语料库语料库是由很多文档组成的。对语料库进行分析，有利于更好地了解词汇、句子及文章的内容含义。对于给定的文本内容，虽然人可以对其内涵进行有效的判断，但是计算机只是将其作为般的符号来处理。计算机本身不含有对文本进行理解的先验知识，因此只能通过“阅读”大量的文档来学习和文本有关的知识。

语料库本质上就是计算机学习的文本资料，让计算机通过“阅读”大量语料库里的文章来学习如何去从语义层面和语法层面对特定层次的文本单元进行理解。广义的语料库除了包括文档的集合，还包含任何计算机都可以进行统计学习的知识集合。

语料库通常分为通用语料库和专用语料库。通用语料库与文本分析的具体应用场景无关， 通常仅按照语言进行划分，如英语的通用语料库、汉语的通用语料库等。通用语料库通常大而全，只要是某种语言的文本，无论什么内容都会有所涉及。通过对通用语料库进行分析，可以了解某种语言最基础的特征，相应的结论可以用于解决一般的、 没有特殊需求的文本分析任务。

专用语料库通常与某种具体的应用场景的相关性较高，与某种特定的文本分析任务密切相关的语科库都被称为专用语料库，其包含的内容五花八门且针对性强，有口语语料库、书面语语料库、方言语料库、在线评论语料库、新闻语料库、医疗语料库等。