Nishant Velagapudi

W251 – HW6

Spark Context

Questions:

1. I obtained the value 3.1406. The value is not exact because it relies on a random sample. We’re essentially sampling randomly within a square with 2 sides of length one with a concentric circle of radius one within it. If the random draw lands inside the concentric circle, we return 1, else 0. We should return 1 with a probability of pi/4 through this method. We then add all results (mapping the function results together) and divide by total samples. This yields an approximate value of pi divided by four, which we multiply by four. Increasing the number of samples will cause this number to converge to pi.
2. The moby10b.txt file has 23244 lines
   1. sc.textFile("/root/coursework/week6/hw/apache\_spark\_introduction/moby10b.txt")
3. The first line is “\*\*The Project Gutenberg Etext of Moby Dick, by Herman Melville\*\*”
4. 1276 lines contain the text “whale”
   1. textFile.filter(line => line.contains("whale")).count()
5. a. Two output files are written.

b. One output file is written. In the previous case, we used the default partitioning (2) as a minimum number and output files as we read them in. In this case, we allow for a minimum partitioning of one, and thus don’t split the task out at all.

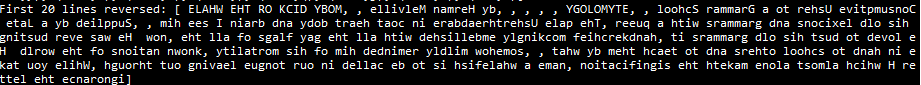
1. Our filtered file has 22933 lines.
2. Our filtered file has 220888 lines.
3. Sorted according to letter:



1. Sorted according to frequency:



1. Top 20 lines mirrored:



**Access Instructions:**

ssh198.11.211.67

root password: W4gse6pX

cd /root/coursework/week6/hw/apache\_spark\_introduction/src

**RUN LOCALLY:** java -cp .:$SPARK\_HOME/jars/\* spark.Hw6JavaCode /root/coursework/week6/hw/apache\_spark\_introduction/moby10b.txt /root/coursework/week6/hw/apache\_spark\_introduction/output/

I wasn’t able to submit this job to the hdfs cluster. I receive the following error message:

