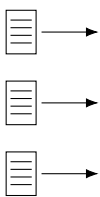
余弦定理和新闻的分类

从空间向量到文本相似性

阅读素材：吴军《数学之美》

新闻怎样变成一道数学题？

新闻 $\longrightarrow$ 一组数 $\longrightarrow$ 向量



$0, 0, 0, 3, 0, \cdots, 28, 0, 0, 0, 3$

$1, 0, 1, 0, \cdots, 8, 0, 10, 0, 2$

$0, 0, 32, 10, \cdots, 2, 0, 0, 3, 1$

怎样用计算判断两篇新闻的主题是否接近？

先只看三个词

固定顺序： **股票、银行、足球**

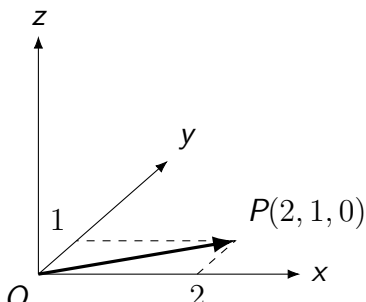
教学示例	股票	银行	足球
新闻 X	2	1	0
新闻 Y	4	2	0
新闻 Z	0	0	3

$$\boldsymbol{x} = (2, 1, 0), \qquad \boldsymbol{y} = (4, 2, 0), \qquad \boldsymbol{z} = (0, 0, 3).$$

每个坐标表示对应词的出现次数。

先猜一猜：哪两篇新闻更接近？

三个数，也可以表示一个空间向量



$$\overrightarrow{OP} = (2, 1, 0)$$

$$|\overrightarrow{OP}| = \sqrt{2^2 + 1^2 + 0^2} = \sqrt{5}$$

新闻中的三个数，

可以借助坐标来研究。

原文：给每个词一个固定位置

编号	词	权重
1	阿	0
2	啊	0.0034
3	阿斗	0
4	阿姨	0.00052
⋮	⋮	⋮
789	服装	0.034
⋮	⋮	⋮
64000	做作	0.075

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_{64000})$$

- 顺序相同，才能逐项比较。
- 没出现的词，对应权重为 0。
- 原文用 TF-IDF 衡量词的重要性。

本课先用词频理解，不展开权重算法。

从三维到更多维

三个词 $\longrightarrow$ (x_1, x_2, x_3)

六个词 $\longrightarrow$ $(x_1, x_2, \ldots, x_6)$

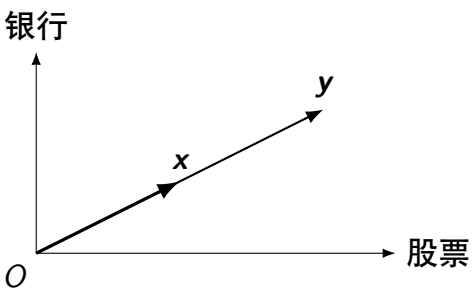
64000 个词 $\longrightarrow$ $(x_1, x_2, \ldots, x_{64000})$

维数 $=$ 选取的词的个数

超过三维时，先把它理解为**按固定顺序排列的一组数**。

为什么要比较方向？

$$\boldsymbol{x} = (2, 1, 0), \qquad \boldsymbol{y} = (4, 2, 0) = 2\boldsymbol{x}$$



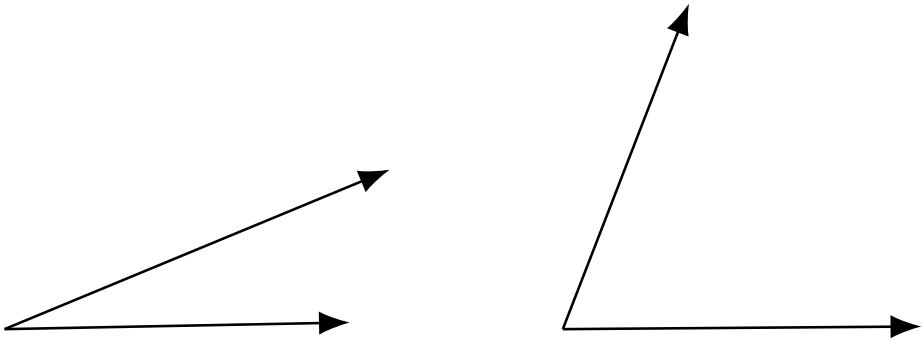
$$|\boldsymbol{y}| = 2|\boldsymbol{x}|$$

$$\angle(\boldsymbol{x}, \boldsymbol{y}) = 0^\circ$$

长度不同，词频比例相同。

整体倍增，不改变方向。

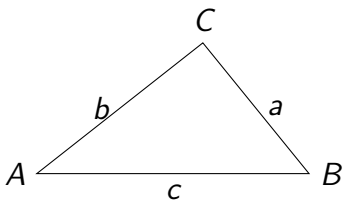
用夹角描述方向的接近程度



夹角小 $\longleftrightarrow$ 方向接近

怎样从坐标求出夹角？

回到熟悉的余弦定理

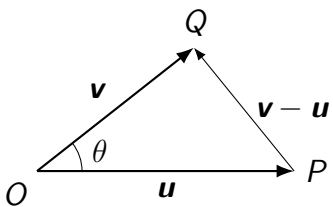


$$a^2 = b^2 + c^2 - 2bc\cos A$$

$$\cos A = \frac{b^2 + c^2 - a^2}{2bc}$$

两个向量移到同一起点，它们与端点连线组成三角形。

把三条边换成向量的长度



$$OP = |\boldsymbol{u}|, \quad OQ = |\boldsymbol{v}|$$

$$PQ = |\boldsymbol{u} - \boldsymbol{v}|$$

$$\cos\theta = \frac{|\boldsymbol{u}|^2 + |\boldsymbol{v}|^2 - |\boldsymbol{u} - \boldsymbol{v}|^2}{2|\boldsymbol{u}||\boldsymbol{v}|}$$

先考虑非零且不共线的向量；同向、反向时也满足上述等式。

用三个坐标展开

$$\boldsymbol{u} = (u_1, u_2, u_3), \qquad \boldsymbol{v} = (v_1, v_2, v_3)$$

$$\begin{aligned} |\boldsymbol{u}|^2 &= u_1^2 + u_2^2 + u_3^2, \\ |\boldsymbol{v}|^2 &= v_1^2 + v_2^2 + v_3^2, \\ |\boldsymbol{u} - \boldsymbol{v}|^2 &= (u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2. \end{aligned}$$

每一项都有相同的消去：

$$u_i^2 + v_i^2 - (u_i - v_i)^2 = 2u_i v_i$$

得到三维向量的夹角公式

$$|\boldsymbol{u}|^2 + |\boldsymbol{v}|^2 - |\boldsymbol{u} - \boldsymbol{v}|^2 = 2(u_1v_1 + u_2v_2 + u_3v_3)$$

$$\cos\theta = \frac{u_1v_1 + u_2v_2 + u_3v_3}{\sqrt{u_1^2 + u_2^2 + u_3^2}\sqrt{v_1^2 + v_2^2 + v_3^2}}$$

分子叫作两个向量的**数量积**（原文称“内积”）：

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_1v_1 + u_2v_2 + u_3v_3 = |\boldsymbol{u}||\boldsymbol{v}|\cos\theta$$

原文的公式：把三项写成更多项

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_{64000}), \qquad \boldsymbol{y} = (y_1, y_2, \ldots, y_{64000})$$

$$\mathrm{sim}(\boldsymbol{x}, \boldsymbol{y}) = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_{64000} y_{64000}}{\sqrt{x_1^2 + x_2^2 + \cdots + x_{64000}^2}\,\sqrt{y_1^2 + y_2^2 + \cdots + y_{64000}^2}}$$

对应项相乘再相加　÷　两个长度的乘积

这就是余弦相似度。

这里采用三维公式的推广；两个向量都须非零。

算一算：验证最初的判断

$$\boldsymbol{x} = (2, 1, 0), \qquad \boldsymbol{y} = (4, 2, 0), \qquad \boldsymbol{z} = (0, 0, 3)$$

$$\mathrm{sim}(\boldsymbol{x}, \boldsymbol{y}) = \frac{2 \times 4 + 1 \times 2 + 0 \times 0}{\sqrt{5}\sqrt{20}} = 1$$

$$\mathrm{sim}(\boldsymbol{x}, \boldsymbol{z}) = \frac{2 \times 0 + 1 \times 0 + 0 \times 3}{\sqrt{5} \times 3} = 0$$

在所选的三个词上，X 与 Y 的词频比例相同。

余弦值怎样读？

词频或非负权重组成的非零向量：

$$0 \leqslant \mathrm{sim}(\boldsymbol{x}, \boldsymbol{y}) \leqslant 1$$

余弦值	方向	特征的含义
1	同向	各词权重成相同的正比例
接近 1	夹角小	用词的权重分布接近
0	垂直	没有共同的非零权重项

特征只保留了部分信息：同向不能推出两篇新闻内容完全相同。

给一篇新新闻分类

仍按“股票、银行、足球”的顺序，设两个类别的代表向量为

$$\boldsymbol{f} = (2, 1, 0) \text{（财经）}, \qquad \boldsymbol{s} = (0, 0, 1) \text{（体育）}.$$

新新闻：$\boldsymbol{q} = (1, 1, 1)$。

分别算出 $\mathrm{sim}(\boldsymbol{q}, \boldsymbol{f})$ 和 $\mathrm{sim}(\boldsymbol{q}, \boldsymbol{s})$。

按相似度最大的规则，它应归入哪一类？

比较两个余弦值

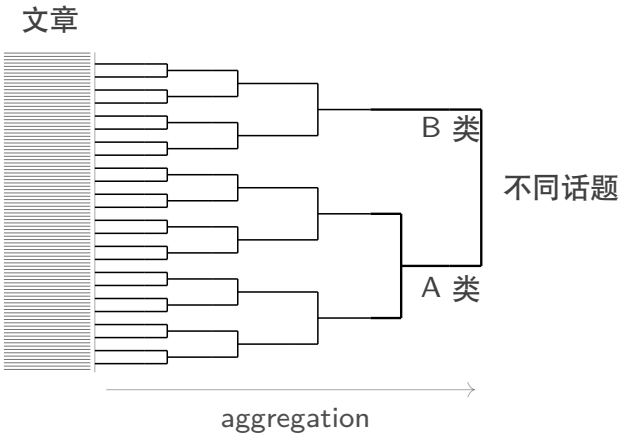
$$\mathrm{sim}(\boldsymbol{q}, \boldsymbol{f}) = \frac{1 \times 2 + 1 \times 1 + 1 \times 0}{\sqrt{3}\sqrt{5}} = \frac{3}{\sqrt{15}} \approx 0.775$$

$$\mathrm{sim}(\boldsymbol{q}, \boldsymbol{s}) = \frac{1 \times 0 + 1 \times 0 + 1 \times 1}{\sqrt{3} \times 1} = \frac{1}{\sqrt{3}} \approx 0.577$$

按这组特征和分类规则，归入**财经类**。

比较余弦即可，无须算出夹角的具体度数。

没有预先给定类别时呢？



原材料中的聚合示意图

相近的新闻先组成小类。

再比较小类，逐步合并。

差异过大时停止合并。

离开课堂前，想一想

1. 把一篇新闻的每个词频都乘 3，
   余弦相似度会改变吗？

2. 所选词一个也没出现，得到零向量，
   还能用这个公式吗？

3. 两篇新闻词频相同，
   能保证表达的意思相同吗？

三个问题的回答

1. 不变：分子、分母同时乘 3。

2. 不能：零向量使分母为 0。
   需要补充特征，或暂不分类。

3. 不能：词频没有记录词的顺序。
   例如“甲战胜乙”与“乙战胜甲”。

把今天的联系串起来

新闻 $\longrightarrow$ 特征向量 $\longrightarrow$ 余弦相似度

余弦定理 $\longrightarrow$ 坐标展开 $\longrightarrow$ 夹角公式

$$\cos\theta = \frac{\boldsymbol{u}\cdot\boldsymbol{v}}{|\boldsymbol{u}||\boldsymbol{v}|}$$

用方向的接近程度，辅助判断主题的接近程度。