Hadoop

Mapper Function:

The main task of mapper function is split input on whitespace and sort the input in the format <word, 1>. The output of mapper function is the input of the Reducer.

Reducer Function:

It has two main tasks: shuffle, and reduce. Shuffle means when the values are same, we put them together. And we set a count to record how many times this value appears. Then reducing. The output format is <key, count>. Also, the output is not sorted. Since our goal is to find the most frequency word, so I set a variable max, if count > max, then max= count at the reduce step. It can find the most frequency word. Last step is writing to the file system.

My Input:

e / Input / **file-1.txt**

First Block | Previous Block | Next Block | Last Block

```
class
database
sql
desk
desk
```

First Block | Previous Block | Next Block | Last Block

e / Input / **file-2.txt**

First Block | Previous Block | Next Block | Last Block

```
jar
sql
sql
sql
bee
desk
database
database
```

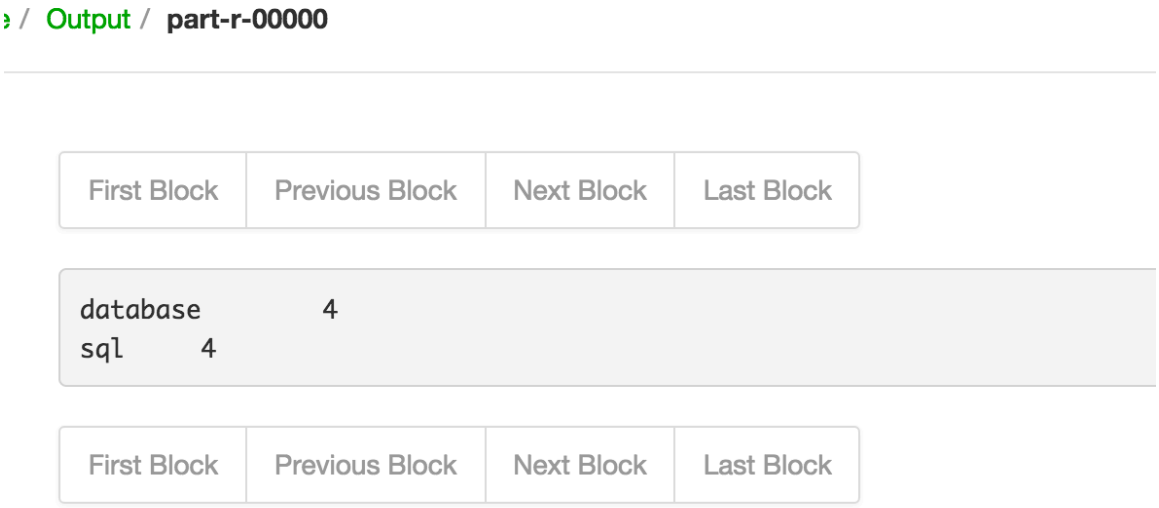
First Block | Previous Block | Next Block | Last Block

Output:

Output / part-r-00000

As we can see, my map and reduce function can find the maximum frequency word "sql".

Also, if I add one more "database" in the Input, we can get the following output:



My code can find the two max frequency words: “database” and “sql”.