Diskrete Stochastik
Klasse 3Ia/5Ibb/5iCbb
HS 2019

Fachhochschule Nordwestschweiz
Hochschule für Technik

Bonusaufgabe
Bloom-Filter

Ziel der Aufgabe ist es, einen Bloom-Filter zu programmieren und die Funktionsweise darzustellen.

Folgende Schritte werden erwartet:

1. Sie verschaffen sich einen Überblick über die Funktionsweise, etwa unter `https://en.wikipedia.org/wiki/Bloom_filter`.
2. Sie erstellen ein Java-Programm, was folgendes leistet:
   (a) Bei einer gegebenen Anzahl n an zu erwartenden Elementen, die in der Datenstruktur gespeichert werden und einer Fehlerwahrscheinlichkeit p wird eine geeignete Filtergrösse m und die optimale Anzahl k an Hashfunktionen berechnet.
   (b) Die Datenstruktur wird implementiert, mit den Methoden zum Einfügen von Strings und dem Test, ob ein String enthalten ist. Als Hash-Funktionen soll `murmur3_128` (etwa in Guava enthalten: `https://github.com/google/guava/wiki/Release19`) verwendet werden mit jeweils einem anderen Seed.
   (c) Lassen Sie das Programm, welches eine Fehlerwahrscheinlichkeit p als Eingabe erhält, die Wörter aus `words.txt` einlesen und in einen Boom-Filter geeigneter Grösse einfügen. Überprüfen Sie die Fehlerwahrscheinlichkeit, in dem Sie für eine grosse Anzahl an nicht enthaltenden Strings testen, ob sie enthalten sind. Die so experimentell bestimmte Fehlerwahrscheinlichkeit soll zusammen mit den Parametern der Datenstruktur ausgegeben werden.
   (d) Bitte räumen Sie ihr Projekt so auf, dass ich keine speziellen Pakete nachladen oder komplizierte Pfadanpassungen machen muss.
3. Sie erstellen mit LaTeX (`https://www.latex-project.org/`) eine Zusammenfassung von ein bis zwei Seiten, welche folgendes enthält:
   (a) Idee des Bloom-Filters, mit Vor- und Nachteilen
   (b) ein konkretes Beispiel aus der Praxis, wo der Bloom-Filter verwendet wird mit kurzer Beschreibung dieses Programms
   (c) Eine Beschreibung, wie Sie die Fehlerwahrscheinlichkeit ihrer Datenstruktur getestet haben und welche Resultate dabei erzielt worden sind. Fügen Sie ein Screenshot der Ausgabe ein.
4. Senden Sie mir bitte per Mail das Java-Projekt, den LaTeX-Quellcode und das fertige pdf zu.

Allgemeine Hinweise:

1. Sie können in Gruppen bis zu drei Personen arbeiten.

2. Bei vollständiger Lösung wird auf die Note des kommenden Tests 0.3 drauf addiert. (Aus systemtechnischen Gründen liegt die Erfahrungsnote zwischen 1.0 und 6.0.) Teilpunkte werden vergeben, wenn nicht allemmier Anforderungen erfüllt sind.

3. Es ist nicht nötig, das Programm hinsichtlich Effizienz zu optimieren.

4. Das Programm sollte verständlich kommentiert sein.

5. Eigentlich gehe ich davon aus, dass Sie aus Fairnessgründen nicht versuchen, zu betrügen. Dennoch werde ich dies (auch mit Hilfe von Tools) kontrollieren. Falls dabei ein Täuschungsversuch festgestellt wird (also: (verschleierte) Kopien von Teilen existierender Programme (Internet oder Kollegen)), wird die Note des nächsten Tests auf 1.0 gesetzt.

Abgabe: 08.12.2019