Document Similarity Measure

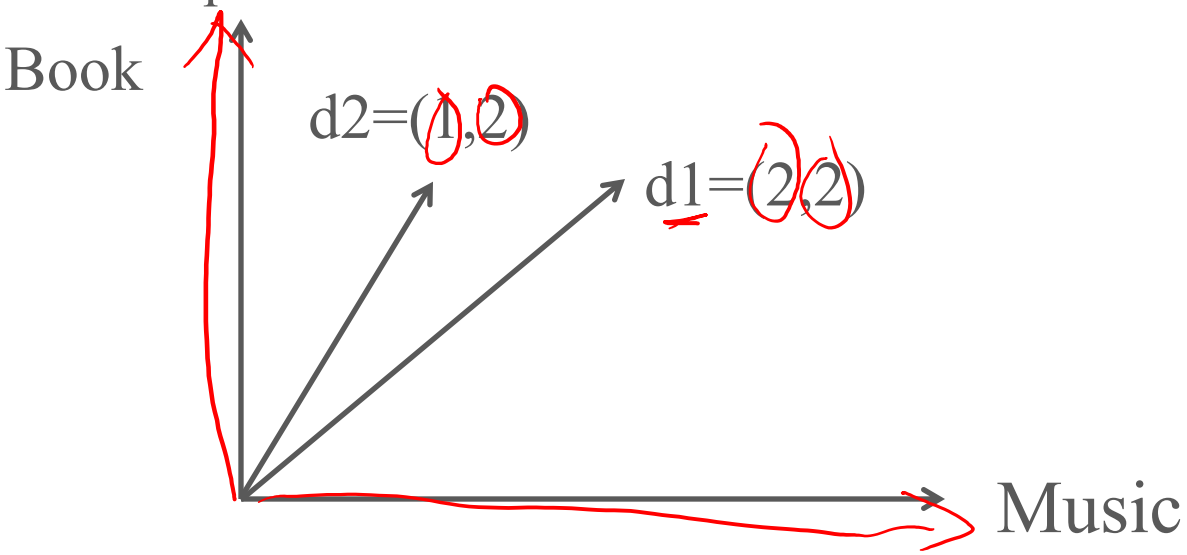
School of Information Studies
Syracuse University

Vector Representation

Bag-of-Words documents

- D1: "book, book, music, music"
- D2: "music, book, book"

Vectors in 2D-space



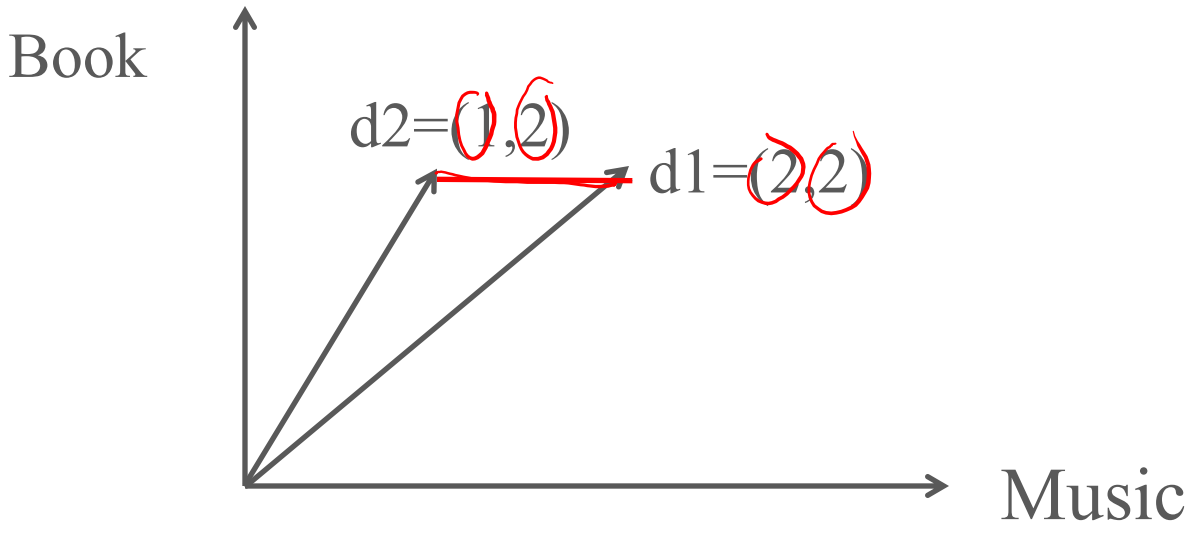
Distance/Similarity Between Two Documents

Distance/similarity measures

- Euclidean distance

$$d = X - Y = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

$$= \sqrt{(1-2)^2 + (2-2)^2} = 1$$



Distance/Similarity Between Two Documents

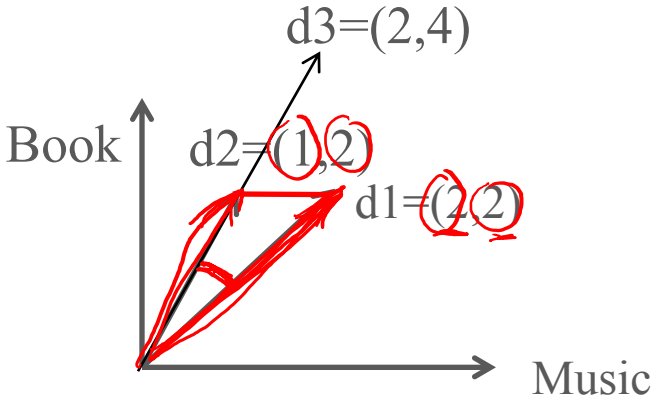
Distance/similarity measure

- Cosine similarity: the cosine value of the angle between the two vectors (0–90°)

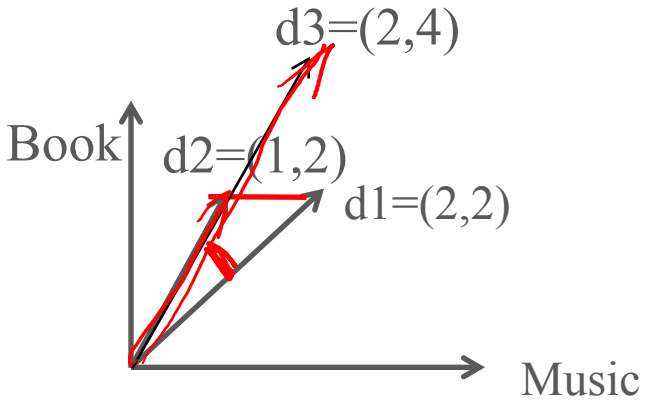
$$\cos(d_1, d_2) = \frac{x \cdot y}{|x||y|} = \frac{x_1 y_1 + x_2 y_2}{\sqrt{x_1^2 + x_2^2}\ \sqrt{y_1^2 + y_2^2}}$$

$$= \frac{1 \cdot 2 + 2 \cdot 2}{\sqrt{1^2 + 2^2}\ \sqrt{2^2 + 2^2}} = \frac{6}{\sqrt{5}\ \sqrt{8}} = 0.95$$

The greater the angle (distance), the smaller the cosine similarity



Does Vector Length Matter?



- (d1, d2) and (d1, d3) have the same angle
- The cosine similarity has normalized by vector norm
- Therefore, cos_sim(d1,d2) = cos_sim(d1,d3)