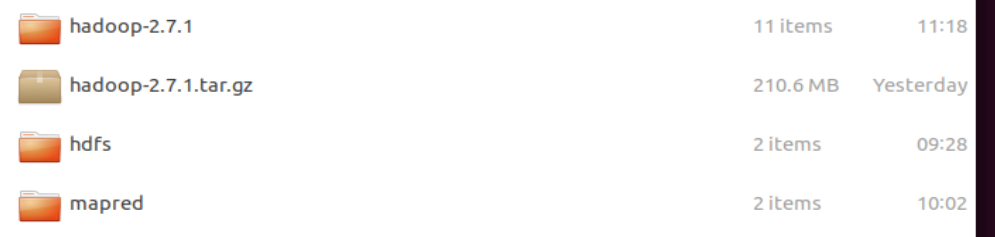
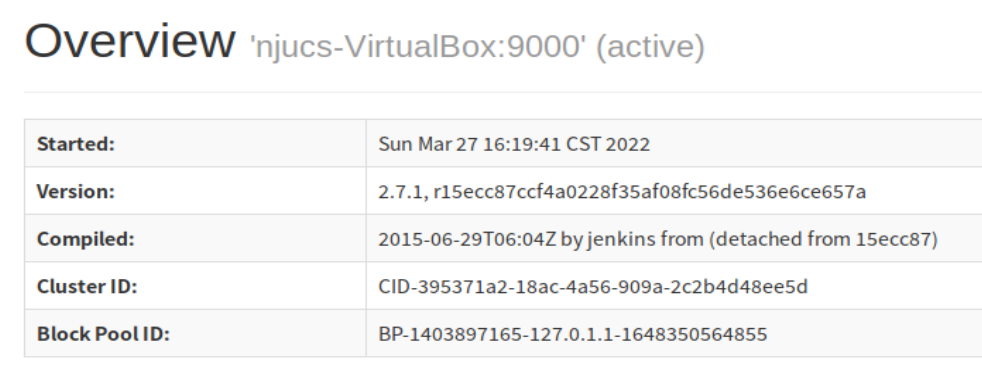
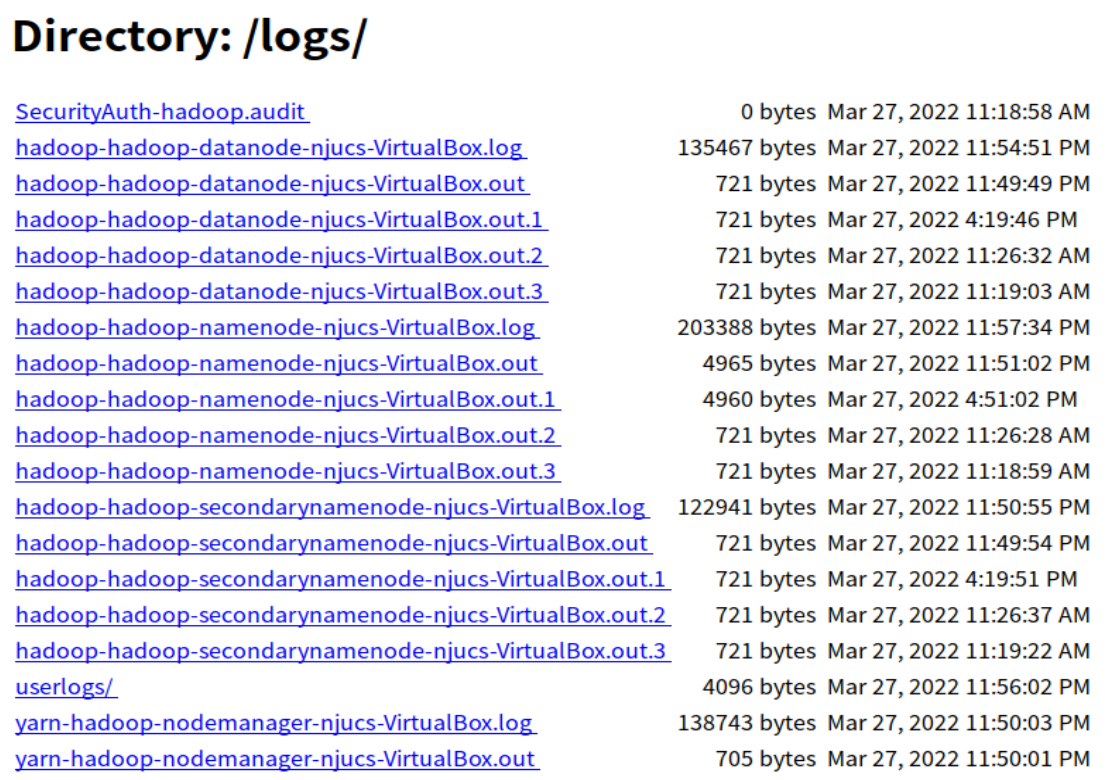
实验1 191220154 张涵之

1. 系统安装运行情况
   1. 虚拟机为计算机网络实验的njucs-VirtualBox，版本为Ubuntu 18.04.3
   2. 安装的Java版本为1.7.0\_79（之前听说版本太高会不支持，再加上开始没仔细看课件，不小心下了个1.6.x，装Hadoop的时候不兼容报错，又老老实实卸了重装。另外发现每次解压以后source /etc/profile在同一个终端里java -version显示正常，开个新终端又command not found，于是试了一般解压、直接执行.bin、rpm，算是对软件的不同安装方式有了全新的了解，最后发现不管哪种只要重启一下虚拟机就好了）。
   3. IntelliJ：idea-IC-213.7172.25
   4. Maven：apache-maven-3.8.5
   5. Java、IntelliJ、Maven都安装在/usr目录下
   6. Hadoop：hadoop-2.7.1，安装在/home/hadoop/hadoop\_installs下



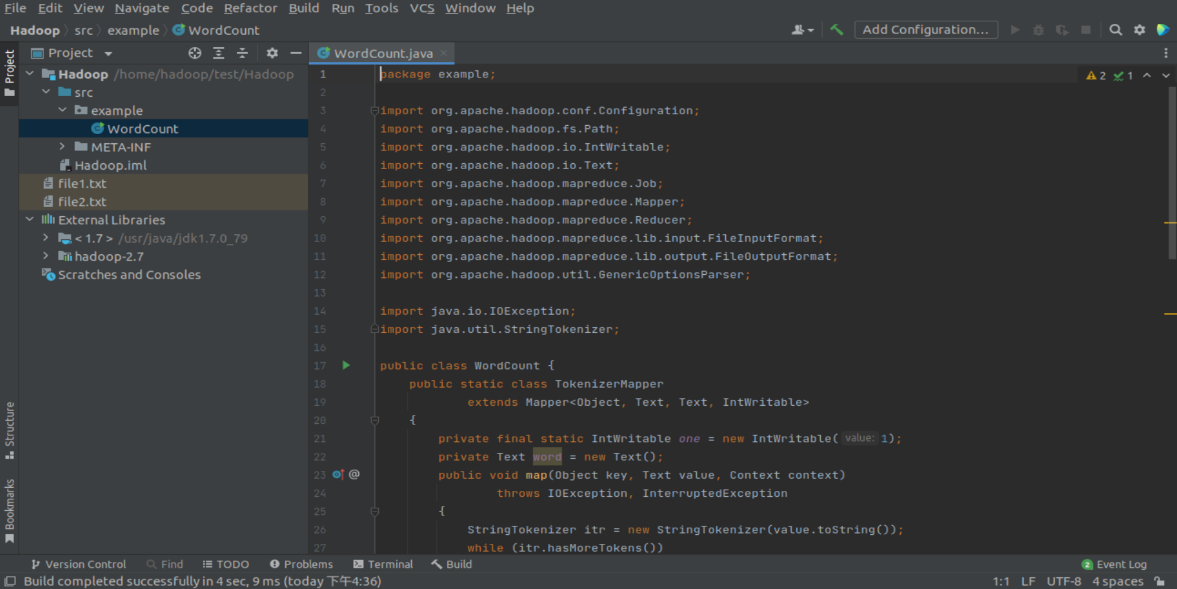






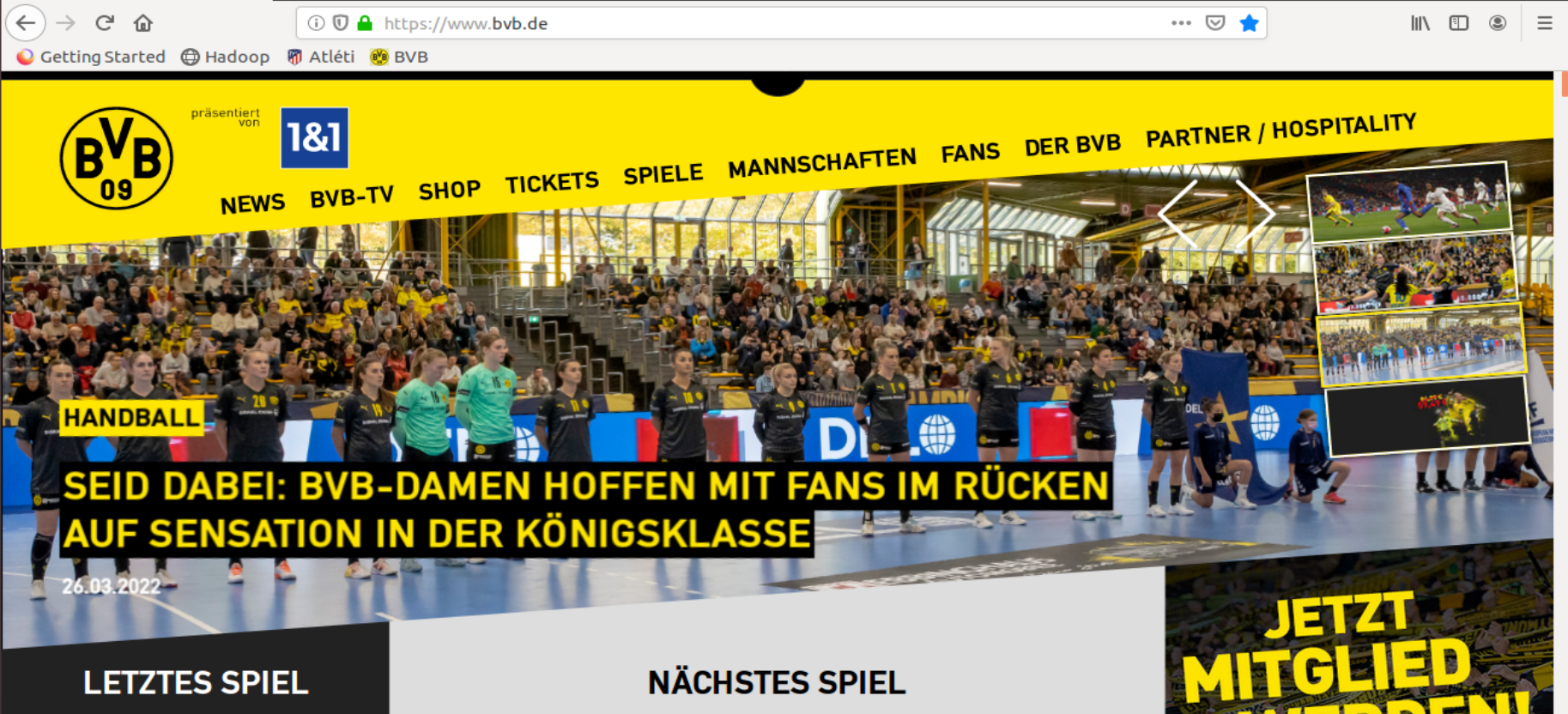
遇到的问题：

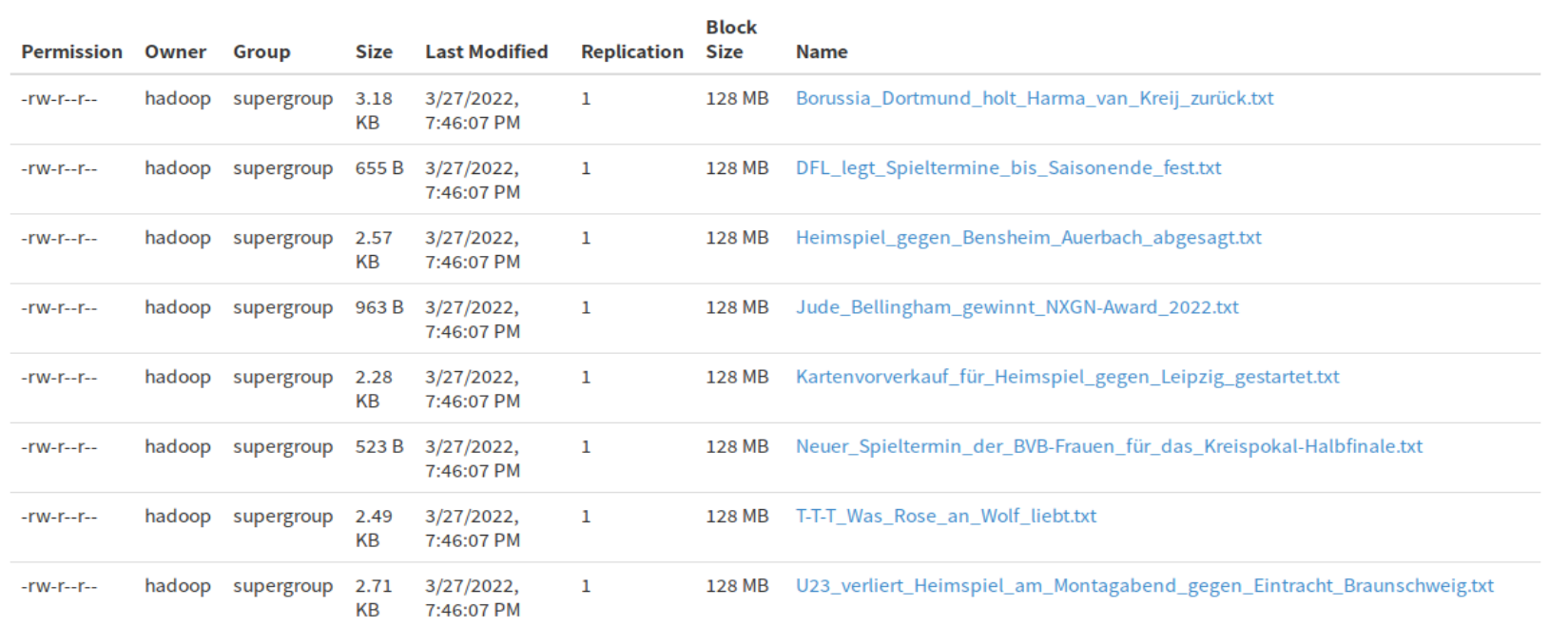
1. 开始装的时候，对绝对和相对路径的理解一塌糊涂，时而加“/”时而不加，还经常纳闷为什么找不到文件，像一个绝望的电脑盲。经过一系列操作，终于明白了平时打开终端“:”和“$”中间的“~”是啥。
2. 又听说装机配环境的时候最好经常备份，装一个软件来一个快照，于是最后弄出了十几个备份，跟玩游戏似的，一言不合就读档。
3. 刚开始装还做一步截一个图，做一步截一个图，再后来发现老是做错改来改去又读档，就完全懒得截图了，只想赶紧装完。
4. 复制报错信息到搜索引擎，得到的结果非常良莠不齐。
5. 不知道要导入哪些外部依赖jar包才能正常编译，觉得很麻烦。
6. 生成jar包的时候可以指定一个主类MainClass作为默认执行的类，然后使用hadoop jar xxx.jar wordcount /input /output运行的时候参数报错，去掉wordcount可以正常运行，猜测Hadoop包提供的example jar含有很多类且没有指定主类，所以运行的时候才需要加上wordcount。



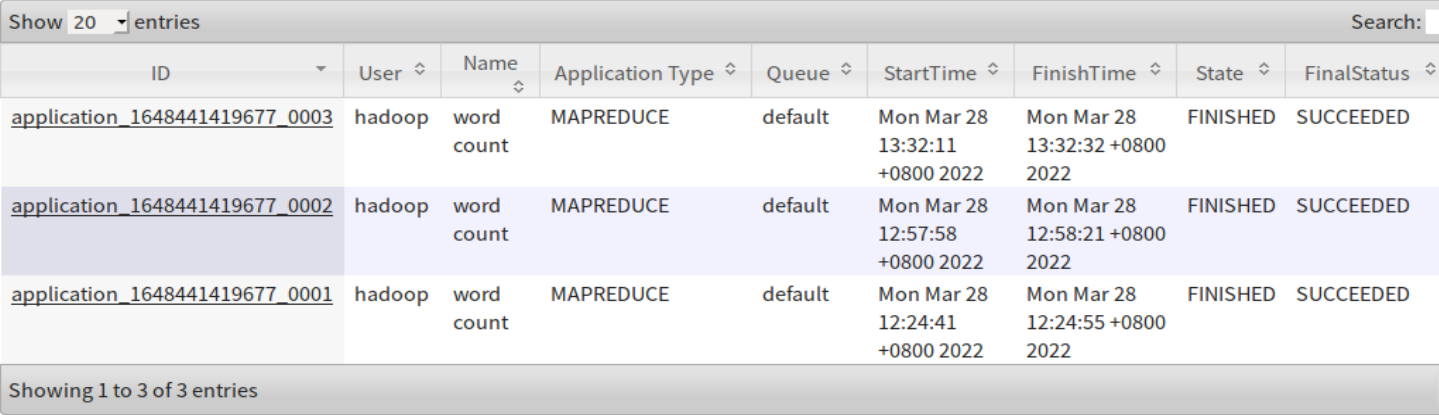
1. 实验数据说明

从德甲足球队多特蒙德的官网<https://www.bvb.de/>选取最近的十余条新闻，试图直接下载HTML文件，但发现那样会混杂很多与新闻正文无关的文本如页面顶部菜单、侧边栏、底部的其他链接等。所以最后还是采取了原始的复制文章粘贴到纯文本（.txt）文件里，最后得到11个.txt文件。



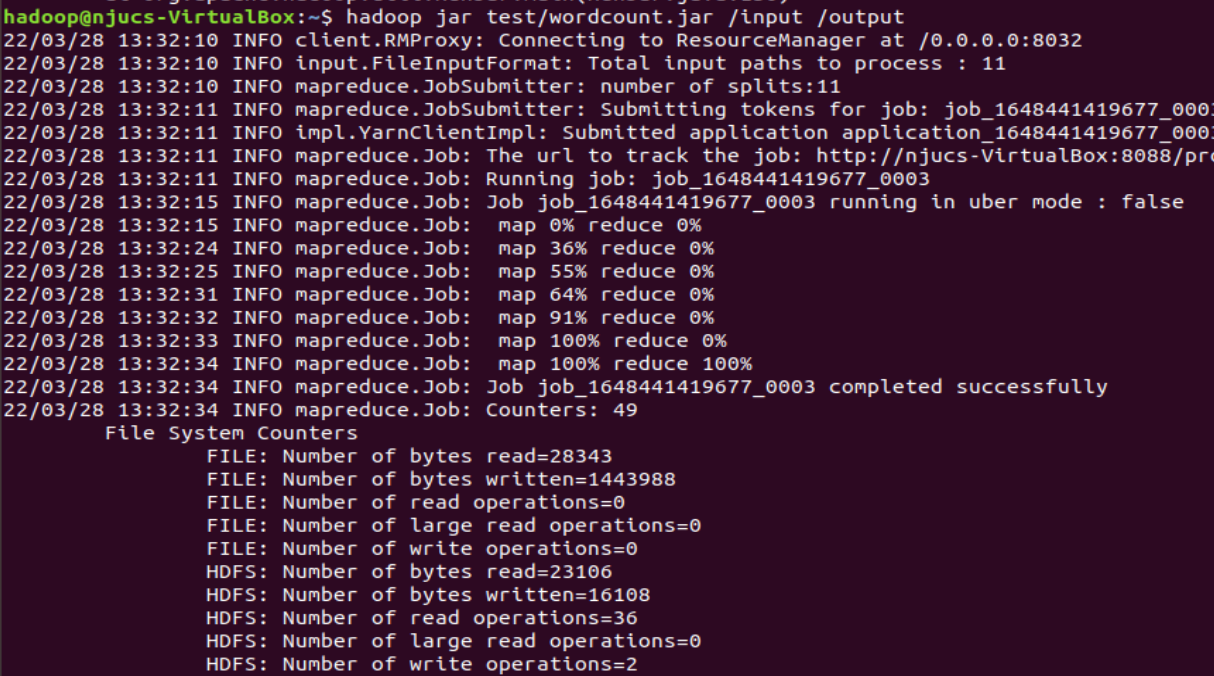


1. 程序运行后在Hadoop Web作业状态查看界面上的作业运行状态屏幕拷贝

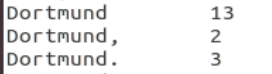


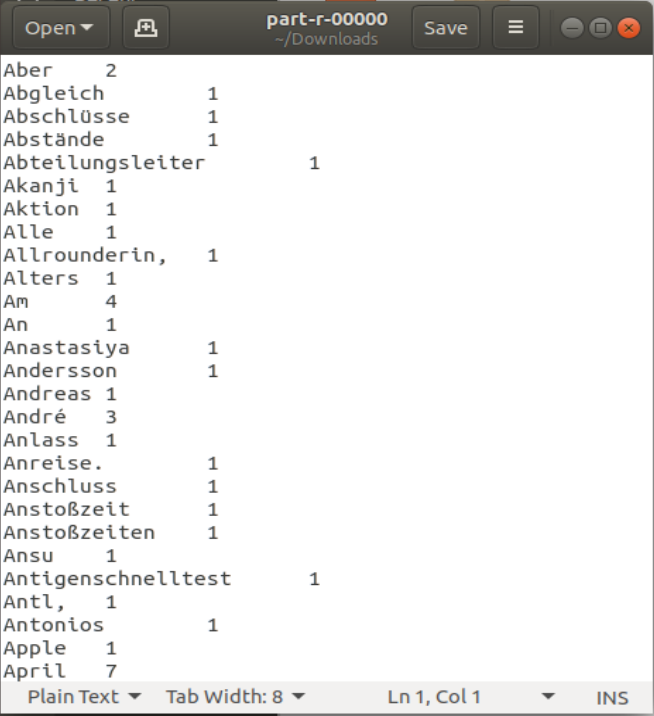
三个作业分别是安装过程中测试时用Hadoop自带的example jar包统计少量简单的文件（课件中的file1.txt和file2.txt），用Hadoop自带example统计德语文章，和用自己编写的WordCount统计德语文章，后面两个结果相同。

1. 实验输出结果开头部分的屏幕拷贝



注意到这个word count程序不是很智能，仅以空格作为单词之间切分的标准而没有考虑标点符号和大小写，所以前后带有标点的同一个单词会分开统计多次，如“Dortmund”、“Dortmund,”、“Dortmund.”和“(Dortmund”。





1. 实验体会
   1. 课件和网络教程里的命令和配置文件别急着复制，看清楚干啥的，免得有些文件路径和IP地址也无脑复制了，后续白天找不出原因。
   2. 配环境的时候，多搞几个快照存档非常有用，之前配置文件没弄好只在本地跑网站上找不到作业记录，改了一会DataNode消失了，又改了一会索性连NodeManager也消失了，赶紧读档从头再来。