# 2022-2023学年第2学期

# 《大数据分析与领域建模》期末课程报告

# 任务书&指导书

**一、选题要求：**

**1.综合设计（共60分）：**

请在大型数据集上结合Spark各类技术进行综合分析，从数据预处理入手，经过整理与清洗后在Hadoop + Spark集群上对数据集进行充分分析，最后结合数据可视化技术展现分析结果。具体过程如下：

（1）对数据集进行数据预处理（比如选取部分字段、进行格式转换等），然后保存到HDFS中；可以使用任意编程语言；

（2）使用Spark对数据进行分析（只能使用Scala语言），可以任意使用SparkCore、SparkSQL、SparkStreaming和SparkMLlib组件，只要使用了Spark编程知识即可；如果有需要，分析结果也可以保存到MySQL或HBase或MongoDB中；

（3）对分析结果进行可视化呈现，可以任意选择可视化方法（比如Python可视化、网页可视化以及其他可视化方法），可以使用任意语言（包括Python、Java等在内的任意语言）。

可选数据集包括但不限于（如有大数据相关参赛经历并采用最新竞赛数据集将有创新加分）：

[1)泰坦尼克号生还数据集](https://dblab.xmu.edu.cn/blog/9281-6/)

[2)美剧《权力的游戏》剧集数据集](https://dblab.xmu.edu.cn/blog/9281-5/)

[3)Covid-19传播数据集](https://dblab.xmu.edu.cn/blog/9281-4/)

[4)DOTA2比赛数据集](https://dblab.xmu.edu.cn/blog/9281-2/)

[5)咖啡连锁店数据集](https://dblab.xmu.edu.cn/blog/9281-3/)

[6)气象监测数据集](https://dblab.xmu.edu.cn/blog/3525-6)

[7)淘宝数据集](https://dblab.xmu.edu.cn/blog/3525-3/)

[8)电信客户数据集](https://dblab.xmu.edu.cn/blog/3525-2)

[9)NBA球员数据集](https://dblab.xmu.edu.cn/blog/3525-1)

10)豆瓣电影数据集

11)奥运会数据集

12)音乐专辑数据集

注意：为了避免作业内容重复与雷同，每个班级内，针对每个数据集不可以有超过3人选用。

**2.科技写作与翻译（共20分）：**

（1）请针对上述大型数据集的处理、分析与可视化过程撰写一篇课程报告(纸质版，双面打印)，字数不少于1500字，报告结构请参考“设计要求”中关于课程报告的相关要求。

（2）根据报告内容撰写中文摘要（不少于400字），并将其翻译成英文。

### 二、设计要求：

（1）认真进行分析与设计，并将分析和设计中间过程写入课程报告中；

（2）根据选题要求，确定设计的内容，设计出正确、合理、有实用价值的应用程序，并在课程报告中给出关键步骤及运行结果截图。

（3）使用Scala语言编程实现算法步骤，在Spark集群模式中运行。

（4）课程报告的封面格式请参考附件1，报告正文结构框架及具体形式可以自己组织，建议包含以下内容：

①选题名称

②中英文摘要

③分析、设计过程

④关键步骤、中间过程以及程序运行结果截图

⑤关键知识点总结与思考

⑥参考文献（书籍、网络资料等）

### 四、评分标准：

期末设计考核将综合考虑学生的学习态度、设计思想和设计方案的正确性，设计效果以及课程设计报告书的质量。具体评分标准如下：

设置5个评分点：

（1）设计方案正确，具有可行性、创新性；（25%）

（2）程序运行结果正确，整体效果较好；（25%）

（3）设计报告规范、质量高、篇幅合理（正文不少于1500字，摘要不少于400字）；（15%）

（4）报告内容完整、逻辑清晰，概念明确。（10%）

（5）英文摘要翻译准确流畅、书写规范；（25%）

按上述5项分别记分后求和，总分按60%的比例记入总评成绩。

优秀（100～90分），良好（80～89分），中等（70～79分），及格（60～69分），不及格（0～59分）。

封面样式请参考附件1。

附件1

**2022-2023学年第二学期**

**《大数据分析与领域建模》**

**期末课程报告**

**班 级**

**姓 名**

**学 号**

**指导教师**

**日 期**