**远程大数据结课报告**

陈宣吉

2019/8/18

**课程总结**

远程大数据课程以为期1个月，每周3个小时的网络课程形式开展。导师张帆围绕数据调用和数据库存储技术对学生进行普及和指导。因为课程时间短，课程之间有一周间隔，课程导师安排了大量课余任务和资源让学生自主学习当周的内容以及预习下一周的主体概念。同时，课程对学生的系统编程能力提出了要求，在课程结束后两个星期以内通过GitHub上传一份自主完成的MNIST项目：通过简易神经网络训练一个高精度的0至9的数字图像识别模型然后通过容器技术将它部署以来实现简洁地传输以及方便地调运于非本地环境。在此之后，它需要在通过curl xpost命令对模型输入手写体并且输出对手写体的数字判断。同时，输出结果需要以表格形式同时间信息和手写体图片信息存储于数据库内。这个项目的完成需要学生掌握各个网络传输工具之间的数据调用，对编程语言和操作环境的熟练和应用，对于未有跨平台程序开发基础的学生具有挑战性。

课程于2019年7月9日开课，第一课的内容主要以巩固和学习课程基础：python和Linux操作系统。课程上重点复习了python中的数据格式，老师对学生对于list, tuple, dictionary的数据格式区别和共性提出了问题：输入{True: “up”, 1： “down”, 1.0：“upanddown”}，为什么在输出输入信息是变成了{True: “upanddown”}? 这源于dictionary的键是唯一的，而True和1以及1.0在涵义上被定义为同等，所以后面的值被第一个键所继承。数据格式的运用在MNIST项目里面十分重要，在训练模型中的softmax算法的代码实现基于python中的灵活性：程序中占位符的宽度或者长度能够以输入数据的格式决定。同时，在课上老师基于GitHub的功能和代码上传做出了演示，作为在课程结束后MNIST的上传做出了要求。

随后的一星期里张导师对于虚拟机技术和网络调用指令的基本技术和应用进行了指导。容器技术相对于虚拟机更加方便在不同开发环境或者使用环境里传输以及兼容。相比虚拟机技术需要运行整个操作系统和其他无关应用软件的笨重，虚拟机技术更适用于开发程序部署与多环境调用。随后，老师就URL调用指令CURL做了解释和演示。这些技术都指向最终MNIST程序的编写。

张老师在第三节课对REST API (Representational State Transfer Application Programming Interface), 数据库格式和Cassandra数据库的概念解释和应用展示。在主要凸显了NOSQL对比SQL的优劣势后，理解了相对于安全和严谨的改写，对于普通程运用而言，NOSQL的行业不普遍性和它的极端快速特点产生了鲜明对比。这节课内同时也介绍了Hadoop MapReduce和spark的效率对比。课后老师对于学生提供了Cassandra Query Language的指导网页。在最后一节课，对于行业内部和MNIST项目所用函数调用的主流工具和程序老师汇总地评价对于它们的优缺点。

作为一个网络课程，这次的学习经历具有深远意义。它提供了了解和运用神经网络及监督学习算法的机会还有调用以及处理大量数据的途径；关键的是，能够让开发端和测试端通过远程调用的方式连接起来使用数据带来的模型。