词袋模型

1. 什么是词袋模型

词袋模型（Bag-of-words）就是描述文档中单词出现的文本的一种表示形式。它涉及到两个方面：

* 已知词汇的集合
* 测试已知单词的存在

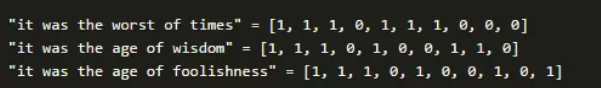
该模型只关注文档中是否出现已知的单词，并不关心词出现的顺序，也不关注不在词汇集合中的单词。

词袋模型的例子：

* 步骤1：收集数据。以下是一个例子
  + It was the best of times,
  + it was the worst of times,
  + it was the age of wisdom,
  + it was the age of foolishness,”

对于这个示例，我们将每一行视为一个单独的文档，将4行视为整个文档。

* 步骤2：设计词汇，下面我们可以列出我们的模型词汇表中的所有单词：it, was, the, best, of, times, worst, age, wisdom, foolish ness。这是一个由包括24个词组成的语料库，其中共包含10个词汇。
* 步骤3：创建文档向量：下一步是在每个文档中记录单词，目的是将自由文本的每个文档转换为一个文本向量，这样我们就可以将其用作机器学习模型的输入或输出。因为我们知道词汇有10个，所以我们可以使用固定长度为10的文档，来表示向量中的每一个单词的位置。单词出现则值为1，不出现为0.文档向量如下：



缺陷：会产生很多稀疏向量。