# CTPS-315 — Homework-01

**Charles Norden, #011606177**

---

## Analytic Part (40pts)

### Question #1

1. The *absolute* support of set $\{A, B\}$ is 4.
2. The *relative* support of set $\{A, B\}$ is 0.6.
3. The confidence of association rule $A => B$ is:

$$\frac{supp(\{A, B\})}{supp(\{A\})} = \frac{4}{6} = 0.6$$

### Question #2

1. Given a dataset of size $n = 20$ and pair[7,8], the actual location in the ragged 1-Dim array is:

$$(i - 1)(n - \frac{i}{2}) + j - i) = [99]$$

2. Suppose we know that only 10% of the total pairs will have a non-zero count; it is ideal that we use the tabular method. Tabular method beats triangular matrix when *at most 1/3 of all pairs have a non-zero count.* In this case, we know for sure that 1/10 of all pairs have non-zero counts, thus we should use tabular method.

### Question #3

Given the six items $\{1, 2, 3, 4, 5, 6\}$ and the 12 support baskets, and the support threshold $supp = 4$. First we count the absolute supports for all single-item sets where the $setsize == 1$.

a. Absolute supp: supp(1)=4, supp(2)=5, supp(3)=8, supp(4)=8, supp(5)=6, supp(6)=4

Relative supp: supp(1)=0.36, supp(2)=0.45, supp(3)=0.72, supp(4)=0.72, supp(5)=0.54, supp(6)=0.36

Absolute supp: supp({1,2})=2, supp({1,3})=3,

supp({1,4})=2, supp({1,5})=1,

supp({2,3})=3, supp({2,4})=4,

supp({2,5})=2, supp({2,6})=1,

supp({3,4})=4, supp({3,5})=3,

supp({3,6})=2,

supp({4,5})=3, supp({4,6})=3,

supp({5,6})=2

Relative supp: supp({1,2})=0.18, supp({1,3})=0.27,

supp({1,4})=0.18, supp({1,5})=0.09,

supp({2,3})=0.27, supp({2,4})=0.36,

supp({2,5})=0.18, supp({2,6})=0.09,

supp({3,4})=0.36, supp({3,5})=0.27,

supp({3,6})=0.18,

supp({4,5})=0.27, supp({4,6})=0.27,

supp({5,6})=0.18

b. buck 1 {2,6} {3,4}

buck 2 {1,2} {4,6}

buck 3 {1,3}

buck 4 {1,4} {3,5}

buck 5 {1,5}

buck 6 {2,3}

buck 7 {3,6}

buck 8 {2,4} {5,6}

buck 9 {4,5}

buck 10 {2,5}

c. Pairs in bucket 1, 2, 4, 8 are counted on the second pass.

**Question #4**

Digital copies and plagiarism were the chief motivation for the author of the paper to investigate and develop a efficient algorithm to check for fingerprints, i.e. whether particular documents contain identitical snippets of text information. The authors pointed out the problem with identifying the integrity in parital documents; comparing idenitical copies are easy but it's a completely different story with snippets laced within a larger documents along with other unrelated content.

The other discussed various techniques and their weakness and finally introduced their own solution, winowing, which has a performance that stays within 33

## Programming & Experimental Part (60pts)

**Solution:**

See enclosed source code.