

Big Data Analytics in Theorie und Praxis

Prof. Johann-Christoph Freytag, Ph.D. und Prof. Dr. Nicole Schweikardt
Sommersemester 2015

Theorie-Übungsblatt 3

Zu bearbeiten bis 2. Juli 2015

Aufgabe 1: **(25 Punkte)**

Seien $a, z, p \in \mathbb{N}$ mit $a < p$ und $z < p$. Beschreiben Sie einen Algorithmus mit möglichst geringer Laufzeit, der $z^a \bmod p$ mit $\mathcal{O}(\log p)$ Speicherbits berechnet. Weisen Sie nach, dass Ihr Algorithmus tatsächlich $z^a \bmod p$ korrekt bestimmt und höchstens $\mathcal{O}(\log p)$ Speicherbits benutzt. Bestimmen Sie die Laufzeit, die Ihr Algorithmus benötigt.

Aufgabe 2: **(25 Punkte)**

Sie erhalten als Eingabe einen Datenstrom der Länge $n \in \mathbb{N}_{\geq 1}$ und der Form a_1, a_2, \dots, a_n mit $a_i \in \mathbb{N}_{\geq 1}$ für alle $i \in \{1, \dots, n\}$. Sei $z = a_1 + a_2 + \dots + a_n$ die Summe aller Elemente des Datenstroms, wobei z vor dem Einlesen des Datenstroms nicht bekannt ist.

Beschreiben Sie einen Datenstromalgorithmus, der genau ein Tupel (i, a_i) mit $i \in \{1, \dots, n\}$ bestimmt. Hierbei soll die Wahrscheinlichkeit, dass das Tupel (i, a_i) ausgewählt wird, exakt $\frac{a_i}{z}$ betragen. Außerdem soll Ihr Algorithmus einen Zwischenspeicher von maximal $\mathcal{O}(\log z)$ Bits benötigen.

Weisen Sie nach, dass Ihr Algorithmus korrekt arbeitet.

Aufgabe 3: **(25 Punkte)**

Beweisen Sie den Satz zum Algorithmus Reservoir-Sampling- s aus der Vorlesung.

Hinweis: Führen Sie einen Beweis per Induktion nach n durch.

Auf der Rückseite finden Sie eine weitere Aufgabe.

Aufgabe 4:**(25 Punkte)**

Sie erhalten einen Datenstrom mit Tupeln der Form

(Universität, Veranstaltungsnummer, Matrikelnummer, Note),

welche jeweils für einen Studierenden, mit der angegebenen Matrikelnummer, an der zugehörigen Universität, in der entsprechenden Veranstaltung, die erreichte Note angeben. Beachten Sie hierbei, dass verschiedene Universitäten unterschiedliche Veranstaltungen mit gleicher Veranstaltungsnummer haben können. Innerhalb einer Universität können die Veranstaltungsnummern eindeutig einer Veranstaltung zugeordnet werden. Entsprechendes gilt auch für die Matrikelnummer, welche nur innerhalb einer Universität eindeutig einem Studierenden zugeordnet werden können. Sei ferner $Note \in \{1, 2, 3, 4, 5\}$.

Es sollen Anfragen auf Grundlage von etwa $\frac{1}{20}$ der Tupel aus dem Datenstrom approximativ beantwortet werden. Folgende drei Anfragen werden gestellt:

- (a) Für jede Universität: Welche durchschnittliche Teilnehmerzahl hat die Universität pro Veranstaltung?
- (b) Wie groß ist der Anteil der Studierenden mit einem Notendurchschnitt von 2,0 oder besser?
- (c) Wie groß ist der Anteil der Veranstaltungen, in denen mindestens die Hälfte der Teilnehmer die Note 1 bekommen haben?

Beschreiben Sie jeweils, wie auf sinnvolle Weise etwa $\frac{1}{20}$ der Tupel aus dem Datenstrom ausgewählt werden können. Welche Teile der Tupel bilden dabei für welche der drei Anfragen die Schlüssel für die Auswahl?