УДК 004

Извлечение текстовых данных из документов формата PDF, DOCX (DOC) с помощью сторонних библиотек

Ширяев В.В., Турчановская А.В.
Алтайский государственный университет, г. Барнаул
asmuddi628@gmail.com, turcanovskaa@gmail.com

Аннотация

В статье проведен сравнительный анализ библиотек таких языков программирования как: C#, Java, Python. В сравнении уделяется особое внимание возможности извлечения текстовых данных из большого количества однотипных документов формата PDF и DOCX (DOC). Рассмотрены основные проблемы применения библиотек.

Ключевые слова: извлечение текстовых данных, библиотеки для работы с текстовыми документами, обработка текстовых документов, сбор данных.

Для того, чтобы провести некоторые манипуляции над данными сначала необходимо их извлечь и структурировать. Особенно важен этот этап при сборе большого количества информации, так как он напрямую влияет на качество выполненной работы. Одним из основных форматов передачи и хранения данных является Portable Document Format – популярный кроссплатформенный формат электронных документов, использующий в своей основе язык PostScript. PDF изначально текстовый, а не бинарный формат. В его основе лежит несколько базовых типов данных с вставками из бинарного кода.

Одной из проблем при работе с форматом PDF является его неполная совместимость. Например, использование “небезопасных шрифтов” или большого количества векторной графики может стать причиной как неправильного отображения страниц, так и невозможности относительно простого извлечения данных. Именно поэтому при выборе инструмента для извлечения нужно исходить прежде всего из особенностей исходных документов, а после выбирать сам инструмент.

Другой важной деталью является то, что такой тип документов не содержит структуру таблиц в явном виде, и способ определения нужной позиции столбца или строки будет различаться в зависимости от содержимого документа. Большинство библиотек имеет набор инструментов для работы с таблицами, но как правило их результат неточен и требует поправок. В общем случае, текст, как и остальные объекты в PDF документе теряет структуру и хранится как векторное изображение, состоящее из текстовых символов.

Файлы формата DOC и DOCX имеют существенные различия, особенно если попытаться извлечь текстовые данные напрямую. DOC представляет собой расширение в формате двоичного файла, DOCX – написано на языке разметки XML, что позволяет распаковать файл, получить компоненты XML и использовать их без привлечения библиотек. При работе с этими файлами следует сначала воспользоваться возможностями конвертации файлах перед основным извлечением данных.

Наиболее частым подходом при сборе данных является использование библиотеки с готовым набором функций и методов, позволяющим сосредоточиться на извлечении информации, а не на написании примитивного синтаксического анализатора или “парсера”.

Цель работы – проанализировать существующие решения для извлечения текстовых данных, выявить их достоинства и недостатки.

Рассмотрим несколько популярных библиотек. iText 7 Community – это библиотека PDF с открытым исходным кодом, состоящая из версий на языках Java и .NET. Она

может использоваться только с лицензией AGPL. iText 7 Community – это простая, производительная и расширяемая библиотека, которая может справляться с проблемами современного цифрового документооборота. Также существуют несколько различных дополнений, реализованных на коммерческой основе, в том числе и pdf2Data, позволяющая извлекать данные по настраиваемым шаблонам. Данная библиотека поддерживает все основные функции по работе с pdf и имеет исчерпывающую документацию. Из достоинств данной библиотеки стоит отметить постоянную поддержку продукта, доступность для двух языков. Основной недостаток – достаточно сложная структура классов.

PDFsharp – еще один мощный инструмент с полным набором функций для работы с PDF документами, опубликован под лицензией MIT. Из преимуществ стоит выделить простоту использования. Для PDFsharp и iText 7 библиотек существует необходимость в дополнительной подготовке документа перед прочтением или перекодировки данных, если эти данные представлены кириллицей.

PDF Mosaic – полностью бесплатная библиотека, не требующая сторонних приложений или библиотек. Её основной особенностью является частичное чтение документа, так называемый “ленивый парсинг”. Такой подход позволяет ускорить процесс открытия и работы с многостраничными документами. Тестирование скорости открытия большого документа (1300 страниц) с помощью PDF Mosaic и библиотекой iTextSharp 0.9 и 27.5 секунд соответственно.

Для работы с файлами Microsoft Office на платформе .NET можно воспользоваться библиотекой DocX, которая не требует установленных пакетов Microsoft Office и является наиболее простой среди остальных библиотек, обновляется, а также имеет лицензию Ms-PL.

Apache PDFBox – широкий набор инструментов для создания и изменения PDF на языке Java, под лицензией Apache 2. Кроме основных возможностей извлечения текста, соединения и заполнения файлов, существует поддержка текста в стандарте кодирования символов Unicode, что позволяет сразу работать с кириллицей. Один из недостатков состоит в том, что извлечение данных из таблиц полностью не реализовано, библиотека помечает таблицы как неопределенные структуры.

Apache POI – свободный API, который позволяет использовать файлы MS Office в среде Java. Компоненты HWPF и XWPF позволяют полноценно работать с расширениями DOC и DOCX соответственно.

Для извлечения текста из таблиц документа лучше всего использовать библиотеки, которые изначально разрабатывались для этого. Одной из таких библиотек на языке Java является tabula-java, распространяемся под MIT лицензией. Также существует версия и для языка Python. Главным недостатком библиотеки является невозможность разделения разных таблиц по отдельным структурам, так, как это можно увидеть при просмотре документа.

pdfminer – популярная библиотека для извлечения текста на языке Python, которая до сих пор развивается. Существует несколько дополнительных пакетов на основе pdfminer, упрощающих процесс извлечения, т.к. для этой библиотеки нет подробной документации. Также невозможен процесс извлечения таблиц, без дополнительной конвертации документа в другие форматы.

Другая похожая библиотека – PyPDF2, имеющая документацию. Позволяет быстро извлекать текст для последующей обработки, что хорошо подходит для случаев, когда необходимо точечно выделить данные.

python-docx – один из модулей для работы с Microsoft Word файлами. Возможность использования этого модуля вместе с применением регулярных выражений позволяют извлекать данные из файла так, если бы это был обычный текстовый файл.

Результат любой работы, связанной с анализом или использованием больших данных

полностью зависит от успешности этапа извлечения и консолидации данных, скорости извлечения, а в некоторых случаях важно учитывать особенности хранения данных в том или ином расширении, поэтому нельзя недооценивать использование библиотек.

Таким образом, существует множество библиотек для извлечения информации, каждая из которых может эффективно использоваться несмотря на сложность и изменчивость PDF документов, а также устройство файлов Microsoft Word. Выбрав любой инструмент для сбора информации, требуется предварительная обработка для структуризации полученных данных в зависимости от задачи и требований к собираемым данным.