# 2.1.3 文本嵌入模型

想要造航母，仅仅一个大语言模型是不够的。

假设我们想要创造一个《红楼梦》聊天机器人应用，并且询问关于这本书的知识。一个大语言模型就够了吗？

还有我们上传专业的论文，长达几十页的临床医学实验报告，或者使用搜索或外部数据的东西，跟真实的数据连接起来，大语言模型能处理吗？

虽然最新的模型一直在突破极限，但是仍然不能满足我们想要大模型阅读大型文档、书籍的需求。

面对这些挑战，我们可能需要总结或寻找其他方法来应对。因为大型语言模型可以传递的标记数量是有限的，大多数模型最多可以传递1024到16K个标记，尽管一些新的模型可以处理更多的标记。这就意味着，我们需要思考如何有效地利用这些有限的标记，以达到我们的应用需求。

此外，理解模型的类型和其各自的特点也同样重要。模型在高层次上可以分为两种不同类型：语言模型和文本嵌入模型。嵌入模型将文本转换为向量标记，让我们可以将文本标记为向量特征，显示在向量空间中。了解这两种模型的特性和差异，能帮助我们更好地选择和使用模型，进一步优化我们的应用。

#### 嵌入模型将文本转换为向量标记的例子

在自然语言处理（NLP）中，嵌入模型是将文本转换为向量的重要工具，这样我们可以在向量空间中表示文本，以便进行后续的机器学习或深度学习任务。举个例子，我们可以用 Word2Vec，一个常见的嵌入模型，来演示这个过程。

假设我们有以下的语料库：

1. I love learning.  
2. I like reading books.  
3. Books are great for learning.

首先，我们需要建立一个词汇库，也就是列出所有在语料库中出现过的独立单词：

"I", "love", "learning", "like", "reading", "books", "are", "great", "for"

接下来，我们用 Word2Vec 来训练这个语料库。Word2Vec 是一个预测模型：它试图从上下文预测目标单词，或者从目标单词预测上下文。训练过程中，Word2Vec 会学习到一个单词和它的上下文单词之间的关系，并将这些信息编码到向量中。

训练完成后，每个单词都会被赋予一个向量。例如，“books”的向量可能是：

[0.1, -0.2, 0.3, 0.5, -0.1]

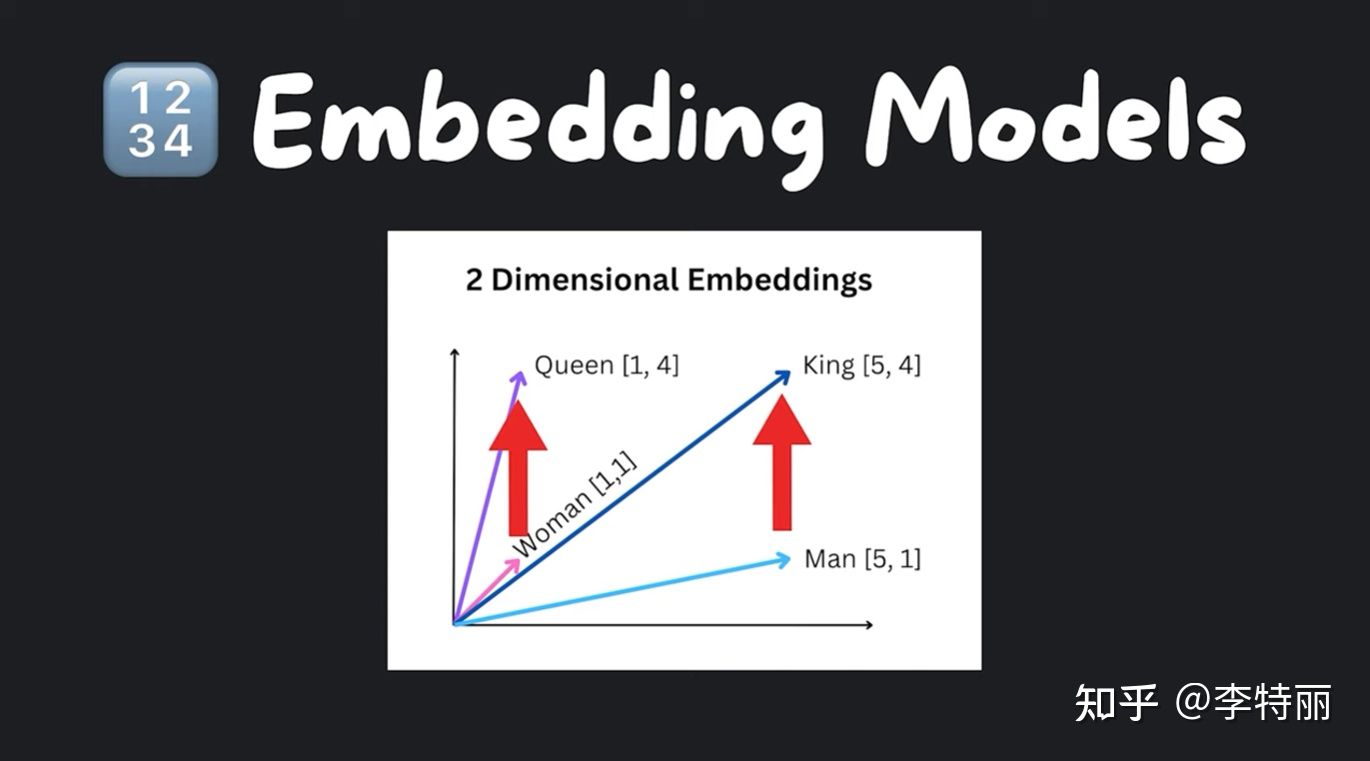
而“learning”的向量可能是：

[-0.1, 0.3, -0.2, 0.1, 0.4]

这些向量现在就代表了对应的单词在向量空间中的位置，且具有一些有趣的属性。比如，语义上相似的单词在向量空间中的位置会比较接近，我们可以通过计算两个向量之间的余弦相似度来衡量这种接近程度。

以上就是嵌入模型将文本转换为向量标记的一个基本例子。

#### 嵌入模型的原理



在上面这个图像中，我们可以看到在一个二维空间中，“man”是“king”，“woman”是“queen”，它们代表不同的事物，但我们可以看到一种模式。这个模式就是可以在向量空间中寻找最相似的文本片段，实现语义搜索。

例如，OpenAI 的文本嵌入模型可以精确地嵌入大段文本，具体而言，8100 个标记，根据它们的词对标记比例 0.75，大约可以处理 6143 个单词。它输出 1536 维的向量。

我们可以使用 LangChain 与多个嵌入提供者进行接口交互，例如 OpenAI 和 Cohere 的 API，但我们也可以通过使用 Hugging Faces 的开源嵌入在本地运行，以达到 免费和数据隐私 的目的。

现在，您可以使用仅 4 行代码在自己的计算机上创建自己的嵌入。但是，维度数量可能会有所不同，嵌入的质量可能会较低，这可能会导致检索不太准确。

#### 快速入门

#### 安装和设置密钥

我们需要安装 Langchain Python 包：

pip install openai langchain

#### 使用 Text Embedding Models 的最简单方法

导入 OpenAIEmbeddings：

from langchain.embeddings import OpenAIEmbeddings  
  
embeddings = OpenAIEmbeddings(openai\_api\_key = "YOUR\_OPENAI\_API\_TOKEN")  
result = embeddings.embed\_query("This is an apple.")

打印 result ，我们就可以看到 “This is an apple.” 这句话的向量空间表达数组。

[0.012008527293801308,  
 -0.001835253438912332,  
 0.00145026296377182,  
 -0.0030227703973650932,  
 -0.005661344155669212,  
 0.005086636636406183,...]

打印出的数字并不能直观看到这些词在向量空间中的位置，但如果你对向量可视化感兴趣，可以访问OpenAI的官方网站，那里有许多关于词嵌入（embedding）可视化的实践场所（playground）供你探索和学习。