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

Data Science 2
Blatt 4, Abgabe am 13.11.2024 um 12:00

Bitte geben Sie Ihren Code (als .py-Datei(en) und ggf. 1 Notebook) zusammen mit einem PDF ab.

Aufgabe 1. *Sie haben bereits zwei explorative Datenanalysen durchgeführt auf dem vorigen Übungsblatt. Nachdem Sie sich mit Kommiliton*innen ausgetauscht haben, was können Sie in Ihrer Analyse ergänzen? Wenn Ihre Analyse bereits sehr umfangreich war, müssen Sie hier nichts weiter tun.*

Aufgabe 2. *Auf dem vorigen Übungsblatt sollten Sie einen Plan aufstellen zur Implementierung von MinHash. Nachdem Sie sich mit Kommiliton*innen ausgetauscht haben, was können Sie an Ihrem Plan ändern? Können Sie ihn detaillierter ausarbeiten?*

Aufgabe 3. *Setzen Sie Ihren Plan um, zunächst um MinHash ohne MapReduce zu nutzen. Evaluieren Sie auf dem Housing-Datensatz. Sind die MinHash-ähnlichen Instanzen auch Jaccard-ähnlich? Was liefert die manuelle Inspektion der Strings? Sind sie ähnlich?*

Vergleichen Sie den Output mit Ihrer MRJob-Implementierung von MinHash.

Tipp: Ihre zuvor erzeugten Spielzeugdaten sind gut geeignet, um zu testen dass der Programmablauf fehlerfrei funktioniert. Da darin nur wenig ähnliche Instanzen enthalten sein werden, eignen sie sich nicht zur finalen Evaluation. Sie können aber mit dem Output Ihrer MinHash-Ähnlichkeitssuche neue Spielzeugdatensätze erzeugen, die sowohl sehr ähnliche als auch sehr unähnliche Instanzenpaare enthalten, und somit als guter Test für die nächste Aufgabe dienen können.

Aufgabe 4. *Implementieren Sie Locality Sensitive Hashing auf Grundlage Ihrer MinHash-Implementierungen (mit und ohne MRJob). Evaluieren Sie auf dem Enron-Email-Datensatz. Beschreiben Sie in Worten, was Sie dabei beobachten.*

Wenn Sie die Aufgaben alle bearbeiten, aber Ihre MRJob-Implementierung nicht korrekt funktioniert (die ohne MRJob aber schon), können Sie die korrekte MRJob-Implementierung immer noch in der folgenden Woche nachreichen und dieses Übungsblatt gilt dann als 'bestanden'.