Midterm 1
Date: March 5, 2018 (in class)
Time: 4:00 ~ 5:15 PM (75 minutes)
Total: 20 points
8 points: 16 True/False questions: 0.5 points each
12 points: 4 groups of questions with 4-6 sub questions per group

20% of the total score will be directly related to the quizzes.

Closed book, NO computer, No cheat sheets

A. Topics covered
1. Data Collection, Sampling, and Preprocessing
   - Types of analytics
   - Types of data sources
   - Sampling
   - Types of data elements
   - Outliers

2. MapReduce
   - Summarization Patterns (numerical summarization, inverted index)
   - Filtering Patterns (Bloom filter, Top 10, Distinct)
   - Data Organization Patterns (Partitioning, Total Order Sorting)
   - Join Patterns (Replicated join, composite join)

3. How MapReduce Works
   - Managing Job (submission, execution, and monitoring)
   - Fault tolerance
   - Shuffle and Sort
   - Combiner and Partitioner
   - Input/output Format (InputSplits, record reader)

4. Link Analyses
   - Inverted index
   - Regular PageRank algorithm
   - PageRank algorithm with Taxation
   - Calculation PageRank algorithm with Dead ends
• Using MapReduce to calculate PageRank values (Matrix/vector multiplication, handling sparse matrix)
• Understanding link farm and link spam
• Spam mass and TrustRank

5. Clustering
• kMeans clustering algorithm
• Canopy algorithm
• Implementing kMeans with MapReduce

B. Sample “Group Problems”

Sample Question A.

Suppose that 10 items are registered to the initial Bloomfilter B using the same set of hash functions. Assume that the current bloomfilter has the state depicted below.

\[
\begin{align*}
   h1(x) &= x \mod 15 \\
   h2(x) &= (x+3) \mod 15 \\
   h3(x) &= (x+2) \mod 15
\end{align*}
\]

\[
\begin{array}{cccccccccccc}
   15 & 14 & 13 & 12 & 11 & 10 & 09 & 08 & 07 & 06 & 05 & 04 & 03 & 02 & 01 & 00 \\
   0 & 1 & 0 & 1 & 11 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1
\end{array}
\]

(A) The integer 4 is not a member of the set. \hspace{1cm} (True/False) \hspace{1cm} (answer. True)
(B) The integer 9 may be a member of the set. \hspace{1cm} (True/False) \hspace{1cm} (answer. True)
(C) The integer 1 must be a member of the set. \hspace{1cm} (True/False) \hspace{1cm} (answer. False)
Sample Question B.

Consider that you are calculating PageRank values for web pages. There are 10 Billion web pages and you have created a 10 Billion x 10 Billion transition matrix $M$. As a part of iterative computations, you use the MapReduce computing framework without Taxation. The $k^{th}$ iteration of the MapReduce job will create a vector $v$ with 10 Billion items. The $j^{th}$ item in $v$ is calculated using the following formula:

$$v_j^{(k+1)} = \sum_j m_{ij}v_j^{(k)}$$

**Question 1.** What are the values $m_{ij}$ stored in the transition matrix $M$? (Answer: b)

a. The total number of times that web page $i$ has been visited  

b. The probability that page $i$ is to be visited from the $j^{th}$ page  

c. The page $i$’s page rank value after $j^{th}$ iteration  

d. Random number generated by server

**Question 2.** What are the values $v$ for the $k^{th}$ iteration? (Answer: b)

a. The average PageRank value of the page $j$ after the $k^{th}$ step  

b. The probability that the surfer was at the node $j$ at the $(k-1)^{th}$ step  

c. The highest PageRank value of the page $j$ after the $k^{th}$ step  

d. The lowest PageRank value of the page $j$ after the $k^{th}$ step