Assignment 1: Basics and Map Reduce

Group 1 Jin Yin a1761810. Yingyao Lu a1784870

Exercise1, Suspected Pairs (10 points)

1, The number of days of observation is 5000; and

2, The number of people observed was raised to 5 billion (and there were therefore 500, 000 hotels); and

3, We only reported a pair as suspect if they were at the same hotel at the same time on four different days.

In this case, the chance that they will visit the same hotel on one given day is:

Four days the, the possibility is

The number of pairs in the people is

The number of pairs of days is

So, the amount of the people is

Exercise 2, Hadoop (20 points)

Standalone mode:

The output is in the zip files

Exercise 3, Friend Recommendation System (Stanford) (35 points)

1, Source code is in the zip

2, Algorithm to tackle this problem:

The basic idea is that if A have 5 friends, let’s they are V, W, X, Y, Z, every two of these 5 friends must have a common friend which is A, so every pair of these 5 people could be each other’s recommended friend, such as recommend V to W, and recommend W to V.

Meanwhile, if A have these 5 friends, so although A and one of these 5 persons appear in another friend list, these two people can’t be recommended to each other. For example, A and V are common friends of B. A should not be recommended to V, vice versa.

In this case, we can use one person as a key, recommended friend and mutual friend number as value to Map task. If the friendship is confirmed, it means they don’t need recommend each other, we mark it as -1. Otherwise, we mark the mutual friend number as 1.

Basing on the friendship graph is undirected, if V in A’s friend list, A must be in the V’s list, so we only need to record on (A, {V, -1}) once.

However, basing on the friend list, V and W may recommend to each other. so (V, {W, 1} and (W, {V, 1}) should be sent to Reduce task.

In Reduce task, we count the number of recommended friends. If there is -1 appear in mutual friend number, they should not be recommended to the person.

During the Reduce task, sort algorithm is implemented in SortedMap.

At last, the output of the <k, v> pair will be the result.

3, The recommendations for the users with following user IDs:

|  |  |
| --- | --- |
| ID | Recommended friend |
| 924 | 439,2409([926]),6995([925]),11860([926]),15416([926]),  43748([926]),45881([926]) |
| 8941 | 8943,8944([8938, 8939]),8940, |
| 8942 | 8939,8940,8943([8938]),8944([8938]), |
| 9019 | 9022,317,9023([320]), |
| 9020 | 9021,9016,9017([320, 9019]),9022([9018, 320]),317,9023([320]) |
| 9021 | 9020,9016,9017([320, 9019]),9022([9018, 320]),317,9023([320]), |
| 9022 | 9019,9020([320, 9018]),9021([9018, 320]),317,  9016([320]),9017([320]),9023([320]), |
| 9990 | 13134,13478([35667]),13877([35667]),34299([35667]),  34485([35667]),34642([35667]),37941([35667]), |
| 9992 | 9987,9989([9994, 9990, 9988, 9993]),35667,9991, |
| 9993 | 9991,13134,13478([35667]),13877([35667]),34299([35667]),  34485([35667]),34642([35667]),37941([35667]), |

s

Exercise 4, MapReduce (15 points)

Part 1:

Output is in the zip

Part 2:

Q1: How many words are there with length 10 in FirstInputFile?

There are **11459** words with length 10 in the FirstInputFile.

Q2: How many words are there with length 4 in FirstInputFile?

There are **205024** words with length 4 in the FirstInputFile.

Q3: What is the longest length between words and what is its frequency in FirstInputFile?

**33** is the longest length, and the frequency is **2**

Q4: How many words are there with length 2 in SecondInputFile?

There are **63500** words with length 2 in the SecondInputFile

Q5: How many words are there with length 5 in SecondInputFile?

There are **35025** words with length 5 in SecondInputFile

Q6: What is the most frequent length and what is its frequency in SecondInputFile?

the most frequent length is **3**, and its frequency is **72428**

Part 3: The output is in the zip

Part 4:

Q7: How many words are there with length 10 in FirstInputFile?

There are **2274** words with length 10 in the FirstInputFile.

Q8: How many words are there with length 4 in FirstInputFile?

There are **1940** words with length 4 in the FirstInputFile.

Q9: What is the most frequent length and what is its frequency in FirstInputFile?

The most frequent length is **7**, and the frequency is **4910**

Q10: How many words are there with length 5 in SecondInputFile?

There are **1824** words with length 5 in SecondInputFile

Q11: How many words are there with length 2 in SecondInputFile?

There are **89** words with length 2 in SecondInputFile

89

Q12: What is the second-most frequent length and what is its frequency in SecondInputFile?

The second-most frequent length is **8**, and its frequency is **2801**

Exercise 5, Summary of 2.4 and 2.5 (10 +10 points) (Postgraduate Students (COMP SCI 7306) only)