Configurations & Instructions

Download SampleData:
http://zhidong.weebly.com/uploads/8/3/7/6/8376282/sampledata.zip

First we should do is putting the SampleData.txt into your HDFS.

1 Distribution Count:

Compile the three java files:

DistributionCount.java
DistributionCountMapper.java
DistributionCountReducer.java

It is very like our first assignment.

After compiling the files in to one jar, Suppose you have a DistributionCount.jar
run it with:

```
hadoop DistributionCount.jar DistributionCount /user/yourusername/SampleData.txt /user/yourusername/OutputFileName
```

SampleData.txt is the small sample data provided.
OutputFileName is specified in your HDFS, then use the getmerged statement to copy it to somewhere.

2 Gene's beta value ranking:

BetaValueAdderMapper.java
BetaValueAdderReducer.java
BetaValueRanking.java

Note that we use a temp file to store the intermediate data of sorting, a non-trivial change should be made in BetaValueRanking.java:

Path tempath=new Path("/user/zhidong/temp");//Here is the location of temp file, must be changed to your user folder: Path tempath=new Path("/user/yourusername/temp");

After compiling the files in to one jar, Suppose you have a BetaValueRanking.jar
Then run it with:

```
hadoop BetaValueRanking.jar BetaValueRanking /user/yourusername/SampleData.txt /user/yourusername/OutputFileName
```

SampleData.txt is the small sample data provided.
OutputFileName is specified in your HDFS, then use the getmerged statement to copy it to somewhere.

3 Distribution Version K-Means:

Before some compiling some changes should be made:
In the KMeansMapper.java file:

Change
String BASE_PATH = "/user/zhidong/TCGAkmeans";
To your user name: "/user/yourusername/TCGAkmeans";

In the KMeansMain.java file:

Change:
String BASE_PATH = "/user/zhidong/TCGAkmeans";
String CEN_PATH = "/user/zhidong/TCGAkmeans/center/centroid.txt";
To:
String BASE_PATH = "/user/ yourusername /TCGAkmeans";
String CEN_PATH = "/user/ yourusername /TCGAkmeans/center/centroid.txt";

Compile the java files:
KMeansMain.java
KMeansMapper.java
KMeansReducer.java

To one jar: KMeansMain.jar

```
hadoop KMeansMain.jar KMeansMain /user/yourusername/SampleData.txt
```

The result will be in the folder of

/user/ yourusername /TCGAkmeans/clustering/depth_**iteration**/part-r-00000

The iteration a number which is defined in my code, I make it 2, so result is in /user/ yourusername /TCGAkmeans/clustering/depth_2 /part-r-00000

You can define it as what you like. The intermediate result is stored in depth_"xx" where the "xx" less than iteration. You are expected to use the getmerged method to copy to your folder.
Feel free to contact me if you have any problem.