5.14 星期日

按照老师给的流程进行操作和安装，搭建 windows 环境下的 Spark 环境，成功搭建 windos10/11 下的 spark 环境，并成功在 vscode 中跑出测试代码。

5.15 星期一

根据给定的啤酒销售数据和去年同期销量数据，把 xlsx 文件转换为 TXT 文件，针对 11 月份啤酒销售数据，并通过编程进行数据处理和计算。主要是基于 spark 进行 RDD 的一些基础编程。

本次实验的最主要收获即为熟悉了 spark 环境以及 RDD 对象的操作。从最开始对于大数据环境的一无所知，到现在能够使用其解决一些实际问题，这无疑是收获巨大的，也同时告诉我们其实大数据没有想象的那么难操作，只要有学习的决心，实际上其和普通的编程没有什么区别。

5.16 星期二

主要包括三个内容：SparkSQL 的基本操作、RDD 到 DataFrame 的转换、使用 DataFrame 读写 mySQL 数据库。在安装 mysql 的驱动时，发现仅仅在 spark 的安装目录下存放是不够的，还需要在相应的 conda 环境的 pyspark 的 jars 文件夹下存放，才可运行。

熟悉了 sparkSQL 的操作以及 RDD 和 DataFrame 的转换等，对于大数据的相关操作有了更加深刻的理解。

5.18 星期四

从文件中导入数据，并转化为 DataFrame；训练决策树模型，用于预测居民收入是否超过 50K；对 Test 数据集进行验证，输出模型的准确率。

机器学习虽然有着比较复杂的运算逻辑，但是经过 Ml 库包装之后变成了非常结构化、简洁化、公式化的流程。事实上，只要我们将数据处理成标准格式，后面我们只需要按部就班地创建决策树模型，进行训练，进行预测，精度评定即可。

5.19 星期五

搭建 Kafka，配置对应的 Spark，并验证是否正确运行。编写车辆位置数据模拟生成程序。从车辆坐标文件中，获取车辆轨迹信息，然后定时将数据发送到指定 Kafka 的消息队列的车辆位置 topic 中。基于 Spark Structured Streaming，编

写车辆轨迹处理程序，实时计算车辆进行速度。假设出发是车速为 0，每收到一条对应车辆的坐标信息，就根据收到的坐标点和上一次的坐标点计算之间距离，然后距离除以时间差，作为当前车速。

5.20 星期六

完成车辆轨迹的分析（速度计算、停留点分析、运动状态分析），在计算距离函数中使用 sin、cos 等一系列的数学公式时，我开始调用的是 pyspark.sql 的 functions 中的函数，但由于其操作对象是 column 而报错（当时就是说操作对象不为 column）。解决方法是调用 math 中的相关函数。

通过本次实验，主要是复习巩固了 spark 的基础知识，将所学内容运用到实际当中。