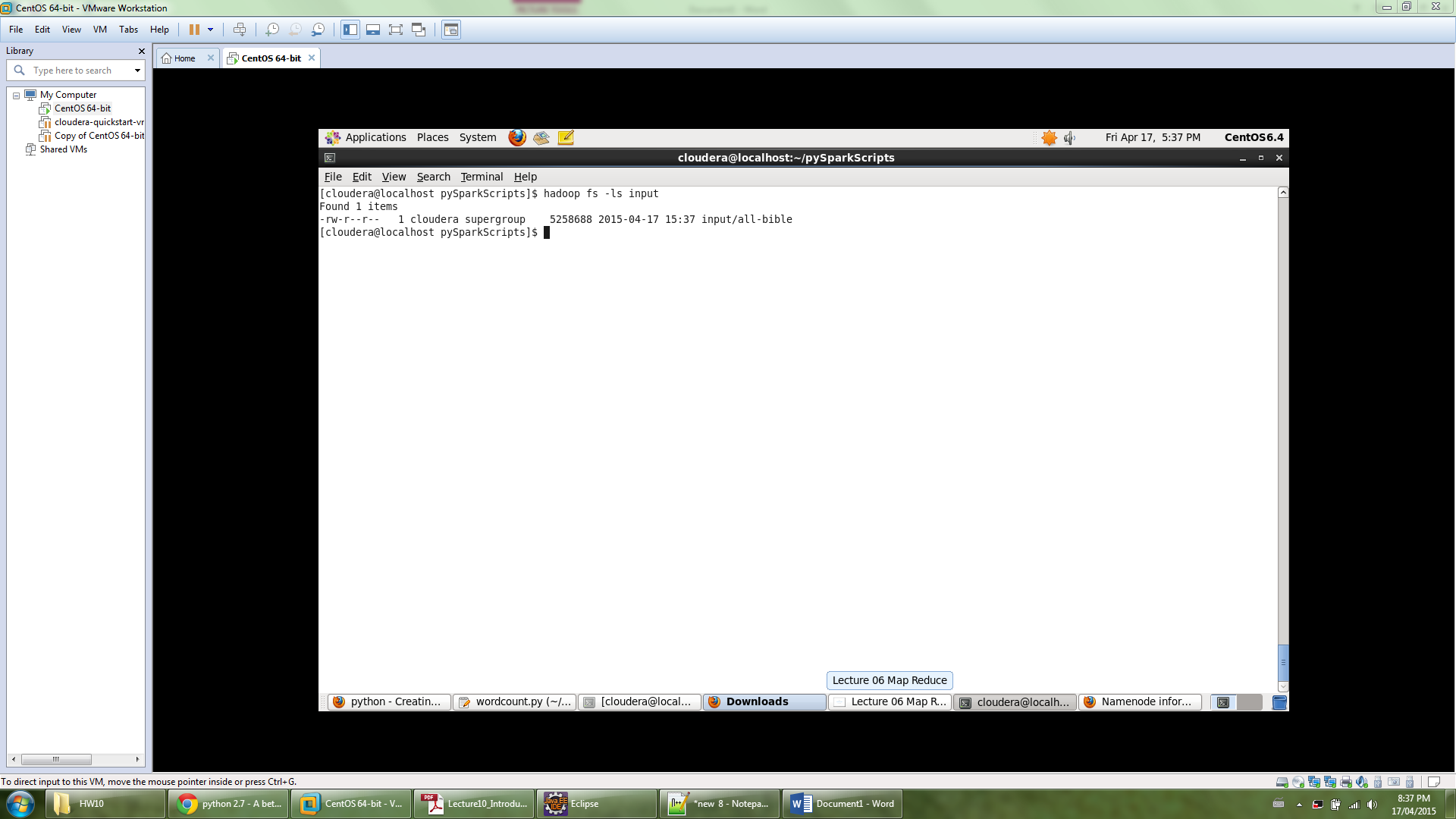
HU Extension **Assignment 10** E-63 Big Data Analytics

**Problem1**

Put input file in hdfs



import string

from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("YourApp")

sc = SparkContext(conf = conf)

file = sc.textFile("hdfs://localhost:8020/user/cloudera/input/all-bible")

counts = file.flatMap(lambda line: line.encode('utf-8','ignore').translate(None, string.punctuation).lower().split()).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)

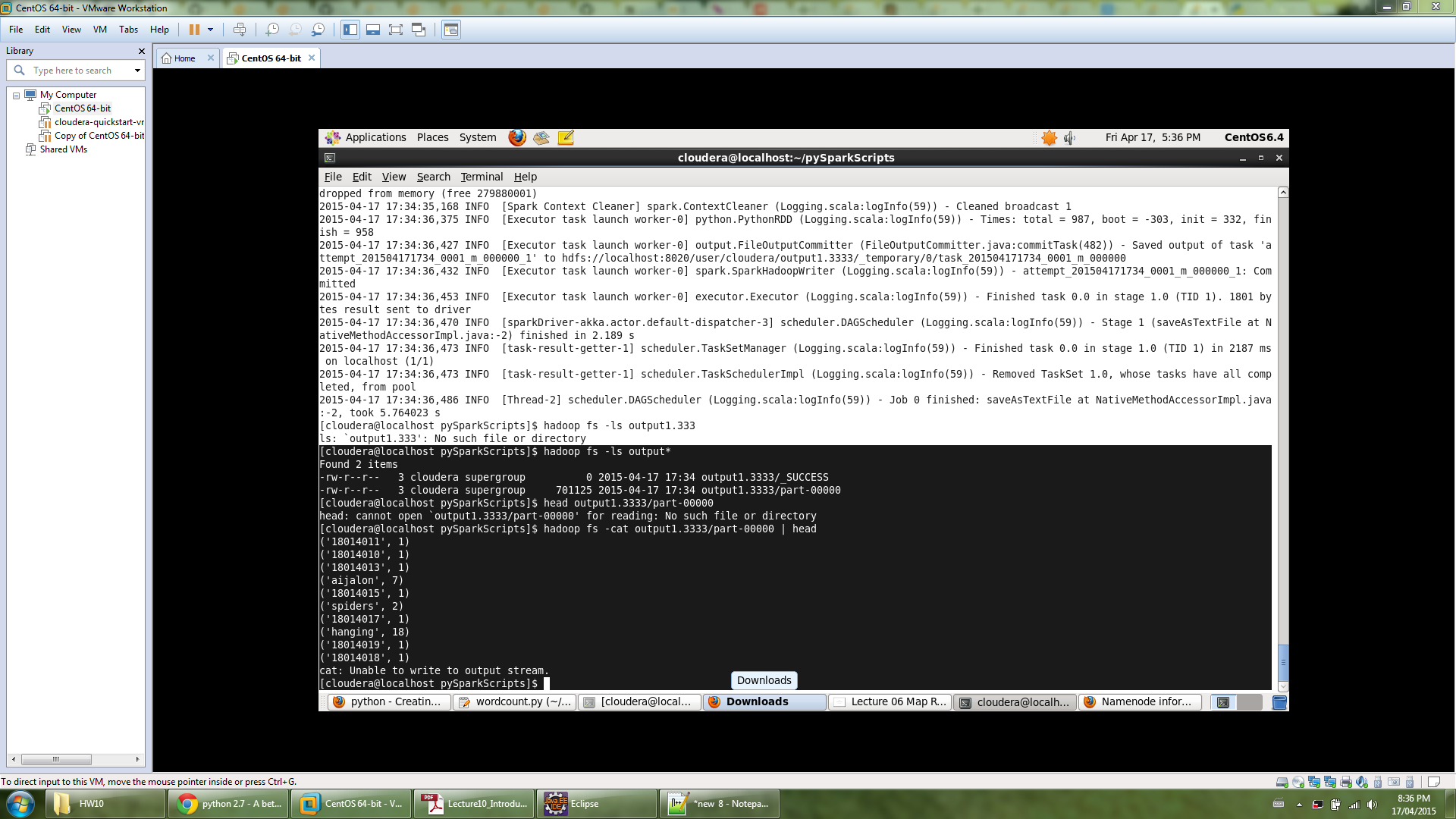
counts.saveAsTextFile("hdfs://localhost:8020/user/cloudera/output1.3333")

Highlighted sections –

Use hdfs://localhost:8020/user/cloudera/ to read from/write to hdfs,

remove punctuation using translate(None, string.punctuation) option in python

and convert string to lower case using lower()

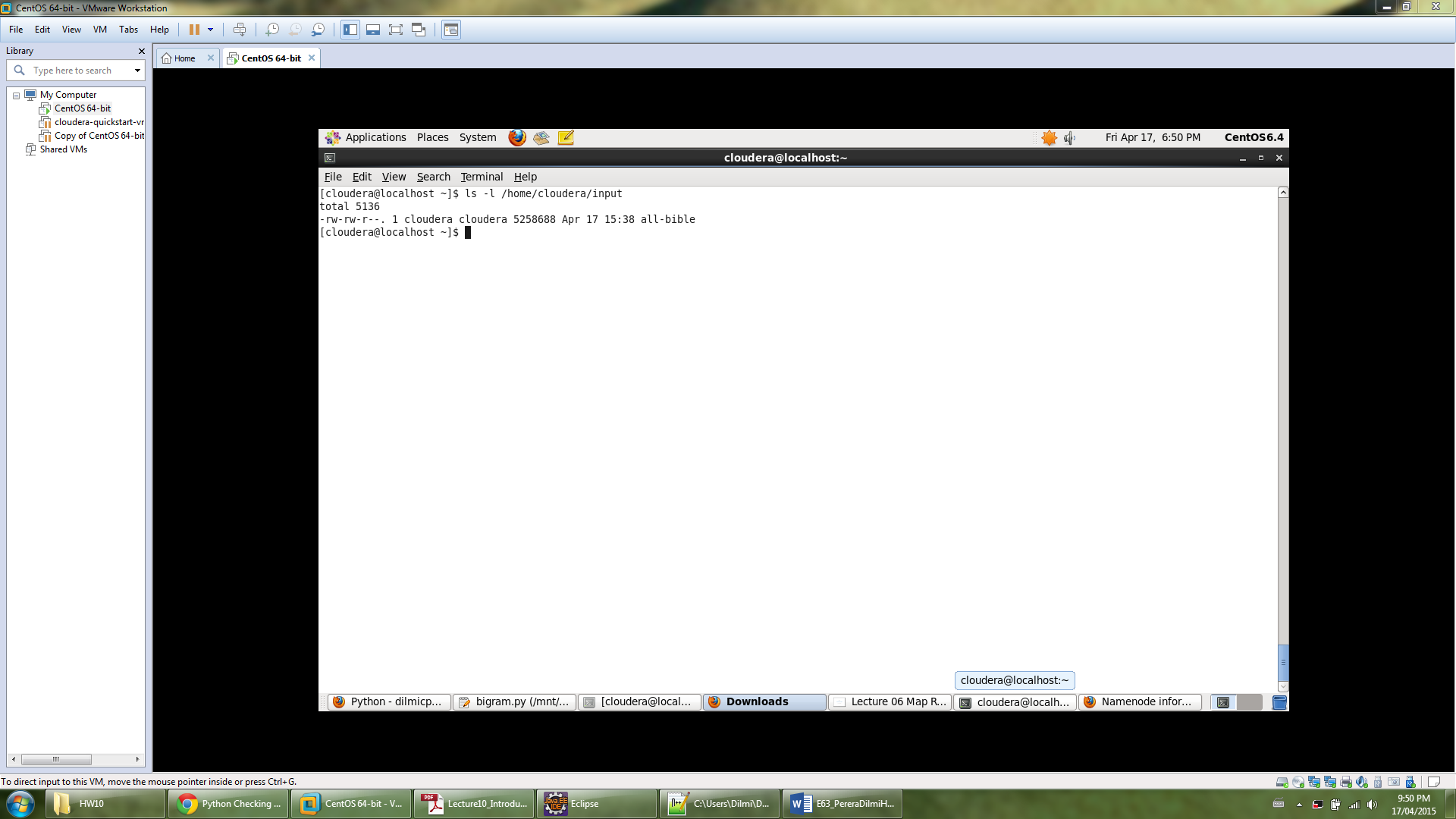


Output file and wordcount.py file attached

**Problem 2**

Read input from local OS using

file:/home/cloudera/input/all-bible



import string

from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("YourApp")

sc = SparkContext(conf = conf)

file = sc.textFile("file:/home/cloudera/input/all-bible")

counts = file.flatMap(lambda line: ['{0} {1}'.format(i[0], i[1]) for i in zip(line.split()[:-1], line.split()[1:]) if not i[0].endswith('.')]).map(lambda word: (word.encode('utf-8','ignore').translate(None, string.punctuation).lower(), 1)).reduceByKey(lambda a, b: a + b)

counts.saveAsTextFile("file:/home/cloudera/output2.0")

Split line into bigrams while filtering out the bigrams that have an end of sentence full stop in the middle. These bygrams will be the keys.

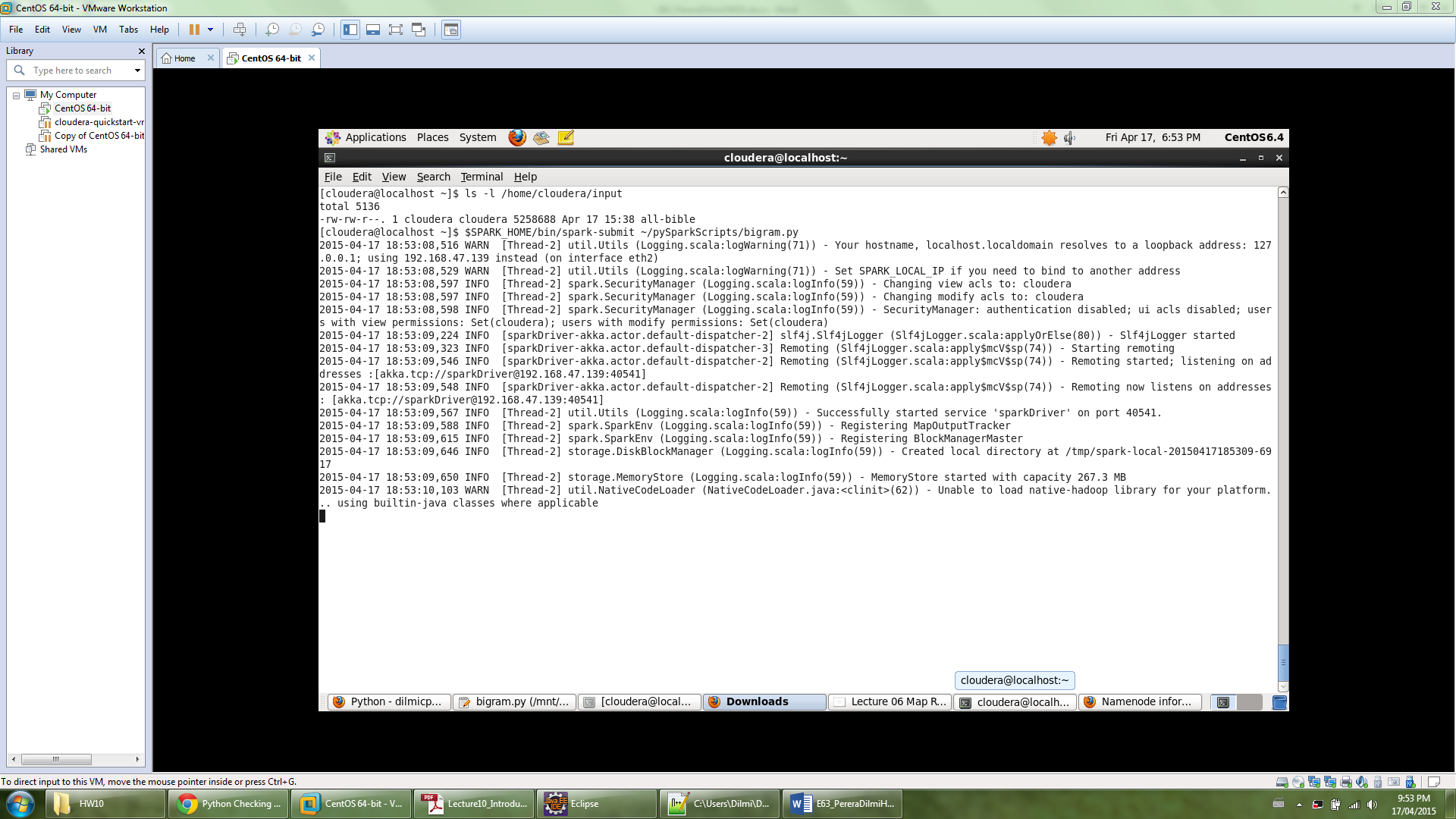
lambda line: ['{0} {1}'.format(i[0], i[1]) for i in zip(line.split()[:-1], line.split()[1:]) if not i[0].endswith('.')]).

Format keys by taking out punctuations and converting to lower case in the map process.

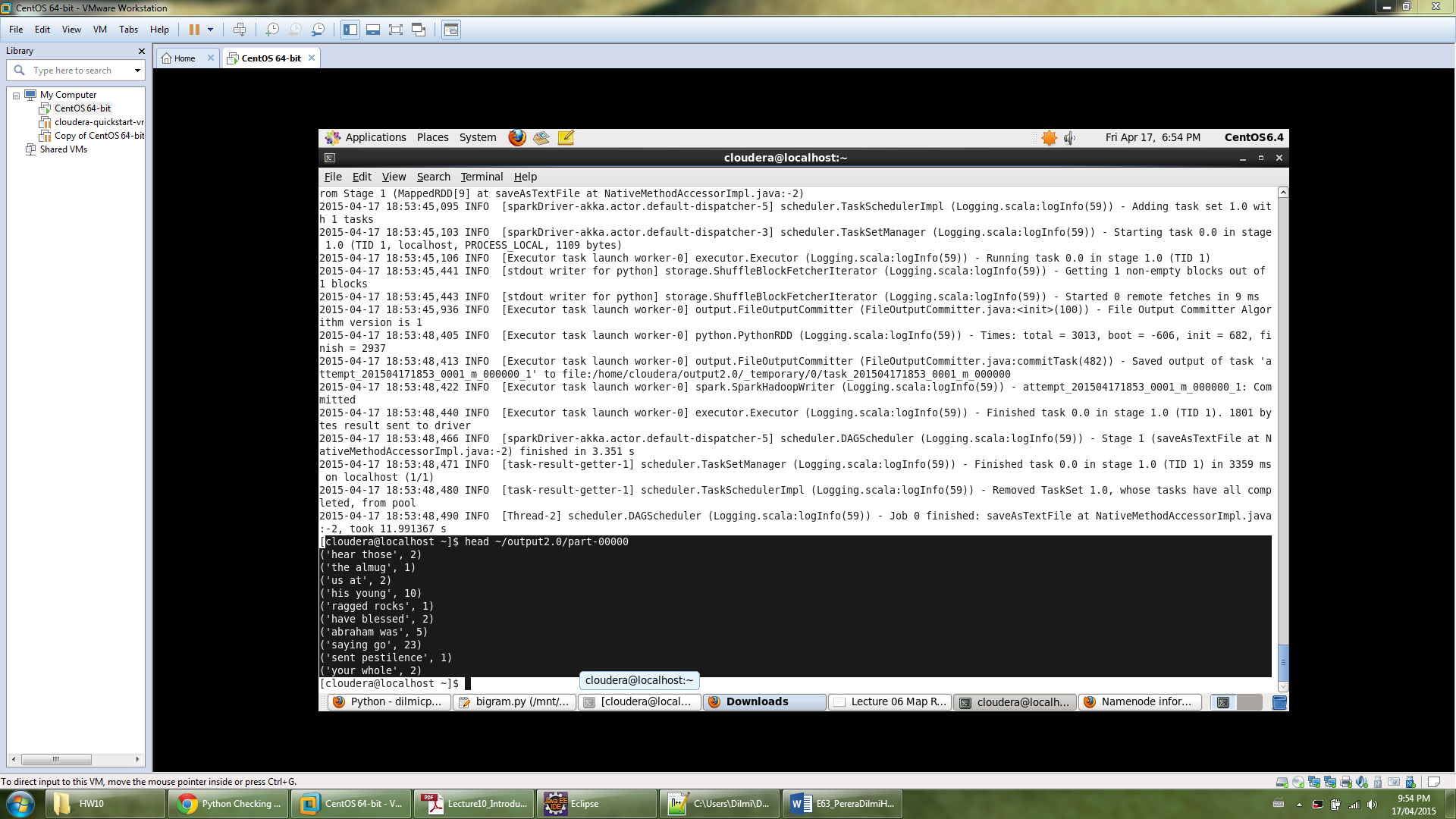
lambda word: (word.encode('utf-8','ignore').translate(None, string.punctuation).lower(), 1)).

Write output to local OS file system

file:/home/cloudera/input/output2.0



**Output**



Output file and bigram.py files attached