Monica Santana – Week 4 Exercise

SparkPi Output

```
santanamonica@dscbigdata: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
36402 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.resource.ResourceProfile  - Default ResourceProfile created, executor resources: Map(cores -> name: cores, amoun
t: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
38161 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnDriverEndpoint  - Registered executor NettyRpcEndpointRef(spark-clien
t://Executor) (172.28.1.2:55700) with ID 2
38520 [dispatcher-BlockManagerMaster] INFO  org.apache.spark.storage.BlockManagerMasterEndpoint  - Registering block manager worker1:7002 with 366.3 MiB RAM, BlockManagerId(2, w
orker1, 7002, None)
38677 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 0.0 in stage 0.0 (TID 0, worker1, executor 2, partition 0, PROCESS_LOC
AL, 7404 bytes)
39790 [dispatcher-BlockManagerMaster] INFO  org.apache.spark.storage.BlockManagerInfo  - Added broadcast_0_piece0 in memory on worker1:7002 (size: 1816.0 B, free: 366.3 MiB)
41692 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 1.0 in stage 0.0 (TID 1, worker1, executor 2, partition 1, PROCESS_LOC
AL, 7404 bytes)
41747 [task-result-getter-0] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 0.0 in stage 0.0 (TID 0) in 3118 ms on worker1 (executor 2) (1/10)
41894 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 2.0 in stage 0.0 (TID 2, worker1, executor 2, partition 2, PROCESS_LOC
AL, 7404 bytes)
41910 [task-result-getter-1] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 1.0 in stage 0.0 (TID 1) in 218 ms on worker1 (executor 2) (2/10)
42056 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 3.0 in stage 0.0 (TID 3, worker1, executor 2, partition 3, PROCESS_LOC
AL, 7404 bytes)
42060 [task-result-getter-2] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 2.0 in stage 0.0 (TID 2) in 169 ms on worker1 (executor 2) (3/10)
42141 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 4.0 in stage 0.0 (TID 4, worker1, executor 2, partition 4, PROCESS_LOC
AL, 7404 bytes)
42143 [task-result-getter-3] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 3.0 in stage 0.0 (TID 3) in 88 ms on worker1 (executor 2) (4/10)
42261 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 5.0 in stage 0.0 (TID 5, worker1, executor 2, partition 5, PROCESS_LOC
AL, 7404 bytes)
42262 [task-result-getter-0] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 4.0 in stage 0.0 (TID 4) in 122 ms on worker1 (executor 2) (5/10)
42385 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 6.0 in stage 0.0 (TID 6, worker1, executor 2, partition 6, PROCESS_LOC
AL, 7404 bytes)
42390 [task-result-getter-1] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 5.0 in stage 0.0 (TID 5) in 129 ms on worker1 (executor 2) (6/10)
42480 [task-result-getter-2] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 6.0 in stage 0.0 (TID 6) in 103 ms on worker1 (executor 2) (7/10)
42492 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 7.0 in stage 0.0 (TID 7, worker1, executor 2, partition 7, PROCESS_LOC
AL, 7404 bytes)
42566 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 8.0 in stage 0.0 (TID 8, worker1, executor 2, partition 8, PROCESS_LOC
AL, 7404 bytes)
42570 [task-result-getter-3] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 7.0 in stage 0.0 (TID 7) in 77 ms on worker1 (executor 2) (8/10)
42683 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 9.0 in stage 0.0 (TID 9, worker1, executor 2, partition 9, PROCESS_LOC
AL, 7404 bytes)
42686 [task-result-getter-0] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 8.0 in stage 0.0 (TID 8) in 120 ms on worker1 (executor 2) (9/10)
42765 [task-result-getter-1] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 9.0 in stage 0.0 (TID 9) in 83 ms on worker1 (executor 2) (10/10)
42769 [dag-scheduler-event-loop] INFO  org.apache.spark.scheduler.DAGScheduler  - ResultStage 0 (reduce at SparkPi.scala:38) finished in 7.314 s
42774 [task-result-getter-1] INFO  org.apache.spark.scheduler.cluster.YarnScheduler  - Removed TaskSet 0.0, whose tasks have all completed, from pool
42789 [dag-scheduler-event-loop] INFO  org.apache.spark.scheduler.DAGScheduler  - Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
42800 [dag-scheduler-event-loop] INFO  org.apache.spark.scheduler.cluster.YarnScheduler  - Killing all running tasks in stage 0: Stage finished
42810 [main] INFO  org.apache.spark.scheduler.DAGScheduler  - Job 0 finished: reduce at SparkPi.scala:38, took 7.637181 s
Pi is roughly 3.140815140815141
42882 [main] INFO  org.sparkproject.jetty.server.AbstractConnector  - Stopped Spark@46b695ec{HTTP/1.1,[http/1.1]}{172.28.1.1:4040}
42888 [main] INFO  org.apache.spark.ui.SparkUI  - Stopped Spark web UI at http://localhost:4040
42896 [YARN application state monitor] INFO  org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend  - Interrupting monitor thread
43039 [main] INFO  org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend  - Shutting down all executors
43040 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnDriverEndpoint  - Asking each executor to shut down
43069 [main] INFO  org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend  - YARN client scheduler backend Stopped
43127 [dispatcher-event-loop-1] INFO  org.apache.spark.MapOutputTrackerMasterEndpoint  - MapOutputTrackerMasterEndpoint stopped!
43267 [main] INFO  org.apache.spark.storage.memory.MemoryStore  - MemoryStore cleared
43268 [main] INFO  org.apache.spark.storage.BlockManager  - BlockManager stopped
43379 [main] INFO  org.apache.spark.storage.BlockManagerMaster  - BlockManagerMaster stopped
43384 [dispatcher-event-loop-0] INFO  org.apache.spark.scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint  - OutputCommitCoordinator stopped!
43418 [main] INFO  org.apache.spark.SparkContext  - Successfully stopped SparkContext
43437 [shutdown-hook-0] INFO  org.apache.spark.util.ShutdownHookManager  - Shutdown hook called
43441 [shutdown-hook-0] INFO  org.apache.spark.util.ShutdownHookManager  - Deleting directory /tmp/spark-a40f37fd-ff27-4c49-a2ad-f7cc0260110a
43457 [shutdown-hook-0] INFO  org.apache.spark.util.ShutdownHookManager  - Deleting directory /tmp/spark-10f24bbc-c05c-4b87-b8f1-b80c6bac404c
bash-5.0#
```

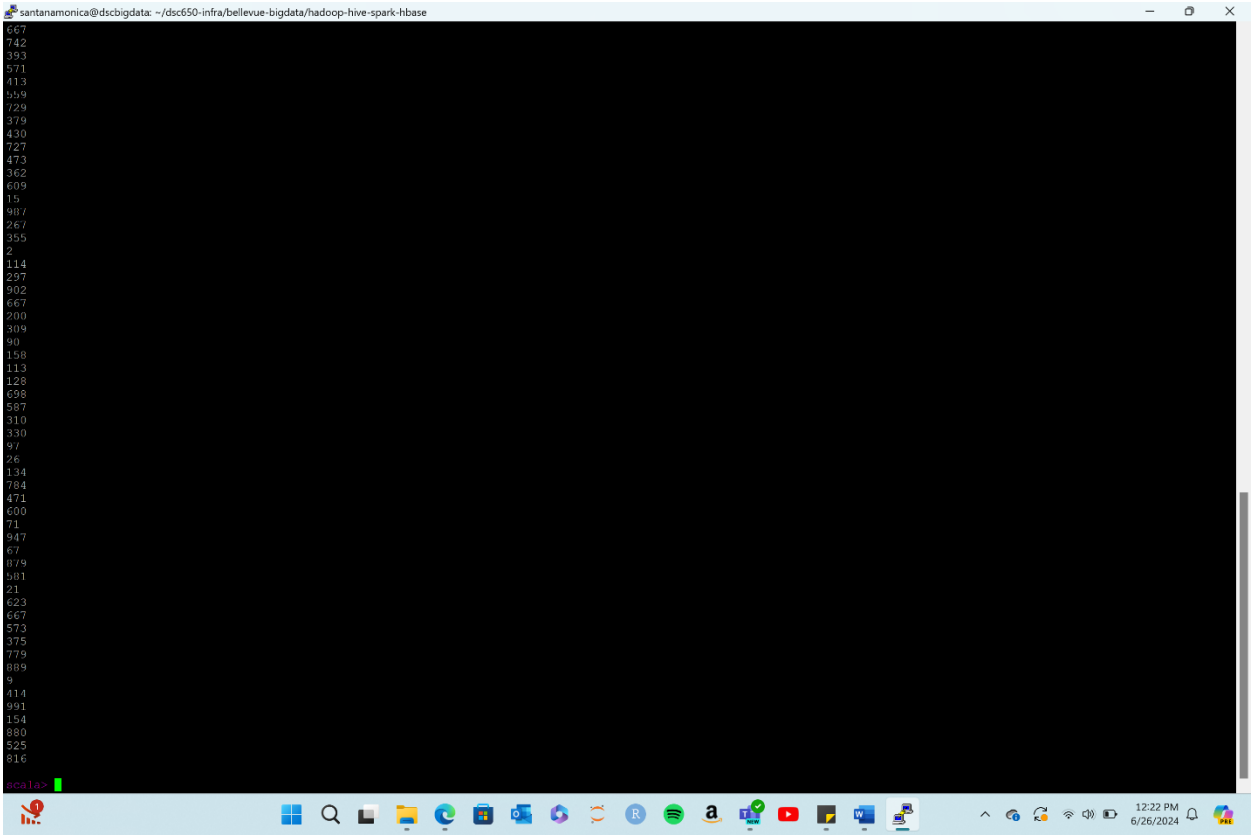
First 100 generated random numbers – First half

```
santanamonica@dscbigdata: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
numbers: scala.collection.immutable.IndexedSeq[Int] = Vector(862, 78, 126, 159,
562, 848, 678, 91, 848, 197, 87, 386, 675, 913, 811, 563, 249, 110, 547, 291, 79
3, 775, 798, 576, 285, 313, 58, 106, 831, 1, 828, 6, 680, 310, 490, 664, 252, 60
4, 69, 405, 329, 610, 356, 667, 742, 393, 571, 413, 559, 729, 379, 430, 727, 473
, 362, 609, 15, 987, 267, 355, 2, 114, 297, 902, 667, 200, 309, 90, 158, 113, 12
8, 698, 587, 310, 330, 97, 26, 134, 784, 471, 600, 71, 947, 67, 879, 981, 21, 62
3, 667, 573, 375, 779, 889, 9, 414, 991, 154, 880, 525, 816, 205, 729, 939, 662,
 379, 172, 645, 176, 892, 37, 830, 332, 415, 422, 10, 711, 830, 384, 930, 53, 17
6, 141, 554, 134, 584, 661, 287, 685, 834, 402, 308, 86, 718, 642, 707, 372, 861
, 930, 0, 996, 560, 866, 873, 561, 873, 592, 954, 8...

scala> val numbersRDD = sc.parallelize(numbers)
numbersRDD: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallel
ize at <console>:26

scala> numbersRDD.take(100).foreach(println)
862
78
126
159
562
848
678
91
848
197
87
386
675
913
811
563
249
110
547
291
793
775
798
576
285
313
58
106
831
1
828
6
680
310
490
664
252
604
69
405
329
610
356
```

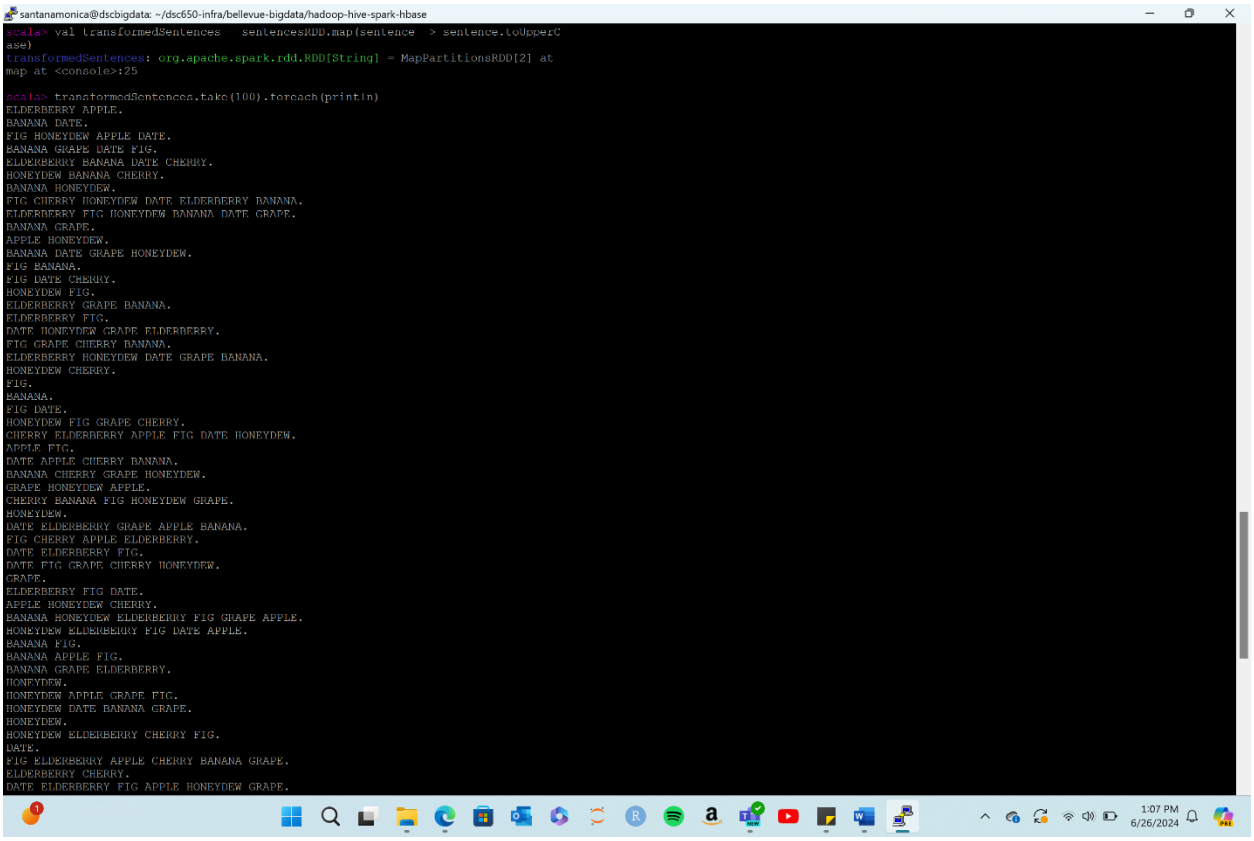
Second half


santanamonica@dscbigdata: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
667
742
393
571
413
559
729
379
430
727
473
362
609
15
987
267
355
2
114
297
902
667
200
309
90
158
113
128
698
587
310
330
97
26
134
784
471
600
71
947
67
879
581
21
623
667
573
375
779
889
9
414
991
154
880
525
816
scala>
12:22 PM
6/26/2024

Custom transformations: 1st one is changing the words to uppercase

CODE:

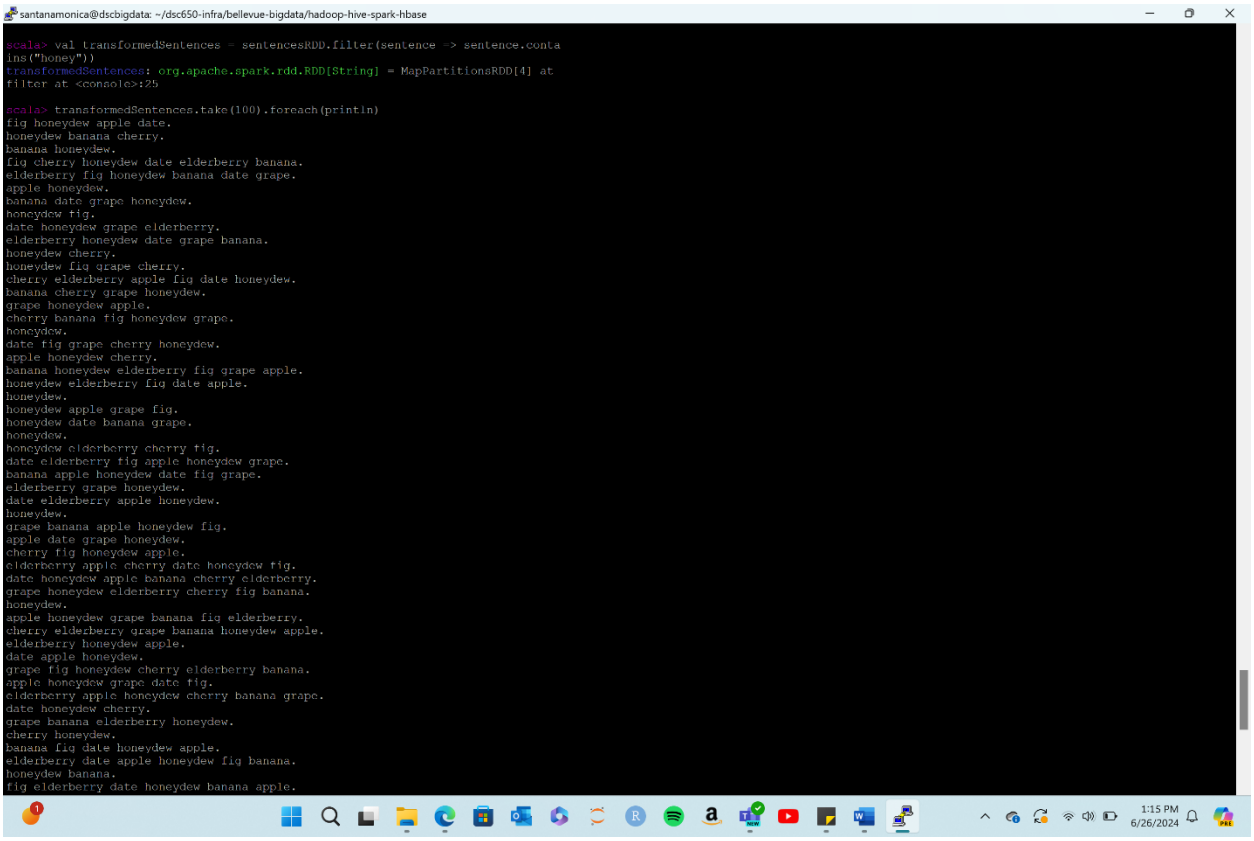
```
val transformedSentences = sentencesRDD.map(sentence => sentence.toUpperCase)
transformedSentences.take(100).foreach(println)
```



2nd transformation: Using the filter() and contains() to filter out only the words that have honey in them

CODE:

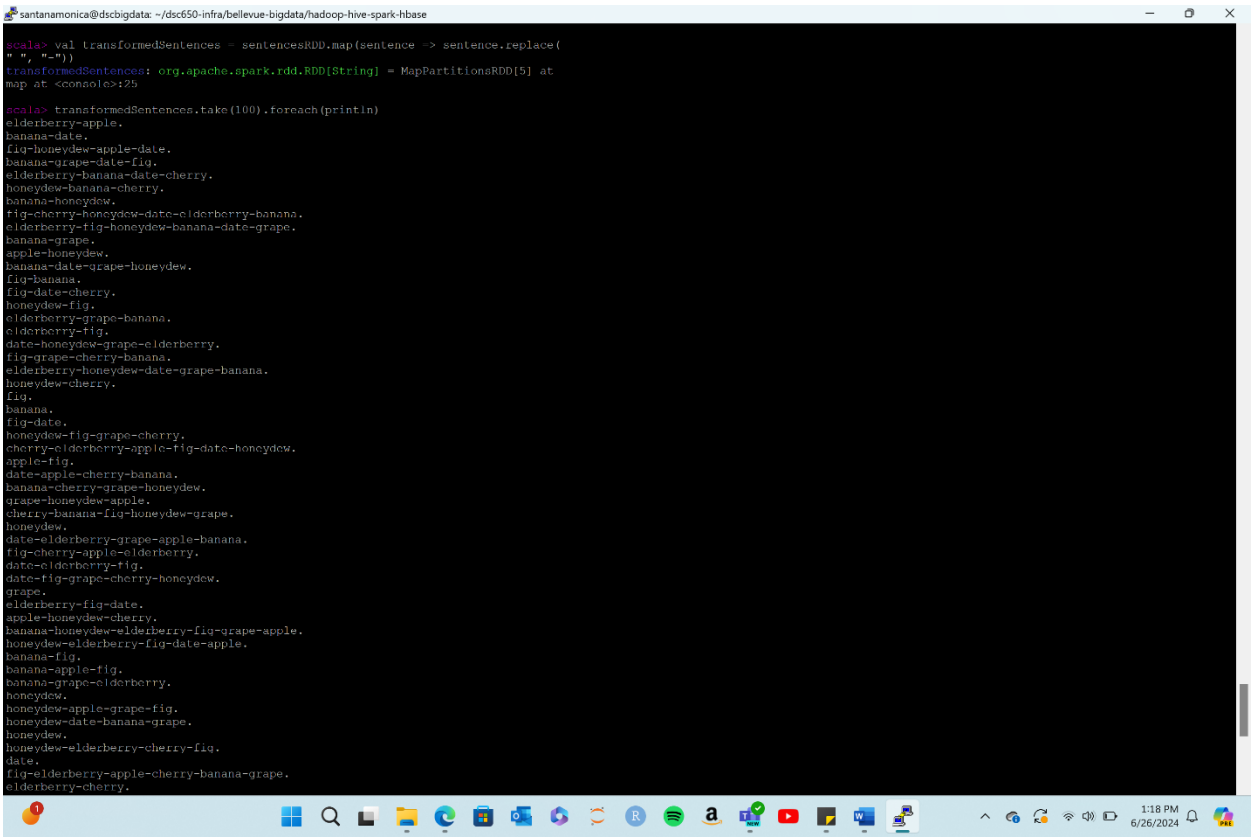
```
val transformedSentences = sentencesRDD.filter(sentence => sentence.contains("honey"))
transformedSentences.take(100).foreach(println)
```



3rd transformation: Using map() and the replace() to change the spaces to hyphens in the words

CODE:

```
val transformedSentences = sentencesRDD.map(sentence => sentence.replace(" ", "-"))
transformedSentences.take(100).foreach(println)
```



4th transformation: Using filter() and startsWith() to filter out the words only starting with date and prints a true or false. True if the word starts with date and false if it doesn't

CODE:

```
val transformedSentences = sentencesRDD.filter(sentence => sentence.startsWith("date"))
transformedSentences.take(100).foreach(println)
```