Web Scale Data Mining homework 1

Karol Kubicki

8.12.2014

2.1) You are analysing a stream which contains s singletons (items occurring once) and d pairs of items, in some random order. You are sampling the stream using Algorithm 2, which uses a pseudorandom number generator and takes each item for further processing with probability 1/100 . What is the expected value of the fraction returned by the algorithm, as a function of s and d? Hint: follow the analysis in Section 4.2.1 of [10] (references in the textbook).

The correct answer would be: $\frac{d}{s+d}$. The answer from algorithm 2:

- number of obtained singletons: $\frac{s}{100}$,
- number of doubles obtained twice: $\frac{d}{100*100} = \frac{d}{10000}$,
- number of doubles obtained once: $d * 2 * \frac{1}{100}\frac{99}{100} = \frac{198d}{10000}$,
- expected value of the fraction returned by the algorithm: $\frac{d}{100s+199d}$.