TI2736-C Assignment 4

Joost Pluim, jwpluim, 4162269
Pascal Remeijsen, premeijsen, 4286243

March 17, 2016

A Priori Algorithm

1 Question 5

1.1 Question 5.1

The frequent doubletons are: [[and, dog], [cat, dog], [a, cat], [a, dog], [and, cat]]

1.2 Question 5.2

We have to construct the filtered candidates for every size $1 \ldots k$. This means we have to pass through the data k times.

1.3 Question 5.3

If we would take this approach and lets say we have n words in our dataset, this means constructing n choose k combinations. This would quickly become extremely large. Taking the A Priori approach would exclude items of size k which aren't frequent in the subsequent step, which reduces the number of elements every round.

PCY Algorithm

2 Question 3

2.1 Question 3.1

The difference is that in A Priori we test 10 pairs, while in PCY we test 9. This is because in PCY before we even consider a pair, we check that count of the hash of the pair is higher than the support threshold.

2.2 Question 3.2

You have to consider less pairs, since in step $k - 1$ in which we coincidentally already look into all combinations, we keep track of the counts of pairs. This means less candidates will be constructed thus less time.

2.3 Question 3.3

If the bucket size is too low, more buckets will pass the threshold, so the PCY algorithm doesn't gain much compared to the A Priori algorithm. If it's too big, you just spoil a lot of memory.

References