Processamento de Linguagem Natural

Flávio Codeço Coelho
FGV-EMAp
(Dated: September 6, 2019)

CONTENTS

I. CONTANDO PALAVRAS

Neste capítulo iremos começar a interagir com os textos no nível da linguagem, por meio das ferramentas do Processamento de Linguagem Natural (PLN). Vamos progredir gradualmente nossa representação da linguagem a partir da morfologia, passando pela sintaxe e chegando à semântica.

Nesta etapa, faremos uso de biliotecas especializadas em PLN como o NLTK e a Spacy.

```
In [11]: import nltk
         import spacy
         import pandas as pd
```

Bibliotecas de PLN requerem o carregamento de modelos de liguagem para funcionar de maneira apropriada: para este capitulo iremos carregar os modelos específicos da lingua portuguesa. Para isso precisamos executar comandos no terminal do sistema operacional:

```
In [2]: !python3 -m spacy download pt
```

```
Requirement already satisfied: pt_core_news_sm==2.1.0 from
https://github.com/explosion/spacy-models/releases/download/pt_core_news_sm-2.1.0/pt_c
ore_news_sm-2.1.0.tar.gz#egg=pt_core_news_sm==2.1.0 in /usr/local/lib/python3.6/dist-
packages (2.1.0)
 Download and installation successful
You can now load the model via spacy.load('pt_core_news_sm')
 Couldn't link model to 'pt'
Creating a symlink in spacy/data failed. Make sure you have the required
permissions and try re-running the command as admin, or use a virtualenv. You
can still import the model as a module and call its load() method, or create the
symlink manually.
/usr/local/lib/python3.6/dist-packages/pt_core_news_sm -->
/usr/local/lib/python3.6/dist-packages/spacy/data/pt
 Download successful but linking failed
Creating a shortcut link for 'pt' didn't work (maybe you don't have admin
permissions?), but you can still load the model via its full package name: nlp =
spacy.load('pt_core_news_sm')
```

```
In [4]: nltk.download()
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```

[4]:
```
True
```

Para contar as palavras de um texto, é preciso primeiro separá-las uma a uma. A este processo dá-se o nome de tokenização, e é tipicamente fácil de fazer mal-feito e difícil de fazer bem-feito.

Primeiramente precisaremos carregar os verbetes com a função que criamos no capítulo 1.

In [15]:
```
import glob, os
caminho = "../../dhbb/text/*.text"
arquivos  = glob.glob(caminho)

def tabula_verbete(n=None):
    """
    Carrega todos os verbetes disponíveis, ou os primeiros n.
    n:  número de verbetes a tabular
    """
    if n is None:
        n = len(arquivos)
    linhas = []
    for a in arquivos[:n]:
        with open (a, 'r') as f:
            verbete = f.read()
        cabeçalho = verbete.split('---')[1]
        campos = {l.split(':')[0].strip() :l.split(':')[1].strip() for l in
cabeçalho.split('\n')[:4] if l}
        campos['arquivo'] = os.path.split(a)[1]
        campos['cargos'] = cabeçalho.split('cargos:')[1]
        campos['corpo'] = verbete.split('---')[2]
        linhas.append(campos)
    tabela = pd.DataFrame(data = linhas, columns=['arquivo','title',␣
↪'natureza', 'sexo',
'cargos', 'corpo'])
    return tabela
```

I.1. Tokenização com o nltk

In [16]:
```
tabela = tabula_verbete(n=10)
```

In [19]:
```
palavras = nltk.word_tokenize(tabela.corpo[0])
palavras[:10]
```

[19]:
```
['ń',
 'José',
 'Machado',
 'Coelho',
 'de',
 'Castro',
 'ż',
 'nasceu',
 'em',
 'Lorena',
 '(',
 'SP',
 ')',
 '.',
 'Estudou',
```

```
'no',
'Ginásio',
'Diocesano',
'de',
'São',
'Paulo',
'e',
'bacharelou-se',
'em',
'1910',
'pela',
'Faculdade',
'de',
'Ciências',
'Jurídicas',
'e',
'Sociais',
'.',
'Dedicando-se',
'à',
'advocacia',
',',
'foi',
'promotor',
'público',
'em',
'Cunha',
'(',
'SP',
')',
'e',
'depois',
'delegado',
'de',
'polícia',
'no',
'Rio',
'de',
'Janeiro',
',',
'então',
'Distrito',
'Federal',
'.',
'Iniciou',
'sua',
'vida',
'política',
'como',
'deputado',
'federal',
'pelo',
'Distrito',
'Federal',
',',
'exercendo',
```

```
'o',
'mandato',
'de',
'1927',
'a',
'1929',
'.',
'Reeleito',
'para',
'a',
'legislatura',
'iniciada',
'em',
'maio',
'de',
'1930',
',',
'ocupava',
'sua',
'cadeira',
'na',
'Câmara',
'quando',
',',
'em',
'3',
'de',
'outubro',
',',
'foi',
'deflagrado',
'o',
'movimento',
'revolucionário',
'liderado',
'por',
'Getúlio',
'Vargas',
'.',
'Ligado',
'ao',
'governo',
'federal',
',',
'encontrava-se',
'ao',
'lado',
'do',
'presidente',
'Washington',
'Luís',
',',
'no',
'palácio',
'Guanabara',
',',
```

```
'no',
'momento',
'de',
'sua',
'deposição',
'no',
'dia',
'24',
'de',
'outubro',
'.',
'Junto',
'com',
'outros',
'companheiros',
'também',
'solidários',
'ao',
'regime',
'deposto',
'e',
'que',
'se',
'haviam',
'asilado',
'em',
'embaixadas',
'e',
'legações',
',',
'foi',
'enviado',
'em',
'novembro',
'para',
'o',
'estrangeiro',
'.',
'Em',
'outubro',
'de',
'1932',
',',
'estava',
'presente',
'no',
'porto',
'de',
'Alcântara',
',',
'em',
'Lisboa',
',',
'para',
'receber',
'os',
```

```
'revolucionários',
'constitucionalistas',
'exilados',
'pelo',
'governo',
'de',
'Getúlio',
'Vargas',
'após',
'a',
'derrota',
'da',
'revolução',
'irrompida',
'em',
'julho',
'desse',
'ano',
'em',
'São',
'Paulo',
'.',
'Com',
'a',
'redemocratização',
'do',
'país',
'em',
'1945',
',',
'candidatou-se',
'pelo',
'estado',
'de',
'São',
'Paulo',
',',
'na',
'legenda',
'do',
'Partido',
'Social',
'Democrático',
'(',
'PSD',
')',
',',
'às',
'eleições',
'para',
'a',
'Assembléia',
'Nacional',
'Constituinte',
'(',
'ANC',
```

```
')',
'realizadas',
'em',
'dezembro',
'desse',
'ano',
'.',
'Obteve',
'uma',
'suplência',
'e',
',',
'em',
'julho',
'de',
'1946',
',',
'foi',
'convocado',
'para',
'participar',
'dos',
'trabalhos',
'constituintes',
'.',
'Com',
'a',
'promulgação',
'da',
'nova',
'Carta',
'(',
'18/9/1946',
')',
'e',
'a',
'transformação',
'da',
'Constituinte',
'em',
'Congresso',
'ordinário',
',',
'integrou',
'a',
'Comissão',
'Permanente',
'de',
'Obras',
'Públicas',
'da',
'Câmara',
'Federal',
',',
'tendo',
'votado',
```

```
 'em',
 'janeiro',
 'de',
 '1948',
 'a',
 'favor',
 'da',
 'cassação',
 'dos',
 'mandatos',
 'dos',
 'parlamentares',
 'comunistas',
 '.',
 'Deixou',
 'a',
 'Câmara',
 'em',
 'janeiro',
 'de',
 '1951',
 '.',
 'Foi',
 'ainda',
 'presidente',
 'da',
 'Companhia',
 'de',
 'Cimento',
 'Vale',
 'do',
 'Paraíba',
 '.',
 'Faleceu',
 'no',
 'Rio',
 'de',
 'Janeiro',
 'no',
 'dia',
 '17',
 'de',
 'maio',
 'de',
 '1975',
 '.']
```

In [ ]: