CSE4/587 Final Exam Review Spring 2017

1 GRAPH PROCESSING USING MR (15 POINTS)

Chapter 5 in Lin and Dyer's text. Given a graph data, trace the MR execution for, say, shortest path. This is like the discussion in the lecture and the accompanying handout of numerical representation of a small graph.

2 MAPREDUCE ALGORITHMS (20 POINTS)

Word co-occurrence; Page Rank; given a problem with data you will trace the execution of the algorithm and provide the values of the intermediate steps and the final result. Given a problem write a MR solution similar to PageRank. Chapter 5 of Lin and Dyer.

3 MR ABSTRACTION THROUGH PIG DATA FLOW (15 POINTS)

Given a problem you will write a PIG script. Given a problem provide a simple PIG script. Some foundational operations of PIG.

4 SPARK ARCHITECTURE AND DATA FLOW (20 POINTS)

Spark programming model; RDD: read the paper; transformations and actions on RDD; SC operations. Spark eco systems: supporting APIs. How fault-tolerance is accomplished vs Hadoop. Solving problem data flow approach.

5 NAÏVE BAYES (15 POINTS)

Chapter 3 from Doing Data Science text. Given a problem solve it using Naïve Bayes.

6 LOGISTIC REGRESSION (15 POINTS)

Chapter 3 from Doing data Science text. Given a problem and needed tables solve it using Logistic regression.

7 DATE, TIME, LOCATION

Location: NSC 225

Date 5/15/2016

Time: 7.00 – 10.00PM

Closed book exam.