Tabula 1.2.1

	Pdf-file Table Extractor	
	Systeemvereisten:	Windows, Mac, Docker image, of Java Library
	Ontwikkeld door:	tabula.technology
Open Source, MIT License	Contactpersoon:	Koen.Vanderkimpen@Smals.be

Functionaliteiten

Ik denk dat ik niet de enige ben die af en toe een pdf-bestand heeft met een tabel erin, waarbij ik de gegevens van de tabel met zo weinig mogelijk gedoe in een Excel sheet wil krijgen. Tabula is een tool die daar kan bij helpen.

Fig. 1: Het welkomstscherm, klaar om een pdf te importeren

Deze kleine tool is een gratis opensource project met een permissieve licentie. We bespreken in deze QR de Windows versie (die lokaal op je pc een webserver start, waardoor je in de browser zal werken), maar er bestaan o.a. ook nog een command-line tool en een docker container voor de web-versie. Alles is gebaseerd op een herbruikbare Java-bibliotheek.

In tegenstelling tot sommige (veelal betalende) SaaS oplossingen, werkt Tabula enkel met tekst-gebaseerde pdf bestanden. OCR is dus niet ondersteund. Dit is echter minder en minder een probleem aan het worden: voor mij komen nu de meeste pdf's als tekst-pdf's binnen, via e-mail; het zijn geen ingescande documenten meer.

Het gebruik is eenvoudig: je importeert de te converteren pdf, je laat tabula zoeken waar er tabellen zijn die je eruit wil halen óf je duidt deze zelf aan, je klikt op de preview-knop om te zien wat het resultaat zal zijn, en daarna kan je nog kiezen op welke manier je de gegevens wil exporteren: er is keuze tussen CSV, TSV, en JSON formaten. Daarnaast kan je een template maken met de locaties van de tabellen, voor wanneer je vaak hetzelfde type document wil converteren.

Conclusies & Aanbevelingen

Tabula is een handige, eenvoudige tool met maar één doel: de tabellen uit een pdf bestand omzetten in bruikbare gegevens zonder gedoe met copy-paste. Hierin blinkt het dan ook uit.

Testen & Resultaten

Om de tool te testen maken we gebruik van een mij via e-mail toegestuurde Collect&Go rekening. We zullen proberen de boodschappen, met eenheidsprijs, aantallen en totalen hieruit te extraheren.

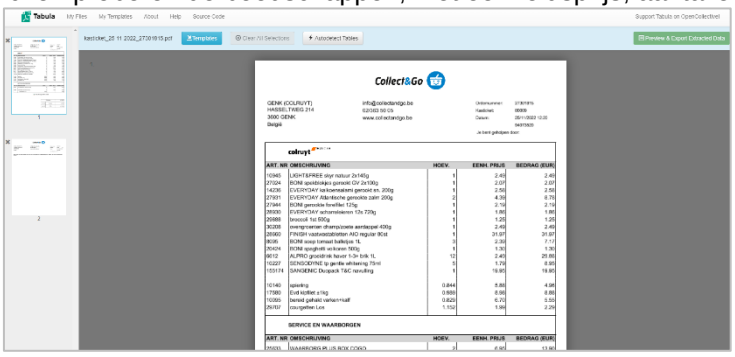


Fig. 2: Het pdf bestand, geïmporteerd

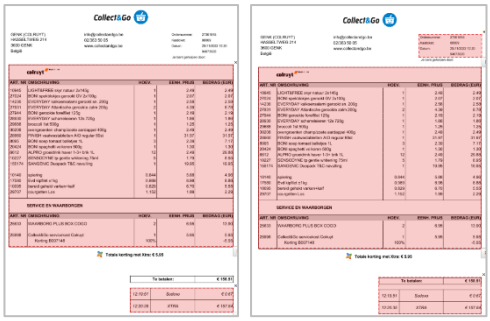
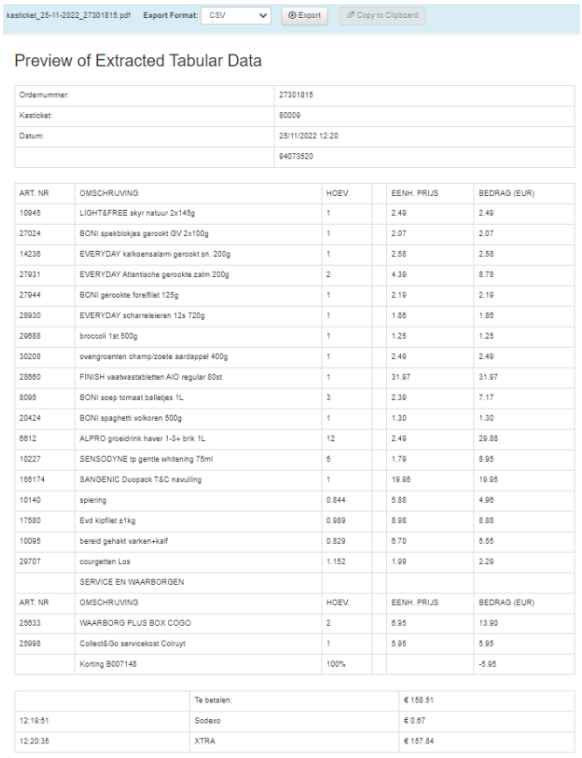


Fig. 3: autodetect en eigen selectie

Na het importeren (eigenlijk een upload naar de lokaal werkende webserver) zien we een preview van de pdf. Om te beginnen kunnen we op de “Autodetect Tables” knop klikken. Het resultaat zie je in figuur 3. We kiezen ervoor om de onderste tabel uit te breiden met het kleine stukje erboven, om ook het totaal mee te hebben, en we selecteren ook nog een extra tabel rond een deel van de hoofding (zo hebben we ook b.v. het kasticket en de datum). Onze uiteindelijke selectie is eveneens te zien in figuur 3.



Preview of Extracted Tabular Data

Ordernummer:	27301815
Kasticket:	80009
Datum:	25/11/2022 12:20
	94073520

ART. NR	OMSCHRIJVING	HOEV.	EENH. PRIJS	BEDRAG (EUR)
10945	LIGHT&FREE skyr natuur 2x145g	1	2.49	2.49
27024	BONI spekblokjes gerookt GV 2x100g	1	2.07	2.07
14236	EVERYDAY kalkoensalami gerookt sn. 200g	1	2.58	2.58
27931	EVERYDAY Atlantische gerookte zalm 200g	2	4.39	8.78
27944	BONI gerookte forelfilet 125g	1	2.19	2.19
28930	EVERYDAY scharreleieren 12s 720g	1	1.86	1.86
29688	broccoli 1st 500g	1	1.25	1.25
30208	ovengroenten champ/zoete aardappel 400g	1	2.49	2.49
28660	FINISH vaatwastabletten AIO regular 80st	1	31.97	31.97
8095	BONI soep tomaat balletjes 1L	3	2.39	7.17
20424	BONI spaghetti volkoren 500g	1	1.30	1.30
6612	ALPRO groeidrink haver 1-3+ brik 1L	12	2.49	29.88
10227	SENSODYNE tp gentle whitening 75ml	5	1.79	8.95
155174	SANGENIC Duopack T&C navulling	1	19.95	19.95
10140	spiering	0.844	5.88	4.96
17580	Evd kipfilet ±1kg	0.989	8.98	8.88
10095	bereid gehakt varken+kalf	0.829	6.70	5.55
29707	courgetten Los	1.152	1.99	2.29
	SERVICE EN WAARBORGEN			
ART. NR	OMSCHRIJVING	HOEV.	EENH. PRIJS	BEDRAG (EUR)
25633	WAARBORG PLUS BOX COGO	2	6.95	13.90
25998	Collect&Go servicekost Colruyt	1	5.95	5.95
	Korting B007148	100%		-5.95

	Te betalen:	€ 158.51
12:19:51	Sodexo	€ 0.67
12:20:36	XTRA	€ 157.84

Fig. 4: resultaat van extractie

Daarna kunnen we de “Preview” knop gebruiken en dan komen we in het volgende scherm terecht, waar we kunnen zien dat de tabellen netjes geëxtraheerd zijn (figuur 4). Hier kunnen we het formaat kiezen waarin we de gegevens willen exporteren, en wanneer we dan op de “export” knop klikken, wordt een bestand van dat formaat gedownload. Wij kiezen voor CSV (comma seperated value) en zullen deze dan importeren in Excel. Hierbij moet je er enkel op letten dat je aangeeft dat de gegevens met komma’s zijn gescheiden, en eventueel kan je ook nog aangeven dat decimalen worden aangegeven door een punt (wat het geval is in mijn rekening). Indien er komma’s in de tabel voorkomen, kunnen we beter opteren voor het TSV (tab separated value) formaat, hetgeen ook gemakkelijk is te importeren in Excel. Voor een geautomatiseerde verwerking van de gegevens is het JSON formaat allicht het beste.

Gebruiksvoorwaarden & Budget

Tabula is gratis te downloaden en gebruiken. Ook het gebruik ervan in je eigen software is nagenoeg zonder voorwaarden. Tabula wordt zelden onderhouden, maar werkt naar behoren.