Yaffa Atkins

Data Mining

Homework 3

Suppose we want to use a MapReduce framework to compute minhash signatures and that the matrix is chunked by rows, so that a Map task is given the hash functions and a set of rows to work on. Design Map and Reduce functions to exploit MapReduce with data in this form. Provide a pseudocode for your design. Please attach your answer in PDF.

```
import numpy as np

def main()
    #initialize docs chuncked by rows
    data = [[1,0,0,1],
            [0,0,1,0],
            [0,1,0,1],
            [1,0,1,1],
            [0,0,1,0]]

    #initialize signature matrix to infinity
    sigMatrix = np.matrix(np.ones((2,len(data[0])))*np.inf

    return data.map(row).reduce()

#define minhash functions
def h1(x):
    return x+1 %5
def h2(x):
    return 3*x + 1 % 5

#map is called on each row seperatly: if an element is one, store its doc number
along with its minhash function outputs
def map(i):
    ones = []
    for j in data[i]:
        if data[i][j] == 1
            ones.append([j, h1(i), h2(i)])
    return ones

# total ones =
[[1,1,1],[4,1,1],[3,2,4],[2,3,2],[4,3,2],[1,4,0],[3,4,0],[4,4,0],[3,0,3]]

#for each document, assign the lowest minhash function outputs
def reduce(ones):
    for i in range(len(data[0])):
```

```
        for element in ones:
            if element[0] == i:
                 sigMatrix[0][i] = min(sigMatrix[0][i],element[1])
                 sigMatrix[1][i] = min(sigMatrix[1][i],element[2])

if __name__ == "__main__":
    main()
```