Nomes: Douglas Costa, Jader Fróes, Patrick Guimarães

CRF para NER

Extração:

Os datasets Tweets e Leis, foram tratados para se adequarem de acordo com o código base, onde possuía a estrutura de colunas como:

	Sentence #	Word	POS	Tag
0	Sentence: 0	@xx	JJ	NOUN
1	Sentence: 0	tem	NN	VERB

Sendo assim possível a fácil aplicação sobre o modelo base.

Para a obtenção do dataset do twitter, foi utilizada a API Tweepy. Sendo necessário para acesso aos dados do Twitter uma chave específica obtida por usuários desenvolvedores, cadastrados no https://developer.twitter.com.

Já para o dataset das leis foi unificada as palavras em suas respectivas frases para tratamento, onde que cada palavra da frase estava na primeira coluna de cada linha.

Features e Resultados:

Twitter:

Para a tokenização deste tipo de dado, foi pensado em utilizar o *TweetTokenizer* do nltk o qual mantém palavras com o @ e o # não separando-os. Mas para a utilização das Features como as apresentadas a seguir a tokenização normal do nltk (*word_tokenizer*) fazia mais sentido.
Para tentar obter melhorias no treinamento do Twitter foram criadas duas features, Arroba(@) e hashtag(#), as quais analisavam a presença de um arroba e hashtag na palavra presente, passada e futura. Não foram obtidos melhores resultados na aplicação delas.

f1-score: `0.8923532657893002`

	precision	recall	f1-score	support
ADJ	0.76	0.57	0.65	61
ADP	0.99	0.96	0.98	288
ADV	0.95	0.89	0.92	297
AUX	1.00	1.00	1.00	1
CCONJ	1.00	1.00	1.00	5
DET	0.99	0.97	0.98	213
NOUN	0.76	0.87	0.81	900
NUM	0.96	0.96	0.96	46
PRON	0.99	0.99	0.99	144
PROPN	0.66	0.64	0.65	207
PUNCT	0.99	0.96	0.98	404
SCONJ	1.00	0.97	0.99	78
VERB	0.90	0.85	0.87	852
X	1.00	1.00	1.00	7
accuracy			0.88	3503
macro avg	0.93	0.90	0.91	3503
weighted avg	0.88	0.88	0.88	3503

Leis:

Foram adicionadas 2 novas features: Traço, que representa o caractere ‘-’ que é muito presente no início de várias frases. Dois pontos, que possibilitaria identificar, características como substantivos após ele.

Nenhuma das duas features gerou grandes melhorias na acurácia

f1-score = 0.9840735179497669

	precision	recall	f1-score	support
ADJ	0.98	0.96	0.97	1448
ADP	1.00	1.00	1.00	6743
ADV	0.99	0.98	0.99	2973
AUX	1.00	0.93	0.96	14
CCONJ	1.00	1.00	1.00	26
DET	1.00	1.00	1.00	2122
NOUN	0.97	0.98	0.98	11606
NUM	0.95	0.95	0.95	1128
PRON	1.00	1.00	1.00	553
PROPN	0.96	0.93	0.95	2693
PUNCT	1.00	1.00	1.00	8189
SCONJ	1.00	1.00	1.00	1103
VERB	0.98	0.98	0.98	6367
X	1.00	0.96	0.98	27
accuracy			0.98	44992

```
   macro avg       0.99      0.98      0.98     44992
weighted avg       0.98      0.98      0.98     44992
```

Link para o código: https://github.com/patrickguima/PLN