Dr. Konrad Völkel
Heinrich-Heine-Universität Düsseldorf
Wintersemester 2024/25

Data Science 2

Blatt 6, Abgabe am 27.11.2024 um 12:00

Bitte geben Sie eine PDF-Datei ab.

Aufgabe 1. *Bestimmen Sie zwei verschiedene Eigenvektoren der Matrix*

$$A := \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \end{pmatrix}$$

Welche Eigenwerte hat die Matrix A? Gibt es noch weitere Eigenvektoren als die beiden, die Sie gefunden haben?

Aufgabe 2. *Sie haben bereits auf dem 2002er-Google-WebGraph `web-Google.txt.gz` eine explorative Analyse durchgeführt. Laden Sie den Graph in einen `networkx.DiGraph`. Erstellen Sie einen Teilgraphen, der alle Nachfolger (Successors) von Knoten 11342 und deren Nachfolger enthält und visualisieren Sie diesen mit `matplotlib` und `networkx.draw`. Berechnen Sie von dem Teilgraph mit `networkx.pagerank` den sogenannten PageRank, das ist für jeden Knoten ein float. Bestimmen Sie die zwei Knoten im Teilgraph mit dem größten PageRank-Wert. Notieren Sie die Namen/Indizes dieser Knoten.*

Bonusaufgabe: Verfahren Sie analog mit dem größeren `wiki-topcats.txt.gz`-Datensatz und dem Knoten 46.

Aufgabe 3. *Ein Bloom Filter kann genutzt werden, um schnell Elemente aus einer vorgegebenen Menge S aus einen Datenstrom zu filtern. Dabei könnte es zu falsch positiven oder falsch negativen Entscheidungen kommen. Erklären Sie (knapp!) wie sich diese beiden Fehler bei einem Bloom Filter verhalten und wie sie (mit welchem Preis) gesenkt werden können. Machen Sie dabei keine quantitative Analyse sondern nur eine qualitative.*

Aufgabe 4. *Was ist die wesentliche Eigenschaft der Hashfunktionen, die bei Flajolet-Martin zum Einsatz kommen? Anders gefragt: wieso ist ein Stream mit Ganzzahlen in einem bekannten Bereich nicht "gut genug" und auch dabei muss im Allgemeinen der ganze Stream gehasht werden, damit Flajolet-Martin funktioniert?*