In-Class Quiz 5 Solution (Big Data)

1. What is Mapreduce? How does it work? Name a few applications of Mapreduce.

   Solution: in slides

2. Assume we want to compute the difference of two sets in a database using map reduce framework. Write map and reduce functions for computing $R \setminus S = \{x \in R | x \notin S\}$.Write map and reduce functions for computing $R \setminus S = \{x \in R | x \notin S\}$.

   Solution:

   difference
   - $R \setminus S$: rows of R that are not in S
   - map function : for each $t \in R$ produce $(t, \bar{r})$,
     for each $t \in S$ produce $(t, \bar{s})$

     map tasks are assigned chunks from either R or S
     $\bar{r}$ is a special symbol to mark that a tuple is in table R, and
     $\bar{s}$ is a special symbol to mark that a tuple is in table S
   - reduce function : if $(t, [\bar{r}])$ produce (t, t), o/w nothing

3. Compare Hadoop and Spark. Describe advantages and disadvantages of each one.

   Solution: `https://data-flair.training/blogs/spark-vs-hadoop-mapreduce/`

4. What are advantages and disadvantages of using bayesian methods in big data frameworks?

   Solution: answer in paper.[1]

[1] Zhu, Jun, et al. "Big learning with Bayesian methods." National Science Review 4.4 (2017): 627-651.