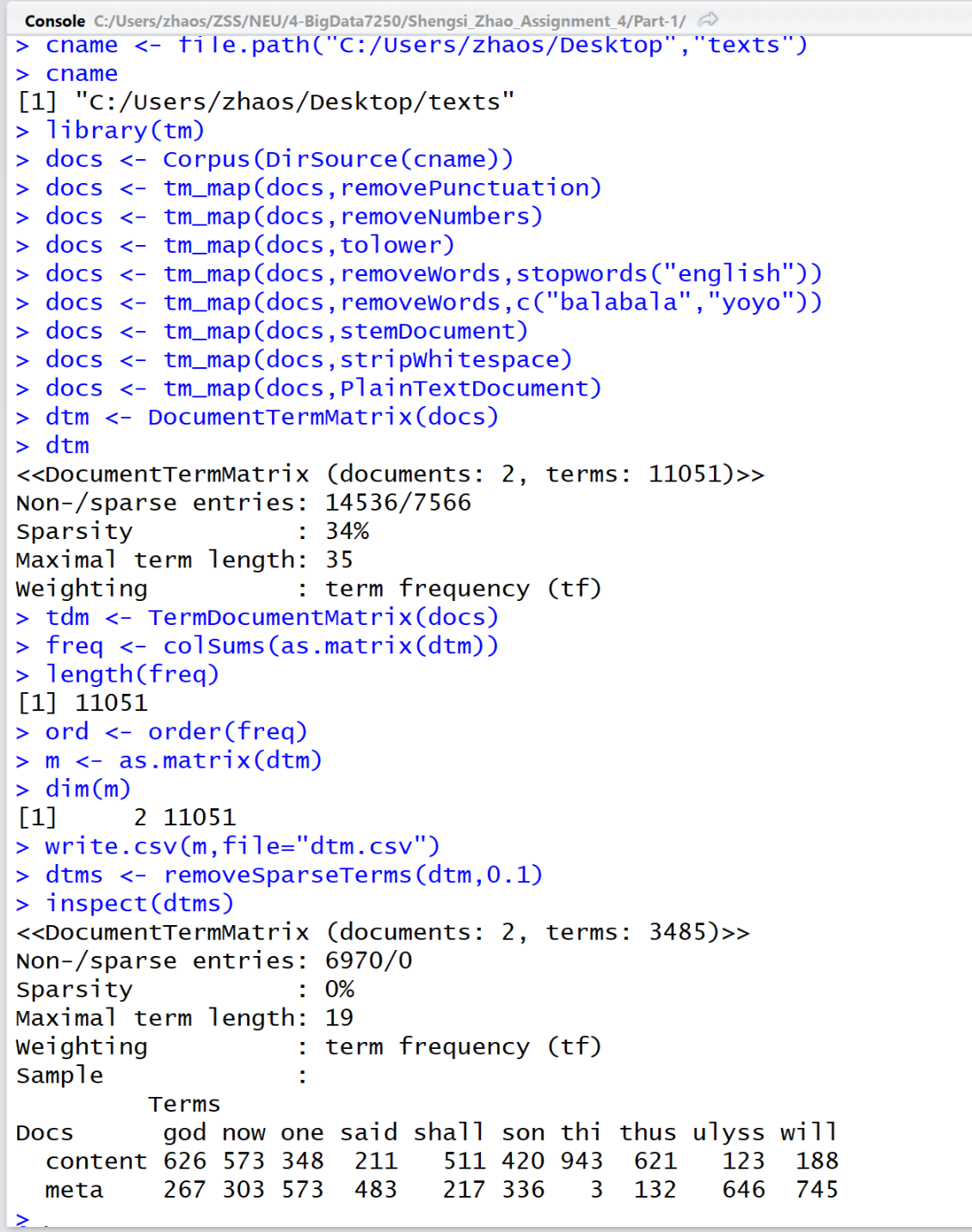
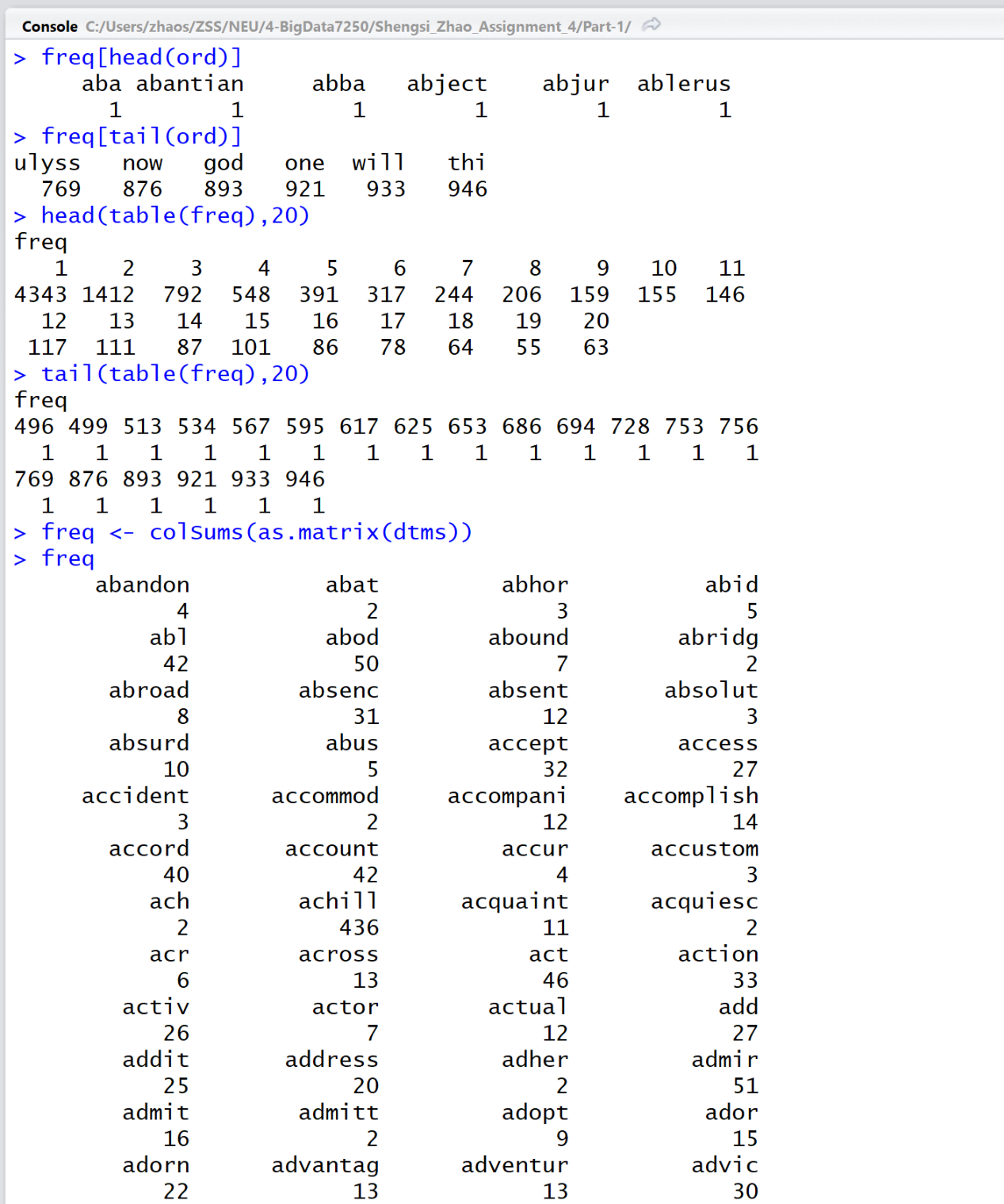
Summary Report

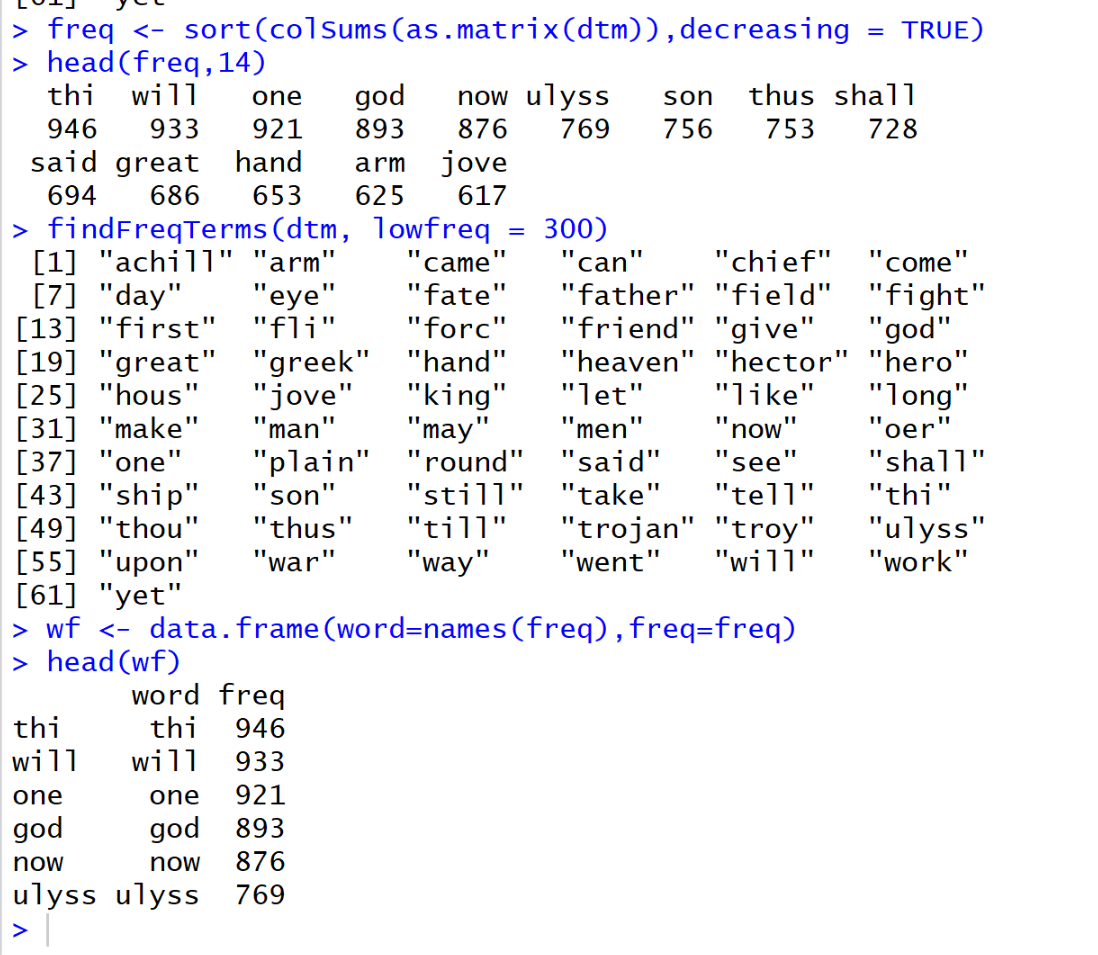
Homework 4

**PART-1:**

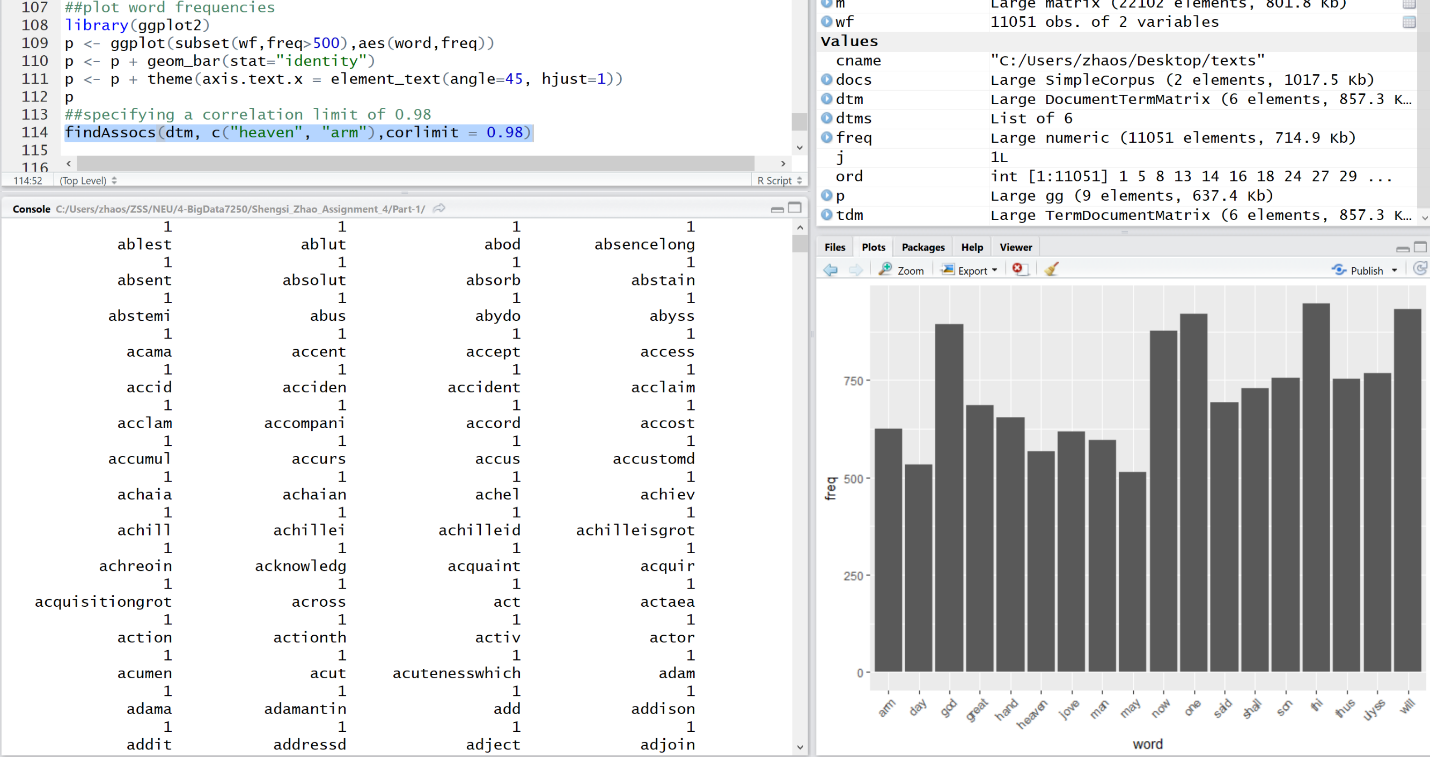
Finish the Text Mining in R Lab and generate the Word Cloud as well as Clustering graph.



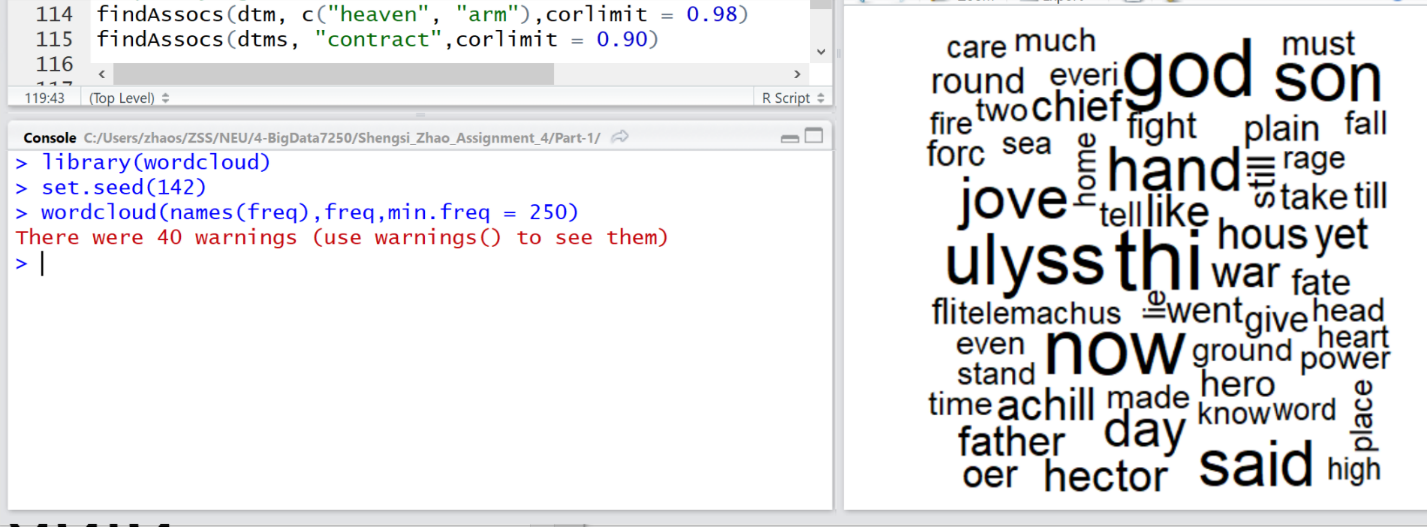




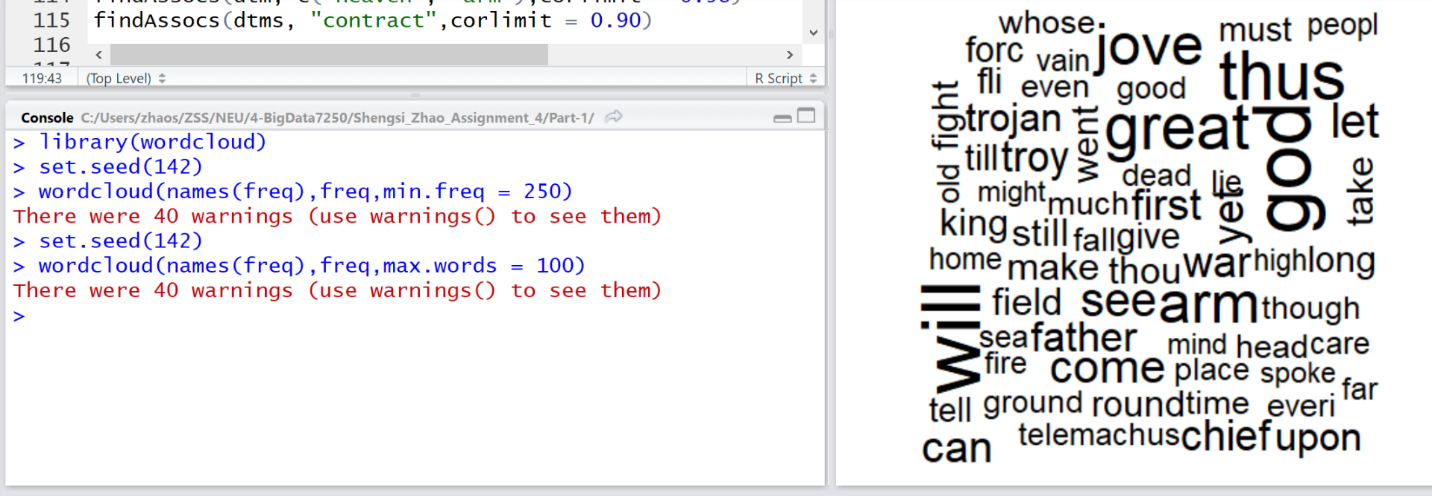
Here I choose the freq>500 to show the plot word frequencies.



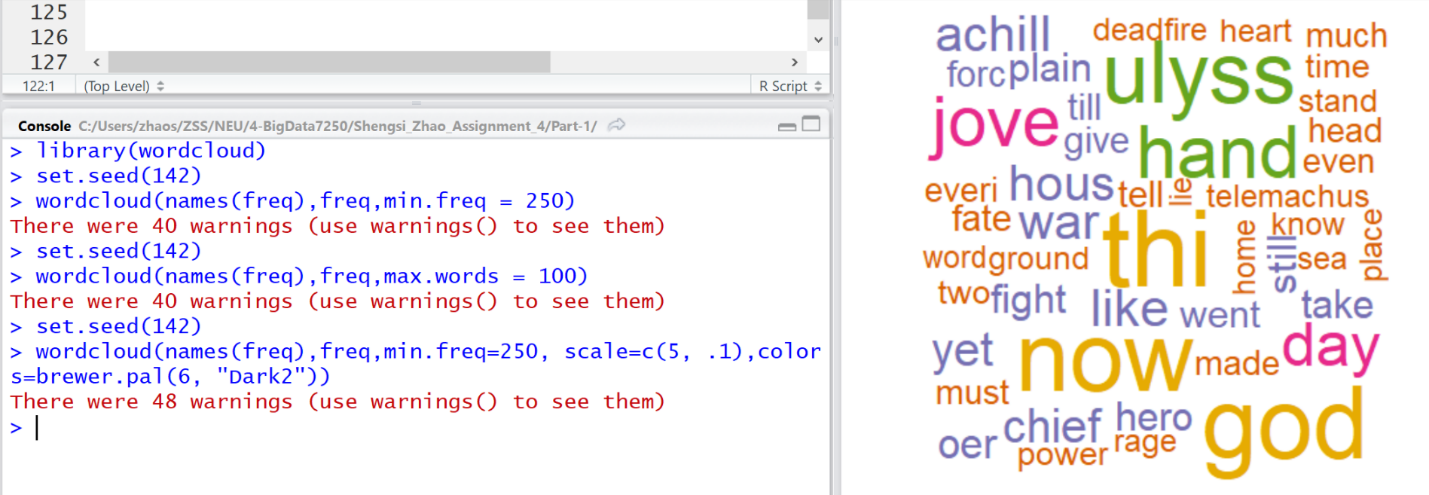
Plot words that occur at least 250 times.



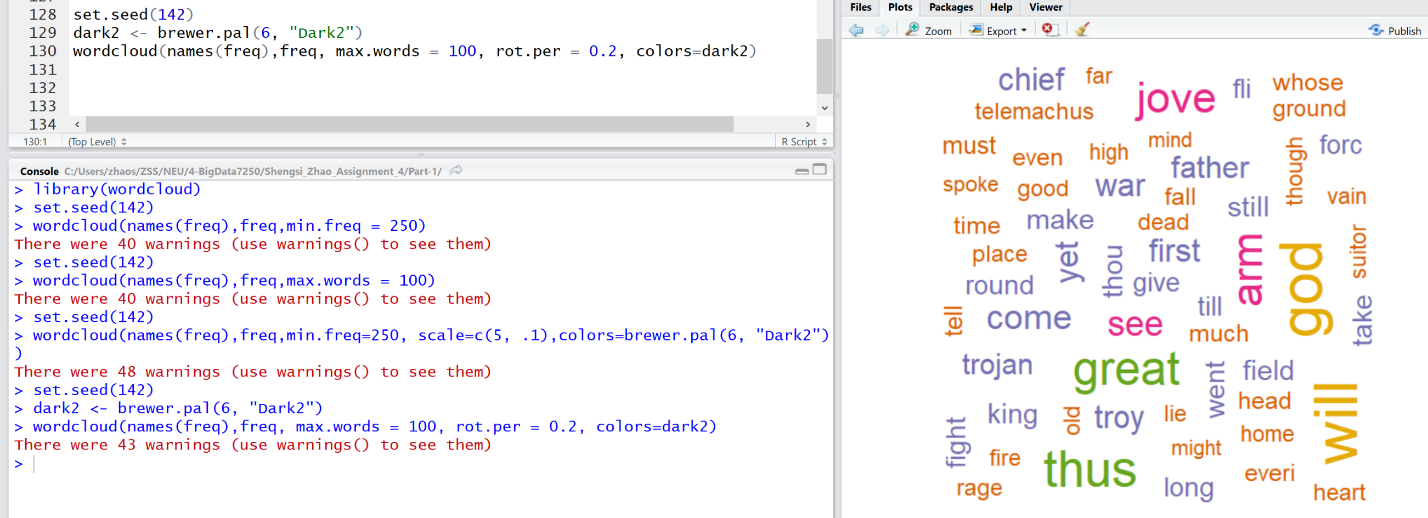
Plot the 100 most frequently used words.



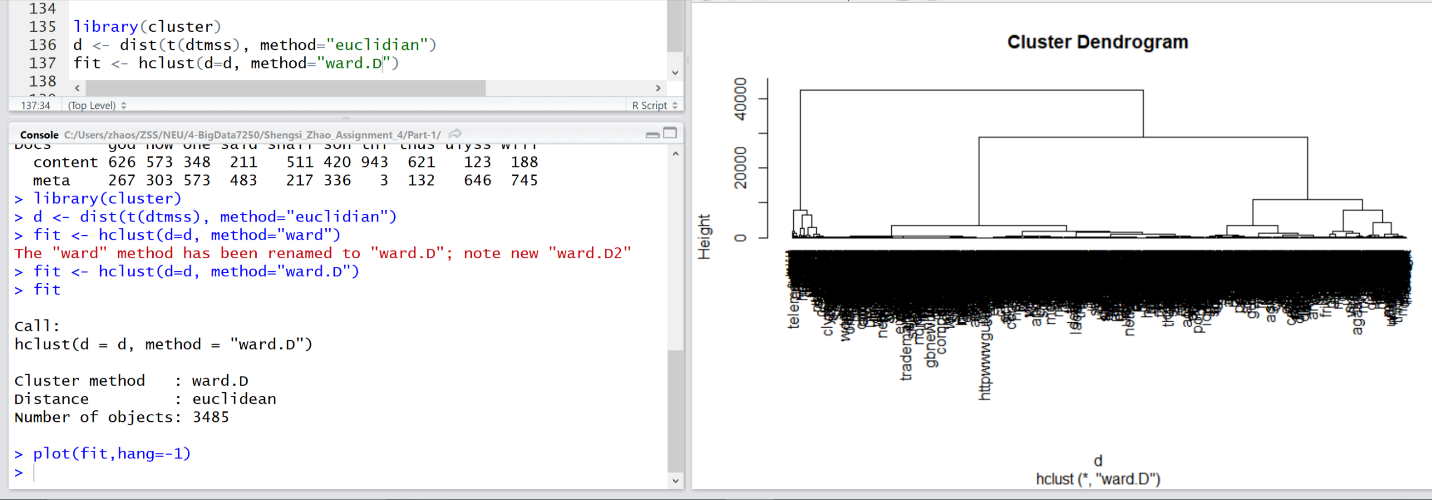
Add some color and plot words occurring at least 250 times.

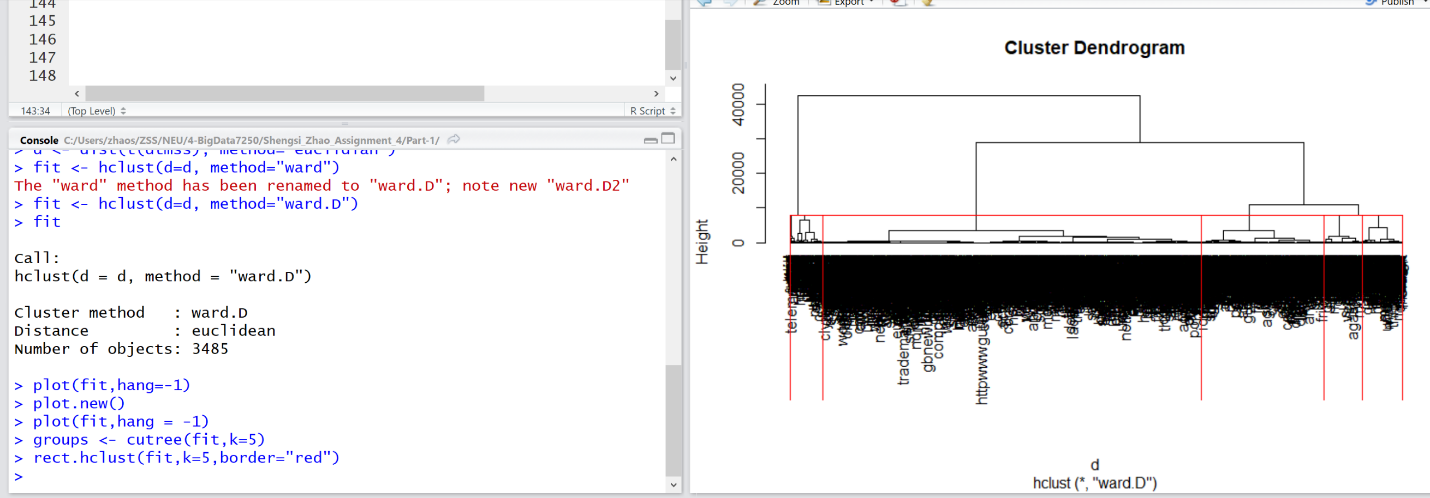


Plot the 100 most frequently occurring words.

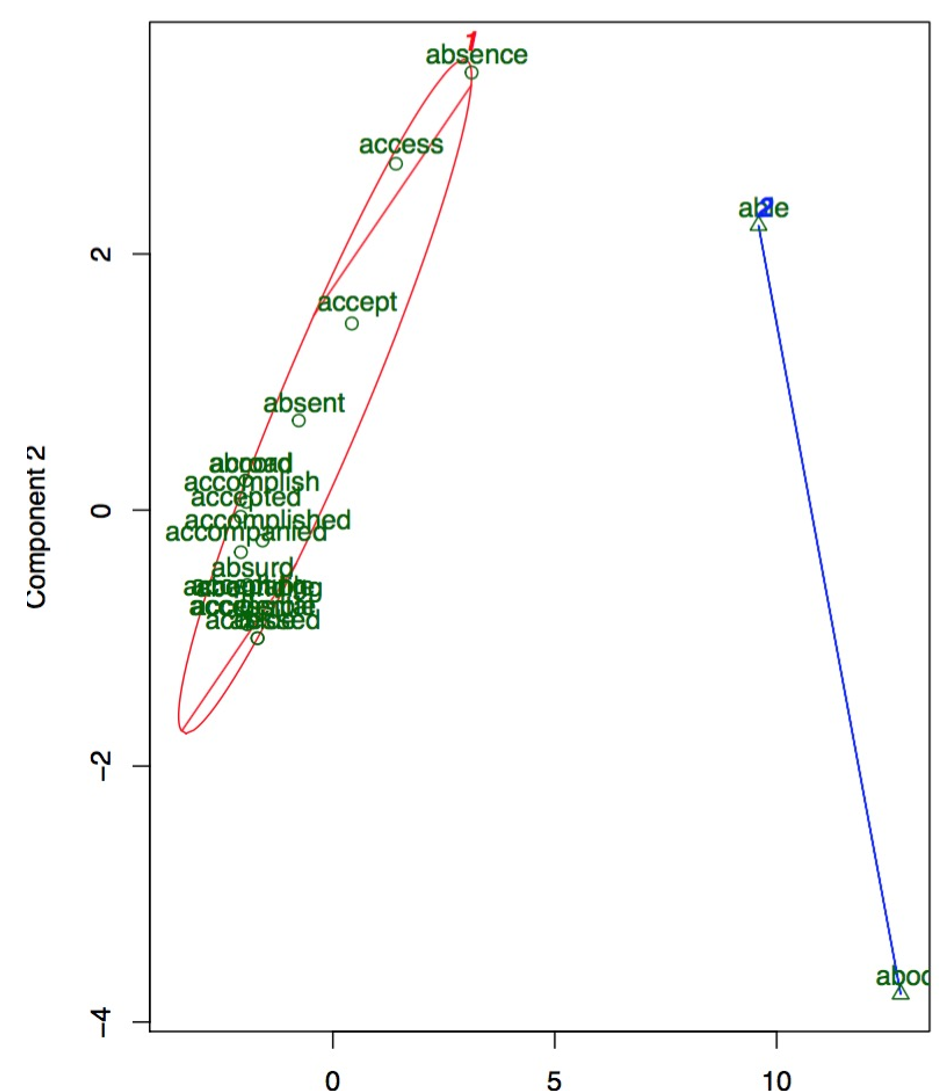


Hierarchal Clustering.





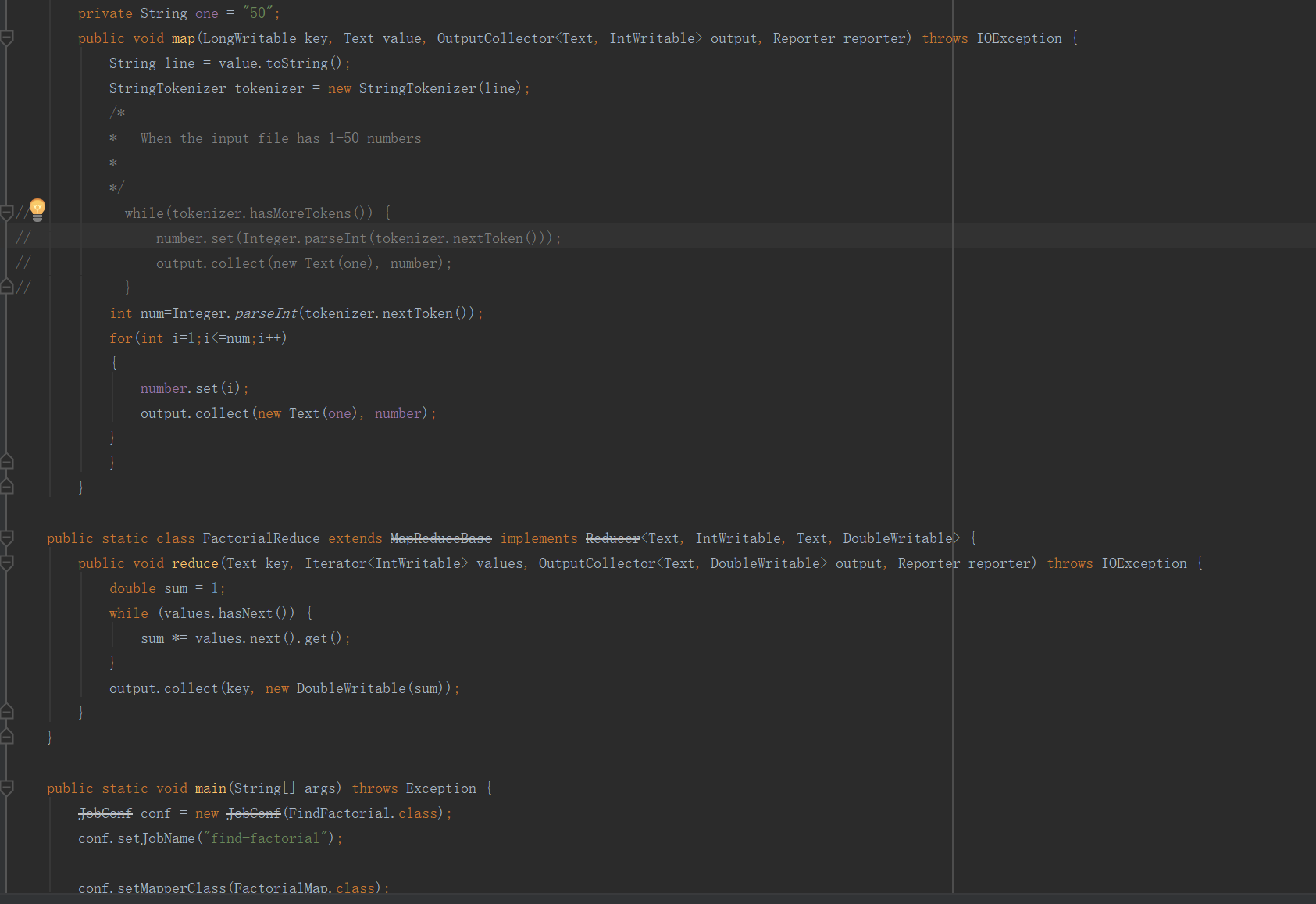
Cluster:



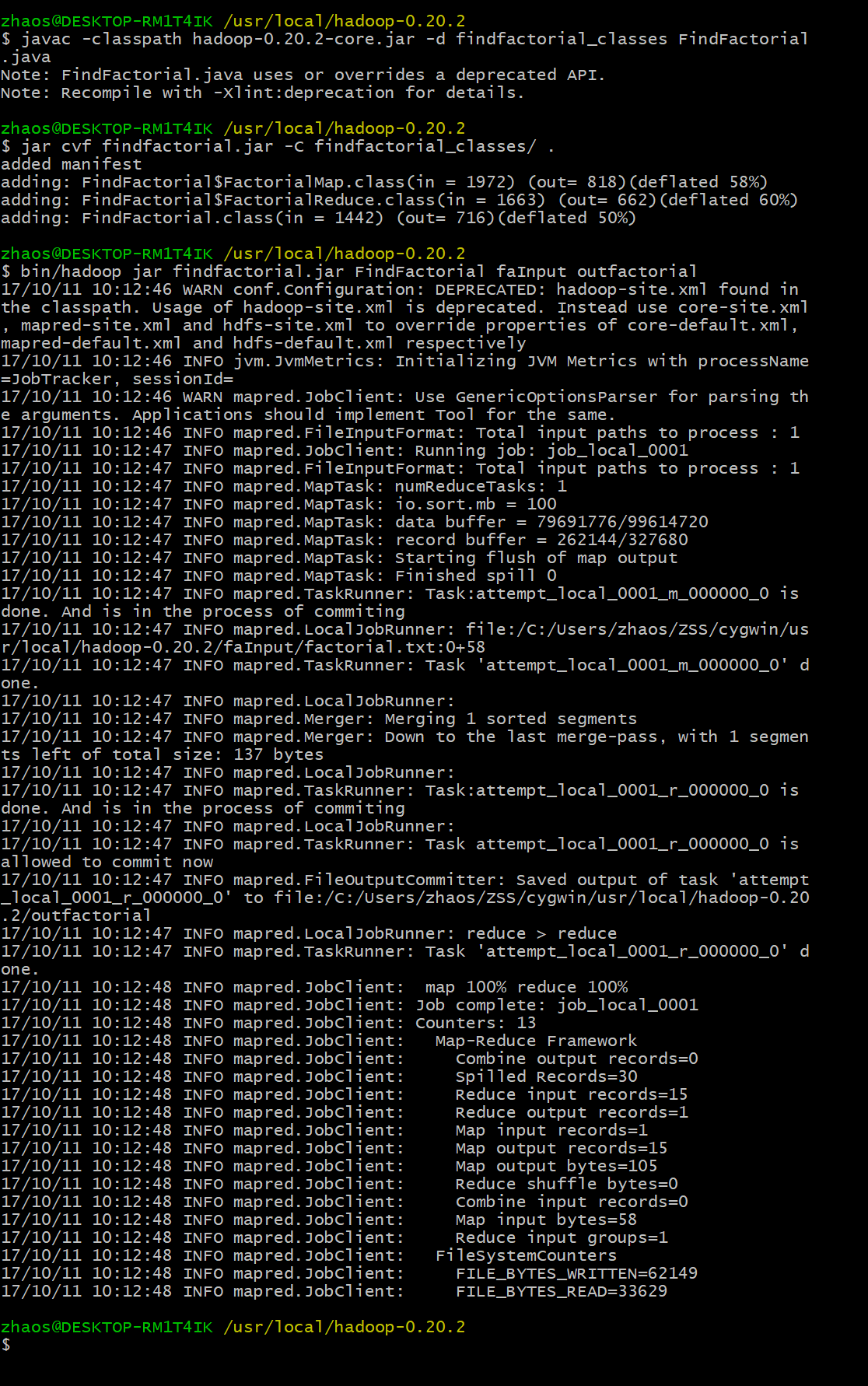
**PART-2:**

Find the Factorial product in Hadoop way. Considering the scale of this problem, you don't have to multiply exactly from 1 to a million. A smaller number, like 50, would be good enough as long as you are using the correct methodology.

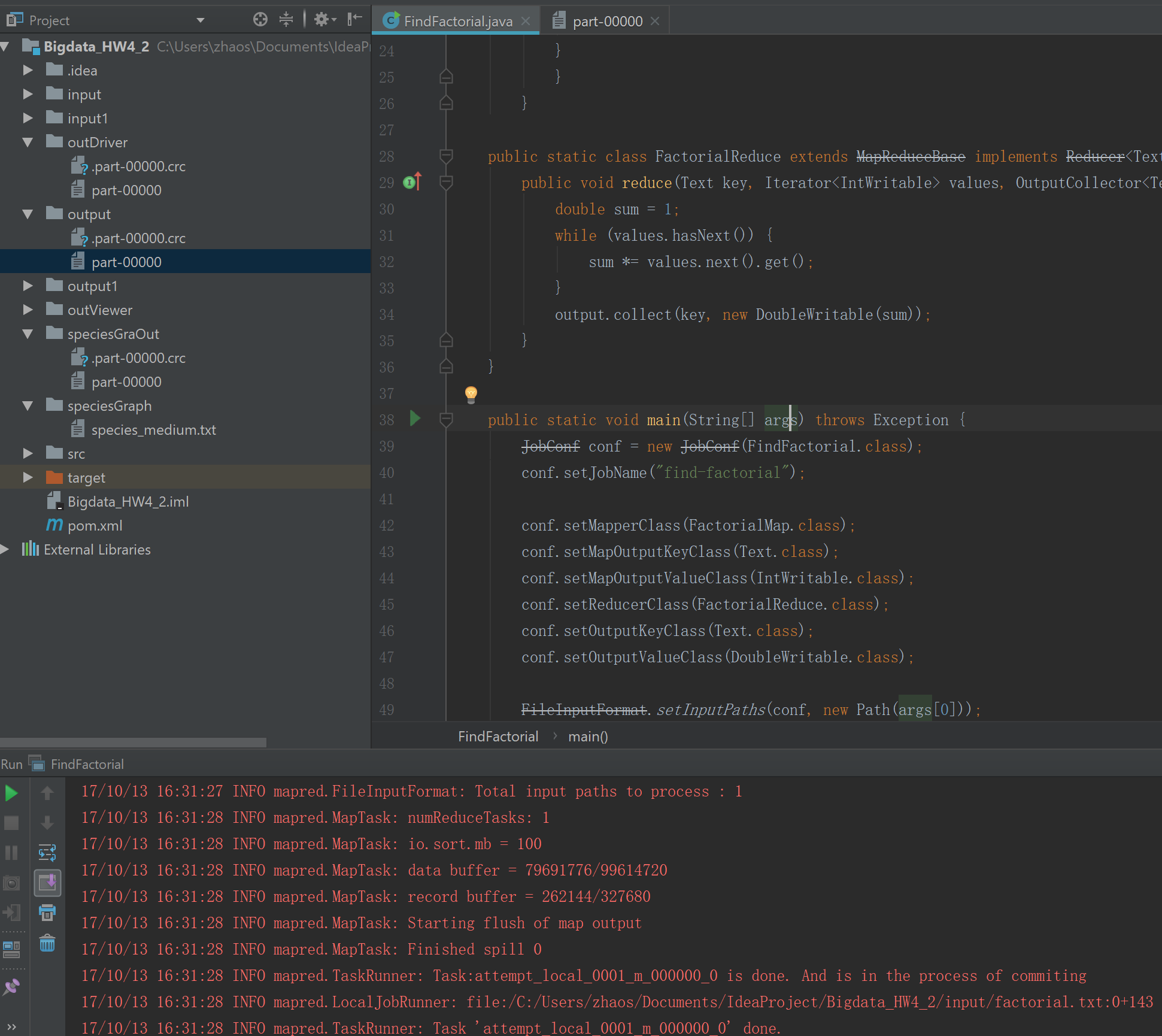
The code:



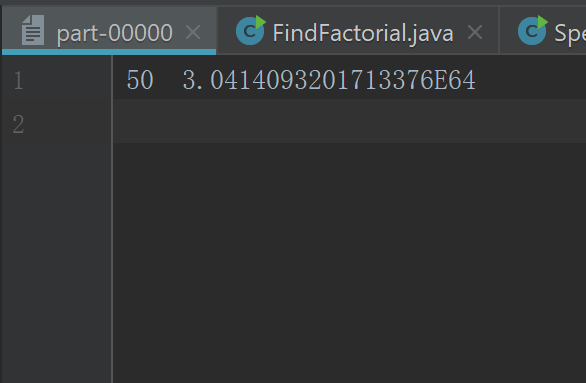
Run in hadoop



Run in IntelliJ:



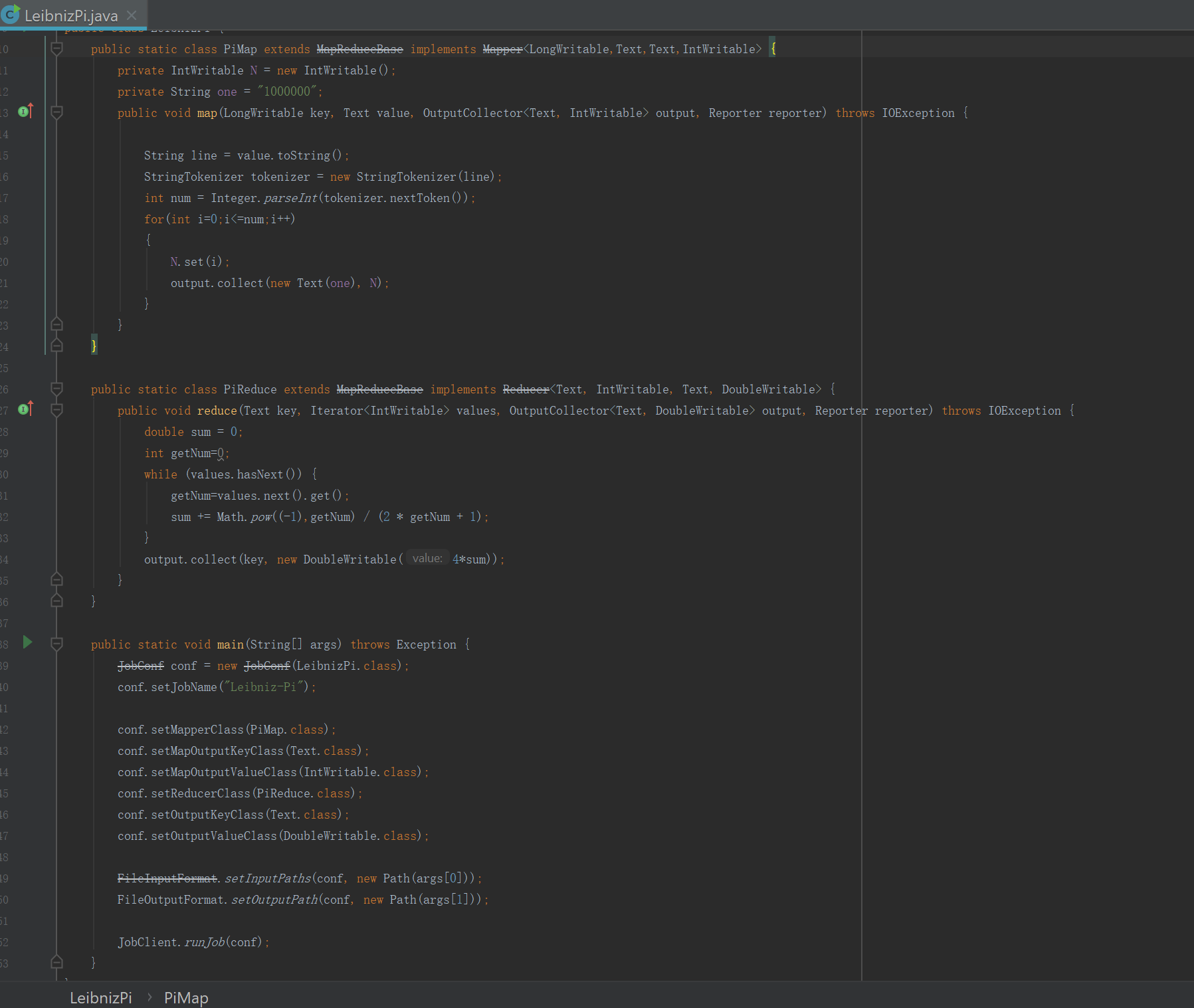
The result:



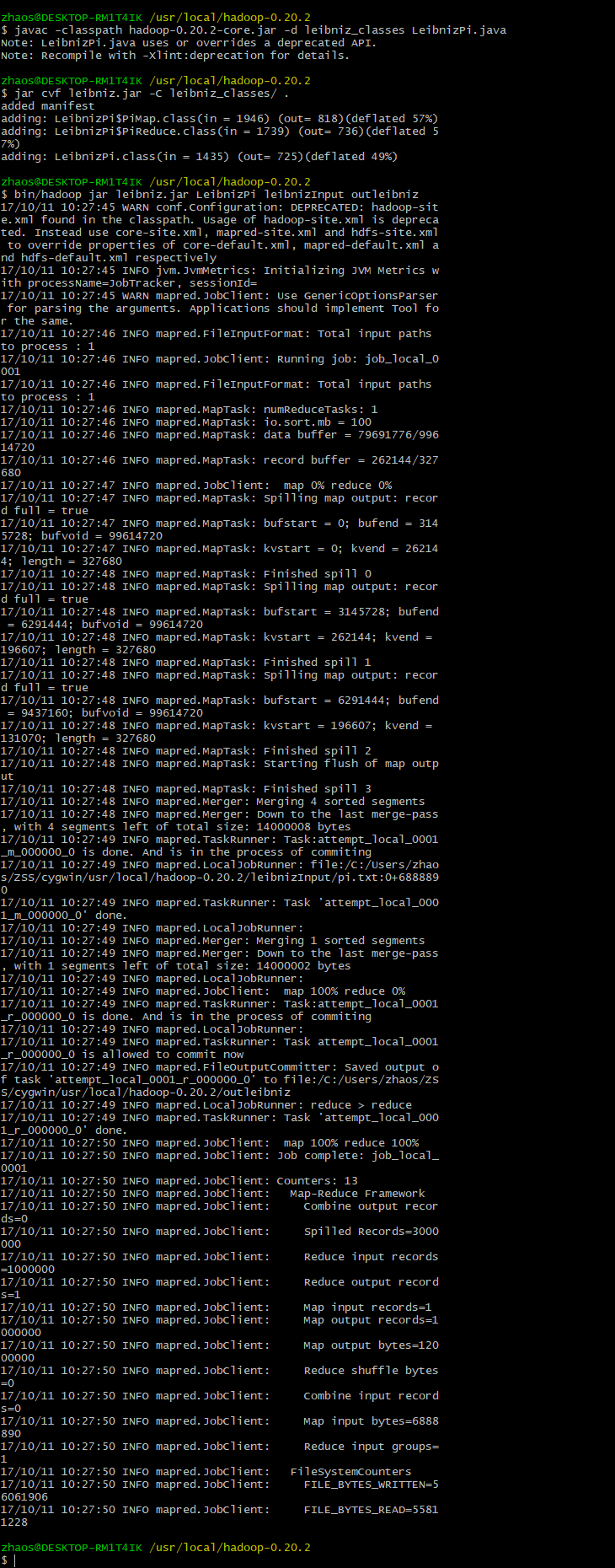
**PART-3:**

Compute Pi using Leibniz formula in Hadoop way. Choose appropriate N and make sure the result converges.

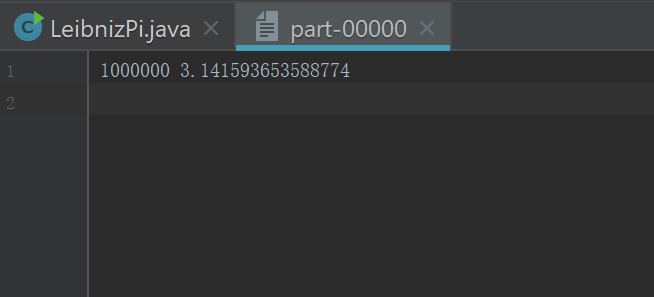
The code:



Run in Hadoop:



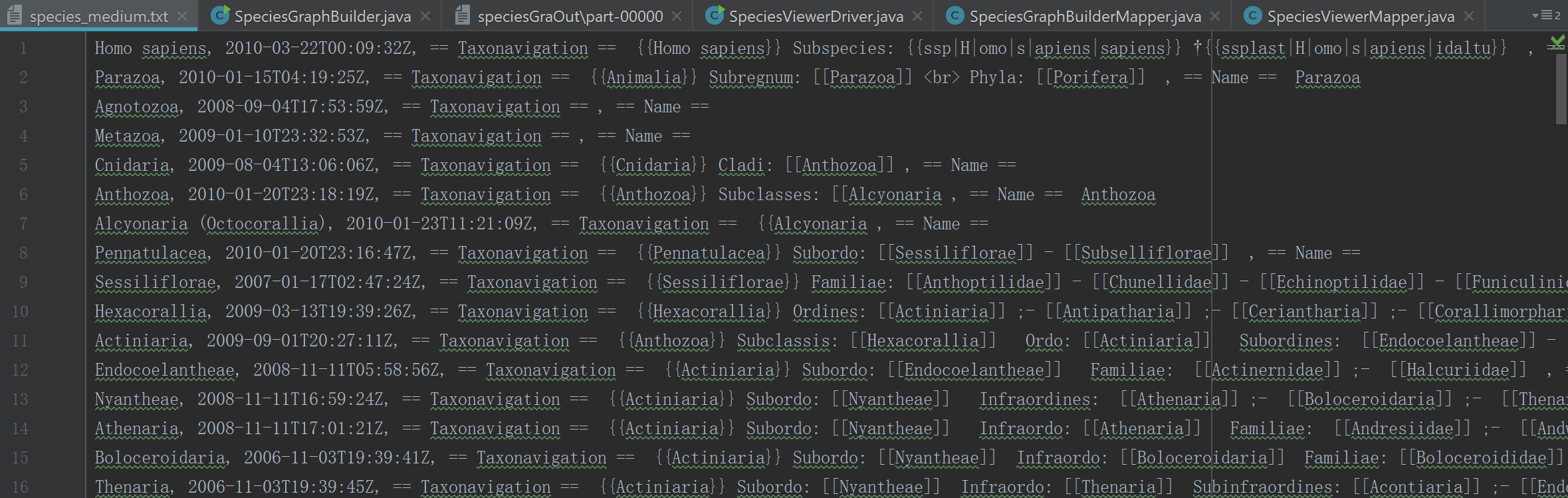
The output:



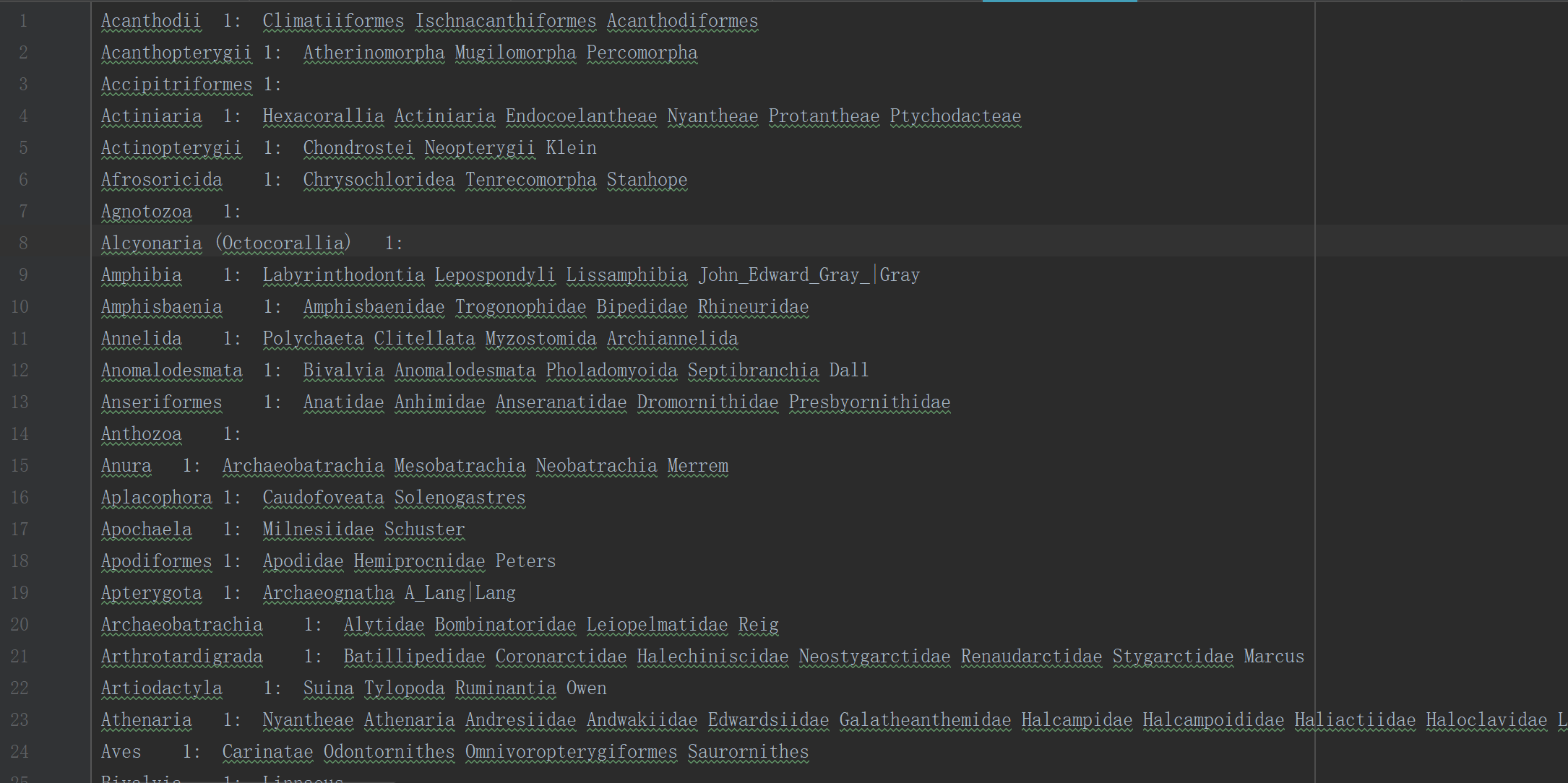
**PART-4:**

PageRank the WikiSpecies websites using the code that Professor provided and the sharded species\_medium.txt data. Please be careful the code might contain errors and bugs. You are expected to find out the abnormality from the output and explain why it happens.

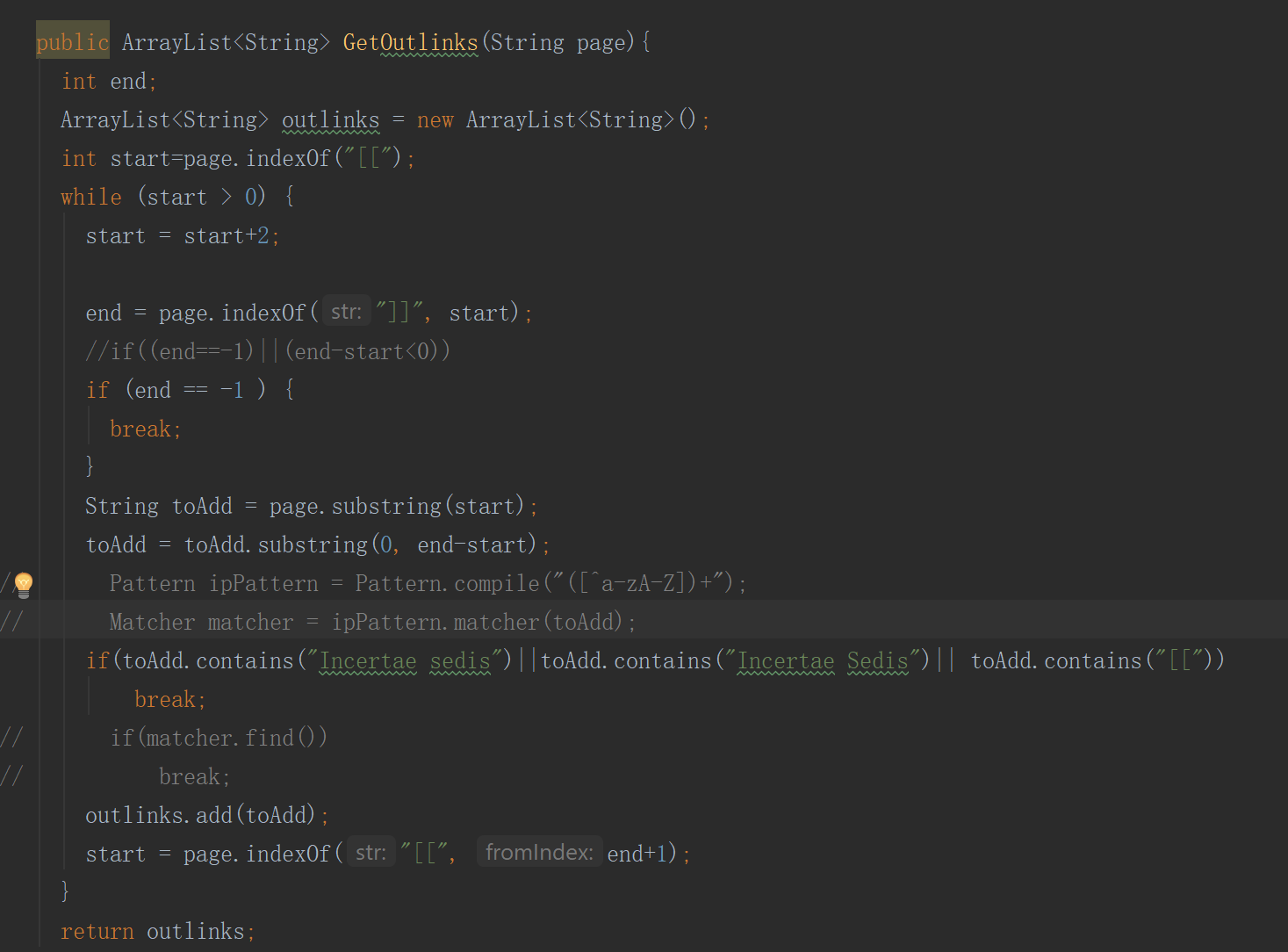
The input file:



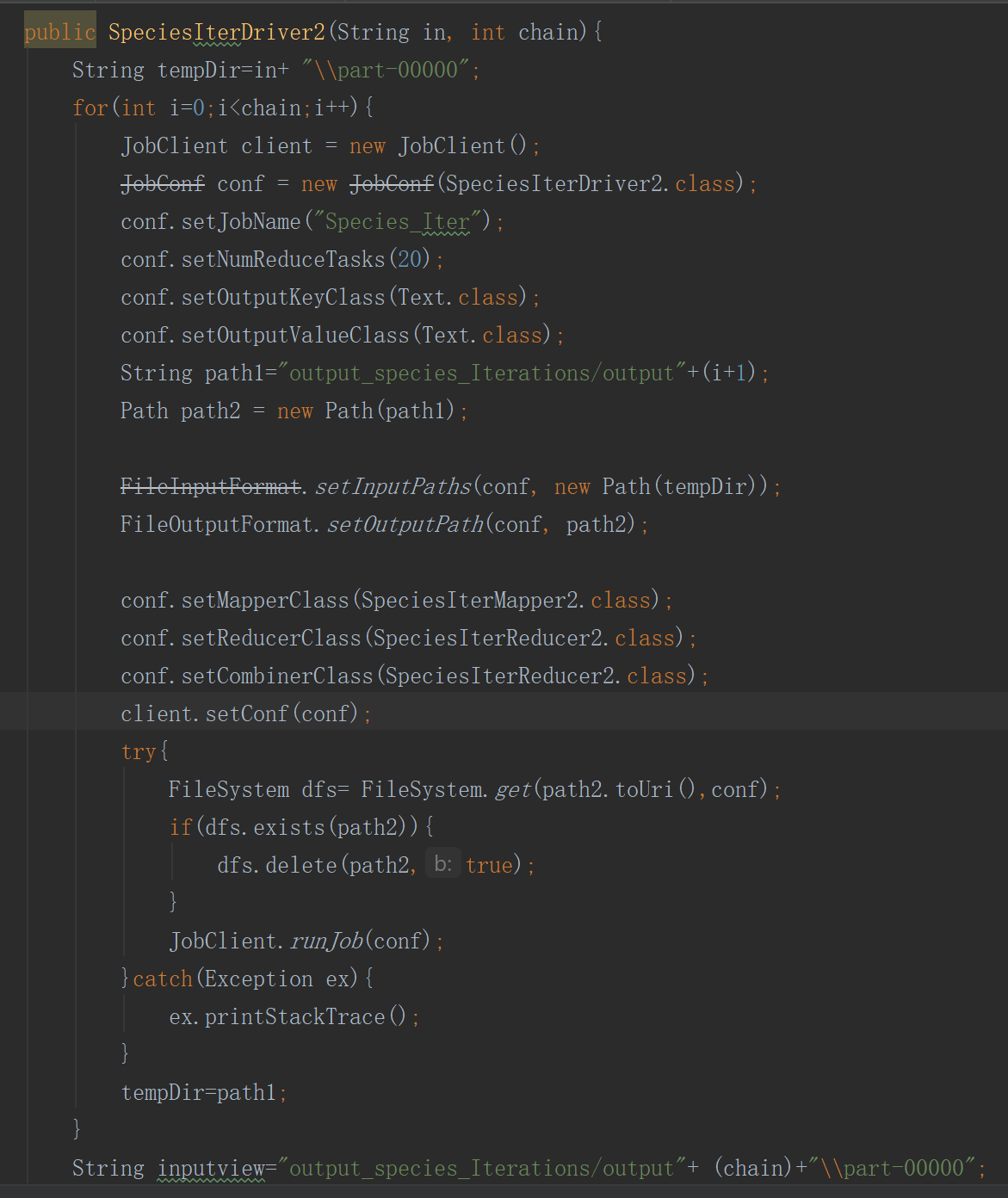
The output file of SpeciesGraphBuilder:



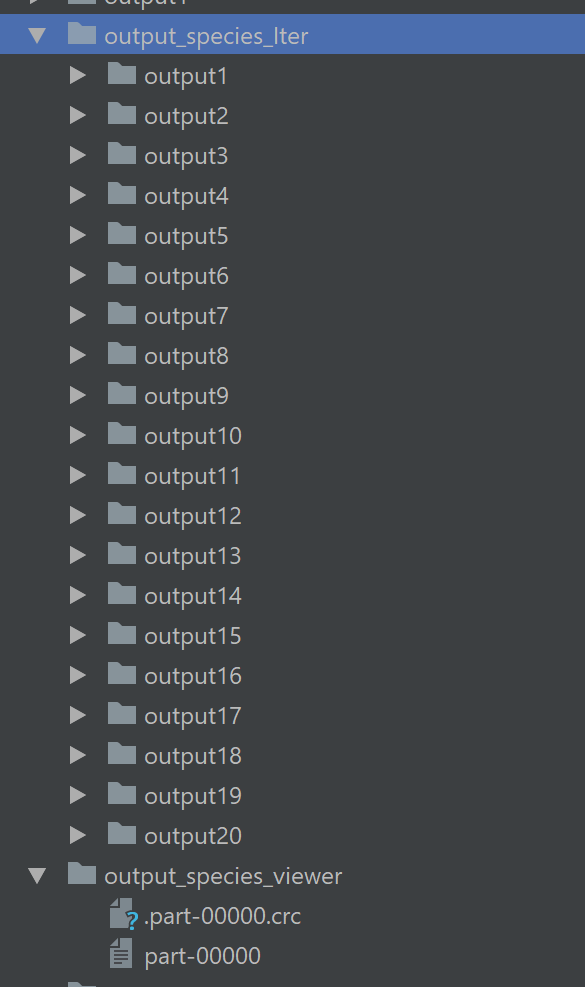
In SpeciesGraphBuilderMapper, I changed like that to reduce the error in output file, because in original mapper, it will add some name not in “[[]]”, but in “[[”, and have some unnamed species, so I made a filter.



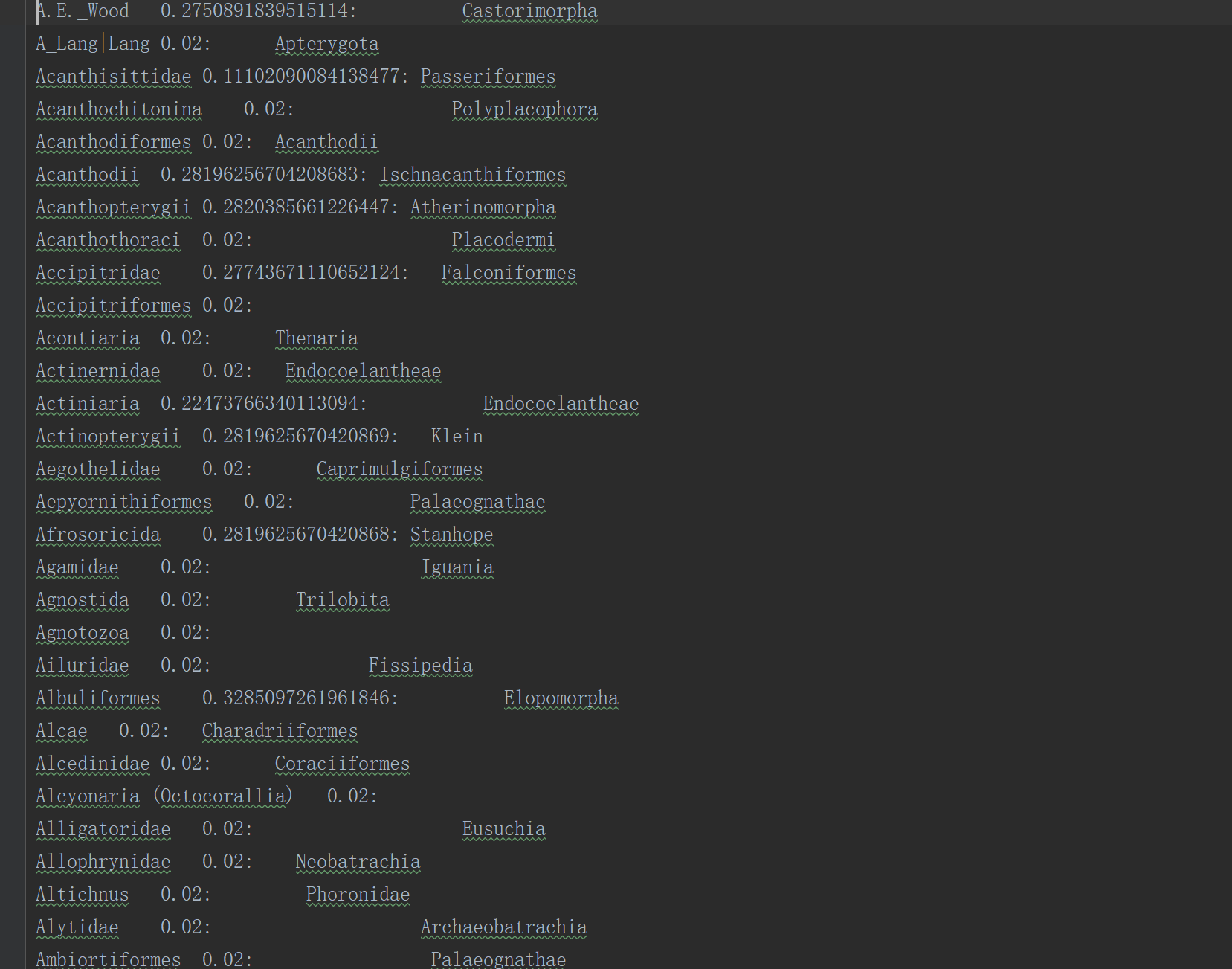
In SpeciesIterDriver2, I changed the code to let it run 20 times, and run the SpeciesViewerDriver by it.



The output file of SpeciesIterDriver2:



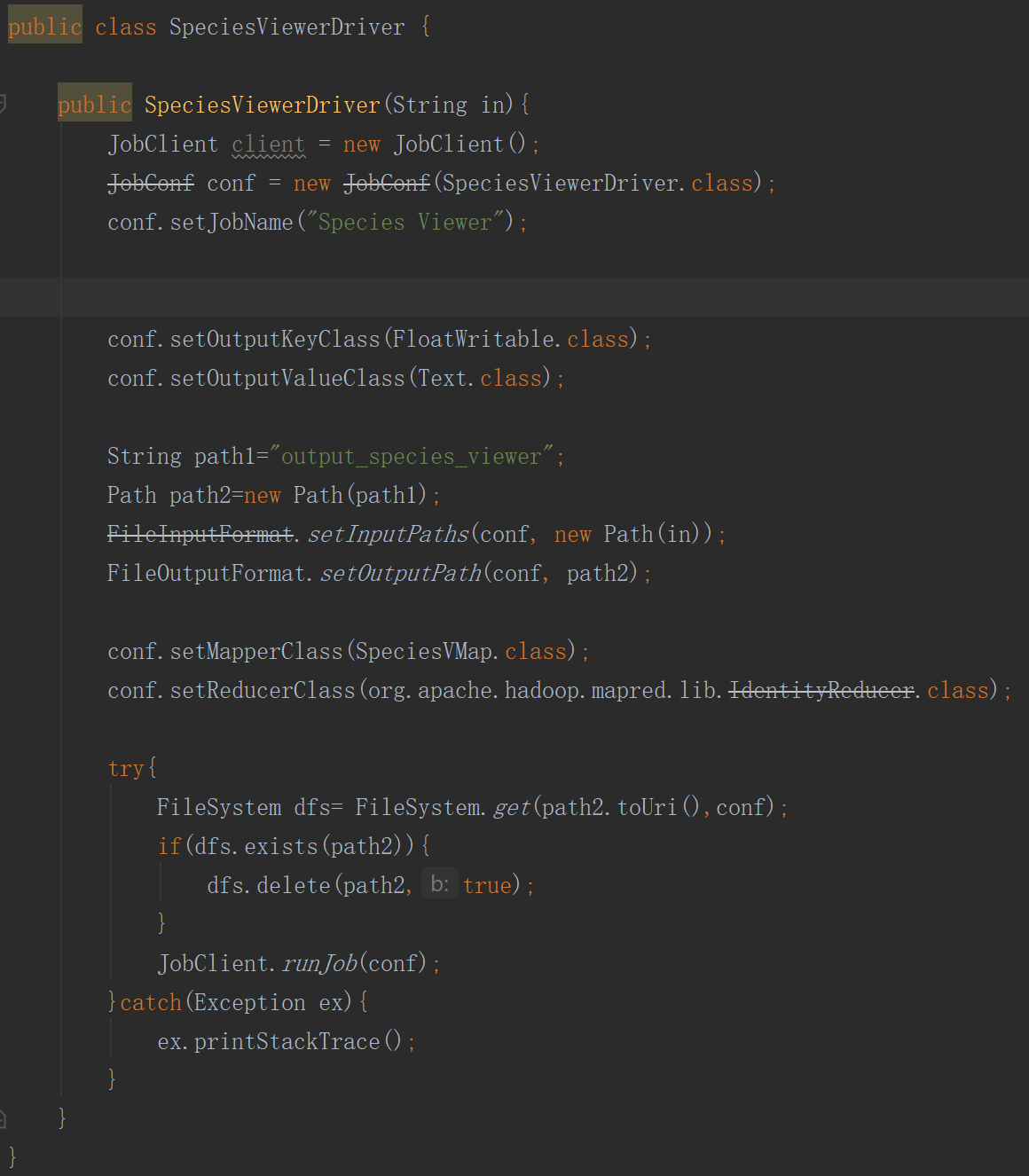
The 20th outfile:



The output file of SpeciesViewerDriver:



I changed the SpeciesViewerDriver like that:



In SpeciesViewerMapper, I changed the damping factor.