

Exercise 1) Distance Functions:

Which of the following distance functions are metrics? Provide a proof for your claim. Note that distance = 1 – similarity

- a. Cosine similarity
- b. Jaccard similarity
- c. $\max(x, y)$ = the larger of x and y.
- d. $\text{sum}(x, y) = x + y$.

Exercise 2) Reservoir sampling

Prove that reservoir sampling guarantees that when there are n elements and reservoir size r each element is kept with the probability r/n.

Exercise 3) Document similarity

Take the following three text examples:

1. “Unlike classification or prediction, which analyzes data objects with class labels, clustering analyzes data objects without consulting a known class label. The class labels are not in the data because they are not known.”
2. “Classification can be used for prediction of class labels of data objects. However, in many applications, prediction of missing or not known data values rather than class labels is performed to fit data objects into a schema.”
3. “Sun Salutation, a ritual performed in the early morning, combines seven different postures. The sun, the life generator, is invoked by this Yogic exercise, an easy way to keep fit.”

Construct vectors based on the frequency of each word, ignoring the following “stop words given below. Use the cosine similarity function from the lecture to determine their mutual similarities:

Stopwords = { a, an, are, be, because, by, can, for, however, in, into, is, keep, many, not, of, or, rather, than, the, they, this, to, unlike, used, way, which, with, without }