# 基于招聘网站的离线统计及实时分析系统教学指导手册

**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：本案例基于大数据技术对招聘网站的离线统计及实时分析。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，实现了对岗位缺口的数据分析，运用大数据对信息人才就业的赋能。

（2）涉及知识点：爬虫；数据预处理；Hadoop；Spark；推荐算法。

（3）分析路径：首先抛出国内外经济状况，当前就业背景，并进行数据集展示，介绍相关字段含义。通过爬虫爬取某招聘网站上的六种热门编程语言的全国程序员岗位信息，并将爬取数据进行一些预处理。然后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，主要包括Hadoop集群搭建、Spark集群搭建、Zookeeper、Kafka、Flume、MongoDB、Redis搭建。最后构建了符合分布式大数据体系的一体化的岗位信息统计、推荐系统，包含了统计以及离线推荐与实时推荐体系，综合利用了协同过滤算法以及基于内容的推荐方法来提供混合推荐。提供了从分布式部署、大屏展示应用、业务系统、算法设计实现等多方位的闭环的大数据业务实现。

**启发思考题:**

离线推荐服务除了综合用户所有历史数据，使用协同过滤算法以及基于内容的推荐方法来提供混合推荐，还能想到其他的解决方法吗？（参考答案：基于规则推荐、基于效用推荐、基于知识推荐）

**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：随着当前网络信息不断累加，使得很多信息出现了过饱和的现象，但信息的增多不能很好的服务用户，反而使用户产生巨大信息的迷茫之中。这节课我们的项目注重于新时代社会信息人才的缺口，抓住用户的人才需求、就业需求的痛点、痒点，着力于运用大数据对信息人才的就业赋能。
5. 结束总结：这节课我们通过爬虫爬取某招聘网站上的六种热门编程语言的全国程序员岗位信息，并将爬取数据进行一些预处理。然后使用到了多种工具进行数据的存储、计算。最后构建了符合分布式大数据体系的一体化的岗位信息统计、推荐系统。我们通过平时课程所学知识去解决了实际工业生产中的问题，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）。