**知乎大数据之青年群体择偶观分析教学指导手册**

**教学适用的课程：**《大数据架构与技术》

**教学适用的专业：**电子信息

**教学目的与用途：**理解实际工业场景下的大数据应用；掌握大数据预处理、数据分析相关知识点；掌握大数据处理与分析的基本流程。

**教学内容:**

（1）理论依据：本案例基于知乎大数据，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。案例整合了常用的数据挖掘方法和大数据处理技术，进行数据预处理、大数据平台搭建、项目可视化展示，分析出了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。

（2）涉及知识点：爬虫；数据预处理；Hadoop；Spark；文本聚类。

（3）分析路径：首先抛出择偶观对青年的价值观体系的重要性。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，并将爬取数据进行一些预处理。然后搭建配置环境，主要包括Hadoop集群搭建、Spark集群搭建。最后按照用户画像、词频统计、KMeans 文本聚类的思路进行了数据分析。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。

**启发思考题:**

如何更精细刻画出用户画像？（参考答案：多维度）

**建议课堂计划：**

1. 时间安排：3课时
2. 学生学习准备：了解常见的爬虫、数据预处理、大数据处理平台的理论知识。
3. 分组及讨论内容：分组讨论启发思考题。
4. 案例开场白：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。这节课我们基于知乎大数据，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。
5. 结束总结：这节课我们通过爬虫抓取了知乎有关择偶问题的相关回答，然后基于python在腾讯云搭建了Spark分布式平台，最后进行用户画像、文本聚类等大数据分析。我们通过平时课程所学知识去解决了实际生活中的问题，课后感兴趣的同学可以搜集相关资料，进一步了解大数据在其他行业的应用，相信你们会收获更多。
6. 案例引导建议：在教师简要分析应用场景后，鼓励学生自主考虑如何使用所学知识解决该场景问题。

**参考文献:**

[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

**其他教学支持材料：**

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）。