1. Explain the differences between RDD and a traditional relational database system.

   An RDD is a dataset formed of a collection of items that are distributed in a cluster. An RDD is immutable so any calculations done on the RDD creates new RDDs. Datasets in RDDs are partitioned across many servers. If a large dataset is loaded into an RDD, it will only load the data when an action is applied. The same concept goes for transformations. Spark only applies those when an action that is called requires the transformations.

   A traditional relational database work as storage systems It is a relational model where data is stored as tables that have relationships between them. This data can be queried and saved as views or reports.

2. I ran the scripts in the Pyspark terminal, but the assignment said to submit a .py file, so I copied the scripts into a python file.

```
>>> file = sc.textFile("textfile.txt")
>>> counts = file.flatMap(lambda line: line.split(" ")).filter(lambda word:len(word)>3).map(lambda word:(word, 1)).reduceByKey(lambda x,y: x + y)
>>> output = counts.collect()
>>> for(word, count) in output:
...     print("%s: %i" % (word,count))
...
1604: 1
TRAGEDY: 1
HAMLET,: 1
PRINCE: 1
DENMARK: 1
William: 1
Shakespeare: 1
Dramatis: 1
Personae: 1
Claudius,: 2
King: 34
Denmark.: 7
Marcellus,: 4
Officer.: 1
Hamlet,: 27
former,: 1
nephew: 3
present: 4
king.: 3
Polonius,: 6
Lord: 13
Chamberlain.: 1
Horatio,: 15
friend: 4
Hamlet.: 25
Laertes,: 16
Polonius.: 14
Voltemand,: 4
courtier.: 7
Cornelius,: 2
Rosencrantz,: 6
Guildenstern,: 5
Osric,: 4
Gentleman,: 1
Priest.: 3
officer.: 2
Bernardo,: 2
```