

CSE 8331 - Fall 2006
Similarity Measures

Book (p 58)

$$\text{Dice: } \text{sim}(t_i, t_j) = \frac{2 \sum_{h=1}^K t_{ih} t_{jh}}{\sum_{h=1}^K t_{ih}^2 + \sum_{h=1}^K t_{jh}^2}$$

$$\text{Jaccard: } \text{sim}(t_i, t_j) = \frac{\sum_{h=1}^K t_{ih} t_{jh}}{\sum_{h=1}^K t_{ih}^2 + \sum_{h=1}^K t_{jh}^2 - \sum_{h=1}^K t_{ih} t_{jh}}$$

$$\text{Cosine: } \text{sim}(t_i, t_j) = \frac{\sum_{h=1}^K t_{ih} t_{jh}}{\sqrt{\sum_{h=1}^K t_{ih}^2 \sum_{h=1}^K t_{jh}^2}}$$

$$\text{Overlap: } \text{sim}(t_i, t_j) = \frac{\sum_{h=1}^K t_{ih} t_{jh}}{\min\left(\sum_{h=1}^K t_{ih}^2, \sum_{h=1}^K t_{jh}^2\right)}$$

Note: Here t_{ih}, t_{jh} are vectors of numeric values. They may be normalized to $[0, 1]$

$$t_i = \langle t_{i1}, \dots, t_{ik} \rangle$$

Alternative Definitions based on sets

$$\text{Dice: } \frac{2 |X \cap Y|}{|X| + |Y|}$$

$$\text{Jaccard's: } \frac{|X \cap Y|}{|X \cup Y|}$$

$$\text{Ex: } \begin{aligned} X &= \{1, 2, 3, 4\} \\ Y &= \{3, 4, 5, 6, 7\} \end{aligned}$$

$$\text{Dice} = \frac{2 \cdot 2}{4 + 5} = \frac{4}{9}$$

$$\text{Jaccard} = \frac{2}{7}$$

From the book's defn we get completely different answers. As a matter of fact we can't really even use these since we are not working with vectors of the same size.

To really compare, let's convert to bit maps:

$$\langle x_1, x_2, \dots, x_7 \rangle$$

where $x_i = 0$ or 1 depending on whether i is in the set.

$$\text{Thus } X = \langle 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0 \rangle$$

$$Y = \langle 0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \rangle$$

Now let's calculate using book definitions:

$$\text{Dice} = \frac{2 \cdot (0+0+1+1+0+0+0)}{4 + 5} = \frac{4}{9}$$

$$\text{Jaccard} = \frac{2}{4 + 5 - 2} = \frac{2}{7}$$

Thus when looking at simple sets, these are the same.

When using these you must be careful. What is the data you are looking at? Is it a set of data? Is it a vector of values??