作业说明（35 分）

一、作业描述：给定类似如下的 SQL 查询和对应的查询计划，使用机器学习算法估算其查询规模（cardinality）

二、所需软件（仅推荐）

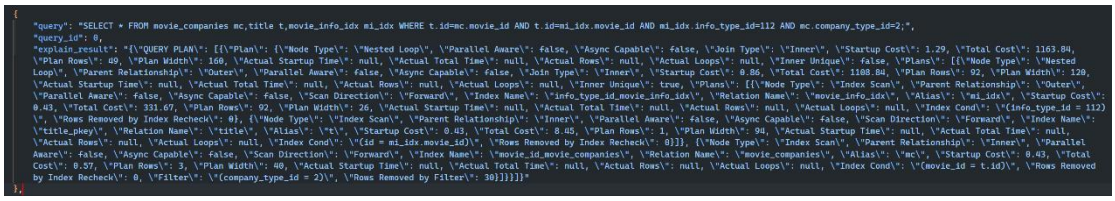
1. Anaconda
2. Pytorch/TensorFlow
3. Sklearn，numpy，pandas

三、数据说明

1. 提供训练样本 60000 条，待预测样本 1070 条
2. 需要大家提供待预测样本的预测结果，并以如下方式提供：

命名要求： 预测结果_姓名_学号.csv（如 预测结果_张三_2021000000.csv）提交首行需要包括对应的列名如下。

```
Query ID,Predicted Cardinality
0,0
1,0
2,2
3,6
4,12
5,20
```



3. 数据格式：

提供 JSON 格式的训练和测试数据，query 表示对应的 SQL 查询，query_id 表示查询的 id，explain_result 表示查询计划，可以用 Python 中的 json 包处理（json.loads）

4. column_min_max_vals.csv 表示数据库各个列的最大值、最小值、基数和列所包含不同值的数量。

四、作业提交

1. 提交内容：
   a) 实验报告：实现代码及具体实现报告（不超过 10 页）。
   b) 源代码。
   c) 预测结果文件。
   d) 验收 PPT
2. 提交地点：报告、代码及最优结果提交到课堂派
3. 提交截止时间点：2025 年 5 月 29 日 24 点（晚一天扣 1 分，本次作业分值扣完为止）
4. 预计验收日期：2025 年 5 月 30 日上课时间

五、模型评测

大家可以通过 10.77.110.133:19052 评测自己的预测效果（需内网访问）

在网站评测自己的结果时，务必每次提交都使用同样的文件名（预测结果_姓名_学号.csv），需要大家自己记录每次提交的不同结果，否则后台文件太多会影响大家评测效率。