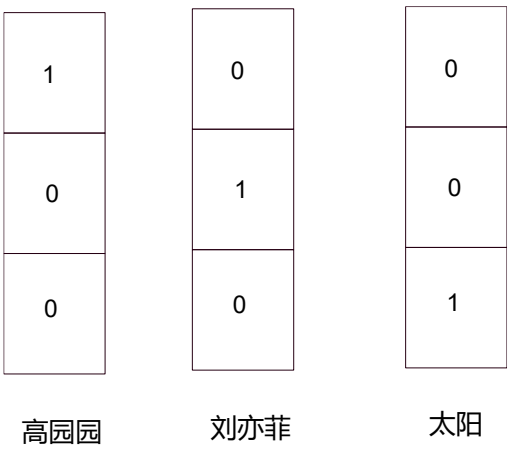
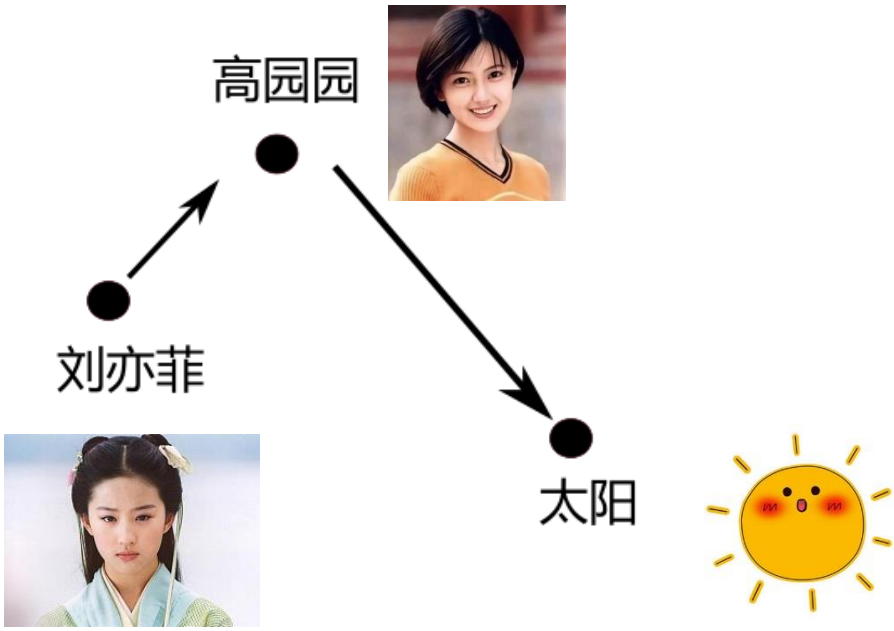
什么是词向量

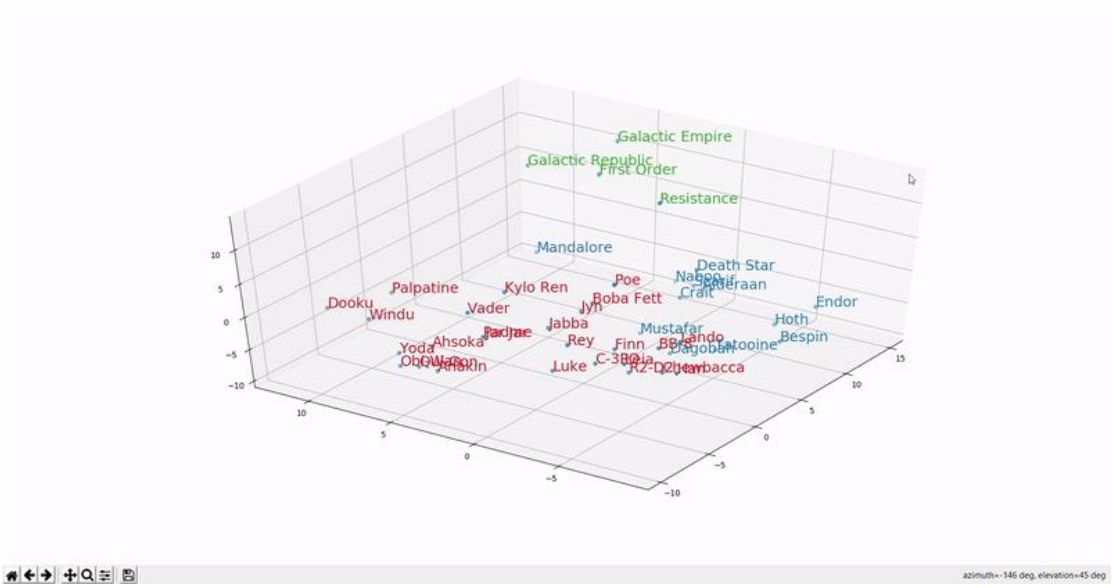
如何表示一个单词的意义？对人来说，一般用解释法，用一段话来解释词的含义。如“太阳”在新华字典中的释义是“太阳系的中心天体。银河系的一颗普通恒星。”然而，这样的解释计算机是听不懂的，必须用更简洁的方式来对词义进行表示。

传统上，计算机用一种称为“独热向量“的方式来表示单词。假设词表里一共有 100 个词，则用 100 维的向量来表示这些单词。对每个单词，只有一个维度值为 1，其余维度都为 0，因此称为独热向量。这种表示方法把每个单词当成孤立的个体，词与词之间没有”距离“的概念，因此只能认为是一个记号，而不是语义表示。 如图所示，和”太阳“相比，”刘亦菲“显然和”高园园“更接近一些，但在独热表示中是没有区别的。





为了解决这个问题，科学家们提出用词向量来表示单词的语义。和独热表示不同，词向量是个连续向量，且两个词向量的距离与对应单词之间的语义相似程度相关：越相似的单词，词向量间的距离越近，越不相关的单词，词向量之间的距离越远。如图所示，相比太阳，高园园和刘亦菲更相似，因此我们调整这些词的词向量，使得刘亦菲的词向量接近高园园，而太阳的词向量远离高园园。通过反复调整，得到词向量就可以反应词与词之间的语义关系。



现在就差最后一个问题：如何定义两个单词语义上是否接近呢？通常采用上下文相关法，如果两个单词同时出现在一个语言环境中（如同一个上下文窗口内），则认为二者的语义相关。当然也有其它定义相关性的方法，如在知识图谱中有边相连，或发音相似，这时生成的词向量可能代表各种信息，而非单纯的语义。因此，词向量的提出不仅极大提高了语义表达能力，更重要的是拓展了人们的思路，为离散对象的神经学习打开了大门。

Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013 。
https://arxiv.org/pdf/1301.3781.pdf%C3%AC%E2%80%94%20%C3%AC%E2%80%9E%C5%93