哈爾濱工業大學（深圳）
HARBIN INSTITUTE OF TECHNOLOGY

实验报告

开课学期：2020 秋季

课程名称：大数据导论

实验名称：Hadoop 环境配置与基本操作

实验性质：设计型

实验学时：2 地点：T2608

学生班级：1801101

学生学号：180110115

学生姓名：方澳阳

评阅教师：

报告成绩：

实验与创新实践教育中心制

2020 年 7 月

一、 实验目的

1. 熟悉 Hadoop 单机环境和伪分布式环境的配置方法；
2. 熟悉命令行运行 Mapreduce 作业的原理和操作。

二、 实验内容

1. 搭建 Hadoop 单机环境和伪分布式环境；
2. 使用 MapReduce 实现 wordCount 任务

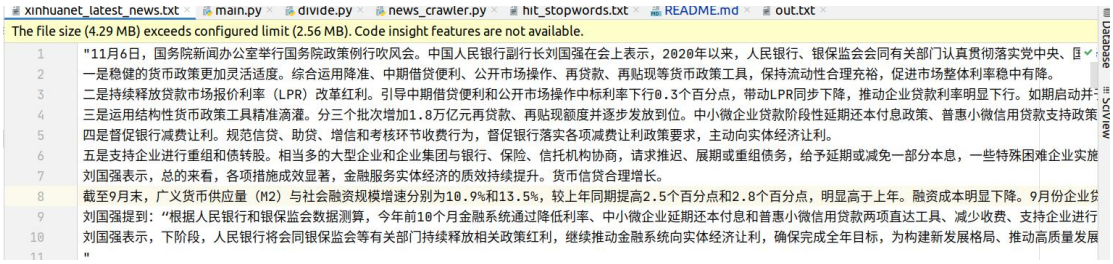
三、实验环境

由于自己已有笔记本中的 ubuntu 系统，故使用自己的笔记本进行实验。
1. Ubuntu 20.04
2. Hadoop 2.7.2

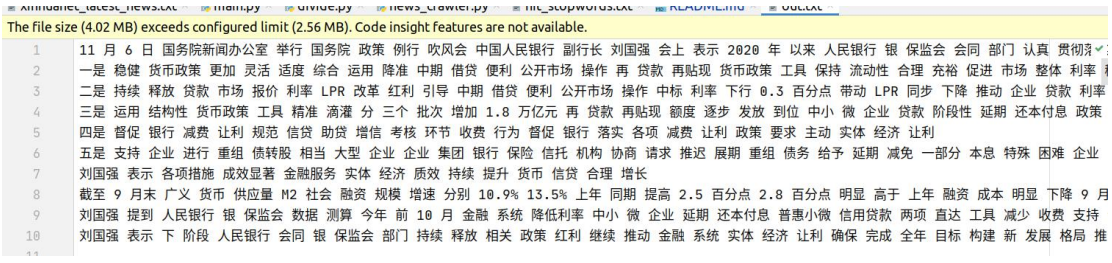
四、实验过程

实验准备阶段

使用软件： pycharm2020.2
使用 python 爬虫爬取新华网的新闻，保存为 txt 文件。



使用 jieba 对上述文件进行分词处理



正式实验

使用软件： vscode+shell
将上述分词好的 txt 文件进行统计

单机模式：

由于一开始没有截图。改成伪分布式之后再修改回去较为困难，所以以伪分布式的截图为准。

伪分布式：

```
llincyaw@llincyaw-TM1703 [10:55:37] [~/Documents/hadoop]
-> % hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.2.jar wordcount input output
20/11/13 10:55:41 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/11/13 10:55:42 INFO input.FileInputFormat: Total input paths to process : 2
20/11/13 10:55:42 INFO mapreduce.JobSubmitter: number of splits:2
20/11/13 10:55:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1605236134784_0001
20/11/13 10:55:43 INFO impl.YarnClientImpl: Submitted application application_1605236134784_0001
20/11/13 10:55:43 INFO mapreduce.Job: The url to track the job: http://llincyaw-TM1703:8088/proxy/application_1605236134784_
0001/
20/11/13 10:55:43 INFO mapreduce.Job: Running job: job_1605236134784_0001
20/11/13 10:55:50 INFO mapreduce.Job: Job job_1605236134784_0001 running in uber mode : false
20/11/13 10:55:50 INFO mapreduce.Job:  map 0% reduce 0%
20/11/13 10:55:56 INFO mapreduce.Job:  map 50% reduce 0%
20/11/13 10:55:58 INFO mapreduce.Job:  map 100% reduce 0%
20/11/13 10:56:03 INFO mapreduce.Job:  map 100% reduce 100%
20/11/13 10:56:03 INFO mapreduce.Job: Job job_1605236134784_0001 completed successfully
20/11/13 10:56:03 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=582403
                FILE: Number of bytes written=1517431
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4021068
                HDFS: Number of bytes written=423498
                HDFS: Number of read operations=9
```

```
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=9363
                Total time spent by all reduces in occupied slots (ms)=3591
                Total time spent by all map tasks (ms)=9363
                Total time spent by all reduce tasks (ms)=3591
                Total vcore-milliseconds taken by all map tasks=9363
                Total vcore-milliseconds taken by all reduce tasks=3591
                Total megabyte-milliseconds taken by all map tasks=9587712
                Total megabyte-milliseconds taken by all reduce tasks=3677184
        Map-Reduce Framework
                Map input records=17083
                Map output records=572858
                Map output bytes=6292334
                Map output materialized bytes=582409
                Input split bytes=231
                Combine input records=572858
                Combine output records=41644
                Reduce input groups=41640
                Reduce shuffle bytes=582409
                Reduce input records=41644
                Reduce output records=41640
                Spilled Records=83288
```

```
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=393
                CPU time spent (ms)=6410
                Physical memory (bytes) snapshot=727441408
                Virtual memory (bytes) snapshot=5688266752
                Total committed heap usage (bytes)=523239424
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=4020837
        File Output Format Counters
                Bytes Written=423498
(base)
```

由于 cat 命令显示过长，无法显示指令，因此截取一小段。
因为使用 head|tail 管道命令报错 Unable to write to output stream.

```
领导力	1
领导小组	9
领导核心	1
领导班子	2
领投	1
领涨	34
领玺	1
领罚	1
领航	3
领航员	1
领衔	1
领衔主演	1
领袖	2
领跌	12
领跑	19
领跑者	1
领路	1
颇	13
颇丰	4
颇为	11
颇具	5
颇受	2
颇受欢迎	1
颇感	2
颇显	1
颇重	1
颐海	3
频	2
频上	1
频出	15
频发	5
频引	1
```

导出后再查看文件：

```
llincyaw@llincyaw-TM1703 [11:10:19] [~/Documents/hadoop]
-> % bin/hdfs dfs -get output ~/Documents/Big-data-Project/
(base)
llincyaw@llincyaw-TM1703 [11:10:44] [~/Documents/hadoop]
-> % cat ~/Documents/Big-data-Project/output/part-r-00000 | head -n 50
(BIS),	1
(ECCN)	1
(TSU)	1
(see	1
0	545
0%	4
0.00035	1
0.0007	1
0.008	2
0.01	3
0.01%	3
0.0175%	1
0.0177%	1
0.02%	3
0.0275%	1
```

三、 实验结果与分析

xinhuanet_latest_news.txt

```
24979	旅业 11
24980	旅企 2
24981	旅客 123
24982	旅客列车 5
24983	旅客量 1
24984	旅居 2
24985	旅投 3
24986	旅检 2
24987	旅游 403
24988	旅游业 41
24989	旅游业者 1
24990	旅游圈 1
24991	旅游局 5
24992	旅游景点 3
24993	旅游点 2
24994	旅游线 1
24995	旅游者 4
24996	旅游部 14
```

在 pycharm 中查看输出的文件，可以看出已经成功输出了对应的统计结果。

这次实验的主要目的是让我们熟悉如何配置 hadoop 环境，难度其实并不高。考验的是学生对命令行，linux 等掌握程度。

由于我自己并不是在学校机房的电脑上进行的实验，在配置环境的时候指导书上的配置方法并不适合我。

例如我使用的 shell 是 zsh，因此修改的配置文件并不是去修改 etc/profile。一开始我并没有意识到这一点，在我修改了 profile 之后没有出现效果。因此我去搜索了相关的信息，网上说需要重启。在当我重启之后，我的电脑就无法登录了，一直重复循环登录界面。还好自己以前也遇到过这样的情况，因此使用命令行登录之后把 profile 修改回去就可以登录了。

再经过一番搜索之后才知道应该在.zshrc 中添加环境变量。那按道理如果是用 bash 的话应该是在.bashrc 中添加环境变量，那么指导书为什么是在/etc/profile 中修改呢？我系统中的 profile 看起来也不像能够在这里添加环境变量的。

这些仍需要去细细体会。

打开(O) profile /etc 保存(S)

```
# /etc/profile: system-wide .profile file for the Bourne shell (sh(1))
# and Bourne compatible shells (bash(1), ksh(1), ash(1), ...).

if [ "${PS1-}" ]; then
  if [ "${BASH-}" ] && [ "$BASH" != "/bin/sh" ]; then
    # The file bash.bashrc already sets the default PS1.
    # PS1='\h:\w\$ '
    if [ -f /etc/bash.bashrc ]; then
      . /etc/bash.bashrc
    fi
  else
    if [ "`id -u`" -eq 0 ]; then
      PS1='# '
    else
      PS1='$ '
    fi
  fi
fi

if [ -d /etc/profile.d ]; then
  for i in /etc/profile.d/*.sh; do
    if [ -r $i ]; then
      . $i
    fi
  done
  unset i
fi
```

个人签名 ：

年　　月　　日