Как показано на рисунке 1 выше, извлеченный текст печатается на постоянной основе. Здесь нет ни абзацев, ни разделений предложений. Как указано в документации по PyPDF2, все текстовые данные возвращаются в том порядке, в котором они представлены в потоке содержимого страницы, и их использование может привести к неожиданностям. Это в основном зависит от внутренней структуры документа PDF и от того, как поток инструкций PDF был создан процессом записи PDF.

Извлечение изображений из PDF с помощью PyMuPDF

PyMuPDF упрощает извлечение изображений из документов PDF с использованием метода getPageImageList(). Листинг 3 основан на примере из вики-страницы PyMuPDF и извлекает и сохраняет все изображения из PDF в формате PNG постранично. Если изображение имеет цветовое пространство CMYK, оно будет сначала преобразовано в RGB.

Листинг 3: Извлечение изображений.