的相关理念，并对此产生了兴趣，希望从事相关工作。然而在求职过程中，招聘网站的信息多而纷杂，难以了解数据分析相关职位整体的情况。

通过对猎聘网数据分析职位网页的爬取与分析，形成这份数据分析求职指南，从地域、行业、公司薪酬等方面为数据分析的求职提供参考，并从需求技能层面进行分析，为想要从事数据分析的学习者提供方向。

分析思路

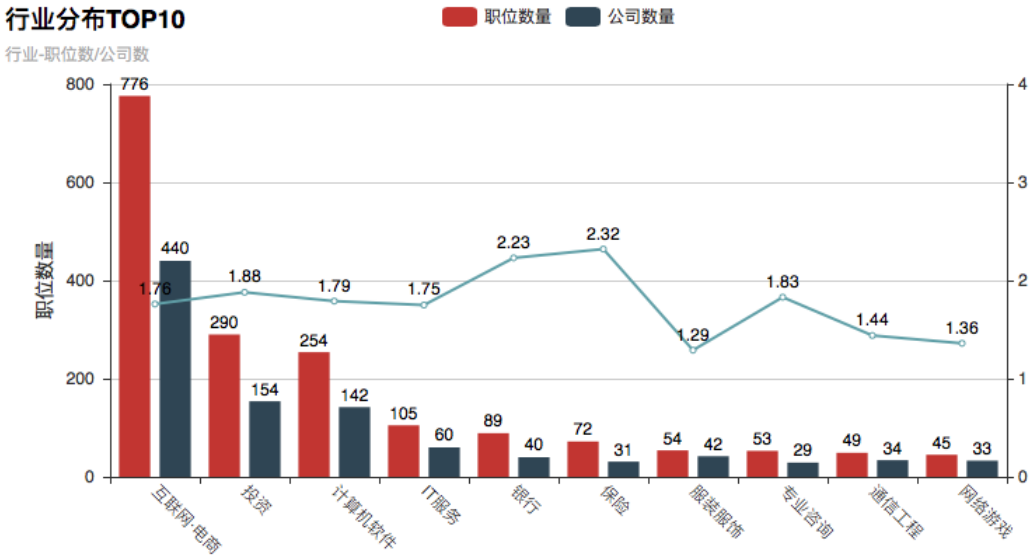

数据分析求职指南
招聘需求分布
行业分布
行业-职位数/公司数
行业-职位数/公司数占比
分布+原因
需求量TOP公司
地域分布
地域-职位数/公司数
分地域职位数/公司数占比
聚集地区
需求量TOP公司
公司规模分布
规模-职位数/公司数
规模-职位数/公司数占比
职位类别分布
类别-职位数占比
分行业类别-职位数占比
薪酬与福利
工作年限-薪酬
福利
技能需求
数据分析
数据运营
数据挖掘
大数据

招聘需求分布

职位集中于计算机、金融领域，互联网•电商行业职位数占比达 1/3，保险、银行行业每家公司平均的职位需求最多

数据分析的行业分布图显示，职位需求 TOP10 中除服装服饰行业外，均为计算机、金融、咨询等日常工作中会产生以及需要处理大量数据的领域，与通常的印象相符。

图中折线各点数值为相应行业职位数量/公司数量，衡量各行业平均每家公司对数据分析职位的需求程度。该指标与各行业需求职位数量的总数并不成正比，保险、银行大于 2，明显高于平均水平，而服装服饰、通信工程和网络游戏行业需求程度较低。



TOP10 中服装服饰行业需求超过专业咨询排到第 7 位令人意外，具体研究该行业的职位，发现岗位职责基本是对货品、销售数据进行整理分析，一个典型的服装服饰行业数据分析职位的职责如下所示。

凡人优品 - 数据分析主管

1、依据公司年度目标和年度计划，协助上级制定、规划、实施运营分析部的年度工作目标和年度工作计划，规划全年商品数据销售统筹及分解实施，制定新品上市计划；

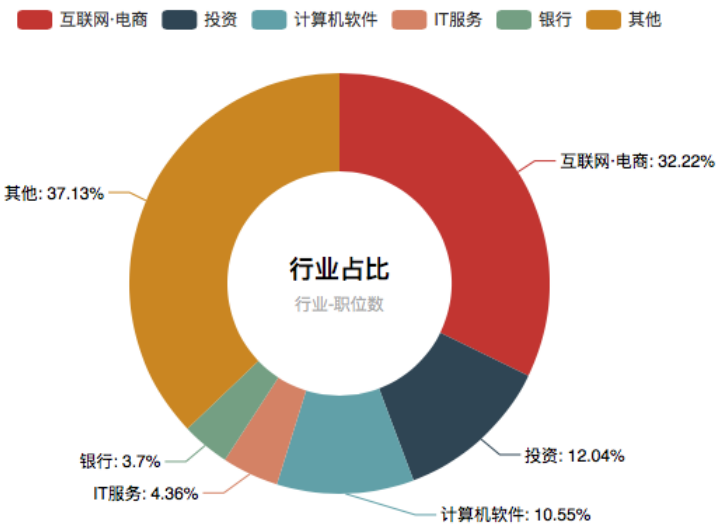
2、协助上级完成每季产品发展策略、产品结构需求、品类计划、产品销售预测、销售周期及降价期，与商品前期企划工作紧密对接，建立并完善商品流通管理流程及管理体系，确保大商品链的正常运作；

3、直营店的具体销售数据（产销率、库存率和库存款式、终端断款率和断款款式、爆款率和爆款款式畅销款式等）及实际情况，结合产品生命周期，负责各类数据的统计和分析，并分析各店铺历史销售数据，有针对性制定各店铺的上货计划；

4、负责各类销售数据的统计与分析，每月、每周提交销售数据分析报告及解决方案（促销、清货、调配、换货）；

5、商品报表分析（消化率、周转率、折扣率、毛利率分析报表）。

行业-职位数占比表明，职位的行业集中度高（互联网·电商一个行业占据了32.22%，TOP5 占比 62.87%，TOP10 占比 74.21%），求职时应重点考虑占比排名靠前的行业。

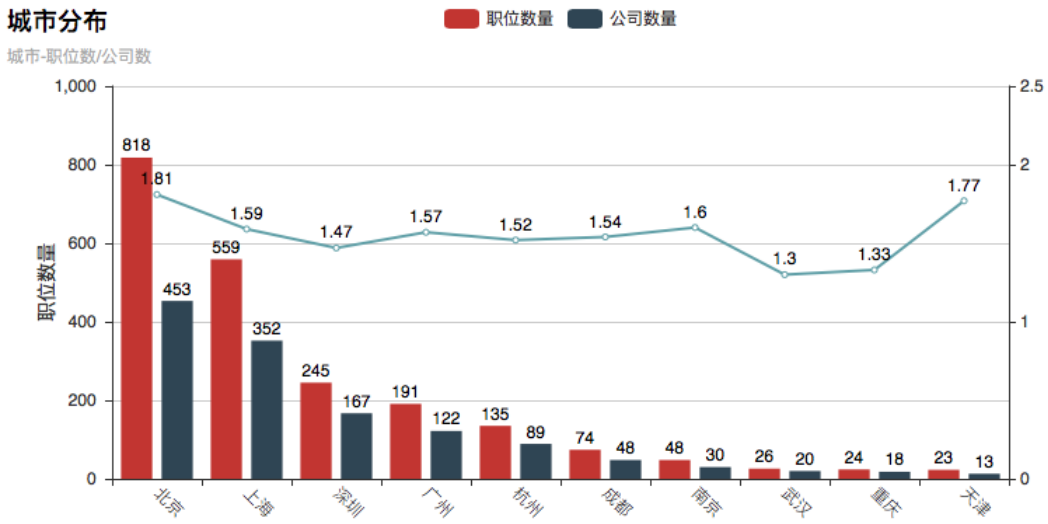


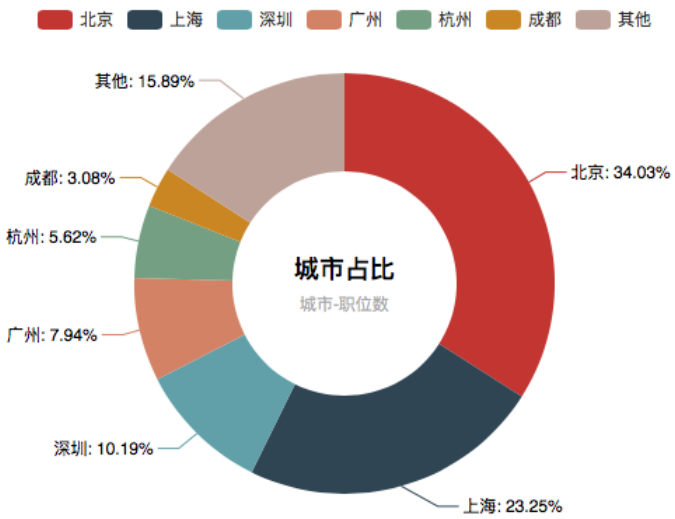
各行业需求最多的前三个公司分别为：

行业	职位需求 TOP 3
互联网·电商	京东、阿里巴巴、Baidu
投资	马上消费金融股份有限公司、深圳平安综合金融服务有限公司、宜信公司
计算机软件	寄云鼎城科技、中国电信云公司、文思海辉
IT 服务	思特奇信息技术、神州数码医疗科技股份有限公司、青岛海信网络科技股份有限公司
银行	广州银行、交银企服张江高科技园区分公司、江苏苏宁银行股份有限公司(分支机构)

职位集中于一线城市，北京、上海占据总量的近 60%

北上深广四个 1 线城市占据前四位，而北京、上海分别以 818 个职位，34.03%的占比和 559 个职位、23.25%的占比遥遥领先。



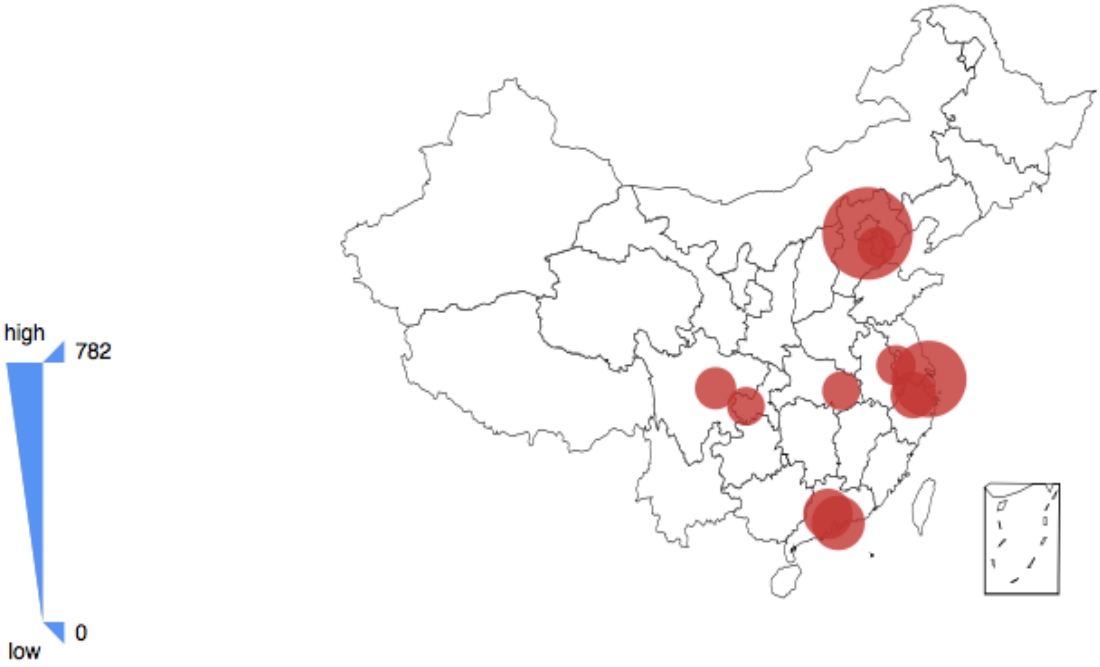


职位需求大的城市平均每家公司的需求量也较高，但天津市奇高。对天津市各公司的职位需求数量进行分组统计，发现捷信中国与捷信消费金融目前对数据分析职位有大量的需求，而天津有数据分析需求的公司数量较小，因此明显拉高了其职位/公司比。

company	count(DISTINCT jobtitle)
北京万方数据	1
天津修源皮具有限公司	1
天津光电集团有限公司	1
天津光电高斯通信工程技术股份有限公司	1
天津嘉业财富投资管理有限公司	1
天津链家宝业房地产经纪有限公司	1
天软时代	1
恒都集团	2
捷信中国	4
捷信消费金融有限公司	5
清华大学天津电子信息研究院	2
美腾科技	2
英业达集团(天津)电子技术有限公司	1

需求聚集在京津冀、长三角、珠三角与川渝四个地区，与当前互联网行业的分布基本一致

从地区层面来说，职位主要分布在京津冀、长三角、珠三角与川渝四个地区，这样的分布特征也与当前互联网行业的分布呈现明显相关性，与职位的行业分布情况吻合。



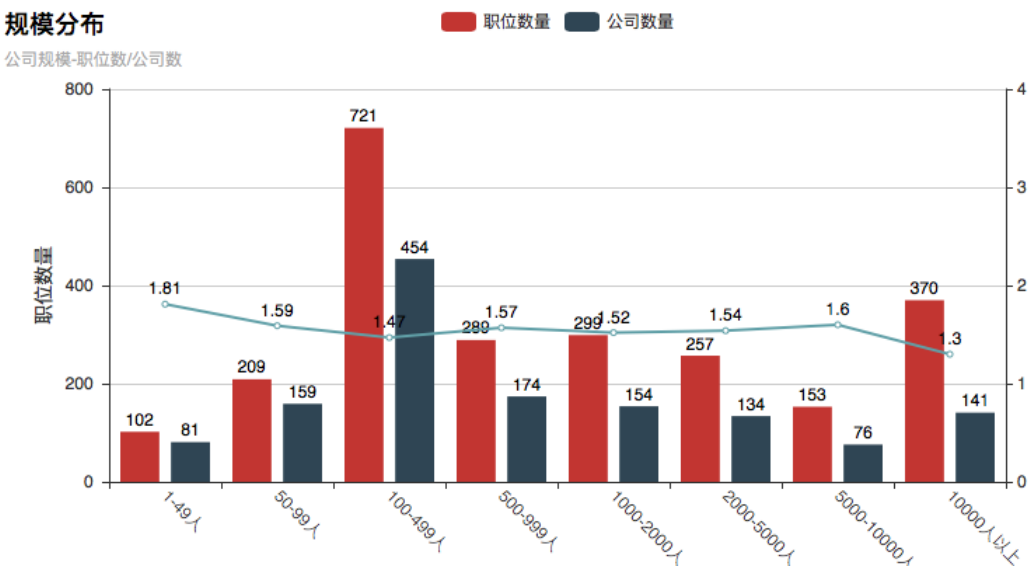
主要城市职位需求的 TOP 3 公司为，Baidu（15）与阿里巴巴（18）不出所料在自己的大本营拔得头筹，然而腾讯却不在榜中，各城市中腾讯的职位需求总和为 5。

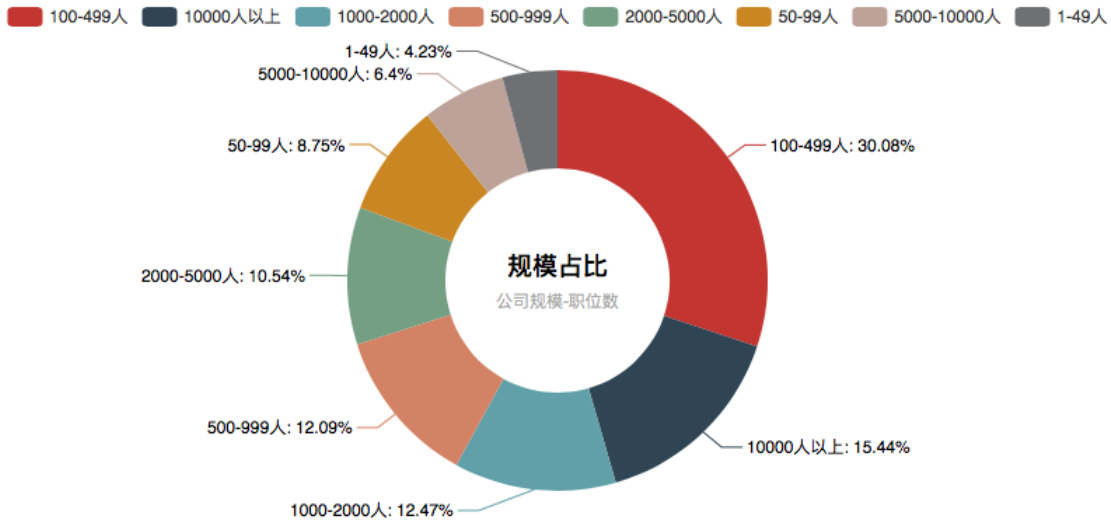
城市	TOP 3
北京	Baidu、京东、掌众科技
上海	深圳平安综合金融服务有限公司、游族、交银企服张江高科技园区分公司
深圳	深圳平安综合金融服务有限公司、深圳市城市交通规划设计研究中心有限公司、大数金融
广州	广州银行、UC 优视(UC 浏览器)、合生元集团
杭州	阿里巴巴、浙江华为、网新科技
成都	普惠快捷、正合联众、成都运力科技有限公司

中等规模公司需求量最大，1-49 人初创公司的平均招聘需求最旺盛

100-400 人中等规模的公司的需求数量最大。

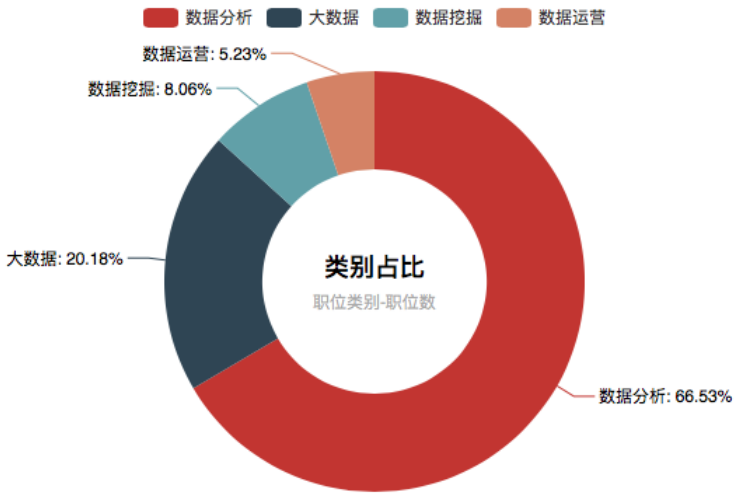
1-49 人规模的小公司平均职位需求最旺盛，具体观察发现，这一规模的公司主要以天使轮与 A 轮融资的初创公司组成，人员流动性大，导致招聘需求旺盛。相应的，10000 人以上的公司已发展到稳定期，因此员工流动性较小，职位数/公司数的值较低。



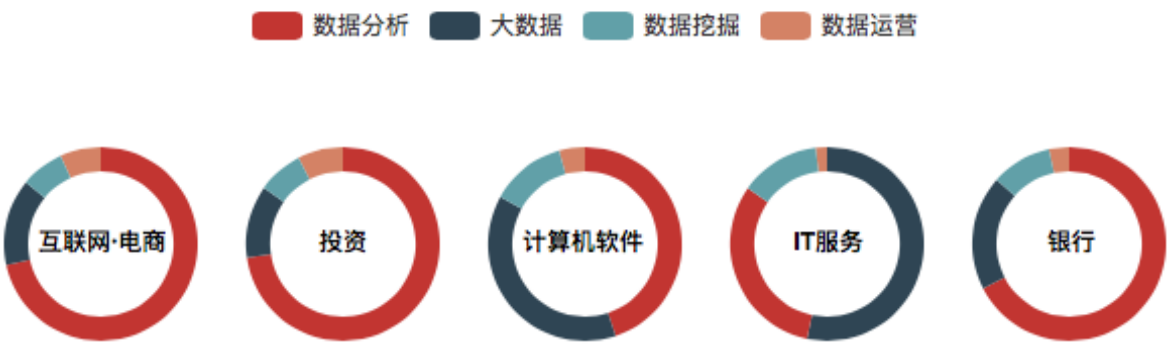


未细化的数据分析职位占 2/3，其后依次为大数据、数据挖掘与数据运营

大多数数据分析职位未进行细分，占比 66.53%。



需求量较大的行业中，IT 服务与计算机软件两行业大数据类型的职位占比较高，主要由许多 toB 的软件与服务提供商组成。



薪酬与福利

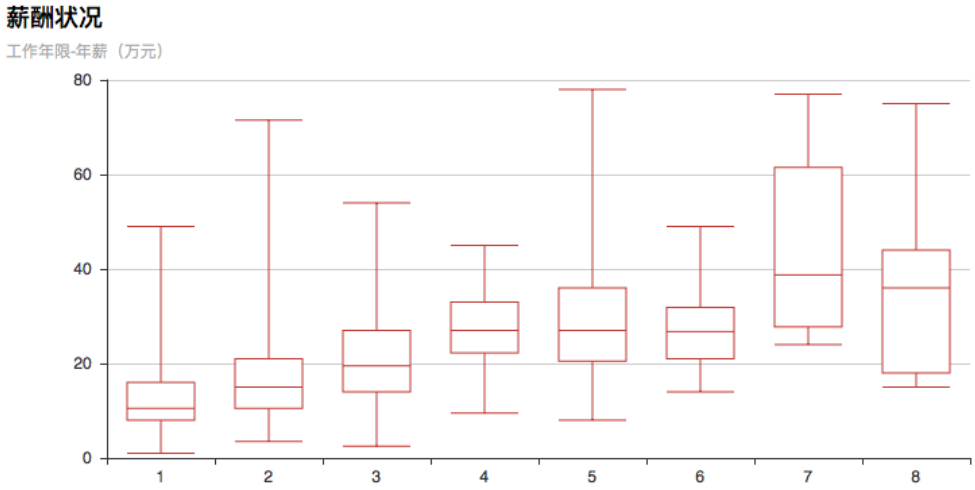
数据分析薪酬在前四年稳步增长，4-6 年出现瓶颈期，此后出现较大分化

企业职位中 8 年以上的职位数据较少，因此仅对前 8 年的薪酬进行分析，发现较为明显的分为三个阶段：

1-4 年：薪酬 30%-50%的速度增长

4-6 年：薪酬出现瓶颈，大部分处于 20-35 万之间

6 年后：较第 6 年上升一个台阶，但此阶段分化程度很大，第 7 年 Q3/Q1 = 2.22，Q3-Q1 = 33.75



福利水平较高，社保、假期与发展空间相关词出现最频繁

对职位标签进行分词统计，获得词云图，并对 TOP 10 标签进行了统计：

五险一金，带薪年假，发展空间大，绩效奖金，岗位晋升，定期体检，领导好，节日礼物，午餐补助，技能培训

技能需求

对四类职位（数据分析、数据运营、数据挖掘、大数据）的岗位描述进行聚合并分词，获得词云。

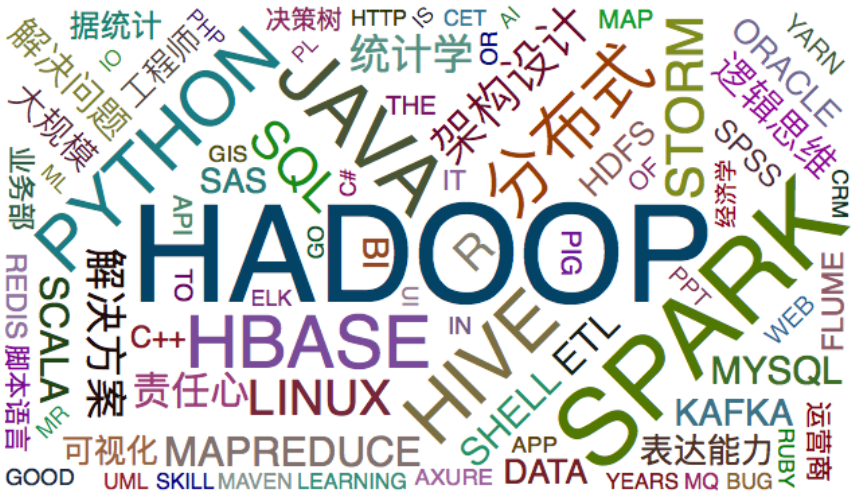
由词云可以看出，职位的技能需求从数据分析和运营偏重业务与技术结合的层面，到数据挖掘的编程、机器学习，再到大数据的各类构架这一较倚重技术的方向。必备技能为 Excel 和 SQL，其他方面在做求职准备时，可以根据自身情况，选择职业类型，并根据所需技能进行重点的准备。

数据分析：SQL、EXCEL、统计学

数据运营：EXCEL、分析报告、SQL

数据挖掘：PYTHON、统计学、R

大数据：HADOOP、SPARK、JAVA



结论

1. 在进行数据分析相关岗位求职时，应优先选择互联网或金融相关领域，并根据自身专业与经历选择适当的细分行业；
2. 北上深广尤其是北京、上海职位多，机会多，求职时应作为重点考虑城市；
3. 工作 3-4 年薪酬一般能够达到 25-30 万，但此时需要加强学习，挣脱瓶颈，争取在第 6 年实现收入的飞跃；
4. Excel 和 SQL 是各类数据分析职位的基本功和必备技能，进阶的诸如 Python、R、Hadoop 等可根据自身定位进行学习。