1. How I can copy content from the terminal and paste content into the terminal?
    a) Putty:
        i. Copy from terminal:
            1. select the content you want to copy, it will be automatically copied
        ii. Paste into terminal:
            1. Copy the content
            2. In the putty terminal, click mouse “right” button
    b) Windows subsystem Linux:
        i. Copy from terminal:
            1. Select the content
            2. Ctrl + shift + C
        ii. Paste into terminal:
            1. Copy the content
            2. Ctrl + shift +V
        iii. For Mac Terminal: https://superuser.com/questions/62609/how-can-i-copy-on-select-in-the-os-x-terminal-like-putty-does
        iv. For Ubuntu Terminal: https://sourcedigit.com/13930-copy-paste-text-linux-ubuntu-terminal/

2. Why I cannot connect to my VM?

    There are several things you need to check:

    a) Check your ssh client setup
    b) Check your connection
        i. If the error is like “connection denied”, the problems can be the two conditions below:
            1. If you enable the firewall without open the “22” port for SSH, it may become a problem because you need port “22” to connect to the server. If this is the problem, please send the TA an email with your condition, your VM IP, and the hostname of your VM.
            2. If you start the Hadoop cluster without open the firewall to block the possible attacks, your VM may be hacked from the WebApp of Hadoop. If this is the condition, please send your condition, your VM IP, and your VM hostname to the TA.

c) Check your ssh key pair:

i. If you use mac or linux, you need to make sure the private key file “key_student” can be only modified by the current user by the command “chmod 600 key_sutdent”. Then try to connect by “ssh” again.

ii. If you use windows putty client, please make sure you choose the “student.ppk” file in the putty setting

3. Why after I format the namenode, the HDFS cannot run?

   This is mostly because you do not delete the HDFS directory in the local disk. Every time you format the HDFS, it will generate a unique clusterID and stores it in the HDFS directory for both the datanodes and namenode. However, if you format the namenode again, it will gernerate a new clusterID. But for the datanodes, it also uses the old clusterID to connect to the namenode which will throw out an exception that the clusterID is not matched. So for this kind of problems, you can follow the steps below:

   a) Stop the HDFS services: sbin/stop-dfs.sh

   b) Delete “/tmp/hadoop-student” directory on the master node: rm -rf /tmp/hadoop-student

   c) Delete “/tmp/hadoop-student” directory on all the slave nodes: rm -rf /tmp/hadoop-student

   d) Reformat HDFS by: bin/hadoop namenode -format

   e) Restart the HDFS services by: sbin/start-dfs.sh

4. Which Hadoop version should I choose?

   Any version you satisfy. But the latest stable version can be a good choice.

5. What should I do if the VMs meet a memory issues when running some MapReduce examples?

   You can either reduce the number of slaves (For example, if you have 3 VMs, you can set the workers to the two VMs which are not the master node) or try an example with smaller input (sample the input to a smaller file).

6. Should I do the performance test for the n-gram Part 3 with a randomly generated dataset?

   No, you have not to do that. This part is recommended but not mandatory.

7. If I have messed up, what should I do if I want to start from the beginning?

   The steps you need to do are:

i. Stop all the Hadoop services

ii. Delete the Hadoop directory

iii. Delete the HDFS directory

a) If the HDFS directory is not set, the default directory of HDFS, it uses /tmp/Hadoop-student as the default derectory

iv. Delete the template files in /tmp directory for the SeconderyNameNode

v. Re-check any configuration files you may edited before, they can be the files as below but not restricted to:

a) /etc/hosts

b) /etc/hostname

c) ~/.bachrc

d) /etc/environment

e) /etc/ssh/sshd_config

f) /etc/profile

g) …

vi. Re-extract the Hadoop package to the directory

vii. Re-do the Hadoop cluster setup from the beginning

8. In the Docker setup, what should I submit?

   You need to submit the Dockerfile and the configuration files(e.g. core-site.xml, hdfs-site.xml, ...) for a pseudo-distributed (single node) Hadoop, which is much simpler than what you did in the real Hadoop cluster setup. The test of it can be done on your own computer.

   If you have no idea of how to start, you can refer to other Hadoop docker image such as: https://github.com/sequenceiq/hadoop-docker

9. In Dockerfile, how can we type a “yes” when builds the Docker image with some installation steps like “ apt-get install ssh”?

   You just need an automatical acceptance installation command like :

   `apt-get install –y ssh`

10. How can I send files to the VMs?

There are many ways you can do that:

a) If you use linux or mac “ssh” command line tool, you can just use “scp” to copy the file.

b) If you use windows, you can try “WinSCP” to transfer the files.

11. How to debug my wordcount program?

You can debug it in several ways:

a) Compile the java files into class files and package them into a jar file. Then upload it to the cluster to run it. You can check: https://hadoop.apache.org/docs/r3.2.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html for some help

b) If you want to debug it on your own laptop with some IDEs, you can try the tutorials below (not guarantee it works, because for different Hadoop versions, the setting will be a little different):

i. With Eclipse: https://dzone.com/articles/running-hadoop-mapreduce

ii. With IDEA: https://mrchief2015.wordpress.com/2015/02/09/compiling-and-debugging-hadoop-applications-with-intellij-idea-for-windows/