

Scott Berry

- 1) Explain the differences between RDD and a traditional Relational Database System.

A RDD is a memory structure created by a Spark job. This data is loaded from a file or a database and organized for the context of a spark job to execute. Once the job is complete, the temporary RDD is deleted. On the other hand, a Relational Database System is a storage system that is permanently accessible after Spark jobs are complete.

- 2) Using pyspark create a word count application of all the words of the file. Avoid counting trivial words such as vowels and pronouns.

```
Sat Feb 12 20:06:49 berry@DESKTOP-SCOTT MINGW64 ~/OneDrive/Documents/Indiana University/Spring 2022/Data Science On Ramp/DsOnRampSpark (master)
$ python code_snippet.py
2022-02-12 20:06:52,813 WARN util.Shell: Did not find winutils.exe: {}
java.io.FileNotFoundException: Could not locate Hadoop executable: C:\Hadoop\hadoop-3.3.1\bin\winutils.exe -see https://wiki.apache.org/hadoop/WindowsProblems
    at org.apache.hadoop.util.Shell.getQualifiedBinInner(Shell.java:619)
    at org.apache.hadoop.util.Shell.getQualifiedBin(Shell.java:592)
    at org.apache.hadoop.util.Shell.<clinit>(Shell.java:689)
    at org.apache.hadoop.util.StringUtils.<clinit>(StringUtils.java:79)
    at org.apache.hadoop.conf.Configuration.getTimeDurationHelper(Configuration.java:1886)
    at org.apache.hadoop.conf.Configuration.getTimeDuration(Configuration.java:1846)
    at org.apache.hadoop.conf.Configuration.getTimeDuration(Configuration.java:1819)
    at org.apache.hadoop.util.ShutdownHookManager.getShutdownTimeout(ShutdownHookManager.java:183)
    at org.apache.hadoop.util.ShutdownHookManager$HookEntry.<init>(ShutdownHookManager.java:207)
    at org.apache.hadoop.util.ShutdownHookManager.addShutdownHook(ShutdownHookManager.java:304)
    at org.apache.hadoop.util.ShutdownHookManager.install(ShutdownHookManager.scala:181)
    at org.apache.spark.util.ShutdownHookManager$.shutdownHooks$lzycompute(ShutdownHookManager.scala:50)
    at org.apache.spark.util.ShutdownHookManager$.shutdownHooks(ShutdownHookManager.scala:48)
    at org.apache.spark.util.ShutdownHookManager$.addShutdownHook(ShutdownHookManager.scala:153)
    at org.apache.spark.util.ShutdownHookManager$.<init>(ShutdownHookManager.scala:58)
    at org.apache.spark.util.ShutdownHookManager$.<clinit>(ShutdownHookManager.scala)
    at org.apache.spark.util.Utils$.createTempDir(Utils.scala:335)
    at org.apache.spark.deploy.SparkSubmit.prepareSubmitEnvironment(SparkSubmit.scala:344)
    at org.apache.spark.deploy.SparkSubmit.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:898)
    at org.apache.spark.deploy.SparkSubmit.doRunMain$1(SparkSubmit.scala:180)
    at org.apache.spark.deploy.SparkSubmit.submit(SparkSubmit.scala:203)
    at org.apache.spark.deploy.SparkSubmit.doSubmit(SparkSubmit.scala:90)
    at org.apache.spark.deploy.SparkSubmit$$anon$2.doSubmit(SparkSubmit.scala:1043)
    at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:1052)
    at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)

Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
2022-02-12 20:06:53,022 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
1604: 1
TRAGEDY: 1
PRINCE: 1
Shakespeare: 1
Dramatis: 1
Personae: 1
Claudius: 2
Officer: 1
son: 11
nephew: 3
Polonius: 6
Lord: 13
Chamberlain: 1
Hamlet: 25
Laertes: 16
Voltemand: 4
courtier: 7
Rosencrantz: 6
Gentleman: 1
officer: 2
Bernardo: 2
Francisco: 2
soldier: 2
Reynaldo: 2
Two: 4
Clowns: 2
Fortinbras: 6
Prince: 1
Norway: 1
Norwegian: 1
English: 2
Ambassadors: 2
daughter: 8
Ghost: 7
ladies: 1
ACT: 4
platform: 3
before: 13
```

There was more than I could screenshot. I noticed that this function would be improved by interpreting all words as lower case and additionally stripping out punctuation marks.