# 向量空间中词表征的有效估计

# 摘要

我们提出了两种用于计算来自非常大的数据集的单词的连续向量表示的新型模型架构。 这些表示的质量在单词相似性任务中测量，并且将结果与基于不同类型的神经网络的先前最佳表现的技术进行比较。 我们以低得多的计算成本观察到准确性的大幅度改进，即从16亿字数据集中学习高质量字矢量所需的时间不到一天。 此外，我们显示这些向量在我们的测试集中提供了最先进的表现，用于测量句法和语义词相似性。

# 简介

许多当前的NLP系统和技术将单词视为原子单元 - 在词之间没有相似性的概念，因为它们被表示为词汇表中的索引。 这个选择有几个好的原因 - 简单性，鲁棒性，以及对大量数据训练的简单模型优于复杂系统训练数据较少的观察。 一个例子是用于统计语言建模的流行的N-gram模型 - 今天，可以在几乎所有可用数据上训练N-gram（万亿字[3]）

## 本文的目的

本文的主要目的是引入可用于从具有数十亿字的巨大数据集学习高质量字矢量的技术，并在词汇中使用数百万个单词。 据我们所知，以前提出的架构中没有一个已经被成功地训练了超过几百万的单词，而在50-100之间的字向量的适度维度。

我们使用最近提出的技术来测量所得到的矢量表示的质量，期望不仅相似的词往往彼此接近，而且这个词可以具有多个相似度[20]。 这在早期的变形语言中已被观察到 - 例如，名词可以有多个单词结尾，如果我们在原始向量空间的子空间中搜索类似的单词，则可以找到具有相似结尾的单词[13 ，14]。

# 模型结构

# 新的对数线性模型

# 结论