91.673 Advanced Database Systems
Homework2

Chang Liu
chang_liu@student.uml.edu

October 4, 2015

Problem1. Design MapReduce algorithms to take a very large file of integers and produce as output:

(a) The largest integer.
(b) The average of all the integers.
(c) The same set of integers, but with each integer appearing only once.
(d) The count of the number of distinct integers in the input.

Solution: (a) First generate all the pairs and then sum up the same-key pair, that will generate a unique set. After that traverse all the pairs and find the max key, which is the max integer in the original file.

```
map(key=file, value=contents):
        for each integer in contents, emit (integer, "1")

reduce(key=integer, values=uniq_counts):
        sum all the pairs that has the same key
        emit all the pairs after the summing

max = pair[0].key
find_max(all pairs):
        for pair in pairs:
                if pair.key() > max:
                        max = pair.key()
```

(b) Input consists of many values, here we define the algorithm as map() and reduce():

```
map(key=file, value=contents):
        for each integer in contents, emit (integer, "1")
```

```
reduce(key=integer, values=uniq_counts):
        sum all (key,value) pairs in the out list
        emit result pair (total_integer_value, total_count)
```

The average number is just the total_integer_value/total_count.

NOTE: since in the above algorithm, during the reduce process the sum will sum all the count for "1", which is the number of integers(=total_count). And for each pair their is a value, so the sum of all is just the sum for all the integers (=total_integer_value), so the result of the division is the average number.

(c) The process is just to first get all the pairs when mapping, then reduce the repeated pair in reducing, as follows:

```
map(key=file, value=contents):
        for each integer in contents, emit (integer, "1")

reduce(key=integer, values=uniq_counts):
        sum all the pairs that has the same key
        emit all the pairs after the summing
```

At last, select all the key once from the pairs after the reducing process, because after reducing there are only unique integers in the pairs, select their key is to select the unique integer and forming a set will have such a unique property.

(d) The core idea is to sum up all the pair after mapping, then when the result pair's count is 1, this is a distinct number, sum up all these pair's count.

```
map(key=file, value=contents):
        for each integer in contents, emit (integer, "1")

reduce(key=integer, values=uniq_counts):
        sum all the pairs that has the same key
        emit all the pairs after the summing
```

At last their will be many pairs with different (key, value), the value field is just the count that the integer appears, count all the pair that their value field is "1", then the result count is just the count of distinct integer.

Problem2. Compute the Jaccard similarities of each pair of the following three sets: {1, 2, 3, 4}, {2, 3, 5, 7}, and $\{2, 4, 6\}$.

Solution: Let $A = \{1, 2, 3, 4\}$, $B = \{2, 3, 5, 7\}$, $C = \{2, 4, 6\}$, then we use $J(A, B), J(B, C) and J(A, C)$ to denote their Jaccard similarities.

1) The equation is as follows:

$$\begin{aligned} J(A,B) &= \frac{A \bigcap B}{A \bigcup B} \\ &= \frac{count(\{2,3\})}{count(\{1,2,3,4,5,7\})} \\ &= \frac{2}{6} \\ &= \frac{1}{3} \end{aligned}$$

2) The equation is as follows:

$$\begin{aligned} J(B,C) &= \frac{B \bigcap C}{B \bigcup C} \\ &= \frac{count(\{2\})}{count(\{2,3,4,5,6,7\})} \\ &= \frac{1}{6} \end{aligned}$$

3) The equation is as follows:

$$\begin{aligned} J(A,C) &= \frac{A \bigcap C}{A \bigcup C} \\ &= \frac{count(\{2,4\})}{count(\{1,2,3,4,6\})} \\ &= \frac{2}{5} \end{aligned}$$