Information Extraction from PDFs



Sarah Böning Researcher German Aerospace Center (DLR) – Institute of Data Science, Jena





Which data do l use ...



- PDF files
- Currently two use cases:
 - Scientific papers describing motors (~1k PDFs)
 - Datasheets of satellite parts (~ 600 PDFs)
- Challenges with PDFs:
 - Digital but not machine-understandable
 - Combination of structured (tables) & unstructured (text) data
 - Very different layouts, even in just 1 document class

		G	ilsu Chei	T. M. Jahny
			-in grant and	joine () mp. 1 in. mit.
			University of We trie Machines and For- Madama Wi	erenin - Madous w Electronics Consentions (WEMPEC) 1 51706 USA
			red design or chooses	II Monetres and Derma Sciences of a Vietna
10 MW Wind Turbina	Direct Drive Gener	tor	to in electric tablets	The prediction of vehicle perferences to very important
TO WW WING TURDING	Direct-Drive Genera	101	the volume's detring	determining the most efficient operating regions of an ele-
Design with Pitch or Ac	tive Speed Stall Cor	ntrol	onhighting of cloud-	of the potentian can be determined based as
			eters and muchles	presentationed time thresholds of motor mappe and up Operating memory is which shorter valuation have the host
H. Polinder, D. Barg R.P.J.O.M.	van Rooij A.S. McDonald, M	A. Maeller	optimize the succluster,	efficiency will be identified by means of vehicle performs
Electrical Proof Processing (OUWEND Wind Energy Rus	arch/DURIND Institute for Energy	n Sources	issue of such of the	stadyon, making it provides to design electric powerballs
Mikelweg 4, M/8 CD Delli Kitoverweg L	2629 HS Delli Kings Buildings, Ma	oficial Road	tener cherriedates	conditions in which the velocity operators must desparelly.
The Nationals The Nati	edads Elioborgh, EHV 32, U	inkal Kingdom	supplies in calculated	The basic sugarenting principles governing valuate tour
			interesting that the	effer are conserved in the published literature ().
Abasad-The objectives of this paper are to incodigate the	- The secole nomic performance in wirel up	been snote down	and the successful	vehicle driving up a dope and the tractive effort trancait
leadhlifty of a 18 NOV generator for a detect-drive wind technic and its constance the removator systems for each control and her	is limited by actively controlling the m	star spead. This	permittent magnet	to the ground through the days wheels.
active speed staff control. The idea behind the active speed staff	rand the amorphic militian in their an	to values above		The total tractore effect of a vehicle is illustrated in 3
control concept is to make a robust that is as simple as possible, and therefore users robust and sample for affiders wind	bathing so that not mare than taked power	in produced	and the second second	1 . 1 . 1 . 1 . 1
tarbles. This is along by removing the piloh control of the	It is not new to use wind tableter witho	at pitch control.	(270) duras mont	10-11-11-11-11
costrolling the pitch, but he controlling the roker speed: the other	constant useoff wird rathers with moury h	ingly helper 1.5	pisse of the sev	A CARDON
pred is so much reduced that the second-manic power is limited	MW with shall control, which recars that the	o blades con met	ill accustely predict	Carton
driv generator design is presented, holicating that such a	pich [1,2]. However, the sendymenc per	investor of the	ch opening point.	to the
preventor's loweline. It is shown that for a thorough realmation of	the black of these furthers is not optimum occur	ue ne design of	to provide a structure of	and a second
changes in the wind sport. However, a considerable increase in	officiency at word speech below mite	tonnal bas b	hers arrestgood in	a manufacture and a road
generator system cost is accounty to stable active speed shall control.	sendytanic dTaxing treading to a neighb	constant output	the design of Fid	FIND PAR VEDDA RANG COD (PILAD
	speed and control concurs is that he never	is compiled by	ectived very lasied	will at a subscript of the
Andre Dorsterviere internet erentere deren deren	active speed cantrol.			1
	It is also not new to use a distributive pe	stranged magnet	all a design approach	1 thomas manufactures
1 Istronetton	is a sial tabler will a deputier to	CARLENDER NY	e efficient spentag	10
The objectives of this paper set to investigate the feasibility of a 13 MW assessment for a disert-drive used radius and ra-	government. Howevers, it is now that the general	nor discussed in	thenance metrics of	Lord In the state
compare the generator overants for pitch control and for active	this paper is used to control the rotor speel	I and the pervet	deags process. In	= IN North March Mill Vision
speel stall circled.	In this paper, the consequences of the se-	the loop site	to composed through	
To investigate the fourbolity of a 10 MW databality	control design choices are incontgoted. Mast	the generator be	otel energy required.	this (w) and a second rank and the one and
multisic, or is it so mendality large and expension that it is not	over-dimensioned in order to be able to me	ske the required	inperiod to the EPA's	ringhouts 1200 scentred with groups, from up to believe
vide? To investigate this, both a medianical americation	a be ever-dracestored? Will the annual of	onengy yield he	tion VII.	10) jellandi met comprent subjellandi quanti
one as contrangence dougs an proposal. The idea behind the active around and control expected in the	influenced by desc dougs choices, and if y	ors, here made?	10000 C	
make a wind rathing that is an simple as possible, and	Then is easing down by comparing the discu-	days generated	1	
descion very robust and watable for efficient wind parks.	The contributions of this paper are the	A R PERSON A	1	
Therefore, a number of design chelces has been made.	tranhity andy of 30 MW detti-drive gener	mators and that it		
pack created.	compares the generator systems for prich-	cannol and for	HADDER	
- The blades are sheeped to measure accelerated	The paper is organized in follows. First, i	Dc wind turbing	Contraction of the	New conversion of the second second second
 A description remains interval assessments and 	characteristics and the properties of the two-	different control	which and Plannight South	marked on Daymondar 20,2020 of 17 10/30 1712 how \$757 liphon Teaching
	concepts are encounted. Next, the generator of	makes dougs it	1	
This work was supported in part to the European Community's Noth Ensemble December (1991) and a contrast on \$19847 (1998), as becaused	porturnance of the two different control	at another in	1	
Project cannot (17937-3), and in part to Sente Version on the constant months'	discussed Finally, some conclusions are draw		1	
and a second state and second			1	
			1	

L	SPECIFICATIONS		μSTAI	R Tra	cker						Jenaoptronil
a Defama Production Act little III.	 Lesses and examine database and 		APPLICATIONS						Tax & Pass	in the second second second	
L example the second seco	 Montal the Transition of a Statistic performance with value from units of a statistical mathematical statistics and a statistical statistics. 		 Sale Ma All GEO and L 	Factor and Me ECD Sedelliter	is Determinal Dista	oe.			Pass	20090-014	mangaris an every Recovers. 1.181
discussions course spectro-and second is an assured to by USS -	 Tracks with mean to believe or new Tage rate people has no College and with reduced 		SOFTWARK PE	TURKS					Bangling Typetron Energy Byrine Denome Countribut	Machine Rockwards Advice Tala 1974 - 1974 - Martine Tala	agentications, belowing, and the option of the of
ю нароль нись, внортаят, нерань. Чатоя	 Tool Wert22 serviced avoid a barrant SpecaWer option 		 She identified in the second se	Setternetic De	on Pymerici C me Connection	ode Allows for Hig In for Descel	h Accumey		Part of Your Sectors	MAN NO.	enale 45 O Millord ethenderen
right accuracy the tracker heritige spoked assemble CY 0220 related executional final/clining a	 Similar costs attracts and put and hull have been entranced to be over the base with the control of the base of the base of the base 		 Extension K Less Sension 	alman Filler Ive in Dourin	Produces All	ale and films	Coll runner.		Programme Broger	STARSBO	Art Different Render
4 is kind high reliais nor	 MOC provides efficiency statistical and TOC provides efficiency statistic detector meng paratol. 		CONTICUENTO	N OPTIONS					Annual Following	4.7. 199	
Oride Severencements/MDNJ pedicacellin the US accellually for	 Manhair and Salak Manhair and Anna Andrea and Anna Bonn and Anna Anna 		Cashes.	MOT	uble-1008	w10+-00044	10 Star 2001	1/30ar-400M	Canitri A ta	 Lance (Le) intercherdight Rance (Le) for english 	adabs.000.000.00
51	 Imputed LED plants to an The endor of an other in the advectment of 		HIN. Annany Dist	Rading Television	1062 5-22 asses	1.450 1.16 annua	< Large	1060 Elsonius	Augustar have	· Conception and a set of the second	dect and in all the second second second
Installey CI-XDD provides of operations fails automotions	dested work Main Dig Prove NW		Terring: From T	11.00	116	8.50% 12.50	11 Hz	- 16 m 5 m Hz	 Georgian Automotive	 E.S. anglosi, C.E.S. anglosif S.S. anglosi, C.E.S. anglosif S.S. anglosi, C.E.B. anglosif S.S. anglosi, C.E.B. anglosif 	Contraction of the second second line (1997) and the second line of the second line (1997) and the second second line of the se
wei, fri auförhömsen att fude mede, e sträte hest socurallet foltre	Novince option Novince 100 V power, -100 V -5 Cogeline		CHI Massing	(hingsted (hing)	62	10	13	2.1	 Sectory Sectors	Schellendersten der Mitte	encel these between
van nature an of-otti	Construction of the construction of the second	100	Cortes Alenhou	as for a valiabilit	v				 (haping) (Ken 25.00 period angle Sam 12.000	half and depending on white a gas are findle distant on which the
to any protocol operations where the tracket's anti-over one rith clouding systems	 Monte of education ME-MITTue of MME engineering to Complete set of decompositions and analysis evaluate with production 	and the second	Total lavibing Single Event	Done (110) Letchap (SE	11 - 28	0 and 300 km McPhropism	n (sprilon)		arterfectus Line	NU. 00110018 MIG17	aprice a select a fair, interview demonst
			Nuclears + 2x10 ¹² mine ²							NPC control of NPC co	approval advector for all relations and an exception of percept particular the last in the last of the last of the after malages an electronic
	SIMILARCOMPONENTS	Sec. S	The µSTAR" fe destruction to en	nhanes prover roure accurate	t, high-perfor c. roliable fan	iserces, rectable isonality = the	e iterdened is horsh spore	oporting previourcent.	 Press Conversion Press Scholler Interimentary	* ENLINE CENTRE DIF + LENLINE CONFERENCE	word the
n, find has delivered the highest- orrane, she taking available as	A Salar		The Processor	· spece co	mpoler is fligh	-proven, high	apeed, and he	14201	Advance land	CTU POINTAN BUSIC	excert the
c tarcol rear the CT-2020 to out log in attached, demette die U.S. runtet. 7 the CT-2020 is fourth quarter	I amon	14	issues and to per- selected with processing data p g 200 Miller with	commercial commercial rosewing with h SEU mitige	In the centre of	rings. E to a g I go 200 Mina I consilion	Name of Street, Party of Street, Party of Street, Stre	e tameet ampeler ter 500 MPLCPS	Telepite Telepite Tyronologia	NOLTE LANDY Desi Sandra Too Antorina Milate Desiritation (Mit Charlo Coll (1919) Filly in Material States with King State Prove Data (1917) Real States parameter test proceeding Mit 1
	Rectard Colling					-	SPACE	MICRO		Party 1000080 with 100 Med	
1927-6004 - Hes-Rokener - wordinateoren	MARKET Standard Local Advances (2011)	100 an 110 an	watana saw	ons Crange	NUTON NOTINE			470/2003	CERT		Colorest March 1998 Annual Annual State of Colorest Descent Name and March 2001 2021 State Free and March 2020 St Development of Colorest Data State Stat



Which methods do I apply ...



- Python 3¹ (incl. nltk², OpenCV³, Camelot⁴, pytesseract⁵)
- Configurable, universial pipeline, easy to adapt
- Simple CV-based table detection + OCR for image-based tables
- Separate processing of unstructured (text) & structured (tables) data
- Extracted data post-processed to remove artifacts
- Normalization of text to simplify extraction with NLP methods
- Domain knowledge-based information extraction (e.g. via ontology)
 1st use case: (key, value, unit)-tuples

preload files	extract table coordinates	extract text & tables	classify text	normali- zation	KVU extraction text	KVU extraction tables	merge KVU tuples
------------------	---------------------------------	-----------------------------	------------------	--------------------	---------------------------	-----------------------------	------------------------

https://www.python.org/
 https://www.nltk.org/
 https://pypi.org/project/opencv-python/
 https://pypi.org/project/camelot-py/
 https://pypi.org/project/pytesseract/
 All accessed 14.10.2021

Key	MatchedSynonym	Value	Unit	
Rotor Speed	Rotor Speed	10	rpm	
Number Of Pole Pairs	Number Of Pole Pairs	160	none	
Air Gap Thickness	air gap	10	mm	
Number Of Poles	Number Of Poles	4	none	
Air Gap Thickness	Air-gap	4	mm	
Rated Power	Rated Power	550	kW	
Rated Speed	nominal speed	4800	rpm	
Electromagnetic Efficiency	efficiency	89.8	%	
Specific Power	Specific Power	0,22224	kW/kg	
Number Of Pole Pairs	pole pairs	10	none	
Rated Torque	Rated Torque	120	Nm	
Rated Power	Rated Power	10	kW	
Rated Speed	Rated Speed	800	rpm	
Number Of Phase	Number Of Phase	3	none	
Number Of Poles	Number Of Poles	6	none	
Air Gap Thickness	Air-gap	60	mm	
Air Gap Thickness	Air-gap	1	mm	



What is the goal of my work ...



- Digitalization of documents
- Develop a universal method to extract any information from PDFs that can be easily adapted & specified for any domain
- Make PDFs machine-understandable & –processable
- Also comprehensible for humans
- Make information within PDFs accessible, findable, & easy-to-use for further automatic processing

