Chapter 1

使用 NLTK 进行文本分类和词性标注

自然语言工具包（Natural Language Toolkit，NLTK）是一个用于 NLP 任务的 Python 库，其功能涉及分词、分句、执行进阶任务（如语法分析和文本分类），等等。NLTK 提供了一些针对自然语言的模块和接口，可用于执行诸如文档主题识别、词性标注、情感分析等任务。为了实验各种 NLP 任务，NLTK 还提供了各种文本语料库的模块，从基本的文本集合到带标签的结构化文本（如 WordNet）。

1.1 文本预处理及探索性分析

文本预处理步骤涉及如分词、词干提取和去除停用词之类的任务。对准备好的文本数据进行探索性分析可以了解其主要特征，包括文本的主题和词频分布。

1.1.1 分词

单词词元（token）是任何 NLP 任务都会涉及的文本基本单元。处理文本时，第一步就是将文本拆分为词元。NLTK 为此提供了不同类型的分词器。

为实现基于标点和空格的文本分割，NLTK 也提供了能同时标注出标点符号的 `wordpunct_tokenize` 分词器。

我们也可以使用 NLTK 的正则表达式分词器实现自定义分词。

1.1.2 词干提取

词干提取是一种文本预处理任务，将单词的相关或相似变体（例如 walking）转换为其基本形式（例如 walk），因为它们具有相同的含义。词干提取转换的基本操作之一是将单词的复数形式还原为单数形式，例如将 apples 还原为 apple。尽管这是一个非常简单的转换，但确实存在更加复杂的操作。

1.1.3 去除停用词

常用英文单词（例如 the、is 和 he 等）通常称为停用词。其他语言也有类似的常用词，同属这一类别。去除停用词是 NLP 应用中另一个常见的预处理步骤。在此步骤中，我们将删除那些对文档没有任何意义的词，例如语法冠词和代词。诸如 a、an、he 和 her 等单词都是需要去除的。停用词在整个文本中频繁出现，但它们本身可能不会对 NLP 任务（例如文本分类或搜索）产生任何影响。除英语外，NLTK 还为 21 种语言提供了停用词语料库。

1.1.4 探索性分析

获得词元数据后，常用的基本分析之一是对单词或词元及其在文档中的分布进行计数，从而更多地了解文档中的主要话题。