Mining massive Datasets WS 2017/18

Problem Set 10

Rudolf Chrispens, Marvin Klaus, Daniela Schacherer

January 11, 2018

Exercise 01

Let the input to the hash functions and thus the elements in S be binary strings s. Then possible hash functions for a bloom filter could be

- h1 takes every third position of s starting from position 0, treats them as a number and computes modulo 11
- h2 takes every third position of s starting from position 1, treats them as a number and computes modulo 11
- h1 takes every third position of s starting from position 2, treats them as a number and computes modulo 11

These hash functions are independent from each other as they use different elements of s for computing the hash value.

Exercise 02

a) The probability that a random element ($m = 1$) gets hashed to a given bit in the bit array with $n = 5$ can be computed by the following formula:

$$1 - (1 - \frac{1}{n})^{n(m/n)} = 1 - e^{-m/n} = 1 - e^{-1/5} = 0.1813$$

For $h_1(x)$ each bit is equally likely to be hit. For $h_2(x)$ this is not the case as $2x + 3$ always results in an odd number which will then be taken modulo 5.
Bit array state: | 1 | 0 | 0 | 0 | 1 |

b) With $k = 2, n = 5$ and $m = 1$ unknown the probability for false positives is:

$$(1 - e^{-km/n})^k = (1 - e^{-2/5})^2 = 0.109$$

Table 1 – Example Completeness

price-series-name	timestamp-gap-start	timestamp-gap-end
t2.micro___Linux-...	2017-11-29 02:00:26	2017-11-28 02:00:06

Exercise 03

With the same formula as in Exercise 2b we receive with $n = 8$ billion, $m = 1$ billion and $k = 3$

$$(1 - e^{-km/n})^k = (1 - e^{-3*1/8})^3 = 0.0306$$

with $k = 4$ we get

$$(1 - e^{-4*1/8})^4 = 0.024$$

Exercise 03

```
a)
see function "checkCoherency" in ex10_6.py
b)
see function "checkCompleteness" in ex10_6.py
c) timeseries related to prices- ca-central
checkCoherency
number of found inconsistencies:  27
checkCompleteness
number of found inconsistencies:  37
```