Project TS

31/10/2017

Luke Evans

Jania Okwechime

London UK

Financials

	2015	2016	2017
Revenue	£17000000	18000000	2000000
Headcount	100	150	200

PDFMiner is a tool for extracting information from PDF documents. Unlike other PDF-related tools, it focuses entirely on getting and analyzing text data. PDFMiner allows one to obtain the exact location of text in a page, as well as other information such as fonts or lines. It includes a PDF converter that can transform PDF files into other text formats (such as HTML). It has an extensible PDF parser that can be used for other purposes than text analysis.

Continued

- **Features**
- Written entirely in Python. (for version 2.4 or newer)
- Parse, analyze, and convert PDF documents.
- PDF-1.7 specification support. (well, almost)
- CJK languages and vertical writing scripts support.
- Various font types (Type1, TrueType, Type3, and CID) support.
- Basic encryption (RC4) support.
- PDF to HTML conversion (with a sample converter web app).
- Outline (TOC) extraction.
- Tagged contents extraction.
- Reconstruct the original layout by grouping text chunks.