ADEPTIA

443 North Clark St, Suite 350
Chicago, IL 60610
Phone: (312) 229-1727

PDF to Text Conversion Usage Guide

Overview

This guide describes the details about the entire API that is used to convert a PDF file into a tab delimited text file. PDF represents unstructured data and in order to get the data from PDF in structured format it is interpreted according to the screen graphics(x and y coordinates).

Details

Method "**extractText**" for an object of Class "**ConvertPDFtoText**" when called with **debug** flag set will generates an output text file containing position(x and y co-ordinates) information for all the data as shown below

```
Crs ID(51,177)  Course Title(87,177)    Mark(149,177)   Cred Att/Cmp(176,177)
Desert Mountain High School(81,187)
Yr:(50,197) 2008(64,197)    Month:(109,197) 12(131,197) Grade:(163,197) 09(190,197)
2318(55,207)    Wrld Hist/Wrld Geog(87,207) A(159,207)  0.50(185,207)   /(199,207)  0.50(203,207)
H(37,215)   3112(55,216)    English I Honors(87,215)    A(159,216)  0.50(185,215)   /(199,216)  0.50(203,215)
H(37,224)   4524(55,224)    Geometry/Trigonomet Honors(87,224)  A(159,224)  0.50(185,224)   /(199,224)  0.50(203,224)
H(37,232)   5412(55,233)    Biology I Honors(87,232)    A(159,233)  0.50(185,232)   /(199,233)  0.50(203,232)
6010(55,241)    Spanish I(87,241)   A(159,241)  0.50(185,241)   /(199,241)  0.50(203,241)
N(37,249)   7450(55,250)    Intro Physical Education(87,249)    A(159,250)  0.50(185,249)   /(199,250)  0.50(203,249)
Cred Att:(47,258)   3.00(73,258)    Cred Cmp:(109,258)  3.00(141,258)   GPA:(178,258)   4.500(195,258)
Desert Mountain High School(81,267)
Yr:(50,277) 2009(64,277)    Month:(109,277) 5(131,277)  Grade:(163,277) 09(190,277)
2319(55,287)    Wrld Hist/Wrld Geog(87,286) A(159,287)  0.50(185,286)   /(199,287)  0.50(203,286)
H(37,295)   3113(55,295)    English I Honors(87,295)    A(159,295)  0.50(185,295)   /(199,295)  0.50(203,295)
H(37,303)   4525(55,304)    Geometry/Trigonomet Honors(87,303)  A(159,304)  0.50(185,303)   /(199,304)  0.50(203,303)
H(37,312)   5413(55,313)    Biology I Honors(87,312)    A(159,313)  0.50(185,312)   /(199,313)  0.50(203,312)
6011(55,321)    Spanish I(87,321)   A(159,321)  0.50(185,321)   /(199,321)  0.50(203,321)
N(37,329)   7451(55,330)    Intro Physical Education(87,329)    A(159,330)  0.50(185,329)   /(199,330)  0.50(203,329)
Cred Att:(47,338)   3.00(73,338)    Cred Cmp:(109,338)  3.00(141,338)   GPA:(178,338)   4.500(195,338)
```

The above file generated need to be used to fill up the parameters values listed below as needed in order to get the data in structured format. Below are the lists of these parameters

- **headerXPosition** – This is an array containing a set of x-coordinates pointing to various data columns. For example the row marked below in red can be used to get a list of x-coordinate for all data columns.

```
Crs ID(51,177)  Course Title(87,177)    Mark(149,177)   Cred Att/Cmp(176,177)
Desert Mountain High School(81,187)
Yr:(50,197) 2008(64,197)    Month:(109,197) 12(131,197) Grade:(163,197) 09(190,197)
2318(55,207)    Wrld Hist/Wrld Geog(87,207) A(159,207)  0.50(185,207)   /(199,207)  0.50(203,207)
H(37,215)   3112(55,216)    English I Honors(87,215)    A(159,216)  0.50(185,215)   /(199,216)  0.50(203,215)
H(37,224)   4524(55,224)    Geometry/Trigonomet Honors(87,224)  A(159,224)  0.50(185,224)   /(199,224)  0.50(203,224)
H(37,232)   5412(55,233)    Biology I Honors(87,232)    A(159,233)  0.50(185,232)   /(199,233)  0.50(203,232)
6010(55,241)    Spanish I(87,241)   A(159,241)  0.50(185,241)   /(199,241)  0.50(203,241)
N(37,249)   7450(55,250)    Intro Physical Education(87,249)    A(159,250)  0.50(185,249)   /(199,250)  0.50(203,249)
Cred Att:(47,258)   3.00(73,258)    Cred Cmp:(109,258)  3.00(141,258)   GPA:(178,258)   4.500(195,258)
```

  Please note that the row selected for getting the x-coordinates value should have more data filled in. Like the row marked in blue above has lesser data as compared to row marked in red.

ADEPTIA

443 North Clark St, Suite 350
Chicago, IL 60610
Phone: (312) 229-1727

- **xDiff** – This refer to the difference in the x-coordinates for considering the data to belong to same column. For example if the data for a column is coming as shown below

```
Desert Mountain High School(81,347)
Yr:(50,357)     2009(64,357)    Month:(109,357) 12(131,357)     Grade:(163,357) 10(190,357)
H(37,366)       1160(55,367)    English IIH/AmAz Hist Pre Dipl(87,366)  A(159,367)      0.50(185,366)   /(199,367)      0.50(203,366)
H(37,375)       2160(55,375)    AM/AZ Hist Pre Diploma(91,375)  A(159,375)      0.50(185,375)   /(199,375)      0.50(203,375)
H(37,383)       4422(55,384)    Algebra II Honors(87,383)       A(159,384)      0.50(185,383)   /(199,384)      0.50(203,383)
H(37,392)       5512(55,392)    Chemistry I Honors(87,392)      A(159,392)      0.50(185,392)   /(199,392)      0.50(203,392)
6026(55,401)    Spanish II - Pre Diploma(87,400)        A(159,401)      0.50(185,400)   /(199,401)      0.50(203,400)
N(37,409)       6752(55,409)    Advanced Guitar(87,409) A(159,409)      0.50(185,409)   /(199,409)      0.50(203,409)
Cred Att:(47,418)       3.00(73,418)    Cred Cmp:(109,418)      3.00(141,418)   GPA:(178,418)   4.666(195,418)
```

  In order to consider the data marked above as same column the value of xdiff here will be 91 – 87 = 4. Please note that column value referred under "**headerXPosition**" should always be the greater one. So for the above case the value mentioned in "**headerXPosition**" will be 91 instead of 87.

- **yDiff** –This refer to the difference in the y-coordinates for considering the data to belong to same row. For example if the data for a row is coming as shown below

```
Desert Mountain High School(81,347)
Yr:(50,357)     2009(64,357)    Month:(109,357) 12(131,357)     Grade:(163,357) 10(190,357)
H(37,366)       1160(55,367)    English IIH/AmAz Hist Pre Dipl(87,366)  A(159,367)      0.50(185,366)   /(199,367)      0.50(203,366)
H(37,375)       2160(55,375)    AM/AZ Hist Pre Diploma(91,375)  A(159,375)      0.50(185,375)   /(199,375)      0.50(203,375)
H(37,383)       4422(55,384)    Algebra II Honors(87,383)       A(159,384)      0.50(185,383)   /(199,384)      0.50(203,383)
H(37,392)       5512(55,392)    Chemistry I Honors(87,392)      A(159,392)      0.50(185,392)   /(199,392)      0.50(203,392)
6026(55,401)    Spanish II - Pre Diploma(87,400)        A(159,401)      0.50(185,400)   /(199,401)      0.50(203,400)
N(37,409)       6752(55,409)    Advanced Guitar(87,409) A(159,409)      0.50(185,409)   /(199,409)      0.50(203,409)
Cred Att:(47,418)       3.00(73,418)    Cred Cmp:(109,418)      3.00(141,418)   GPA:(178,418)   4.666(195,418)
```

  In order to consider the data marked above as same row the ydiff here will be 384-383 = 1.

- **mergeRowWithyDiff** – This is Boolean flag that if enabled will merge two data rows whose y-coordinate is within range of yDiff. For example if the data is coming as shown below

```
Cred Att:(227,338)  3.00(253,338)   Cred Cmp:(289,338)  3.00(321,338)   GPA:(358,338)   5.000(375,338)
Cred Att:(227,375)  1.00(253,375)   Cred Cmp:(289,375)  1.00(321,375)   GPA:(358,375)   0.000(375,375)
5040(235,421)   IB English IV (Language A1 HL)(267,421) 0.00(383,421)
H(217,420)  0.50(365,421)   /(379,421)
H(217,429)  5050(235,430)   IB Extended Essay(267,430)  0.50(365,429)   /(379,429)  0.00(383,429)
5094(235,438)   IB HL Chemistry II- Year 2(267,438) 0.00(383,438)
H(217,437)  0.50(365,438)   /(379,438)
H(217,446)  5110(235,447)   IB 20th Cent Wld Hist (HL)(267,447) 0.50(365,446)   /(379,446)  0.00(383,446)
5220(235,455)   IB Mathematics II (HL)(267,455) 0.00(383,455)
H(217,454)  0.50(365,455)   /(379,455)
H(217,463)  5622(235,464)   AP Physics II(267,464)  0.50(365,463)   /(379,463)  0.00(383,463)
Cred Att:(227,472)  3.50(253,472)   Cred Cmp:(289,472)  GPA:(358,472)
```

  In this case if this flag is set then the above two rows will be merged into one row in the generated text file.

- **wrapping** –This is a Boolean flag which if enabled will merge rows containing multiline column data. For example consider an example as shown below

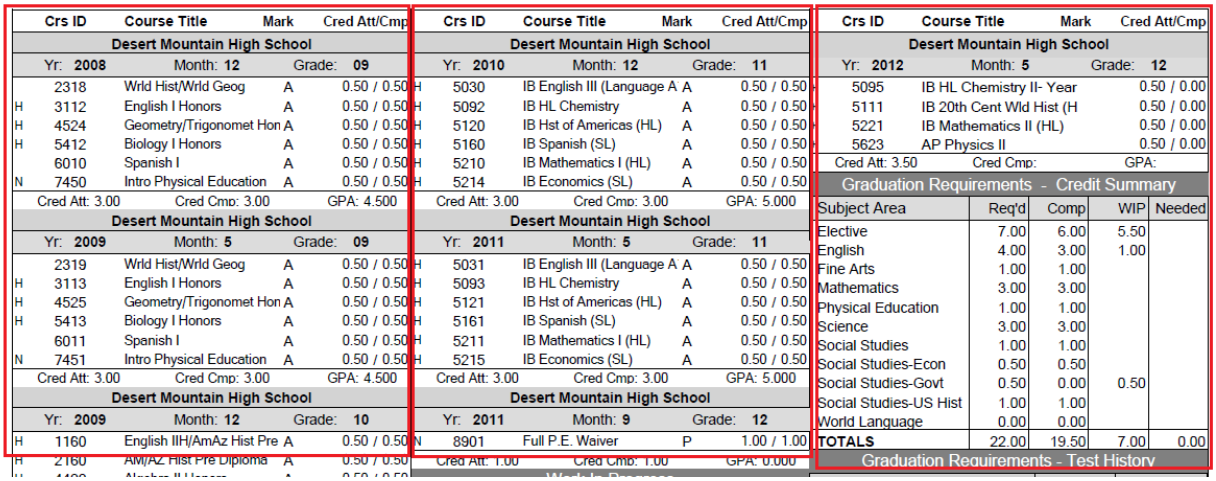
```
08-09  Huron High School - Grade 9(36,134)
English 9 Intensive(40,141) S1(108,141) A(127,141)  0.500(150,141)
English 9(40,150)   S2(186,150) A(204,150)  0.500(228,150)
History and(40,159)
Geography, World(40,168)
S1(108,159) A(127,159)  0.500(150,159)  S2(186,159) A(204,159)  0.500(228,159)
*(36,177)   Algebra II AC(40,177)   S1(108,177) C(127,177)  0.500(150,177)  S2(186,177) C(204,177)  0.500(228,177)
Biology(40,186) S1(108,186) A(127,186)  0.500(150,186)  S2(186,186) A(204,186)  0.500(228,186)
```

  The data marked here represents a single data row. On enabling this flag the above data will be merged into a single data row in the generated text file.

ADEPTIA

443 North Clark St, Suite 350
Chicago, IL 60610
Phone: (312) 229-1727

- **sorting** –This is a Boolean flag which if enabled will sort the data rows in ascending order of y-coordinates. Data from PDF usually are retrieved in the order in which they are written which is completely different from what it is displayed.
- **blockStartXposition** – This is an array containing a list of x-coordinates representing a starting position of a block. The data in the PDF can be divided into blocks as shown below



Crs ID	Course Title	Mark	Cred Att/Cmp
Desert Mountain High School			
Yr: 2008	Month: 12	Grade: 09	
2318	Wrld Hist/Wrld Geog	A	0.50 / 0.50
H 3112	English I Honors	A	0.50 / 0.50
H 4524	Geometry/Trigonomet Hon	A	0.50 / 0.50
H 5412	Biology I Honors	A	0.50 / 0.50
6010	Spanish I	A	0.50 / 0.50
N 7450	Intro Physical Education	A	0.50 / 0.50
Cred Att: 3.00	Cred Cmp: 3.00		GPA: 4.500
Desert Mountain High School			
Yr: 2009	Month: 5	Grade: 09	
2319	Wrld Hist/Wrld Geog	A	0.50 / 0.50
H 3113	English I Honors	A	0.50 / 0.50
H 4525	Geometry/Trigonomet Hon	A	0.50 / 0.50
H 5413	Biology I Honors	A	0.50 / 0.50
6011	Spanish I	A	0.50 / 0.50
N 7451	Intro Physical Education	A	0.50 / 0.50
Cred Att: 3.00	Cred Cmp: 3.00		GPA: 4.500
Desert Mountain High School			
Yr: 2009	Month: 12	Grade: 10	
H 1160	English IIH/AmAz Hist Pre	A	0.50 / 0.50
H 2160	AM/AZ Hist Pre Diploma	A	0.50 / 0.50

Crs ID	Course Title	Mark	Cred Att/Cmp
Desert Mountain High School			
Yr: 2010	Month: 12	Grade: 11	
H 5030	IB English III (Language A	A	0.50 / 0.50
H 5092	IB HL Chemistry	A	0.50 / 0.50
H 5120	IB Hst of Americas (HL)	A	0.50 / 0.50
H 5160	IB Spanish (SL)	A	0.50 / 0.50
H 5210	IB Mathematics I (HL)	A	0.50 / 0.50
H 5214	IB Economics (SL)	A	0.50 / 0.50
Cred Att: 3.00	Cred Cmp: 3.00		GPA: 5.000
Desert Mountain High School			
Yr: 2011	Month: 5	Grade: 11	
H 5031	IB English III (Language A	A	0.50 / 0.50
H 5093	IB HL Chemistry	A	0.50 / 0.50
H 5121	IB Hst of Americas (HL)	A	0.50 / 0.50
H 5161	IB Spanish (SL)	A	0.50 / 0.50
H 5211	IB Mathematics I (HL)	A	0.50 / 0.50
H 5215	IB Economics (SL)	A	0.50 / 0.50
Cred Att: 3.00	Cred Cmp: 3.00		GPA: 5.000
Desert Mountain High School			
Yr: 2011	Month: 9	Grade: 12	
N 8901	Full P.E. Waiver	P	1.00 / 1.00
Cred Att: 1.00	Cred Cmp: 1.00		GPA: 0.000

Crs ID	Course Title	Mark	Cred Att/Cmp
Desert Mountain High School			
Yr: 2012	Month: 5	Grade: 12	
5095	IB HL Chemistry II- Year		0.50 / 0.00
5111	IB 20th Cent Wld Hist (H		0.50 / 0.00
5221	IB Mathematics II (HL)		0.50 / 0.00
5623	AP Physics II		0.50 / 0.00
Cred Att: 3.50	Cred Cmp:		GPA:

Graduation Requirements - Credit Summary

Subject Area	Req'd	Comp	WIP	Needed
Elective	7.00	6.00	5.50	
English	4.00	3.00	1.00	
Fine Arts	1.00	1.00		
Mathematics	3.00	3.00		
Physical Education	1.00	1.00		
Science	3.00	3.00		
Social Studies	1.00	1.00		
Social Studies-Econ	0.50	0.50		
Social Studies-Govt	0.50	0.00	0.50	
Social Studies-US Hist	1.00	1.00		
World Language	0.00	0.00		
TOTALS	22.00	19.50	7.00	0.00

Graduation Requirements - Test History

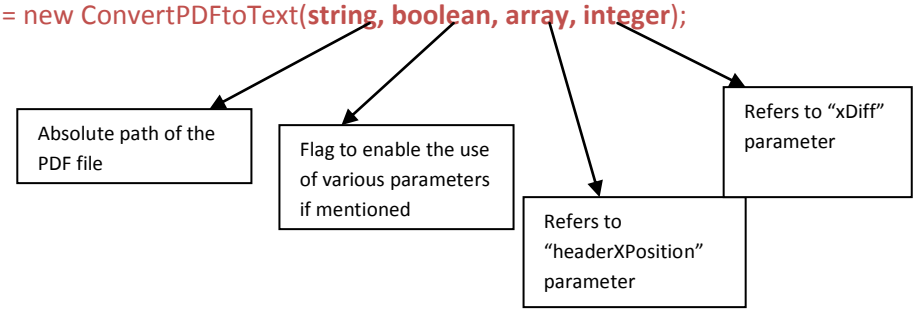
This helps in dividing the data rows into separate rows otherwise a single row can have different sets of data.

- **spaceDiff** - For PDF where each column data is separated by equal number of spaces this parameters can be used to structure the PDF data into a tab delimited file. The numeric value here refers to the fixed number of spaces that separates each data columns

Sample Codes

- Convert PDF to Text

```
ConvertPDFtoText obj = new ConvertPDFtoText(string, boolean, array, integer);
```



Example -

```
ConvertPDFtoText obj = new ConvertPDFtoText(pdfFilePath, true, new int[] { 37, 55, 87, 159, 199, 203, 217, 235, 267, 339, 365, 379, 383, 397, 415, 447, 519, 545, 559, 563 }, 4);
```

ADEPTIA

443 North Clark St, Suite 350
Chicago, IL 60610
Phone: (312) 229-1727

- Set yDiff

  { ConvertPDFtoText Class Instance}.setyDiff(integer);

  Example - obj.setyDiff(1);

- Enable wrapping

  { ConvertPDFtoText Class Instance}. setHandleWrapping(boolean);

  Example - obj.setHandleWrapping(true);

- Enable sorting

  { ConvertPDFtoText Class Instance}. setSortRecords(boolean);

  Example - obj.setSortRecords (true);

- Enable mergewithydiff

  { ConvertPDFtoText Class Instance}. setMergeRowWithyDiff (boolean);

  Example - obj.setMergeRowWithyDiff (true);

- Set blockStartXPosition

  { ConvertPDFtoText Class Instance}. setBlockStartXPositions (integer array);

  Example - obj.setBlockStartXPositions(new long[] { 37, 217, 397 });

- Set spaceDiff

  { ConvertPDFtoText Class Instance}. setSpaceDiff (integer);

  Example - obj. setSpaceDiff (4);

Note – "spaceDiff" parameter can be used even if the second Boolean parameter for the constructor of class "**ConvertPDFtoText**" is set false. This is not true for other parameters.