第四届“泰迪杯”
全国数据挖掘挑战赛

优秀作品

作品名称：网络招聘信息的数据挖掘与综合分析

荣获奖项：特等并获企业冠名奖

作品单位：北京林业大学

作品成员：孙海锋 郑中枢 杨武岳

指导老师：崔晓晖

网络招聘信息的分析与挖掘

摘要

近年来，随着互联网的广泛应用和网络招聘的迅速发展，网络招聘信息平台已成为招聘者获取信息的主要渠道。因此，运用网络文本分析和数据挖掘技术对网络招聘信息的研究具有重大的意义。

对于问题 1，通过 PositionId 对招聘信息表、职位描述表进行去重，得到不重复的招聘职位信息。利用 jieba 中文分词工具对岗位描述信息进行分词，并通过 TF-IDF 算法提取每个职位描述的前 5 个关键词。再利用 TF-IDF 算法得到每个职位描述的 TF-IDF 权重向量，采用 K-means 对 TF-IDF 权重向量进行聚类，得到 7 个质心。分别求出距离各个质心最近的 5 个职位，结合招聘信息表的 PositionFirstType 字段，根据 KNN 算法，为各个类加上行业性质标签。再分别对各个职业类型的 PositionName 进行统计分析，得出各个职业类型对应的专业领域。

对于问题 2，通过利用 excel 对去重后的招聘信息表对行业领域、工作地域、职位分类三个项目进行分类筛选，对各个项目的各类内容进行计数汇总统计，根据计数多的内容去定于热门的行业、地域、职位。

对于问题 3，根据数据挖掘与分析的职位特征，将新兴的职位定义为两大类并分别筛选出来。利用发散性思维，再分别对筛选出来的结果按照城市（city）、公司阶段（financestage）、学历要求（Education）、薪资（Salary）四个方面对其进行多方面系统地统计，结合图表进行分析预测相关职位的需求。

对于问题 4，通过寻找 it 职位对应的 id 的职业描述，并对其分词和 it 专业语义库构建，在此基础上筛选出所有的 it 职位。对附件 1 进行数据预处理，在预处理得到的数据上进行数据初步筛选出 it 行业的职位。对筛选出的 it 职位对应的职业 id 找到职位描述表的职位描述，对该描述构建 it 专业语义库。判断职业描述表中职位是否符合 it 职业，通过判断与专业语义库的交集长度来确定是否为 it 职业并统计地域。

对于问题 5，根据研究结果，通过分析目前的主要职业类型、职业要求、热门行业及地域、工作经验及就业现状等问题，给在校大学生的就业规划提出可行性的建议。

关键词：去重　中文分词　K-means 聚类　KNN 算法　TF-IDF 算法　预测相关职位

Network Recruitment Information Analysis and Mining

Abstract

In recent years, with the wide application of Internet and the rapid development of Internet recruitment, recruitment information network platform has become the main channel for interviewers to obtain information.Therefore, using the network text analysis and data mining technology to network recruitment information of the research is of great significance.

Aiming at the problem of the first, the recruitment information table, by PositionId job description table to heavy, don't repeat job information.Using jieba Chinese word segmentation tools to participle of job description information, and through the TF - IDF algorithm to extract each job description of the top five keywords.Reusing the TF - IDF algorithm for each job description of the TF - IDF weight vector, the K - means of TF - IDF weight vector clustering, get seven centers of mass.Respectively calculated from the center of mass of recent 5 position, combination of recruitment information table PositionFirstType fields, based on KNN algorithm, for each class with nature of the industry.Then respectively the statistical analysis of various professional types of PositionName, drawing the corresponding professional career type.

Aiming at the problem of the second, by using excel to go after heavy recruitment information table (IndustryField) to industry field, work area (City), the position classification (PositionFirstType) classify three projects selection, all kinds of content to calculate summary statistics for each project, according to calculating more than content to industry and region, due to be popular, position.

Aiming at the problem of the third, according to the characters of the position of data mining and analysis, definsing the position of emerging as two categories and filtered, respectively.Using divergent thinking, and then would get results of screening out respectively according to the City (City), phase (Financestage), Education (Education), compensation (Salary) from four aspects on the various statistics systematically, with the demand of chart analysis forecast related position.

Aiming at the problem of the forth, by looking for the it position corresponding to the id of the job description, and the word segmentation and built it professional semantic library on the basis of screening all it position.To annex 1 for data preprocessing in data preprocessing the data on a preliminary screening the position of the it industry**.**To screen out the it positions the corresponding professional id to find the job description table in the job description, description on the build it professional semantic repository.Determining the job description in the table position is in line with the it profession, through the judgment and professional semantic repository to determine whether the intersection of length for the it professional and statistical area.

Aiming at the problem of the fifth, according to the research results, through the analysis of the current main professional type, the professional requirements, popular industry and region, work experience, and the problem of employment

situation for college students employment planning and feasibility Suggestions are put forward.

Keywords: to heavy Chinese participle K-means clustering TF - IDF weighted KNN algorithm Predict related position

目录

1、挖掘目标.....6
2、分析方法与过程.....6
2.1 问题 1 分析方法与过程.....7
2.1.1 流程图.....7
2.1.2 数据预处理.....7
2.1.3 职业类型的分类.....9
2.1.4 Knn 最邻近分类算法[2].....11
2.2 问题 2 分析方法与过程.....12
2.2.1 数据筛选.....12
2.2.2 数据统计.....12
2.3 问题 3 分析方法与过程.....12
2.3.1 问题 2 流程图.....12
2.3.2 数据预处理.....13
2.4 问题 4 分析方法与过程.....13
2.4.1 数据预处理.....13
2.4.2 数据对照筛选分析.....14
2.5 问题 5 分析方法与过程.....14
3. 结果分析.....14
3.1 问题 1 结果分析.....14
3.1.1 聚类中心分类结果.....14
3.1.2 职业领域分类.....15
3.2 问题 2 结果分析.....15
3.2.1 对热门行业的分析.....15
3.2.2 对热门领域的分析.....16
3.2.3 对热门职位的分析.....17
3.3 问题 3 结果分析.....17
3.3.1 按城市地域进行划分.....17
3.3.2 按公司发展阶段进行划分.....18
3.3.3 按学历进行筛选.....19
3.3.4 按 salary（月薪）进行统计.....20

3.4 问题 4 结果分析.....20
3.5 结合研究结果，给在校大学生就业规划提建议.....22
4 结论.....24
5 参考文献.....24

1、挖掘目标

本次建模目标是利用网络信息平台系统发布的网络招聘信息数据，利用 jieba 中文分词工具对职位描述进行分词、K-means 聚类的方法及 KNN 算法，达到以下三个目标：

1） 利用文本分词和文本聚类的方法对非结构化的数据进行文本挖掘，根据聚类结果，结合招聘职位工作性质和内涵分析现阶段所需的职业类型、专业领域；结合招聘单位的特点分析目前热门行业走向。

2） 根据新兴数据挖掘行业的职位体系的数据，预测未来的人才走向及相关的职位要求。分析 IT 行业人才市场的供求现状，了解其未来的发展趋势。

3） 根据研究的目前人才情况、热门行业、未来人才需求走向等结果，给大学生的就业规划提供真实可靠的建议。

2、分析方法与过程

总体流程图

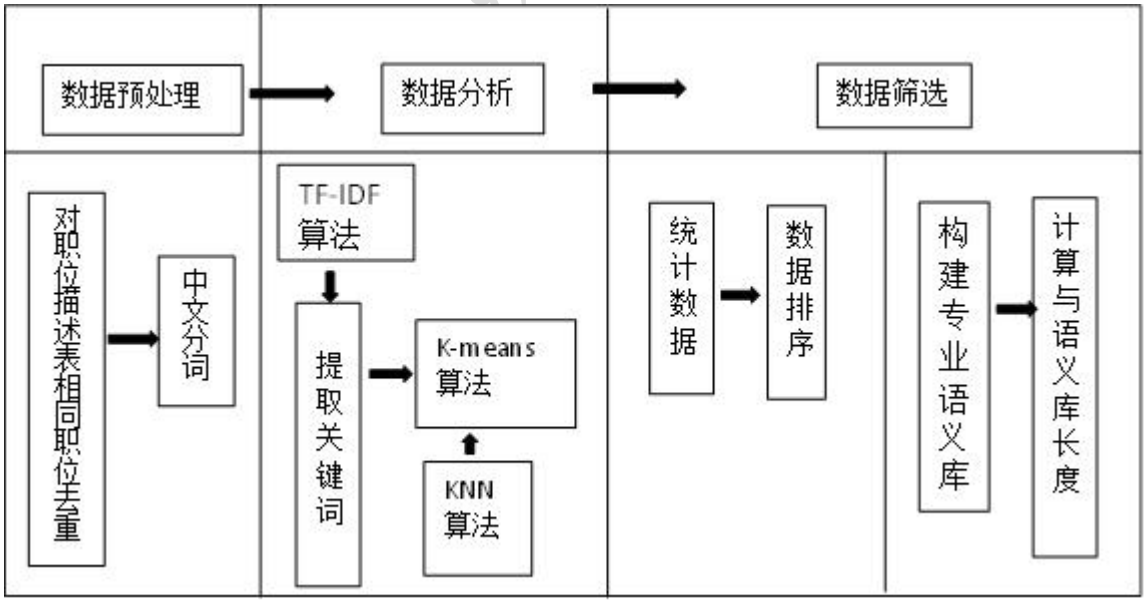


图 1：总体流程图

本用例主要包括如下步骤：

步骤一：数据预处理，在题目给出的数据中，出现了很多重复的招聘数据，在原始的数据上进行去重处理，在此基础上进行中文分词。

步骤二：数据分析，在对职位描述信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，找出每个职位描述的关键词，把职位描

述信息转换为权重向量。采用 K-means 算法对职业进行分类，利用 Knn 算法找出与各中心相似的元素，根据个数多的判定所属类别。
步骤三：数据筛选，统计相关数据，分类筛选汇总，预测热门行业的问题、人才需求走向和相关职业的需求情况等。
步骤四：利用步骤一的结果构建专业语义库，通过计算与语义库的距离，找出对应的 IT 职业的 ID，统计地域分布情况。

2.1 问题 1 分析方法与过程

2.1.1 流程图

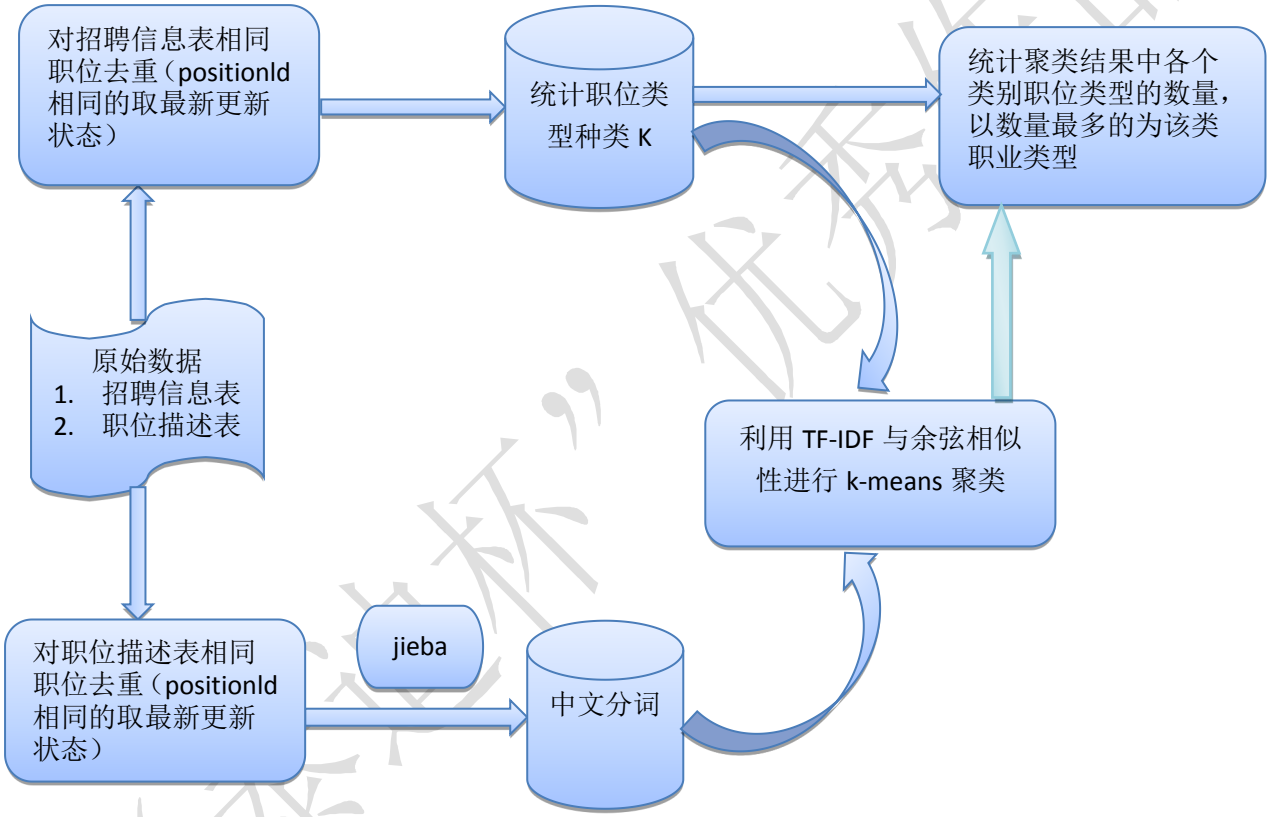


图 2：问题 1 流程图

2.1.2 数据预处理

2.1.2.1 招聘信息的去重、去空

在题目给出的数据中，出现了很多重复的招聘数据。例如招聘信息表跟职位描述表中出现了很多重复的职位信息。考虑到公司招聘人才时可能每天都会对要招聘的职位进行更新，因此在去重的时候应该取更新时间最晚的记录，去掉历史

记录。考虑到 python 中的字典在保存数据时，key 相同的内容，value 取值为最后更新的值。因此在读取数据时，按时间升序把招聘信息的 PositionId 作为 key，把整个招聘信息作为 value 保存在 value 中。最后再将字典中的内容写入文本即可。同时在职位描述表中出现了职位描述为空的记录，干扰了问题的分析，采取直接滤过方法，从文本中删除。对招聘数据去重的 python 程序见附件 duplicatedetection.py。去重、去空后的数据分别保存在附件 1 去重.csv、附件 2 去重.csv、附件 3 去重.csv 中。

2.1.2.2 对职位信息表进行中文分词

在对招聘信息进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件职位描述表中，以中文文本的方式给出了数据。为了便于转换，先要对这些职位描述信息进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)，同时采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果。

在分词的同时，采用了 TF-IDF 算法，抽取每个职位描述中的前 5 个关键词，这里采用 jieba 自带的语义库。

2.1.2.3 TF-IDF 算法

在对职位描述信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，把职位描述信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）。

$$\text{词频（TF）} = \text{某个词在文本中出现的次数} \tag{1}$$

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$\text{词频（TF）} = \frac{\text{某个词在文本中的出现次数}}{\text{文本的总词数}} \tag{2}$$

或

$$\text{词频（TF）} = \frac{\text{某个词在文本中的出现次数}}{\text{该文本出现次数最多的词的出现次数}} \tag{3}$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$\text{逆文档频率（IDF）} = \log\left(\frac{\text{语料库的文本总数}}{\text{包含该词的文本数} + 1}\right) \tag{4}$$

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

TF - IDF = 词频（TF）× 逆文档频率（IDF）　　(5)

实际分析得出 TF-IDF 值与一个词在职位描述表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的职位描述表中文本的关键词。

2.1.2.4 生成 TF-IDF 向量

生成 TF-IDF 向量的具体步骤如下：
（1）使用 TF-IDF 算法，找出每个职位描述的前 5 个关键词；
（2）对每个岗位描述提取的 5 个关键词，合并成一个集合，计算每个岗位描述对于这个集合中词的词频，如果没有则记为 0；
（3）生成各个岗位描述的 TF-IDF 权重向量，计算公式如下：

TF-IDF=词频（TF）×逆文档频率（IDF）　　（6）

2.1.3 职业类型的分类

生成职位描述的 TF-IDF 权重向量后，根据每个职位的 TF-IDF 权重向量，对职业进行分类。这里采用 K-means 算法把职业类型分成 7 类。

K-mean 聚类的原理如下：

假设有一个包含 n 个 d 维数据点的数据集 $X = \{x_1, x_2, \dots, x_i, \dots, x_n\}$，其中 $x_i \in R^d$，K-means 聚类将数据集 X 组织为 K 个划分 $C = \{c_k, i = 1,2, \dots, K\}$。每个划分代表一个类 c_k，每个类 c_k 有一个类别中心 μ_i。选取欧式距离作为相似性和距离判断准则，计算该类内个点到聚类中心 μ_i 的距离平方和

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \qquad (1)$$

聚类目标是使各类总的距离平方和 $J(C) = \sum_{k=1}^{K} J(c_k)$ 最小，

$$J(C) = \sum_{k=1}^{K} J(c_k) = \sum_{k=1}^{K} \sum_{x_i \in C_i} \|x_i - \mu_i\|^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} d_{ki} \|x_i - \mu_i\|^2 \qquad (2)$$

其中，$d_{ki} = \begin{cases} 1, \text{若} x_i \in c_i \\ 0, \text{若} x_i \notin c_i \end{cases}$，所以根据最小二乘法和拉格朗日原理，聚类中心 μ_k

应该取为类别c_k类各数据点的平均值。

K-mean 聚类的算法步骤如下：

1、从X中随机取K个元素，作为K个簇的各自的中心。

2、分别计算剩下的元素到K个簇中心的相异度，将这些元素分别划归到相异度最低的簇。

3、根据聚类结果，重新计算K个簇各自的中心，计算方法是取簇中所有元素各自维度的算术平均数。

4、将X中全部元素按照新的中心重新聚类。

5、重复第 4 步，直到聚类结果不再变化。

6、将结果输出。

K-mean 聚类的算法流程图如下：

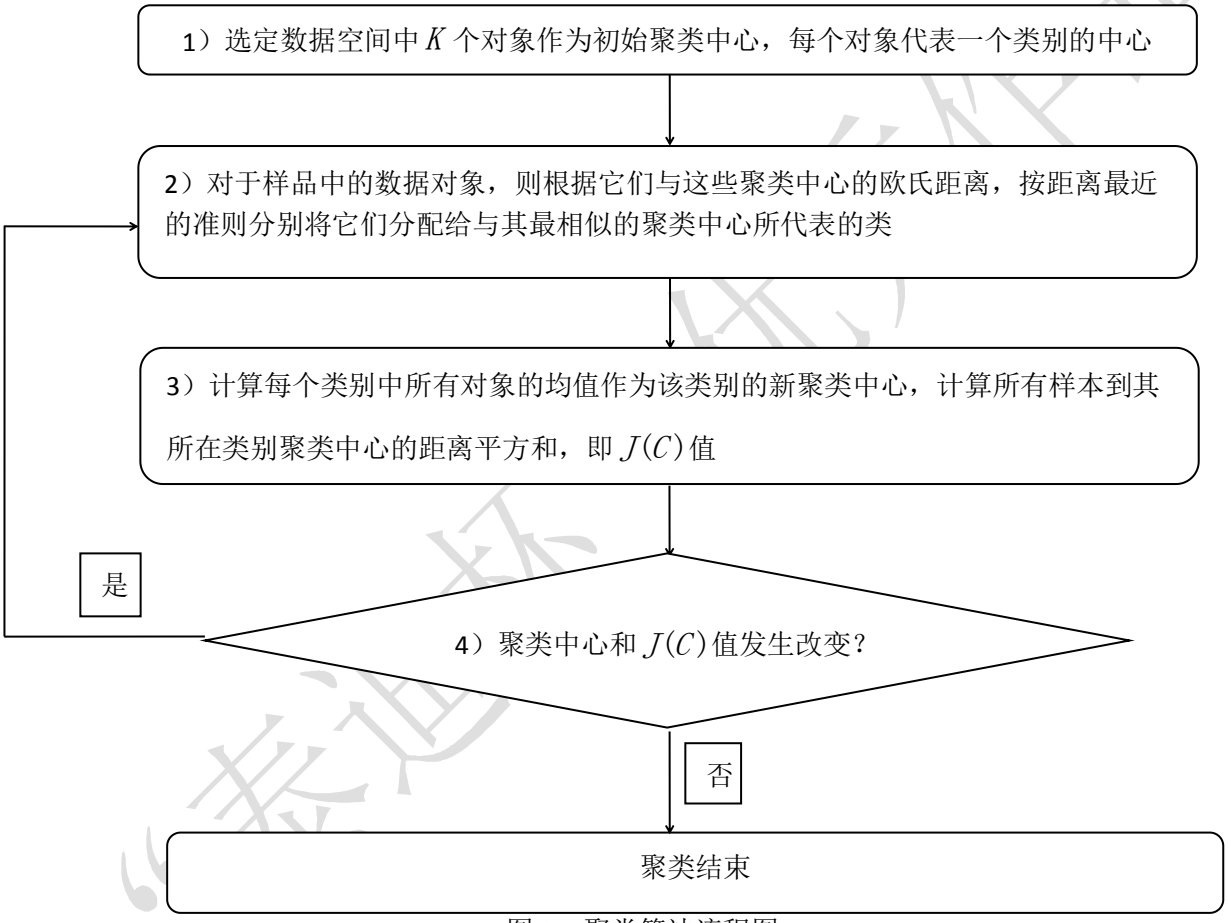


图 3：聚类算法流程图

由于职位描述表给出了 539216 条记录，去重后还有 402727 条记录，如果把所有的职位都用来挖掘分析，会占用很大的机器性能跟时间。为了节省机器性能跟时间，获得结果。从 402727 条记录数据中随机抽取 40000 条记录，抽样 python 程序见附件 sampling.py，抽样结果保存在抽样样本.csv 文件里面。然后利用抽样样本进行分词、求 TF-IDF 向量，并利用 K-mean 聚类，把样本分成 7 类，程序见附件 kmean.py，得出来的七个聚类中心保存在 centroids.csv 中（由于维度较高，打开时请用文本编辑器打开），每步迭代的$J(C)$值保存在附件 clusterAssment.csv 中。

2.1.4 Knn 最邻近分类算法[2]

由 K-Means 分类得到聚类中心，利用 Knn 算法找出与各中心相似的元素，根据个数多的判定所属类别。根据向量空间模型，将每一类别文本训练后得到该类别的中心向量记为$C_j(W_1,W_2,\cdots,W_n)$，将待分类文本T表示成n维向量的形式$T(W_1,W_2,\cdots,W_n)$，则文本内容被形式化为特征空间中的加权特征向量，即$D = D(T_1,W_1;T_2,W_2;\cdots T_n,W_n)$。对于一个测试文本，计算它与训练样本集中每个文本的相似度，找出K个最相似的文本，根据加权距离和判断测试文本所属的类别。具体算法步骤如下：

（1） 对于一个测试文本，根据特征词形成测试文本向量。

（2） 计算该测试文本与训练集中每个文本的文本相似度，计算公式为：

$$Sim(d_i,d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{\sum_{k=1}^{M} W_{ik}^{\ 2}}\sqrt{\sum_{k=1}^{M} W_{jk}^{\ 2}}}$$

式中，d_i为测试文本的特征向量，d_j为j类的中心向量；M为特征向量维数；W_k为向量的第k维。k值的确定一般先采用一个初始值，然后根据实验测试K的结果来调整K值。

（3） 按照文本相似度，在训练文本集中选出与测试文本最相似的k个文本。

（4） 在测试文本的k个近邻中，以此计算每类的权重，计算公式如下：

$$P(X,C_j) = \begin{cases} 1, \text{若} \sum_{d \notin Knn} Sim(x,d_i)y(d_i,C_j) - b \geq 0 \\ 0, \text{其他} \end{cases}$$

式中，x为测试文本的特征向量；$Sim(x,d_i)$为相似度计算公式；b为阈值，有待于优化选择；而$y(d_i,C_j)$的值为 1 或 0，如果d_i属于C_j，则函数值为 1，否则为 0。

（5） 比较类的权重，将文本分到权重最大的那个类别中。

2.1.5 分析职业类型和初步定义职业领域

对附件 **3** 根据 K-means 聚类方法和 Knn 最邻近分类得出 7 个点和每个点周围 100 个 id，根据这些 id 对照附件 1 所属的职业 PositionFirstType，包括技术、职能、市场与销售、产品、运营、设计和金融七大职业分类，统计数量最多的即为目前企业最需要的职业类型，并定义相关职业领域。

2.2 问题 2 分析方法与过程

2.2.1 数据筛选

（1）根据招聘信息表对不同行业领域进行分类筛选，得到 020 行业、电子商务、分类信息、广告营销等 21 个不同行业领域。
（2）根据招聘信息表对不同工作地域进行分类，得到 299 个不同的地域。
（3）根据招聘信息表对不同的职位所属大类分类，分为技术、运营、市场与销售、设计、职能、产品、金融七大类。

2.2.2 数据统计

（1）对各个行业领域出现的招聘次数进行计数，通过排序得出排名前 10 行业，并定义热门行业；
（2）对各个地域进行分类计数，通过排序得出排名前 10 的地域，并定义热门的地域；
（3）对七大职位所属大类进行分类计数，通过排序得出各大职位的需求情况，并定义热门的职位。

2.3 问题 3 分析方法与过程

2.3.1 问题 2 流程图

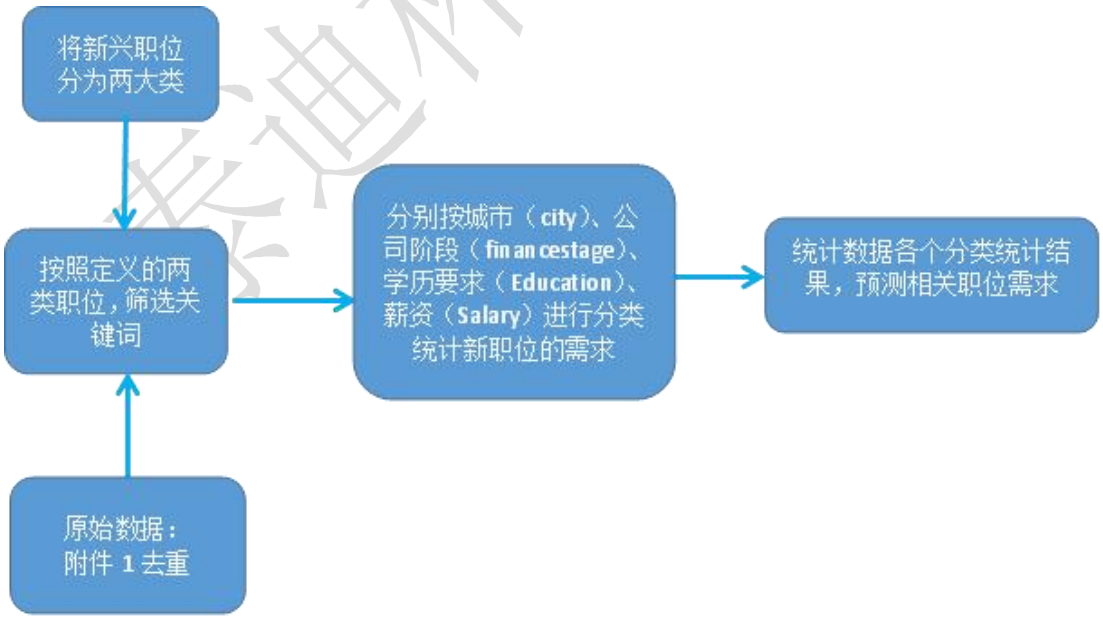


图 4：问题 3 流程

2.3.2 数据预处理

2.3.2.1 定义新兴职位

定义新兴职位，并将其分为两类：
第一类：数据分析
第二类：大数据、数据挖掘、机器学习、人工智能

2.3.2.2 数据筛选

根据定义的两大类新兴职位，对招聘信息表中 positionName（职位名称）这一列进行按关键词定义为大数据、数据挖掘、机器学习、人工智能与数据分析这两类，进行筛选分析；（筛选结果见附件 3）

2.3.2.3 统计做图表分析

利用数据透视表，对两类职业所在城市（city）、公司阶段（financestage）、学历要求（Education）、薪资（Salary）进行计数，并且进行降序处理（具体结果见附件 3）

2.4 问题 4 分析方法与过程

2.4.1 数据预处理

2.4.1.1、数据初步筛选

粗略筛选出 it 行业的职位。通过招聘信息的 PositionType 字段筛选出 dba、it 支持、java、sqlserver、web 前端、测试工程师、后端开发、架构师、嵌入式、网络安全、网络工程师、网页产品设计师、移动开发、运维工程师这些选项的职位,记为集合C。把C保存在 it.csv 文件中。

2.4.1.2、it 职位分词与 it 专业语义库构建

利用 2.4.1 选出来职位的 id，得到每个职位在附件 3 中对应的职位描述，分别对这些职位描述进行分词、去停用词、去重，得到一个语义库Y

2.4.1.3、筛选出所有的 it 职位

假设 z 是职位描述表中的一条记录，对 z 的职位描述内容进行分词，得到集合S，如果集合S跟 it 专业语义库Y的交集长度大于等于 2（改大一点进行比较严格的筛选）、则可以认为 z 属于 it 职业，记录 z 的 PositionId。

遍历所有的职位描述，可以筛选出所有的 it 职位跟对应的职位描述，程序见附件 itposition.py ，结果分别保存在 it 职位.csv 跟 it 专业职位描述.csv 中。

2.4.2 数据对照筛选分析

对于选出的所有 IT 职位，根据招聘信息表的 PositionName 对每一个 IT 职位进行筛选，得出各个 IT 职位的信息。根据招聘信息表中的 City、IndustryField 和 Education 分析所有 IT 职位的地域分布情况、人才的专业和学历层次，。再根据 CompanySize、提供的 Salary 和除 IT 的其他职位在这段时间的需求量。公司规模大，说明所需要的人才越高；工资越高说明职位需求量高，受欢迎程度越高；在同一段时间相比于其他行业需求量大，说明有发展前景；以此分析了 IT 人才市场的供求现状及未来的发展趋势。

2.5 问题 5 分析方法与过程

根据本题研究结果，对目前的就业市场行情和职位供需情况进行简要的概况，再此基础上从学生自身特点、专业特点、热门行业与地域、工作经验、能力这 5 方面给大学生的就业提供建议，最后鼓励大学生做好就业规划，未雨绸缪。

3. 结果分析

3.1 问题 1 结果分析

3.1.1 聚类中心分类结果

通过去重后对文本进行分词，提取五个关键词后由 K-Means 分类得到聚类中心，利用 KNN 算法找出离各个聚类中心最近的前 5 个元素，根据“少数服从多数”判定聚类中心所属类别。KNN 算法的大致步骤如下：

1、算距离：给定聚类中心，计算它与样本中的每个 TF-IDF 权重向量的距离

2、找邻居：圈定距离最近的 15 个样本，作为聚类中心的近邻

3、做分类：根据这 5 个近邻归属的主要类别，来对聚类中心进行分类

结合抽样样本，分别找出七个聚类中心的 5 个近邻样本点所属的职业类型（程序见附件 knn.py）。结果如下表所示为：

表 1：KNN 分类表

聚类中心	市场与销售	技术	运营	职能	设计	所属类型
第一个聚类中心	3	2	0	0	0	市场与销售
第二个聚类中心	0	1	3	0	0	运营
第三个聚类中心	0	1	0	4	0	职能
第四个聚类中心	1	4	0	0	0	技术
第五个聚类中心	2	3	0	0	0	技术
第六个聚类中心	2	3	0	0	0	技术
第七个聚类中心	0	1	1	0	3	设计

从 KNN 分类表可以看出：七个聚类中心可分为：市场与销售、运营、职能、技术、设计四大类。

3.1.2 职业领域分类

从职位信息表中筛选出属于技术类的职业，由职位类型可以得到：

1、市场与销售领域集中在采购、高端市场职位、公关、供应链、市场/营销、投资、销售。

2、运营领域集中在编辑、高端运营职位、客服、网点运营、设计。

3、职能领域集中在财务、法务、高端职能职位、行政、人力资源。

4、技术类的专业领域集中在 dba、测试、高端技术职位、后端开发、企业软件、前端开发、项目管理、移动开发、硬件开发、运维。

3.2 问题 2 结果分析

3.2.1 对热门行业的分析

通过对 21 个行业领域排序计数，选取排名前 10 的行业领域进行分析，（见表 1）分析得出移动互联网这一行业在前 10 个行业所占比例为 60%，电子商务行业占了 13%，金融行业占了 9%。（见图 1）可以发现移动互联网这一行业正在蓬勃发展，在最近几年里，移动通信和互联网成为当今世界发展最快、市场潜力最大、前景最诱人的两大业务，它们的增长速度都是任何预测家未曾预料到的，所以移动互联网可以预见将会创造怎样的经济神话。根据发展的需要，招聘单位对该类人才的需求量也逐渐增加，使移动互联网迅速成为热门行业之一。

表 2：排名前 10 的行业领域

行业领域	网络信息招聘次数	排序
移动与互联网	2066778	1
电子商务	44513	2
金融	31877	3
企业服务	15977	4
020 行业	13188	5
数据服务	9380	6
教育	8615	7
游戏	7682	8
文化娱乐	4767	9
其他	4707	10

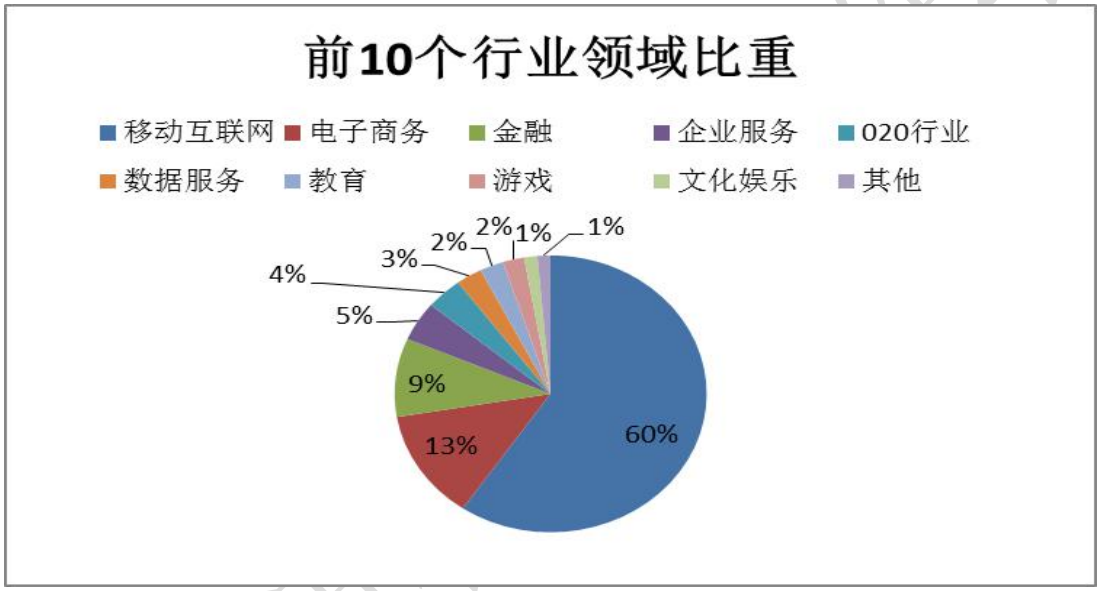


图 5：前 10 个行业领域比重

3.2.2 对热门领域的分析

随着经济发展，城市发展日新月异，一线城市在生产、服务、金融、创新、就业、流通等全国社会活动中起到引领的主导功能。研究表明，各大招聘单位提供的工作地域为北京、上海、深圳、广州、杭州等这些经济发达的城市，（图 3 和图四）大城市对人才的需求量高，而大城市新兴行业的兴起，为求职者提供更多的就业岗位。

城市	招聘个数
总计	402627
北京	152245
上海	66610
深圳	49336
广州	34746
杭州	30270
成都	11920
武汉	7434
南京	6327

表 3：排名前 10 的地域数据

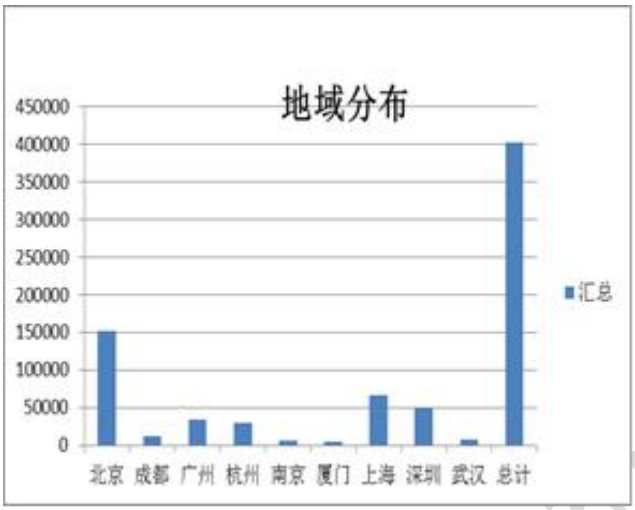


图 6：地域分布排名前 10 的地域分布

3.2.3 对热门职位的分析

通过对职位所属大类的 7 个类别进行分类技术，得到技术类职位是目前需求量的职位，技术类职位包括 MySQL 数据库工程师（DBA）、高端技术职位、PHP 开发等技术型的职位，其发展前景可观，具有一定的发展优势。仅次于技术型职位是市场与销售职位，促进了经济的发展，体现了供求关系，发展前景也是较为可观的。

职位大类	计数
技术	163692
市场与销售	83117
运营	67340
设计	28792
职能	26789
产品	25130
金融	7018

表 4：职位大类计数图

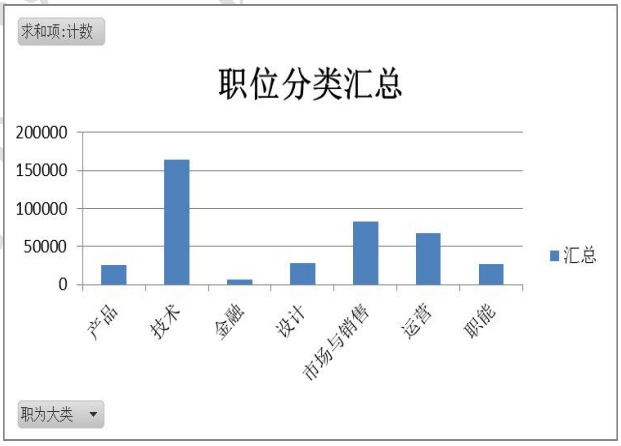


图 7：职位大类分类汇总图

3.3 问题 3 结果分析

3.3.1 按城市地域进行划分

这里，截取前 10 个城市进行分析，得到：

表 5 不同城市对这两类新兴职位的需求情况表

第一类		第二类	
城市	计数求和	城市	计数求和
北京	1248	北京	1718
上海	463	上海	502
深圳	243	深圳	351
杭州	170	杭州	240
广州	159	广州	168
成都	36	成都	85
南京	31	南京	60
武汉	29	武汉	51
长沙	18	厦门	24
苏州	17	苏州	19

如表 1 所示：

作为中国首都的北京对这两类职位的需求量位居榜首，通过计算可得，北京对第一类职位的需求为 49%，对第二类——数据分析领域的职位需求为 51%，超出中国所有城市的一半！而享有“东方巴黎”的上海紧随其后，需求数量也是达到 18%和 15%。并且，诸如深圳、杭州、广州等一线城市对其需求量也都是过百的。可以看到：经济越发达，科技越进步的一线城市对这两类新兴的数据分析与挖掘的职位需求量越大。

进入 21 世纪信息爆炸的时代，各种各样的信息满天飞，有用资讯和没用资讯混为一潭，很多数据信息需要我们去整理并且发现其中的规律，社会迫切需要这种人才将各种各样的数据转换成有用的信息和知识，数据分析与挖掘这一行业便孕育而生了。

作为一线城市北京、上海、深圳经济发达，有比较早的 IT 基础，拥有高超的 IT 技术，在 IT 行业一直作为领头羊领跑全中国。而杭州近几年软件发展也很迅速，这也从侧面推动了 IT 行业的蓬勃发展。而数据分析与挖掘作为新兴的 IT 行业，与传统的 IT 行业对比，拥有更加鲜活的时代性与科技含量，势必在这个信息爆炸的时代掀起一番热潮。而也只有这些一线城市有这样的本事，利用自身优越的科技与人才资源，庞大的数据系统，推动着数据分析与挖掘行业的蓬勃发展。所以可以预测：未来对数据挖掘与分析的职位需求应该集中在这些一线城市。

3.3.2 按公司发展阶段进行划分

表 6： 不同公司发展阶段对这两类职位的需求

第一类		第二类	
公司阶段	汇总	公司阶段	汇总
上市公司	483	上市公司	792
成长型(A 轮)	360	成长型(B 轮)	399
成长型(B 轮)	296	初创型(未融资)	322
成熟型(D 轮及以上)	219	成熟型(C 轮)	304
成熟型(C 轮)	206	成熟型(D 轮及以上)	278
成长型(不需要融资)	189	初创型(天使轮)	267
初创型(天使轮)	180	成长型(不需要融资)	251
成熟型(不需要融资)	166	成熟型(不需要融资)	194
初创型(不需要融资)	42	初创型(不需要融资)	64

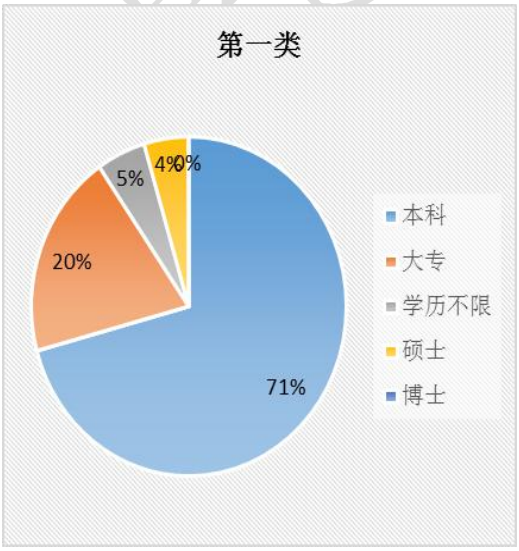
如表 2 所示：

作为龙头老大的上市公司，对这两类新兴职位的需求量上都是稳居第一。在筛选过程中发现，这些上市公司中，诸如阿里巴巴、JD 京东商城、当当网等这些电子商务平台，凭借自身海量的数据，玩转数据分析与挖掘市场，推动着数据分析与挖掘蓬勃发展，需要大量的数据分析人才，为这两类行业提供了大量的职位。除此之外，还有很多诸如知网、凤凰网、腾讯这些知名的上市公司也对这两大类职位需求比重也是很高的。可以看出：信息时代，数据分析与挖掘行业已经在大型企业的公司发挥举足轻重的作用，成为衡量一个公司规模的指标。

按融资和为融资分，不难发现融资型公司需要的人才量更多。广义上的融资也叫金融，就是货币资金的融通，当事人通过各种方式到金融市场上筹措或贷。所以，数据分析与挖掘跟金融这一方面关系密切。在新时代下，数据分析与挖掘被赋予重大的使命，在金融业中也发挥了重大的作用。

3.3.3 按学历进行筛选

通过对学历的筛选，我们得到如下饼图：



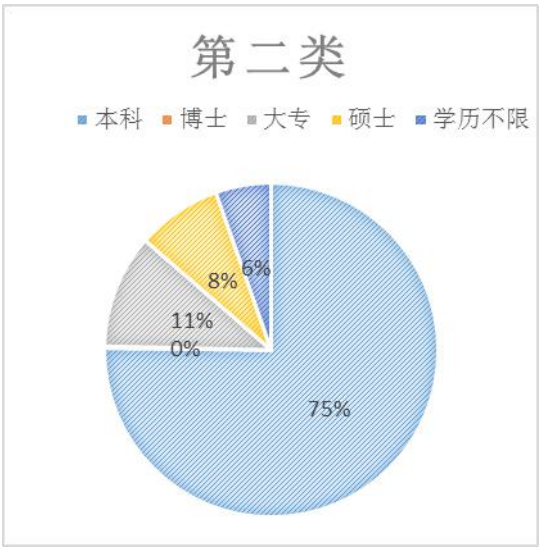


图 8：两类职位对应的学历要求情况图

由图 1 可知：

对于从事这两大类职位的学历要求其实不高，要求本科毕业高达 70%。并且这些职位也为大专生提供了就业机会，甚至有些招聘是不限学历的。相反的，在这两类新兴招聘中，要求硕士，博士与硕士的比例低于 10%。

3.3.4 按 salary（月薪）进行统计

研究发现这两大类职位的薪资都集中在 15k—30k，综合学历要求看出，这两大类职位属于学历门槛不高，但是薪资相对比较丰厚的职位，这两大新兴职位势必会成为求职心目中一个很理想的职位，将会受到人们的追捧。

通过对 salary（月薪）进行统计（详见附件 3），以下截取人数最多的前 10 名分析：

表 7： 两类职位的不同月薪集中范围

第一类		第二类	
月薪	人数	月薪	人数
10k-20k	247	15k-30k	456
8k-15k	158	10k-20k	351
15k-25k	137	15k-25k	327
15k-30k	137	20k-40k	209
8k-10k	129	10k-15k	139
10k-15k	121	8k-15k	131
5k-10k	89	20k-30k	130
4k-6k	85	15k-20k	89
6k-12k	73	8k-16k	73
6k-10k	68	20k-35k	55

通过表 3，发现这数据分析这一职位的薪资集中在 10k—20k，而第二类主要月薪范围更高，集中在 15k—30k，与其他普通的职业相比，数据分析与挖掘所在行业相对工资更高。综合学历要求看出，这两大类职位属于学历门槛不高，但是薪资相对比较丰厚的职位，这两大新兴职位势必会成为求职心目中一个很理想的职位，将会受到人们的追捧。

3.4 问题 4 结果分析

IT 人才市场的供求现状及未来的发展趋势

对筛选出来的所有 it 职位，按城市统计每个城市 it 职位招聘的个数，并按招聘职位个数进行降序排列，取前面 10 个城市，结果如 it 职位招聘地区排行表所示：

表 8：it 职位招聘地区排行表

城市	职位个数	排序
北京	30478	1
上海	14224	2
深圳	9735	3
广州	6668	4
杭州	6652	5
成都	2866	6
武汉	1830	7
南京	1638	8
西安	924	9
长沙	900	10

由上表可知：对 it 人才需求较大的十大城市分别为：北京、上海、深圳、广州、杭州、成都、武汉、南京、西安、长沙。北京对 it 人才的需求居于首位，而且远远高于排名第二的上海跟第三的深圳。从地域分布上，可以看出，对 it 人才需求较大的城市大部分集中在东部沿海一带，而且都是大城市。事实上“北上广深”地区 it 行业比较发达，对 it 人才的需求量较大，从而形成“聚集”效应，预计在未来很长一段时间里，对 it 人才需求量较大的还会集中在这些地区。

对筛选出来的所有 it 职位，按学历统计每个学历层次 it 职位招聘的个数，结果如 it 招聘职位学历分布表所示：

表 9：it 招聘职位学历分布表

学历	个数	排序
本科	84775	1
大专	66189	2
学历不限	26558	3
硕士	1479	4
博士	34	5
高中	7	6
中专	4	7
初中	1	8

由 it 招聘职位学历分布表可以看出，对 it 人才的要求不是很高，大部分集中在大专、本科学历，甚至出现了 26558 个不限学历的招聘职位，而硕士、博士以上学历的却非常少。事实上 it 行业属于“吃青春饭”的行业，很多做 it 的到最后都转型，当上管理层。这是一个属于年轻人的职业。

对筛选出来的所有 it 职位，按职位类型统计职位类型职位招聘的个数，并按招聘职位个数进行降序排列，取前面 10 个职位类型，结果如下表所示：

表 10：it 职位排名前 10 的职位类型

专业领域	计数	排序
后端开发	24122	1
移动开发	10542	2
前端开发	10091	3
测试	5967	4
运维	3204	5
高端计数职位	1917	6
dba	1574	7
企业软件	907	8
硬件开发	875	9
项目管理	703	10

由上表可知：it 行业需求较大的专业领域集中在后端开发、移动开发、前段开发、测试跟运维。

3.5 结合研究结果，给在校大学生就业规划提建议

网络招聘信息平台现已成为招聘者发布招聘信息和应聘者获取职位信息的主要渠道。如今是互联网时代，互联网日新月异，网上的招聘信息层出不穷，这给我们求职者找工作提供了一条方便快捷并且有效的途径。特别是在校大学生，往往可以通过结合自身专业条件和能力水平选择适合自己的岗位，投递简历联系应聘。

随着每年毕业生总量压力进一步增大，很多大学生难以在社会中找到一份适合自己的职位，被迫失业，而也有很多企业高新招聘却找不到合适的人才。这种供需矛盾的现象要求我们应该要有对自己将来所从事的行业有一定的规划准备以下为大学生就业规划的建议：

（1）提高自身能力储备，明确所学专业

大学生应该结合个人发展的需要，选择适合自己的学习内容、学习方法和学习方式，形成自己的学习目标并提高自己学习能力。要树立正确的职业理想，大学生一旦确定自己理想的职业，就会依据职业目标规划自己的学习和实践，并为获得理想的职业做好积极准备。大学生要明确了解自己所学的专业，包括专业的要求、专业适合的岗位、专业的职业类型、专业发展的前景等，才能对自己所学专业保持一个全面正确的了解，对学习内容有一定的认识，对就业有一个更有目的性的规划。

（2）了解专业的职位类型和职位要求

职业类型是一个较大的概念，一般按一定的规则、标准及方法，按照职业的性质和特点，把一般特征和本质特征相同或相似的社会职业，统一归纳为同一职业类型。所以不同的职位可以同属一个职业类型，如技术类职业类型，其职位可以是前端工程师、MySQL 数据库工程师、运维开发工程师等。大学生了解自己专业所属的职位类型可以知道与自己专业性质相类似的其他职位，进而可以对自己专业有更深刻的认识。也可以根据个人的需要和特点，选择相临近的职业。

另一方面，了解专业的职位要求，对自己的学习内容、专业素养、能力培养

都有一个初步的对照，根据职位要求进而制定就业规划。要明确不同招聘单位对职位的要求不同，如同为数据分析师，有些单位只要求掌握对数据的筛选；有些只需要求前端开发；有些则要求是广告平台数据分析师-java Hadoop，甚至是要求高级 Java 软件工程师。因此，在校大学生要严格要求自己，全面学习相关专业知识，全面掌握相关专业技能和岗位要求，形成多元化全面的发展，以便符合不同的岗位的工作要求。

（3）了解热门的行业和地域

本题通过分析研究得出热门行业基本为 IT 行业、数据分析师、Java 等，这些热门行业主要分布的地域为北京、上海、广州、深圳这些经济相当发达的大城市，对专业人才的需求量高，对专业人才的能力要求高，且薪资方面相对于小城市也高了许多。大学生要根据自己所学专业的前景去判定要不要多学专业外的热门行业的相关知识，以便遇到自己所学专业招聘岗位已饱和或发展前景不好能有另外的出路，有应聘别的工作的机会。大学生应该选择适合自己的发展区域，根据不同区域对工作岗位的工作内容、工作要求和工作能力等不同来要求自己，使自己具备成为一名合格工作者的标准。

（4）积累工作经验

根据网上招聘信息得知不同工作职位需要的应聘者要求是实习生，3-5 年或 5-10 年的工作，这需要在校大学生通过社会实践去积累工作经验，争取在校期间到各大知名企业去应聘实习生的工作。开始步入社会工作，有利于以后工作任职的顺利进行。

大学生认识要超前，要认识到单有理论知识没有实践经验，将来走向社会很难得到社会的认可。现在有不少用人单位明确表示不招刚毕业的大学生，确实有他们的理由，因为企业要招的人是不通过培训招进来就能立即上岗的人，而不是缺乏工作经验处理事情毫无条理的人，然而这却是很多应届大学毕业生的“软肋”。大学生可以利用寒暑假到各大企业、社会单位实习，如从大一就开始，平均拿三年的寒暑假计算，大概有 9 个月的时间可以积累到很多工作经验，特别要珍惜临近毕业的这一年实习机会。大学生要根据自己所学专业或毕业后打算从事哪方面的职业来确定实习的单位和实习的内容，不能毫无目的的到任何性质的单位实习，没有对自己以后从事的工作积累有用的经验，这样的社会实践对将来的就业帮助不会很大。

（5）不断提升自己，用能力说话

当今社会人才辈出，就业形势严峻，很多大学生都面临着一毕业就失业的窘境，归咎起来，都是一个能力问题。所以，大学生在校期间一定要把握好四年时光，努力学习，培养良好的心理素质，并且不断提升自己的能力，只有这样，将来在求职时才不至于处处碰壁，才能在艰难险阻中立于不败之地。首先，在校大学生应该多多参加一些学校举办的活动，多多上台去表现自己，不断提高自己的表达能力与胆识。许多高校会定期举办一些模拟就业的活动，大学生可以多多参与，从中获得一些经验体悟。

其次，一些学生缺乏市场意识，缺乏择业经验，很多学生在就业过程中主体意识薄弱，一是求职过程中过多依赖学校老师和家长，求职准备和主动性不够；二是，一些学生的查找资料和获取信息的能力太差，不懂得如何获取有效的网络招聘信息。所以，作为在校大学生，在刚踏入大学校门的那一刻开始，就应该培养自己的独立意识，而不应该过多地依赖别人，对于求职，应当有足够的准备和主动性，学会利用互联网的时代背景优势，获取有利的资源和信息。

大学生要时刻关注招聘信息，有目的的学习，应该从大一开始未雨绸缪，在大学期间不断提升自己，丰富社会实践工作，制定好就业计划，努力将自己打造成社会栋梁！

4 结论

对网络招聘信息进行分析研究，了解社会和相关行业的需求特点与趋势，对广大的求职者有重大意义，同时也是文本分析的一个课题、一个难题。传统的文本解读已经不能满足数据量庞大的网络招聘信息。本文采用根据 K-means 聚类方法和 Knn 最邻近分类，统计目前企业最需要的职业类型，并定义相关职业领域，深入分析人才市场的供需现状。

由分析结果可以看出，网络招聘中所需要的人才可以分为技术、职能、市场、销售、产品、运营、设计和金融七大职业分类，对各个行业领域出现的招聘次数进行计数，从而定义热门的地域，可以看出北上广发达地区需求量较大，通过排序得出各大职位的需求情况，并定义热门的职位，可以发现热门行业为移动与互联网相关职业。

统计新兴行业在公司在不同的发展阶段对这四类新兴职位的需求量，得出上市型和成长型需求量较大，另外分析了 IT 人才市场的供求现状及未来的发展趋势，可以看出发展前景相对较好。

5 参考文献

[1] 赵琳瑛.基于隐马尔科夫模型的中文命名实体识别研究.西安电子科技大学.2007
[2]翟东海，鱼江，高飞，于磊等。最大距离法选取初始簇中心的 K_means 文本聚类算法的研究.西南交通大学.2014
[3]朱志远.基于数据挖掘的网络招聘系统是设计与实现.电子科技大学.硕士学位论文.2013
[4]王千，王成，冯振元，叶金凤.K-means 聚类算法研究综述.2012
[5]张晓辉，李莹，王华勇等.应用特征聚合进行中文文本分类的改进 KNN 算法.东北大学.2003
[6] 卜凡军.KNN 算法的改进及其在文本分类中的应用.江南大学.硕士学位论文.2009
[7]曹卫峰.中文分词关键技术研究.南京理工大学.硕士学位论文.2009
[8]杨虎.面向海量短文文本去重技术的研究与实现.国防科学技术大学.2007