## Appendix D. FAQ

### 1. Versions of tools (Scala, Python, Spark…).

If your computer is not M1 or M2 MAC. You'd better install the same version of tools as Appendix C to avoid some errors. If your computer is M1 or M2 MAC, please refer to question 3.

- Java8
- Scala2.12.4
- Maven3.5.2
- Python3.6 (If you want to use PyCharm, you need to install python3.6 or higher version)
- Spark2.2.1

### 2. Can I use Windows to set up all of the environments needed for this lab?

Yes. But it's more difficult. So it's better to use the virtual machine to get the Ubuntu system.

### 3. I cannot install Virtual Box on Mac.

Virtual Box doesn't support Mac Book with M1 or M2. Although the developer preview version of VirtualBox for Mac with M1 or M2 has been released, it's only for beta test. It's unstable with many strange errors. So you need to install the stand-alone Spark or connect to the server according to Appendix E.

### 4. Can I use dataframe or other methods as the solution?

No, you need to implement the labs with RDDs.

### 5. The Ubuntu image already has python3.5, could I upgrade it to python3.6?

No, you need to keep Python3.5 and additionally install Python3.6.

### 6. The order of the outcome of my example code (wordcount) is different from Appendix C.

It doesn't matter. The example code doesn't sort the result, so the order of the outcome depends on which worker finishes his job first, which could be random.

**7. There are several files rather than one in my output folder.**

Please refer to https://mungingdata.com/apache-spark/output-one-file-csv-parquet/.

**8. I cannot copy files from Virtual Box to Windows/Mac.**

Maybe you need to configure the shared folder first. You can refer to https://helpdeskgeek.com/virtualization/virtualbox-share-folder-host-guest/ or google the question.

**9. Notes to process the dataset.**

- You would better use the parameter "allowBackslashEscapingAnyCharacter=True" when reading json file with read.json function. (**Important**)
- "brand" in some records is none or empty, just drop out these records.

**10. The virtual machine sometimes runs quite low and may pause. What can I do?**

You could allocate more CPU processors and memory to the virtual machine.

**11. I was running appendix C but I don't see any output file even though it ran successfully.**

Maybe you saved the output to HDFS rather than the local computer.

**12. Which IDE should I use?**

You'd better choose which you are familiar with. If you are new to CS and are using M1 or M2 Mac, it's more convenient to use VScode to connect to remote Soc cluster. Otherwise, it's more convenient to implement and debug with PyCharm.

**13. How could I get a high mark?**

Grading of lab1 is based on the result and documentation for the code, rather than the quality of the code or other things. But the program of lab1 is simple, therefore, it should get the result within one or two minutes. If we cannot get the result in 5 minutes, we will give some penalty to you.

14. **Do documentation, is the code documentation suffices, or do we need to put one readme.text file for further explanation?**

Both are ok. Remember that documentation is needed for important steps.

15. **If you are using Jupyter Notebook, remember to convert it to a python file before submitting.**

16. **If you have a question, please state it on Canvas so that other students can also see it. (Important)**

17. **Please read the instruction of lab1 and the Appendix carefully before submitting.**