|  |  |  |
| --- | --- | --- |
| unitbv.emf | **Universitatea *Transilvania* din Brașov**  **Facultatea de Inginerie Electrică și Știința Calculatoarelor** | iesc.emf |

**Extragerea conținutului de tip text dintr‐un fișier PDF**

**[PROIECT PCLP3]**

|  |  |
| --- | --- |
| Serbanoiu Ciprian  *Gr. 4421, AIA II* |  |

**BRAȘOV, 2013**

**Extragerea conținutului de tip text dintr‐un fișier PDF**

**1.Fisiere PDF(Portable Document Format)**



# Portable Document Format(PDF) este un format de fisier creat de Adobe System, in 1993, care acum este un standard deschis formal, mentinut incepand cu 1 iulie 2008 de Organizatia Internationala pentru Standardizare (ISO). Scopul definit al formatului PDF este de a permite portabilitatea documentelor, sau cu alte cuvinte accesul la documente indiferent de sistemul de operare, calculatorul folosit si programele detinute de utilizatori. Fiecare fisier PDF include o descriere completa a documentului in format 2D care include texte, fonturi, imagini si grafica vectoariala 2D, si mai tarziu, desenele 3D.

In anul 1990 compania Adobe avea un cuvant greu de spus in domeniul graficii datorita limbajului Postscript inventat.Limbajul Postscript fiind un limbaj interpretat de platforma pe care ruleaza. Compania Adobe dezvolta un driver de imprimanta care face din PS un standard pentru grafica profesionala.In 1991 Jhon Warnock pune bazele unui proiect grandios numit "Camelot" prin care viseaza la un birou fara hartie.Scopul acestui proiect este gasirea unui format de fisier care sa permita oameniilor sa poata partaja informatia independent de platforma.Acest proiect sub numele de "Carousel" apare pentru prima data in 1992 si este premiat la Comdex.In 1993 apare prima versiune a programului Adobe Reader.

Formatul PDF reuseste sa devina un format independent de platforma datorita inrudirii cu limbajul Postscript.Adobe adauga formatului PDF mai mult dinamism permitand includerea in formatul PDF a elementelor multimedia.In 1994 se lanseaza renumitul software Adobe Acrobat Reader care exista pentru orice platforma software si este gratuit.

Funcția principală a acestui program este de a permite oricărui utilizator vizualizarea fișierelor PDF și apoi de a interacționa cu ele. Datorită ultimelor tehnologii incluse în formatul de fișiere PDF utilizatorii programului gratuit [Adobe Reader](http://www.adobe.com/products/acrobat/readstep2.html) pot comenta și adnota fișierele PDF sau pot completa formularele din fișerele PDF. Comentariile sau informațiile introduse în formulare se pot întoarce la creatorul fișierului PDF permițând astfel colaborarea sau conlucrarea în cadrul unor echipe. Formatul PDF este denumit din această cauză "document inteligent". Navigarea prin fișierul PDF se face liniar pagina cu pagina sau neliniar în funcție de opțiunile utilizatorului. Cel ce crează un fișier PDF poate pune la dispoziția utilizatorului diverse interacțiuni: butoane, navigare neliniară, sunet, video, calcule pe baza informațiilor introduse, pagini generate dinamic sau link-uri gen web. Astfel parcurgerea unui fișier PDF poate deveni o experiență unică relativ la celelalte formate de fișier.

Deaorece toate aceste metode de interacțiune presupun creații unice devine importantă ideea respectării drepturilor de autor și securitatea informațiilor conținute în aceste fișiere. În prezent formatul PDF permite asigurarea securității informațiilor cu mai multe metode în funcție de scopul și destinația fișierelor PDF. Soluțiile sunt multiple și acoperă toate nevoile utilizatorilor la multiple nivele.

Am amintit mai sus de programul [Adobe Reader](http://www.adobe.com/products/acrobat/readstep2.html) care este gratuit și permite vizualizarea, comentarea fișierelor și completarea formularelor conținute în fișierele PDF. Programul de bază care permite crearea fișierelor PDF din aplicațiile tip office este Adobe Acrobat Standard. Acesta simplifică foarte mult crearea fișierelor PDF din cele mai utilizate aplicații de tip office (Word, Excel, Acces, Powerpoint, Outlook, Internet Explorer, Visio, etc) și în principiu din orice aplicație care poate tipări pe imprimantă. Pe lângă posibilitatea producerii fișierelor PDF din toate aceste surse se mai pot crea și formulare sau adăuga o mulțime de alte elemente care fac din PDF un format dinamic. Pachetul software mai conține programul Adobe Acrobat Distiller care face conversia fișierelor PS în format PDF. La instalarea programului Adobe Acrobat mai apare în sistem o imprimantă virtuală care permite crearea unor fișiere PDF din orice aplicație.

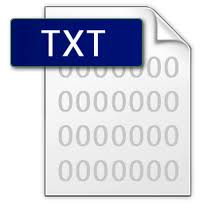
Profesioniștii din industria grafică au nevoie de funcții speciale de analiză, modificare și funcții speciale tiparului. Pentru ei există software-ul [Adobe Acrobat Professional](http://www.adobe.com/products/acrobatpro/) care le pune la dispoziție toate aceste unelte specializate. Această versiune a familiei Adobe Acrobat conține și programul Adobe Designer destinat creării unor formulare inteligente care se pot integra apoi în workflow-ul unei companii.

Familia [Adobe Acrobat](http://www.adobe.com/products/acrobat/?ogn=EN_US-gntray_prod_acrobat_family_home) s-a mărit prin apariția unei noi versiuni:[Adobe Acrobat 3D](http://www.adobe.com/products/acrobat3d/). După cum îi spune și numele această versiune este destinată în principal arhitecților și inginerilor care până acum erau nevoiți să-și exprime creațiile doar în două dimensiuni. De acum există unelte special destinate transformării acestor creații în fișiere PDF devenind astfel disponibile publicului larg indiferent de platformă.

Achiziția companiei Macromedia de către Adobe a deschis noi perspective formatului PDF. Integrarea cu formatul Flash a adus o nouă interfață programelor din familia Acrobat și posibilitatea utilizării tehnologiei Flash în interiorul formatului PDF.

Formatul PDF este un standard aprobat de ISO atât ca format de fișier pentru managementul documentelor cât și ca bază pentru alte domenii de activitate. Există mai multe standarde internaționale bazate pe formatul PDF. Industria grafică a adoptat standardul PDF/X (cu mai multe variante în funcție de workflow-ul adoptat) pentru schimbul de fișiere grafice. PDF/A este standardul adoptat de ISO pentru arhivarea documentelor.

**2.Fisere TEXT**



**Un fișier text** este un tip de [fișier](http://ro.wikipedia.org/wiki/Fi%C8%99ier) în care datele sunt stocate ca o secvență de caractere, într-o codificare predefinită (de obicei [ASCII](http://ro.wikipedia.org/wiki/ASCII), dar mai recent și [Unicode](http://ro.wikipedia.org/wiki/Unicode)). Este un tip canonic de fișier unde datele sunt stocate ca "secvențe de [biți](http://ro.wikipedia.org/w/index.php?title=Bi%C8%9Bi&action=edit&redlink=1)" (binar). Nu orice fișier binar este un fișier text. Prin definire, un fișier text este codificat unitar, conținutul său fiind lizibil și editabil direct, prin intermediul unui editor simplu de text.

Într-un fișier text, tipul de date este considerat explicit ca fiind textul neformatat. Organizarea unui astfel de fișier este pe rânduri (care se generează la apăsarea tastei Enter/Return), fiecare rând de text fiind delimitat de unul sau mai multe caractere de sfârșit de rând (caractere [EOL](http://ro.wikipedia.org/w/index.php?title=EOL&action=edit&redlink=1)). Acestea diferă în funcție de sistemul de operare folosit pentru crearea și editarea fișierului. Astfel, Windows folosește două caractere de control (ASCII 13 urmat de ASCII 10, sau, mai scurt CR+LF) pentru a semnaliza sfârșitul unei linii de text, pe când sistemele tip [UNIX](http://ro.wikipedia.org/wiki/UNIX) (incluzând aici și[Linux](http://ro.wikipedia.org/wiki/Linux) și [Mac OS X](http://ro.wikipedia.org/wiki/Mac_OS_X)) folosesc numai caracterul LF, iar sistemele [Mac OS](http://ro.wikipedia.org/wiki/Mac_OS) pre-Unix (versiunile 9 sau mai vechi), folosesc doar caracterul CR. Astăzi, majoritatea sistemelor și editoarelor de text sunt perfect capabile să se folosească și să convertească automat fișierele text între diverse arhitecturi. O excepție notabilă în acest sens este, însă, editorul [Notepad](http://ro.wikipedia.org/wiki/Notepad), care, în cazul în care este pus să deschidă un fișier text generat pe o altă platformă, va afișa caracterul de control LF sau CR după fiecare rând, neîmpărțind textul în rânduri.

Codificarea fișierelor text se face în general ori în [ASCII](http://ro.wikipedia.org/wiki/ASCII), ori în [Unicode](http://ro.wikipedia.org/wiki/Unicode) (standardul [UTF-8](http://ro.wikipedia.org/w/index.php?title=UTF-8&action=edit&redlink=1), compatibil invers cu ASCII) permițând o portabilitate ridicată între diverse tipuri de sisteme.

Pentru utilizator, un astfel de fișier, vizualizat ca atare, apare ca text neformatat și nearanjat în pagină. Fișierele text, per se, nu pot stoca decât caractere printabile (litere și cifre) și un număr foarte limitat de caractere de control. Astfel, în structura însăși a fișierului nu se pot stoca elemente de formatare precum culori sau imagini. Fișierele text pot astfel fi folosite pentru activități simple, precum luatul de notițe. Cu toate acestea, un fișier text poate stoca date de intrare pentru un program terț. Astfel de utilizări, datorită portabilității uriașe a formatului (neținând seama de problema sfârșitului de linie) și datorită faptului că conținutul fișierului este ușor lizibil și editabil ca text a dus la apariția multor standarde care codifică informații de diverse feluri ca text lizibil stocat în fișiere text, și care procesate cu ajutorul unor alte programe iau altă formă. Mai mult, [codul sursă](http://ro.wikipedia.org/w/index.php?title=Codul_surs%C4%83&action=edit&redlink=1) al oricărui program, în orice [limbaj de programare](http://ro.wikipedia.org/wiki/Limbaj_de_programare), este salvat pe disc sub formă de fișier text.

3. Convertirea PDF----->TEXT



Pentru Conversia din format PDF in format TEXT am utilizat libraria IText Sharp.

iTextSharp este o librărie ideală pentru programatorii care vor sa îmbunătăţească aplicaţiile web sau alte tipuri de aplicaţii inserând documente dinamice de tip PDF. iTextSharp nu poate fi folosit ca o aplicaţie desktop precum Acrobat Reader sau alte aplicaţii de vizualizare a fişierelor PDF, ci se foloseşte la crearea si manipularea automată a fişierelor PDF din propriile aplicaţii.

iTextSharp poate fi folosit la:  
-generarea de documente şi rapoarte bazate pe date din fişiere xml sau din baze de date  
-adăugarea de bookmarks, page numbering, watermarks, etc.  
-separarea sau concatenarea paginilor unor fişiere PDF existente.  
-completarea de formulare PDF, adăugarea semnaturilor digitale,etc.  
-vizualizarea fişierelor pdf în browsere web.

**4.Codul programului pentru conversia din PDF in TEXT**

using System;

using System.Collections.Generic;

using System.ComponentModel;

using System.Data;

using System.Drawing;

using System.Linq;

using System.Text;

using System.Windows.Forms;

using System.IO;

using iTextSharp.text;

using iTextSharp.text.pdf;

using iTextSharp.text.pdf.parser;

namespace WindowsFormsApplication1

{

public partial class Form1 : Form

{

public Form1()

{

InitializeComponent();

}

private void Form1\_Load(object sender, EventArgs e)

{

}

private void button1\_Click(object sender, EventArgs e)

{

OpenFileDialog dlg = new OpenFileDialog();

string filePath;

dlg.Filter = "PDF Files(\*.PDF)|\*.PDF|All Files(\*.\*)|\*.\*";

if (dlg.ShowDialog() == DialogResult.OK)

{

filePath = dlg.FileName.ToString();

string strText = string.Empty;

try

{

PdfReader reader = new PdfReader(filePath);

for (int page = 1; page <= reader.NumberOfPages; page++)

{

ITextExtractionStrategy its = new iTextSharp.text.pdf.parser.LocationTextExtractionStrategy();

String s = PdfTextExtractor.GetTextFromPage(reader, page, its);

s = Encoding.UTF8.GetString(ASCIIEncoding.Convert(Encoding.Default, Encoding.UTF8, Encoding.Default.GetBytes(s)));

strText = strText + s;

richTextBox1.Text = strText;

}

reader.Close();

}

catch (Exception ex)

{

MessageBox.Show(ex.Message);

}

}

}

private void richTextBox1\_TextChanged(object sender, EventArgs e)

{

}

private void button2\_Click(object sender, EventArgs e)

{

StreamWriter File = new StreamWriter(@"C:\Users\Ciprian\Desktop\Proiect C#\Pdf\_to\_text.txt");

System.Diagnostics.Process.Start(@"C:\Users\Ciprian\Desktop\Proiect C#\Pdf\_to\_text.txt");

File.Write(richTextBox1.Text);

File.Close();

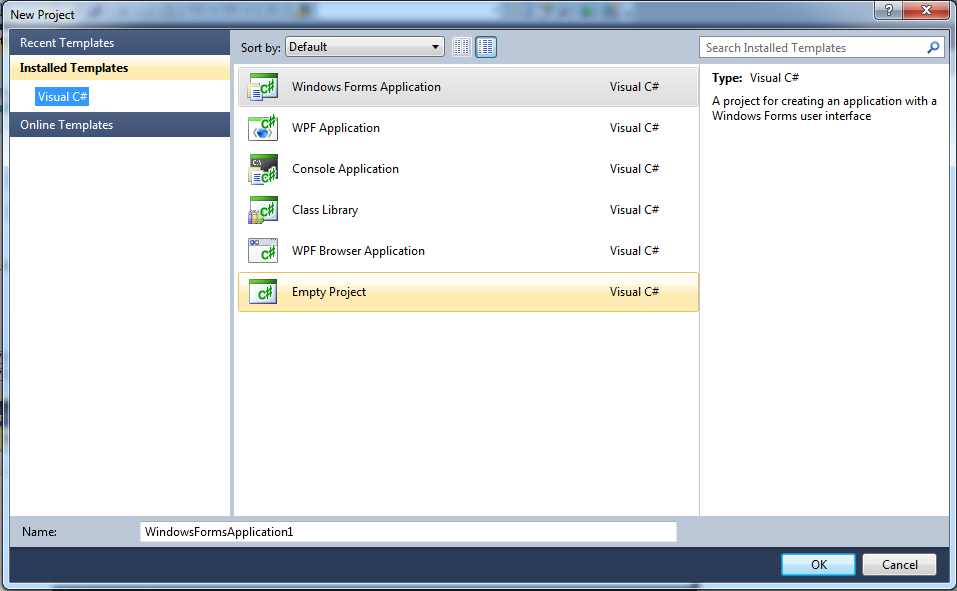
}

}

}

**5.Creerea programului****pentru conversia din PDF in TEXT**

*Pasul 1:Petru a creea programul de convertit din pdf in text am utilizat programul Microsoft Visual Studio C# 2010*

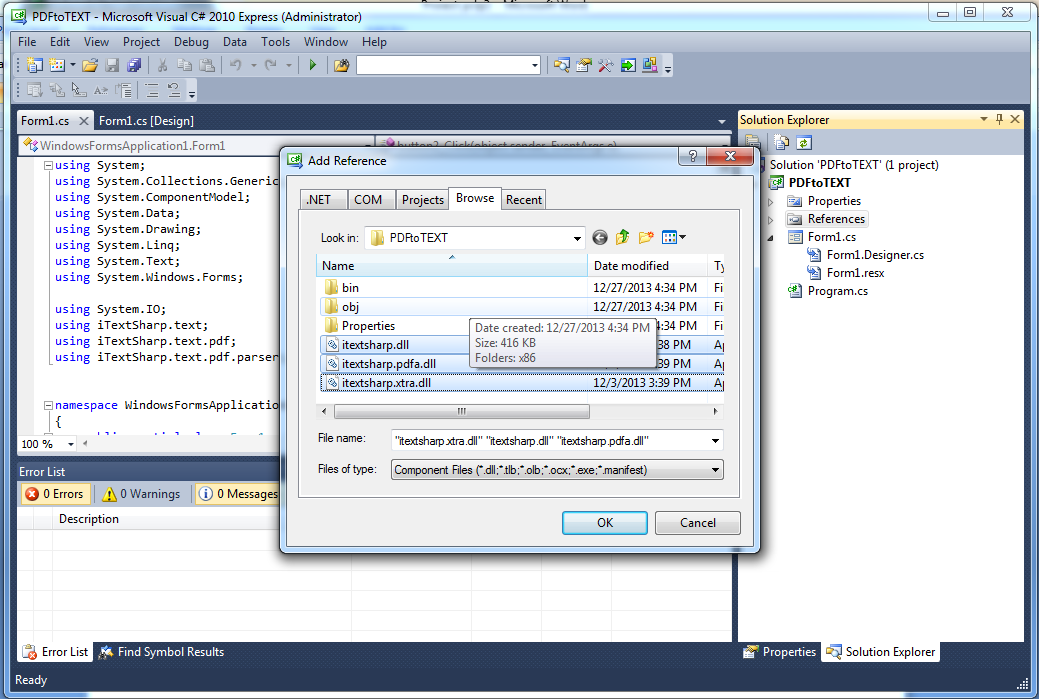
*Pasul 2:Am creat o noua windows forms application*****

*Pasul 3:Am descarcat libraria iTextSharp de pe site-ul:*

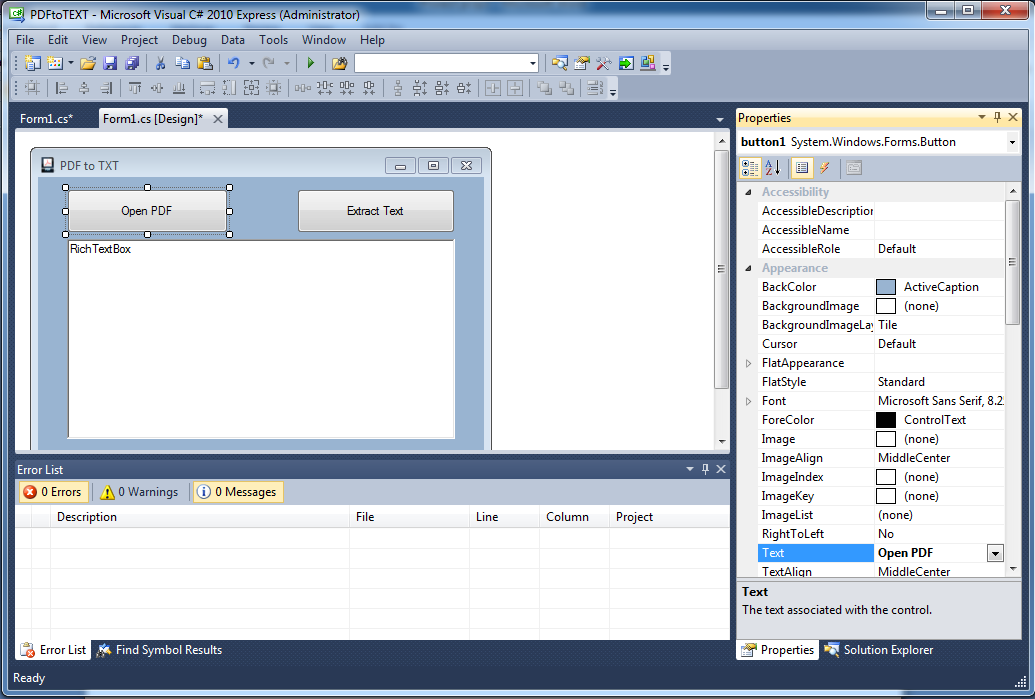
[*http://sourceforge.net/projects/itextsharp/*](http://sourceforge.net/projects/itextsharp/)

*Pasul 4:Am extras fisierele itextsharp.dll,ITextSharp.xml,itextsharp.pdfa.dll, itextsharp.xtra.dll din arhivele itextsharp-dll-core.zip, itextsharp-dll-pdfa.zip respectiv itextsharp-dll-xtra.zip in folderul unde am creeat proiectul.*

*Pasul 5:Din fereastra din dreapta a programului C# "Solution Explorer" click dreapta pe References , add reference apoi mergem la Browse si selectam fisierele itextsharp.dll,itextsharp.pdfa.dll, itextsharp.xtra.dll si dam ok.*

****

*Pasul 6:Am creat 2 butoane si o caseta text "richTextBox" un buton pentru a alege fisierul pdf pe care dorim sa il convertim pe care il numim "Open PDF",caseta text unde ne va aparea continutul text din fisierul pdf si cel de-al doilea buton pe care il numim "Extract Text" pe care il utilizam pentru a crea un fisier text cu continutul fisierului pdf.*

**

*Pasul 7:Facem dublu click pe butonul Open PDF care ne va duce in codul programului unde mai intai in topul programului trebuie sa importam:*

using System.IO;

using iTextSharp.text;

using iTextSharp.text.pdf;

using iTextSharp.text.pdf.parser;

*Iar in interiorul primului buton vom scrie urmatorul cod:*

OpenFileDialog dlg = new OpenFileDialog();

string filePath;//variabila string

dlg.Filter = "PDF Files(\*.PDF)|\*.PDF|All Files(\*.\*)|\*.\*";

\\Acest filtru este utilizat pentru a ne arata doar fisierele cu extensie pdf atunci cand ne alegem fiserul pe care vrem sa il convertim

if (dlg.ShowDialog() == DialogResult.OK)

{

filePath = dlg.FileName.ToString();

\\ pana aici acest cod este folosit pentru a ne alege fiserul pdf si sa il transforme intr-o variabila string

string strText = string.Empty;

try

{

PdfReader reader = new PdfReader(filePath);

for (int page = 1; page <= reader.NumberOfPages; page++)

{

ITextExtractionStrategy its = new iTextSharp.text.pdf.parser.LocationTextExtractionStrategy();

String s = PdfTextExtractor.GetTextFromPage(reader, page, its);

s = Encoding.UTF8.GetString(ASCIIEncoding.Convert(Encoding.Default, Encoding.UTF8, Encoding.Default.GetBytes(s)));

strText = strText + s;

richTextBox1.Text = strText;

}

reader.Close();

}

catch (Exception ex)

{

MessageBox.Show(ex.Message);

}

}

}

\\in blocul try catch este codul pentru a transforma converti textul din fisierul pdf in casuta richtextbox cu ajutorul itxextsharp.

*In interiorul celui de-al doilea buton scriem urmatorul cod:*

StreamWriter File = new StreamWriter(@"C:\Users\Ciprian\Desktop\Proiect C#\Pdf\_to\_text.txt");\\calea unde fa fi creat fisierul text

System.Diagnostics.Process.Start(@"C:\Users\Ciprian\Desktop\Proiect C#\Pdf\_to\_text.txt");\\pentru a deschide fisierul txt

File.Write(richTextBox1.Text);

*\\acest cod este folosit pentru a transforma tot ce este scris in richtextbox intr-un fisier text.*