INSTITUTO BUTANTAN

centre of excellence in new target discovery

Seminários de Python

Tipos de Arquivos – como acessar

PhD Flavio Lichtenstein

Bioinformatics, Systems Biology, and Biostatistics

CENTD – Instituttuto Butantan

setembro/2020

Tipos de arquivos

Há vários tipos de arquivos:

- Texto ou Doc: txt, doc(x), pdf, ...
- Imagem: jpeg, png, tiff, gif, eps, svg (vetorizada), … + com/sem perda
- Videos: MP4, FLV, AVI, WMV, WEBM, QuickTime, 3GP, OGG, ...
- Tabelas: xls(x), csv (comma separated values), tsv (tab sv), …
- Arquivos texto extruturados:
  - HTML – fracamente estruturado
  - XML – estruturado e coeso
  - JSON (jason) – estruturado e coeso
  - RDF – estruturado, coeso e semântico

Arquivos CSV ou TSV

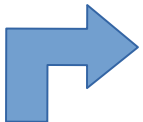
Arquivos tipo Tabela:

- Excel: arquivo binário, compactado, baseado em XML (eXtensible Markup Language)
  - Pode ser estruturado ou não (colunas não formatadas)

- CSV: arquivo texto, separado por vírgulas
- TSV: arquivo texto, separado por tabulações
  - Podem conter:
    - Cabeçalho (ou não) – header
    - Aspas nos conteúdos

Processando arquivos texto

```
import os
fname = '../data/exemplo_estranho.txt'
os.path.exists(fname)

with open(fname, 'r') as f:

    while True:
        line = f.readline()
        if line == '':
            break

        line = line.strip() # tire os espaços antes e depois
        mat = line.split('-')

        termo0 = mat[0]
        mat = mat[1].split(' ')
        termo1 = mat[0]
        termo2 = mat[1]

        print(termo0, '\t', termo1, '\t', termo2)


#-- \n == carriage return
print("\n---- fim ----")
```



https://github.com/flalix/curso_python/tree/master/lecture01%20-%20python%20b%C3%A1sico

Arquivos disponíveis na web

Arquivos disponíveis na WEB:

- WHO - https://www.who.int/tb/country/data/download/en/
  - Baixando: TB outcomes (TB_outcomes_2020-09-11.csv)
  - Acessando via pandas

- CSV: arquivo texto, separado por vírgulas
- TSV: arquivo texto, separado por tabulações
  - Podem conter:
    - Cabeçalho (ou não) – header
    - Aspas nos conteúdos

Como abrir um arquivo CSV ou TSV

```
# Dados de TB da WHO - https://www.who.int/tb/country/data/download/en/
fname = "../data/TB_outcomes_2020-09-11.csv"
os.path.exists(fname)

# carregando a biblioteca/pacote pandas
import pandas as pd

df = pd.read_csv(fname)
# linhas e colunas == shape
print(df.shape)

df.head()
```

Out[22]:

	country	iso2	iso3	iso_numeric	g_whoregion	year	rep_meth	new_sp_coh	new_sp_cur	new_sp_cmplt	...	mdr_coh	mdr_succ	mdr_fail	mdr_die
0	Afghanistan	AF	AFG	4	EMR	1994	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	Na
1	Afghanistan	AF	AFG	4	EMR	1995	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	Na
2	Afghanistan	AF	AFG	4	EMR	1996	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	Na
3	Afghanistan	AF	AFG	4	EMR	1997	100.0	2001.0	786.0	108.0	...	NaN	NaN	NaN	Na
4	Afghanistan	AF	AFG	4	EMR	1998	100.0	2913.0	772.0	199.0	...	NaN	NaN	NaN	Na

5 rows × 72 columns

Como abrir um arquivo remotamente CSV ou TSV

```
url = '
https://apps.who.int/gho/athena/data/GHO/WHOSIS_000004?filter=COUNTRY:-;REGION:*&x-sideaxis=REGION;YEAR&x-topaxis=GHO;SEX&profile=crosstable&format=csv
'

#-- adult mortality
dfam = pd.read_csv(url)
print(dfam.shape)

dfam.head()
```

Out[35]:

	Unnamed: 0	Unnamed: 1	Adult mortality rate (probability of dying between 15 and 60 years per 1000 population)	Adult mortality rate (probability of dying between 15 and 60 years per 1000 population).1	Adult mortality rate (probability of dying between 15 and 60 years per 1000 population).2
0	WHO region	Year	Both sexes	Male	Female
1	Global	2016	142	171	112
2	Global	2015	144	174	114
3	Global	2014	146	176	115
4	Global	2013	148	178	117

Padrão HTML (hypertext markup language)

(veja https://www.w3schools.com/html/default.asp)

Padrão HTML: adotado para troca de mensagens na internet (browser)

```
<!DOCTYPE html>
<html>
<head>
  <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
  <title>Aqui vai um titulo qualquer</title>
</head>
<body>

<h1>Primeiro cabeçalho</h1>
<h3>Primeiro cabeçalho</h3>
<p>Parágrafo simples ...</p>

</body>
</html>
```

Exercício: salve este texto num arquivo texto, nomeie-o “arquivo.html”, arraste no browse (navagador)

Padrão XML - eXtensible Markup Language: veja

https://www.w3schools.com/xml/xml_whatis.asp

Padrão XML:

- Padrão mais rígido e seguro para troca de mensagens

- Estruturalmente parecido com XML
- Serve para armazenar e transportar dados
  - Arquivos de configuração
  - Arquivos que auxiliam na troca de mensagens
- Consistente:
  - Caso uma tag não feche, não é processado
  - Pode-se usar um código Hash para certificar integridade via checksum

Padrão XML

```
<mensagem_screta>
  <depto>Laboratório de vacinas</depto>
  <to>João</to>
  <from>Pedro</from>
  <heading>Descoberta de vacina da Dengue</heading>
  <body>
     <issue1>
       <item1>assunto 01</item1>
     </issue1>
     <issue2>
       <item2>assunto 02</item2>
     </issue2>
     <dateline>
           <location>Laboratório Clark Kent</location>
           <date>2020-12-31 12:00</date>
      </dateline>
  </body>
</mensagem_screta>
```

Padrão JSON – JavaScript Object Notation

https://www.w3schools.com/js/js_json_intro.asp ; https://json.org/example.html

Padrão JSON: mais moderno – troca de mensagens na Internet e entre aplicativos

```
{
    "glossary": {
        "title": "example glossary",
                "GlossDiv": {
            "title": "S",
                        "GlossList": {
                "GlossEntry": {
                    "ID": "SGML",
                                        "SortAs": "SGML",
                                        "GlossTerm": "Standard Generalized Markup Language",
                                        "Acronym": "SGML",
                                        "Abbrev": "ISO 8879:1986",
                                        "GlossDef": {
                        "para": "A meta-markup language, used to create markup languages such as DocBook.",
                                                "GlossSeeAlso": ["GML", "XML"]
                    },
                                        "GlossSee": "markup"
                }
            }
        }
    }
}
```

Padrão RDF resource description framework

https://www.w3.org/RDF/ ; https://www.w3schools.com/xml/xml_rdf.asp

Padrão RDF: adotado para WEB semantica

- É o padrão mais completo entre os que aqui relatamos
- Este padrão que permite conter informações semânticas - ontologia
- Utilizado para descrever recursos na WEB
  - Não deve ser lido por humanos (só por máquinas)
  - Utiliza o padrão XML, porém sempre é complexo (e grande)
- É mais que consistente:
  - Consegue definir semântica:
    - Este é o objeto 01
    - Este é o objeto 02
    - O objeto01 contém o objeto02

Padrão RDF resource description framework

https://www.w3.org/RDF/ ; https://www.w3schools.com/xml/xml_rdf.asp

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cd="http://www.recshop.fake/cd#">

<rdf:Description
rdf:about="http://www.recshop.fake/cd/Empire Burlesque">
  <cd:artist>Bob Dylan</cd:artist>
  <cd:country>USA</cd:country>
  <cd:company>Columbia</cd:company>
  <cd:price>10.90</cd:price>
  <cd:year>1985</cd:year>
</rdf:Description>

<rdf:Description
rdf:about="http://www.recshop.fake/cd/Hide your heart">
  <cd:artist>Bonnie Tyler</cd:artist>
  <cd:country>UK</cd:country>
  <cd:company>CBS Records</cd:company>
  <cd:price>9.90</cd:price>
  <cd:year>1988</cd:year>
</rdf:Description>
…. </rdf:RDF>
```

Obrigado

Dúvidas?

PhD Flavio Lichtenstein

Bioinformatics, Systems Biology, and Biostatistics

CENTD – Institututo Butantan

setembro/2020