同义

词之间具有以下五种关系：同义、近义、反义、关联、内涵

注：·不存在绝对同义的词，即使在某方面含义一致，也会在俚语、语气、风格等方面有一定差异

·词的关联性指那些可以通过语义框架或语义场联系的词，如"汽车"与"驾驶"，"汽车"与"充油"等

向量语义学

用向量表示词的含义，把词放进向量空间，用向量距离表示语义相似度

依据与假说：词的含义=它在语言中的用法
上下文一样的词，认为是同义词

向量语义：用词周围出现的词（上下文）来表示它
把词义映射到一个向量空间，每个词=空间内的一个点/向量，相似的词在空间上挨得近
这种将词义按某种规则映射为向量的操作称作"词嵌入"，而这个向量便被称作"词向量"

优点：1、把语义转化为可计算的数值
2、可泛化相似词义，不局限于完全相同的词

两种嵌入方法：Tf-idf：稀疏向量，基于计数的基线模型
词被表示为一个邻接词对数量的函数

Word2vec：密集向量，基于神经网络预测
通过训练一个分类器区分近邻和非近邻词汇得到词的表示

词的向量表示法

词-文档共生矩阵

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

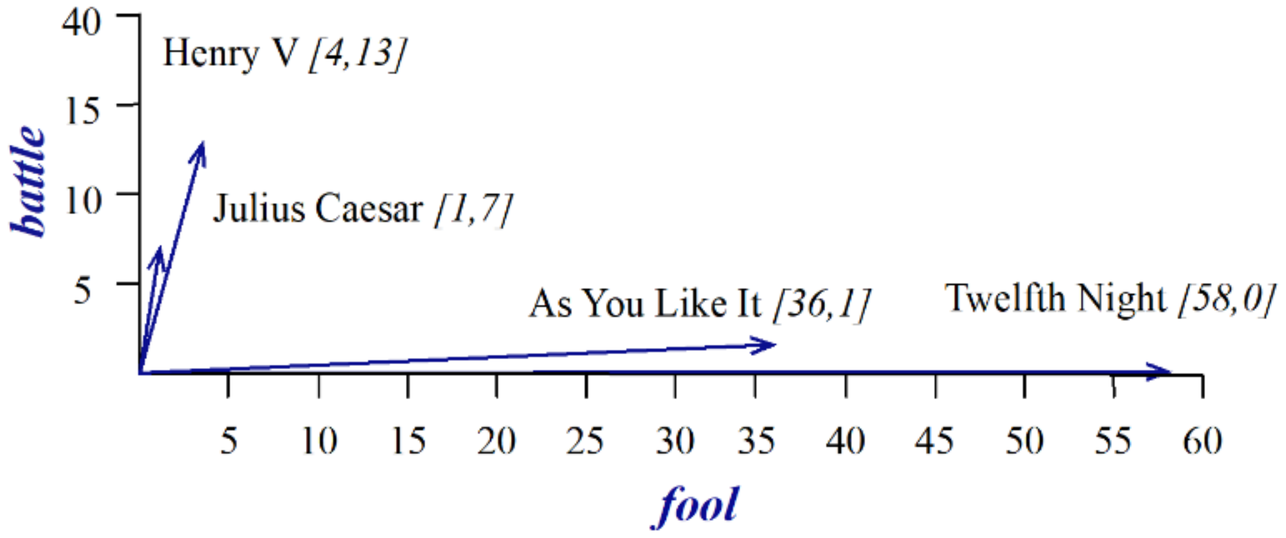
行：特征词
列：文档
值：该词在该文档出现次数

由这个矩阵可知，两个喜剧（即前两列）的向量是相似的，而与两个历史剧（后两列）不同。喜剧有更多的 fool 与 wit，而 battle 几乎不出现

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

battle 更可能出现在历史剧中，而 fool 更可能出现在喜剧中

可根据以上信息刻画出几部文档的相似程度：

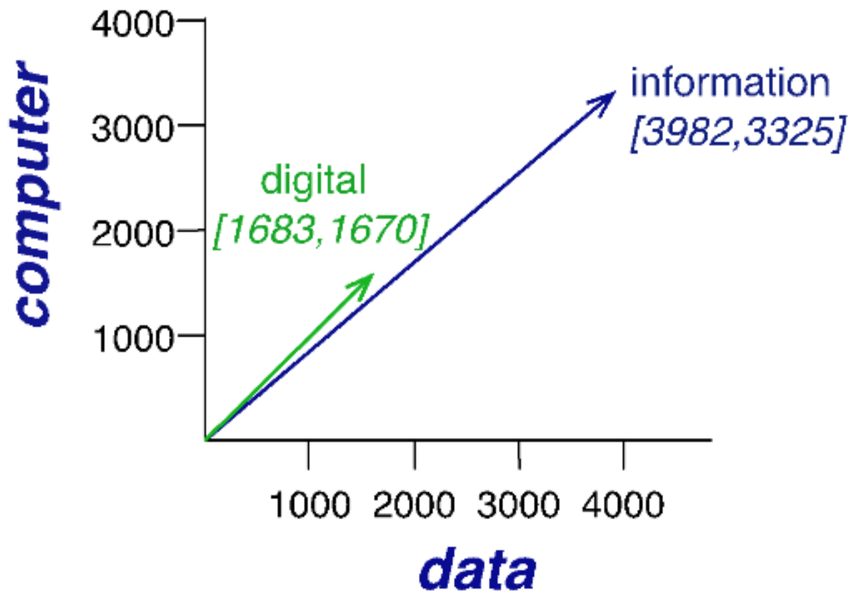


词-词共生矩阵

假设前提：两个词相似，则具有相似的上下文

is traditionally followed by	**cherry**	pie, a traditional dessert	
often mixed, such as	**strawberry**	rhubarb pie. Apple pie	
computer peripherals and personal	**digital**	assistants. These devices usually	
a computer. This includes	**information**	available on the internet	

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...



这便刻画出两个词的语义关系

向量相似度量：$dot\text{-}product(\vec{v},\vec{w}) = \vec{v}\cdot\vec{w} = \sum_{i=1}^{N} v_i w_i$

向量长度：$|\vec{v}| = \sqrt{\sum_{i=1}^{N} v_i^2}$

余弦度量：$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\,\vec{w}}{|\vec{v}||\vec{w}|}$

注：$\vec{v},\vec{w}$：两个词的向量表示

$\cos(\vec{v},\vec{w})$：$\vec{v}$与$\vec{w}$余弦相似度

v_i：词v在上下文i上的数量

$$\cos(\vec{v},\vec{w}) = \begin{cases} -1 & \text{向量方向相反，无实际含义} \\ 0 & \text{正交，语义无关} \\ 1 & \text{平行，同义} \end{cases}$$

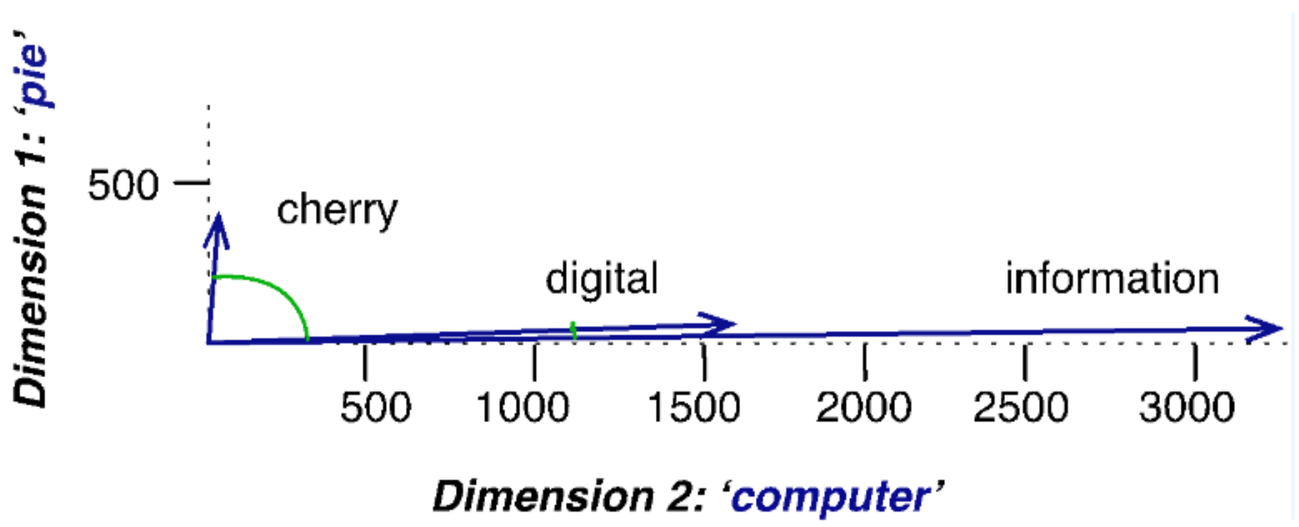
注：$\cos=-1$并不指代v与w互为反义，因为反义词实际上联系紧密，$\cos$为负并无实际含义，应用中也几乎不出现$\cos$为负的两个向量

例：

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{cherry},\text{information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = .017$$

$$\cos(\text{digital},\text{information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = .996$$



TF-IDF词向量

频率表示方法的缺点：频率对于诸如"it""the"这样的词而言属于无效信息，它们的上下文对它们本身几乎无意义

TF-IDF核心思想：压低常见词，抬高稀有词

一个词在当前文档出现越多，说明其区分度低，但若一个词在整个语料库中越常见，说明其区分度越高

公式：1、TF（Term Frequency 词频）

$tf_{t,d} = count(t,d)$ 词t在文档d中出现次数

一般会做对数化处理：

$tf_{t,d} = log_{10}(count(t,d)+1)$ OR

$$tf_{t,d} = \begin{cases} 1 + \log_{10}(count(t,d)) & count(t,d)>0 \\ 0 & else \end{cases}$$

2、DF（Document Frequency 文档频率）

$df(t)$ = 包含t的文档数量

3、IDF（Inverse Document Frequency 逆文档频率）

$idf_t = \log_{10}\left(\frac{N}{df_t}\right)$　　N是文档总数

用来表示t不稀有（idf越大t越稀有）

4、TF-IDF：

$w_{t,d} = tf_{t,d} \times idf_t$

表示文档d中词t的TF-IDF值，越大说明t对当前文档的区分度越强

如："this"它的tf很高，但 $idf_t \approx 0$，

它的 $w_{t,d} \approx 0$，说明其区分度低

例、

- 频次：　$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- tf-idf:

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

向量特点：·属于稀疏向量（大部分是0）

·维度很高（=词表大小）

·是词向量的基线模型

注意：TF-IDF衡量的是某个词的区分度而非其重要性，假设给出的文档围绕"AI"进行讨论，那"AI"的tf-idf值为0，这仅仅说明"AI"一词没有区分文档的能力，不

能说明"AI"不重要、不核心.

TF-IDF会低估这种高频常用词，这是其局限，也是其特点，比如：

is traditionally followed by **cherry** pie, a traditional dessert
often mixed, such as **strawberry** rhubarb pie. Apple pie
computer peripherals and personal **digital** assistants. These devices usually
a computer. This includes **information** available on the internet

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

- tf-idf?

若按传统的统计频次法，"information"一词在"computer"
"data"等词出现次数特别多，这能说明它相比"digital"
与计算机科技的相关性更强吗？

而若按TF-IDF，"Information"会被压得非常低（IDF很小）
但"digital"不会。这导致前者的词向量被拉远，而
"digital"则更靠近"computer"与"data"
这体现了"digital"与"computer"、"data"的强共现性.
"digital"与"Information"被区分开 （专业 vs 通用）
"cherry"与"strawberry"被聚在一起 （甜品领域）
"digital"与"cherry"被彻底分开 （领域不同）

TF-IDF会放大这些"特定领域的共现"，使词向量
更能体现语义领域，即"区分度"

Word2vec词向量.

上面提到，TF-IDF是稀疏向量，会很费大量资源，于是便有了
密集向量（维度少，大部分元素非0），密集向量更适合输入机器，
计算快，它含更少参数有利于避免过拟合.