**知乎大数据之青年群体择偶观分析**

**作者**：黄旺辉、向优、姚培、刘欣，黄嘉翔，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：Spark；用户画像；文本聚类。

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 在青年的价值观体系中，择偶观是不可忽视的一个重要方面，对择偶问题的关注也成为青年生活的重要内容之一。在新世纪和多元时代下成长起来的"90后"在恋爱观、择偶观和婚姻观上有着自己独特鲜明的看法。以往通过人工问卷调查、采访等统计分析方式不够全面、细致，对青年群体择偶观科学分析存在一定的局限性。知乎作为主要由青年群体组成的社交问答网站，是很好的价值观平台。本案例基于知乎大数据，通过爬虫抓取全站择偶观相关问答及相关信息，采用Kmeans聚类等多种大数据处理算法，对青年群体的择偶标准即恋爱价值观进行详尽多元化的数据分析。最后本案例还利用了多种大数据可视化策略包括词云、图表分析等，形象直观地展示了分析结果，广大青年择偶提供一个科学参考依据。

**关键词**：Spark；分布式；用户画像；文本聚类

## 1引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为基于招聘网站的离线统计及实时分析系统，需引导学生进行的主要内容有：（1）使用爬虫工具从知乎网站爬取关于“当代男女择偶观”的问题回答以及相关的用户信息；（2）搭建大数据平台环境；（3）进行用户画像、文本聚类等大数据分析。

## 2背景介绍

择偶观是人们对选择配偶的看法和态度。古往今来，选择什么样的配偶，如何选择配偶，是人们一直在寻求的最佳模式。根据多项采访结果发现，对于所有年龄层面来讲，关于选择未来配偶，人们往往更重视对方的个人品质而不是个性魅力，例如忠诚、可靠性、善良等个人特征，希望对方和自己有共同的价值标准、类似的生活态度、性格以及宗教信仰。当然，生活在密集地区和高消费地区的人类往往要求更多的资源，比如社会地位、经济能力、家庭状况、外貌特征等。

针对青年这一特殊而重要的人类群体,他们生理、心理的发展使他们产生对恋爱、交友的需要。在青年的价值观体系中,择偶观是不可忽视的一个重要方面，对择偶问题的关注也成为青年生活的重要内容之一。对现代青年择偶观的差异分析不仅能从一个侧面揭示当代青年婚恋价值观的变动特点,还能对其潜在的价值取向隐患起到预估作用。由于他们的择偶观是异于其他年龄层面的群体，所以，想要对青年群体的择偶标准即恋爱价值观进行相应的数据分析，分别分析出女性和男性各自择偶的关注点、侧重点，为广大青年择偶提供一个参考依据。

## 3内容

该案例的主要内容主要分为三个小节，分别为数据集收集、大数据平台环境搭建，以及用户画像、文本聚类等大数据分析。

3.1数据集收集

数据集主要是来源于知乎网站，对于关于“当代男女择偶观”的问题回答以及相关的用户信息进行爬虫，一共是有80000多条数据。

本案例主要使用的Python语言进行数据爬虫，采用的是Selenium+Requests+MySQL模拟用户登录来爬取数据，并储存在本地数据库。

Selenium:一个用于Web应用程序的自动化测试工具，底层框架采用JavaScript模拟真实用户对浏览器进行操作，使浏览器可以按照脚本代码自动对于点击、输入、验证、打开等等操作做出反应。同时，它的使用覆盖面比较广，可以在Windows、Linux、Macintosh上的IE、GoogleChrome等不同平台使用。

Requests:Requests是一个Python的HTTP库，可以允许用户发送原始的HTTP请求，不需要自己手动操作。既不用为URL手动添加查询字串，也不需要对于POST进行表单编码。

MySQL:开源的关系数据库管理系统，当前主流的中小型数据存储软件。

3.2 大数据平台环境搭建

在项目的实现过程中使用到了多种工具进行数据的存储、计算，下面介绍所需搭建的环境，主要包括Hadoop集群搭建、Spark集群搭建。

所需安装软件版本：

|  |  |
| --- | --- |
| 软件 | 版本号 |
| Jdk | 1.8 |
| Hadoop | 2.7.7 |
| Spark | 2.1.1 |

3.3用户画像、文本聚类等大数据分析

（1）用户画像分析实现过程如下：

a.读取 MySQL 表中数据，如：性别、职业、地区等。

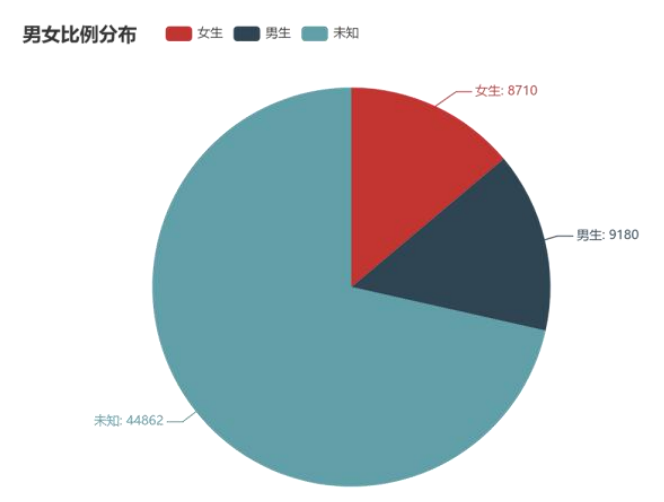
b.对数据进行清洗过滤，如：一些无效数据等。

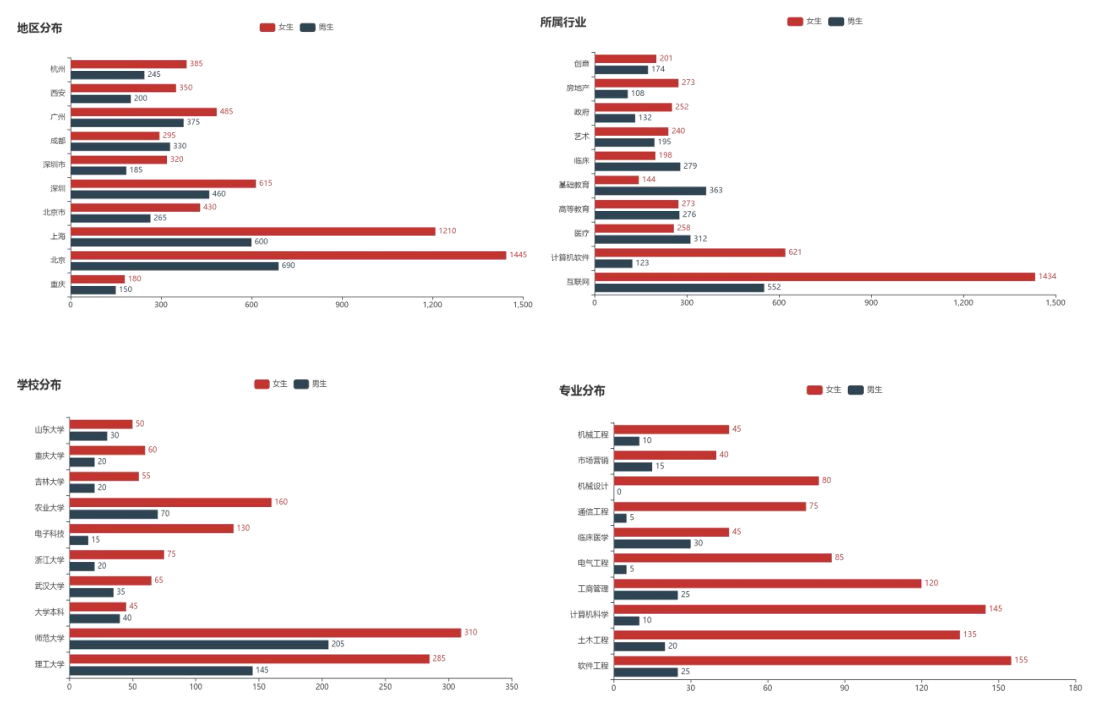
c.对读取的的数据进行统计分析。

d.使用 Echars 对用户进行结果可视化

根据数据分析，我们得出部分分析结果。绝大多用户不太愿意透露自己的性 别，在已知性别的用户中，参加此问题男生和女生的比例差不多为１：１。用户 地区分布以一线城市为主，如：北京、深圳、广州等第；所属行业以计算机、软 件、土木居多；毕业院校多为知名院校，如中山大学、重庆大学、电子科技大学 等。可见用户群体素质较高，以工科为主，拥有较高学历和视野。

实验用户画像分析结果见下：





（2）词频统计

首先，从 MySQL 中读取数据到 Spark 集群中；然后采用结巴分词的精确模 式，对用户回答内容精确分词，提取其中可以成词的所有词语；结合 Spark 的 RDD 操作，统计所有词语出现的频率，最后利用 WordCloud 对结果进行可视化。



所有回答词频统计结果见下图，可以发现：身高、性格、父母、家庭、学历出现频率占据了前五位。、



所有回答男、女生回答词频统计结果见下图。可以发现男、女生词频统计结果差别不大，依据是身高、性格、家庭、学历、父母等占据了最高比重。但二者统计也有区别，比如在女生回答出现有关身高数值170、175，但是在男生回答中出现的数值是 160、165，可见男、女需求侧重点还是有所不同。



所有重庆回答男、女生回答词频统计结果见下图。显而易见这和所有男、女生回答词频差别不大，但重庆占据了词频出现最高值，这也间接说明我们统计结果具有较高可信度。



（3）文本聚类

KMeans 是一种无监督的文本聚类算法，具体应用到我们项目进行文本聚类算法算法设计流程如下：

**a. 分词**

对所有答案进行分词，返回所有回答分词结果list列表。



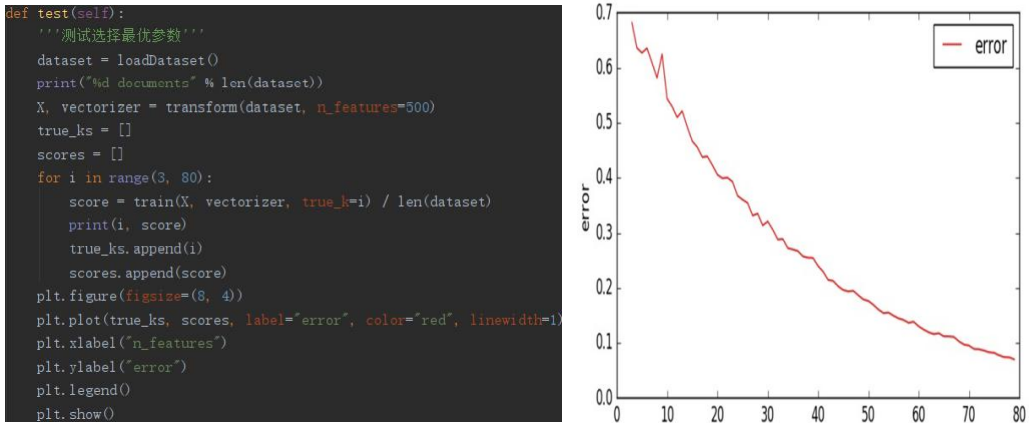
**b. 计算 TD-DIF 矩阵**

获取回答 list 的 TD-DIF 矩阵，用于聚类。



**c. 测试聚类簇数**

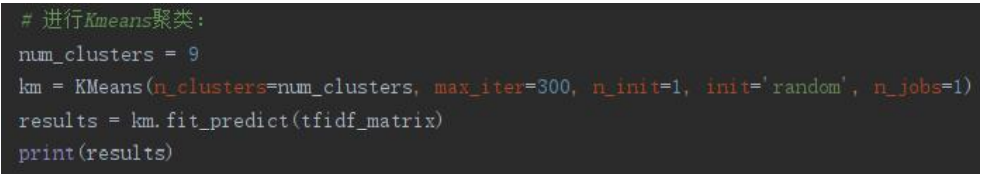
KMeans 虽然是一种简单的无监督聚类算法，但是到底将答案聚为几类是需 要经过进行测试选取最佳聚类簇数。



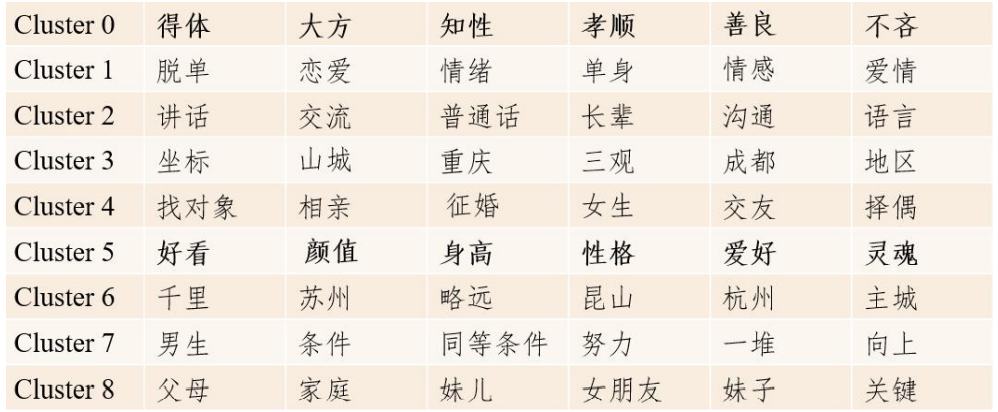
经过测试我们发现，从上图中在 k=9 处出现一个较明显的拐点，因此选择 k=9 作为中心点的个数。

**d. KMeans 聚类**

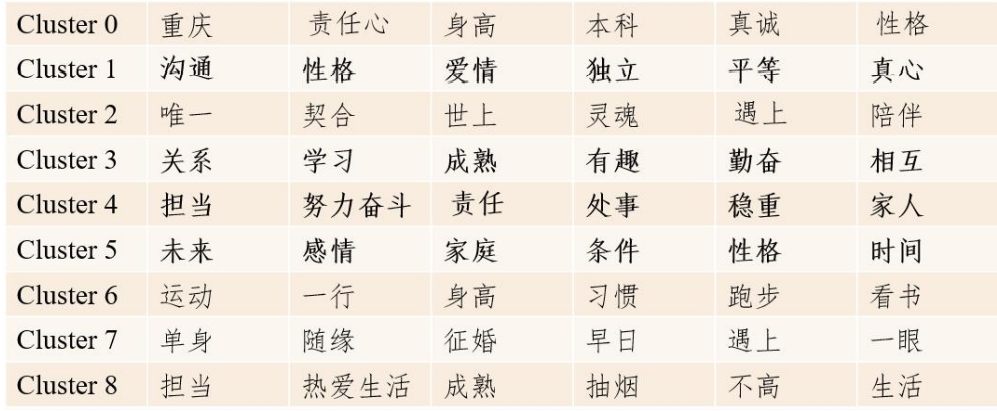
主要通过调用 sklearn 库中 KMeans()方法进行聚类。



重庆男生聚类结果见下图，可看出重庆男生对于女生的择偶要求基本在于个人品质、颜值外形方面，对于配偶的地区职业无明显要求。



重庆女生聚类结果见下图，可看出重庆女生对于男生的要求更多的在于男生个人的内在品质、未来发展以及家庭情况，对于外在条件没有无明显要求。



## 4小结

本案例主要对知乎用户进行了青年群体择偶观大数据分析，需要使用爬虫工具从知乎网站爬取关于“当代男女择偶观”的问题回答以及相关的用户信息，搭建大数据平台环境，对数据进行用户画像、文本聚类等分析。该案例主题新颖，结合了工业界现实需求与大数据分析与挖掘的多种理论与技术，可以充分增强学生的实践能力与理论基础。另外，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、用户画像完成任务。

## 附录

1.本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2.本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），scikit-learn（机器学习算法库），matplotlib（可视化绘图库）。

3.本案例参考文献如下：

[1]周志华.机器学习[M].北京:清华大学出版社,2016.

参考内容：数据预处理理论；数据挖掘