Assignment 2

Course Name: Big Data Analysis, Module 1, Fall Semester 2020, PHBS

Assignment Type: group work

Version: v1.2

Submission Requirement:

1. Each group should only submit one electronic copy to both TA's and the professor's email inbox.
2. The email's title should be **2020.M1.BigDataAnalysis.HW2.GroupX** , where the **X** is your group number.
3. In the main body of the email, please write down explicitly your group information, including your group member's names and student IDs.
4. In the attachment of the email, please put all your codes and necessary documentation files in one zip/rar file named with exactly the same name of your email's title.
5. In your code submission (it is a part of your zipped file), you should make sure you create one folder for one map reduce job, your code submission should contain both your map reduce output folder as well as the screenshots of your map reduce execution, each of your screenshot **must** show the system clock of your device.

Submission Deadline: **2020 Nov.1 11:59pm(Beijing Time Zone)**

Tips:

1. **In this homework, you are not required to upload your code to the cluster for execution, the homework can be accomplished locally.**
2. Generally, before you submit your job to the cluster, always check the status of resource manager & HDFS by `hadoop-namenode:9082` & `hadoop-namenode:8088` (if you configured the 'host' file) or `10.0.50.139:9082` & `10.0.50.139:8088` and make sure you are @ PHBS.
3. Anytime, debug your code locally on a small piece of data firstly, see the end of this file for more information.
4. Practically, map reduce can be realized with other programming languages, refer to MATLAB for example.

Part 1. Air Quality Data Set

In this part, you possibly only use the following columns of the data: **No**, **year**, **month**, **day**, **hour**, **PM2.5**, **station**. One city only has one station.

The PM2.5 column contained in `airQuality` file is a numerical data, however, for the convenience of classification, the following definition is applied: every hour, if the PM2.5 of a station is **smaller than or equal to 35**, the air quality of this hour of the city the station is in is defined as '**good**'; if the PM2.5 is **greater than 35 and smaller than or equal to 75**, the air quality is defined as '**medium**'; **otherwise**, the air quality is defined as '**bad**'.

Questions

Using map reduce, Java, you are asked to complete the following tasks:

1. Using the air quality classification rule defined by PM2.5, summarize **(1)** ratio of cities/stations has "good" air quality each hour, **(2)** ratio of cities/stations has "medium" air quality each hour, **(3)** ratio of cities/stations has "bad" air quality each hour.
2. Using the air quality classification rule defined by PM2.5, summarize **(1)** for each city, how long (count by hour) does a city have "fine"(= "good" or "medium") air quality **(2)** for each city, how long (count by hour) does a city have "bad" air quality.

Hint: when dealing with the data, check the quality of the data firstly, pay special attention to NA value.

End of assignment 2

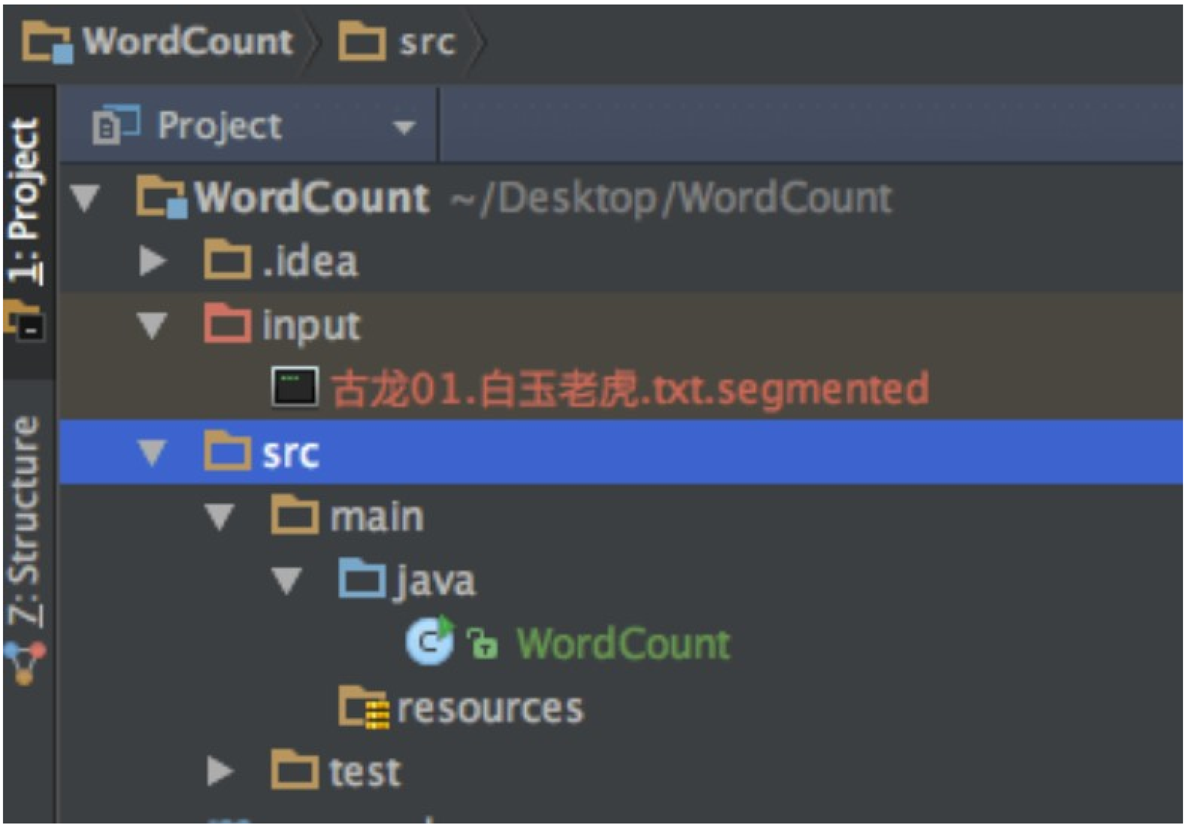
Debug map reduce locally with a handy data

Material: IntelliJ IDEA.

Note: the screenshot is only for reference!

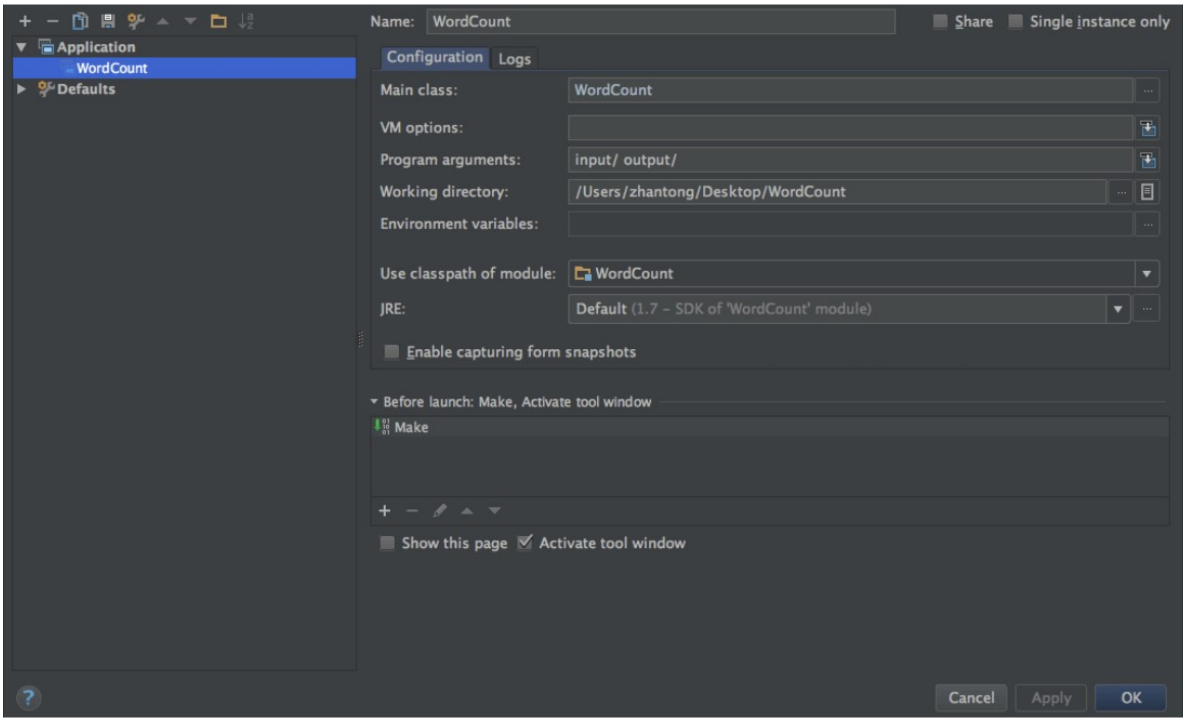
Reference

1. Suppose you have a started a java project, create a folder called `input` in the same level of the `src` folder, put your piece of data into the `input` folder.



2. Select `Run`--`Edit Configurations`, click `+` to add new `Application`, set the `Main class` to be your main class, set `Program arguments` to be `input/ output/`.

   Note: you don't need to create `output` folder since it will be created automatically and you need to clear the content in `output` folder if there is any before running the codes.



Name: WordCount

Main class: WordCount

VM options:

Program arguments: input/ output/

Working directory: /Users/zhantong/Desktop/WordCount

Environment variables:

Use classpath of module: WordCount

JRE: Default (1.7 - SDK of 'WordCount' module)

□ Enable capturing form snapshots

□ Show this page ☑ Activate tool window

3. Once you run the program, the program will create the `output` folder automatically and you can check `part-r-00000` for results directly in the IntelliJ IDEA.