共现矩阵（co-occurrence matrix）

Stephen CUI

2023 年 8 月 26 日

考虑如何基于分布式假设使用向量表示单词，最直截了当的实现方法是对周围单词的数量进行计数。具体来说，在关注某个单词的情况下，对它的周围出现了多少次什么单词进行计数，然后再汇总。这里，我们将这种做法称为“基于计数的方法”，在有的文献中也称为“基于统计的方法”。

考虑下面的句子：You say goodbye and I say hello. 我们将窗口大小设为 1.

表 1: 单词 you 的上下文中包含的单词的频数

	you	say	goodbye	and	i	hello	.
you	0	1	0	0	0	0	0

对其他单词做同样的处理，可以得到：

表 2: 汇总各个单词的上下文中包含的单词的频数

	you	say	goodbye	and	i	hello	.
you	0	1	0	0	0	0	0
say	1	0	1	0	1	1	0
goodbye	0	1	0	1	0	0	0
and	0	0	1	0	1	0	0
i	0	1	0	1	0	0	0
hello	0	1	0	0	0	0	1
.	0	0	0	0	0	1	0

这个表格的各行对应相应单词的向量。因为表格呈矩阵状，所以称为共现矩阵（co-occurrence matrix）。