Adrien Xie, 451852 Guangda Ouyang, 467428 xie.z

Problem 1

(a) Please refer to the files in the SVN repository
(b) The top 10 list running by the sum of the ratings are below

103701	Ferris Bueller's Day Off
95216	Rain Man
94398	Seven
92377	The Godfather
92029	The Incredibles
90891	Pretty Woman
88670	As Good as It Gets
82862	The Italian Job
81889	Terminator 2: Extreme Edition
80580	Good Morning, Vietnam

Problem 2

(a) The code added to the driver is written below. The output does not change.

```
job.setCombinerClass(TopNReducer.class);
```

(b) The counters are shown in the table blow for a top-10 map reduce job

	Without Combiner	With Combiner
FILE number of bytes read	37092255	23520
FILE number of bytes written	74472267	118024
HDFS number of bytes read	58524138	58524138
HDFS number of bytes written	231	231

The number of bytes written and read from the HDFS does not change. For the reading counter, it does not change because the input data needs to be read from the HDFS and the input data is the same for both map reduce job. Similarly, the writing counter is same because both map reduce job performed top 10 list with same input data.

The number of the bytes written and read from the HDFS decreases when combiner is used. The Combiner class is used in between the Map class and the Reduce class to reduce the volume of data transfer between Map and Reduce. Therefore the number of files read and written is smaller when the combiner is used.

(c) The reason why we can use the sum reducer as the combiner is because 1. The summation as the reduce step is commutative and associative, and 2. The output data type matches.

(d) Associativity means that the calculation gives the same result regardless of the way the numbers are grouped. Median is not associative because it depends on the overall sorted order of the data input, so a reducer that finds the median cannot be used as a combiner.