Word Embeddings是一种对文本进行学习表示的一种技术，其中单个单词在预定义的向量空间中表示为一个向量。每个单词都被映射到一个向量上，每个单词与向量空间中的一个点相关联，特征向量表示单词的不同方面。方法的关键是为每个单词使用密集的分布式表示，分布式表示是基于单词的使用来学习的。这使得以相似方式使用的单词具有相似的表示。

Word Embedding中，每个单词都由实值向量表示，通常是几十维或几百维。这与稀疏单词表示所需的成千上万个维度形成了对比，避免了向量空间模型中的矩阵稀疏问题。

Word2Vec是一种统计方法，可以有效地通过语料库训练词向量。它是由Google的Tomas Mikolov等人于2013年为了提高基于神经网络的训练效率而开发的。该方法基于一个前馈的、完全连接的架构，但是仅包含输入层、隐藏层和输出层三层结构，并且对隐藏层和输出层进行了优化，使模型的训练效率更高。

word2vec训练模型包括CBOW模型和skip-gram模型。CBOW模型通过根据上下文预测当前单词来学习Word Embedding。Skip-gram模型通过预测给定当前单词的周围单词来学习。skip-gram和CBOW模型分别如图8和图9所示。



图8 skip-gram模型示意图

