一、实验目的

- 掌握熟悉 Linux 操作系统基础操作，掌握在 Linux 下安装 Java 的技能
- 掌握 Hadoop 分布式环境配置方法，理解 MapReduce 作业的原理和操作方法
- 掌握 Hadoop 集群的启动与关闭操作，并在 Hadoop 上执行词频统计程序

二、实验环境

- CentOS 6.10
- Hadoop 2.6.4
- Java 1.8

三、实验内容

- 在 Linux 下安装 Java，修改 Hadoop 配置文件，独立搭建 Hadoop 分布式环境
- 上传预处理后的实验文本数据至 HDFS
- 启动关闭 Hadoop 集群
- 使用 MapReduce 实现 WordCount 程序任务

四、实验过程

1、配置集群机器，下载安装包。

如图 1 所示，在终端输入 initnetwork 命令初始化网络。

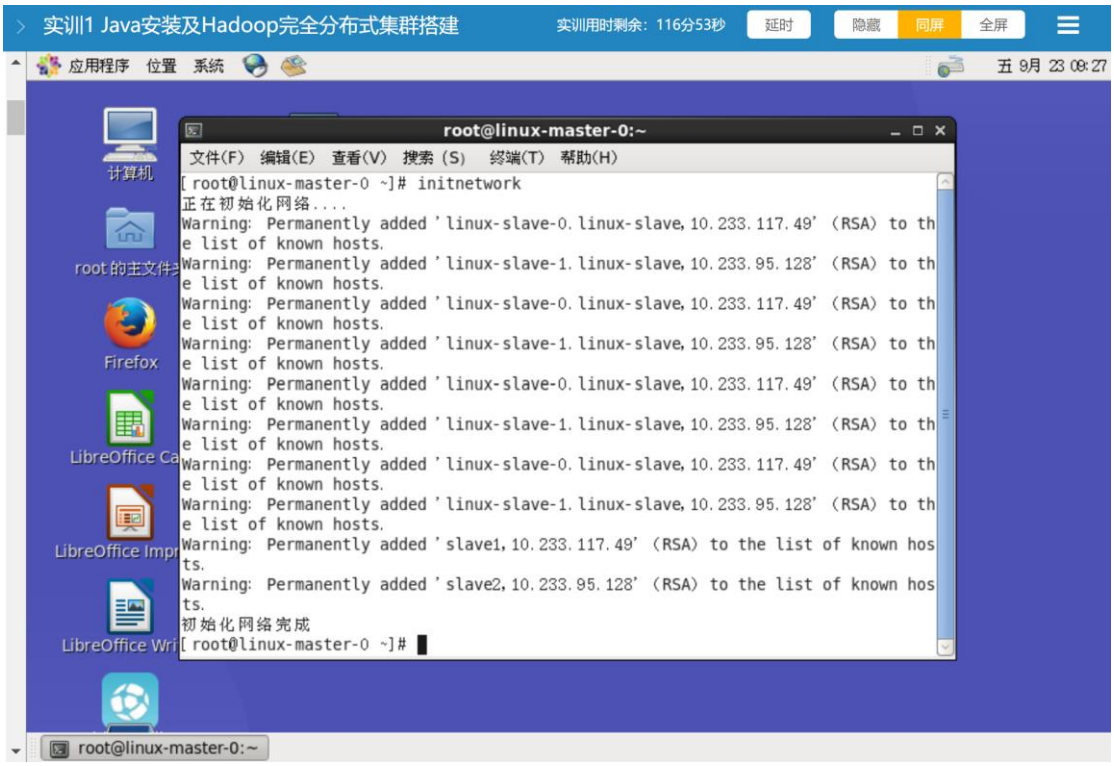

图 1

如图 2 所示，在 Linux 终端执行命令 wget -P /opt/software http://datasrc.tipdm.net:81/bigdata/hadoop/software/hadoop-2.6.4.tar.gz 和命令 wget -P /opt/software http://datasrc.tipdm.net:81/bigdata/hadoop/software/jdk-8u151-linux-x64.rpm，从而分别将 jdk-8u151-linux-x64.rpm 和 hadoop-2.6.4.tar.gz 下载到 Linux 本地/opt/software 目录。

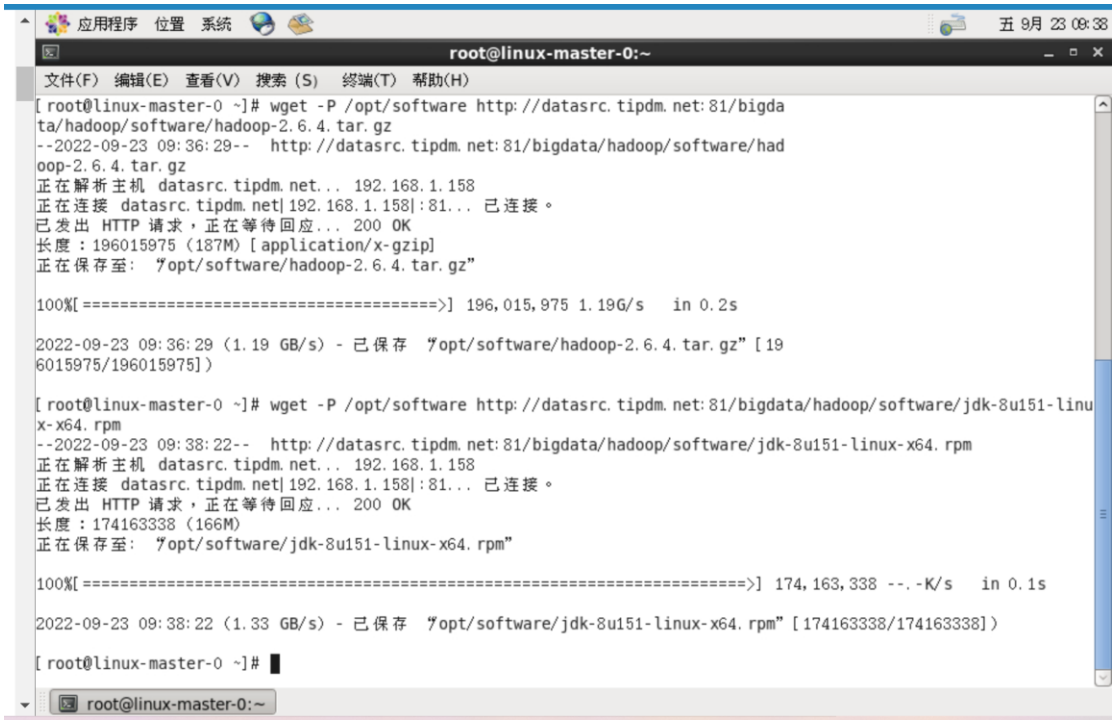

图 2

如图 3 所示，首先将 Java 安装包发送到子节点上，然后在终端输入命令 java -version。由此检测到当前环境并没有安装 Java。因此，进入/opt/software 目录安装 JDK。

图 3

通过 vi 命令修改/etc/profile 文件，在 master 节点添加 Java 的环境变量。然后，在两个子节点 slave1、slave2 上面执行相同的步骤，安装 Java 并添加环境变量，完成所要求版本 JDK 的安装，如图 4、图 5 所示。

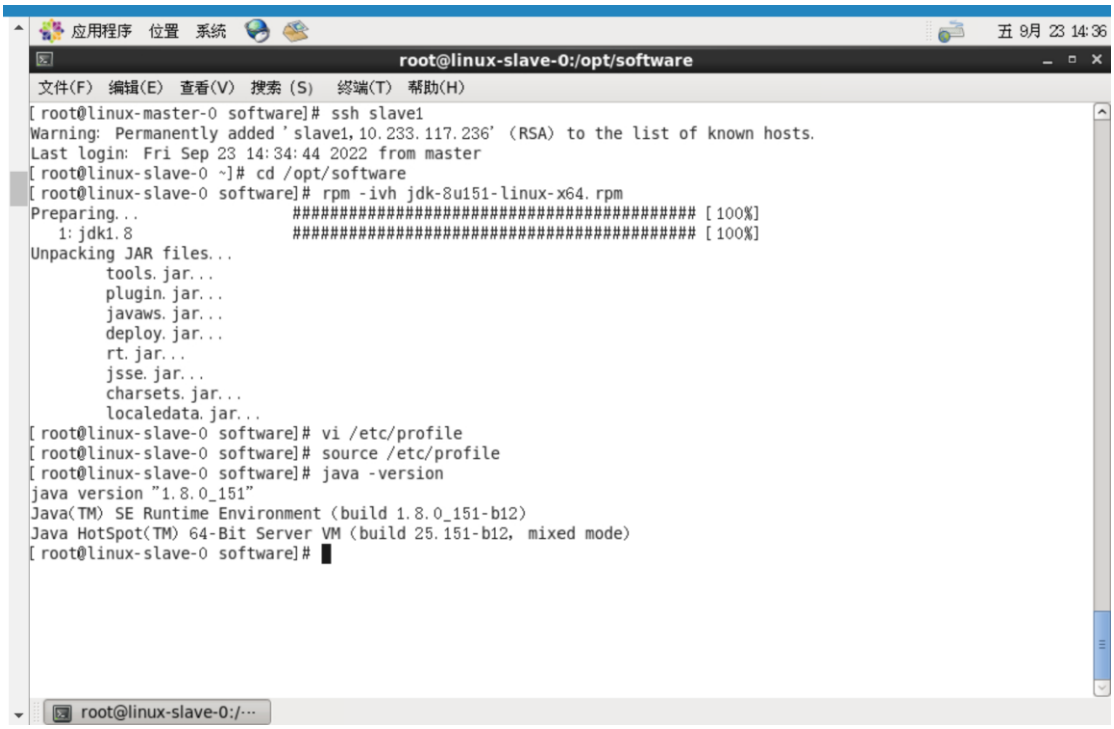

图 4

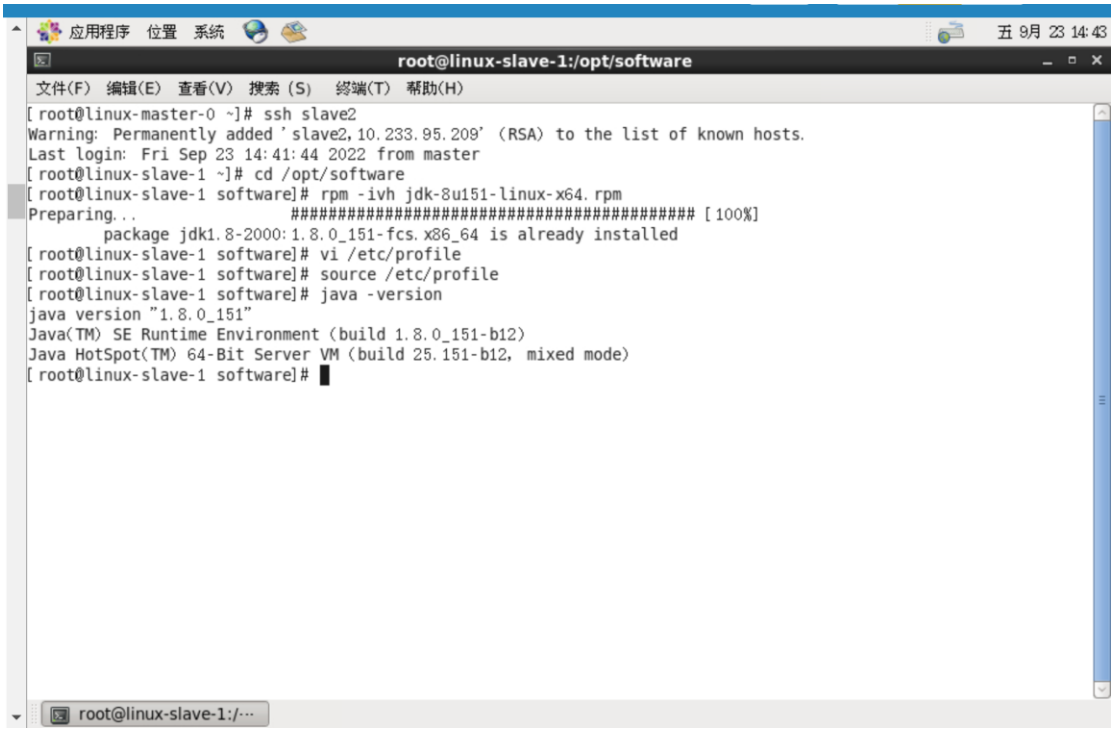

图 5

2、修改 Hadoop 配置文件。

首先执行 tar -zxf /opt/software/hadoop-2.6.4.tar.gz -C /usr/local，将 Hadoop 安装包解压至 master 的/usr/local 目录下，随后进入/usr/local/hadoop-2.6.4/etc/hadoop 目录，用 vi 命令修改相关配置文件，如图 6 所示。

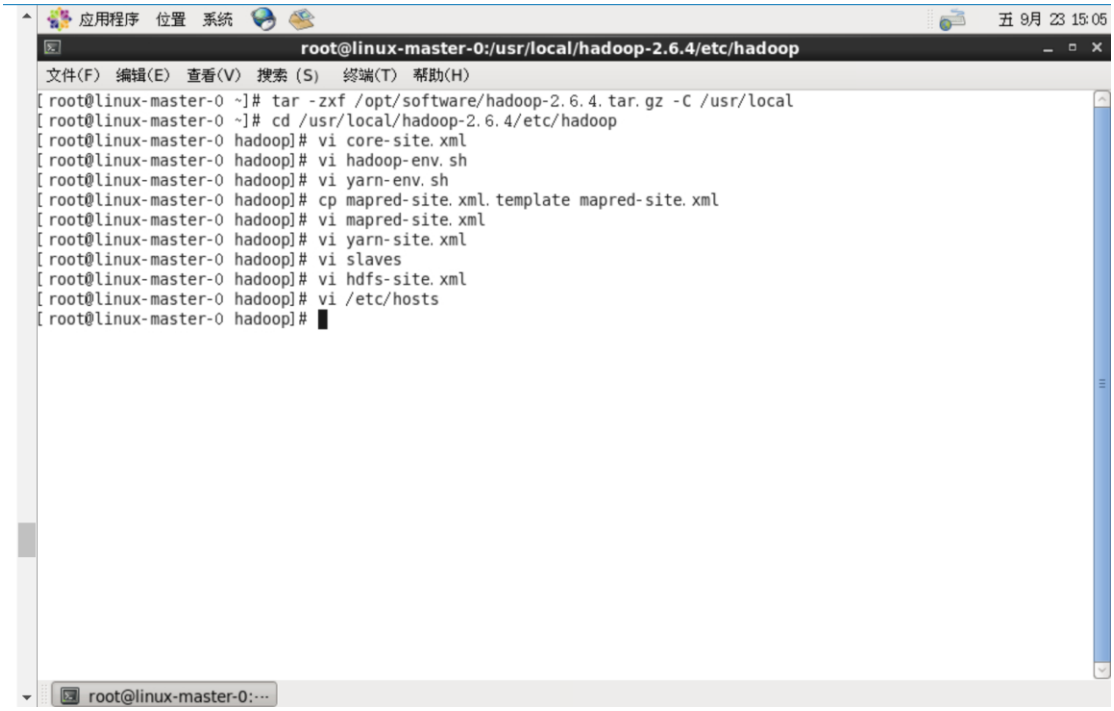

图 6

随后，在 master 上修改/etc/profile 文件，以配置 Hadoop 的环境变量，使用命令 source /etc/profile 使配置生效。将 Hadoop 安装包、/etc/profile 文件复制传输到 slave1、slave2 节点，如图 7 所示。

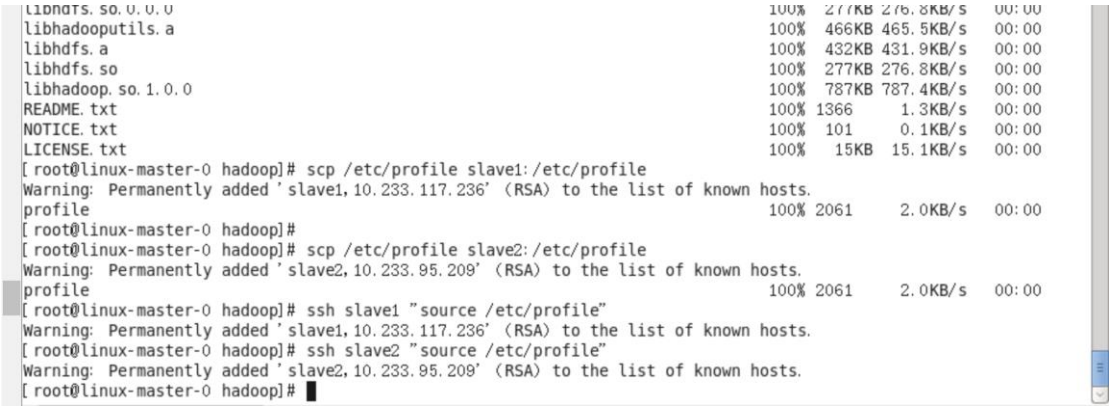

图 7

在 master 上执行命令 hdfs namenode -format 以进行格式化，观察到终端输出提示 Storage directory hadoop/hdfs/name has been successfully formatted，说明已经格式化成功。如图 8 所示。

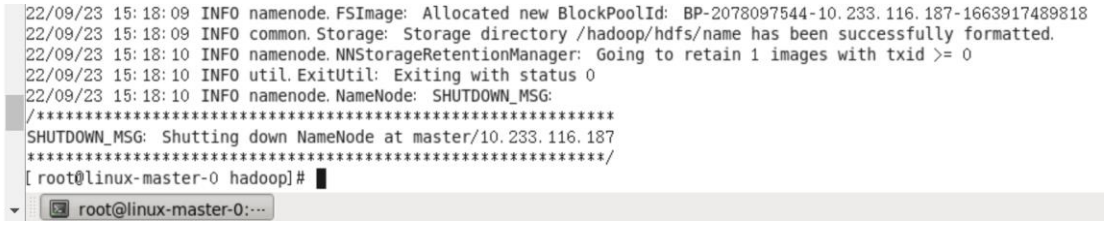

图 8

3.启动、关闭 Hadoop 集群

在 master 节点，进入 Hadoop 安装目录，启动 HDFS 相关服务、YARN 相关服务和日志相关服务，如图 9 所示。

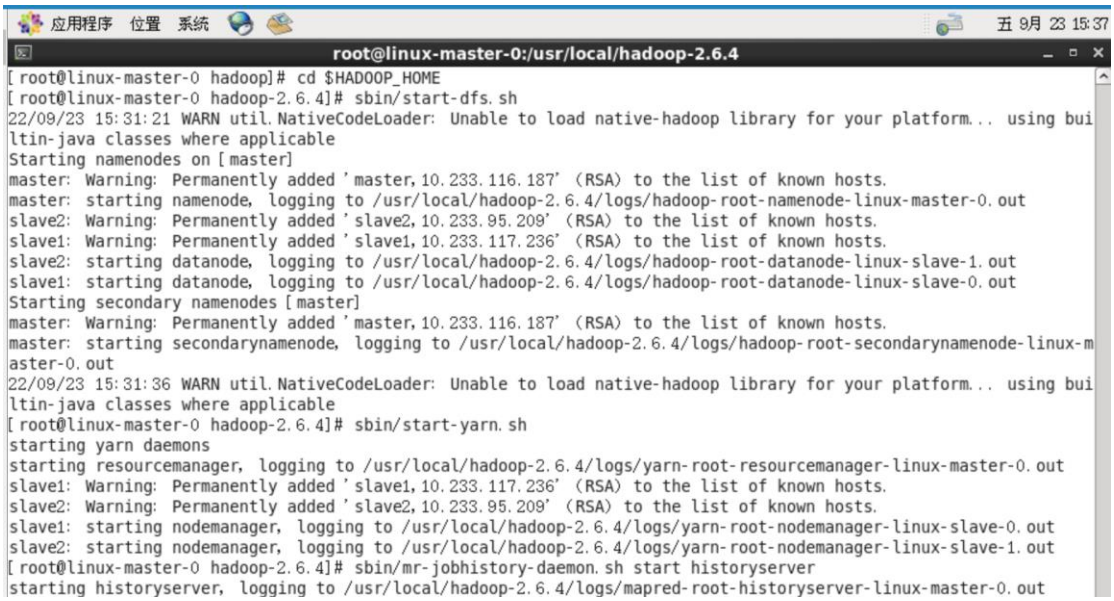

图 9

集群启动之后，在主节点 master，子节点 slave1、slave2 分别执行 jps，结果如图 10 所示。这说明集群的启动没有问题。

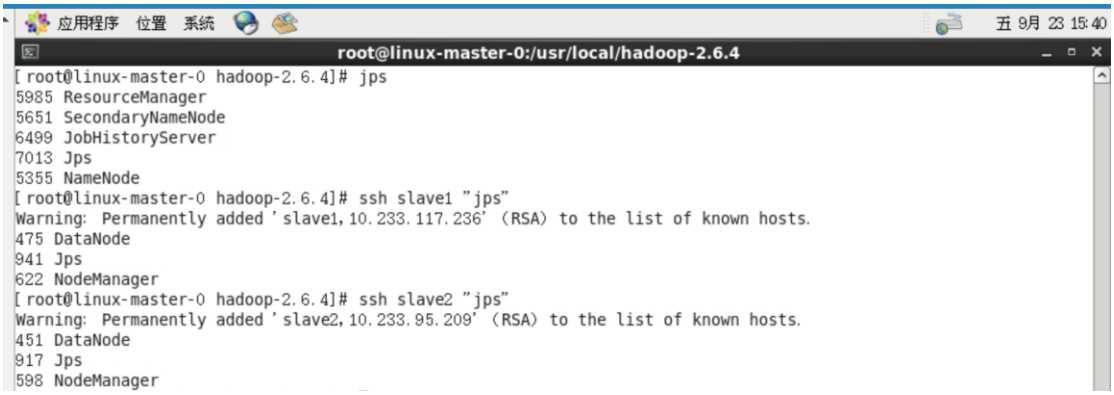

图 10

对 Hadoop 集群进行关闭，如图 11 所示。

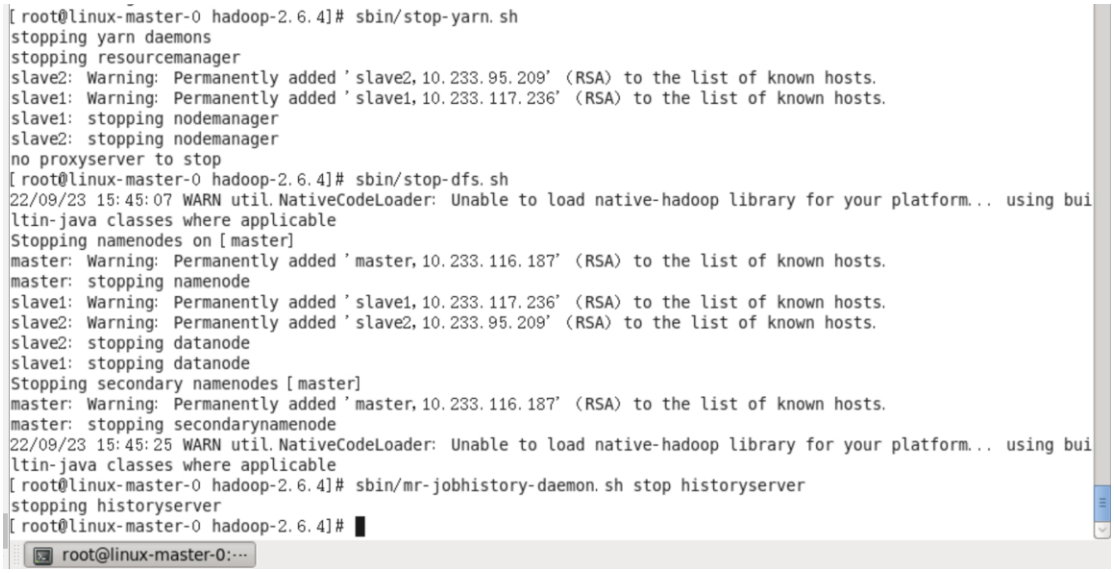

图 11

3、使用 MapReduce 完成 WordCount 程序任务。

如图 12 所示，进入“实训 3 运行首个 MapReduce 任务”实训环境，首先将预处理后得到的文本上传至实训环境的/data 目录下。

图 12

如图 13、图 14 所示，打开终端，使用命令 initnetwork 初始化网络，并启动 Hadoop 集群。

```
[ root@base-master-0 ~]# initnetwork
正在初始化网络....
Warning: Permanently added 'base-slave-0.base-slave,10.233.116.216' (RSA) to the
 list of known hosts.
Warning: Permanently added 'base-slave-1.base-slave,10.233.99.177' (RSA) to the
list of known hosts.
Warning: Permanently added 'base-slave-0.base-slave,10.233.116.216' (RSA) to the
 list of known hosts.
Warning: Permanently added 'base-slave-1.base-slave,10.233.99.177' (RSA) to the
list of known hosts.
Warning: Permanently added 'base-slave-0.base-slave,10.233.116.216' (RSA) to the
 list of known hosts.
Warning: Permanently added 'base-slave-1.base-slave,10.233.99.177' (RSA) to the
list of known hosts.
Warning: Permanently added 'base-slave-0.base-slave,10.233.116.216' (RSA) to the
 list of known hosts.
Warning: Permanently added 'base-slave-1.base-slave,10.233.99.177' (RSA) to the
list of known hosts.
Warning: Permanently added 'slave1,10.233.116.216' (RSA) to the list of known ho
sts.
Warning: Permanently added 'slave2,10.233.99.177' (RSA) to the list of known hos
ts.
初始化网络完成
```

图 13

```
root@base-master-0:~
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
[ root@base-master-0 ~]# $HADOOP_HOME/sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
22/09/24 09:50:37 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Starting namenodes on [master]
master: Warning: Permanently added 'master,10.233.116.42' (RSA) to the list of k
nown hosts.
master: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenod
e-base-master-0.out
slave1: Warning: Permanently added 'slave1,10.233.89.218' (RSA) to the list of k
nown hosts.
slave2: Warning: Permanently added 'slave2,10.233.116.243' (RSA) to the list of
known hosts.
slave2: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanod
e-base-slave-1.out
slave1: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanod
e-base-slave-0.out
Starting secondary namenodes [master]
master: Warning: Permanently added 'master,10.233.116.42' (RSA) to the list of k
nown hosts.
master: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-roo
t-secondarynamenode-base-master-0.out
22/09/24 09:50:52 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-root-resourcema
nager-base-master-0.out
slave2: Warning: Permanently added 'slave2,10.233.116.243' (RSA) to the list of
known hosts.
root@base-master-0:~
```

图 14

如图 15 所示，首先在 HDFS 新建目录/user/root，然后将/data 目录下的待统计文本文件 new_result.txt 上传至 HDFS。

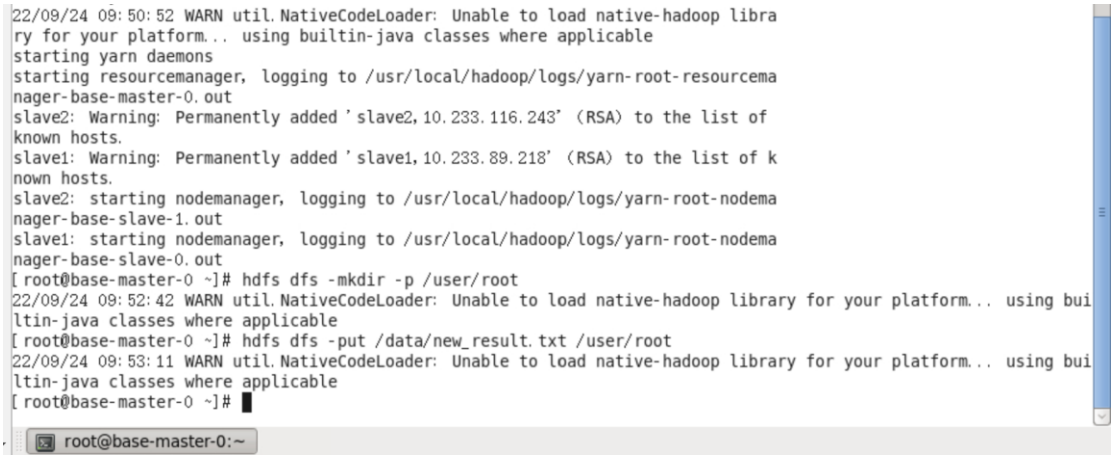

图 15

如图 16 所示，提交 MapReduce 任务，对所上传的 new_result.txt 文件执行词频统计程序。

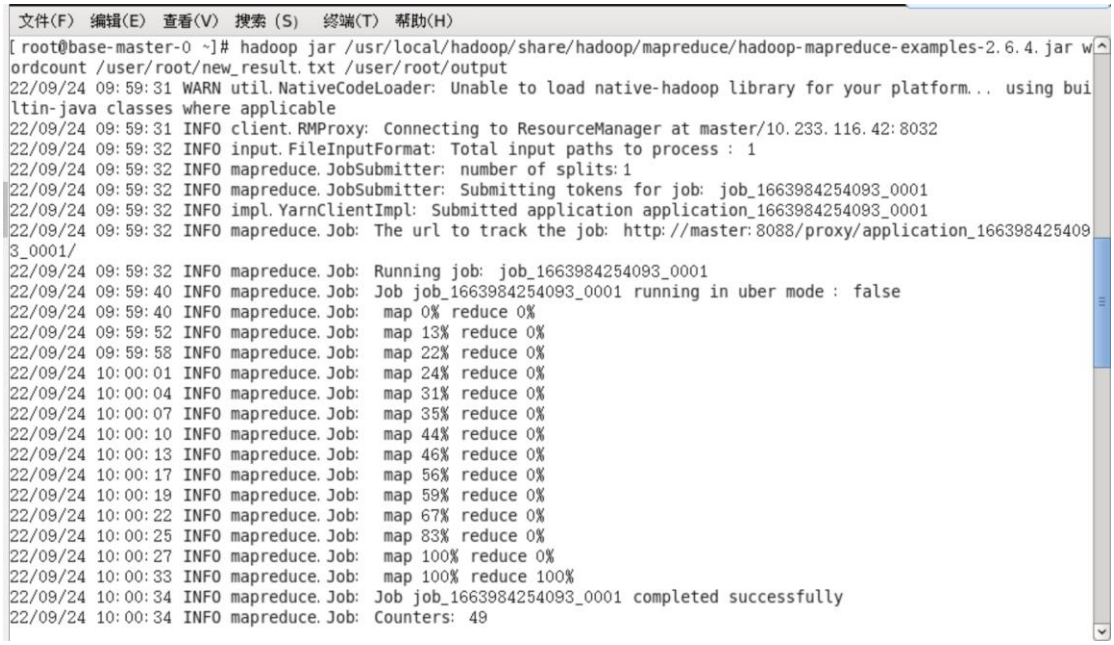

图 16

如图 17 所示，可以在/user/root/output 目录下观察到新生成的文件。其中文件_SUCCESS 表示任务执行完成，而 part-r-00000 则为任务执行完成后产生的结果文件。

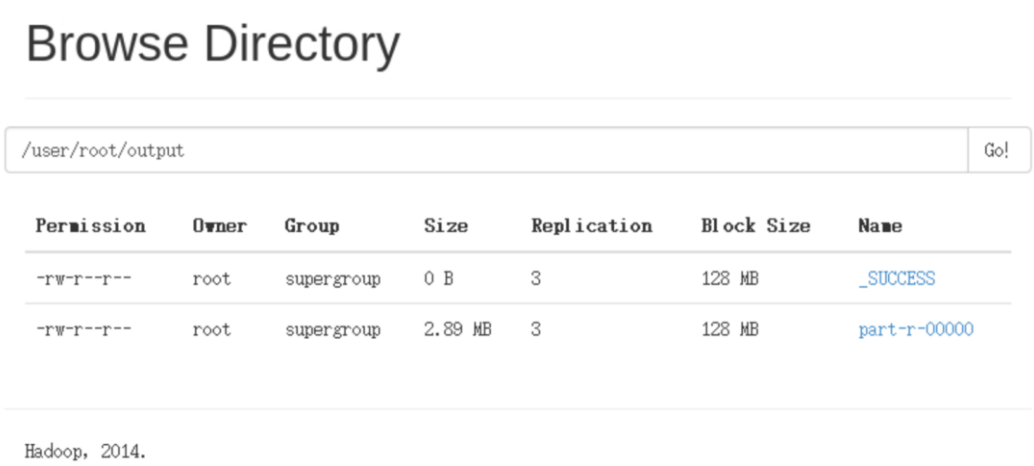

图 17

下载 part-r-00000 文件（如图 18），并更名为 statistical_result.txt，即可得到分词统计结果。将该文件从实训环境中导出，并在 Visual Studio Code 中查看，结果如图 19 所示。

图 18

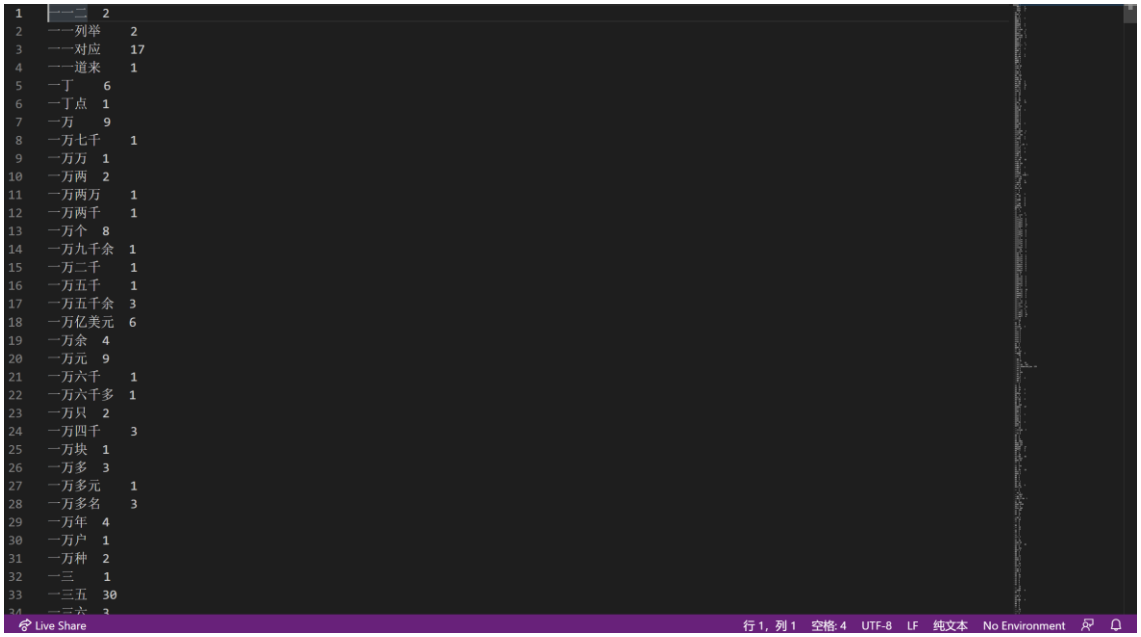

图 19

五、实验结果

WordCount 程序任务的完整运行命令、执行结果如下图 20、21、22 所示。

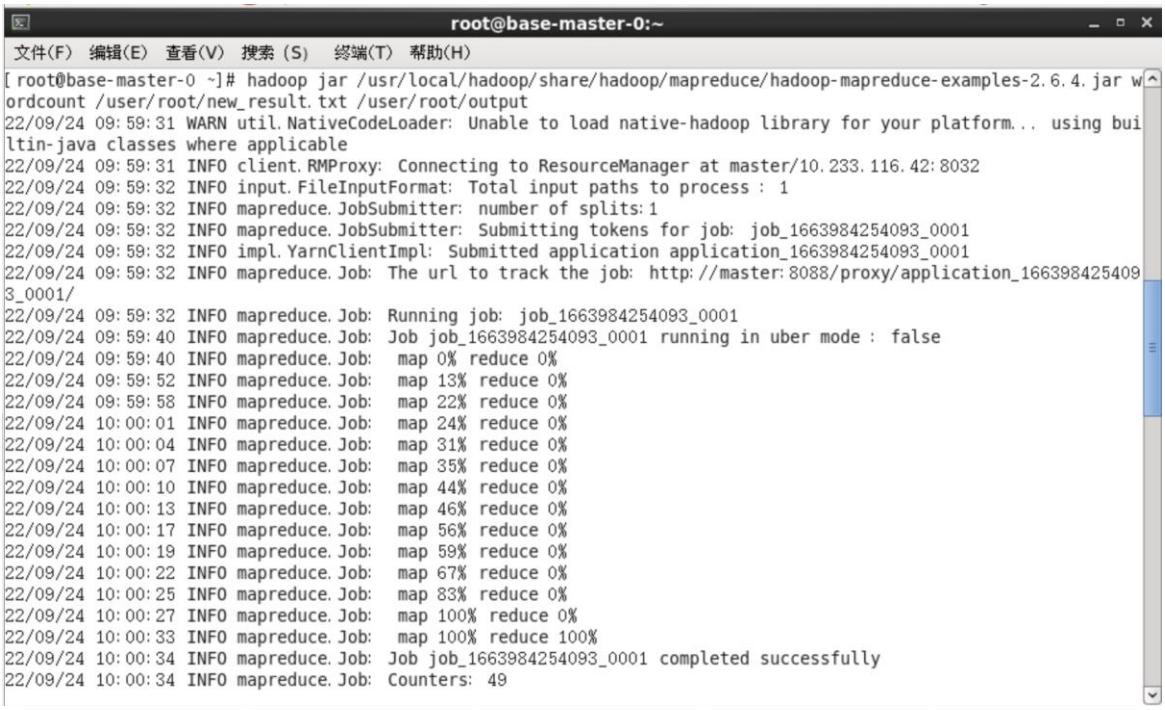

图 20

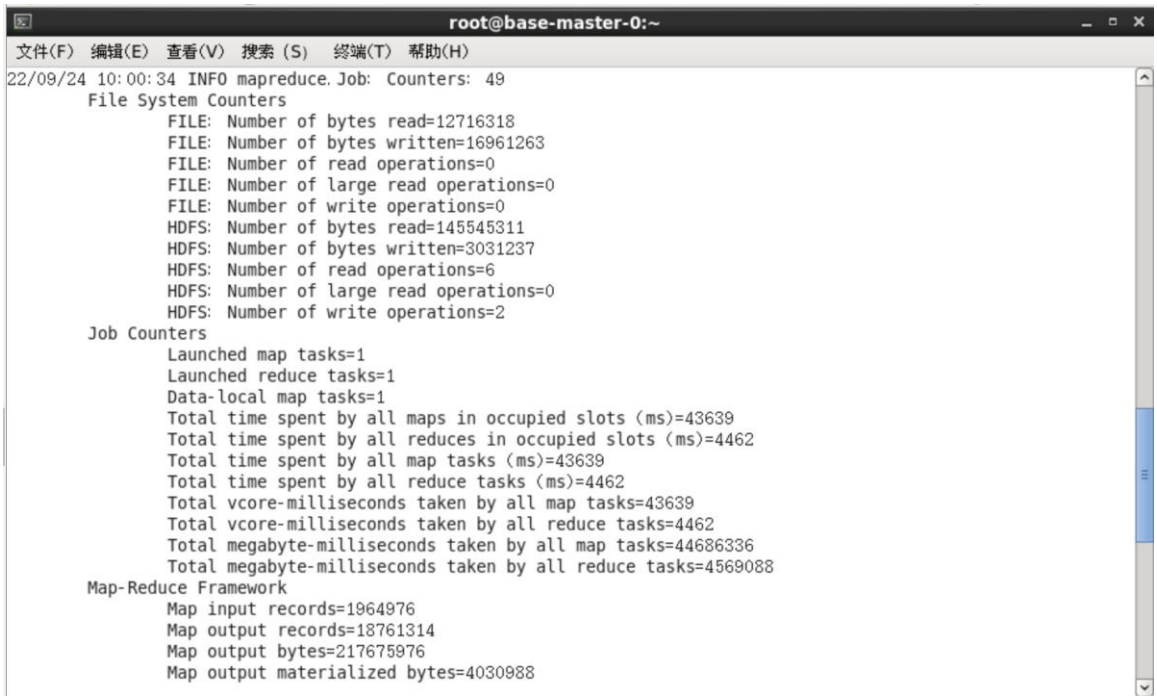

图 21

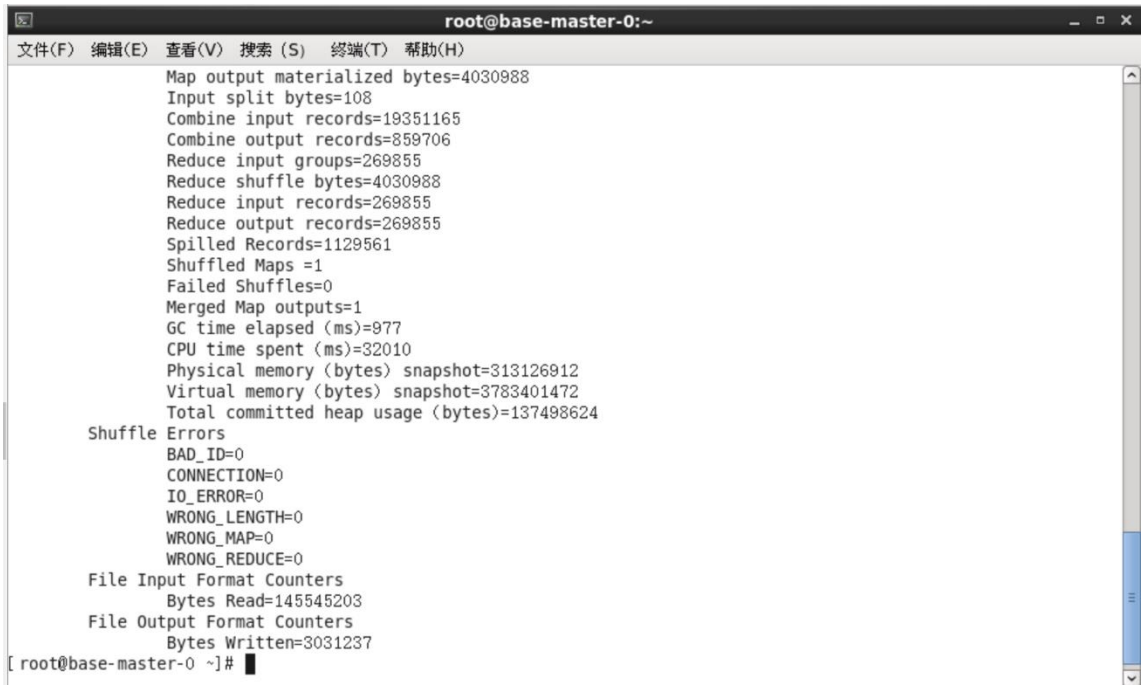

图 22

文件 statistical_result.txt 存储了 WordCount 程序的运行输出（使用 UFT-8 编码）。部分词频统计结果如图 23、24 所示。

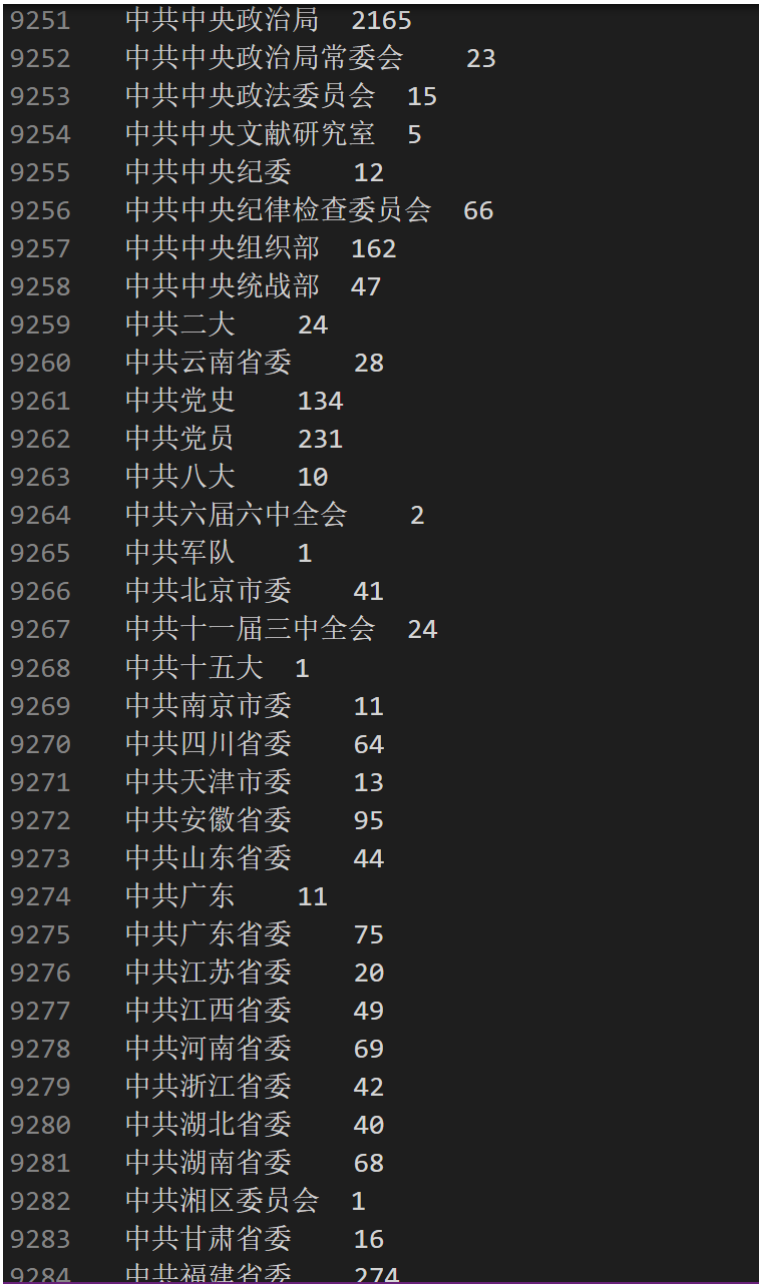
9251 中共中央政治局 2165
9252 中共中央政治局常委会 23
9253 中共中央政法委员会 15
9254 中共中央文献研究室 5
9255 中共中央纪委 12
9256 中共中央纪律检查委员会 66
9257 中共中央组织部 162
9258 中共中央统战部 47
9259 中共二大 24
9260 中共云南省委 28
9261 中共党史 134
9262 中共党员 231
9263 中共八大 10
9264 中共六届六中全会 2
9265 中共军队 1
9266 中共北京市委 41
9267 中共十一届三中全会 24
9268 中共十五大 1
9269 中共南京市委 11
9270 中共四川省委 64
9271 中共天津市委 13
9272 中共安徽省委 95
9273 中共山东省委 44
9274 中共广东 11
9275 中共广东省委 75
9276 中共江苏省委 20
9277 中共江西省委 49
9278 中共河南省委 69
9279 中共浙江省委 42
9280 中共湖北省委 40
9281 中共湖南省委 68
9282 中共湘区委员会 1
9283 中共甘肃省委 16
9284 中共福建省委 274

图 23

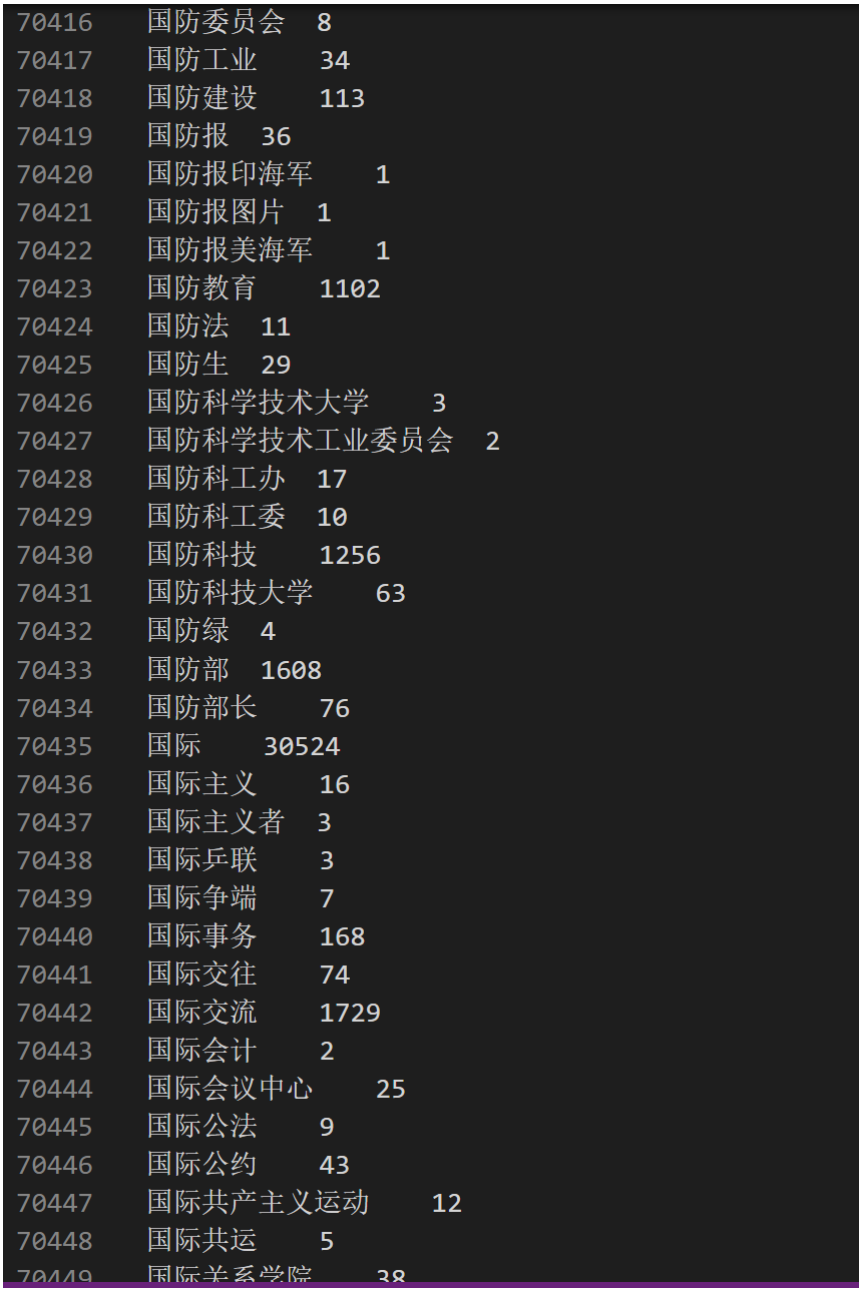
70416 国防委员会 8
70417 国防工业 34
70418 国防建设 113
70419 国防报 36
70420 国防报印海军 1
70421 国防报图片 1
70422 国防报美海军 1
70423 国防教育 1102
70424 国防法 11
70425 国防生 29
70426 国防科学技术大学 3
70427 国防科学技术工业委员会 2
70428 国防科工办 17
70429 国防科工委 10
70430 国防科技 1256
70431 国防科技大学 63
70432 国防绿 4
70433 国防部 1608
70434 国防部长 76
70435 国际 30524
70436 国际主义 16
70437 国际主义者 3
70438 国际乒联 3
70439 国际争端 7
70440 国际事务 168
70441 国际交往 74
70442 国际交流 1729
70443 国际会计 2
70444 国际会议中心 25
70445 国际公法 9
70446 国际公约 43
70447 国际共产主义运动 12
70448 国际共运 5

图 24