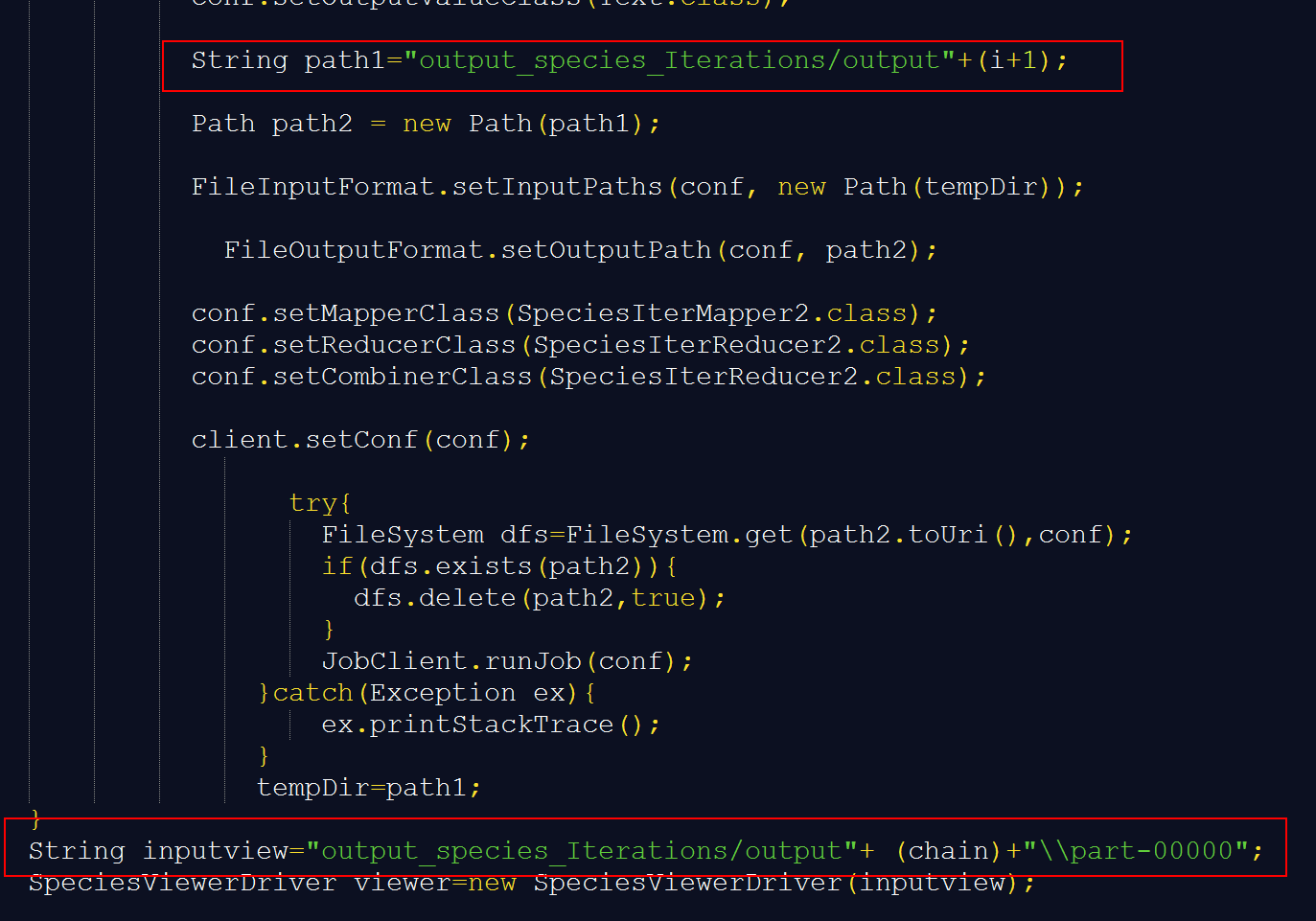
**Summary**

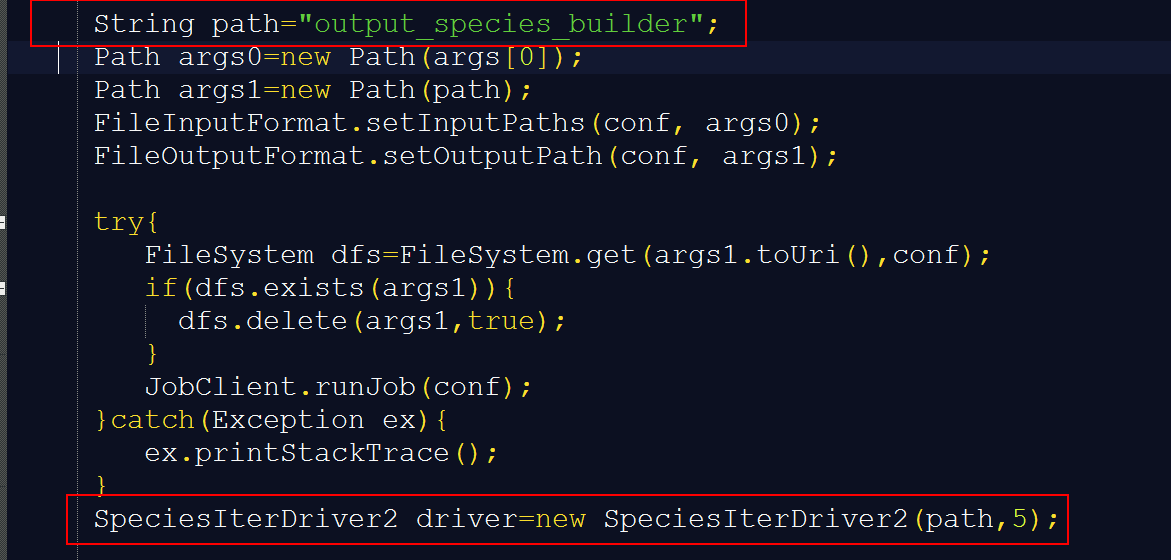
Part 1

First, follow the step on the Species Lab Slide and make 9 code file for the Page Rank,

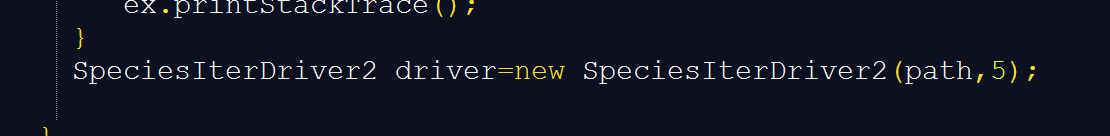
Separate them in 9 different java file.

Run the file in the Cygwin and get the first output.

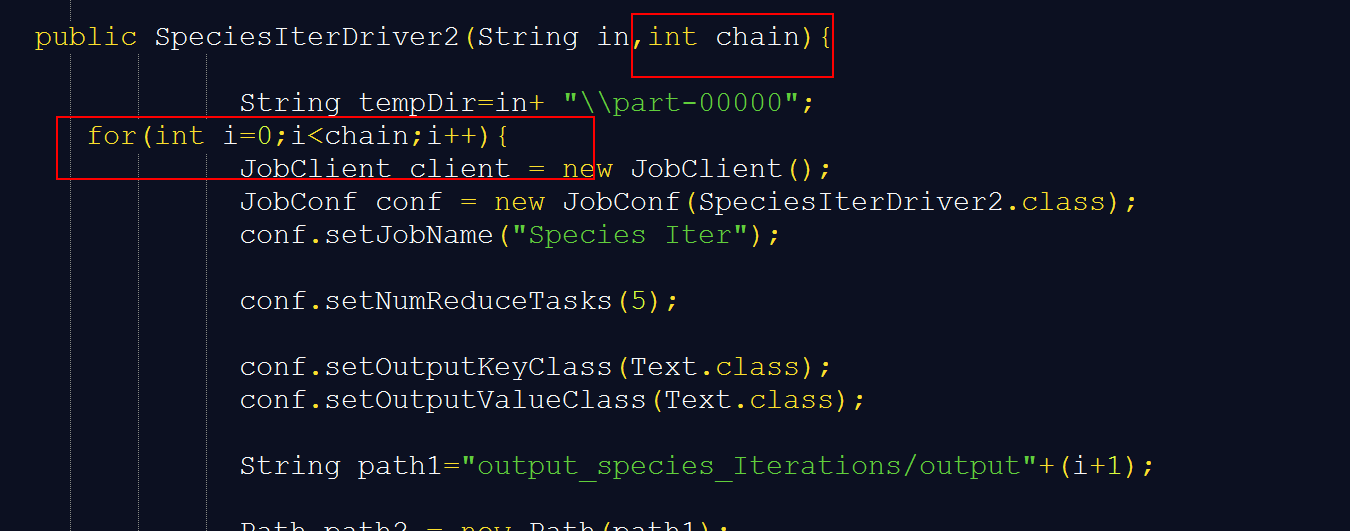
In my code, I have already set the input file and output file ‘s directions and every driver or runner can read other’s output file automatically :



and my iterations run for 5 times:



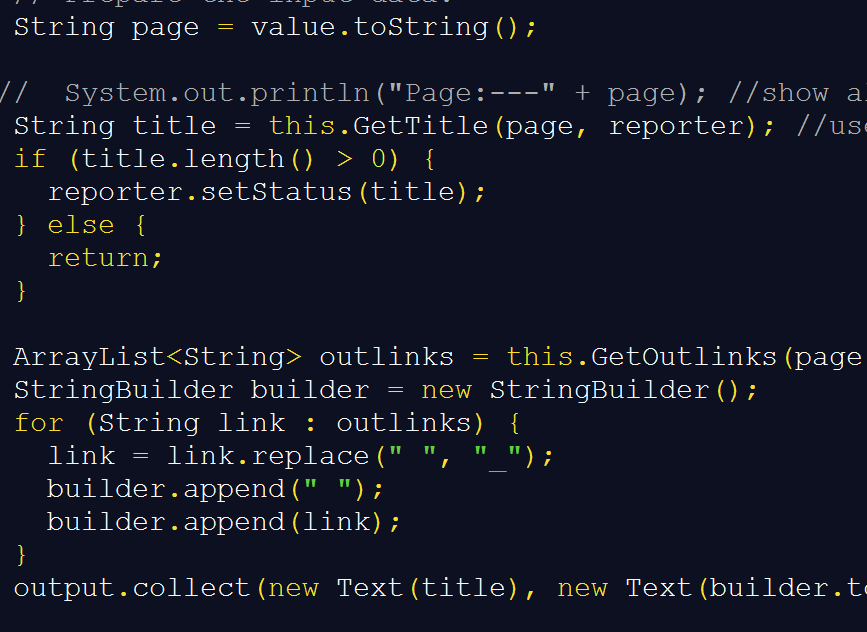
5 is the parameter for the function SpeciesIterDriver class because I have made a for loop for the iterations:



In the Mapper part：

We can see that in the code we can recognized about the xml files and split it than transferred it into the output file.

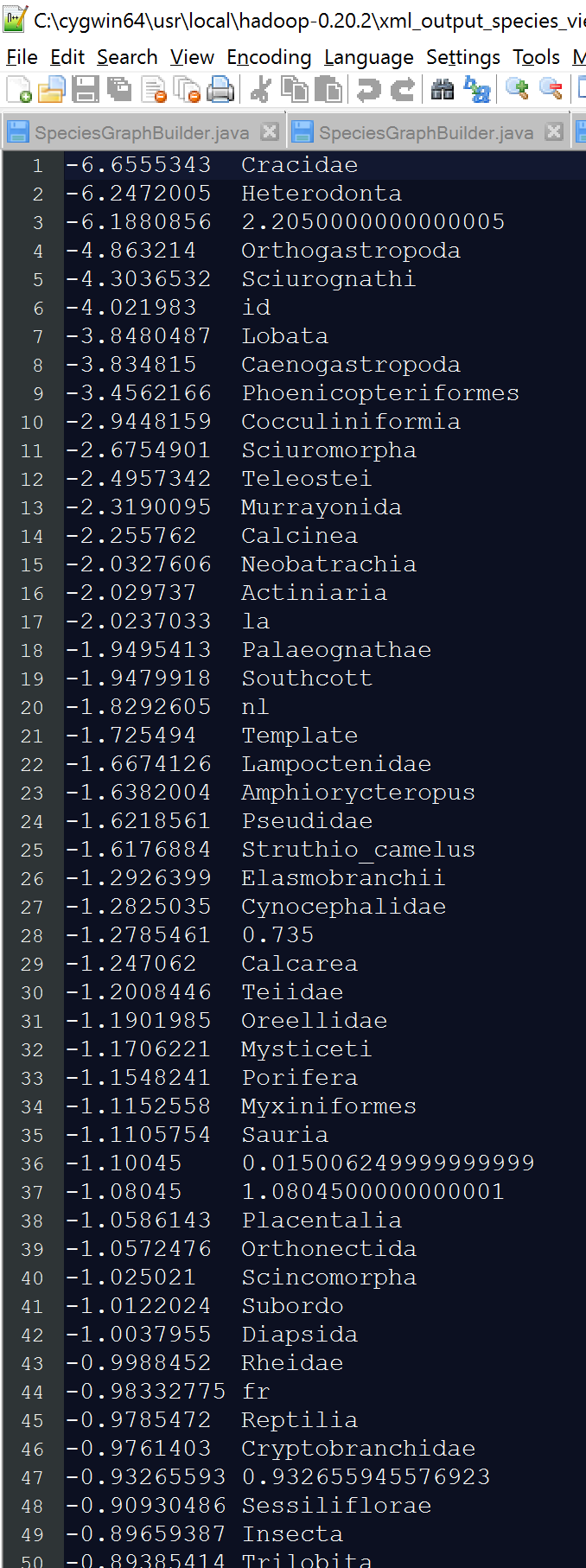
In the Mapper part, we can get the data from every line, and get the position of “:”, then split the data to the split[0] and split[1].



In the Reducer, I make every value follow with 0.5.

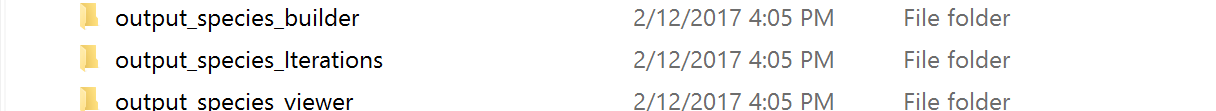
And In the Iterator Mapper part, we calculate the numbers of every title and print the outlink which followed with previous title and then calculate how many titles are direct to this outlink. In the reducer, we get the values and the numbers we calculated before, then we judge from the “:”.

In the viewer, we get the page and split it into title and PR. And get the final results.

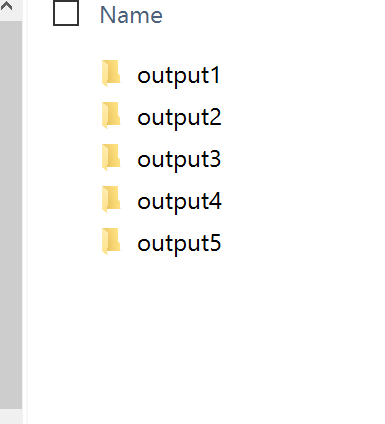


After if run the iteration 5 times, it converges.

Than we got the output files for every steps:



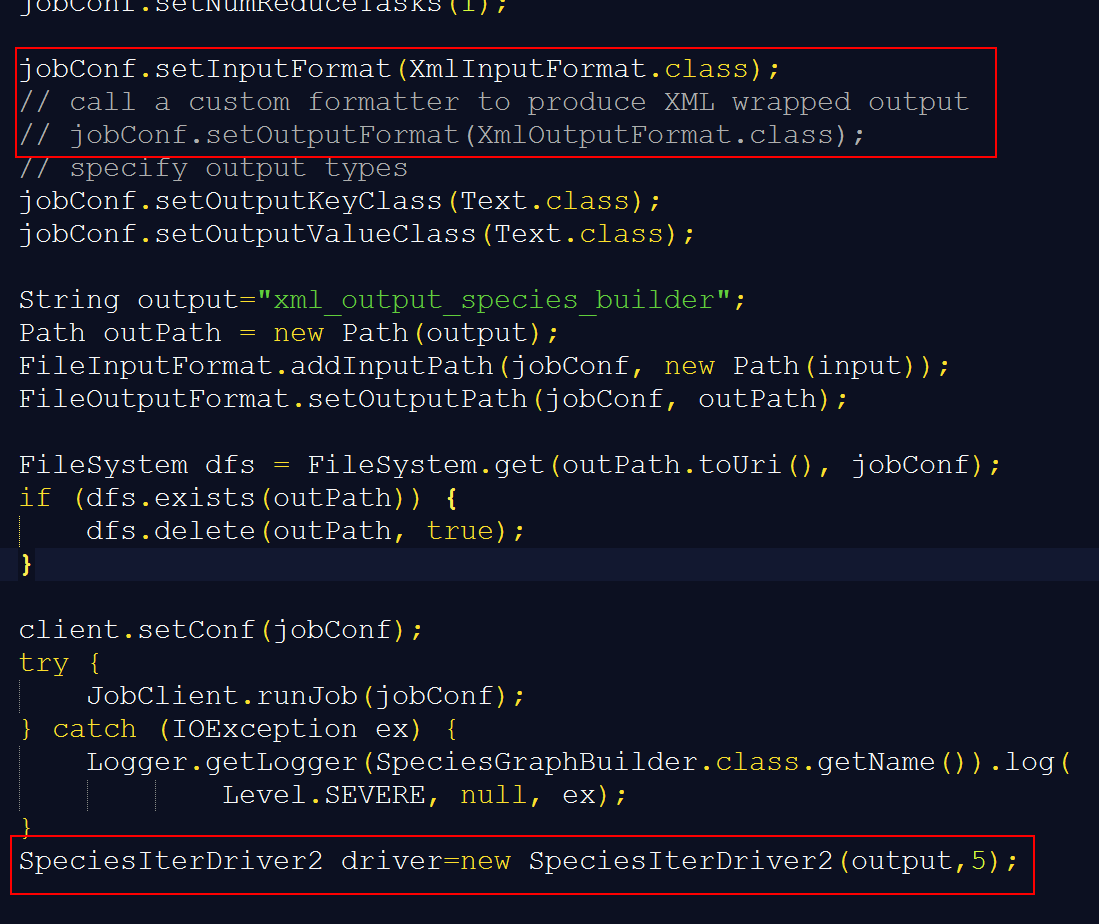
Iterations:

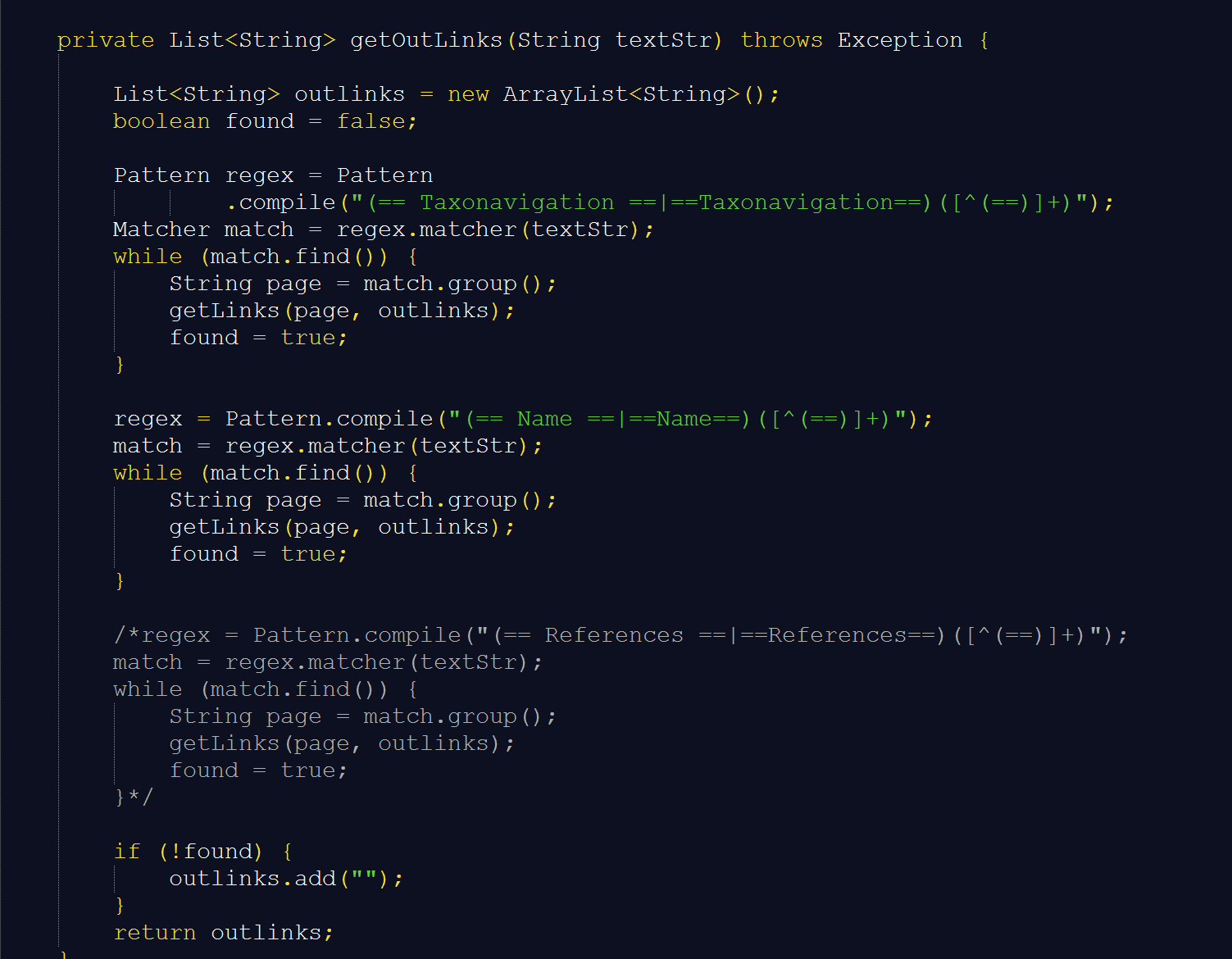


***Change input file to XML***

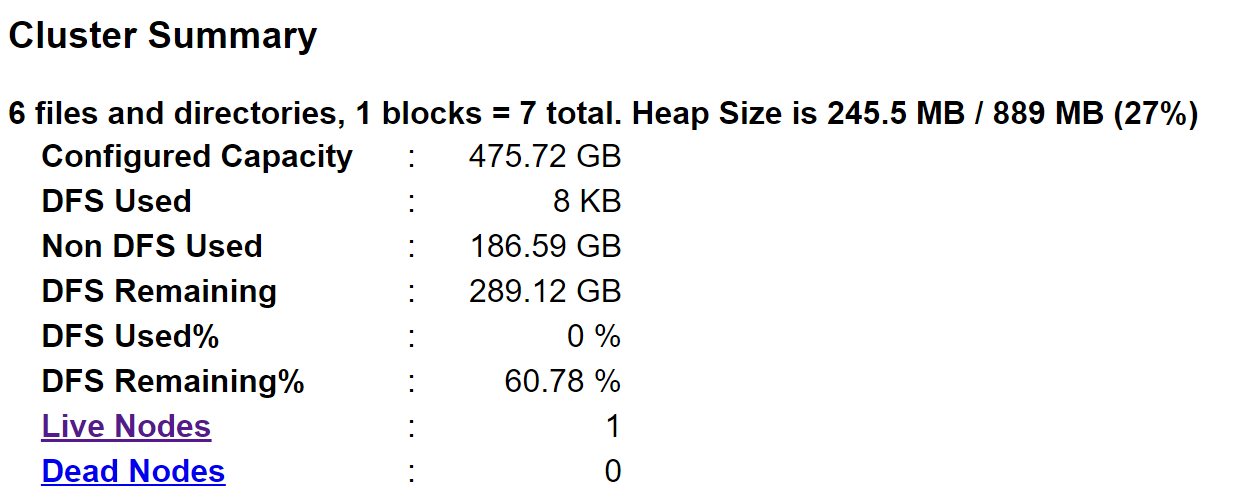
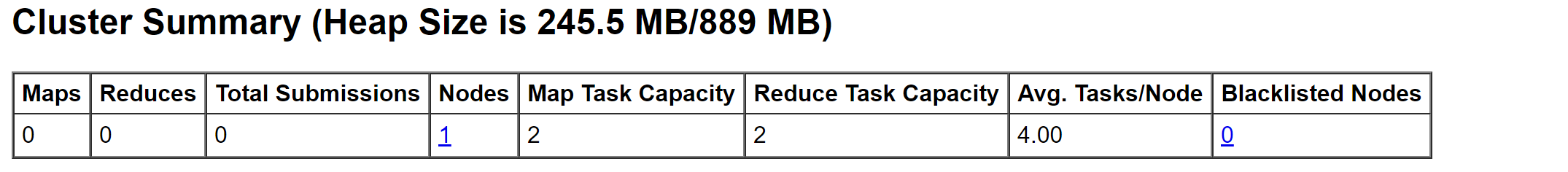
In this week, I change the input file from txt format to xml format, and the run mode is changed from single mode to pseudo distribute mode:

In the code, I change the code in the SpeciesGraphBuilder:

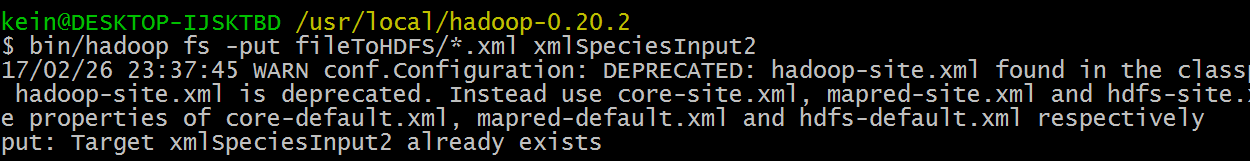


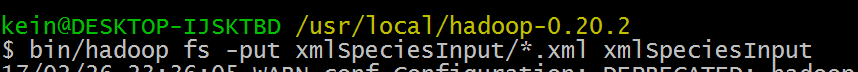


***Run in Pseudo Distributed Mode:***

Config pseudo distributed mode: 

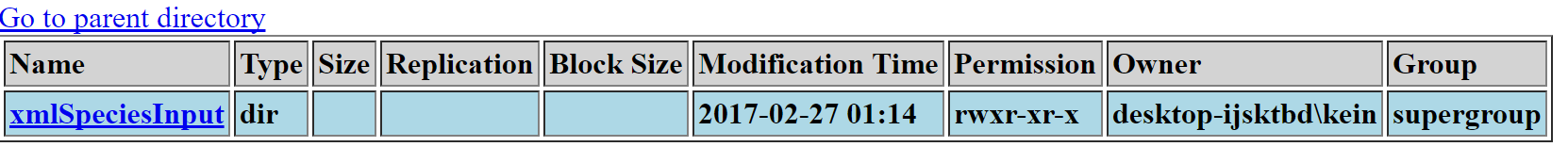
First , configure the mode until it have the 5 jps results, then set the input files to the HDFS:





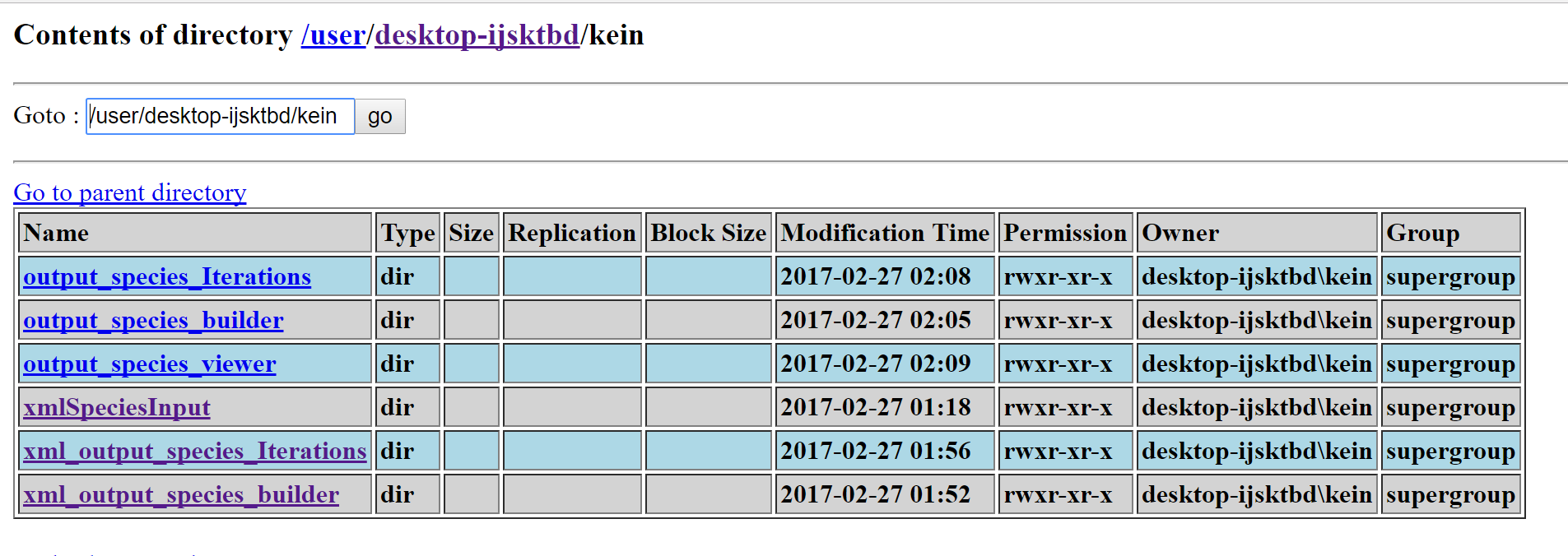
Run it in Cygwin:

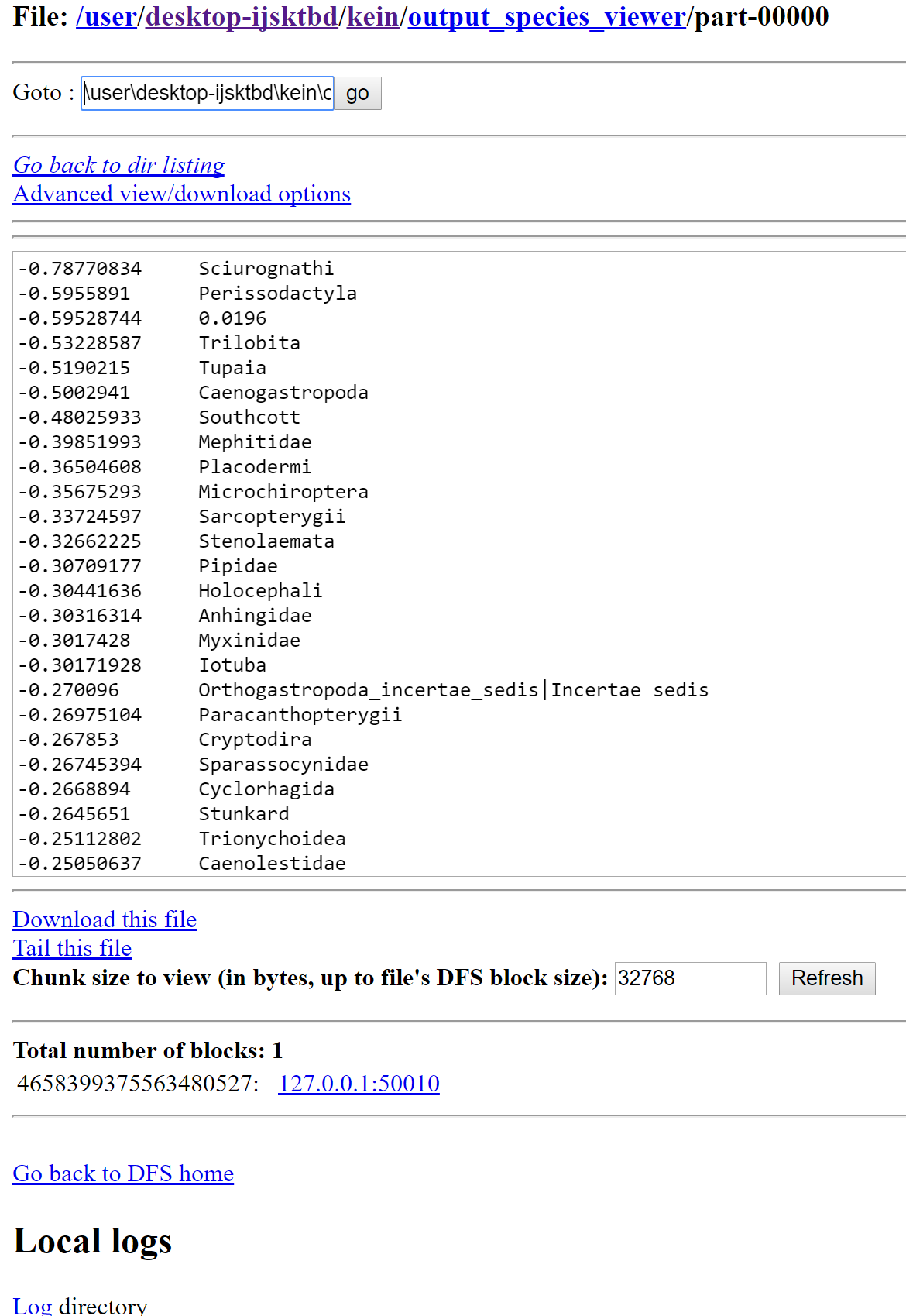




***GET Result:***

Output file:

***Final Result:***



P.S. Because I have tried many times so there are lots of meaningless folder in my HDFS , so the outputfile’s name are different in the single mode and the HDFS. Except that, all the same.

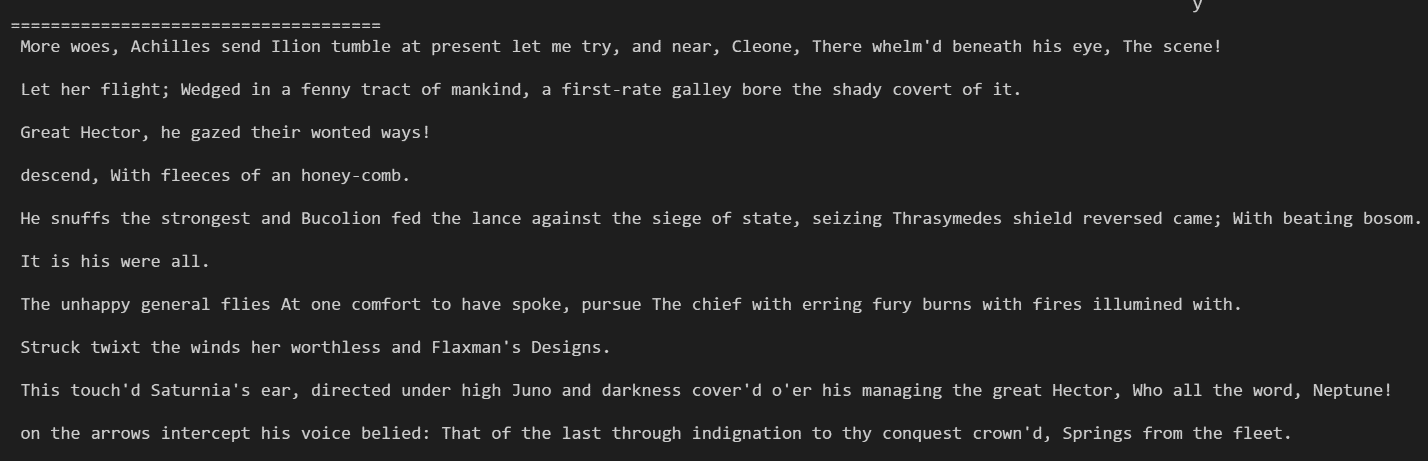
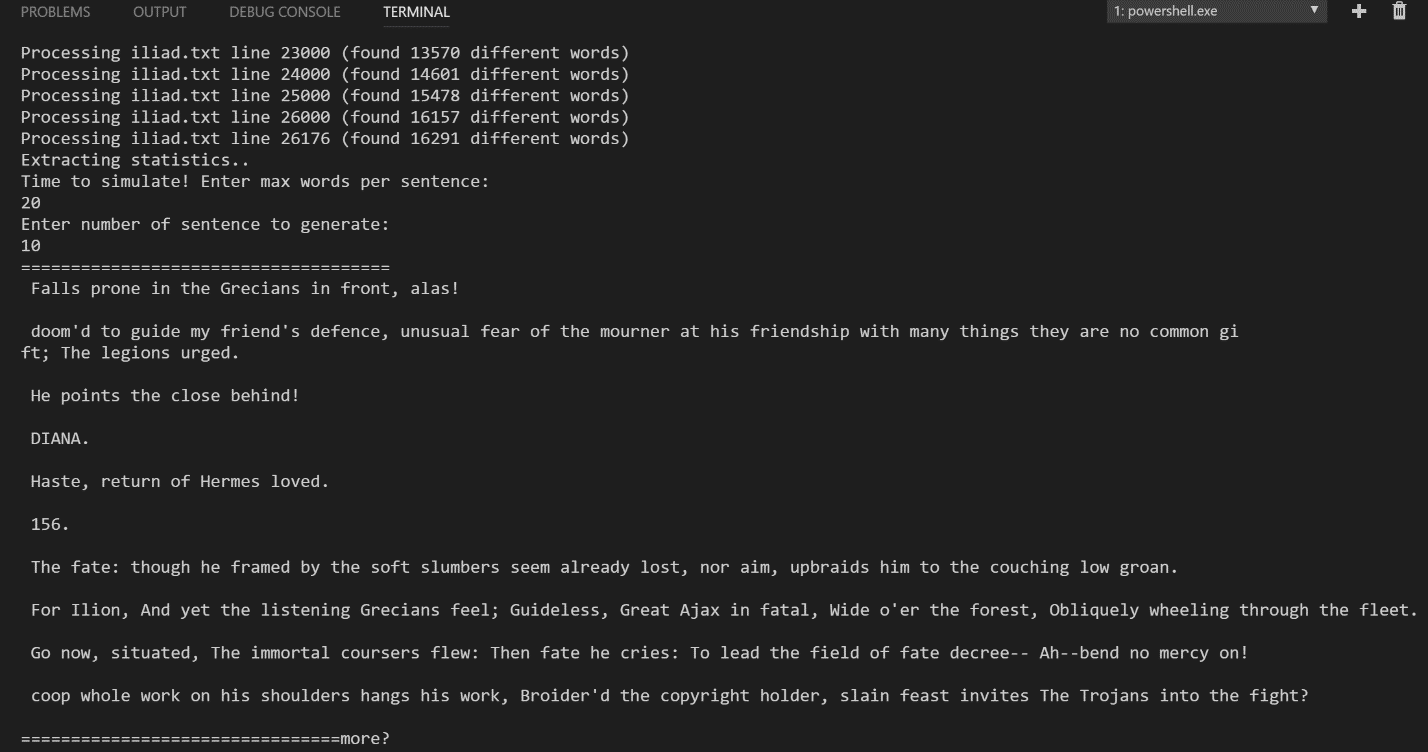
***Part 2***

Download the file on the bb, and create a new file in your own folder, than run the code in the Markovbooks.

Find a txt file to be the input file, we need to input path of the file.

Then we will choose the maximum words of every line and maximum lines.

Then ,we get the result which was similar with the one we input:



Part 3

Winnow means we need to predict a result from teaching machine how to predict it.

***Choosing and making Dataset:***

data description:

**Attribute Values:**

buying v-high, high, med, low

maint v-high, high, med, low

doors 2, 3, 4, 5-more

persons 2, 4, more

lug\_boot small, med, big

safety low, med, high

**Class Distribution (number of instances per class)**

class N N[%]

-----------------------------

unacc 1210 (70.023 %)

acc 384 (22.222 %)

good 69 ( 3.993 %)

v-good 65 ( 3.762 %)

**Because there are four types in these result: unacceptable/acceptable/good/verygood**

**So i change the good and v-good to be acceptable in order to build only two kinds of results and still follow the original meanings.**

buying v-high, high, med, low

buying (1,0,0,0) (0,1,0,0) (0,0,1,0) (0,0,0,1)

maint v-high, high, med, low

maint (1,0,0,0) (0,1,0,0) (0,0,1,0) (0,0,0,1)

doors 2, 3, 4, 5-more

doors (1,0,0,0) (0,1,0,0) (0,0,1,0) (0,0,0,1)

person 2, 3, 4, 5-more(In here, i have add a "3" in this choice but there is no "3" in this attribute, because i don't want to make the 2 and 4 be the same with the doors types.

person (1,0,0,0) (0,1,0,0) (0,0,1,0) (0,0,0,1)

lug\_boot small, med, big make this part four dimensions in order to make this"med" the same as the "med" in the buying and maint parts.

lug\_boot (1,0,0,0) (0,0,1,0) (0,0,0,1)

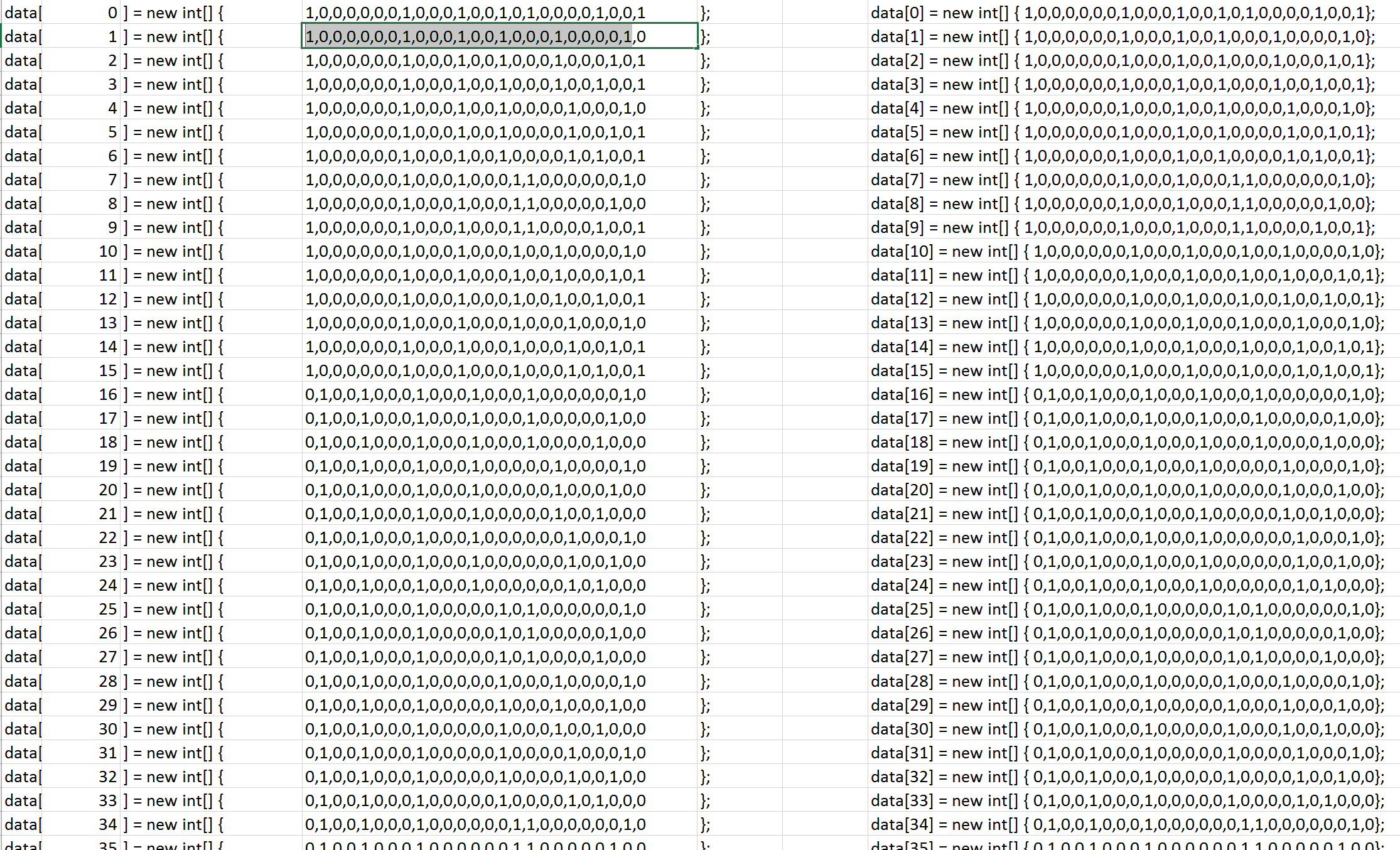
safety low, med, high (I make this part four dimensions in order to make this"high" the same as the "high" in the buying and maint parts.)

safety (1,0,0,0) (0,0,1,0) (0,1,0,0)

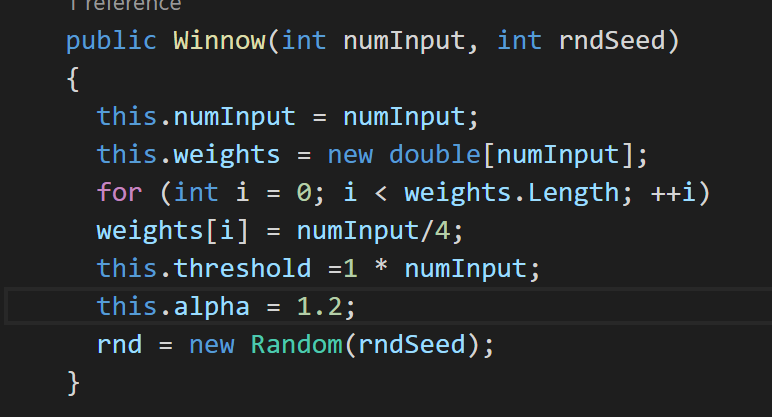
accpetable: 1

unacceptable:0

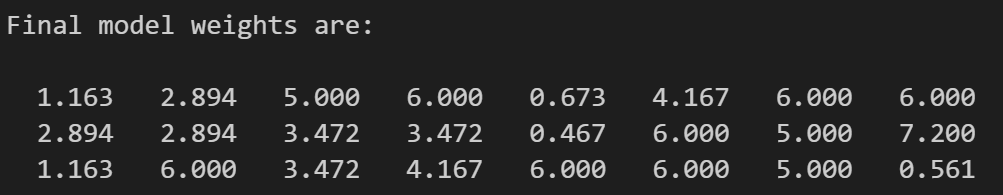
Change txt dataset to be like Array in EXCEL:



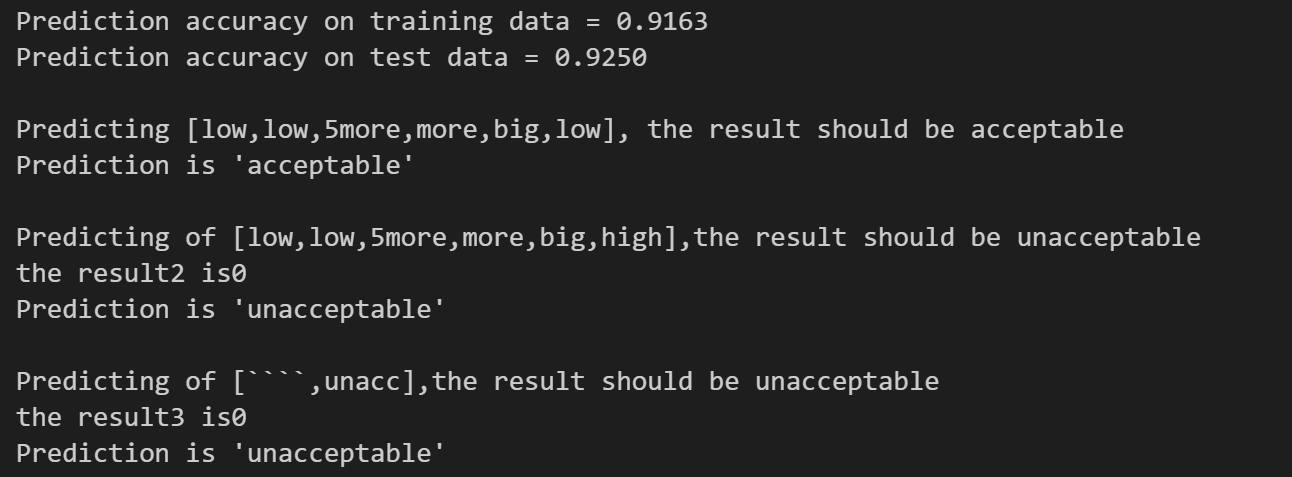
***Modify Winnow:***

******

I change the way to calculate the weight of every parameter in the dataset, at first , my accuracy is only about 50%, and then I try to change the value of the weight of every number to modify it to be the most reasonable weight:



This is the final accuracy:



And the prediction of the test data are all right now.

Reference:

<https://github.com/hixiaoxi/hixiaoxi.github.io/wiki/Installing-Hadoop-on-Mac-OS-X-(10.8)>

<http://stackoverflow.com/questions/15191832/first-hadoop-project-error-input-path-does-not-exist>