Mining Big Data - Map-Reduce

Assignment 1 - Hadoop

M. Vincent & A. Phansalkar
a1148120 &

School of Computer Science, The University of Adelaide

March 20, 2020

Report submitted for **COMP SCI 3306 Mining Big Data** at the School of Computer Science, University of Adelaide

Exercise 1

In this exercise, we consider an attempt to identify suspected evil-doers from a population of 5 billion people, observed over a period of 5,000 days. For a pair of individuals to be suspected of evil-doing, they must visit the same hotel at the same time, on four different days.

The number of hotels is set at 1% of the population, hence there are $500,000$ hotels. The chance of two individuals being at the same hotel on the same day is .0001. The chance that they will visit the same hotel on the same day, is this divided by the number of hotels, which is 2^{-10}. The chance that a pair will visit the same hotel on 4 different days is this to the power of four, which is 1.6^{-39}.

Using the fact that $\binom{n}{2}$ can be approximated by $n^2/2$, we find that the number of pairs of people is $\binom{5^{10}}{2} = 1.25^{19}$, and the number of pairs of days is $\binom{5000}{2} = 1.25^{7}$. The expected number of events that look like evil-doing is then calculated as,

$$1.25^{19} \times 1.25^{7} \times 1.6^{-39} = 2.5^{-13} \tag{1}$$

The expected number of events that look like evil-doing is effectively zero, using the above definition of suspected evil-doing.

Exercise 2

Exercise 3

Exercise 4

Exercise 5