EE412 Foundation of Big Data Analytics, Fall 2021
HW1

Name: Cao Viet Hai Nam

Student ID: 20200817

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

Answer to Problem 1:

* Time elapsed : 87.8545594215393 s

* Explanation about the algorithm:

- Since we only care about those who are not friend of each other, we will map user and his friend to value 0 (map (user, friend): 0) and map each couple in user's friends list to 1 for counting consideration since they might not know each other ((friend_i, friend_j) : 1).
- After flatMap and group by key, we filter out all those couples having 0s in the list (meaning they know each other). The remaining are just unknown couples and count of their mutual friends.
- Using python sort function, sort all the tuple ((a,b), count) based on count in descending order and if tie, sort by the first user ID integer in ascending order and then the second ID. (time complexity: O(nlogn))
- Pick the top 10 from the sorted list and print it out.

Answer to Problem 2:

a, - Since there are N frequent items, 4N bytes are needed to store N intergers in the main memory in pass 2.

- If we use triangular-matrix method, we need $4\frac{N^2}{2} = 2N^2$ bytes.

- If we use hash table, since there are M frequent pairs and 1,000,000 frequent items, there will be $(M + 10^6)$ tripples, which takes $12(M + 10^6)$ bytes (each tripple takes 12 bytes).

Thus, the minumum bytes a memory needed to execute A-piori algorithm is $\min(4N + 2N^2, 4N + 12(M + 10^6))$

b,

***** Time elapsed: 0.9260931015014648 s

Answer to Problem 3

a, Let the family of minhash function be (d1,d2, p1, p2)-sensitive
Applying:

- A 2-way AND construction followed by a 3-way OR construction.

Result function: (d1,d2, $(1-(1-p_1^2)^3, (1-(1-p_2^2)^3$) - sensitive

- A 3-way OR construction followed by a 2-way AND construction.

Result function: (d1,d2, $((1-(1-p_1)^3)^2, ((1-(1-p_2)^3)^2$) - sensitive

- A 2-way AND construction followed by a 2-way OR construction, followed by a 2-way AND construction.

Result function: (d1,d2, $(1-(1-p_1^2)^2)^2$, $(1-(1-p_2^2)^2)^2$ - sensitive

- A 2-way OR construction followed by a 2-way AND construction, followed by a 2-way OR construction followed by a 2-way AND construction.

Result function: (d1,d2, $(1-(1-(1-(1-p_1)^2)^2)^2)^2, 1-(1-(1-(1-p_1)^2)^2)^2)^2$) - sensitive

b,

* Time elapsed: 24.060283184051514s