DataFrame of <span, entity type> pairs

Entity spans in document context

DataFrame of raw token lables

Data entry pane

CoNLL_View_Doc.ipynb

	span	ent_type
	35]: 'SAN MARINO'	LOC
	46]: 'GRAND PRIX'	MISC
	[63, 68]: 'IMOLA'	LOC
	[70, 75]: 'Italy'	LOC
	[129, 139]: 'San Marino'	LOC
5	[159, 169]: 'Grand Prix'	MISC
6	[174, 188]: 'Michael Doohan'	PER
7	[190, 199]: 'Australia'	LOC
8	[201, 206]: 'Honda'	ORG
9	[228, 245]: 'Jean-Michel Bayle'	PER
10	[247, 253]: 'France'	LOC
11	[255, 261]: 'Yamaha'	ORG
12	[274, 286]: 'Norifumi Abe'	PER
13	[288, 293]: 'Japan'	LOC
14	[295, 301]: 'Yamaha'	ORG
15	[314, 327]: 'Luca Cadalora'	PER
16	[329, 334]: 'Italy'	LOC
17	[336, 341]: 'Honda'	ORG
18	[354, 367]: 'Alex Criville'	PER
19	[369, 374]: 'Spain'	LOC
20	[376, 381]: 'Honda'	ORG
21	[394, 407]: 'Scott Russell'	PER
22	[409, 422]: 'United States'	LOC

35   [590, 595]: 'Honda'

[7]: doc_df["span"].values.repr_   offsets = False
      doc_df["span"].values

[7]:
-DOCSTART-
MOTORCYCLING- SAN MARINO GRAND PRIX PRACTICE
TIMES.
IMOLA, Italy 1996-08-30
Practice times set on Friday
for Sunday's San Marino 500cc motorcycling Grand Prix:
1. Michael Doohan (Australia) Honda one minute 50.250
2. Jean-Michel Bayle (France) Yamaha 1:50.727
3. Norifumi Abe (Japan) Yamaha 1:50.858
4. Luca Cadalora (Italy) Honda 1:51.006
5. Alex Criville (Spain) Honda 1:51.075
6. Scott Russell (United States) Suzuki 1:51.287
7. Tadayuki Okada (Japan) Honda 1:51.528
8. Carlos Checa (Spain) Honda 1:51.588
9. Alexandre Barros (Brazil) Honda 1:51.784
10. Shinichi Itoh (Japan) Honda 1:51.857

[8]: # Dataframe of tokens for finding offsets
      toks_df = corpus_raw[fold][doc_offset]
      toks_df

	span	ent_iob	ent_type	sentence	line_
0	[0, 10]: '-DOCSTART-'	O	None	[0, 10]: '-DOCSTART-'	216
1	[11, 23]: 'MOTORCYCLING'	O	None	[11, 62]: 'MOTORCYCLING-SAN MARINO GRAND PRIX...'	216

# Dataframe of tokens for finding offsets
toks_df = corpus_raw[fold][doc_offset]
toks_df

	span	ent_iob	ent_type	sentence	
0	[0, 10]: '-DOCSTART-'	O	None	[0, 10]: '-DOCSTART-'	
1	[11, 23]: 'MOTORCYCLING'	O	None	[11, 62]: 'MOTORCYCLING- SAN MARINO GRAND PRIX...'	219104
2	[23, 24]: '-'	O	None	[11, 62]: 'MOTORCYCLING- SAN MARINO GRAND PRIX...'	
3	[25, 28]: 'SAN'	B	LOC	[11, 62]: 'MOTORCYCLING- SAN MARINO GRAND PRIX...'	219104
4	[29, 35]: 'MARINO'	I	LOC	[11, 62]: 'MOTORCYCLING- SAN MARINO GRAND PRIX...'	219105
5	[36, 41]: 'GRAND'	B	MISC	[11, 62]: 'MOTORCYCLING- SAN MARINO GRAND PRIX...'	219106
6	[42, 46]: 'PRIX'	I	MISC	[11, 62]: 'MOTORCYCLING- SAN MARINO GRAND PRIX...'	219107
7	[47, 55]: 'PRACTICE'	O	None	[11, 62]: 'MOTORCYCLING- SAN MARINO GRAND PRIX...'	219108
8	[56, 61]: 'TIMES'	O	None	[11, 62]: 'MOTORCYCLING- SAN MARINO GRAND PRIX...'	219109
9	[61, 62]: '.'	O	None	[11, 62]: 'MOTORCYCLING- SAN MARINO GRAND PRIX...'	219110
10	[63, 68]: 'IMOLA'	B	LOC	[63, 86]: 'IMOLA, Italy 1996-08-30'	219112
11	[68, 69]: ','	O	None	[63, 86]: 'IMOLA, Italy 1996-08-30'	219113
12	[70, 75]: 'Italy'	B	LOC	[63, 86]: 'IMOLA, Italy 1996-08-30'	219114
13	[76, 86]: '1996-08-30'	O	None	[63, 86]: 'IMOLA, Italy 1996-08-30'	219115
14	[87, 95]: 'Practice'	O	None	[87, 115]: 'Practice times set on Friday'	219117

CoNLL_3_in_gold.csv

Delimiter:

	num_models	fold	doc_offset	corpus_span	corpus_ent_type	error_type	correct_span	correct_ent_type	notes	time_started	time_stopped	tir
1	0	dev	0	[20, 34]: LEICESTER...	ORG							
2	0	dev	0	[2225, 2235]: 'ex-Engl...	MISC							
3	0	dev	2	[25, 30]: 'ASHES'	MISC							
4	0	dev	2	[87, 92]: 'Ashes'	MISC							
5	0	dev	2	[614, 634]: 'Duke of N...	ORG							
6	0	dev	6	[19, 24]: 'ROTOR'	ORG							

Simple   0   4   Python 3 | Idle   Mem: 4.11 GB   Saving completed   Mode: Command   Ln 1, Col 1   CoNLL_View_Doc.ipynb