Running the WordCount example

Summary

This guide describes how to *compile* and *run* the **WordCount** program, a Java program implementing the *map and reduce functions* used for counting the number of words in the text of **The Miserables**.[1]

Compiling the WordCount program

The WordCount program resides inside the **WordCount folder**.[2] The folder is composed of the following files:

- **WordCountMapper.java**. Contains the *map* function implementation.
- **WordCountReducer.java**. Contains the *reduce* function implementation.
- **WordCount.java**. Contains the code coordinating the execution of the *map and reduce functions*.

Inside order to compile the WordCount program, execute the following commands in the WordCount folder:

```
WordCount J$  javac -cp hadoop-core-1.0.4.jar *.java
WordCount J$  jar cvf WordCount.jar *.class
```

The first command *compiles* the program using the classes developed by Hadoop (i.e., ***hadoop-core-1.0.4.jar***). The second command *creates* a *jar* file called **WordCount.jar** that you will use for running the WordCount program in Hadoop.

Running the WordCount program in Hadoop

Assuming that your are in the folder containing your *Hadoop installation*, execute the following commands

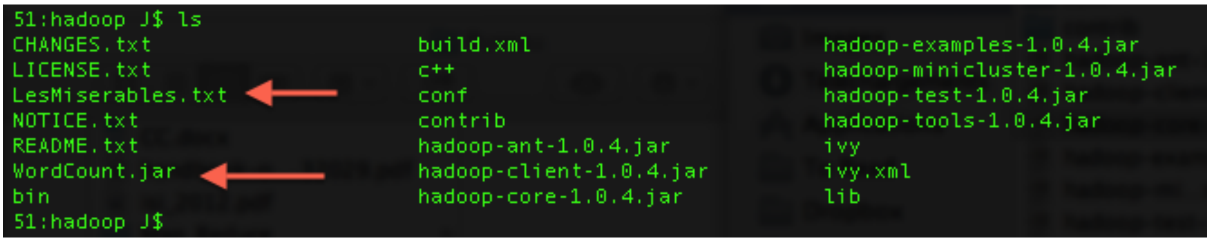
```
hadoop J$  bin/start-all.sh
hadoop J$  ssh localhost
hadoop J$  mkdir input
```

[1] This guide assumes that you already have *The Miserables* inside a file called **TheMiserables.txt**.

[2] You should find the **WordCount** folder next to this guide.

The first command starts the Hadoop services. The second command establishes a secure connection with your machine.[3] The third command creates the directory where you will put file containing **The Miserables**.

Afterwards, copy the **WordCount.jar** and the **TheMiserables.txt** file into the folder containing your *Hadoop installation* (cf. figure below).



```
51:hadoop J$ ls
CHANGES.txt                     build.xml                       hadoop-examples-1.0.4.jar
LICENSE.txt                     c++                             hadoop-minicluster-1.0.4.jar
LesMiserables.txt               conf                            hadoop-test-1.0.4.jar
NOTICE.txt                      contrib                         hadoop-tools-1.0.4.jar
README.txt                      hadoop-ant-1.0.4.jar            ivy
WordCount.jar                   hadoop-client-1.0.4.jar         ivy.xml
bin                             hadoop-core-1.0.4.jar           lib
51:hadoop J$
```

Then prepare the input for the WordCount program:

> *hadoop J$* bin/hadoop dfs -mkdir input
> *hadoop J$* bin/hadoop dfs -put LesMiserables.txt input

The former command creates a directory called **input** in the *Hadoop Distributed File System* (HDFS). The second command will copy **TheMiserables.txt** into the *input* folder in HDFS. Without this command Hadoop cannot find the input file.

Finally execute the following commands:

> *hadoop J$* bin/hadoop jar WordCount.jar *WordCount* **input** output
> *hadoop J$* bin/hadoop dfs -get **output** output

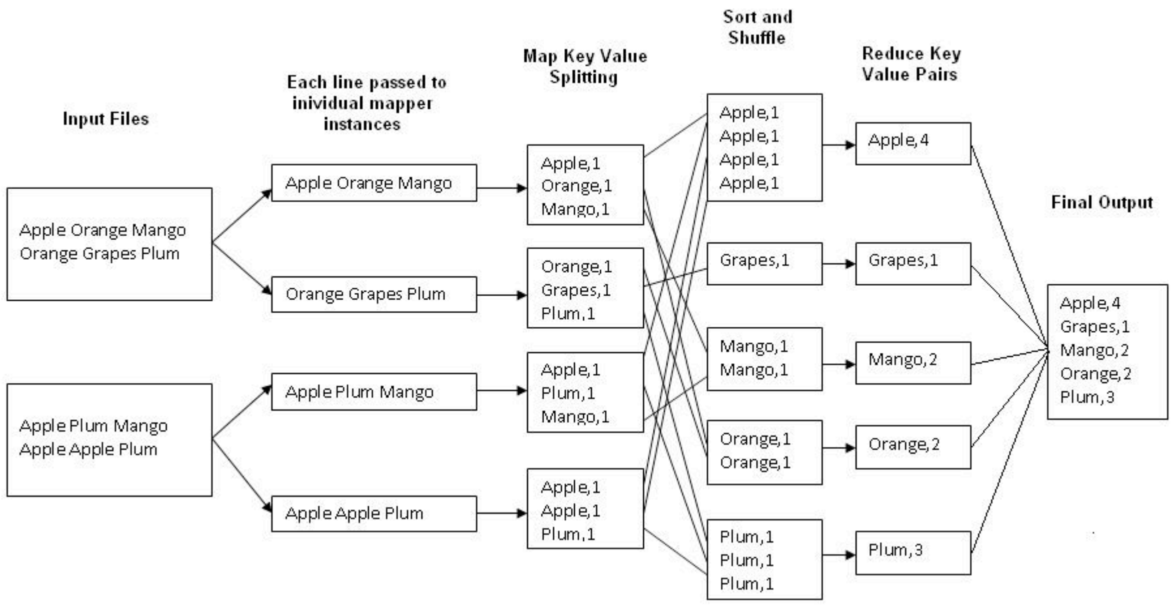
The first command run the *WordCount* program in Hadoop. Note that the command specifies the names of:

- the class where the *main method* resides (cf. the **WordCount.java** file).
- the HDFS folder where the *input* files resides.
- the HDFS folder that will contain the *output* files.

The second command copies the *output* folder from HDFS to your machine. You will find the result of the WordCount program in a file (probably) called **part-00000**.

[3] This is required due to the *implementation* of Hadoop.

For illustration purposes the following image gives a *general overview* of the execution of the WordCount program.



Finally recall that you can monitor the execution of the WordCount program by navigating to the following addresses:

- http://localhost:50070/ – web UI of the NameNode daemon
- http://localhost:50030/ – web UI of the JobTracker daemon
- http://localhost:50060/ – web UI of the TaskTracker daemon