Data Mining and Optimization

Lecture 5: Text Similarity

Liu Yang

Nanjing University

Spring, 2025

文本语义相似度

- 如何衡量两个/多个文本在语义上的相似性？
- 文本聚类、文本分类、关键词提取等文本分析任务都需要衡量文本与文本之间的相似性。
- 文本语义相似度是建立在文本向量化基础上的，两个文本的语义相似度体现为这两个文本对应的文本向量的相似度。

内积与标准化内积

- 记文本d_i与d_j的向量化表示分别为：

$$d_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}, d_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{kj} \end{pmatrix}.$$

- d_i与d_j的内积(inner product)为

$$< d_i, d_j >= x_{1i}x_{1j} + x_{2i}x_{2j} + \cdots + x_{ki}x_{kj}.$$

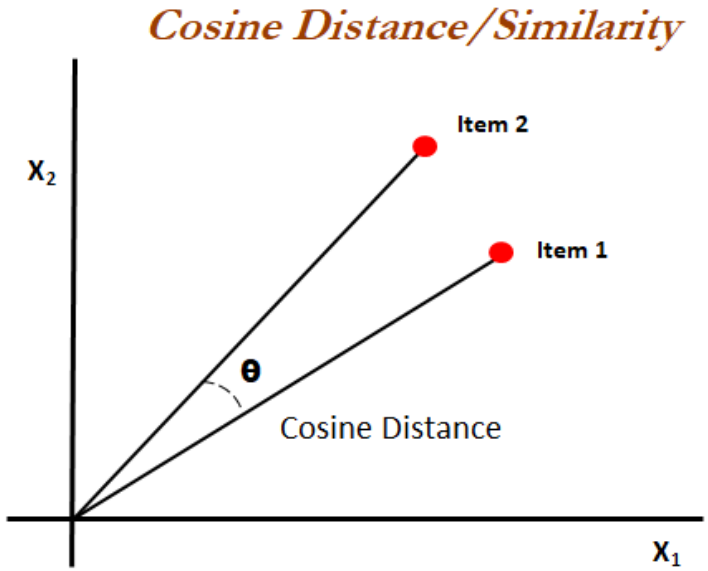
- d_i与d_j的标准化内积为

$$\frac{< d_i, d_j >}{\|d_i\|_2\|d_j\|_2},$$

其中，$\|d_i\|_2$为向量d_i的Euclidean norm。

余弦相似度

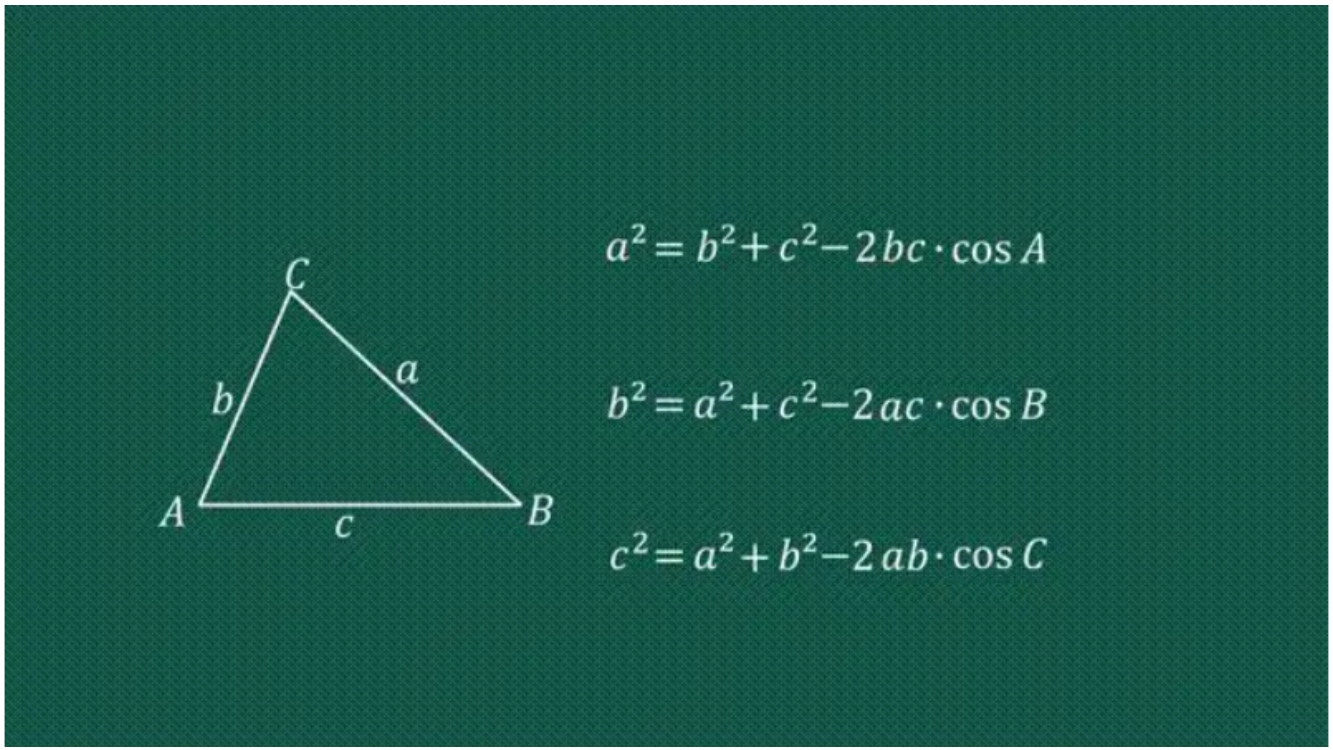
一个向量空间中两个向量夹角间的余弦值作为衡量两个个体之间差异的大小，余弦值接近1，夹角趋于0，表明两个向量越相似，余弦值接近于0，夹角趋于90度，表明两个向量越不相似。



余弦定理

余弦定理

三角形任一边的平方等于其他两边平方和减去这两边与它们夹角的余弦的积的两倍。



余弦相似度

- 如果$k=2$，$d_i=(x_{1i},x_{2i})'$且$d_j=(x_{1j},x_{2j})'$，d_i与d_j夹角的余弦值为

$$\begin{aligned}
\cos(d_i,d_j) &= \frac{\|d_i\|_2^2+\|d_j\|_2^2-\|d_i-d_j\|_2^2}{2\|d_i\|_2\|d_j\|_2} \\
&= \frac{x_{1i}^2+x_{2i}^2+x_{1j}^2+x_{2j}^2-((x_{1i}-x_{1j})^2+(x_{2i}-x_{2j})^2))}{2\|d_i\|_2\|d_j\|_2} \\
&= \frac{x_{1i}^2+x_{2i}^2+x_{1j}^2+x_{2j}^2-(x_{1i}^2+x_{1j}^2-2x_{1i}x_{1j}+x_{2i}^2+x_{2j}^2-2x_{2i}x_{2j})}{2\|d_i\|_2\|d_j\|_2} \\
&= \frac{2x_{1i}x_{1j}+2x_{2i}x_{2j}}{2\|d_i\|_2\|d_j\|_2} = \frac{2<d_i,d_j>}{2\|d_i\|_2\|d_j\|_2} = \frac{<d_i,d_j>}{\|d_i\|_2\|d_j\|_2}.
\end{aligned}$$

- 结论：余弦相似度=标准化内积。