Assignment 2 solutions

Heath Rusby – a1708147 – a1708147@student.adelaide.edu.au
Zhao Ming Soh - a1751699 - a1751699@student.adelaide.edu.au

Exercise 1

1. r=3, b-10

s	$1-(1-s^r)^b$
0.1	0.0100
0.2	0.0772
0.3	0.2394
0.4	0.4839
0.5	0.7369
0.6	0.9123
0.7	0.9850
0.8	0.9992
0.9	1.0000

2. r=6, b=20

s	$1-(1-s^r)^b$
0.1	0.0000
0.2	0.0013
0.3	0.0145
0.4	0.0788
0.5	0.2702
0.6	0.6154
0.7	0.9182
0.8	0.9977
0.9	1.0000

3. r=5, b=50

s	$1-(1-s^r)^b$
0.1	0.0005
0.2	0.0159
0.3	0.1145
0.4	0.4023
0.5	0.7956
0.6	0.9825
0.7	0.9999
0.8	1.0000
0.9	1.0000

Exercise 2

1.

n = 10 billion bits, m = 2 billion members, k = 3 hash functions:

Probability of false-positive = $\left(1 - e^{-\frac{km}{n}}\right)^k = \left(1 - e^{-\frac{3\times2\times10^9}{10x10^9}}\right)^3 = (1 - e^{-0.6})^3 = 9.18\%$

n = 10 billion bits, m = 2 billion members, k = 4 hash functions:

Probability of false-positive = $\left(1 - e^{-\frac{km}{n}}\right)^k = \left(1 - e^{-\frac{4\times2\times10^9}{10x10^9}}\right)^4 = (1 - e^{-0.8})^4 = 9.20\%$

2. Number (k) of hash functions that minimize false-positive rate.

There are two competing factors at play in this optimization problem. Using more hash functions gives us more chances to find a 0 bit for an element that is not a member of S, but using fewer hash functions increases the fraction of 0 bits in S.

Letting the probability of a false-positive be P(k):

$$P(k) = \left(1 - e^{-k\frac{m}{n}}\right)^k$$
$$= e^{\mathrm{kln}\left(1-e^{-k\frac{m}{n}}\right)}$$

Therefor minimising $\mathrm{kln}\left(1 - e^{-k\frac{m}{n}}\right)$ will minimise P(k). Letting $g(k) = \mathrm{kln}\left(1 - e^{-k\frac{m}{n}}\right)$ and keeping in mind that both m and n are positive the minimum of g occurs at g'(k) = 0.

$$\frac{\mathrm{d}}{\mathrm{dk}}\mathrm{k} \cdot \ln\left(1 - e^{-k\frac{m}{n}}\right) = 0$$
$$\therefore \ln\left(1 - e^{-k\frac{m}{n}}\right) + \frac{k}{1 - e^{-k\frac{n}{m}}} \cdot \frac{d}{dk}\left(1 - e^{-k\frac{m}{n}}\right) = 0$$
$$\therefore \ln\left(1 - e^{-k\frac{m}{n}}\right) + \frac{kn}{m}\frac{e^{-k\frac{m}{m}}}{1 - e^{-k\frac{m}{m}}} = 0$$
$$\therefore k = \frac{n}{m}\ln 2$$

From this it can be seen that g'(k) = 0 at $k = \frac{n}{m}\ln 2$, thus this k minimises g(k) which in-turn minimises P(k).

Exercise 3

My program makes use of 3 map-reduce jobs:

Job 1: Initialises input for Job 2 by grouping the neighbours of each node and initialising page ranks

Job 2: Performs PageRank calculation. Is setup up so it's output can be passed back into its input and hence called iteratively until ranks converge. All outputs generated by these iterations are not sorted (due to size constraints I have only included a selection of iteration outputs..

Job 3: Makes use of Map-Reduces inbuilt prioritisation to keys to sort the nodes into ascending order of page rank. i.e. **the higher page ranks are found at the bottom of the final output (found in outputs/rfinal)**

My results can be found below:

Input File: *web-Google.txt* found at http://snap.stanford.edu/data/web-Google.html
Number of Nodes: 875713
$\boldsymbol{\beta}$ = 0.85
Iterations = 50
Runtime: 26 minutes

Top 10 Ranked Pages:

Rank	Page
6.44E-04	597621
6.42E-04	41909
6.31E-04	163075
6.27E-04	537039
5.49E-04	384666
5.34E-04	504140
5.07E-04	486980
5.02E-04	605856
4.97E-04	32163
4.96E-04	558791

*All files relating to this exercise can be found in the PageRank1 folder.

Exercise 4

1. Scenario 1:
   (a) Question 1:

x	h(x) = (2x+1) mod 32	5-bit Binary
3	7	00111
1	3	00011
4	9	01001
6	13	01101
5	11	01011
9	19	10011

   i. Maximum Tail Length = 0

   ii. Estimate Number of Distinct Elements = $2^0 = 1$

   (b) Question 2:

x	h(x) = (3x+7) mod 32	5-bit Binary
3	16	10000
1	10	01010
4	19	10011
6	25	11001
5	22	10110
9	2	00010

   i. Maximum Tail Length = 4

   ii. Estimate Number of Distinct Elements = $2^4 = 16$

   (c) Question 3:

x	h(x) = (4x) mod 32	5-bit Binary
3	12	01100
1	4	00100
4	16	10000
6	24	11000
5	20	10100
9	4	00100

   i. Maximum Tail Length = 4

   ii. Estimate Number of Distinct Elements = $2^4 = 16$

2. Scenario 2:
   (a) Question 4:

x	h(x) = (6x+2) mod 32	5-bit Binary
4	26	11010
5	0	00000
6	6	00110
7	12	01100
10	30	11110
15	28	11100

i. Maximum Tail Length = 5

ii. Estimate Number of Distinct Elements = $2^5 = 32$

(b) Question 5:

x	h(x) = (2x+5) mod 32	5-bit Binary
4	13	01101
5	15	01111
6	17	10001
7	19	10011
10	25	11001
15	3	00011

i. Maximum Tail Length = 0

ii. Estimate Number of Distinct Elements = $2^0 = 1$

(c) Question 6:

x	h(x) = (2x) mod 32	5-bit Binary
4	8	01000
5	10	01010
6	12	01100
7	14	01110
10	20	10100
15	30	11110

i. Maximum Tail Length = 3

ii. Estimate Number of Distinct Elements = $2^3 = 8$