Big Data Analytics

ESSEC

Olga Klopp

Home work 1: solutions

1. **Install Spark:**

2. **Hash Functions**. Suppose hash-keys are drawn from the population of all non-negative integers that are multiples of some constant c, and hash function $h(x)$ is x mod 15. For what values of c will h be a suitable hash function, i.e., a large random choice of hash-keys will be divided roughly equally into buckets?
   Solution: The hash function, $h(x) = x$ mod 15 can give only values from 0 to 14, so we have 15 buckets numbered from $0, 1 \dots 14$. We have to choose c in such a way that all hash-keys are uniformly distributed into all buckets. The suitable c will be all c co-prime with 15. In this case, we have that there exists α and β such that $\alpha \times c + \beta \times 15 = 1$ and, for any $n \in 0, 1 \dots 14$, we have that $\alpha \times n \times c + \beta \times n \times 15 = n$ (Bézout's identity). So, any value of c, except multiple of 3 or 5, will be suitable. If we take c multiple of 3 or 5, then all the hash-keys will be distributed only in buckets multiple of 3 or 5.

3. **The Base of Natural Logarithms**.

   (a) In terms of e, give approximations to

   - $(1.01)^{500} \approx \exp(0.01 \times 500) = e^5 \approx 148,41$
   - $(1.05)^{1000} \approx \exp(0.05 \times 1000) = e^{50} = 5,184705529 \times 10^{21}$
   - $(0.9)^{40} \approx \exp(-0.1 \times 40) = e^{-4} = 0,018315639$

   (b) Use the Taylor expansion of e^x to compute, to three decimal places:

   - $e^{1/10} = 1,105170918 \approx 1 + 0.1 + 0.1^2/2 = 1,105$
   - $e^{-1/10} = 0,904837418 \approx 1 - 0.1 + 0.1^2/2 - 0.1^3/6 = 0,904833333$
   - $e^2 = 7,389056099 \approx 1 + 2 + (2^2)/2 + (2^3)/6 + (2^4)/4! + (2^5)/(5!) + (2^6)/6! + (2^7)/7! + (2^8)/8! + (2^9)/9! + (2^{10})/10! + (2^{11})/11! = 7,389046016$