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

Data Science 2

Blatt 7, Abgabe am 04.12.2024 um 12:00

Bitte geben Sie eine PDF-Datei ab und ein Notebook oder einzelne Python-Skripten.

Aufgabe 1. *Sie haben bereits auf dem 2002er-Google-WebGraph* `web-Google.txt.gz` *eine explorative Analyse durchgeführt und mit NetworkX den 'pagerank' ausrechnen lassen (für einen Teilgraph).*

Implementieren Sie Markovsches spektrales Ranking (so wie in der Vorlesung beschrieben). Vergleichen Sie auf Teilgraphen, so wie dem vom vorigen Übungsblatt, Ihr Ergebnis mit dem PageRank, den NetworkX berechnet. Achtung: die Zahlen werden nicht übereinstimmen.

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

Wenn Sie eine Beobachtung machen, schreiben Sie diese doch gleich mit auf!

Aufgabe 2. *Überlegen Sie eine Strategie, wie die vorige Aufgabe mit MapReduce umgesetzt werden könnte.*

Aufgabe 3. *Verwenden Sie die SSE API (*`https://html.spec.whatwg.org/multipage/server-sent-events.html`*) der Wikipedia (die wir auf dem vorigen Blatt betrachtet haben), also via* `https://wikitech.wikimedia.org/wiki/Event_Platform/EventStreams_HTTP_Service#Python` *um die Rate der eintreffenden Recent-Changes Events über 10 Minuten hinweg zu erfassen.*

Bonusaufgabe: filtern Sie die Events nach einem oder mehreren Keywords.

Tipp: Sie können die moderne Bibliothek `aiosseclient` *nutzen:* `https://github.com/ebraminio/aiosseclient`