HW1 - Data Mining - due by 10 am 30 January 2019

Homework Policy:

1. You can discuss your HW with at most one other group. However, discussions are restricted to oral only. Written similarity is regarded as plagiarism. If you group discusses with other groups, you must acknowledge the discussion in your written solution.
2. Homework submission: one person in the group submitting the homework is ok, but all the names must appear on the paper of the written homework.
3. The written homework must be submitted by the due date. Please turn in a hard copy in class. The hard copy will be graded.
4. This list may be updated further.

Problem 1 (10 points) (Exercise 1.2.1 MMDS book) In this problem, we revisit the example we learn in class with different parameters. Suppose that you have a data of 1 billion people going to 10^5 hotels in 1000 days, and you hope to find "evil doers" in your data. To find "evil doers", you will find pairs of people who went to the same hotel in two different days, assume further that everyone gets to a hotel in 100 days. In class we showed that the expected number of suspected "evil doer" pairs is $250,000$.

(a) [5 points] What would be the number of suspected pairs if the number days was raised to $2,000$.

(b) [5 points] What would be the number of suspected pairs if the number of people observed was raised to 2 billion **and** the number of hotels is $200,000$.

Problem 2 (10 points) (Exercise 1.3.3 MMDS book) Suppose hash-keys are drawn from the population of all non-negative integers that are multiples of some constant c, and hash function $h(x)$ is $h(x) = x \mod 15$. For what values of c will $h(.)$ be a suitable hash function, i.e., a large random choice of hash-keys will be divided roughly equally into buckets?

Problem 3 (10 points) In class, we showed that if the MinHash signature has size r, the probability that two different items I_1, I_2 of similarity x have the same signature is x^r.

(a) [5 points] Prove that if our data structure has b hash tables, each uses signatures of size r as keys, then the probability that I_1 and I_2 are hashed in the same location of *some* hash table is $1 - (1 - x^r)^b$.

(b)[5 points] In class, we discussed how to choose r and b by fixing a threshold x so that $1 - (1 - x^r)^b = \frac{1}{2}$. Prove that in this case, b and r *approximately* satisfy:

$$b = \frac{\ln(2)}{x^r} \tag{1}$$

where ln(.) is the natural logarithm.

For example, when $x = 0.6$ and $r = 6$, we have $b \sim 14$. That's how I choose x, r and b in the programming assignment 1.

Problem 4 (20 points) (Exercise 6.1.1 - 6.1.2-6.1.5 MMDS book) Suppose there are 100 items, numbered 1 to 100, and also 100baskets, also numbered 1 to 100. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, and so on. Basket 12 consists of items $\{1, 2, 3, 4, 6, 12\}$, since these are all the integers that divide 12. Answer the following questions:

(a)[5 points] If the support threshold is 5, which items are frequent?

(b) [5 points] If the support threshold is 5, which pairs of items are frequent?

(c) [5 points] Which basket is the largest?

(d) [5 points]What is the confidence of the following association rules $\{5, 7\} \rightarrow 2$, $\{2, 3, 4\} \rightarrow$ 5.