Assignment 3 for Large Scale Data Mining

Name: Zijian Zhang
Matrikelnr.: 3184680

May 2, 2016

Problem 1

1

The key should be *university*. During the sampling we only maintain tuples that fit the *courseID* demanded and hash each sample with same *university* into same bucket.

2

The key should be the combination of *university* and *studentID*.

3

The key should be the combination of *university* and *courseID*.

Problem 2

(a)

$k = 5 \longrightarrow$ first two buckets from right are counted precisely, the third are only a half token into account.thus $1 + 1 + \frac{1}{2} = 3$. Precise value is also 3, two values are the same.

(b)

$k = 15 \longrightarrow$ the total value of the first four buckets is $1 + 1 + 2 + 4 = 8$, while the half of the fifth bucket is $\frac{4}{2} = 2$, thus in total $8 + 2 = 10$. Correct value is however 9.

Problem 3

If there are three more 1's enter the window:

step 1:

The second and the third 1 bucket are combined together as a 2-size-of bucket.

step 2:

Now we have three bucket of size 2, the second and the third from right of size-2-buckets are combined together as a size-4-bucket.

step 3:

Two of size-4-buckets from left are combined together, now we have from right one of size 1, one of size 2, one of size 4 and at least one of size 8.

Problem 4

Space requirement: $\mathcal{O}(\frac{N}{w}log(w))$, where there are in total $\frac{N}{w}$ buckets and each of which contains a count $= 2^w$.

Maximum error should be $\frac{w}{2}$. Because when the length of $k = tw \pm 1$ where $t \in \mathbb{N}$ the error could be $\frac{w}{2}$ if for the last of the bucket only the half of the count was caculated.