



PDFlib TET

Text Extraction Toolkit



Was ist PDFlib TET?

PDFlib TET (Text Extraction Toolkit) ist ein Entwicklungswerkzeug, mit dem Sie Text zuverlässig aus PDF-Dokumenten extrahieren können. TET stellt den Text eines PDF-Dokuments als Unicode-Strings zur Verfügung und liefert detaillierte Informationen über Schriften und Zeichen sowie die Position auf der Seite.

TET verfügt über einen ausgefeilten Algorithmus zur Inhaltsanalyse und kann damit Wortgrenzen erkennen, Text zu Spalten zusammenfassen oder redundanten Text entfernen, zum Beispiel Schatteneffekte oder künstliche Fettschrift. Mit der pCOS-Schnittstelle können Sie zudem beliebige Objekte aus einem PDF-Dokument abfragen, zum Beispiel Metadaten oder Hypertext. Mit PDFlib TET können Sie:

- ▶ Eine Suchmaschine für PDF-Dateien implementieren
- ▶ Text aus PDFs extrahieren, um ihn zum Beispiel in einer Datenbank zu speichern
- ▶ Texte aus PDF-Dokumenten in andere Formate wie XML konvertieren
- ▶ PDFs abhängig von ihren Inhalten verarbeiten

Vorteile von PDFlib-Software

Zuverlässig

Weltweit arbeiten viele Tausend Programmierer mit unserer Software. PDFlib erfüllt alle Qualitäts- und Geschwindigkeitskriterien für den Einsatz auf großen Servern. Alle PDFlib-Produkte sind für den zuverlässigen, unbeaufsichtigten 24-Stunden-Betrieb ausgelegt.

Schnell und einfach

PDFlib-Produkte sind unglaublich schnell – bis zu Tausenden von Seiten pro Sekunde. Die Programmierschnittstelle ist übersichtlich und einfach zu erlernen.

PDFlib ist überall

Unsere Produkte unterstützen alle internationalen Sprachen sowie Unicode. Sie werden von Kunden in der ganzen Welt eingesetzt.

Professioneller Support

Bei Problemen bietet Ihnen unser Support-Team professionelle Unterstützung. Um den reibungslosen Ablauf unternehmenskritischer Anwendungen zu gewährleisten, können Sie Ihre Software-Lizenz durch einen Supportvertrag ergänzen. Ein Supportvertrag garantiert Ihnen kurze Antwortzeiten und Zugang zu den jeweils neuesten Versionen.

Funktionalität von PDFlib TET

Unterstützte PDF-Eingabe

PDFlib TET verarbeitet alle gängigen Varianten von PDF:

- ▶ Alle PDF-Versionen bis PDF 1.6 (Acrobat 7)
- ▶ Alle Font- und Encodingtypen: Acrobat-Standardschriften sowie TrueType-, PostScript-, OpenType- und CID-Schriften
- ▶ Verschlüsseltes PDF mit 40- und 128-Bit-Verschlüsselung (bei passenden Berechtigungen oder Angabe des Kennworts)

Unicode

Da Textinhalte in PDF üblicherweise nicht in Unicode kodiert sind, normalisiert PDFlib TET sämtlichen Text aus einem PDF-Dokument nach Unicode:

- ▶ TET konvertiert alle Textinhalte nach Unicode. In C und anderen Sprachen, die Unicode nicht unterstützen, wird der Text in den Formaten UTF-8 oder UTF-16 zurückgegeben, in Unicode-fähigen Sprachen dagegen in normalen Strings.
- ▶ Ligaturen und andere zusammengesetzte Zeichen werden als Folge der zugrunde liegenden Unicode-Zeichen ausgegeben.
- ▶ Herstellerabhängig kodierte Unicode-Zeichen in der Private Use Area (PUA) werden erkannt und nach Möglichkeit in den allgemeinen Unicode-Bereich abgebildet.
- ▶ Zeichen ohne eigene Unicode-Zuordnung werden erkannt und auf ein konfigurierbares Ersatzzeichen abgebildet, so dass Fehlinterpretationen verhindert werden.

Unterstützung von chinesischem, japanischem und koreanischem Text

TET bietet umfassende Unterstützung zur Extraktion von chinesischem, japanischem und koreanischem Text. Alle vordefinierten CJK-CMaps (Zeichensätze) werden erkannt; horizontale und vertikale Schreibrichtung werden korrekt behandelt.

Geometrie

TET liefert genaue metrische Daten zum Text, zum Beispiel die Position auf der Seite, die Zeichenbreiten und die Textrichtung. Bei der Textextraktion können bestimmte Seitenbereiche explizit ausgeschlossen oder einbezogen werden, zum Beispiel um Kopf- und Fußzeilen oder Seitenränder zu übergehen.

Inhaltsanalyse und Worterkennung

TET liefert nicht nur grundlegende Zeicheninformationen, sondern beinhaltet auch ausgefeilte Algorithmen zur Inhaltsanalyse:

- ▶ Wortgrenzen werden ermittelt und Wörter korrekt erkannt.
- ▶ Getrennte Silben werden zu Wörtern zusammengesetzt.
- ▶ Redundanter Text wird entfernt, zum Beispiel bei Schatteneffekten oder künstlicher Fettschrift.
- ▶ Absätze werden gemäß der Lesereihenfolge umsortiert.
- ▶ Über die Seite verteilter Text wird umsortiert.
- ▶ Vollständige Textzeilen werden zusammengefügt.

Konfigurationsmöglichkeiten für problematische PDFs

TET verfügt über spezielle Verfahren und Hilfsmittel zur Behandlung verschiedener Klassen von PDFs, deren Text sich mit anderen Produkten nicht korrekt extrahieren lässt. Außerdem bietet TET zahlreiche Konfigurationsmöglichkeiten, mit denen Sie die Verarbeitung problematischer Dokumente erheblich verbessern können:

- ▶ Sie können das Unicode-Mapping steuern, indem Sie eigene Tabellen übergeben, die Zeichencodes oder -namen auf Unicode abbilden.
- ▶ PDFlib FontReporter ist ein kostenloses Tool zur Analyse von Fonts, Encodings und Glyphen in PDF-Dokumenten. Es ist als Plugin für Adobe Acrobat 5-7 auf dem Mac und unter Windows verfügbar.
- ▶ Eingebettete Schriften werden nach Hinweisen durchsucht, die beim Unicode-Mapping nützlich sind. Ist eine Schrift nicht eingebettet, lässt sich die Extraktion durch externe Fontdateien oder Systemschriften weiter verbessern.

Einfacher Zugriff auf PDF-Objekte mit pCOS

TET enthält die Programmierschnittstelle pCOS. PDFlib pCOS bietet Ihnen eine einfache und elegante Methode, um aus PDF-Dokumenten Informationen abzurufen, die nicht zum Seiteninhalt gehören. PDF-Metadaten, Hypertext oder Seitengrößen sind zum Beispiel bequem mit pCOS abfragbar.

TET als Bibliothek oder Kommandozeilen-Tool?

TET wird als Software-Bibliothek (Komponente) für verschiedene Entwicklungsumgebungen sowie als Kommandozeilen-Tool für Batch-Prozesse ausgeliefert. Beide Ausführungen bieten den gleichen Funktionsumfang, eignen sich aber für unterschiedliche Einsatzbereiche.

Die TET-Software-Bibliothek eignet sich...

...zur Integration in Ihre Desktop- oder Server-Anwendungen. Programmierbeispiele für alle unterstützten Sprachbindungen sind im TET-Paket enthalten.

Das TET-Kommandozeilen-Tool eignet sich...

...zur Batch-Verarbeitung von PDF-Dokumenten. Es erfordert keine Programmierung, sondern kann über leistungsfähige Kommandozeilen-Optionen gesteuert und damit in komplexe Arbeitsabläufe integriert werden. Das TET-Kommandozeilen-Tool ergänzt die Bibliothek um folgende Funktionen:

- ▶ Zusätzlich zur Erstellung von reinem Text können PDF-Dokumente nach XML konvertiert werden.
- ▶ Das TET-Kommandozeilen-Tool kann auch in Umgebungen aufgerufen werden, die den Einsatz der TET-Bibliothek nicht unterstützen.

Unterstützte Entwicklungsumgebungen

PDFlib TET läuft überall – auf praktisch allen Computersystemen. Wir unterstützen alle gängigen Varianten von Windows, Mac OS, Linux und Unix sowie IBM eServer iSeries und zSeries.

Der Kern von TET ist in C geschrieben und auf Schnelligkeit und geringen Overhead optimiert. Über ein einfaches API (Application Programming Interface) lässt sich die TET-Funktionalität in zahlreichen Programmiersprachen nutzen:

- ▶ COM für VB, ASP, Borland Delphi, etc.
- ▶ C und C++
- ▶ Java einschließlich Servlets und Java Application Server
- ▶ .NET für C#, VB.NET, ASP.NET, etc.
- ▶ PHP Hypertext Processor
- ▶ RPG (IBM eServer iSeries)

Lizenzierung

Bei der Lizenzierung können Sie zwischen verschiedenen Modellen für Serverlizenzen, Integrations- und Firmenlizenzen sowie Quellcodelizenzen wählen. Ergänzend bieten wir Supportverträge für umfangreichen technischen Support mit kurzen Reaktionszeiten und kostenlosen Software-Aktualisierungen an.

Über PDFlib GmbH

PDFlib GmbH widmet sich seit 1997 ausschließlich der PDF-Technologie. Die PDF-Entwicklung steht im Mittelpunkt all unserer Tätigkeiten, und alle Produkte haben mit PDF zu tun. Wir verfügen über langjährige Erfahrung und können es uns erlauben, auch noch die kleinsten Details des PDF-Formats zu untersuchen. Unsere Produkte werden auf der ganzen Welt vertrieben, wobei Nordamerika, Japan und Europa die wichtigsten Märkte sind.



Kontakt

Evaluierungsversionen mit vollem Funktionsumfang und Dokumentation sowie Beispielen sind für alle unterstützten Systeme auf unserer Website verfügbar. Weitere Informationen erhalten Sie unter:

PDFlib GmbH
 Tal 40, 80331 München
 Tel. +49 • 89 • 29 16 46 87
 Fax +49 • 89 • 29 16 46 86
 sales@pdflib.com
 www.pdflib.com