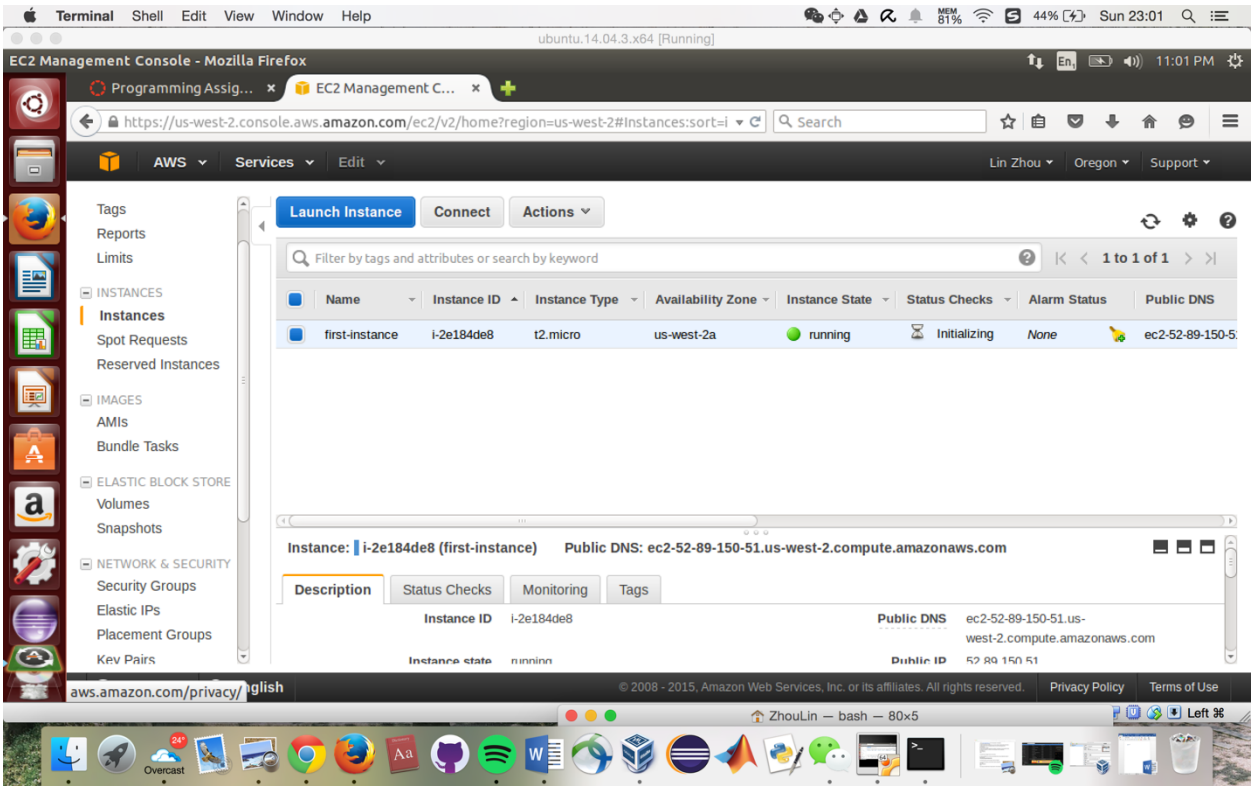
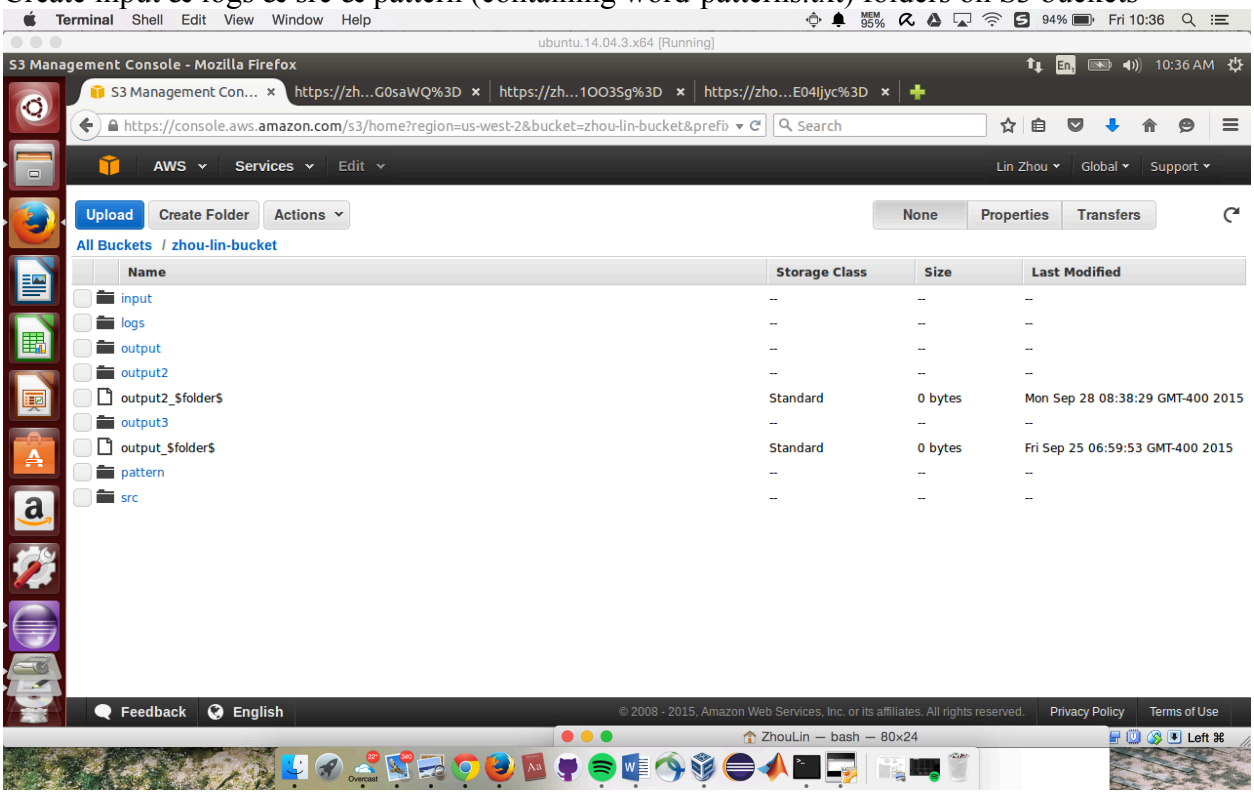
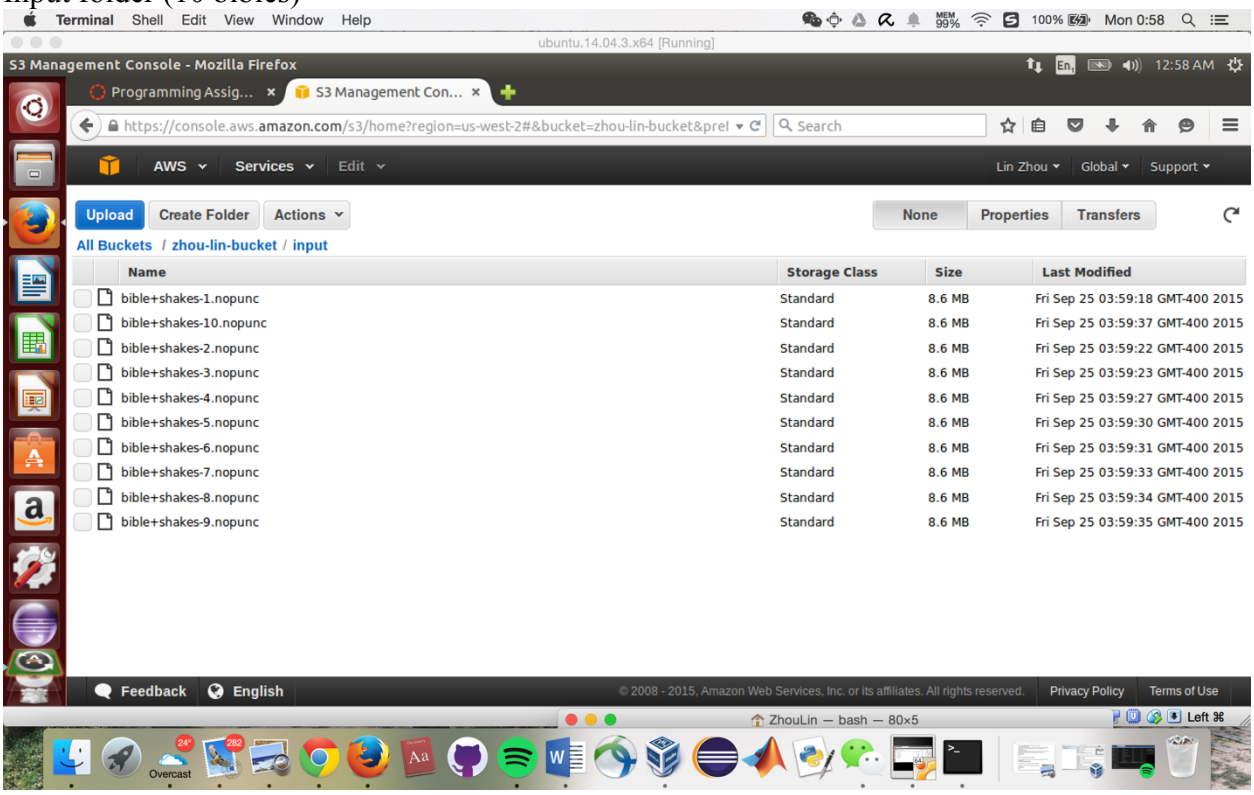
Connect to remote virtual severs on EC2



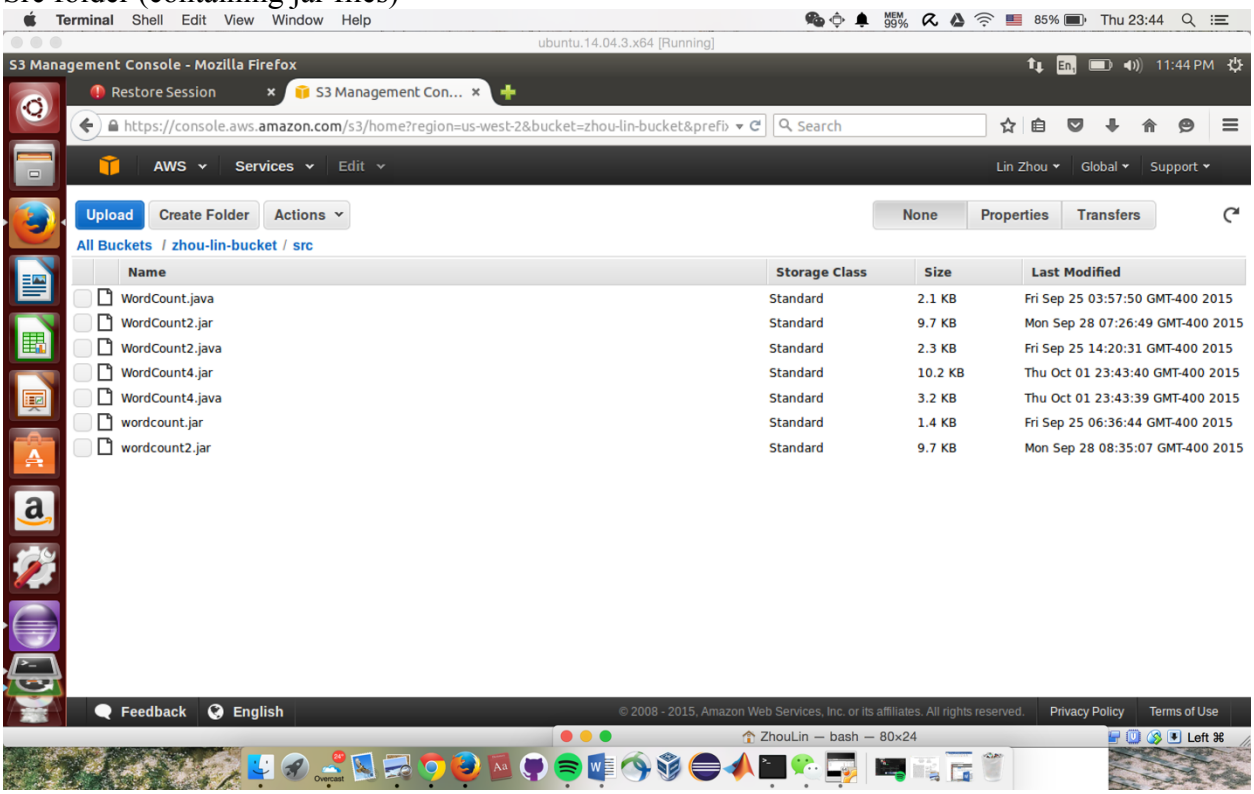
Create input & logs & src & pattern (containing word-patterns.txt) folders on S3 buckets



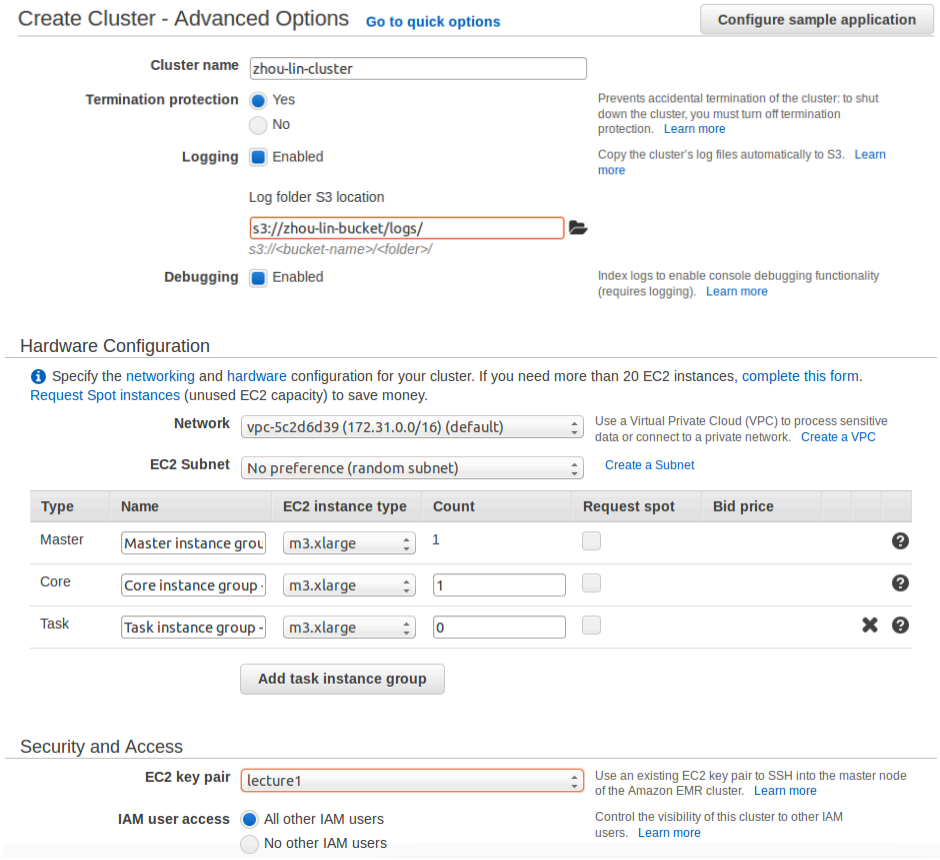
Input folder (10 bibles)

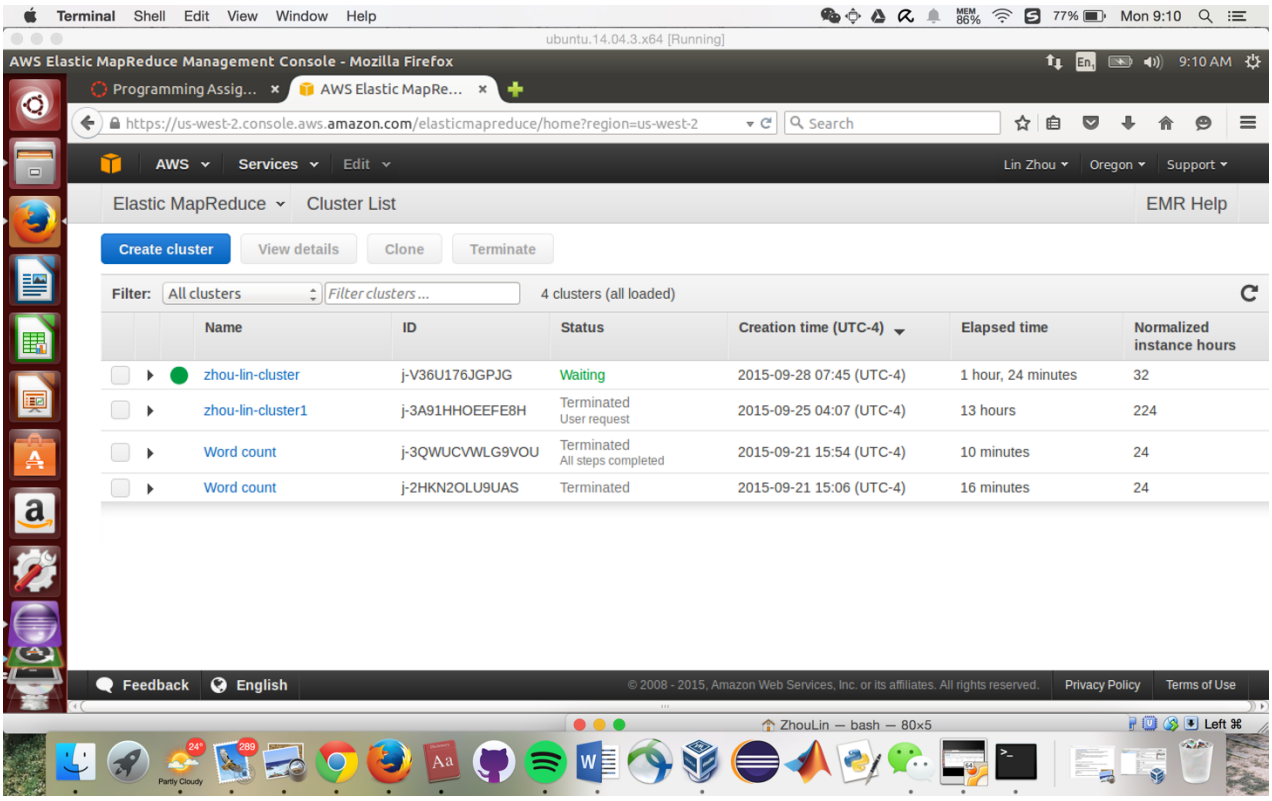


Src folder (containing jar files)

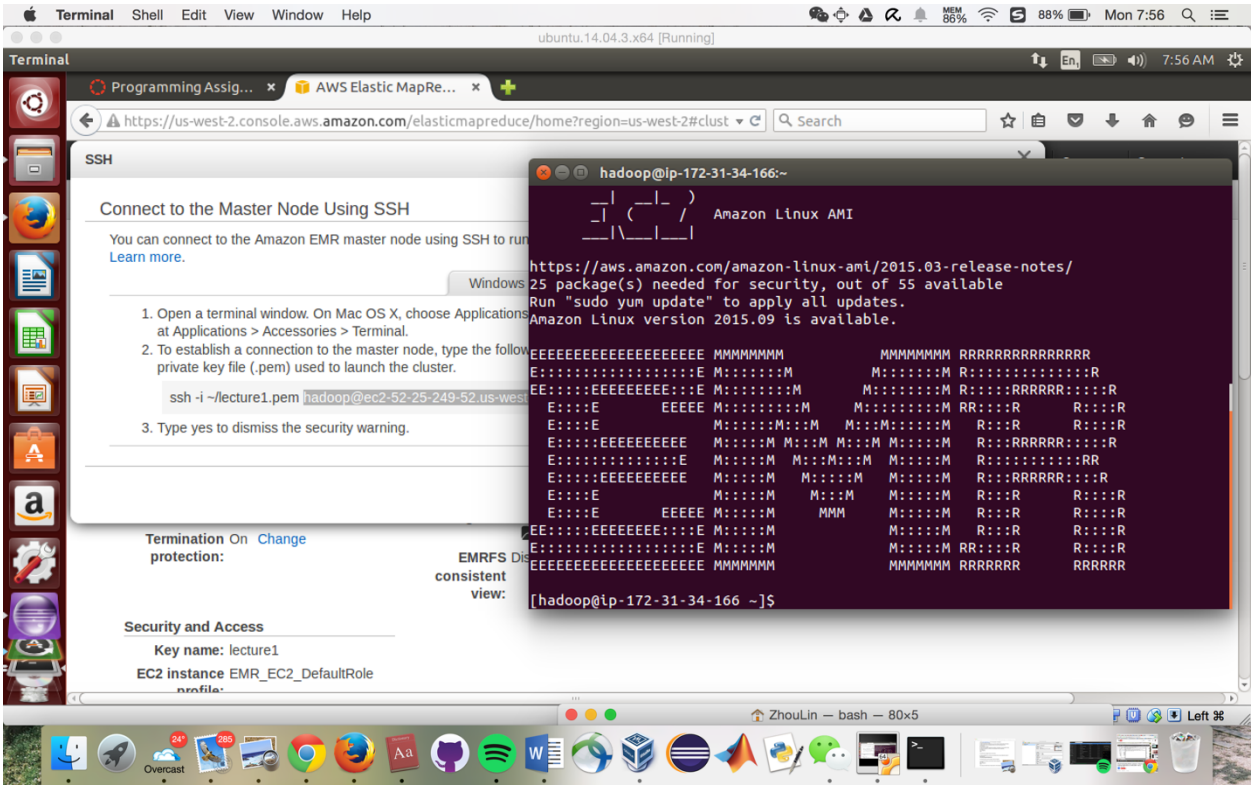


Create a new cluster

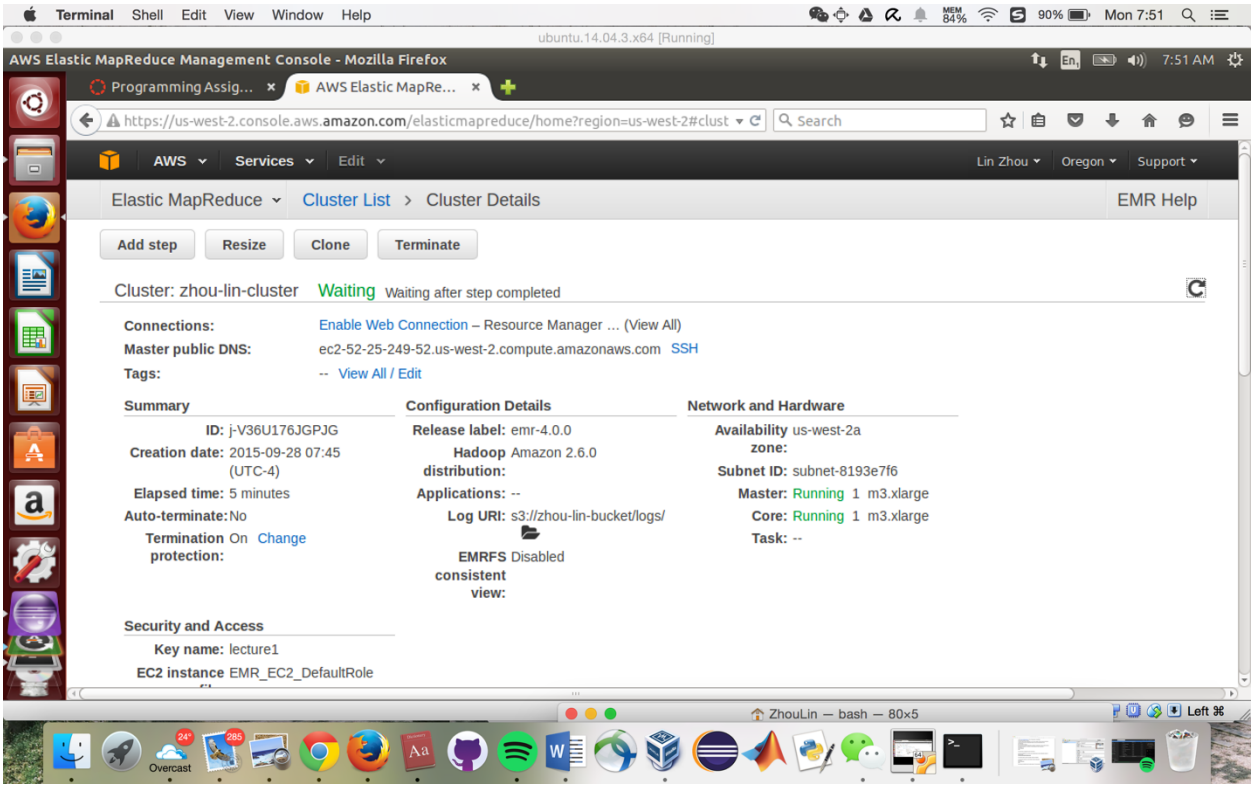




Connect to Master-node

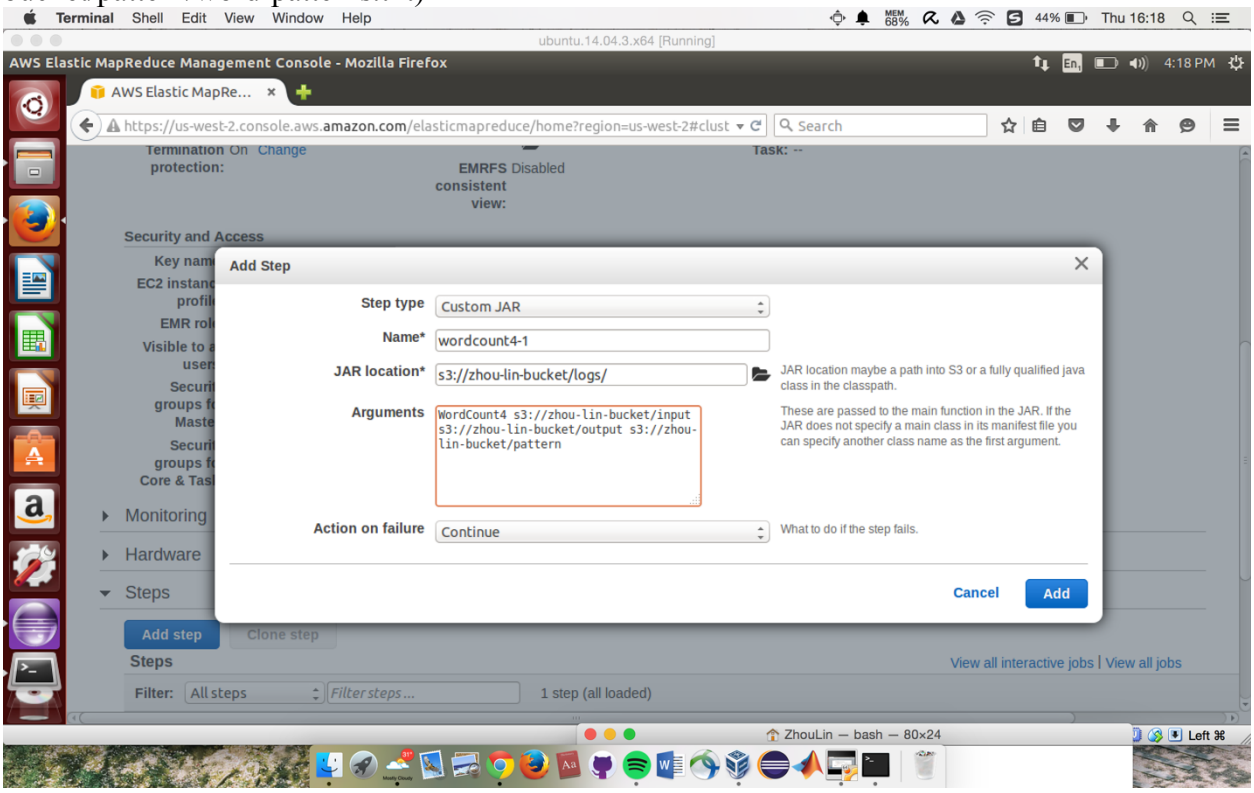


EMR service

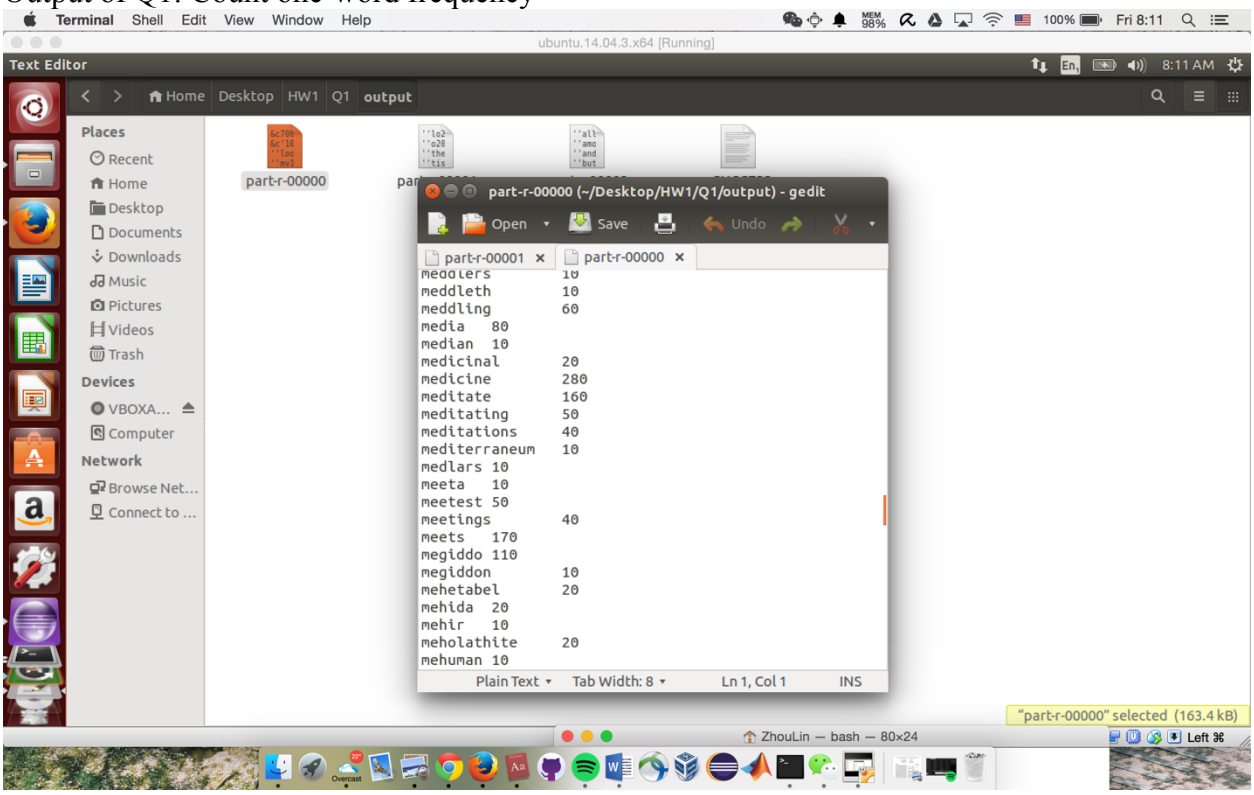


Submit a job to Hadoop cluster (Add step; Configuration of Q3)

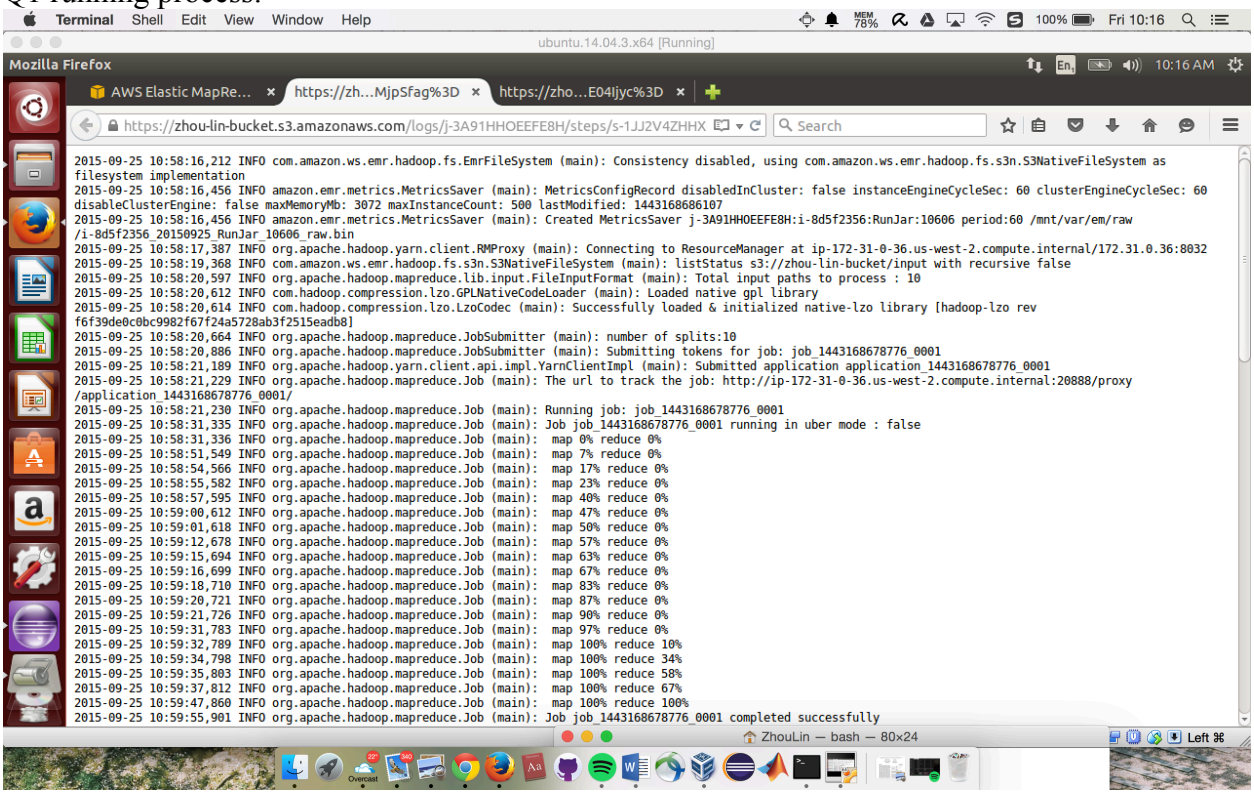
(Arguments are change shown as follows:
WordCount4 s3://zhou-lin-bucket/input s3://zhou-lin-bucket/output3 s3://zhou-lin-bucket/pattern/word-patterns.txt)



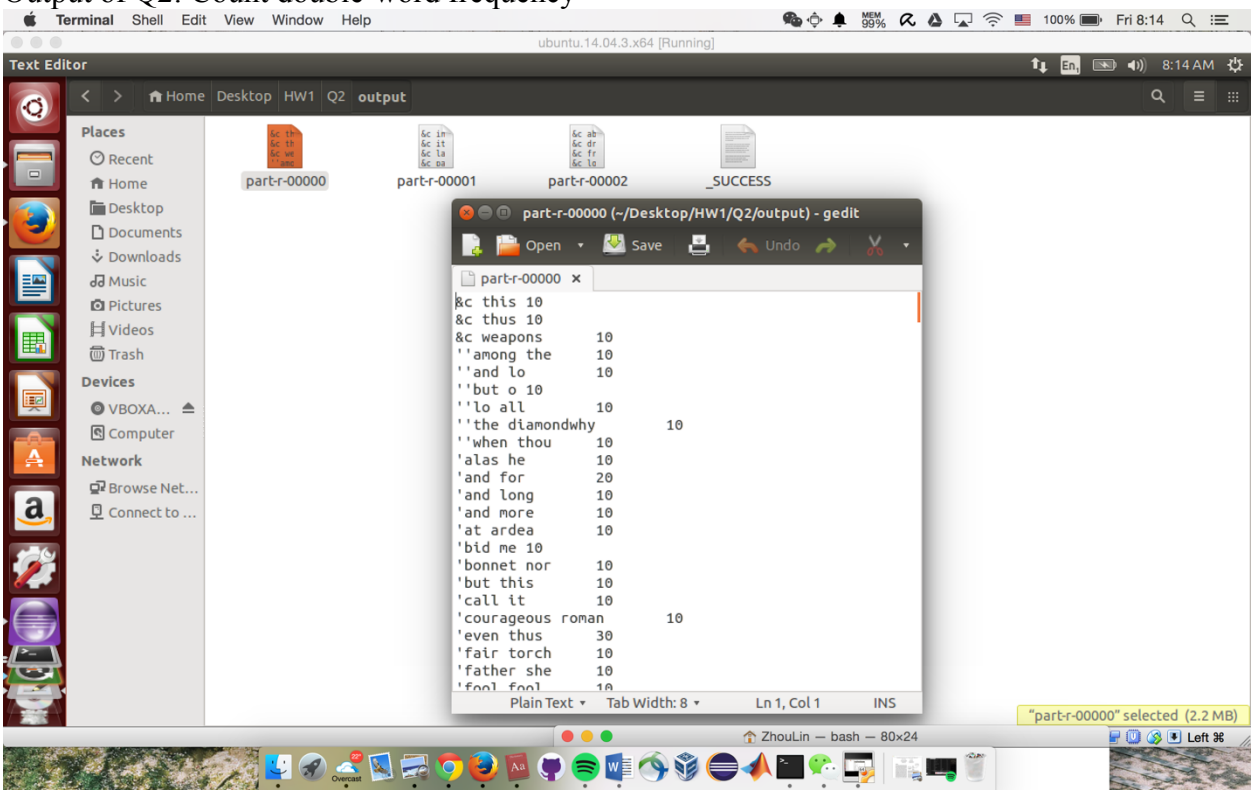
Output of Q1: Count one-word frequency



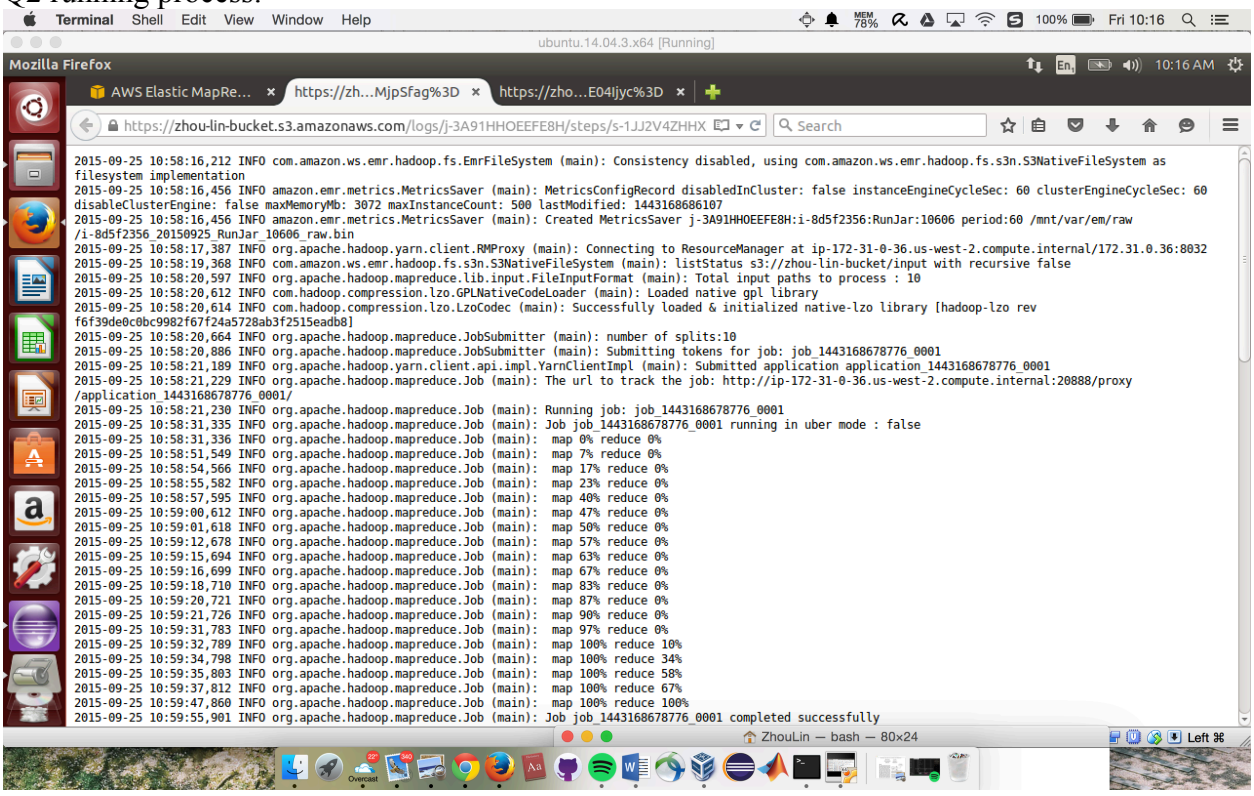
Q1 running process:



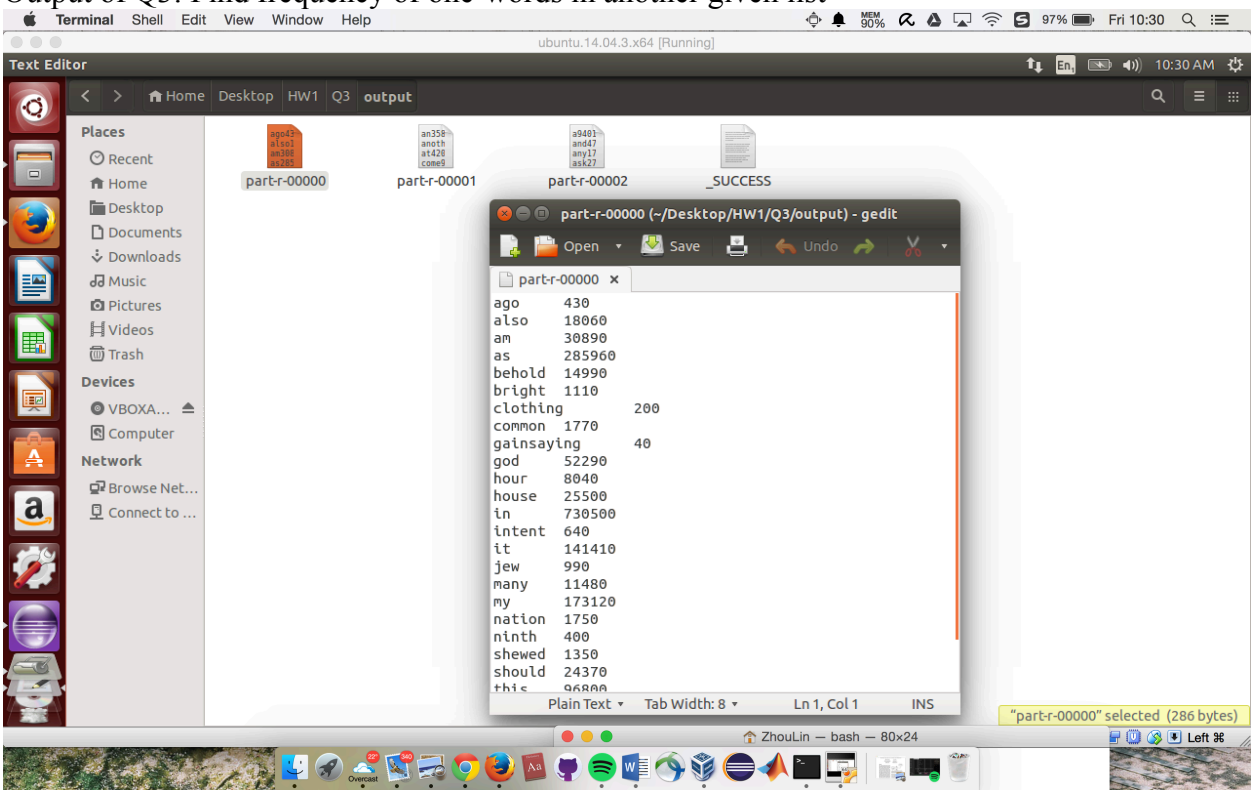
Output of Q2: Count double-word frequency



Q2 running process:



Output of Q3: Find frequency of one-words in another given list



Q3 running process: