Interagindo com Arquivos de Texto

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

CONTENTS

I. ABRINDO ARQUIVOS DE TEXTO

Neste curso de mineração de textos usaremos como material principal de trabalho, os verbetes do Dicionário Histórico e Biográfico do Brasil – DHBB. Estes verbetes stão disponíveis para Download público.

Neste capítulo vamos aprender a interagir com os verbetes no disco e extrair informações simples a partir dos mesmos.

Vamos começar importando alguma bibliotecas que nos serão úteis nesta tarefa:

```
In [2]: import os
        import glob
```

Assumindo que os dados do DHBB já foram baixados para um diretório local, podemos começar inspecionando o diretório e listando o seu conteúdo.

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

```
[3]: 7687
```

Temos 7687 verbetes neste diretório. Vamos agora ver como abrir um destes verbetes e inspecionar o seu conteúdo:

```
In [4]: arquivos[0]
```

```
[4]: '../../dhbb/text/1.text'
```

```
In [5]: with open(arquivos[0],'r') as f:
            verbete = f.read()
            print(verbete)
```

```
---
title: COELHO, Machado
natureza: biográfico
sexo: m
```

```
cargos:
 - dep. fed. DF 1927-1929
 - dep. fed. DF 1930
 - const. 1946
 - dep. fed. SP 1946-1951
---

ńJosé Machado Coelho de Castrożnasceu 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.
```

A variável **verbete** que criamos na célula anterior é uma variável do tipo `string`, que é a otipo usado pelo Python para representar um bloco de texto. Podemos manipular o texto dentro de uma `string` de diversas maneiras:

```
In [6]: print(verbete.split('---')[1])
```

```
title: COELHO, Machado
natureza: biográfico
sexo: m
cargos:
 - dep. fed. DF 1927-1929
```

```
 - dep. fed. DF 1930
 - const. 1946
 - dep. fed. SP 1946-1951
```

Tipos de dados em Python, também conhecidos como objetos, possuem métodos. O método `split` do tipo `string` segmenta uma string nas posicões em que ocorram uma sequência específica de caracteres, retornando um outro tipo de dado, denominado `lista`.

In [7]:
```
type(verbete.split('---'))
```

[7]:
```
list
```

Listas são sequencias de objetos de quaisquer tipos que também apresenta seu conjunto de métodos. Para descobrir os métodos de qualquer objeto, basta colocar um ponto após o nome da variável e pressionar a tecla `<tab>`. Listas são delimitadas por colchetes: `[]` (lista vazia).

In [8]:
```
l = verbete.split('---')
l
```

[8]:
```
['',
 '\ntitle: COELHO, Machado\nnatureza: biográfico\nsexo: m\ncargos:\n - dep. fed.
DF 1927-1929 \n - dep. fed. DF 1930\n - const. 1946\n - dep. fed. SP
1946-1951\n',
 '\n\nńJosé Machado Coelho de Castroż nasceu em Lorena (SP).\n\nEstudou no
Ginásio Diocesano de São Paulo e bacharelou-se em 1910 pela\nFaculdade de
Ciências Jurídicas e Sociais. Dedicando-se à advocacia, foi\npromotor público em
Cunha (SP) e depois delegado de polícia no Rio de\nJaneiro, então Distrito
Federal.\n\nIniciou sua vida política como deputado federal pelo Distrito
Federal,\nexercendo o mandato de 1927 a 1929. Reeleito para a legislatura
iniciada\nem maio de 1930, ocupava sua cadeira na Câmara quando, em 3 de
outubro,\nfoi deflagrado o movimento revolucionário liderado por Getúlio
Vargas.\nLigado ao governo federal, encontrava-se ao lado do
presidente\nWashington Luís, no palácio Guanabara, no momento de sua deposição
no\ndia 24 de outubro. Junto com outros companheiros também solidários
ao\nregime deposto e que se haviam asilado em embaixadas e legações,
foi\nenviado em novembro para o estrangeiro. Em outubro de 1932,
estava\npresente no porto de Alcântara, em Lisboa, para receber
os\nrevolucionários constitucionalistas exilados pelo governo de Getúlio\nVargas
após a derrota da revolução irrompida em julho desse ano em São\nPaulo.\n\nCom a
redemocratização do país em 1945, candidatou-se pelo estado de São\nPaulo, na
legenda do Partido Social Democrático (PSD), às eleições para\na Assembléia
Nacional Constituinte (ANC) realizadas em dezembro desse\nano. Obteve uma
suplência e, em julho de 1946, foi convocado para\nparticipar dos trabalhos
constituintes. Com a promulgação da nova Carta\n(18/9/1946) e a transformação da
Constituinte em Congresso ordinário,\nintegrou a Comissão Permanente de Obras
Públicas da Câmara Federal,\ntendo votado em janeiro de 1948 a favor da cassação
dos mandatos dos\nparlamentares comunistas. Deixou a Câmara em janeiro de
1951.\n\nFoi ainda presidente da Companhia de Cimento Vale do
Paraíba.\n\nFaleceu no Rio de Janeiro no dia 17 de maio de 1975.\n\n']
```

Note que nas strings acima existem várias ocorrências da sequencia de caracteres `'\n'`. Esta sequência identifica quebra de linhas. Podemos então utilizá-la para dividir o cabeçalho do verbete em uma lista de linhas:

In [9]:
```
cabeçalho = verbete.split('---')[1]
cabeçalho.split('\n')
```

[9]:
```
['',
 'title: COELHO, Machado',
 'natureza: biográfico',
 'sexo: m',
 'cargos:',
 ' - dep. fed. DF 1927-1929 ',
 ' - dep. fed. DF 1930',
 ' - const. 1946',
 ' - dep. fed. SP 1946-1951',
 '']
```

Elementos de uma lista podem ser acessado por sua posição na sequência, por exemplo:

In [10]:
```
print(l[2])
```

```
ńJosé Machado Coelho de Castroż 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 [11]:
```
campos = {l.split(':')[0].strip() :l.split(':')[1].strip() for l in
          cabeçalho.split('\n')[:4] if l}
          campos
```

[11]:
```
{'title': 'COELHO, Machado', 'natureza': 'biográfico', 'sexo': 'm'}
```

Na célula acima contruímos um novo tipo de variável chamada *Dicionário*, é basicamente um conjunto de pares, delimitado por `{}`. Estes pares são chamados pares `chave: valor`.

I.1. Abrindo um grande número de documentos texto

Como vimos acima existem 7687 verbetes à nossa disposição no disco, mas não podemos abrir todos ao mesmo tempo pois, em primeiro lugar podem não caber na memória, em segundo lugar raramente precisaremos inpecioná-los todos ao mesmo tempo. O mais comum é analisá-los em sequência. Vamos inspecionar os primeiros 10:

In [12]:
```
for a in arquivos[:10]:
            with open (a, 'r') as f:
                verbete = f.readlines()
            print('Verbete: ', a.split('.text')[0].split('/')[-1])
            print(verbete[1])
```

```
Verbete:  1
title: COELHO, Machado

Verbete:  10
title: ABÍLIO, Armando

Verbete:  100
title: ALEIXO, Pedro

Verbete:  1000
title: CAMPOS, Eduardo

Verbete:  1001
title: CAMPOS, Eleazar Soares

Verbete:  1002
title: CAMPOS, Epílogo de

Verbete:  1003
title: CAMPOS, França

Verbete:  1004
title: CAMPOS, Francisco Machado de

Verbete:  1005
title: CAMPOS, Francisco

Verbete:  1006
title: CAMPOS, Frederico
```

In [13]:
```
arquivos[1]
```

[13]:
```
'../../dhbb/text/10.text'
```

Acima utilizamos uma estrutura de repetição, denominada “laço for” para abrir sequencialmente os arquivos. É importante notar que a cada volta do laço, o arquivo texo é atribuído à mesma variável, o que

significa que nunca há mais do que apenas um verbete na memória. Desta forma poderíamos potencialmente analisar todos os milhares de verbetes ocupando apenas uma quantidade pequena e constante de memória. Outro detalhe do código acima é que, para facilitar a extração do título do verbete, Fizemos a leitura do arquivo com o método `readlines` que retorna o verbete já divido em uma lista de linhas ao invés de uma `string`.

II. EXTRAINDO INFORMAÇÃO ESTRUTURADA

Agora que sabemos como abrir arquivos de texto e ler o seu conteúdo, podemos experimentar a extração de imformações específicas dos verbetes e organizá-la em uma tabela. Para isso vamos lançar mão de uma biblioteca chamada **Pandas** para organizar em uma estrutura tabular, chamada `DataFrame` os dados que vamos extrair.

```
In [14]: import pandas as pd
         pd.set_option("display.latex.repr", True)
```

Nós vimos acima que os verbetes contém uma seção inicial delimitada pelos caracteres `---` vamos utilizar esta característica do texto para guiar nossa extração de informação. Como você pode perceber, já começamos a reutilizar código que escrevemos anteriormente. Para facilitar o reuso e reduzir a necessidade de escrever múltiplas vezes o mesmo código vamos aprender a organizá-lo melhor. Vamos começar definindo uma função.

```
In [28]: 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'])
             return tabela
```

A função acima inclui a maior parte do código que escrevemos anteriormente, só que encapsulado em uma função que nos permite executar a extração e tabulação do cabeçalho para o numero de verbetes que desejarmos. Podemos ver abaixo que na verdade é muito rápido processar todos os verbetes.

```
In [29]: help(tabula_verbete)

Help on function tabula_verbete in module __main__:

tabula_verbete(n=None)
    Carrega todos os verbetes disponíveis, ou os primeiros n.
    :param n:  número de verbetes a tabular
```

```
In [23]: tab = tabula_verbete()
         tab.head()
```

[23]:

	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

Podemos visualizar uma descrição básica da tabela resultante

```
In [24]: tab.describe()
```

[24]:

	arquivo	title	natureza	sexo
count	7687	7687	7687	6722
unique	7687	7596	2	2
top	1541.text	MACHADO, José	biográfico	m
freq	1	3	6724	6517

Por exemplo fica fácil ver que no DHBB predominam biografias de personagens do sexo masculino.

```
In [27]: print(tab.sexo.value_counts())
```

```
m    6517
f     205
Name: sexo, dtype: int64
```

Percebemos também que a natureza predominante dos verbetes é biográfica e que só existem duas naturezas, mas qua a outra?

```
In [30]: print(tab.natureza.value_counts())
```

```
biográfico    6724
temático       963
Name: natureza, dtype: int64
```

III. EXPORTANDO PARA BANCOS DE DADOS

Depois de realizarmos a nossa análise e tabular os resultados, podemos exportar a tabela em vários formatos. Em primeiro lugar, caso queiramos abri nossa trabalho em uma planilha, devemos salvar no formato CSV, ou "comma-separated-values". Este formato pode ser aberto imediatamente em uma planilha.

```
In [31]: tab.to_csv("minha_tabela.csv", sep='|')
```

Acima usamos o caracter "|" como separador para evitar confusões com as virgulas existentes no texto. ## Exportando para um banco de dados relacional Para exportar para um banco relacional, precisamos de uma biblioteca adicional, o SQLAlchemy. Estabiblioteca nos permite interagir com a maioria dos banco relacionais. Aqui vamos usar o banco SQLite.

```
In [32]: from sqlalchemy import create_engine
```

```
In [34]: engine = create_engine('sqlite:///minha_tabela.sqlite', echo=False)
         tab.to_sql('resultados', con=engine, if_exists='append')
```

Uma vez inserido no banco relacional, podemos fazer consultas aos dados usando a linguagem SQL. Abaixo obtemos o resultado da consulta em uma lista.

```
In [39]: engine.execute("select * from resultados where natureza='temático'").
 ↪fetchall()[:10]
```

[39]:
```
[(354, '10989.text', 'Destacamento de Operações e Informações  Centro de
Operações e Defesa Interna (DOI-CODI)', 'temático', None),
 (1027, '11595.text', 'Agência Brasileira de Inteligência (Abin)', 'temático',
None),
 (1028, '11596.text', 'Associação Brasileira de Emissoras de Rádio e Televisão
(ABERT)', 'temático', None),
 (1029, '11597.text', 'Associação Nacional de Jornais (ANJ)', 'temático', None),
 (1030, '11598.text', 'Associação Nacional dos Membros do Ministério Público
(CONAMP)', 'temático', None),
 (1031, '11599.text', 'CAROS AMIGOS', 'temático', None),
 (1034, '11600.text', 'CARTA CAPITAL', 'temático', None),
 (1035, '11601.text', 'Central dos Trabalhadores e das Trabalhadoras do Brasil
(CTB)', 'temático', None),
 (1036, '11602.text', 'Central Geral dos Trabalhadores do Brasil (CGTB)',
'temático', None),
 (1037, '11603.text', 'Conselho de Comunicação Social (CCS)', 'temático', None)]
```

Se quisermos os resultado na forma de um `Dataframe`, podemos usar o **Pandas**.

In [40]:
```
pd.read_sql_query("select * from resultados where natureza='temático'",
        con=engine).head()
```

[40]:

	index	arquivo	title	natureza	sexo
0	354	10989.text	Destacamento de Operações e Informações Cent...	temático	None
1	1027	11595.text	Agência Brasileira de Inteligência (Abin)	temático	None
2	1028	11596.text	Associação Brasileira de Emissoras de Rádio e ...	temático	None
3	1029	11597.text	Associação Nacional de Jornais (ANJ)	temático	None
4	1030	11598.text	Associação Nacional dos Membros do Ministério ...	temático	None

IV. EXERCÍCIOS

1. Construa uma função para buscar apenas verbetes de personagens que tenham ocupado o cargo de deputado federal. Tabule os resultados incluindo o número de mandatos.
2. Construa uma função para buscar o primeiro verbete temático e apresente o seu conteúdo.
3. Construa uma linha do tempo que represente a cobertura histórica do DHBB.

In [ ]: