Compsci 3306 - Assignment 1 - < A1-groups 49 >

Zhao Ming Soh
a1751699

Charlie Hibbert
a1745714

March 31, 2021

Exercise 1

1. Given that the "Probability that 2 people would visit a hotel on any given days" :
   (a) $\frac{1}{100} \cdot \frac{1}{100} = \mathbf{1 \cdot 10^{-4}}$
2. Given that the "Chance that they will visit the same hotel on one given day" :
   (a) $\frac{1 \cdot 10^{-4}}{5 \cdot 10^{5}} = \mathbf{2 \cdot 10^{-10}}$
3. Given that the "Chance that they will visit the same hotel on 4 different days" :
   (a) $(\mathbf{2 \cdot 10^{-10}})^4 = \mathbf{1.6 \cdot 10^{-39}}$
4. The number of suspected pairs would be :
   (a) The approximation for the number of pairs of people:
      i. $\mathbf{n} = (\mathbf{5 \cdot 10^{9}})^2$
      ii. $\frac{\mathbf{n}}{2} = \frac{(5 \cdot 10^{9})^2}{2} = \mathbf{1.25 \cdot 10^{19}}$
   (b) The approximation for the 4 different days :
      i. $\mathbf{n} = (\mathbf{5000})^4$
      ii. $\frac{\mathbf{n}}{4!} = \frac{(5000)^4}{4 \cdot 3 \cdot 2 \cdot 1} = \mathbf{2.6042^{13}}$
   (c) **The result :**
      - $(\mathbf{1.25 \cdot 10^{19}}) \cdot (\mathbf{2.6042^{13}}) \cdot (\mathbf{1.6 \cdot 10^{-39}}) = \mathbf{5.2083 \cdot 10^{-7}}$

Exercise 2

Exercise 3

1. For the friend recommendation system I have used two MapReduce jobs.
   (a) The first mapper takes the input from the provided txt. Splits the 'User' and 'Friends' into the individual (key, value) format = ('User', r = 'recommendedFriend';m=x). Where x = 0 if 'User' and 'recommendedFriend' are already friends. x = 1 if 'User'

and 'recommendedFriend' are mutual friends. (using a nested loop). This is the outputted.

(b) The first reducer takes the input. Using a 2d ArrayList, all mutual friends are tallied up. If m=0 for any input, this person is deleted from the ArrayList as they are already friends and do not need to be in the recommendation system. The output is then in (key, value) format = ('User', r = 'recommendedFriend';m=x). Where x = no. of mutual friends.

(c) In the second mapper there is no change to the format.

(d) The second reducer ouptputs the top 10 recommended friends by most mutual friends. Using a class ArrayList, all recommended friends are added. Using a compare class and the collection.sort() function, the top recomennded friends are ordered in the descending order in the ArrayList. Then the top 10 recommended friends are outputted for each user.

2. The recommendations for the users in with the following user IDs :

   (a) 924 : 439,2409,6995,11860,15416,43748,45881

   (b) 8941 : 8943,8944,8940

   (c) 8942 : 8939,8940,8943,8944

   (d) 9019 : 9022,317,9023

   (e) 9020 : 9021,9016,9017,9022,317,9023

   (f) 9021 : 9020,9016,9017,9022,317,9023

   (g) 9022 : 9019,9020,9021,317,9016,9017,9023

   (h) 9990 : 13134,13478,13877,34299,34485,34642,37941

   (i) 9992 : 9987,9989,35667,9991

   (j) 9993 : 9991,13134,13478,13877,34299,34485,34642,37941

Exercise 4

4.1 Part 1

1. Mapreduce result for pg100.txt :

   (a) Console output is in the Assignment_1/Exercise_4/Part_1

   (b) Output :

```
0       119383
1       35527
10      16471
11      8532
12      3310
13      1748
14      624
15      385
16      149
17      54
18      23
19      10
2       139003
20      9
21      4
22      3
23      3
24      2
27      1
28      2
29      1
3       173743
33      1
34      1
37      2
38      2
4       189110
40      1
5       121442
6       80604
7       61659
8       44175
9       27460
```

Figure 1: part_1_pg100_lowercase.txt mapreduce result

1. Mapreduce result for 3399.txt :
   (a) Console output is in the Assignment_1/Exercise_4/Part_1
   (b) Output :

```
0       3782
1       13417
10      7373
11      4232
12      2305
13      1210
14      560
15      311
16      133
17      48
18      35
19      16
2       61021
20      9
21      5
22      7
23      2
24      3
25      2
3       71353
39      1
4       52898
40      1
5       35874
54      1
6       25171
7       21830
8       14716
9       10098
```

Figure 2: part_1_pg3399_lowercase.txt mapreduce result

4.2 Part 2

(q1) There are 16471 number of words with a length of 10 in the FirstInputFile(pg100.txt).

(q2) There are 189110 number of words with a length of 4 in the FirstInputFile(pg100.txt).

(q3) The longest words length that the FirstInputFile has is 40 and its frequency is 1.

(q4) There are 61021 number of words with a length of 2 in the SecondInputFile(3399.txt).

(q5) There are 35874 number of words with a length of 5 in the SecondInputFile(3399.txt).

(q6) The most frequent length in the SecondInputFile is 3 and its frequency is 71353.

4.3 Part 3

1. Mapreduce result for pg100.txt :
   (a) Console output is in the Assignment_1/Exercise_4/Part_3
   (b) Output :

```
0       1
1       28
10      2264
11      1456
12      780
13      393
14      184
15      83
16      30
17      14
18      13
19      5
2       195
21      2
22      2
23      1
24      1
26      1
27      1
28      2
3       674
30      1
33      2
4       1938
5       3377
6       4367
7       4931
8       4391
9       3344
```

Figure 3: part_3_pg100_lowercase_and_nopunctuation.txt mapreduce result

1. Mapreduce result for 3399.txt :
   (a) Console output is in the Assignment_1/Exercise_4/Part_3
   (b) Output :

```
0	1
1	28
10	1737
11	1179
12	692
13	411
14	198
15	91
16	41
17	24
18	9
19	7
2	89
20	3
21	3
22	1
23	2
3	341
32	1
33	1
4	1209
44	1
5	1824
6	2567
7	2815
8	2801
9	2321
```

Figure 4: part_3_pg3399_lowercase_and_nopunctuation.txt mapreduce result

4.4 Part 4

(q1) There are 2264 number of words with a length of 10 in the FirstInputFile(pg100.txt).

(q2) There are 1938 number of words with a length of 4 in the FirstInputFile(pg100.txt).

(q3) The most frequent length in the FirstInputFile is 7 and its frequency is 4931.

(q4) There are 89 number of words with a length of 2 in the SecondInputFile(3399.txt).

(q5) There are 1824 number of words with a length of 5 in the SecondInputFile(3399.txt).

(q6) The second most frequent length in the SecondInputFile is 8 and its frequency is 2801.