# 基于哔哩哔哩视频网站的热门视频分析案例教学指导手册

**教学适用的课程：**《大数据架构与设计》

**教学适用的专业：**电子信息

**教学目的与用途：**理解实际背景下的大数据应用；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：本案例基于哔哩哔哩视频网站的热门视频分析。案例整合了常用的数据采集与数据挖掘方法，进行了有监督学习和无监督学习，分析了热门视频的共通点。

（2）涉及知识点：数据预处理；回归预测；分类预测；预测评估。

（3）分析路径：本案例首先介绍了短视频时代背景，然后从哔哩哔哩视频网站中随机爬取视频的信息并进行数据存储，并对数据进行预处理。然后基于爬取到的视频信息，从众多视频中筛选出热门视频。对筛选出的热门视频做分析，以视频标题词频统计和视频播放时长两个角度探寻其共通点，并利用可视化对结果进行展示。

**启发思考题:**

除了本案例中所使用的视频标题词频统计和视频播放时长两个角度，还有哪些可以成为热门视频共通点的参考角度？（参考答案：视频弹幕数量）

**建议课堂计划：**

1. 时间安排：3课时
2. 学生学习准备：了解常见的数据预处理、分类、聚类方法的理论知识。
3. 分组及讨论内容：分组讨论启发思考题。
4. 案例开场白：近年来，在哔哩哔哩视频网站等媒体平台上产生了大量具有社会影响力的短视频作品。哔哩哔哩视频网站每日有数以万计的短视频发布，但成为热门的短视频是为凤毛麟角。这节课我们就将走进这些热门视频，探寻背后的共通点。
5. 结束总结：这节课我们基于哔哩哔哩网站采集了视频数据，并通过有监督方法和无监督方法完成了热门视频筛选，并对热门视频进行了视频标题词频统计和视频播放时长两个角度的共通点探寻。通过本堂课，我们对身边产生的现象探寻了其背后的数据支撑，也鼓励大家课后继续探索与实践。
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，KMeans、SVM

1. 吕云翔,张扬编.Python网络爬虫与数据采集[M].人民邮电出版社,2021.

参考内容：数据采集理论

**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）。