Scalable Data Mining
Practice Questions
Hadoop

1. Consider a map-reduce program which takes a collection of words separated by space as input and computes the frequency of each word.

   Complete the following mapper code:

```
Mapper Code:

for line in sys.stdin:
   words = line.split(" ")
   for word in words:
        ------------------- (1)
```

   (a) print(word,1, sep='\t')

   (b) print(line,word, sep='\t')

   (c) print(line,1,sep='\t')

   (d) print(word,line, sep='\t')

2. Following the question above, complete the reducer code:

```
Reducer Code:
count=0
prevword = NULL
for line in sys.stdin:
        word,count = line.split("\t")

       if prevword==word:

          -------------- (2)

         else:
           if prevword != NULL:
             ------------------ (3)
           count=0
           prevword=word
           articlelist.append(articleid)
```

(a) (2) count=count+1
(3) print(prevword, count)

(b) (2) print count
(3) print prevword

(c) (2) print(word,count)
(3) print prevword

(d) (2) count=count+ 1
(3) print(word,count)

3. Suppose we have two mappers with outputs being as follows:

   Mapper 1: (a,8) (b,4)

   Mapper 2: (c,3) (c,6)

   How many key value pairs will be fed as input to the reducer with combiner?

   (a) 4
   (b) 3
   (c) 2
   (d) 1

4. State True or False.

   In a map-reduce framework, mapper/reducer should generate similar number of output key/value pairs it receives on the input.

5. How many mapper outputs are provided as input to a single combiner?

   (a) All of them
   (b) Outputs of one mapper
   (c) As many reducer inputs
   (d) Not fixed

6. Besides improving job completion time in Hadoop, backup tasks also improve fault-tolerance. State whether it is *True* or *False*.