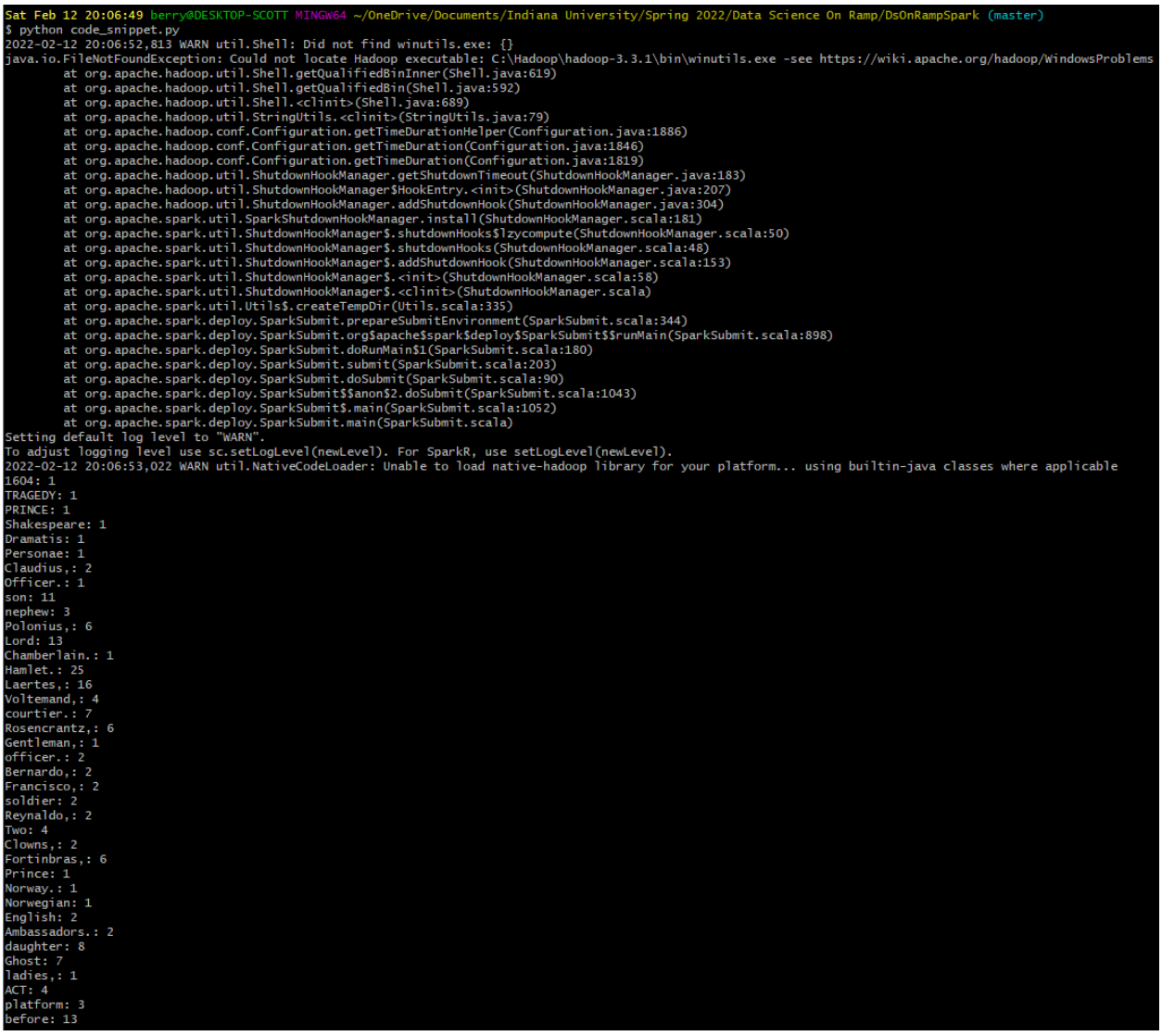
Scott Berry

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.



   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.