DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
10	**document**	**document**		container for document	root	TET
11	**document**	**configuration**		container for configuration	dict	TET
11	document	createdAt		created at (DD.MM.YY HH:MM:SS)	str	system
11	document	createdBy		created by module	str	system
11	**document**	**fonts**		container for fonts	list	TET
11	document	modifiedAt	x	last modified at (DD.MM.YY HH:MM:SS)	str	system
11	document	modifiedBy	x	last modified by module	str	system
11	document	noFonts		number fonts	int	TET
11	document	noLines		number lines in document	int	TET
11	document	noLinesFooter	x	number footer lines	int	LT
11	document	noLinesHeader	x	number header lines	int	LT
11	document	noLinesToc	x	number toc lines	int	LT
11	document	noListsBullet	x	number bulleted lists	int	LT
11	document	noListsNumber	x	number numbered lists	int	LT
11	document	noPages		number pages in document	int	TET
11	document	noParas		number paragraphs in document	int	TET
11	document	noSentences	x	number sentences in document	int	spaCy
11	document	noTables	x	number tables in document	int	LT
11	document	noWords		number words in document	int	TET
11	**document**	**pages**		container for pages	list	TET
11	**document**	**params**		container for parameters	dict	TET
20	**font**	**font**		container for font	root	TET
21	font	embedded		embedded	bool	TET
21	font	fontNo		font number in document	int	TET
21	font	fullName		font full name	str	TET
21	font	id		font identification	str	TET
21	font	italicAngle		font italic angle	float	TET
21	font	name		font name	str	TET
21	font	type		font type	str	TET
21	font	weight		font weight	float	TET
23	**params**	**params**		container for parameters	root	TET
24	params	directoryName		file directory for intermediate files and final result files	str	TET
24	params	documentId	x	document identification	int	TET
24	params	environmentVariant		environment variant	str	TET
24	params	fileNameCurr		name of input file	str	TET
24	params	fileNameNext		name of output file	str	TET
24	params	fileNameOrig		original file name	str	TET
24	params	ltFooterRequired		footer determination required	bool	TET
24	params	ltHeaderRequired		header determination required	bool	TET
24	params	ltHeadingRequired		heading determination required	bool	TET

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
24	params	`ltListBulletRequired`		bulleted list determination required	bool	TET
24	params	`ltListNumberRequired`		numbered list determination required	bool	TET
24	params	`ltTableRequired`		table determination required	bool	TET
24	params	`ltTocRequired`		TOC determination required	bool	TET
24	params	`noPdfPages`		number PDF pages	int	TET
25	**config**	**config**	x	container for configuration	root	TET
26	config	`jsonIndent`			int	TET
26	config	`jsonSortKeys`		Sort the keys in ascending order	bool	TET
26	config	`ltFooterMaxDistance`		maximum Levenshtein distance for a footer line	int	TET
26	config	`ltFooterMaxLines`		maximum number of footers	int	TET
26	config	`ltHeaderMaxDistance`		Maximum Levenshtein distance for a header line	int	TET
26	config	`ltHeaderMaxLines`		Maximum number of headers	int	TET
26	config	`ltHeadingFileInclNoCtx`		number of lines following the heading to be included as context into the JSON file	bool	TET
26	config	`ltHeadingFileInclRegexp`		number of lines following the heading to be included as context into the JSON file	bool	TET
26	config	`ltHeadingMaxLevel`		maximum level of the heading structure	int	TET
26	config	`ltHeadingMinPages`		minimum number of pages to determine the headings	int	TET
26	config	`ltHeadingRuleFile`		file with rules to determine the headings	bool	TET
26	config	`ltHeadingToleranceLlx`		tolerance of vertical indentation in percent	int	TET
26	config	`ltListBulletMinEntries`		minimum number of entries to determine a bulleted list	int	TET
26	config	`ltListBulletRuleFile`		file with rules to determine the bulleted lists	bool	TET
26	config	`ltListBulletToleranceLlx`		tolerance of vertical indentation in percent	bool	TET
26	config	`ltListNumberFileInclRegexp`		if it is set to true, the regular expression for the numbered list is included in the JSON file	bool	TET
26	config	`ltListNumberMinEntries`		minimum number of entries to determine a numbered list	bool	TET
26	config	`ltListNumberRuleFile`		file with rules to determine the numbered lists	bool	TET

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
26	config	ltListNumberToleranceLlx		tolerance of vertical indentation in percent	bool	TET
26	config	ltTableFileInclEmptyColumns		if it is set to true, the empty ells are included in the separate JSON file with the tables	bool	TET
26	config	ltTocLastPage		maximum number of pages for the search of the TOC (from the beginning)	bool	TET
26	config	ltTocMinEntries		minimum number of TOC entries	bool	TET
26	config	spacyIgnoreBracket		ignore the tokens which are brackets ?	bool	TET
26	config	spacyIgnoreLeftPunct		ignore the tokens which are left punctuation marks, e.g. "(" ?	bool	TET
26	config	spacyIgnoreLineTypeFooter		ignore the tokens from line type footer ?	bool	TET
26	config	spacyIgnoreLineTypeHeader		ignore the tokens from line type header ?	bool	TET
26	config	spacyIgnoreLineTypeHeading		ignore the tokens from line type heading ?	bool	TET
26	config	spacyIgnoreLineTypeListBullet		ignore the tokens from line type bulleted list ?	bool	TET
26	config	spacyIgnoreLineTypeListNumber		ignore the tokens from line type numbered list ?	bool	TET
26	config	spacyIgnoreLineTypeTable		ignore the tokens from line type table ?	bool	TET
26	config	spacyIgnoreLineTypeToc		ignore the tokens from line type TOC ?	bool	TET
26	config	spacyIgnorePunct		ignore the tokens which are punctuations ?	bool	TET
26	config	spacyIgnoreQuote		ignore the tokens which are quotation marks ?	bool	TET
26	config	spacyIgnoreRightPunct		ignore the tokens which are right punctuation marks, e.g. ")" ?	bool	TET
26	config	spacyIgnoreSpace		ignore the tokens which consist of whitespace characters ?	bool	TET
26	config	spacyIgnoreStop		ignore the tokens which are part of a "stop list" ?	bool	TET
26	config	spacyTknAttrCluster		brown cluster ID	bool	TET
26	config	spacyTknAttrDep_		syntactic dependency relation	bool	TET
26	config	spacyTknAttrDoc		the parent document	bool	TET
26	config	spacyTknAttrEntIob_		IOB code of named entity tag	bool	TET

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
26	config	spacyTknAttrEntKbId_		knowledge base ID that refers to the named entity this token is a part of, if any	bool	TET
26	config	spacyTknAttrEntType_		named entity type	bool	TET
26	config	spacyTknAttrHead		the syntactic parent, or "governor", of this token	bool	TET
26	config	spacyTknAttrI		the index of the token within the parent document	bool	TET
26	config	spacyTknAttrIdx		the character offset of the token within the parent document	bool	TET
26	config	spacyTknAttrIsAlpha		does the token consist of alphabetic characters?	bool	TET
26	config	spacyTknAttrIsAscii		does the token consist of ASCii characters?	bool	TET
26	config	spacyTknAttrIsBracket		is the token a bracket?	bool	TET
26	config	spacyTknAttrIsCurrency		is the token a currency symbol?	bool	TET
26	config	spacyTknAttrIsDigit		does the token consist of digits?	bool	TET
26	config	spacyTknAttrIsLeftPunct		is the token a left punctuation mark, e.g. "(" ?	bool	TET
26	config	spacyTknAttrIsLower		is the token in lowercase?	bool	TET
26	config	spacyTknAttrIsOov		is the token out-of-vocabulary?	bool	TET
26	config	spacyTknAttrIsPunct		is the token punctuation?	bool	TET
26	config	spacyTknAttrIsQuote		is the token a quotation mark?	bool	TET
26	config	spacyTknAttrIsRightPunct		is the token a right punctuation mark, e.g. ")" ?	bool	TET
26	config	spacyTknAttrIsSentEnd		does the token end a sentence?	bool	TET
26	config	spacyTknAttrIsSentStart		does the token start a sentence?	bool	TET
26	config	spacyTknAttrIsSpace		does the token consist of whitespace characters?	bool	TET
26	config	spacyTknAttrIsStop		is the token part of a "stop list"?	bool	TET
26	config	spacyTknAttrIsTitle		is the token in titlecase?	bool	TET
26	config	spacyTknAttrIsUpper		is the token in uppercase?	bool	TET
26	config	spacyTknAttrLang_		language of the parent document's vocabulary	bool	TET
26	config	spacyTknAttrLeftEdge		the leftmost token of this token's syntactic descendants	bool	TET
26	config	spacyTknAttrLemma_		base form of the token, with no inflectional suffixes	bool	TET

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
26	config	spacyTknAttrLex		the underlying lexeme	bool	TET
26	config	spacyTknAttrLexId		sequential ID of the token's lexical type, used to index into tables, e.g. for word vectors	bool	TET
26	config	spacyTknAttrLikeEmail		does the token resemble an email address?	bool	TET
26	config	spacyTknAttrLikeNum		does the token represent a number?	bool	TET
26	config	spacyTknAttrLikeUrl		does the token resemble a URL?	bool	TET
26	config	spacyTknAttrLower_		lowercase form of the token text	bool	TET
26	config	spacyTknAttrMorph		morphological analysis	bool	TET
26	config	spacyTknAttrNorm_		the token's norm, i.e. a normalized form of the token text	bool	TET
26	config	spacyTknAttrOrth_		verbatim text content	bool	TET
26	config	spacyTknAttrPos_		coarse-grained part-of-speech from the Universal POS tag set	bool	TET
26	config	spacyTknAttrPrefix_		a length-N substring from the start of the token	bool	TET
26	config	spacyTknAttrProb		smoothed log probability estimate of token's word type	bool	TET
26	config	spacyTknAttrRank		sequential ID of the token's lexical type, used to index into tables, e.g. for word vectors	bool	TET
26	config	spacyTknAttrRightEdge		the rightmost token of this token's syntactic descendants	bool	TET
26	config	spacyTknAttrSent		the sentence span that this token is a part of	bool	TET
26	config	spacyTknAttrSentiment		a scalar value indicating the positivity or negativity of the token	bool	TET
26	config	spacyTknAttrShape_		transform of the token's string to show orthographic features	bool	TET
26	config	spacyTknAttrSuffix_		length-N substring from the end of the token	bool	TET
26	config	spacyTknAttrTag_		fine-grained part-of-speech	bool	TET
26	config	spacyTknAttrTensor		the token's slice of the parent doc's tensor	bool	TET
26	config	spacyTknAttrText		verbatim text content	bool	TET
26	config	spacyTknAttrTextWithWs		text content, with trailing space character if present	bool	TET

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
26	config	`spacyTknAttrVocab`		the vocab object of the parent doc	bool	TET
26	config	`spacyTknAttrWhitespace_`		trailing space character if present	bool	TET
30	**page**	**page**		container for page	root	TET
31	page	`lineNoFirst`		number of first line in page	int	TET
31	page	`lineNoLast`		number of last line in page	int	TET
31	**page**	**lines**		container for lines	list	TET
31	page	`pageNo`		page number in document	int	TET
31	page	`paraNoFirst`		number of first paragraph in page	int	TET
31	page	`paraNoLast`		number of last paragraph in page	int	TET
31	**page**	**paras**		container for paragraphs	list	TET
31	page	`sentenceNoFirst`		number of first sentence in page	int	spaCy
31	page	`sentenceNoLast`		number of last sentence in page	int	spaCy
31	**page**	**sentences**		container for sentences	list	spaCy
31	page	`wordNoFirst`		number of first word in page	int	TET
31	page	`wordNoLast`		number of last word in page	int	TET
40	**line**	**line**		container for lines	root	TET
41	line	`lineNo`		line number in document	int	TET
41	line	`lineNoPage`		line number in page	int	TET
41	line	`lineNoPara`		line number in paragraph	int	TET
41	line	`llx`		x coordinate of the lower left corner	float	TET
41	line	`pageNo`		page number in document	int	TET
41	line	`paraNo`		paragraph number in document	int	TET
41	line	`sentenceNo`		sentence number in document	int	spaCy
41	line	`text`		line text	str	TET
41	line	`type`		line type	str	TET, TL
41	line	`urx`		x coordinate of the upper right corner	float	TET
41	line	`wordNoFirst`		number word of first word in line	int	TET
41	line	`wordNoLast`		number word of last word in line	int	TET
50	**para**	**para**		container for paragraph	root	TET
51	para	`lineNoFirst`		number of first line in paragraph	int	TET
51	para	`lineNoLast`		number of last line in paragraph	int	TET
51	para	`pageNo`		page number in document	int	TET
51	para	`paraNo`		paragraph number in document	int	TET
51	para	`paraNoPage`		paragraph number in page	int	TET

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
51	para	sentenceNoFirst		number of first sentence in paragraph	int	spaCy
51	para	sentenceNoLast		number of last sentence in paragraph	int	spaCy
51	para	text		paragraph text	str	TET, spaCy
51	para	wordNoFirst		number of first word in paragraph	int	TET
51	para	wordNoLast		number of last word in paragraph	int	TET
51	**para**	**words**		container for words	list	TET
60	**sentence**	**sentence**		container for sentences	root	TET
61	sentence	sentenceNo		sentence number in document	int	TET
61	sentence	sentenceNoPage		sentence number in page	int	TET
61	sentence	sentenceNoPara		sentence number in paragraph	int	TET
61	sentence	text		sentence text	str	TET
61	sentence	wordNoFirst		number word of first word in sentence	int	TET
61	sentence	wordNoLast		number word of last word in sentence	int	TET
70	**word**	**word**		container for word	root	TET
71	word	font		font identification	str	TET
71	word	lineNo		line number in document	int	TET
71	word	pageNo		page number in document	int	TET
71	word	paraNo		paragraph number in document	int	TET
71	word	sentenceNo		sentence number in document	int	spaCy
71	word	size		font size	float	TET
71	word	spacyTknAttrCluster		brown cluster ID	str	spaCy
71	word	spacyTknAttrDep_		syntactic dependency relation	str	spaCy
71	word	spacyTknAttrDoc		the parent document	str	spaCy
71	word	spacyTknAttrEntIob_		IOB code of named entity tag	str	spaCy
71	word	spacyTknAttrEntKbId_		knowledge base ID that refers to the named entity this token is a part of, if any	str	spaCy
71	word	spacyTknAttrEntType_		named entity type	str	spaCy
71	word	spacyTknAttrHead		the syntactic parent, or "governor", of this token	str	spaCy
71	word	spacyTknAttrI		the index of the token within the parent document	str	spaCy
71	word	spacyTknAttrIdx		the character offset of the token within the parent document	str	spaCy

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
71	word	`spacyTknAttrIsAlpha`		does the token consist of alphabetic characters?	bool	spaCy
71	word	`spacyTknAttrIsAscii`		does the token consist of ASCii characters?	bool	spaCy
71	word	`spacyTknAttrIsBracket`		is the token a bracket?	bool	spaCy
71	word	`spacyTknAttrIsCurrency`		is the token a currency symbol?	bool	spaCy
71	word	`spacyTknAttrIsDigit`		does the token consist of digits?	bool	spaCy
71	word	`spacyTknAttrIsLeftPunct`		is the token a left punctuation mark, e.g. "(" ?	bool	spaCy
71	word	`spacyTknAttrIsLower`		is the token in lowercase?	bool	spaCy
71	word	`spacyTknAttrIsOov`		is the token out-of-vocabulary?	bool	spaCy
71	word	`spacyTknAttrIsPunct`		is the token punctuation?	bool	spaCy
71	word	`spacyTknAttrIsQuote`		is the token a quotation mark?	bool	spaCy
71	word	`spacyTknAttrIsRightPunct`		is the token a right punctuation mark, e.g. ")" ?	bool	spaCy
71	word	`spacyTknAttrIsSentEnd`		does the token end a sentence?	bool	spaCy
71	word	`spacyTknAttrIsSentStart`		does the token start a sentence?	bool	spaCy
71	word	`spacyTknAttrIsSpace`		does the token consist of whitespace characters?	bool	spaCy
71	word	`spacyTknAttrIsStop`		is the token part of a "stop list"?	bool	spaCy
71	word	`spacyTknAttrIsTitle`		is the token in titlecase?	bool	spaCy
71	word	`spacyTknAttrIsUpper`		is the token in uppercase?	bool	spaCy
71	word	`spacyTknAttrLang_`		language of the parent document's vocabulary	str	spaCy
71	word	`spacyTknAttrLeftEdge`		the leftmost token of this token's syntactic descendants	str	spaCy
71	word	`spacyTknAttrLemma_`		base form of the token, with no inflectional suffixes	str	spaCy
71	word	`spacyTknAttrLex`		the underlying lexeme	str	spaCy
71	word	`spacyTknAttrLexId`		sequential ID of the token's lexical type, used to index into tables, e.g. for word vectors	str	spaCy
71	word	`spacyTknAttrLikeEmail`		does the token resemble an email address?	bool	spaCy
71	word	`spacyTknAttrLikeNum`		does the token represent a number?	bool	spaCy
71	word	`spacyTknAttrLikeUrl`		does the token resemble a URL?	bool	spaCy

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
71	word	spacyTknAttrLower_		lowercase form of the token text	str	spaCy
71	word	spacyTknAttrMorph		morphological analysis	str	spaCy
71	word	spacyTknAttrNorm_		the token's norm, i.e. a normalized form of the token text	str	spaCy
71	word	spacyTknAttrOrth_		verbatim text content	str	spaCy
71	word	spacyTknAttrPos_		coarse-grained part-of-speech from the Universal POS tag set	str	spaCy
71	word	spacyTknAttrPrefix_		a length-N substring from the start of the token	str	spaCy
71	word	spacyTknAttrProb		smoothed log probability estimate of token's word type	str	spaCy
71	word	spacyTknAttrRank		sequential ID of the token's lexical type, used to index into tables, e.g. for word vectors	str	spaCy
71	word	spacyTknAttrRightEdge		the rightmost token of this token's syntactic descendants	str	spaCy
71	word	spacyTknAttrSent		the sentence span that this token is a part of	str	spaCy
71	word	spacyTknAttrSentiment		a scalar value indicating the positivity or negativity of the token	str	spaCy
71	word	spacyTknAttrShape_		transform of the token's string to show orthographic features	str	spaCy
71	word	spacyTknAttrSuffix_		length-N substring from the end of the token	str	spaCy
71	word	spacyTknAttrTag_		fine-grained part-of-speech	str	spaCy
71	word	spacyTknAttrTensor		the token's slice of the parent doc's tensor	str	spaCy
71	word	spacyTknAttrText		verbatim text content	str	spaCy
71	word	spacyTknAttrTextWithWs		text content, with trailing space character if present	str	spaCy
71	word	spacyTknAttrVocab		the vocab object of the parent doc	str	spaCy
71	word	spacyTknAttrWhitespace_		trailing space character if present	str	spaCy
71	word	text		word text	str	TET
71	word	type		line type	str	TET, TL
71	word	wordNo		word number in document	int	TET
71	word	wordNoLine		word number in line	int	TET
71	word	wordNoPage		word number in page	int	TET
71	word	wordNoPara		word number in paragraph	int	TET
100	**heading**	**toc**	x	container for toc	root	LT

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
101	**heading**	**entries**		container for heading entries	list	LT
110	**entry**	**entry**		container for heading entry	root	LT
111	entry	ctxtLine1		1. context line	str	LT
111	entry	ctxtLine2		2. context line	str	LT
111	entry	ctxtLine3		3. context line	str	LT
111	entry	level		heading level	int	LT
111	entry	lineNoPage		line number in page	int	LT
111	entry	pageNo		page number in document	int	LT
111	entry	regexp	x	regular expression	str	LT
111	entry	text		heading text	str	LT
200	**lists bullet**	**listsBullet**	x	container for bulleted lists	root	LT
201	lists bullet	**lists**		container for bulleted lists	list	LT
210	**list**	**list**		container for list	root	LT
211	**list**	**entries**		container for list entries	list	LT
211	list	format		number format	str	LT
211	list	listNo		list number in document	int	LT
211	list	noEntries		number entries in list	int	LT
211	list	pageNoFirst		page number first entry	int	LT
211	list	pageNoLast		page number last entry	int	LT
220	**entry**	**entry**		container for bulleted list entry	root	LT
221	entry	entryNo		entry number in list	int	LT
221	entry	lineNoPageFirst		first line in page	int	LT
221	entry	lineNoPageLast		last line in page	int	LT
221	entry	pageNo		page number in document	int	LT
221	entry	paraNo		paragraph number in page	int	LT
221	entry	text		list entry text	str	LT
300	**lists number**	**listsNumber**	x	container for numbered lists	root	LT
301	lists number	**lists**		container for numbered lists	list	LT
310	**list**	**list**		container for list	root	LT
311	list	format		bullet format	str	LT
311	list	listNo		list number in document	int	LT
311	list	noEntries		number entries in list	int	LT
311	list	pageNoFirst		page number first entry	int	LT
311	list	pageNoLast		page number last entry	int	LT
311	list	regexp	x	regular expression	str	LT
319	**list**	**entries**		container for list entries	list	LT
320	**entry**	**entry**		container for numbered list entry	root	LT
321	entry	entryNo		entry number in list	int	LT
321	entry	lineNoPageFirst		first line in page	int	LT
321	entry	lineNoPageLast		last line in page	int	LT
321	entry	pageNo		page number in document	int	LT
321	entry	paraNo		paragraph number in page	int	LT
321	entry	text		list entry text	str	LT
400	**tables**	**tables**	x	container for document tables	root	LT

DCR-CORE Processing Results

lvl	container	tag	opt	description	type	source
401	**tables**	**tables**		container for document tables	list	LT
410	**table**	**table**		container for document table	root	LT
411	table	llxFirstRow		first row: x coordinate of the lower left corner	float	LT
411	table	noColumns		number columns in table	int	LT
411	table	noRows		number rows in table	int	LT
411	table	pageNoFirst		page number first row	int	LT
411	table	pageNoLast		page number last row	int	LT
411	table	tableNo		table number in document	int	LT
411	table	urxFirstRow		first row: x coordinate of the upper right corner	float	LT
419	**table**	**rows**		container for table rows	list	LT
420	**row**	**row**		container for table rows	root	LT
421	row	llxFirstColumn		first column: x coordinate of the lower left corner	float	LT
421	row	rowNo		number row in table	int	LT
421	row	urxFirstColumn		first column: x coordinate of the upper right corner	float	LT
429	**row**	**columns**		container for table columns	list	LT
430	**column**	**column**		container for table column	root	LT
431	column	columnNo		number column in row	int	LT
431	column	lineNoPage		line number in page	int	LT
431	column	llx		x coordinate of the lower left corner	float	LT
431	column	paraNo		paragraph number in page	int	LT
431	column	text		column text	str	LT
431	column	urx		x coordinate of the upper right corner	float	LT