Question 1 :

① Inverted index posting lists for terms 'cat' and 'dog' as follows :

cat : 234569 : 1, 234578 : 1, 234839 : 1
dog : 234569 : 1, 234578 : 1, 234879 : 1

② Here is the compressed form for 'dog' :

dog : 234569 : 1, 9 : 1, 301 : 1 [using gaps]

In binary :

234569 = b111001 0100 0100 1001
1 = b1
9 = b1001
1 = b1
301 = b1 0010 1101
1 = b1

VB encoding :

00001110 00101000 11001001
10000001
10001001
10000001
00000010 10101101
10000001

③ We compute the document vector magnitude as follows

Document ID - 234569

Text : phrase fight cat dog reflect natural tendency relationship two species antagonistic two species friend

maxf : 2 (two, species)

$\therefore$ vector magn : $\sqrt{(\frac{1}{2})^2+(\frac{1}{2})^2+(\frac{1}{2})^2+(\frac{1}{2})^2+(\frac{1}{2})^2+(\frac{1}{2})^2+(\frac{1}{2})^2+(\frac{1}{2})^2+(\frac{2}{2})^2+(\frac{2}{2})^2+(\frac{1}{2})^2+(\frac{1}{2})^2}$

$= 2.12$

Doc ID: 234578

Text : Dog cat bad relationship

maxf : 1

Vec. magn : $\sqrt{4} = 2$

DOC ID : 234839

Text : Cat furry

maxf : 1

Vec. magn : $\sqrt{2} = 1.41$

DOC ID : 234879

Text : Dog man best friend

maxf := 1

Vec. magn : $\sqrt{4} = 2$

q : cat dog

maxf : 1

Vec. magn : $\sqrt{2} = 1.41$

$\therefore \cos(q, d_{234569}) = \frac{(\frac{1}{2}\times\frac{1}{1})+(\frac{1}{2}\times\frac{1}{1})}{1.41 \times 2.12}$

$= 0.33$

$$\cos(q, d_{234578}) = \frac{1+1}{1.41 \times 2}$$

$$= 0.71$$

$$\cos(q, d_{234839}) = \frac{1}{1.41 \times 1.41}$$

$$= 0.50$$

$$\cos(q, d_{234879}) = \frac{1}{1.41 \times 2}$$

$$= 0.35$$

∴ Document 234578 is most similar.

Question 2:

(a) The largest integer:

```
Map(id, i)
    emit(1, i)

Combine(j, list)
    emit(1, max(list))

Reduce(j, list)
    emit(max(list), null)
```

(b) The average of all integers?

```
Map(id, i)
    emit(1, i)

Combine(j, list)          → n = sizeof(list)
    emit(1, (Σ_n i, n))

Reduce(j, list)           → [(Σ_{n1} i, n1), (Σ_{n2} i, n2), ....]
    emit( Σ_m (sum_i × n_i) / Σ_m n_i , null )
                          → m = size of (list)
```

(c) Same set of integers (no duplicate)

```
Map(id, i)
    emit(i, null)

Reduce(j, list)
    emit(j, null)
```

[taking advantage of MapReduce ability to group keys]

(d) output from (c) + 2

```
Map2(id, i)
    emit(1, i)

Reduce2(j, list)          → [distinct values]
    emit(sizeof(list), null)
```

Question 3:

(a) Bag Union:

t is tuple, tid is tuple id, x is bit indicating two Relation (R or S)

```
Map ( tid , (t, x))
    emit ( t, x)

Reduce( t, li )
    lr = [x for x in li if x == R]
    ls = [x for x in li if x == s]
    emit ( t, sum (len(lr), len (ls)))
                ↑
          defined in question
```

~~(b) Bag Intersection: (len(lr), len(ls)) min(len(lr), len(ls))~~

(b) Bag Intersection:

```
Map( tid , (t, x))
    emit (t, x)

Reduce( t, li )
    lr = [x for x in li if x == R]
    ls = [x for x in li if x == s]
    c = min(len(lr), len(ls))
    if(c>0)
        emit (t, c)
```

(c) Bag difference

```
Map (tid, (t,x))
    emit (t,x)

Reduce(t, li) → [R,R,S,S,R,S ···]
    lr = [x for x in li if x==R]
    ls = [x for x in li if x==S]
    c = len(lr) - len(ls)
    if (c>0)
        emit(t, c)
```