Humboldt-Universität zu Berlin
Institut für Informatik
Lehrstuhl Logik in der Informatik
Prof. Dr. Nicole Schweikardt

28. Juni 2016

Big Data Analytics in Theorie und Praxis

Sommersemester 2016

Theorie-Übungsblatt 3

Zu bearbeiten bis 5. Juli 2016

Aufgabe 1: **(25 Punkte)**

Betrachten Sie nun die „Suche nach 2 fehlenden Zahlen“, bei der man als Eingabe einen Datenstrom von $n-2$ unterschiedlichen Zahlen aus der Menge $\{1, 2, \ldots, n\}$ erhält, bei dem diese $n-2$ Zahlen eine beliebige Reihenfolge haben können. Ziel ist, die beiden fehlenden Zahlen aus der Menge $\{1, 2, \ldots, n\}$ zu ermitteln.

- Beschreiben Sie einen Datenstromalgorithmus, der die beiden fehlenden Zahlen bestimmt und dabei möglichst wenig Speicher benutzt.
- Weisen Sie nach, dass Ihr Algorithmus die beiden fehlenden Zahlen korrekt bestimmt.
- Wie viele Speicherbits benötigt Ihr Algorithmus?
- Beweisen Sie, dass *jeder* Datenstromalgorithmus, der dieses Problem löst, mindestens $\lceil \log \binom{n}{2} \rceil$ Speicherbits benutzt.

Aufgabe 2: **(25 Punkte)**

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

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

Aufgabe 3: **(25 Punkte)**

Sie erhalten als Eingabe einen Datenstrom der Länge $n \in \mathbb{N}_{\geq 1}$ und der Form $a_1, a_2, \ldots, a_n$ mit $a_i \in \mathbb{N}_{\geq 1}$ für alle $i \in \{1, \ldots, n\}$. Sei $z = a_1 + a_2 + \cdots + a_n$ 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 $i \in \{1, \ldots, n\}$ 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.

Auf der Rückseite finden Sie eine weitere Aufgabe.

Aufgabe 4: **(25 Punkte)**

Sie erhalten einen Datenstrom mit Tupeln der Form

$$(\textit{Universität},\ \textit{Veranstaltungsnummer},\ \textit{Matrikelnummer},\ \textit{Note}),$$

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 $\textit{Note} \in \{1, 2, 3, 4, 5\}$.

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?