首先，导入 PyPDF2 模块。然后以读二进制模式打开 meetingminutes.pdf，并将它保存在 pdfFileObj 中。为了取得表示这个 PDF 的 PdfFileReader 对象，调用 PyPDF2.PdfFileReader()并向它传入 pdfFileObj。将这个 PdfFileReader 对象保存在 pdfReader 中。

该文档的总页数保存在 PdfFileReader 对象的 numPages 属性中❶。示例 PDF 文档有 19 页，但我们只提取第一页的文本。

要从一页中提取文本，需要通过 PdfFileReader 对象取得一个 Page 对象，它表示 PDF 中的一页。可以调用 PdfFileReader 对象的 getPage()方法❷，向它传入感兴趣的页码（在我们的例子中是 0），从而取得 Page 对象。

PyPDF2 在取得页面时使用从 0 开始的下标：第一页是 0 页，第二页是 1 页，以此类推。事情总是这样，即使文档中页面的页码不同。例如，假定你的 PDF 是从一个较长的报告中抽取出 3 页，它的页码分别是 42、43 和 44，要取得这个文档的第一页，需要调用 pdfReader.getPage(0)，而不是 getPage(42)或 getPage(1)。

在取得 Page 对象后，调用它的 extractText()方法，返回该页文本的字符串❸。文本提取并不完美：该 PDF 中的文本 Charles E.“Chas”Roemer, President，在函数返回的字符串中消失了，而且空格有时候也会没有。但是，这种近似的 PDF 文本内容，可能对你的程序来说已经足够了。