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

Data Science 2

Blatt 2, Abgabe am 30.10.2024 um 12:00

Bitte geben Sie ein PDF ab.

Aufgabe 1. *Sei $M := \{a, b, c, \ldots, x, y, z\}$ die Menge der lateinischen Kleinbuchstaben. Wenn $S, T \subseteq M$ Teilmengen sind, so ist die Jaccard-Ähnlichkeit definiert als $Jac(S,T) := \frac{|S \cap T|}{|S \cup T|}$*

Geben Sie jeweils Mengen S und T an, sodass

1. $Jac(S,T) = 0$
2. $Jac(S,T) = 1$
3. $Jac(S,T) = \frac{1}{2}$

Beweisen Sie außerdem, dass für alle Mengen S und T gilt: $0 \leq Jac(S,T) \leq 1$.

Aufgabe 2. *Sei k eine positive natürliche Zahl und x ein String. Ein zusammenhängender Substring von x der Länge k heißt auch k-Shingle von x.*

Beispiel: x = "hallo", k=2, dann ist "ja" kein Substring ("j" kommt gar nicht vor), also kein 2-Shingle, "allo" zwar ein Substring, aber nicht Länge 2, also kein 2-Shingle, "ao" zwar ein Substring, aber nicht zusammenhängend, also kein 2-Shingle, "ha", "al", "ll", "lo" sind genau alle 2-Shingles.

Geben Sie selbst ein Beispiel eines Strings x der Länge 5 an und zählen Sie alle 3-Shingles auf.

Aufgabe 3. *Führen Sie zwei (kurze!) explorative Datenanalysen durch:*

1. *Das "Rent in Rome Kijiji" Dataset, Wohnungsannoncen die vor einigen Jahren von einer italienischen Plattform gescraped wurden. Ein kleiner Spieldatensatz.*
2. *Das "Enron Emails" Dataset, aus 2002, echte Emails der Enron Corporation.*

Beide Datensätze finden Sie in ILIAS.

Wir möchten diese Daten verwenden, um locality sensitive hashing (MinHash) daran zu üben. Wir sind also an den Texten (nicht so sehr den Metadaten) interessiert und möchten später Paare ähnlicher Texte identifizieren.