Re: PDF Python

From: **Sepehr Ekbatani TeIAS** I sepehrekbatani@teias.institute Tuesday, Sep 20, 2022 at 12:09 PM

To: **Mahdi Mir** I mrmahdimir@gmail.com

Mahdi Salaam,

Following our conversation, I am attaching the couple of files that I am struggling with. I am also attaching the warnings I get and the output I'm getting. Please let me know if the code doesn't run or it's not clear enough.
Thanks for your help.

Best,
Sepehr

```
Got stderr: Sep 20, 2022 11:57:39 AM org.apache.pdfbox.pdmodel.font.PDSimpleFont toUnicode
WARNING: No Unicode mapping for g166 (12) in font FHHCAL+BTitrBold
Sep 20, 2022 11:57:39 AM org.apache.pdfbox.pdmodel.font.PDSimpleFont toUnicode
WARNING: No Unicode mapping for g203 (13) in font FHHCAL+BTitrBold
Sep 20, 2022 11:57:39 AM org.apache.pdfbox.pdmodel.font.PDSimpleFont toUnicode
WARNING: No Unicode mapping for g123 (14) in font FHHCAL+BTitrBold
Sep 20, 2022 11:57:39 AM org.apache.pdfbox.pdmodel.font.PDSimpleFont toUnicode
WARNING: No Unicode mapping for g127 (4) in font FHHCAL+BTitrBold
Sep 20, 2022 11:57:39 AM org.apache.pdfbox.pdmodel.font.PDSimpleFont toUnicode
WARNING: No Unicode mapping for g185 (15) in font FHHCAL+BTitrBold
Sep 20, 2022 11:57:39 AM org.apache.pdfbox.pdmodel.font.PDSimpleFont toUnicode
WARNING: No Unicode mapping for g90 (10) in font FHHCAL+BTitrBold
Sep 20, 2022 11:57:39 AM org.apache.pdfbox.pdmodel.font.PDSimpleFont toUnicode
WARNING: No Unicode mapping for g191 (16) in font FHHCAL+BTitrBold
Sep 20, 2022 11:57:39 AM org.apache.pdfbox.pdmodel.font.PDSimpleFont toUnicode
```

	0	1	2	3	4	5	6	7	8	9	...	22	23	24	25	26	27	28	29	30	31
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	,=/	...	-$),	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	:	NaN	J	NaN	NaN	NaN	NaN	NaN	6839 :	NaN	...	NaN	NaN	NaN	NaN	= 130 = 78	$ #	NaN	7.0	NaN	$%
2	215 : 3	NaN	'()	NaN	!	NaN	NaN	NaN	898 : 2	'()	...	NaN	NaN	NaN	NaN	NaN	A++:	NaN	NaN	NaN	5
3	NaN	NaN	NaN	$%	NaN	NaN	!" !	NaN	NaN	NaN	...	!" !	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	!+,-	NaN	NaN	NaN	#$%	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	*	NaN	'()	&	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	2 783	NaN	NaN	11	19	NaN	7188	NaN	NaN	NaN	...	7834	NaN	NaN	&'	NaN	;- q >	NaN	NaN	NaN	1
7	1 812	NaN	NaN	17	18	NaN	6896	NaN	&'	"	...	7718	NaN	NaN	NaN	NaN	"	NaN	NaN	NaN	2
8	2 842	NaN	NaN	19	17	NaN	7269	NaN	NaN	(V)" +	...	6545	NaN	[	/	;	! "$ *	NaN	NaN	NaN	3
9	2 850	NaN	NaN	18	19	NaN	6623	NaN	")*	NaN	...	6704	NaN	Y	NaN	NaN	*	NaN	NaN	NaN	4
10	1 862	NaN	NaN	16	19	NaN	7280	NaN	&'	NaN	...	6542	NaN	r	NaN	(V)	= C =R	NaN	NaN	NaN	5

مهندسی برق(الکترونیک،بیوالکتریک،قدرت،کنترل،مخابرات، دیجیتال)--دانشگاه صنعتی شریف – تهران

تعداد قبول شدگان کانون ۶۴ نفر از ۱۲۰ نفر | **معادل ۵۳ درصد از کل ظرفیت** | **میانه ی تراز کانون : ۷۸۱۴** | **نوع رشته : روزانه**

بازه : A++ | **میانه ی رتبه در منطقه ی ۱ : ۴۳** | **میانه ی رتبه در منطقه ی ۲ : ۲۹**

ردیف	نام خانوادگی و نام	محل ثبت نام در کانون	میانگین تراز در کانون	معدل	تعداد آزمون	رتبه در منطقه *	سهمیه
۱	فلاحتگر معین	بهشهر	۸۳۱۲	۱۹	۱۹	۱	۲
۲	عباسی احسان	اهواز	۸۱۵۹	۱۹	۱۵	۲	۲
۳	فروزنده شهرکی رامین	اصفهان	۷۹۲۳	۱۹	۱۷	۲	۱
۴	موذنی سجاد	اصفهان	۸۱۹۷	۱۹	۱۸	۳	۱
۵	جعفرنیای جهرمی مهدی	جهرم	۷۷۹۱	۱۹	۱۴	۳	۲
۶	گنجی مهدی	اصفهان	۸۱۶۲	۱۹	۱۷	۴	۱
۷	اریامن خسرو	مراغه	۷۱۷۱	۱۹	۱۷	۶	سایر
۸	رییسی زاده مبارکه امیرحسین	اصفهان	۸۰۸۱	۱۹	۱۹	۷	۱

ردیف	نام خانوادگی و نام	محل ثبت نام در کانون	میانگین تراز در کانون	معدل	تعداد آزمون	رتبه در منطقه *	سهمیه
۲۷	حیدری شعیب	کرمانشاه	۸۰۱۵	۱۹	۱۵	۲۹	۲
۲۸	داوری نژاد احسان	همدان	۷۸۴۸	۱۸	۱۹	۲۹	۲
۲۹	موحدنسب محمد	یزد	۷۹۴۶	۱۹	۱۷	۳۱	۲
۳۰	طالبی شهریار	ملارد	۷۳۹۸	۱۹	۱۸	۳۲	۳
۳۱	اسمعیلی محدثه	ساوه	۷۹۹۶	۱۹	۱۹	۳۶	۲
۳۲	مهربان خمارتاش محمود	قوچان	۸۱۲۰	۱۹	۱۹	۳۷	۲
۳۳	احمدی محمد	قزوین	۷۳۴۳	۱۸	۱۳	۳۹	۲
۳۴	رسولی علی	مراغه	۷۷۷۴	۱۹	۱۷	۴۰	۲

From: **M Sepehr Ekbatani TeIAS** | sepehrekbatani@teias.institute — Tuesday, Aug 30, 2022 at 10:53 AM

Mahdi.
This is awesome! Works like a charm. Thank you for your help, it was a great deal.

Best,
Sepehr

> From: **M Mahdi Mir** | mrmahdimir@gmail.com — Monday, Aug 29, 2022 at 5:43 PM
>
> Salaaam,
>
> The attached Jupiter-notebook will do your job easily. Because these tables are in pdfs that are generated by word and their are not raw images taken from printed tables, they can be easily parsed.
>
> Open the attached notebook in the Google Colab and the notebook do the reset.
> If you want to run it locally you have to have java8+ besides python installed on your machine.
>
> I tested it, it read tables on each page but it needs some cleaning like fixing the table column headers and etc.

In [7]:
```
import tabula

pdf_path = "1393.pdf"

dfs = tabula.read_pdf(pdf_path, stream=True, pages = '5')
# read_pdf returns list of DataFrames
print(len(dfs))
dfs[0]
```

1

Out[7]:

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	مهندسي برق	Unnamed: 8	Unnamed: 9	Unnamed: 10
0	NaN	NaN	چند تا از 10 تا !؟	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	چارك پايين	چارك پايين	چارك پايين	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	چارك پايين	تعداد قبولي	NaN	NaN
3	NaN	NaN	NaN	NaN	ادبيات	رتبه قبولي	رتبه قبولي	رتبه قبولي هاي	ظرفيت	NaN	NaN	NaN
4	شيمي	فيزيك	معارف زبان رياضيات	عربي	NaN	NaN	NaN	NaN	تراز قبولي	هاي كانوني	گرايش-توضيحات	نام دانشگاه
5	NaN	NaN	NaN	NaN	فارسي	هاي كانون در	هاي كانون در	كانون در منطقه	رشته	NaN	NaN	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	هاي كانون	)نفر(	NaN	NaN
7	NaN	NaN	NaN	NaN	NaN	منطقه 3	منطقه 2	1	NaN	NaN	NaN	NaN
8	6	8	7 8 7	8	7	30	62	42	7668 122	150	NaN	صنعتي شريف
9	5	7	6 7 7	7	6	88	201	179	7009 99	120	NaN	دانشگاه تهران
10	5	7	5 7 7	7	6	179	344	449	7001 59	80	NaN	اميركبير
11	4	6	5 6 6	6	5	186	539	760	6541 79	100	NaN	علم و صنعت
12	3	5	4 7 6	6	5	818	NaN	2386	5908 14	25	الكترونيك	دانشگاه شيراز
13	4	6	4 5 6	6	5	320	920	1242	6333 100	120	NaN	خواجه نصير
14	4	5	3 6 5	5	5	NaN	1571	2087	6118 14	18	الكترونيك	دانشگاه تبريز
15	3	5	4 5 6	6	5	511	1193	1459	6162 76	100	NaN	شهيد بهشتي
16	4	5	4 5 6	5	5	569	1357	1362	6026 102	135	NaN	دانشگاه صنعتي اصفهان
17	4	5	4 5 6	5	5	677	1673	2040	6065 16	18	قدرت	دانشگاه تبريز
18	3	5	3 5 6	5	5	NaN	1607	NaN	6166 10	20	كنترل--محل تحصيل اهواز	دانشگاه صنعت نفت
19	3	5	3 4 4	4	4	NaN	NaN	3494	5798 14	20	شبانه	فردوسي مشهد
20	3	5	3 5 5	5	5	784	2468	2266	5816 76	120	NaN	فردوسي مشهد
21	3	5	3 4 6	5	4	898	2334	2735	5707 26	30	مخابرات	دانشگاه شيراز
22	3	5	3 4 6	5	5	NaN	2245	3284	5836 36	60	NaN	دانشگاه شاهد
23	3	5	3 4 6	5	5	623	1854	2631	6063 15	15	قدرت	دانشگاه اصفهان

Best,
Mahdi Mir

> From: **Sepehr Ekbatani TeIAS** I sepehrekbatani@teias.institute Saturday, Aug 27, 2022 at 6:05 PM
>
> Salaam,
> Here is a sample of what I need.
>
> Best,
> Sepehr