1) Which lines from the book war\_and\_peace.txt contain both words “war” and “peace”?

val file = sc.textFile("file:///root/war\_and\_peace.txt")

val sortWar = file.filter(line => line.split(" ").exists(w => w.toLowerCase.matches(".\*\\bwar\\b.\*")))

val warAndPeace = sortWar.filter(line =>line.split(" ").exists(w => w.toLowerCase.matches(".\*\\bpeace\\b.\*")))

warAndPeace.count

Long = 7

2) Approximately how many sentences are in the book war\_and\_peace.txt? HINT: Let’s

define a sentence as a sequence of text that ends with a period and note that an

approximation is ok.

val file = sc.textFile("file:///root/war\_and\_peace.txt")

val words = file.flatMap(line => line.split(" "))

val lines = words.filter(word => word.endsWith("."))

lines.count

Long = 24190

(I also tried val lines = words.filter(word => word.contains(".")) which was giving me answer 26652 . but I just felt ends with makes more sence so I went with word.endsWith(“.”))

3) Save a random sampling of 500 lines from war\_and\_peace.txt either to local or HDFS

storage as text.

val file = sc.textFile("file:///root/war\_and\_peace.txt")

val randomSample = file.takeSample(false, 500)

sc.parallelize(randomSample).saveAsTextFile("file:///root/sample\_war\_and\_peace.txt")

4) From home\_data.csv, how many houses sold were built prior to 1979?

val home\_data = sc.textFile("file:///root/home\_data.csv")

val buildPrior1979 = home\_data.filter(line => line.split(",")(14) < "1979" )

buildPrior1979.count

Long = 11991

5) From home\_data.csv, how many houses sold had a lot size (in sq. ft) that was greater than triple the living space (in sq. ft)? Save this output to file either locally or on HDFS.

Example: lot size 500 living size 100 would count (500 > 300) but 500/200 would not

(500 < 600).

val home\_data = sc.textFile("file:///root/home\_data.csv")

val header = home\_data.first()

val data = home\_data.filter(row => row != header)

val bigLot = data.filter( {line =>((line.split(",")(6).toInt) > (3 \* line.split(",")(5).toInt))})

bigLot.saveAsTextFile("file:///root/lotBiggerThanLiving.txt")

bigLot.count

Long = 14267