(Hao Zhang, WustlKey: h.zhang633, ID: 452003), (Hanming Li, WustlKey: lihanming, ID: 451802)

Homework 7

1. (b) Take the sentence “No, now is no now time” as an example, there is a comma in the sentence, so if we measuring merely word co-occurrence of words directly next to each other, the comma will appear in a co-occurrence of words, which is not what we need.

(Hao Zhang, WustlKey: h.zhang633, ID: 452003), (Hanming Li, WustlKey: lihanming, ID: 451802)

2. (a) The pseudo-code for MapReduce program for the stripes:

```
map (Mk){
        For jth row in Martix v
                vj =v(j)
        For ith row in Matrix Mk
                For jth column in Matrix Mk
                        emit(i, mijvj)

}

Reduce (i, [mi1v1, mi2v2, mi3v3, …]){

        xki = 0

        For mijvj in the input array

                xki = xki + mijvj

        emit(i, xki)

}
```

The pseudo-code for the final MapReduce program:

```
Map (M){

        For Mk is stripes of Martix M

                For ith row in Matrix M

                        emit(i, xki)

}

Reduce (i, [x1i, x2i, x3i, …]){

        xi = 0

        for xki in the input array

                xi = xi + xki

        emit(i, xi)

}
```

(b) The mapper input: [16 2; 5 11; 9 7; 4 14] and [3 13; 10 8; 6 12; 15 1]

The mapper output: (1, 16), (1, 4), (2, 5), (2, 22), (3, 9), (3, 14), (4, 4), (4, 28) and (1, 9), (1, 52), (2, 30), (2,32), (3, 18), (3, 48), (4, 45), (4, 4)

The reducer input: (1, [16, 4]), (2, [5, 22]), (3, [9, 14]), (4, [4, 28]) and (1, [9, 52]), (2, [30, 32]), (3, [18, 48]), (4, [45, 4])

The reducer output: (1, 20), (2, 27), (3, 23), (4, 32) and (1, 61), (2, 62), (3, 66), (4, 49)

The final reducer input: (1, [20, 61]), (2, [27, 62]), (3, [23, 66]), (4, [32, 49])

The final reducer output: (1, 81), (2, 89), (3, 89), (4, 81)

(c) The work I will do to divide the Matrix into k stripes:

```
Divide(M){
        l = M[0].length/k
        for (int i=0;i<l;i++){
                for(int j=0;j<M.length;j++){
                        M1[j][i] = M[j][i]
                        M2[j][i] = M[j][i+l]
                        M3[j][i] = M[j][i+2*l]
                                ...
                        Mk[j][i] = M[j][i+(k-1)*l]
                }
        }
}
```

And the M1 and M2 are the divided 2 matrixes.

(d) The communication cost of this algorithm is:

The number of stripes + (The number of rows * the number of column) + (the number of rows * the number of stripes) * 2

(Hao Zhang, WustlKey: h.zhang633, ID: 452003), (Hanming Li, WustlKey: lihanming, ID: 451802)

3. (a) For the Fig. 5.7, the matrix is M = [1/3 1/2 0; 1/3 0 1/2; 1/3 1/2 1/2]
   V = [1/3; 1/3; 1/3]
   M*v for several times-> v = [3/13; 4/13; 6/13]
   The PageRank is [3/13; 4/13; 6/13]
   (b) v = [1/3; 1/3; 1/3]
   β = 0.8 v = β*M*v + (1- β)*e/n for several times -> v = [7/27; 25/81; 35/81]
   The PageRank is M*v = [13/54; 49/162; 37/81]
   (c) For the Figure 5.4, the transition matrix is:
   M = [0 1/2 0 0 0; 1/3 0 0 1/2 0; 1/3 0 0 1/2 0; 1/3 1/2 0 0 0; 0 0 1 0 0]
   For the Figure 5.7 is:
   M = [1/3 1/2 0; 1/3 0 1/2; 1/3 1/2 1/2]
   This representation is more efficient because it is unnecessary to store the index of row and column, and the data do not need to be stored separately.