

Intelligent Data Management - Exercise 3

Christoph Prinz

November 7, 2017

Assignment 1

Jaccard Similarities formula = $|S \cap T| \div |S \cup T|$

$$a) SIM(\{1, 2, 3, 4\}, \{2, 3, 5, 7\}) = \frac{1}{3}$$

$$SIM(\{1, 2, 3, 4\}, \{2, 4, 6\}) = \frac{2}{5}$$

$$SIM(\{2, 3, 5, 7\}, \{2, 4, 6\}) = \frac{1}{6}$$

$$b) SIM(\{1, 1, 1, 2\}, \{1, 1, 2, 2, 3\}) = \frac{2}{3}$$

$$SIM(\{1, 1, 1, 2\}, \{1, 2, 3, 4\}) = \frac{1}{2}$$

$$SIM(\{1, 1, 2, 2, 3\}, \{1, 2, 3, 4\}) = \frac{3}{4}$$

Assignment 2

a) {The, most, effective}, {most, effective, way}, {effective, way, to}, {way, to, represent}, {to, represent, documents}, {represent, documents, as}, {documents, as, sets}, {as, sets, for}, {sets, for, the}, {for, the, purpose}

b) Stopwords are marked as *italic*: "The most effective way to represent documents as sets, for the purpose of identifying lexically similar documents is to construct from the document the set of short strings that appear within it."

Resulting Shingles:

{The, most, effective}, {way, to, represent}, {to, represent, documents}, {as, sets, for}, {for, the, purpose}, {the, purpose, of}, {of, identifying, lexically}, {is, to, construct}, {to, construct, from}, {the, document, the}, {the, set, of}, {set, of, short }, {of, short, strings}