**Distributed Systems**

**Assignment 7**

**Ashwini Prabhu and Mruganka Londhe**

There are 4 files in the output folder. One with the success indication (\_SUCCESS) and the other three files are parts of the output. Part-0001 has the complete output whereas the part-0000 and part-0002 are empty.

We think that there are only two keys, pos and neg i.e. the output of classifier. The multiple output files are generated because each reducer writes to its own file. This allows the system to scale because, having reducers write to separate files increases aggregate disk write throughput. The generation of empty output files depends on the mapreduce program that is being run and its implementation.

At the end of the program we need the output as a whole since the total is to be displayed after adding the output of all the reducers after the splitting. Hence the output is in the last part only.

Reference : https://forums.aws.amazon.com/thread.jspa?threadID=34905