Capitulo_3

Flavio Codeço Coelho
(Dated: September 10, 2019)

CONTENTS

I. IDENTIFICANDO ENTIDADES NOMEADAS

Neste capítulo vamos treinar um identificador de Entidades nomeadas usando a biblioteca Spacy. A partir deste capítulo vamos importar também funções que já criamos anteriormente, e que encontram-se reproduzidas em dhbbmining.py

In [1]:
```
import os, glob
import spacy
from spacy import displacy
from dhbbmining import *
```

Para utilizar o spacy em um corpus na lingua portuguesa, vamos primeiro importar o modelo liguístico do português

In [2]:
```
nlp = spacy.load("pt_core_news_sm")
```

In [3]:
```
caminho = "../../dhbb/text/*.text"
arquivos  = glob.glob(caminho)
```

In [4]:
```
dhbb = tabula_verbete(arquivos)
dhbb.head()
```

[4]:

	arquivo	title	natureza	sexo	\
0	1.text	COELHO, Machado	biográfico	m	
1	10.text	ABÍLIO, Armando	biográfico	m	
2	100.text	ALEIXO, Pedro	biográfico	m	
3	1000.text	CAMPOS, Eduardo	biográfico	m	
4	1001.text	CAMPOS, Eleazar Soares	biográfico	m	

	cargos	\
0	\n - dep. fed. DF 1927-1929 \n - dep. fed. DF ...	
1	\n - dep. fed. PB 1995-1999\n - dep. fed. PB ...	
2	\n - const. 1934\n - dep. fed. MG 1935-1937\n ...	
3	\n - dep. fed. PE 1995\n - dep. fed. PE 1998-...	
4	\n - magistrado\n - interv. MA 1945-1946\n	

	corpo
0	\n\nńJosé Machado Coelho de Castroż nasceu em ...
1	\n\nńArmando Abílio Vieiraż nasceu em Itaporan...
2	\n\nńPedro Aleixoż nasceu em São Caetano, dist...
3	\n\nńEduardo Henrique Accioly Camposż nasceu e...
4	\n\nńEleazar Soares Camposż nasceu em São Luís...

In [6]:
```
doc = nlp(dhbb.corpo[0])
type(doc)
```

[6]: spacy.tokens.doc.Doc

In [7]:
```
print([w.text for w in doc])
```

```
['\n\n', 'ń', 'José', 'Machado', 'Coelho', 'de', 'Castro', 'ż', 'nasceu', 'em',
'Lorena', '(', 'SP', ')', '.', '\n\n', 'Estudou', 'no', 'Ginásio', 'Diocesano', 'de',
'São', 'Paulo', 'e', 'bacharelou-se', 'em', '1910', 'pela', '\n', 'Faculdade', 'de',
'Ciências', 'Jurídicas', 'e', 'Sociais', '.', 'Dedicando-se', 'à', 'advocacia', ',',
'foi', '\n', 'promotor', 'público', 'em', 'Cunha', '(', 'SP', ')', 'e', 'depois',
'delegado', 'de', 'polícia', 'no', 'Rio', 'de', '\n', 'Janeiro', ',', 'então',
'Distrito', 'Federal', '.', '\n\n', 'Iniciou', 'sua', 'vida', 'política', 'como',
'deputado', 'federal', 'pelo', 'Distrito', 'Federal', ',', '\n', 'exercendo', 'o',
'mandato', 'de', '1927', 'a', '1929', '.', 'Reeleito', 'para', 'a', 'legislatura',
'iniciada', '\n', 'em', 'maio', 'de', '1930', ',', 'ocupava', 'sua', 'cadeira', 'na',
'Câmara', 'quando', ',', 'em', '3', 'de', 'outubro', ',', '\n', 'foi', 'deflagrado',
'o', 'movimento', 'revolucionário', 'liderado', 'por', 'Getúlio', 'Vargas', '.', '\n',
'Ligado', 'a', 'o', 'governo', 'federal', ',', 'encontrava-se', 'a', 'o', 'lado',
'do', 'presidente', '\n', 'Washington', 'Luís', ',', 'no', 'palácio', 'Guanabara',
',', 'no', 'momento', 'de', 'sua', 'deposição', 'no', '\n', 'dia', '24', 'de',
'outubro', '.', 'Junto', 'com', 'outros', 'companheiros', 'também', 'solidários', 'a',
'o', '\n', 'regime', 'deposto', 'e', 'que', 'se', 'haviam', 'asilado', 'em',
'embaixadas', 'e', 'legações', ',', 'foi', '\n', 'enviado', 'em', 'novembro', 'para',
'o', 'estrangeiro', '.', 'Em', 'outubro', 'de', '1932', ',', 'estava', '\n',
'presente', 'no', 'porto', 'de', 'Alcântara', ',', 'em', 'Lisboa', ',', 'para',
'receber', 'os', '\n', 'revolucionários', 'constitucionalistas', 'exilados', 'pelo',
'governo', 'de', 'Getúlio', '\n', 'Vargas', 'após', 'a', 'derrota', 'da', 'revolução',
'irrompida', 'em', 'julho', 'd', 'esse', 'ano', 'em', 'São', '\n', 'Paulo', '.',
'\n\n', 'Com', 'a', 'redemocratização', 'do', 'país', 'em', '1945', ',', 'candidatou-
se', 'pelo', 'estado', 'de', 'São', '\n', 'Paulo', ',', 'na', 'legenda', 'do',
'Partido', 'Social', 'Democrático', '(', 'PSD', ')', ',', 'à', 's', 'eleições',
'para', '\n', 'a', 'Assembléia', 'Nacional', 'Constituinte', '(', 'ANC', ')',
'realizadas', 'em', 'dezembro', 'd', 'esse', '\n', 'ano', '.', 'Obteve', 'uma',
'suplência', 'e', ',', 'em', 'julho', 'de', '1946', ',', 'foi', 'convocado', 'para',
'\n', 'participar', 'dos', 'trabalhos', 'constituintes', '.', 'Com', 'a',
'promulgação', 'da', 'nova', 'Carta', '\n', '(', '18/9/1946', ')', 'e', 'a',
'transformação', 'da', 'Constituinte', 'em', 'Congresso', 'ordinário', ',', '\n',
'integrou', 'a', 'Comissão', 'Permanente', 'de', 'Obras', 'Públicas', 'da', 'Câmara',
'Federal', ',', '\n', 'tendo', 'votado', 'em', 'janeiro', 'de', '1948', 'a', 'favor',
'da', 'cassação', 'dos', 'mandatos', 'dos', '\n', 'parlamentares', 'comunistas', '.',
'Deixou', 'a', 'Câmara', 'em', 'janeiro', 'de', '1951', '.', '\n\n', 'Foi', 'ainda',
'presidente', 'da', 'Companhia', 'de', 'Cimento', 'Vale', 'do', 'Paraíba', '.',
'\n\n', 'Faleceu', 'no', 'Rio', 'de', 'Janeiro', 'no', 'dia', '17', 'de', 'maio',
'de', '1975', '.', '\n\n']
```

In [ ]: