“自然语言处理”实验报告

实验2：命名实体识别

姓名：石翔宇 学号：1190200523

Email：xyu.shi@hit.edu.cn

# 1. 实验概述

本次实验学习使用HMM、ME、CRF和深度学习等不同的命名实体识别方法，并在两个不用的数据集上进行实验。实验需要用到百度AI Studio平台和华为云计算平台。

# 2. 实验目标

* 通过不同方法的结果对比，掌握不同实体识别方法的优缺点。
* 通过对不同数据集的使用，掌握命名实体识别需要的数据预处理、模型训练、模型测试和评价方法。
* 通过对百度AI Studio平台和华为云平台的使用，了解国产主流人工智能平台提供的学习资源和计算资源，并能使用这些平台完成特定的自然语言处理任务。
* 通过实验报告的撰写，提高分析能力、写作能力和表达能力。

# 3. 命名实体识别的评价

本次实验学习使用HMM、ME、CRF和深度学习等不同的命名实体识别方法，并在两个不用的数据集上进行实验。实验需要用到百度AI Studio平台和华为云计算平台。

请在这部分给出：

1）命名实体识别的评价指标及计算公式

2）自己实现的实体级命名实体识别的评价程序的实现思路（用文字、流程图及伪代码进行说明）；

3）自己实现的评价程序的调用方法（含参数说明）

4）对HMM模型在ner\_char\_data目录下的test.txt文件上的识别结果分别按词级别和实体级别进行评价，给出分类别的P、R、F值和总体的P、R、F值，并分析两种评价方式结果差别的原因。

# 4.基于最大熵模型的实体识别

请在这部分给出：

1）最大熵模型原理的简单介绍

2）描述自己实现的基于最大熵模型的实体识别系统

3）详细说明自己实现的基于最大熵模型的实体识别系统利用的特征

4）评价自己实现的基于最大熵模型的实体识别系统的效果

利用ner\_char\_data目录下的train.txt文件训练模型，在test.txt文件上进行测试。对测试结果按词级别和实体级别进行评价，给出分类别的P、R、F值和总体的P、R、F值。

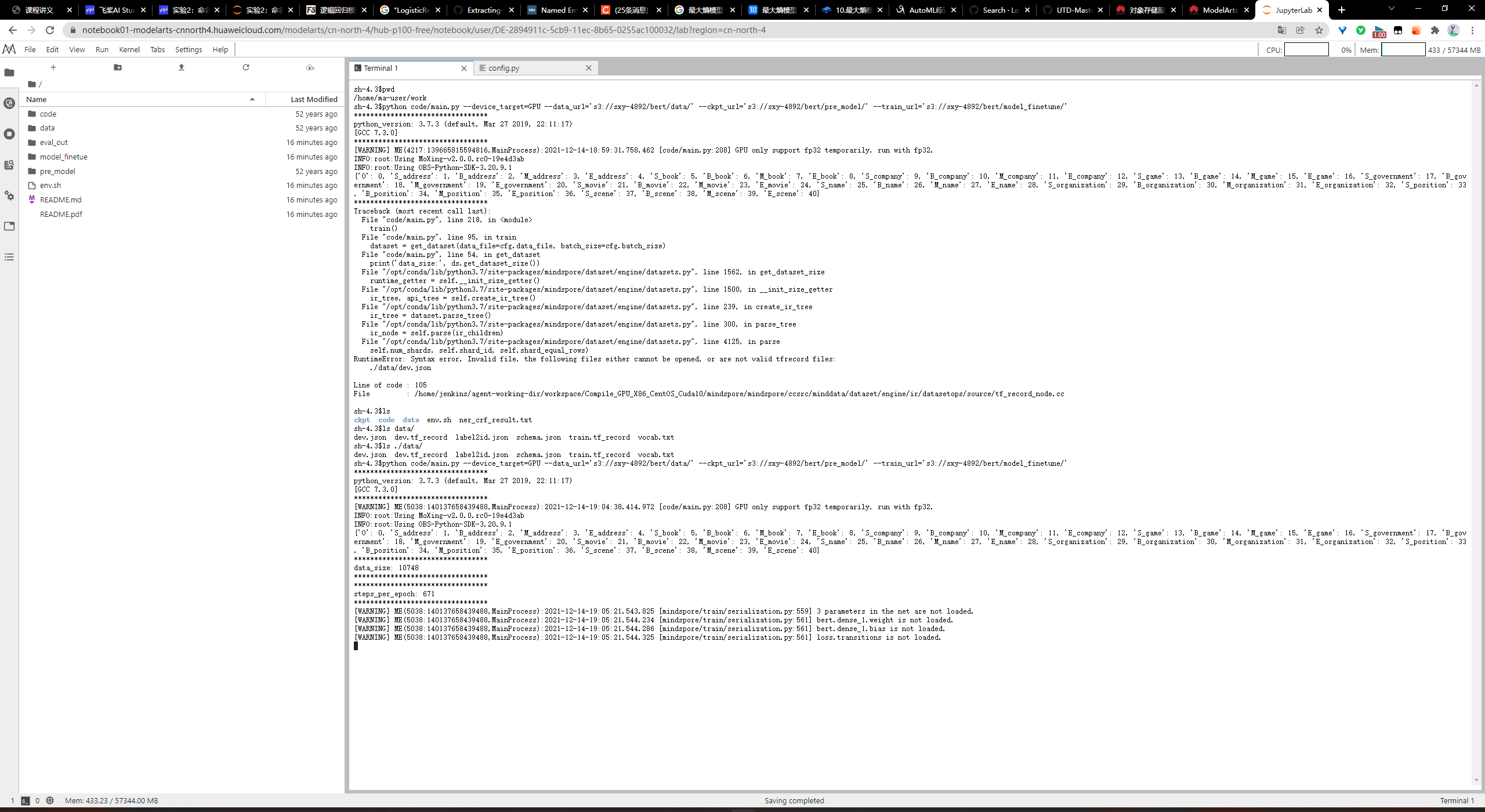
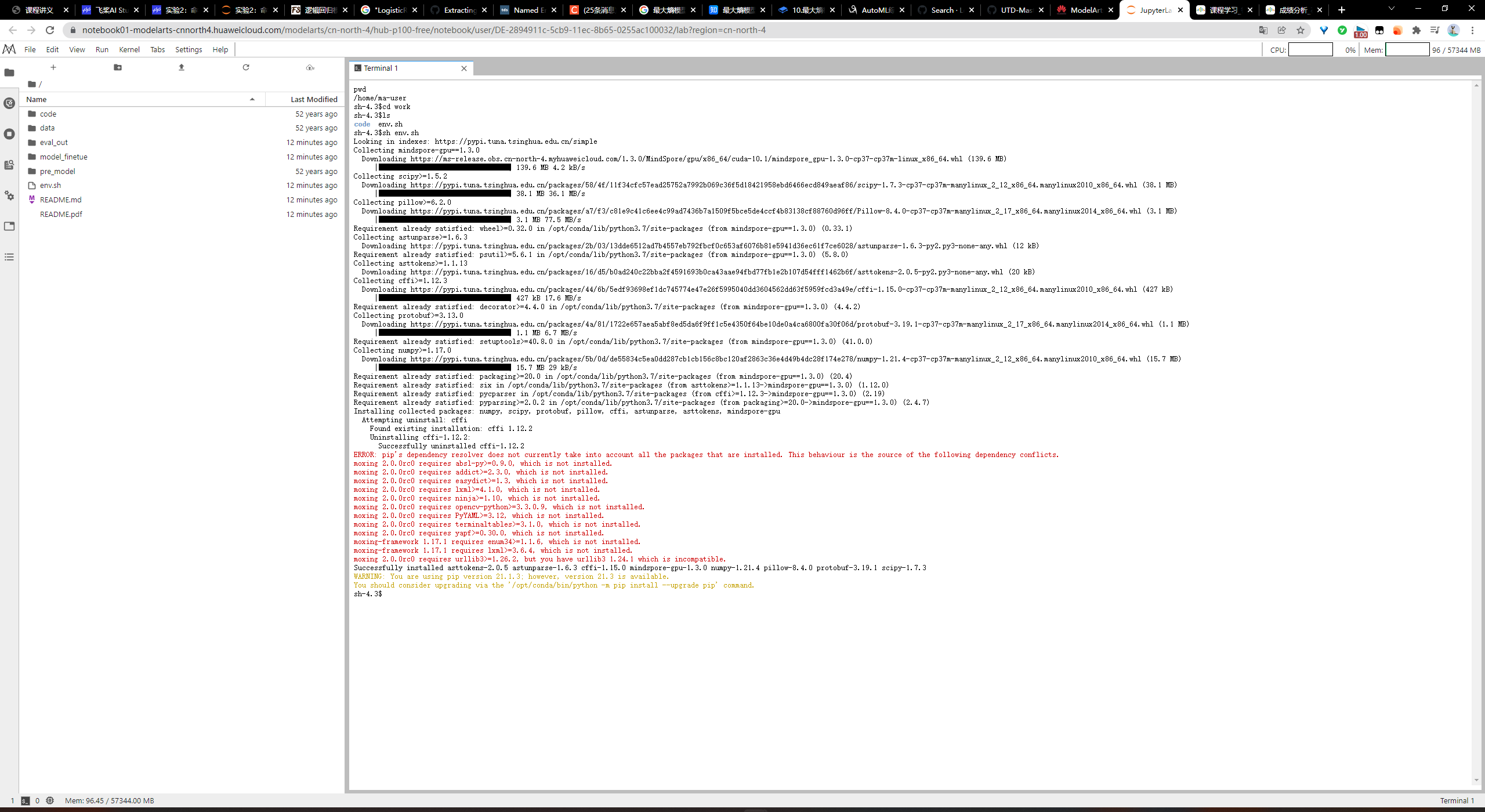
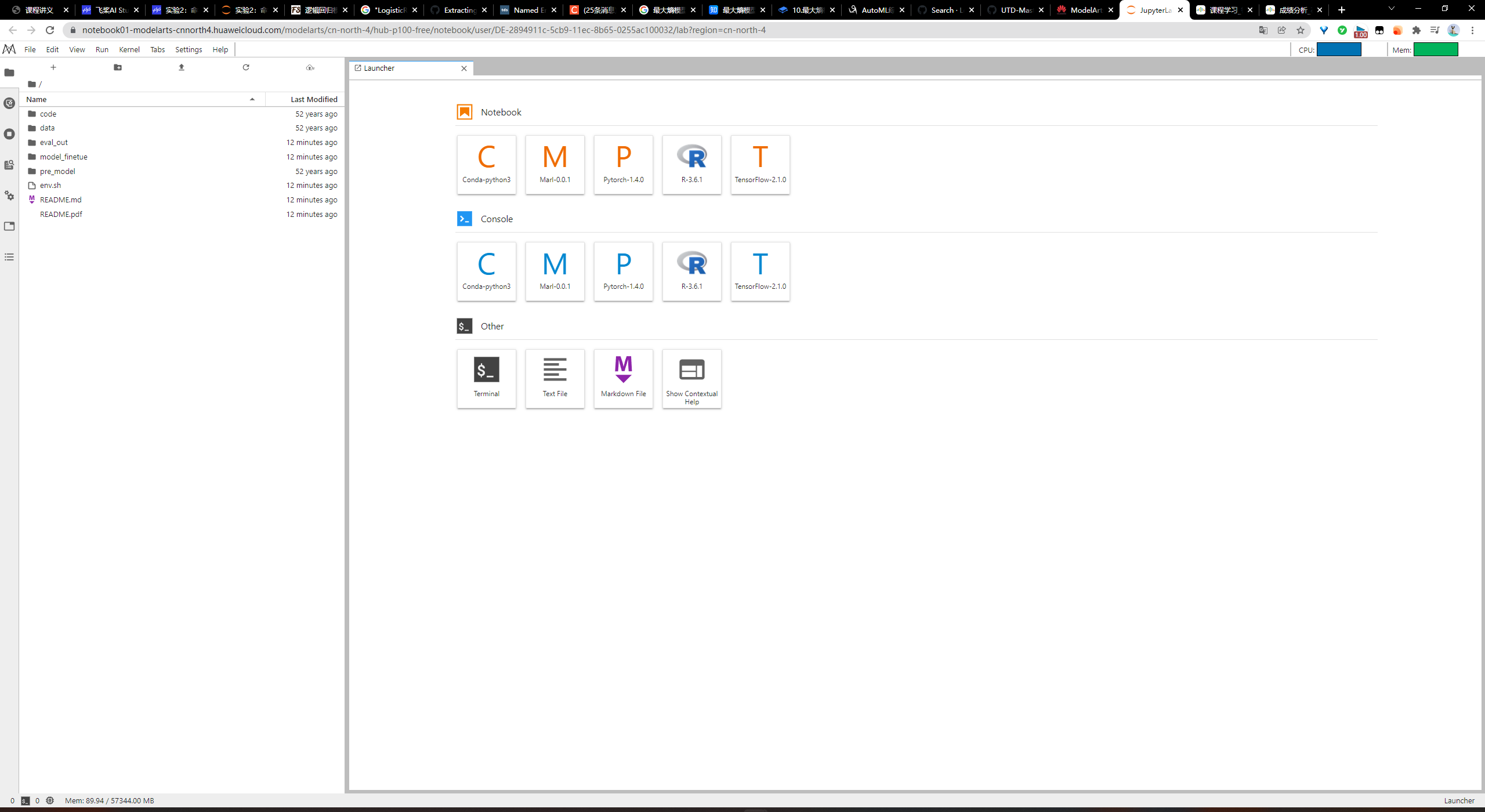
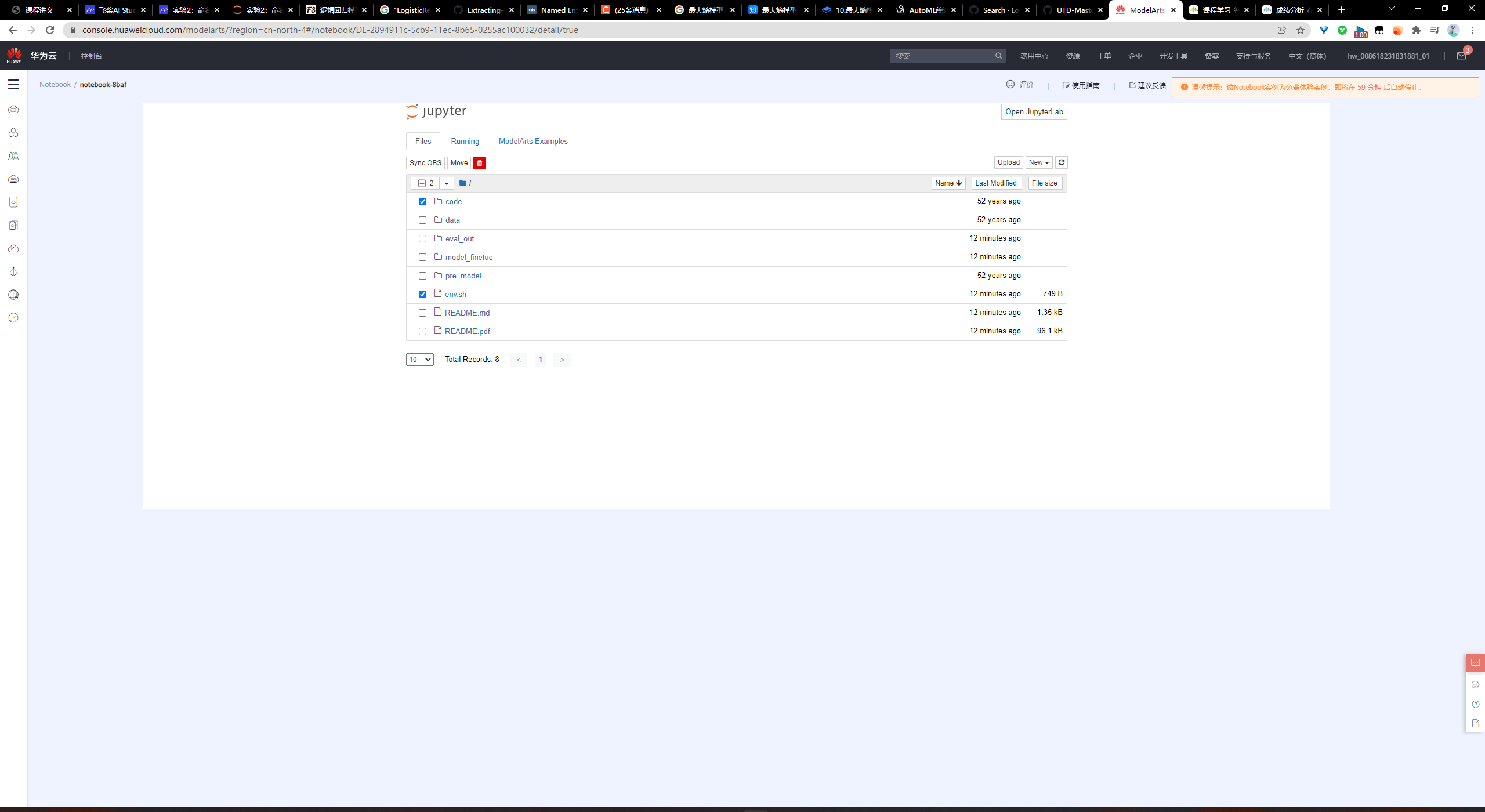
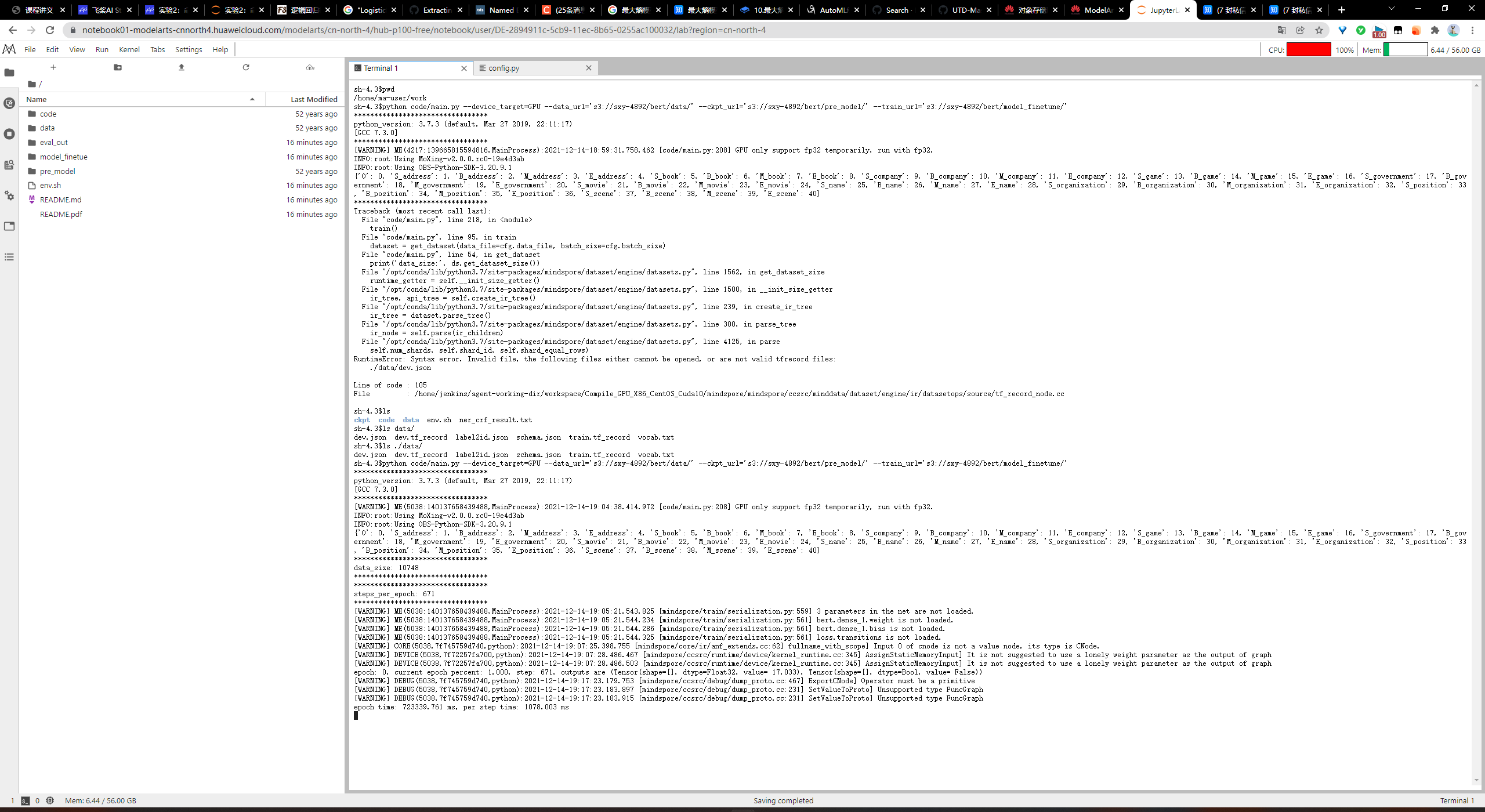
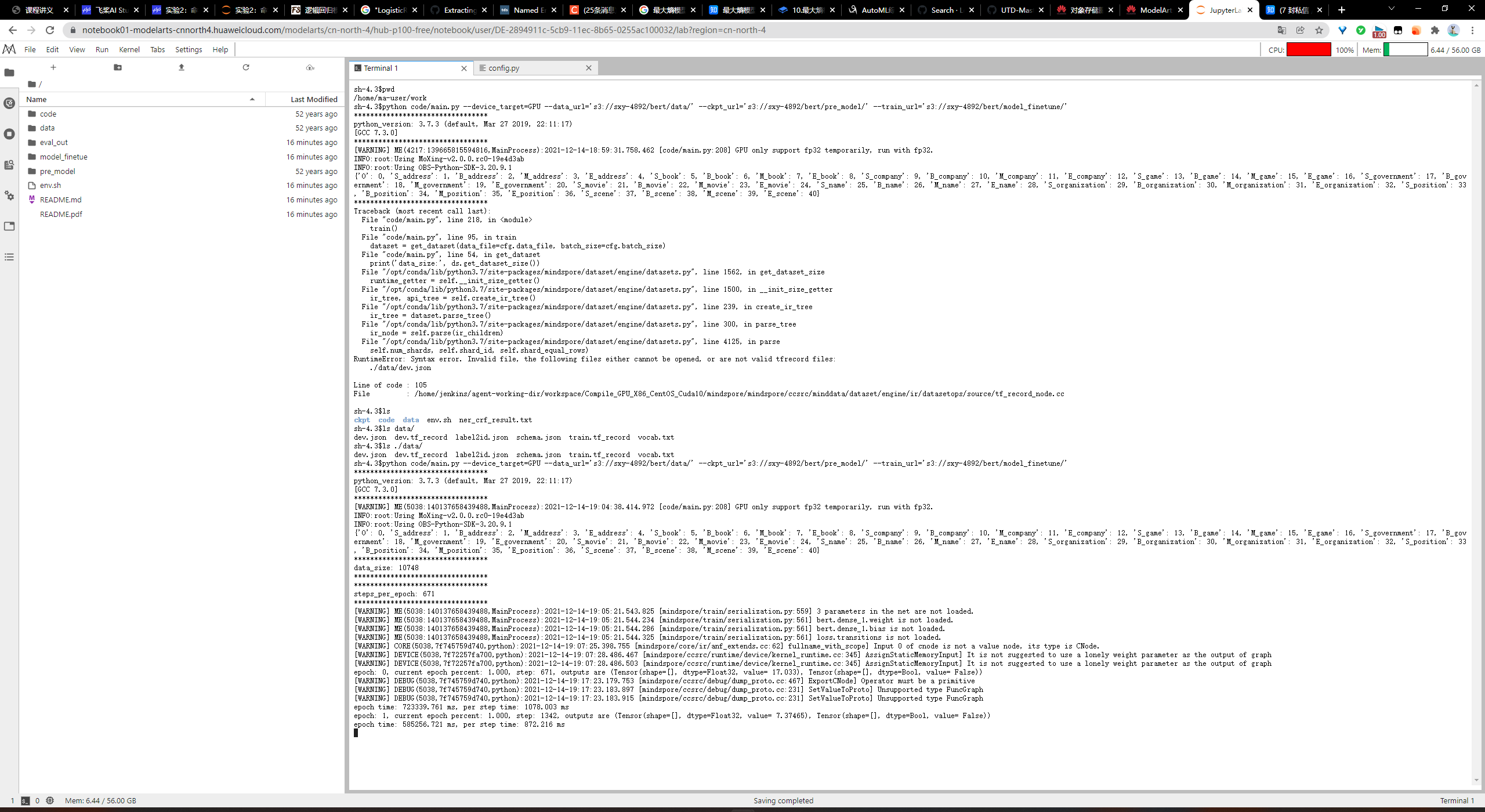
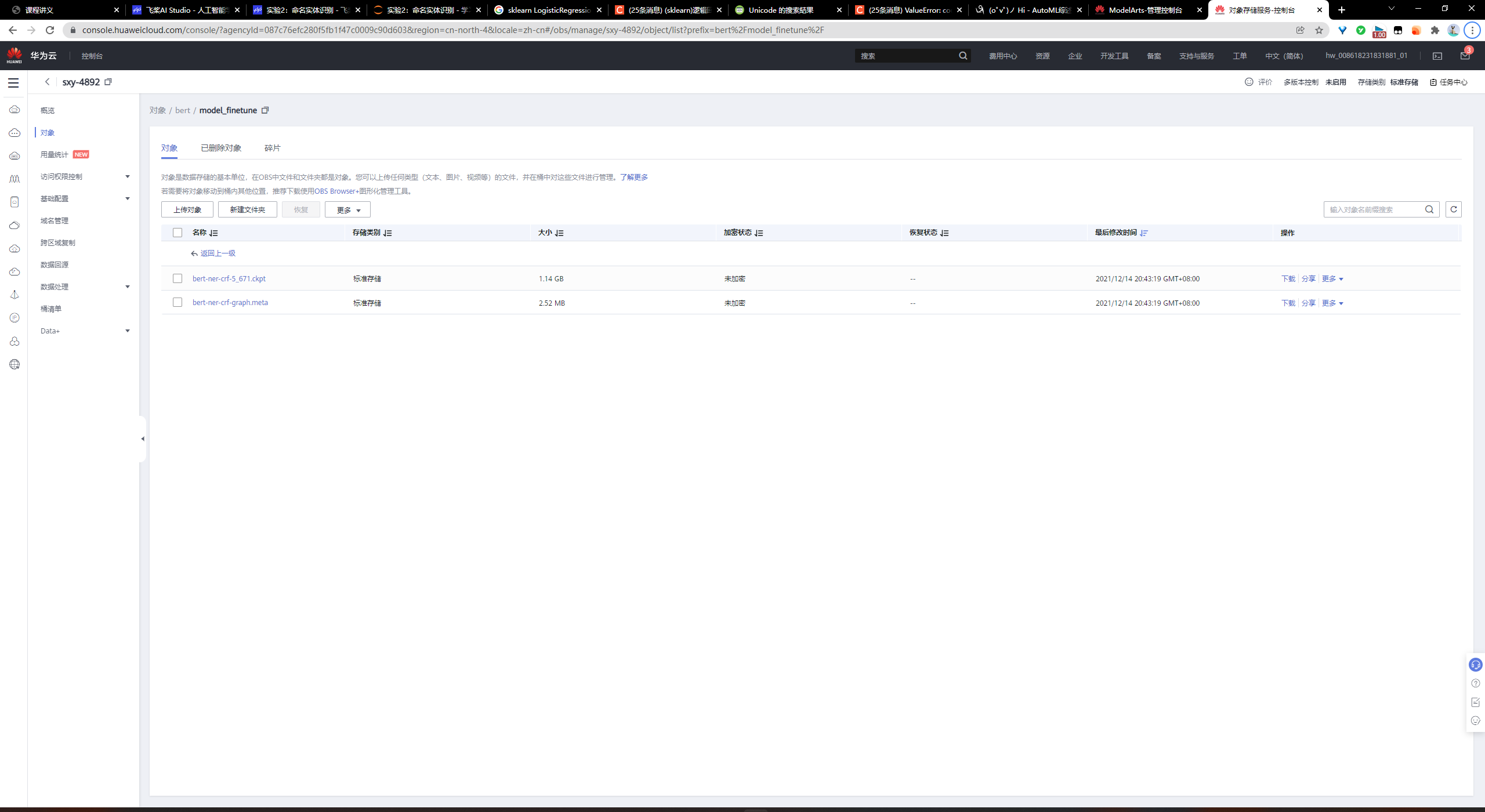
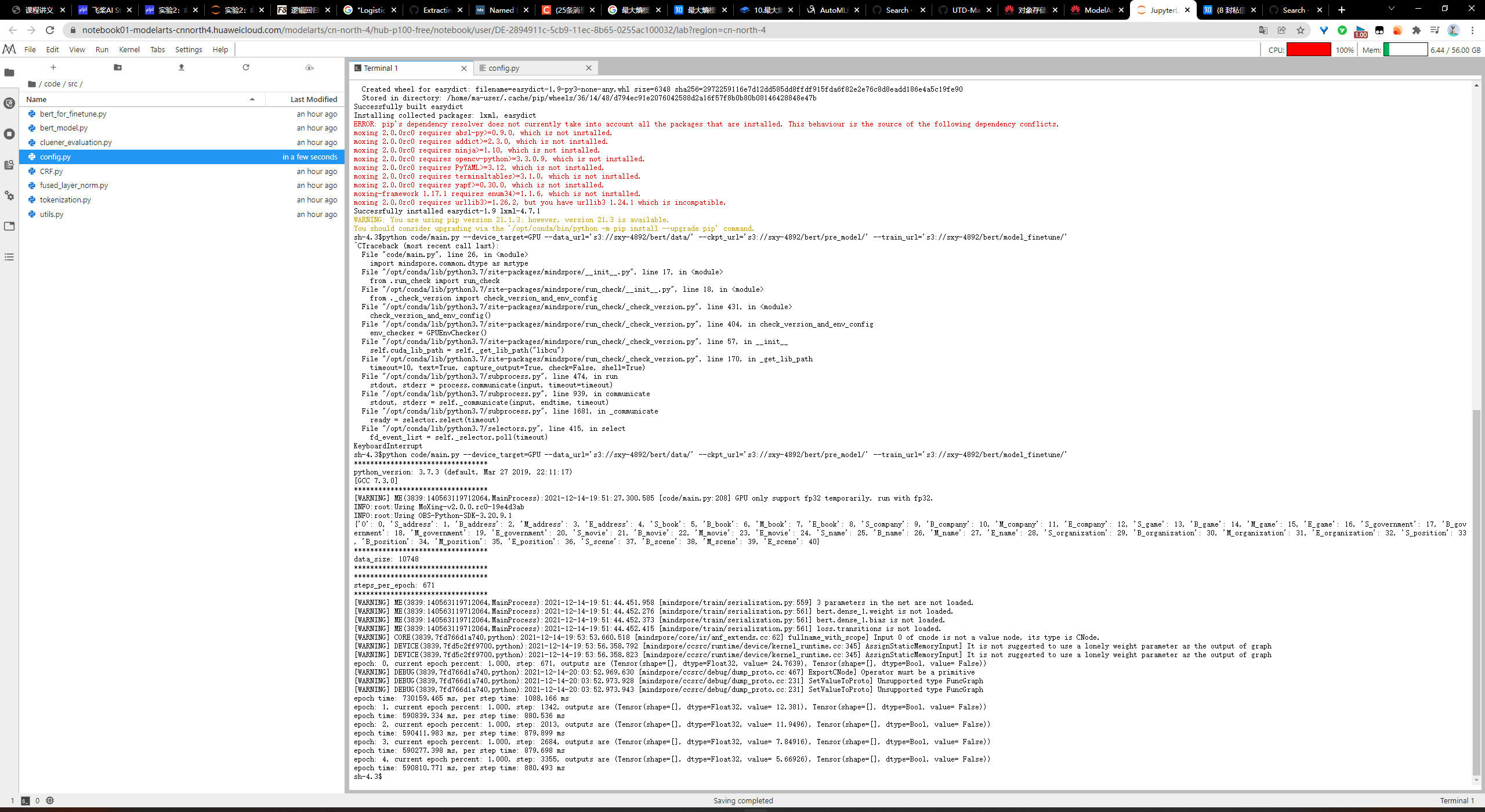
# 5. HMM、ME与CRF的效果对比

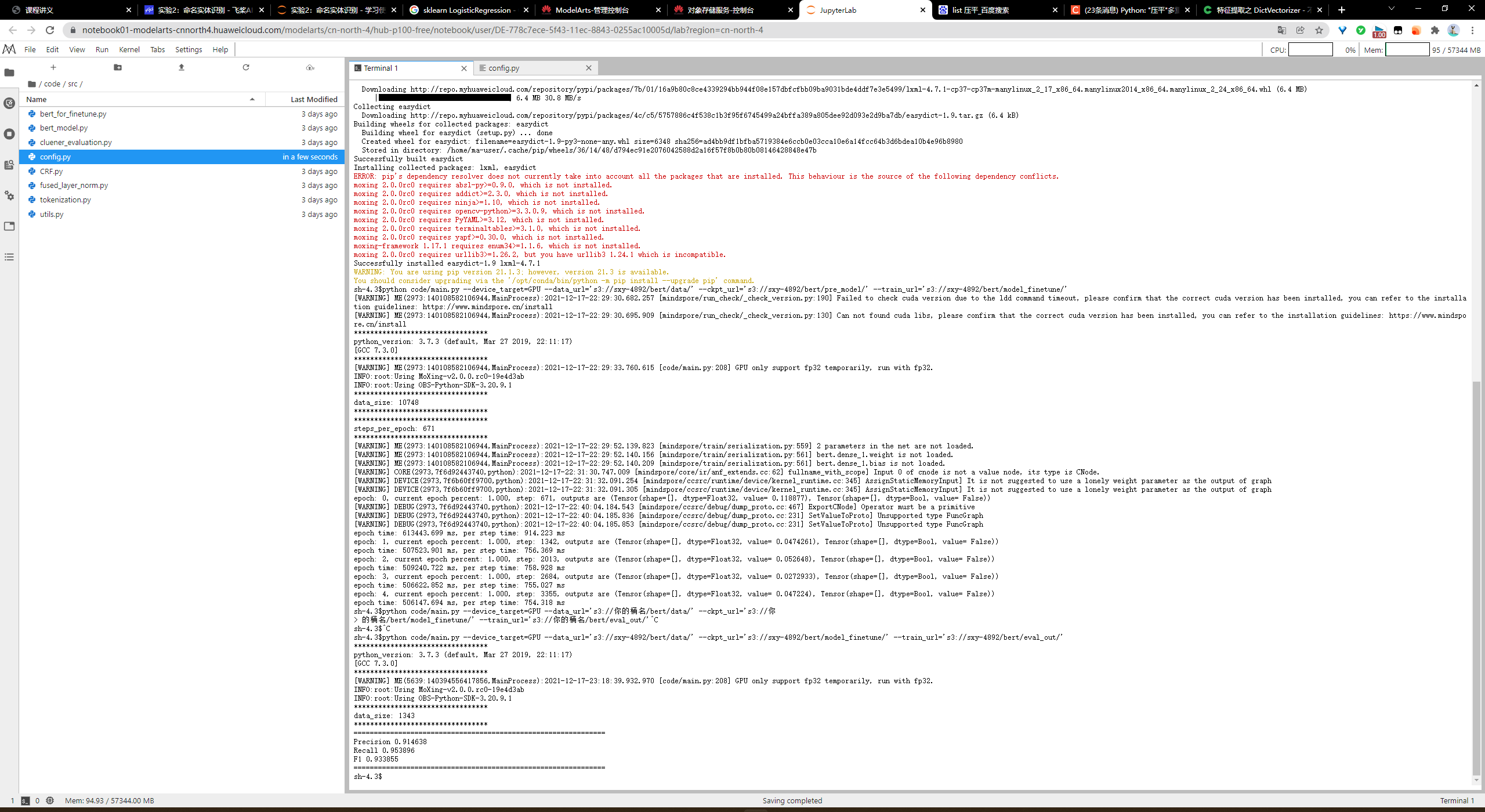
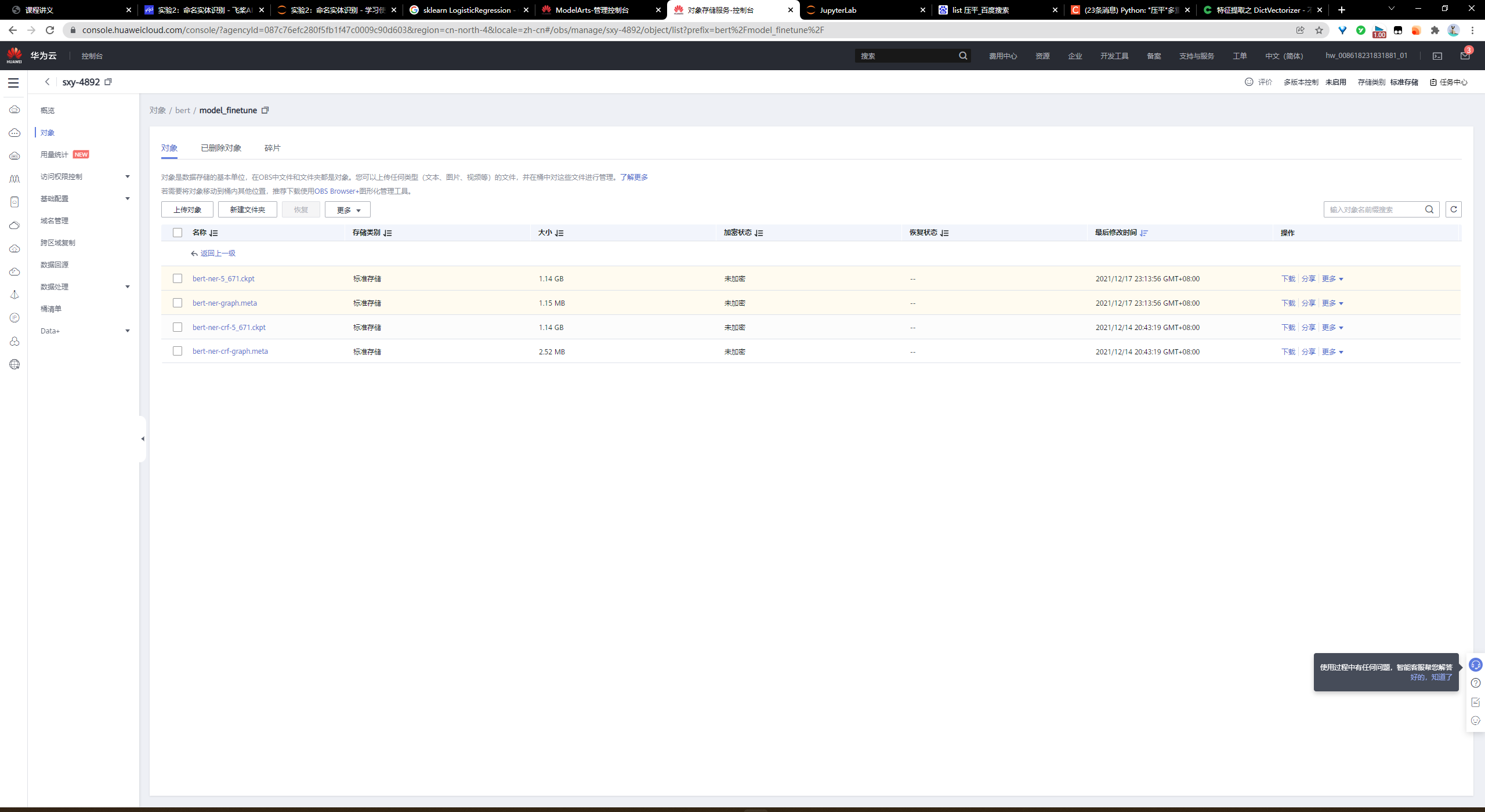
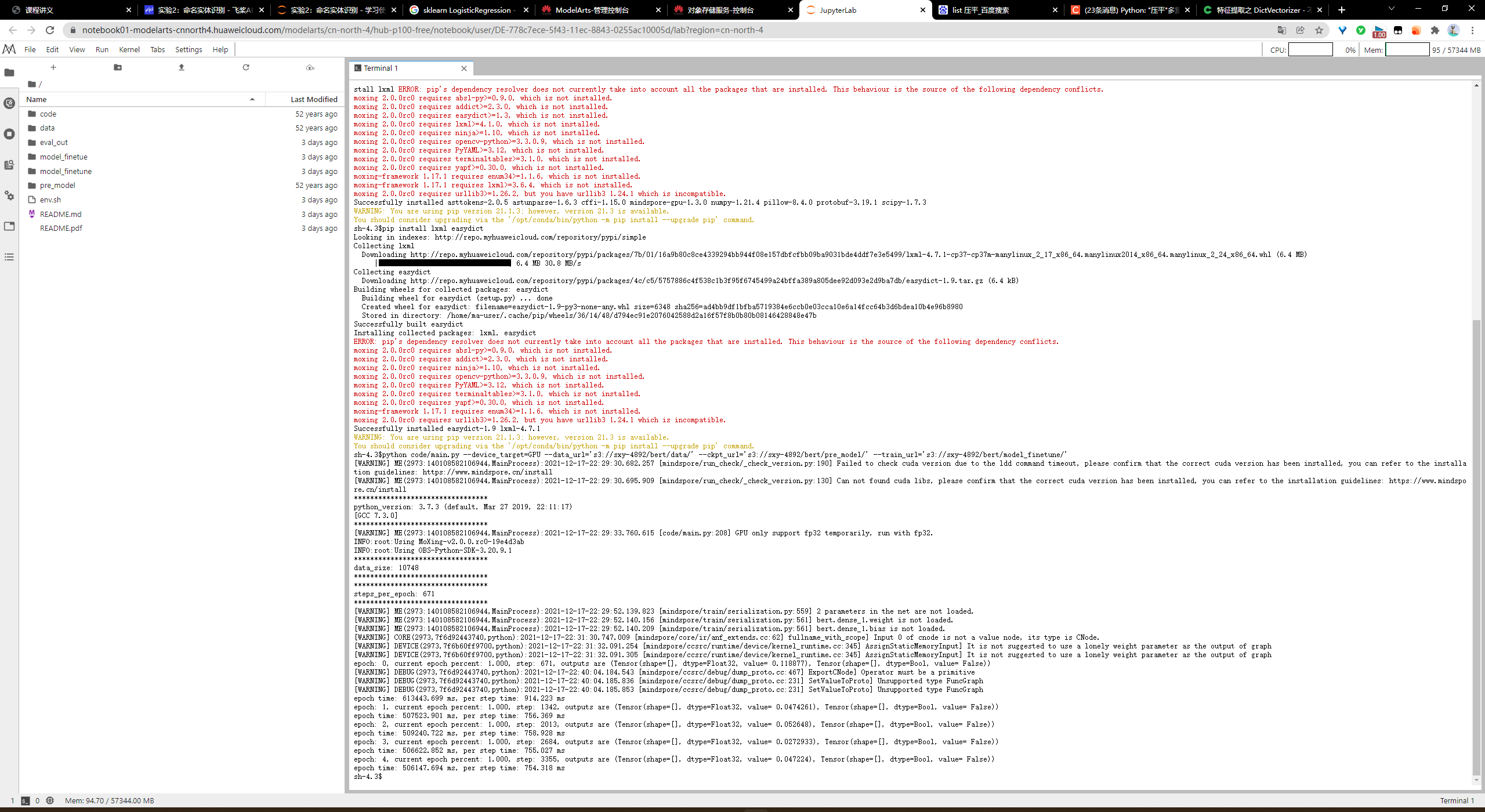
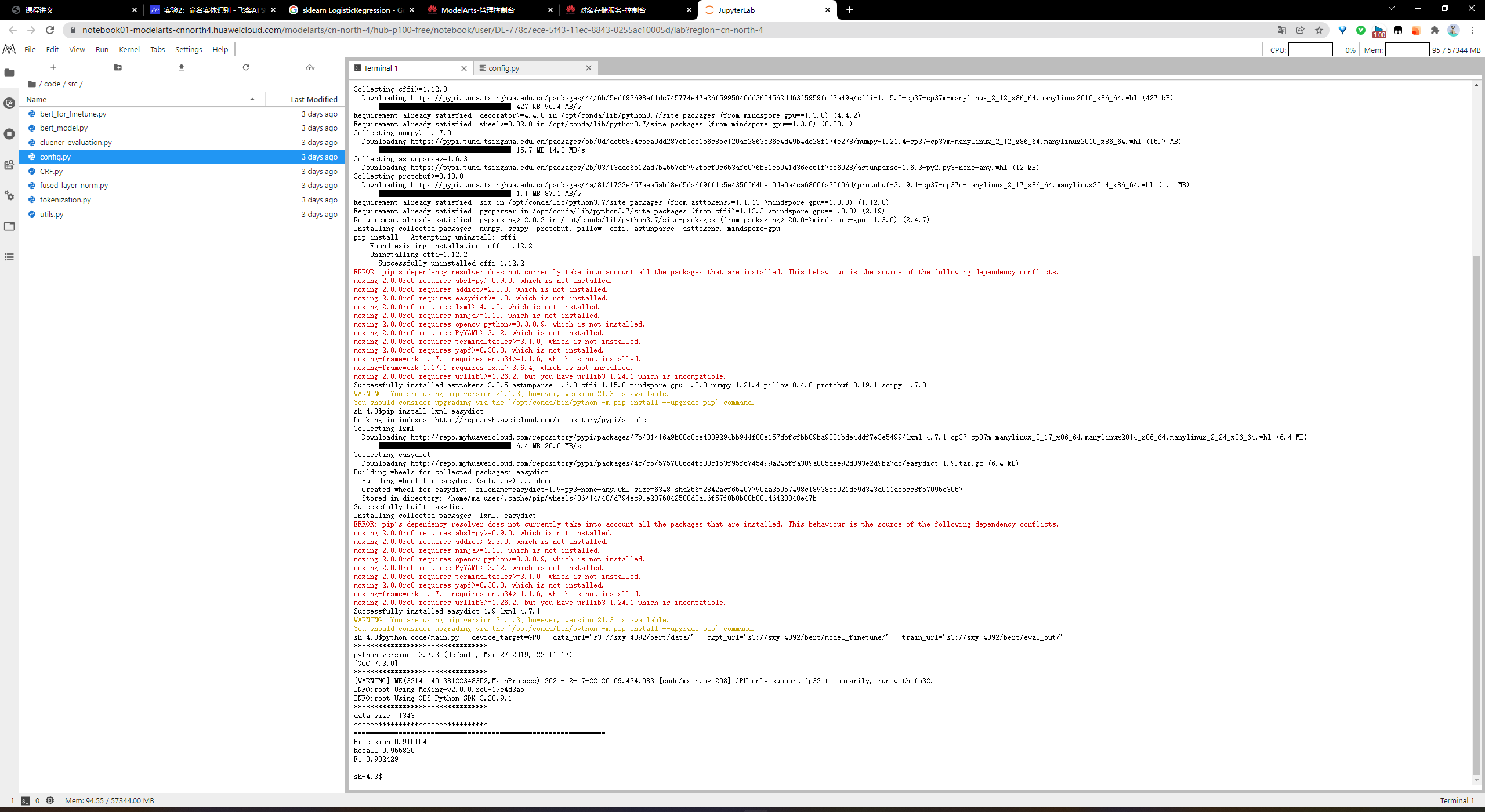
请在这部分给出：

1）说明如何利用给定的ner\_clue\_data目录下的训练数据文件train.txt分别训练HMM、ME和CRF模型并使用dev.txt文件里的数据来测试三个模型。

2）利用表格列出每个模型对应的每种实体以及总体的Precision、Recall和F1值。分析3种模型的结果差异，并给出可能的原因。

# 6. 在华为云上的计算资源进行命名实体识别

请在这部分给出：

1）至少2张包含自己华为云账号信息的运行过程及结果截图，证明自己完成了利用华为云上的计算资源进行命名实体识别的任务。

# 7. 实验的收获和体会

请在这部分给出本次实验的收获和体会。

# 参考文献

在这里列出完成实验用到的参考文献，包括论文、图书、参考的github项目；参考的技术博客。

报告其他要求：

1）字数：不少于3000字。

2）文字通顺、内容完整、逻辑性强。

3）报告中的数据和结果要与提交的代码的运行结果一致。

4）严禁抄袭。

5）提交截止时间：2021年12月29日 晚上10点。提交方式：在百度AI Studio平台，作为作业的附件一起提交。