Taller 8

Métodos Computacionales para Políticas Públicas - URosario

Entrega: viernes 23-abr-2021 11:59 PM

[Juan Diego Castro Rodriguez]

[juand.castro@urosario.edu.co]

Instrucciones:

- Guarde una copia de este *Jupyter Notebook* en su computador, idealmente en una carpeta destinada al material del curso.
- Modifique el nombre del archivo del *notebook*, agregando al final un guión inferior y su nombre y apellido, separados estos últimos por otro guión inferior. Por ejemplo, mi *notebook* se llamaría: mcpp_taller8_santiago_matallana
- Marque el *notebook* con su nombre y e-mail en el bloque verde arriba. Reemplace el texto "[Su nombre acá]" con su nombre y apellido. Similar para su e-mail.
- Desarrolle la totalidad del taller sobre este *notebook*, insertando las celdas que sea necesario debajo de cada pregunta. Haga buen uso de las celdas para código y de las celdas tipo *markdown* según el caso.
- Recuerde salvar periódicamente sus avances.
- Cuando termine el taller:
  1. Descárguelo en PDF. Si tiene algún problema con la conversión, descárguelo en HTML.
  2. Suba todos los archivos a su repositorio en GitHub, en una carpeta destinada exclusivamente para este taller, antes de la fecha y hora límites.

---

1. [1 punto]

Usando expresiones regulares extraiga en una lista todos los números presentes en el siguiente objeto de Python:

ob1 = "JEFF BEZOS, the founder of Amazon, has reached a divorce settlement with his wife, MacKenzie. Mr Bezos will keep all the shares in the Washington Post and Blue Origin, a space-exploration firm, as well as 75% of the couple's Amazon stock. Mrs Bezos will retain a 4% stake in the tech giant, worth nearly $36bn, which is likely to make her the third-richest woman alive when the divorce is finalised."

In [7]:
```
import re
ob1 = "JEFF BEZOS, the founder of Amazon, has reached a divorce settlement with his wife, MacKenzie. Mr Bezos will keep all the shares in the Washington Post
```

In [3]:
```
re.findall("\d+", ob1)
```

Out[3]:
```
['75', '4', '36']
```

2. [1 punto]

Usando expresiones regulares ahora extraiga de *ob1* sólo los números que correspondan a porcentajes.

In [5]:
```
re.findall("\d+%", ob1)
```

Out[5]:
```
['75%', '4%']
```

3. [2 puntos]

Usando expresiones regulares, escriba una función de Python que reciba una fecha en formato **Marzo 7, 2019** y retorne la fecha en formato **2019-07-03**

In [34]:
```
def convertir_mes(mes):
    meses = ["Enero", "Febrero", "Marzo", "Abril", "Mayo", "Junio", "Julio", "Agosto", "Septiembre", "Octubre", "Noviembre", "Diciembre"]
    return meses.index(mes) + 1


def fecha_formato(x):
    a=re.search("\w+", x).group()
    b=re.findall("\d+", x)
    return(str(b[1])+"-"+str(b[0]).zfill(2)+"-"+str(convertir_mes(a)).zfill(2))
```

In [36]:
```
x="Marzo 7, 2019"
fecha_formato(x)
```

Out[36]:
```
'2019-07-03'
```

In [37]:
```
y="Abril 22, 2021"
fecha_formato(y)
```

Out[37]:
```
'2021-22-04'
```

In [38]:
```
z="Diciembre 1, 2020"
fecha_formato(z)
```

Out[38]:
```
'2020-01-12'
```

4. [3 puntos]

ob2 es un string que reune una lista de clases en una universidad. Use expresiones regulares para extraer los códigos de cada una de las clases. Ejemplo: El código de la clase **COMPSCI 143 (Spring 2012): Machine Learning** es 143.

ob2 = "COMPSCI 270 (Spring 2019): Introduction to Artificial Intelligence. COMPSCI 590.2 (Fall 2018): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 223 (Spring 2018): Computational Microeconomics. COMPSCI 570 (Fall 2017): Artificial Intelligence. COMPSCI 590.3 (Fall 2017) / 590.1 (Spring 2018): Ethics and AI. COMPSCI 590.2 (Spring 2017): Computation, Information, and Learning in Market Design. COMPSCI 590.4 (Spring 2016): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 290.4/590.4 (Spring 2015): Crowdsourcing Societal Tradeoffs. COMPSCI 570 (Fall 2014): Artificial Intelligence. COMPSCI 590.4 (Spring 2014): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 590.1 (Fall 2012): Linear and Integer Programming. COMPSCI 173 (Spring 2012): Computational Microeconomics. COMPSCI 296.1 (Fall 2011): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 296.1 (Fall 2010): Linear and Integer Programming. COMPSCI 173 (Spring 2010): Computational Microeconomics. COMPSCI 196.1/296.1 (Fall 2009): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. COMPSCI 170 (Spring 2009): Introduction to Artificial Intelligence. COMPSCI 270 (Fall 2008): Artificial Intelligence. COMPSCI 196/296.2 (Spring 2008): Linear and Integer Programming. COMPSCI 196.2 (Fall 2007): Introduction to Computational Economics. COMPSCI 296.3 (Spring 2007): Topics in Computational Economics. COMPSCI 296.2 (Fall 2006): Computational Game Theory and Mechanism Design."

In [39]:
```
ob2 = "COMPSCI 270 (Spring 2019): Introduction to Artificial Intelligence. COMPSCI 590.2 (Fall 2018): Computational Microeconomics: Game Theory, Social Choice
```

In [42]:
```
re.findall("COMPSCI (\d+)", ob2)
```

Out[42]:
```
['270',
 '590',
 '223',
 '570',
 '590',
 '590',
 '590',
 '290',
 '570',
 '590',
 '590',
 '173',
 '296',
 '296',
 '173',
 '196',
 '170',
 '270',
 '196',
 '196',
 '296',
 '296']
```

5. [5 puntos]

ob3 es un string que reune una lista de publicaciones. Use expresiones regulares para extraer todos los *Journals* en los cuales el autor ha publicado. Ejemplo: El paper **Bail, CA. "The configuration of symbolic boundaries against immigrants in Europe." American Sociological Review 73.1 (January 1, 2008): 37-59. Full Text** fue publicado en el Journal *American Sociological Review*

ob3 = "Bail, CA, Argyle, LP, Brown, TW, Bumpus, JP, Chen, H, Hunzaker, MBF, Lee, J, Mann, M, Merhout, F, and Volfovsky, A. \"Exposure to opposing views on social media can increase political polarization.\" Proceedings of the National Academy of Sciences of the United States of America 115.37 (September 2018): 9216-9221. Full Text Open Access Copy.\n", "Bail, CA, Merhout, F, and Ding, P. \"Using Internet search data to examine the relationship between anti-Muslim and pro-ISIS sentiment in U.S. counties.\" Science Advances 4.6 (June 6, 2018): eaao5948-null. Full Text Open Access Copy.\n", "Bail, CA, Brown, TW, and Mann, M. \"Channeling Hearts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation.\" American Sociological Review 82.6 (December 1, 2017): 1188-1213. Full Text.\n", "Bail, CA. \"Taming Big Data: Using App Technology to Study Organizational Behavior on Social Media.\" Sociological Methods and Research 46.2 (March 1, 2017): 189-217. Full Text.\n", "McDonnell, TE, Bail, CA, and Tavory, I. \"A Theory of Resonance.\" Sociological Theory 35.1 (March 1, 2017): 1-14. Full Text.\n", "Bail, CA. \"Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media.\" Proceedings of the National Academy of Sciences of the United States of America 113.42 (October 2016): 11823-11828. Full Text.\n", "Bail, CA. \"Emotional Feedback and the Viral Spread of Social Media Messages About Autism Spectrum Disorders.\" American journal of public health 106.7 (July 2016): 1173-1180. Full Text.\n", "Bail, CA. \"The public life of secrets: Deception, disclosure, and discursive framing in the policy process.\" Sociological Theory 33.2 (January 1, 2015): 97-124. Full Text.\n", "Bail, CA. \"The cultural environment: Measuring culture with big data.\" Theory and Society 43.3 (January 1, 2014): 465-524. Full Text.\""

In [84]:
```
ob3 = 'Bail, CA, Argyle, LP, Brown, TW, Bumpus, JP, Chen, H, Hunzaker, MBF, Lee, J, Mann, M, Merhout, F, and Volfovsky, A. "Exposure to opposing views on soci
```

In [136…
```
re.findall( "{[\w\s*]+}\s\d+[.]\d+", str(ob3))
```

Out[136…
```
[' Proceedings of the National Academy of Sciences of the United States of America',
 ' Science Advances',
 ' American Sociological Review',
 ' Sociological Methods and Research',
 ' Sociological Theory',
 ' Proceedings of the National Academy of Sciences of the United States of America',
 ' American journal of public health',
 ' Sociological Theory',
 ' Theory and Society']
```

6. [10 puntos]

Vamos a hacer "scraping" a esta página: https://archive.ics.uci.edu/ml/datasets.php, que contiene un listado de 559 bases de datos que hacen parte del repositorio de la Universidad de California, Irvine.

Su tarea consiste en crear un "Pandas dataframe" que contenga 585 filas (una por base de datos) y las siguientes columnas:

- Nombre de la base de datos
- Link a la base de datos
- Tipo de datos
- Tipo de tarea a resolver (default task)
- Tipo de las variables
- Número de observaciones
- Número de variables
- Año
- Descripción de la base (Pista: Utilice la opción list view: https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=list)

Diviértase.

In [142…
```
import re
import requests
from bs4 import BeautifulSoup
import pandas as pd
def get_html(url):
    resp = requests.get(url).text
    resp = resp.replace(" ", " ")
    return BeautifulSoup(resp,"lxml")
def get_table1(html):
    table = html.find('table', attrs={"border":"1", "cellpadding":"5"})
    return table
def get_names(table):
    names = re.findall('(?<=\>)\.*([^>]*)(?=<\/a><\/b><\/p><\/td><\/tr><\/table><\/td>)', str(table))
    return names
def get_links(table):
    links = re.findall('"(datasets/.+?)"', str(table))
    links = links[::2]
    return links
def get_other_data(table):
    all_data = re.findall('<td><p class="normal">(.*?) <\/p><\/td>\n', str(table))
    return all_data
def get_description():
    resp = requests.get("https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=list").text
    html = BeautifulSoup(resp,"lxml")
    table = html.find('table', attrs={"cellpadding":"3"})
    description = re.findall('</a></b>:(.*?)</p><', str(table), re.DOTALL)
    return description
def frame():
    html = get_html('https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table')
    table = get_table1(html)
    names = get_names(table)
    links = get_links(table)
    other_data = get_other_data(table)
    data_type = other_data[::6]
    default_task = other_data[1::6]
    type_of_variables = other_data[2::6]
    observations = other_data[3::6]
    number_of_variables = other_data[4::6]
    year = other_data[5::6]
    description = get_description()
    df = pd.DataFrame({"Names": names, "Links":links, "Data Type":data_type, "Default Task":default_task,
                       "Type of Variables":type_of_variables,"Observations":observations,"Number of variables": number_of_variables, "Year":year, "Description":de
    return df
```

In [143…
```
frame= frame()
```

In [144…
```
frame
```

Out[144…

	Names	Links	Data Type	Default Task	Type of Variables	Observations	Number of variables	Year	Description
0	2.4 GHZ Indoor Channel Measurements	datasets/2.4+GHZ+Indoor+Channel+Measurements	Multivariate	Classification	Real	7840	5	2018	Measurement of the S21,consists of 10 sweeps,...
1	3D Road Network (North Jutland, Denmark)	datasets/3D+Road+Network+%28North+Jutland%2C+D...	Sequential, Text	Regression, Clustering	Real	434874	4	2013	3D road network with highly accurate elevatio...
2	3W dataset	datasets/3W+dataset	Multivariate, Time-Series	Classification, Clustering	Integer, Real	1984	8	2019	The first realistic and public dataset with r...
3	: Simulated Data set of Iraqi tourism places	datasets/%3A+Simulated+Data+set+of+Iraq+touri...	Multivariate	Classification, Clustering		232	16	2020	Simulated Data set of Iraqi tourism places wi...
4	A study of Asian Religious and Biblical Texts	datasets/A+study+of++Asian+Religious+and+Bibli...	Multivariate, Text	Classification, Clustering	Integer	590	8265	2019	Mainly from Project Gutenberg, we combine Upa...
...	...	...	...	...	...	...	...	...	...
580	Youtube cookery channels viewers comments in H...	datasets/Youtube+cookery+channels+viewers+comm...	Multivariate, Text	Classification		9800	3	2019	The datasets are taken from top 2 Indian cook...
581	YouTube Multiview Video Games Dataset	datasets/YouTube+Multiview+Video+Games+Dataset	Multivariate, Text	Classification, Clustering	Integer, Real	120000	1000000	2013	This dataset contains about 120k instances, e...
582	YouTube Spam Collection	datasets/YouTube+Spam+Collection	Text	Classification		1956	5	2017	It is a public set of comments collected for ...
583	Z-Alizadeh Sani	datasets/Z-Alizadeh+Sani		Classification	Integer, Real	303	56	2017	It was collected for CAD diagnosis.
584	Zoo	datasets/Zoo	Multivariate	Classification	Categorical, Integer	101	17	1990	Artificial, 7 classes of animals

585 rows × 9 columns
