| Stream | $h_1(x)$ | $h_1(x)$ binär | $r(h_1(x))$ | $h_2(x)$ | $h_2(x)$ binär | $r(h_2(x))$ | $h_3(x)$ | $h_3(x)$ binär | $r(h_3(x))$ |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 7 | 00111 | 0 | 16 | 10000 | (4) | 12 | 01100 | 2 |
| 1 | 3 | 00011 | 0 | 10 | 01010 | 1 | 4 | 00100 | 2 |
| 4 | 9 | 01001 | 0 | 19 | 10011 | 0 | 16 | 10000 | (4) |
| 1 | 3 | 00011 | 0 | 10 | 01010 | 1 | 4 | 00100 | 2 |
| 5 | 11 | 01011 | 0 | 22 | 10110 | 1 | 20 | 10100 | 2 |
| 9 | 19 | 10011 | 0 | 4 | 00100 | 2 | 6 | 00110 | 1 |
| 2 | 5 | 00101 | 0 | 13 | 01101 | 0 | 8 | 01000 | 3 |
| 6 | 13 | 01101 | 0 | 25 | 11001 | 0 | 24 | 11000 | 3 |
| 5 | 11 | 01011 | 0 | 22 | 10110 | 1 | 20 | 10100 | 2 |

-> $h_1(x) \Rightarrow R = 0$
$h_2(x) \Rightarrow R = 4$
$h_3(x) \Rightarrow R = 4$

Estimated number of distinct elements:

for $h_1(x) \Rightarrow 2^R = 2^0 = 1$
for $h_2(x) \Rightarrow 2^R = 2^4 = 16$
for $h_3(x) \Rightarrow 2^R = 2^4 = 16$

-> For a $h(x)$ of the form $h(x) = (ax+b) \bmod 2^u$ you should be careful about the values "a" and "b"! -> you should insert numbers for them, that also even numbers are computed! -> $h_1(x)$ only generates uneven numbers, consequently there are no trailing 0's!

Exercise 1

## Exercice 2

Stream: 3 4 1 3 4 2 1 2

$x_1$ $x_2$ $x_3$ $x_4$ $x_5$ $x_6$ $x_7$ $x_8$

$x_1$: $x_1.el = 3$ , $x_1.val = 2$

$x_2$: $x_2.el = 4$ , $x_2.val = 2$

$x_3$: $x_3.el = 1$ , $x_3.val = 2$

$x_4$: $x_4.el = 3$ , $x_4.val = 1$

$x_5$: $x_5.el = 4$ , $x_5.val = 1$

$x_6$: $x_6.el = 2$ , $x_6.val = 2$

$x_7$: $x_7.el = 1$ , $x_7.val = 1$

$x_8$: $x_8.el = 2$ , $x_8.val = 1$

-> In large streams it doesn't make sense -> we should split the stream in smalle pieces

## Exercise 3

Beispiel für $v = 5$ , $n = 15$

Stream: 1, 2, 3, 2, 4, 1, 3, 4, 1, 2, 4, 3, 1, 1, 2

a) $v = 5$ random positions: $x_1 = 2$ , $x_2 = 4$ , $x_3 = 5$, $x_4 = 7$, $x_5 = 9$

$k = 2 \rightarrow n \cdot (2 \cdot c - 1)$

| | | |
|---|---|---|
| $x_1.el = 2$ | $x_1.var = 4$ | $15 \cdot (2 \cdot 4 - 1) = 105$ |
| $x_2.el = 2$ | $x_2.var = 3$ | $15 \cdot (2 \cdot 3 - 1) = 75$ |
| $x_3.el = 4$ | $x_3.var = 3$ | $15 \cdot (2 \cdot 3 - 1) = 75$ |
| $x_4.el = 3$ | $x_4.var = 2$ | $15 \cdot (2 \cdot 2 - 1) = 45$ |
| $x_5.el = 1$ | $x_5.var = 3$ | $15 \cdot (2 \cdot 3 - 1) = 75$ |

Estimate S of 2nd moment is:

$105 + 75 + 75 + 45 + 75 = \frac{375}{5} = 75$

$\frac{375}{v}$

$k = 3 \rightarrow n \cdot (3 \cdot c^2 - 3c + 1)$

$x_1$: $15 \cdot (3 \cdot 4^2 - 3 \cdot 4 + 1) = 555$

$x_2$: $15 \cdot (3 \cdot 3^2 - 3 \cdot 3 + 1) = 285$

$x_3$: $15 \cdot (3 \cdot 3^2 - 3 \cdot 3 + 1) = 285$

$x_4$: $15 \cdot (3 \cdot 2^2 - 3 \cdot 2 + 1) = 105$

$x_5$: $15 \cdot (3 \cdot 3^2 - 3 \cdot 3 + 1) = 285$

3rd moment:

$555 + 285 + 285 + 105 + 285 = \frac{1515}{5} = 303$

Fragen:

Difference 3a) / b)?

## Exercise 4

Estimate of $k^{th}$ moment: $n \cdot (c^k - (c-1)^k)$

$\Rightarrow k = 4 \Rightarrow n \cdot (c^4 - (c-1)^4)$

$= n \cdot (c^4 - (c^4 - 4c^3 + 6c^2 - 4c + 1^4))$

$= n \cdot (4c^3 - 6c^2 + 4c - 1)$ $4^{th}$ moment

## Exercise 5

a) The stream is splitted into batches of 1 second.

The batch gets treated by Spark as a RDD. So the input sentence is inserted into a RDD.

In the output you can see the current time as a headline and under the headline you can see a list like: (<word>, <wordcount>).

We got 2 hits, where a comma or dot is appended to a word, for example 'mankind.' and 'man,'. This is what we expected, because the lines are splitted by a blank.

We expected, that the word 'a' gets a wordcount of "3", but it got only a "2", because Spark treats 'a' and 'A' as different words so that ('a', 3) is splitted into ('a', 2)
('A', 1)

↓ what we expected

↳ So it seems, that Spark does not "clean" words, that upper- and lowercased words represents a problem

b) It differs extremly from one second to another. We get for example one RDD with: ('thx', 2803199) ('u', 2803199) ('ba', 5606398)

and the next RDD is: ('thx', 4084292) ('u', 4084292) ('ba', 8168584)

↳ The wordcount relations in the RDD are correct, so that the wordcount of 'bai' is twice the count of 'thx' or 'u', but the wordcounts differs between the RDD's!