Big Data and Data Analysis Hw2

B00902064 宋昊恩

May 18, 2015

1 Basic Part

1.1 Implementation

Firstly, I installed virtualbox and built up a virtual machine in the version of Ubuntu 14.04. Later on, I built up the hadoop in version 2.7.0 instead of version 1.2.1 mentioned in the slides. There were some difference between them, but not a big deal.

After environment built up, I read many hadoop tutorial to get through the main framework of it. The latter part of java code refinement and parsing code implementation in python were comparitively much easier.

2 Bonus Part

2.1 Implementation

I parsed the data in python code again. After that, I simply counted the number of pair appearences in terms of all pair in *plist*, and then filtered the valid pairs.

3 Encountered Problem

The main problems I encountered are all about the spec and powerpoint delivered by the course. I am sorry that I need to complain about it.

First, the newest version of hadoop is 2.7.0, which is far newer than the 1.2.7 version provided in the slides. There are already many differences between this two version, in both APIs and installation. Why can't the TA introduce the newer one? Especially the download link provided are even invalid...

Second, the input and output format are not specified. I am confused about the output format for the sorted result. Should I print the id itself or with the number of appearences? And, what is the order of them? Furthermore, there is no clear defintion on what we have to do with hadoop and what we can do with other source codes. Parsing *EHC_1st_round.log* into *big.list*-like format is acceptable for the homework basic part, but can we do the same things for the bonus part? And, What is the output format for the bonus part? There are just too many questions left for me to ask.

I have been really disappointed at this course since the ridiculous Hw1 spec. I am looking forward to the improvemnt in the following homeworks. Thank you.

4 Execution

There are two python files, *parseID.py*, and *parsePair.py*; three java files, *CountFilter.java*, *Sort.java*, *WordCount.java*; and three shell files, *BonusRun.sh*, *SortCompile.sh*, *SortRun.sh*.

Parsing code in python is mainly used to parse the file *EHC_1st_round.log*, I assumed it should be the same directory as the other source files.

For compiling, execute *SortCompile.sh* will create a folder *my.class* and a jar file *my.jar*.

For executing basic part of homework, execute *SortRun.sh* with two arguments.

For executing bonus part of homework, execute *BonusRun.sh* with three arguments.

The usage of arguments is the same as homework spec.