Taller 8

Métodos Computacionales para Políticas Públicas - URosario

Entrega: viernes 10-abr-2020 11:59 PM

[Santiago Ortiz Ortiz]

[santiago.ortizo@urosario.edu.co (mailto:santiago.ortizo@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.

In [102]:
```
import re
import requests
from bs4 import BeautifulSoup
import pandas as pd
```

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 [1]:
```
ob1 = "JEFF BEZOS, the founder of Amazon, has reached a divorce settlement with h
```

In [2]:
```
ob1
```

Out[2]:
```
'JEFF BEZOS, the founder of Amazon, has reached a divorce settlement with his w
ife, MacKenzie. Mr Bezos will keep all the shares in the Washington Post and Bl
ue Origin, a space-exploration firm, as well as 75% of the couple's Amazon stoc
k. Mrs Bezos will retain a 4% stake in the tech giant, worth nearly $36bn, whic
h is likely to make her the third-richest woman alive when the divorce is final
ised.'
```

Tenemos dos formas:

- La primera solo tomará los números que hayan.
- La segunda tomará el número y adicional (como un elemento aparte) tomará el signo que acompaña al número, para saber de que hablamos.

In [6]:
```
# Solo número
patron = "[\d]+"
re.findall(patron, ob1)
```

Out[6]:
```
['75', '4', '36']
```

In [5]:
```
# Número más símbolo
patron = "([\d]+)([\w%]+)"
re.findall(patron, ob1)
```

Out[5]:
```
[('75', '%'), ('4', '%'), ('36', 'bn')]
```

2. [1 punto]

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

In [7]:
```
patron = "([\d]+)[%]+"
re.findall(patron, ob1)
```

Out[7]:
```
['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**

Pruebas

In [19]:
```
# Prueba
fecha = "Marzo 7, 2019"
patron = "([A-Z][a-z]+)\s([\d]+),\s([\d]{4})"
re.findall(patron, fecha)[0]
```

Out[19]:
```
('Marzo', '7', '2019')
```

In [18]:
```
tupla = re.findall(patron, fecha)[0]
print(tupla[0])
print(tupla[1])
print(tupla[2])
```

```
Marzo
7
2019
```

Solución

In [24]:
```
def get_date(fecha):
    tupla = re.findall("([A-Z][a-z]+)\s([\d]+),\s([\d]{4})", fecha)[0] # Me crea
    if tupla[0] == "Enero":
        mes = "01"
    elif tupla[0] == "Febrero":
        mes = "02"
    elif tupla[0] == "Marzo":
        mes = "03"
    elif tupla[0] == "Abril":
        mes = "04"
    elif tupla[0] == "Mayo":
        mes = "05"
    elif tupla[0] == "Junio":
        mes = "06"
    elif tupla[0] == "Julio":
        mes = "07"
    elif tupla[0] == "Agosto":
        mes = "08"
    elif tupla[0] == "Septiembre":
        mes = "09"
    elif tupla[0] == "Octubre":
        mes = "10"
    elif tupla[0] == "Noviembre":
        mes = "11"
    elif tupla[0] == "Diciembre":
        mes = "12"
    if len(tupla[1]) == 1:
        dia = "0" + tupla[1]
    else:
        dia = tupla[1]
    año = tupla[2]
    nuevo_formato = año+"-"+dia+"-"+mes
    return nuevo_formato
```

In [28]:
```
fecha = "Mayo 18, 2009" # Pruebe con cualquier fecha
get_date(fecha)
```

Out[28]:
```
'2009-18-05'
```

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 [29]:
```
ob2 = "COMPSCI 270 (Spring 2019): Introduction to Artificial Intelligence. COMPSC
```

In [30]:
```
ob2
```

Out[30]:
```
'COMPSCI 270 (Spring 2019): Introduction to Artificial Intelligence. COMPSCI 59
0.2 (Fall 2018): Computational Microeconomics: Game Theory, Social Choice, and
Mechanism Design. COMPSCI 223 (Spring 2018): Computational Microeconomics. COMP
SCI 570 (Fall 2017): Artificial Intelligence. COMPSCI 590.3 (Fall 2017) / 590.1
(Spring 2018): Ethics and AI. COMPSCI 590.2 (Spring 2017): Computation, Informa
tion, 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 201
4): Artificial Intelligence. COMPSCI 590.4 (Spring 2014): Computational Microec
onomics: 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 T
heory, Social Choice, and Mechanism Design. COMPSCI 296.1 (Fall 2010): Linear a
nd Integer Programming. COMPSCI 173 (Spring 2010): Computational Microeconomic
s. 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. COMP
SCI 196/296.2 (Spring 2008): Linear and Integer Programming. COMPSCI 196.2 (Fal
l 2007): Introduction to Computational Economics. COMPSCI 296.3 (Spring 2007):
Topics in Computational Economics. COMPSCI 296.2 (Fall 2006): Computational Gam
e Theory and Mechanism Design.'
```

In [37]:
```
re.findall("COMPSCI\s([\d./]+)\s[(\w\s)]+:\s([A-Za-z\s]+)", ob2)
```

Out[37]:
```
[('270', 'Introduction to Artificial Intelligence'),
 ('590.2', 'Computational Microeconomics'),
 ('223', 'Computational Microeconomics'),
 ('570', 'Artificial Intelligence'),
 ('590.2', 'Computation'),
 ('590.4', 'Computational Microeconomics'),
 ('290.4/590.4', 'Crowdsourcing Societal Tradeoffs'),
 ('570', 'Artificial Intelligence'),
 ('590.4', 'Computational Microeconomics'),
 ('590.1', 'Linear and Integer Programming'),
 ('173', 'Computational Microeconomics'),
 ('296.1', 'Computational Microeconomics'),
 ('296.1', 'Linear and Integer Programming'),
 ('173', 'Computational Microeconomics'),
 ('196.1/296.1', 'Computational Microeconomics'),
 ('170', 'Introduction to Artificial Intelligence'),
 ('270', 'Artificial Intelligence'),
 ('196/296.2', 'Linear and Integer Programming'),
 ('196.2', 'Introduction to Computational Economics'),
 ('296.3', 'Topics in Computational Economics'),
 ('296.2', 'Computational Game Theory and Mechanism Design')]
```

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 [33]:
```
ob3 = '"Bail, CA, Argyle, LP, Brown, TW, Bumpus, JP, Chen, H, Hunzaker, MBF, Lee,
```

In [34]:
```
ob3
```

Out[34]:
```
'"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 [36]:
```
re.findall('"\s([\w\s]+)\s\d+', ob3)
```

Out[36]:
```
['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 (https://archive.ics.uci.edu/ml/datasets.php), que contiene un listado de 468 bases de datos que hacen parte del repositorio de la Universidad de California, Irvine.

Su tarea consiste en crear un "Pandas dataframe" que contenga 468 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 (https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=list))

Diviértase.

In [3]:
```
html = requests.get("https://archive.ics.uci.edu/ml/datasets.php").text
```

In [4]:
```
html
```

```
ss="normal, whitetext"><p class="normal, whitetext"><a href=\'datasets.php?fo
rmat=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameDown&view=table\'><b>Na
me</b></a></p></td>\n\t\t<!-- <td><p class="normal, whitetext"><b>Abstract</b
></p></td> -->\n\t\t<td><p class="normal, whitetext"><a href=\'datasets.php?f
ormat=&task=&att=&area=&numAtt=&numIns=&type=&sort=typeUp&view=table\'><b>Dat
a Types</b></a></p></td>\n\t\t<td><p class="normal, whitetext"><a href=\'data
sets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=taskUp&view=tabl
e\'><b>Default Task</b></a></p></td>\n\t\t<td><p class="normal, whitetext"><a
href=\'datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=attTy
peUp&view=table\'><b>Attribute Types</b></a></p></td>\n\t\t<td><p class="norm
al, whitetext"><a href=\'datasets.php?format=&task=&att=&area=&numAtt=&numIns
=&type=&sort=instUp&view=table\'><b># Instances</b></a></p></td>\n\t\t<td><p
 class="normal, whitetext"><a href=\'datasets.php?format=&task=&att=&area=&nu
mAtt=&numIns=&type=&sort=attUp&view=table\'><b># Attributes</b></a></p></td>
\n\t\t<td><p class="normal, whitetext"><a href=\'datasets.php?format=&task=&a
tt=&area=&numAtt=&numIns=&type=&sort=dateUp&view=table\'><b>Year</b></a></p>
</td>\n\n\t\t<!-- <td><p class="normal, whitetext"><b>Area</b></p></td> -->\n
\n\n\t\t</tr><tr>\n\t\t\t<td><table><tr><td><a href="datasets/Abalone"><img s
rc="assets/MLimages/SmallLarge1.jpg" border=1 /></a> </td><td><p class
="normal"><b><a href="datasets/Abalone">Abalone</a></b></p></td></tr></table>
```

In [24]:
```
soup = BeautifulSoup(html, "lxml")
```

In [25]:
```
soup
```

```
</td>
</tr>
<tr> <td bgcolor="#003366"><p class="whitetext"><b>Data Type</b> </p>
</td>
</tr>
<tr>
<td valign="top"><p class="normal"><a href="datasets.php?format=&amp;task=&amp;att=&amp;area=&amp;numAtt=&amp;numIns=&amp;type=mvar&amp;sort=nameUp&amp;view=table">Multivariate</a> <font color="red">(383)</font><br/><a href="datasets.php?format=&amp;task=&amp;att=&amp;area=&amp;numAtt=&amp;numIns=&amp;type=uvar&amp;sort=nameUp&amp;view=table">Univariate</a> <font color="red">(24)</font><br/><a href="datasets.php?format=&amp;task=&amp;att=&amp;area=&amp;numAtt=&amp;numIns=&amp;type=seq&amp;sort=nameUp&amp;view=table">Sequential</a> <font color="red">(52)</font><br/><a href="datasets.php?format=&amp;task=&amp;att=&amp;area=&amp;numAtt=&amp;numIns=&amp;type=ts&amp;sort=nameUp&amp;view=table">Time-Series</a> <font color="red">(102)</font><br/><a href="datasets.php?format=&amp;task=&amp;att=&amp;area=&amp;numAtt=&amp;numIns=&amp;type=text&amp;sort=nameUp&amp;view=table">Text</a> <font color="red">(57)</font><br/><a href="datasets.php?format=&amp;task=&amp;att=&amp;area=&amp;numAtt=&amp;numIns=&amp;type=dt&amp;sort=nameUp&amp;view=table">Domain-Theory</a> <font color
```

Cuando uno revisa el código HTML y busca la primera base de datos: Abalone, encuentra lo siguiente: href="datasets/Abalone">Abalone.

- Es decir, aparece el link y el nombre de la base de datos. Entonces con el comando "a" buscamos TODOS los links de la página principal, pero solo utilizaremos aquellos que cumplan con esa caracteristica. Así obtendremos el nombre de la base de datos y una parte del código del link que la acompaña.
- Sin embargo, no vienen con: https://archive.ics.uci.edu/ml/ (https://archive.ics.uci.edu/ml/) antes. Esto se puede solucionar luego.

In [27]:
```
links_nombres = soup.find_all("a")
```

In [13]:
```
links_nombres
```

```
 <a href="datasets/Quadruped+Mammals">Quadruped Mammals</a>,
 <a href="datasets/Servo"><img border="1" src="assets/MLimages/SmallLarge87.j
pg"/></a>,
 <a href="datasets/Servo">Servo</a>,
 <a href="datasets/Shuttle+Landing+Control"><img border="1" src="assets/MLima
ges/SmallLarge92.jpg"/></a>,
 <a href="datasets/Shuttle+Landing+Control">Shuttle Landing Control</a>,
 <a href="datasets/Solar+Flare"><img border="1" src="assets/MLimages/SmallLar
ge89.jpg"/></a>,
 <a href="datasets/Solar+Flare">Solar Flare</a>,
 <a href="datasets/Soybean+%28Large%29"><img border="1" src="assets/MLimages/
SmallLarge90.jpg"/></a>,
 <a href="datasets/Soybean+%28Large%29">Soybean (Large)</a>,
 <a href="datasets/Soybean+%28Small%29"><img border="1" src="assets/MLimages/
SmallLarge90.jpg"/></a>,
 <a href="datasets/Soybean+%28Small%29">Soybean (Small)</a>,
 <a href="datasets/Challenger+USA+Space+Shuttle+O-Ring"><img border="1" src
="assets/MLimages/SmallLarge92.jpg"/></a>,
 <a href="datasets/Challenger+USA+Space+Shuttle+O-Ring">Challenger USA Space
Shuttle O-Ring</a>,
```

In [38]:
```
re.findall('"(datasets/[\w+:%,.¿\'()-]+)">([\w\s:,.¿\'%()-]+)', str(links_nombres
# Note que el primer elemento es el link, y el segundo elemento es el nombre de l
```

```
 ('datasets/Molecular+Biology+%28Protein+Secondary+Structure%29',
  'Molecular Biology (Protein Secondary Structure)'),
 ('datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29',
  'Molecular Biology (Splice-junction Gene Sequences)'),
 ('datasets/MONK%27s+Problems', "MONK's Problems"),
 ('datasets/Moral+Reasoner', 'Moral Reasoner'),
 ('datasets/Multiple+Features', 'Multiple Features'),
 ('datasets/Mushroom', 'Mushroom'),
 ('datasets/Musk+%28Version+1%29', 'Musk (Version 1)'),
 ('datasets/Musk+%28Version+2%29', 'Musk (Version 2)'),
 ('datasets/Nursery', 'Nursery'),
 ('datasets/Othello+Domain+Theory', 'Othello Domain Theory'),
 ('datasets/Page+Blocks+Classification', 'Page Blocks Classification'),
 ('datasets/Optical+Recognition+of+Handwritten+Digits',
  'Optical Recognition of Handwritten Digits'),

 ('datasets/Pen-Based+Recognition+of+Handwritten+Digits',
  'Pen-Based Recognition of Handwritten Digits'),
 ('datasets/Post-Operative+Patient', 'Post-Operative Patient'),
 ('datasets/Primary+Tumor', 'Primary Tumor'),
```

In [66]:
```
nombres = re.findall('"(datasets/[\w+:%,.¿\'()-]+)">([\w\s:,.¿\'%()-]+)', str(li
```

In [67]:
```
len(nombres)
```

Out[67]: 497

In [68]:
```
nombres
```

```
  'Online Shoppers Purchasing Intention Dataset'),
 ('datasets/PMU-UD', 'PMU-UD'),
 ('datasets/Parkinson%27s+Disease+Classification',
  "Parkinson's Disease Classification"),
 ('datasets/Electrical+Grid+Stability+Simulated+Data+',
  'Electrical Grid Stability Simulated Data '),
 ('datasets/Caesarian+Section+Classification+Dataset',
  'Caesarian Section Classification Dataset'),
 ('datasets/BAUM-1', 'BAUM-1'),
 ('datasets/BAUM-2', 'BAUM-2'),
 ('datasets/Audit+Data', 'Audit Data'),
 ('datasets/BuddyMove+Data+Set', 'BuddyMove Data Set'),
 ('datasets/Real+estate+valuation+data+set', 'Real estate valuation data se
t'),
 ('datasets/Early+biomarkers+of+Parkinson%E2%80%99s+disease+based+on+natural+
connected+speech+Data+Set+',
  'Early biomarkers of Parkinson'),
 ('datasets/Somerville+Happiness+Survey', 'Somerville Happiness Survey'),
 ('datasets/2.4+GHZ+Indoor+Channel+Measurements',
  '2.4 GHZ Indoor Channel Measurements'),
```

In [69]:
```
len(nombres)
```

Out[69]: 497

In [70]:
```
url = []
for i in range(len(nombres)):
    url.append(nombres[i][0])
```

In [71]:
```
url
```

Out[71]:
```
['datasets/Abalone',
 'datasets/Adult',
 'datasets/Annealing',
 'datasets/Anonymous+Microsoft+Web+Data',
 'datasets/Arrhythmia',
 'datasets/Artificial+Characters',
 'datasets/Audiology+%28Original%29',
 'datasets/Audiology+%28Standardized%29',
 'datasets/Auto+MPG',
 'datasets/Automobile',
 'datasets/Badges',
 'datasets/Balance+Scale',
 'datasets/Balloons',
 'datasets/Breast+Cancer',
 'datasets/Breast+Cancer+Wisconsin+%28Original%29',
 'datasets/Breast+Cancer+Wisconsin+%28Prognostic%29',
 'datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29',
 'datasets/Pittsburgh+Bridges',
 'datasets/Car+Evaluation',
 'datasets/Census+Income',
 'datasets/Chess+%28King-Rook+vs.+King-Knight%29',
 'datasets/Chess+%28King-Rook+vs.+King-Pawn%29',
 'datasets/Chess+%28King-Rook+vs.+King%29',
 'datasets/Chess+%28Domain+Theories%29',
 'datasets/Bach+Chorales',
 'datasets/Connect-4',
 'datasets/Credit+Approval',
 'datasets/Japanese+Credit+Screening',
 'datasets/Computer+Hardware',
 'datasets/Contraceptive+Method+Choice',
 'datasets/Covertype',
 'datasets/Cylinder+Bands',
 'datasets/Dermatology',
 'datasets/Diabetes',
 'datasets/DGP2+-+The+Second+Data+Generation+Program',
 'datasets/Document+Understanding',
 'datasets/EBL+Domain+Theories',
 'datasets/Echocardiogram',
 'datasets/Ecoli',
 'datasets/Flags',
 'datasets/Function+Finding',
 'datasets/Glass+Identification',
 'datasets/Haberman%27s+Survival',
 'datasets/Hayes-Roth',
 'datasets/Heart+Disease',
 'datasets/Hepatitis',
 'datasets/Horse+Colic',
 'datasets/ICU',
 'datasets/Image+Segmentation',
 'datasets/Internet+Advertisements',
 'datasets/Ionosphere',
 'datasets/Iris',
 'datasets/ISOLET',
 'datasets/Kinship',
 'datasets/Labor+Relations',
```

```
'datasets/LED+Display+Domain',
'datasets/Lenses',
'datasets/Letter+Recognition',
'datasets/Liver+Disorders',
'datasets/Logic+Theorist',
'datasets/Lung+Cancer',
'datasets/Lymphography',
'datasets/Mechanical+Analysis',
'datasets/Meta-data',
'datasets/Mobile+Robots',
'datasets/Molecular+Biology+%28Promoter+Gene+Sequences%29',
'datasets/Molecular+Biology+%28Protein+Secondary+Structure%29',
'datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29',
'datasets/MONK%27s+Problems',
'datasets/Moral+Reasoner',
'datasets/Multiple+Features',
'datasets/Mushroom',
'datasets/Musk+%28Version+1%29',
'datasets/Musk+%28Version+2%29',
'datasets/Nursery',
'datasets/Othello+Domain+Theory',
'datasets/Page+Blocks+Classification',
'datasets/Optical+Recognition+of+Handwritten+Digits',
'datasets/Pen-Based+Recognition+of+Handwritten+Digits',
'datasets/Post-Operative+Patient',
'datasets/Primary+Tumor',
'datasets/Prodigy',
'datasets/Qualitative+Structure+Activity+Relationships',
'datasets/Quadruped+Mammals',
'datasets/Servo',
'datasets/Shuttle+Landing+Control',
'datasets/Solar+Flare',
'datasets/Soybean+%28Large%29',
'datasets/Soybean+%28Small%29',
'datasets/Challenger+USA+Space+Shuttle+O-Ring',
'datasets/Low+Resolution+Spectrometer',
'datasets/Spambase',
'datasets/SPECT+Heart',
'datasets/SPECTF+Heart',
'datasets/Sponge',
'datasets/Statlog+Project',
'datasets/Student+Loan+Relational',
'datasets/Teaching+Assistant+Evaluation',
'datasets/Tic-Tac-Toe+Endgame',
'datasets/Thyroid+Disease',
'datasets/Trains',
'datasets/University',
'datasets/Congressional+Voting+Records',
'datasets/Water+Treatment+Plant',
'datasets/Waveform+Database+Generator+%28Version+1%29',
'datasets/Waveform+Database+Generator+%28Version+2%29',
'datasets/Wine',
'datasets/Yeast',
'datasets/Zoo',
'datasets/Undocumented',
'datasets/Twenty+Newsgroups',
'datasets/Australian+Sign+Language+signs',
```

```
'datasets/Australian+Sign+Language+signs+%28High+Quality%29',
'datasets/US+Census+Data+%281990%29',
'datasets/Census-Income+%28KDD%29',
'datasets/Coil+1999+Competition+Data',
'datasets/Corel+Image+Features',
'datasets/E.+Coli+Genes',
'datasets/EEG+Database',
'datasets/El+Nino',
'datasets/Entree+Chicago+Recommendation+Data',
'datasets/CMU+Face+Images',
'datasets/Insurance+Company+Benchmark+%28COIL+2000%29',
'datasets/Internet+Usage+Data',
'datasets/IPUMS+Census+Database',
'datasets/Japanese+Vowels',
'datasets/KDD+Cup+1998+Data',
'datasets/KDD+Cup+1999+Data',
'datasets/M.+Tuberculosis+Genes',
'datasets/Movie',
'datasets/MSNBC.com+Anonymous+Web+Data',
'datasets/NSF+Research+Award+Abstracts+1990-2003',
'datasets/Pioneer-1+Mobile+Robot+Data',
'datasets/Pseudo+Periodic+Synthetic+Time+Series',
'datasets/Reuters-21578+Text+Categorization+Collection',
'datasets/Robot+Execution+Failures',
'datasets/Synthetic+Control+Chart+Time+Series',
'datasets/Syskill+and+Webert+Web+Page+Ratings',
'datasets/UNIX+User+Data',
'datasets/Volcanoes+on+Venus+-+JARtool+experiment',
'datasets/Statlog+%28Australian+Credit+Approval%29',
'datasets/Statlog+%28German+Credit+Data%29',
'datasets/Statlog+%28Heart%29',
'datasets/Statlog+%28Landsat+Satellite%29',
'datasets/Statlog+%28Image+Segmentation%29',
'datasets/Statlog+%28Shuttle%29',
'datasets/Statlog+%28Vehicle+Silhouettes%29',
'datasets/Connectionist+Bench+%28Nettalk+Corpus%29',
'datasets/Connectionist+Bench+%28Sonar%2C+Mines+vs.+Rocks%29',
'datasets/Connectionist+Bench+%28Vowel+Recognition+-+Deterding+Data%29',
'datasets/Economic+Sanctions',
'datasets/Protein+Data',
'datasets/Cloud',
'datasets/CalIt2+Building+People+Counts',
'datasets/Dodgers+Loop+Sensor',
'datasets/Poker+Hand',
'datasets/MAGIC+Gamma+Telescope',
'datasets/UJI+Pen+Characters',
'datasets/Mammographic+Mass',
'datasets/Forest+Fires',
'datasets/Reuters+Transcribed+Subset',
'datasets/Bag+of+Words',
'datasets/Concrete+Compressive+Strength',
'datasets/Hill-Valley',
'datasets/Arcene',
'datasets/Dexter',
'datasets/Dorothea',
'datasets/Gisette',
'datasets/Madelon',
```

```
'datasets/Ozone+Level+Detection',
'datasets/Abscisic+Acid+Signaling+Network',
'datasets/Parkinsons',
'datasets/Character+Trajectories',
'datasets/Blood+Transfusion+Service+Center',
'datasets/UJI+Pen+Characters+%28Version+2%29',
'datasets/Semeion+Handwritten+Digit',
'datasets/SECOM',
'datasets/Plants',
'datasets/Libras+Movement',
'datasets/Concrete+Slump+Test',
'datasets/Communities+and+Crime',
'datasets/Acute+Inflammations',
'datasets/Wine+Quality',
'datasets/URL+Reputation',
'datasets/p53+Mutants',
'datasets/Parkinsons+Telemonitoring',
'datasets/Demospongiae',
'datasets/Opinosis+Opinion+%26frasl%3B+Review',
'datasets/Breast+Tissue',
'datasets/Cardiotocography',
'datasets/Wall-Following+Robot+Navigation+Data',
'datasets/Spoken+Arabic+Digit',
'datasets/Localization+Data+for+Person+Activity',
'datasets/AutoUniv',
'datasets/Steel+Plates+Faults',
'datasets/MiniBooNE+particle+identification',
'datasets/YearPredictionMSD',
'datasets/PEMS-SF',
'datasets/OpinRank+Review+Dataset',
'datasets/Relative+location+of+CT+slices+on+axial+axis',
'datasets/Online+Handwritten+Assamese+Characters+Dataset',
'datasets/PubChem+Bioassay+Data',
'datasets/Record+Linkage+Comparison+Patterns',
'datasets/Communities+and+Crime+Unnormalized',
'datasets/Vertebral+Column',
'datasets/EMG+Physical+Action+Data+Set',
'datasets/Vicon+Physical+Action+Data+Set',
'datasets/Amazon+Commerce+reviews+set',
'datasets/Amazon+Access+Samples',
'datasets/Reuter_50_50',
'datasets/Farm+Ads',
'datasets/DBWorld+e-mails',
'datasets/KEGG+Metabolic+Relation+Network+%28Directed%29',
'datasets/KEGG+Metabolic+Reaction+Network+%28Undirected%29',
'datasets/Bank+Marketing',
'datasets/YouTube+Comedy+Slam+Preference+Data',
'datasets/Gas+Sensor+Array+Drift+Dataset',
'datasets/ILPD+%28Indian+Liver+Patient+Dataset%29',
'datasets/OPPORTUNITY+Activity+Recognition',
'datasets/Nomao',
'datasets/SMS+Spam+Collection',
'datasets/Skin+Segmentation',
'datasets/Planning+Relax',
'datasets/PAMAP2+Physical+Activity+Monitoring',
'datasets/Restaurant+%26+consumer+data',
'datasets/CNAE-9',
```

```
 'datasets/Individual+household+electric+power+consumption',
 'datasets/seeds',
 'datasets/Northix',
 'datasets/QtyT40I10D100K',
 'datasets/Legal+Case+Reports',
 'datasets/Human+Activity+Recognition+Using+Smartphones',
 'datasets/One-hundred+plant+species+leaves+data+set',
 'datasets/Energy+efficiency',
 'datasets/Yacht+Hydrodynamics',
 'datasets/Fertility',
 'datasets/Daphnet+Freezing+of+Gait',
 'datasets/3D+Road+Network+%28North+Jutland%2C+Denmark%29',
 'datasets/ISTANBUL+STOCK+EXCHANGE',
 'datasets/Buzz+in+social+media+',
 'datasets/First-order+theorem+proving',
 'datasets/Wearable+Computing%3A+Classification+of+Body+Postures+and+Movements
+%28PUC-Rio%29',
 'datasets/Gas+sensor+arrays+in+open+sampling+settings',
 'datasets/Climate+Model+Simulation+Crashes',
 'datasets/MicroMass',
 'datasets/QSAR+biodegradation',
 'datasets/BLOGGER',
 'datasets/Daily+and+Sports+Activities',
 'datasets/User+Knowledge+Modeling',
 'datasets/Reuters+RCV1+RCV2+Multilingual%2C+Multiview+Text+Categorization+Test
+collection',
 'datasets/NYSK',
 'datasets/Turkiye+Student+Evaluation',
 'datasets/ser+Knowledge+Modeling+Data+%28Students%27+Knowledge+Levels+on+DC+El
ectrical+Machines%29',
 'datasets/EEG+Eye+State',
 'datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure',
 'datasets/seismic-bumps',
 'datasets/banknote+authentication',
 'datasets/USPTO+Algorithm+Challenge%2C+run+by+NASA-Harvard+Tournament+Lab+and+
TopCoder++++Problem%3A+Pat',
 'datasets/YouTube+Multiview+Video+Games+Dataset',
 'datasets/Gas+Sensor+Array+Drift+Dataset+at+Different+Concentrations',
 'datasets/Activities+of+Daily+Living+%28ADLs%29+Recognition+Using+Binary+Senso
rs',
 'datasets/SkillCraft1+Master+Table+Dataset',
 'datasets/Weight+Lifting+Exercises+monitored+with+Inertial+Measurement+Units',
 'datasets/SML2010',
 'datasets/Bike+Sharing+Dataset',
 'datasets/Predict+keywords+activities+in+a+online+social+media',
 'datasets/Thoracic+Surgery+Data',
 'datasets/EMG+dataset+in+Lower+Limb',
 'datasets/SUSY',
 'datasets/HIGGS',
 'datasets/Qualitative_Bankruptcy',
 'datasets/LSVT+Voice+Rehabilitation',
 'datasets/Dataset+for+ADL+Recognition+with+Wrist-worn+Accelerometer',
 'datasets/Wilt',
 'datasets/User+Identification+From+Walking+Activity',
 'datasets/Activity+Recognition+from+Single+Chest-Mounted+Accelerometer',
 'datasets/Leaf',
 'datasets/Dresses_Attribute_Sales',
```

```
 'datasets/Tamilnadu+Electricity+Board+Hourly+Readings',
 'datasets/Airfoil+Self-Noise',
 'datasets/Wholesale+customers',
 'datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis',
 'datasets/Combined+Cycle+Power+Plant',
 'datasets/Urban+Land+Cover',
 'datasets/Diabetes+130-US+hospitals+for+years+1999-2008',
 'datasets/Bach+Choral+Harmony',
 'datasets/StoneFlakes',
 'datasets/Tennis+Major+Tournament+Match+Statistics',
 'datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings',
 'datasets/Gesture+Phase+Segmentation',
 'datasets/Perfume+Data',
 'datasets/BlogFeedback',
 'datasets/REALDISP+Activity+Recognition+Dataset',
 'datasets/Newspaper+and+magazine+images+segmentation+dataset',
 'datasets/AAAI+2014+Accepted+Papers',
 'datasets/Gas+sensor+array+under+flow+modulation',
 'datasets/Gas+sensor+array+exposed+to+turbulent+gas+mixtures',
 'datasets/UJIIndoorLoc',
 'datasets/Sentence+Classification',
 'datasets/Dow+Jones+Index',
 'datasets/sEMG+for+Basic+Hand+movements',
 'datasets/AAAI+2013+Accepted+Papers',
 'datasets/Geographical+Original+of+Music',
 'datasets/Condition+Based+Maintenance+of+Naval+Propulsion+Plants',
 'datasets/Grammatical+Facial+Expressions',
 'datasets/NoisyOffice',
 'datasets/MHEALTH+Dataset',
 'datasets/Student+Performance',
 'datasets/ElectricityLoadDiagrams20112014',
 'datasets/Gas+sensor+array+under+dynamic+gas+mixtures',
 'datasets/microblogPCU',
 'datasets/Firm-Teacher_Clave-Direction_Classification',
 'datasets/Dataset+for+Sensorless+Drive+Diagnosis',
 'datasets/TV+News+Channel+Commercial+Detection+Dataset',
 'datasets/Phishing+Websites',
 'datasets/Greenhouse+Gas+Observing+Network',
 'datasets/Diabetic+Retinopathy+Debrecen+Data+Set',
 'datasets/HIV-1+protease+cleavage',
 'datasets/Sentiment+Labelled+Sentences',
 'datasets/Online+News+Popularity',
 'datasets/Forest+type+mapping',
 'datasets/wiki4HE',
 'datasets/Online+Video+Characteristics+and+Transcoding+Time+Dataset',
 'datasets/Chronic_Kidney_Disease',
 'datasets/Machine+Learning+based+ZZAlpha+Ltd.+Stock+Recommendations+2012-201
4',
 'datasets/Folio',
 'datasets/Taxi+Service+Trajectory+-+Prediction+Challenge%2C+ECML+PKDD+2015',
 'datasets/Cuff-Less+Blood+Pressure+Estimation',
 'datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transi
tions',
 'datasets/Mice+Protein+Expression',
 'datasets/UJIIndoorLoc-Mag',
 'datasets/Heterogeneity+Activity+Recognition',
 'datasets/Educational+Process+Mining+%28EPM%29%3A+A+Learning+Analytics+Data+Se
```

```
t',
 'datasets/HEPMASS',
 'datasets/Indoor+User+Movement+Prediction+from+RSS+data',
 'datasets/Open+University+Learning+Analytics+dataset',
 'datasets/default+of+credit+card+clients',
 'datasets/Mesothelioma%E2%80%99s+disease+data+set+',
 'datasets/Online+Retail',
 'datasets/SIFT10M',
 'datasets/GPS+Trajectories',
 'datasets/Detect+Malacious+Executable%28AntiVirus%29',
 'datasets/Occupancy+Detection+',
 'datasets/Improved+Spiral+Test+Using+Digitized+Graphics+Tablet+for+Monitoring+
Parkinson%E2%80%99s+Disease',
 'datasets/News+Aggregator',
 'datasets/Air+Quality',
 'datasets/Twin+gas+sensor+arrays',
 'datasets/Gas+sensors+for+home+activity+monitoring',
 'datasets/Facebook+Comment+Volume+Dataset',
 'datasets/Smartphone+Dataset+for+Human+Activity+Recognition+%28HAR%29+in+Ambie
nt+Assisted+Living+%28AAL%29',
 'datasets/Polish+companies+bankruptcy+data',
 'datasets/Activity+Recognition+system+based+on+Multisensor+data+fusion+%28ARe
M%29',
 'datasets/Dota2+Games+Results',
 'datasets/Facebook+metrics',
 'datasets/UbiqLog+%28smartphone+lifelogging%29',
 'datasets/NIPS+Conference+Papers+1987-2015',
 'datasets/HTRU2',
 'datasets/Drug+consumption+%28quantified%29',
 'datasets/Appliances+energy+prediction',
 'datasets/Miskolc+IIS+Hybrid+IPS',
 'datasets/KDC-4007+dataset+Collection',
 'datasets/Geo-Magnetic+field+and+WLAN+dataset+for+indoor+localisation+from+wri
stband+and+smartphone',
 'datasets/DrivFace',
 'datasets/Website+Phishing',
 'datasets/YouTube+Spam+Collection',
 'datasets/Beijing+PM2.5+Data',
 'datasets/Cargo+2000+Freight+Tracking+and+Tracing',
 'datasets/Cervical+cancer+%28Risk+Factors%29',
 'datasets/Quality+Assessment+of+Digital+Colposcopies',
 'datasets/KASANDR',
 'datasets/FMA%3A+A+Dataset+For+Music+Analysis',
 'datasets/Air+quality',
 'datasets/Epileptic+Seizure+Recognition',
 'datasets/Devanagari+Handwritten+Character+Dataset',
 'datasets/Stock+portfolio+performance',
 'datasets/MoCap+Hand+Postures',
 'datasets/Early+biomarkers+of+Parkinson%92s+disease+based+on+natural+connected
+speech',
 'datasets/Data+for+Software+Engineering+Teamwork+Assessment+in+Education+Setti
ng',
 'datasets/PM2.5+Data+of+Five+Chinese+Cities',
 'datasets/Parkinson+Disease+Spiral+Drawings+Using+Digitized+Graphics+Tablet',
 'datasets/Sales_Transactions_Dataset_Weekly',
 'datasets/Las+Vegas+Strip',
 'datasets/Eco-hotel',
```

```
 'datasets/MEU-Mobile+KSD',
 'datasets/Crowdsourced+Mapping',
 'datasets/gene+expression+cancer+RNA-Seq',
 'datasets/Hybrid+Indoor+Positioning+Dataset+from+WiFi+RSSI%2C+Bluetooth+and+ma
gnetometer',
 'datasets/chestnut+%E2%80%93+LARVIC',
 'datasets/Burst+Header+Packet+%28BHP%29+flooding+attack+on+Optical+Burst+Switc
hing+%28OBS%29+Network',
 'datasets/Motion+Capture+Hand+Postures',
 'datasets/Anuran+Calls+%28MFCCs%29',
 'datasets/TTC-3600%3A+Benchmark+dataset+for+Turkish+text+categorization',
 'datasets/Gastrointestinal+Lesions+in+Regular+Colonoscopy',
 'datasets/Daily+Demand+Forecasting+Orders',
 'datasets/Paper+Reviews',
 'datasets/extention+of+Z-Alizadeh+sani+dataset',
 'datasets/Z-Alizadeh+Sani',
 'datasets/Dynamic+Features+of+VirusShare+Executables',
 'datasets/IDA2016Challenge',
 'datasets/DSRC+Vehicle+Communications',
 'datasets/Mturk+User-Perceived+Clusters+over+Images',
 'datasets/Character+Font+Images',
 'datasets/DeliciousMIL%3A+A+Data+Set+for+Multi-Label+Multi-Instance+Learning+w
ith+Instance+Labels',
 'datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++',
 'datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent+++',
 'datasets/APS+Failure+at+Scania+Trucks',
 'datasets/Wireless+Indoor+Localization',
 'datasets/HCC+Survival',
 'datasets/CSM+%28Conventional+and+Social+Media+Movies%29+Dataset+2014+and+201
5',
 'datasets/University+of+Tehran+Question+Dataset+2016+%28UTQD.2016%29',
 'datasets/Autism+Screening+Adult',
 'datasets/Activity+recognition+with+healthy+older+people+using+a+batteryless+w
earable+sensor',
 'datasets/Immunotherapy+Dataset',
 'datasets/Cryotherapy+Dataset+',
 'datasets/OCT+data+%26+Color+Fundus+Images+of+Left+%26+Right+Eyes',
 'datasets/Discrete+Tone+Image+Dataset',
 'datasets/News+Popularity+in+Multiple+Social+Media+Platforms',
 'datasets/Ultrasonic+flowmeter+diagnostics',
 'datasets/ICMLA+2014+Accepted+Papers+Data+Set',
 'datasets/BLE+RSSI+Dataset+for+Indoor+localization+and+Navigation',
 'datasets/Container+Crane+Controller+Data+Set',
 'datasets/Residential+Building+Data+Set',
 'datasets/Health+News+in+Twitter',
 'datasets/chipseq',
 'datasets/SGEMM+GPU+kernel+performance',
 'datasets/Repeat+Consumption+Matrices',
 'datasets/detection_of_IoT_botnet_attacks_N_BaIoT',
 'datasets/Absenteeism+at+work',
 'datasets/SCADI',
 'datasets/Condition+monitoring+of+hydraulic+systems',
 'datasets/Carbon+Nanotubes',
 'datasets/Optical+Interconnection+Network+',
 'datasets/Sports+articles+for+objectivity+analysis',
 'datasets/Breast+Cancer+Coimbra',
 'datasets/GNFUV+Unmanned+Surface+Vehicles+Sensor+Data',
```

```
 'datasets/Dishonest+Internet+users+Dataset',
 'datasets/Victorian+Era+Authorship+Attribution',
 'datasets/Simulated+Falls+and+Daily+Living+Activities+Data+Set',
 'datasets/Multimodal+Damage+Identification+for+Humanitarian+Computing',
 'datasets/EEG+Steady-State+Visual+Evoked+Potential+Signals',
 'datasets/Roman+Urdu+Data+Set',
 'datasets/Avila',
 'datasets/PANDOR',
 'datasets/Drug+Review+Dataset+%28Druglib.com%29',
 'datasets/Drug+Review+Dataset+%28Drugs.com%29',
 'datasets/Physical+Unclonable+Functions',
 'datasets/Superconductivty+Data',
 'datasets/WESAD+%28Wearable+Stress+and+Affect+Detection%29',
 'datasets/GNFUV+Unmanned+Surface+Vehicles+Sensor+Data+Set+2',
 'datasets/Student+Academics+Performance',
 'datasets/Online+Shoppers+Purchasing+Intention+Dataset',
 'datasets/PMU-UD',
 'datasets/Parkinson%27s+Disease+Classification',
 'datasets/Electrical+Grid+Stability+Simulated+Data+',
 'datasets/Caesarian+Section+Classification+Dataset',
 'datasets/BAUM-1',
 'datasets/BAUM-2',
 'datasets/Audit+Data',
 'datasets/BuddyMove+Data+Set',
 'datasets/Real+estate+valuation+data+set',
 'datasets/Early+biomarkers+of+Parkinson%E2%80%99s+disease+based+on+natural+con
nected+speech+Data+Set+',
 'datasets/Somerville+Happiness+Survey',
 'datasets/2.4+GHZ+Indoor+Channel+Measurements',
 'datasets/EMG+data+for+gestures',
 'datasets/Parking+Birmingham',
 'datasets/Behavior+of+the+urban+traffic+of+the+city+of+Sao+Paulo+in+Brazil',
 'datasets/Travel+Reviews',
 'datasets/Tarvel+Review+Ratings',
 'datasets/Rice+Leaf+Diseases',
 'datasets/Gas+sensor+array+temperature+modulation',
 'datasets/Facebook+Live+Sellers+in+Thailand',
 'datasets/Parkinson+Dataset+with+replicated+acoustic+features+',
 'datasets/Metro+Interstate+Traffic+Volume',
 'datasets/Query+Analytics+Workloads+Dataset',
 'datasets/Wave+Energy+Converters',
 'datasets/PPG-DaLiA',
 'datasets/Alcohol+QCM+Sensor+Dataset',
 'datasets/Divorce+Predictors+data+set',
 'datasets/Incident+management+process+enriched+event+log',
 'datasets/Opinion+Corpus+for+Lebanese+Arabic+Reviews+%28OCLAR%29',
 'datasets/MEx',
 'datasets/Beijing+Multi-Site+Air-Quality+Data',
 'datasets/Online+Retail+II',
 'datasets/Hepatitis+C+Virus+%28HCV%29+for+Egyptian+patients',
 'datasets/QSAR+fish+toxicity',
 'datasets/QSAR+aquatic+toxicity',
 'datasets/Human+Activity+Recognition+from+Continuous+Ambient+Sensor+Data',
 'datasets/WISDM+Smartphone+and+Smartwatch+Activity+and+Biometrics+Dataset+',
 'datasets/QSAR+oral+toxicity',
 'datasets/QSAR+androgen+receptor',
 'datasets/QSAR+Bioconcentration+classes+dataset',
```

```
 'datasets/QSAR+fish+bioconcentration+factor+%28BCF%29',
 'datasets/A+study+of++Asian+Religious+and+Biblical+Texts',
 'datasets/Real-time+Election+Results%3A+Portugal+2019',
 'datasets/Bias+correction+of+numerical+prediction+model+temperature+forecast',
 'datasets/Bar+Crawl%3A+Detecting+Heavy+Drinking',
 'datasets/Kitsune+Network+Attack+Dataset']
```

In [72]:
```
url_completa = []
for i in range(len(url)):
    pedazo = 'https://archive.ics.uci.edu/ml/'+url[i]
    url_completa.append(pedazo)
```

In [73]:
```
url_completa
```

```
 'https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%2
9',
 'https://archive.ics.uci.edu/ml/datasets/Quality+Assessment+of+Digital+Colpo
scopies',
 'https://archive.ics.uci.edu/ml/datasets/KASANDR',
 'https://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysi
s',
 'https://archive.ics.uci.edu/ml/datasets/Air+quality',
 'https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition',
 'https://archive.ics.uci.edu/ml/datasets/Devanagari+Handwritten+Character+Da
taset',
 'https://archive.ics.uci.edu/ml/datasets/Stock+portfolio+performance',
 'https://archive.ics.uci.edu/ml/datasets/MoCap+Hand+Postures',
 'https://archive.ics.uci.edu/ml/datasets/Early+biomarkers+of+Parkinson%92s+d
isease+based+on+natural+connected+speech',
 'https://archive.ics.uci.edu/ml/datasets/Data+for+Software+Engineering+Teamw
ork+Assessment+in+Education+Setting',
 'https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities',
 'https://archive.ics.uci.edu/ml/datasets/Parkinson+Disease+Spiral+Drawings+U
sing+Digitized+Graphics+Tablet',
```

In [74]:
```
len(url_completa)
```

Out[74]: 497

In [75]:
```
names = []
for i in range(len(nombres)):
    names.append(nombres[i][1])
names
```

Out[75]:
```
['Abalone',
 'Adult',
 'Annealing',
 'Anonymous Microsoft Web Data',
 'Arrhythmia',
 'Artificial Characters',
 'Audiology (Original)',
 'Audiology (Standardized)',
 'Auto MPG',
 'Automobile',
 'Badges',
 'Balance Scale',
 'Balloons',
 'Breast Cancer',
 'Breast Cancer Wisconsin (Original)',
 'Breast Cancer Wisconsin (Prognostic)',
 'Breast Cancer Wisconsin (Diagnostic)',
 'Pittsburgh Bridges',
 'Car Evaluation',
```

In [76]:
```
len(names)
```

Out[76]:
```
497
```

In [34]:
```
datos = soup.find_all("p")
datos
```

```
 <p class="normal">Categorical, Integer, Real </p>,
 <p class="normal">690 </p>,
 <p class="normal">15 </p>,
 <p class="normal"> </p>,
 <p class="normal"><b><a href="datasets/Japanese+Credit+Screening">Japanese C
redit Screening</a></b></p>,
 <p class="normal">Multivariate, Domain-Theory </p>,
 <p class="normal">Classification </p>,
 <p class="normal">Categorical, Real, Integer </p>,
 <p class="normal">125 </p>,
 <p class="normal"> </p>,
 <p class="normal">1992 </p>,
 <p class="normal"><b><a href="datasets/Computer+Hardware">Computer Hardware
</a></b></p>,
 <p class="normal">Multivariate </p>,
 <p class="normal">Regression </p>,
 <p class="normal">Integer </p>,
 <p class="normal">209 </p>,
 <p class="normal">9 </p>,
```

In [42]:
```
re.findall('class="normal">([\w\s,-]+)<', str(datos))
```

Out[42]:
```
['\n',
 'Multivariate\xa0',
 'Classification\xa0',
 'Categorical, Integer, Real\xa0',
 '4177\xa0',
 '8\xa0',
 '1995\xa0',
 'Multivariate\xa0',
 'Classification\xa0',
 'Categorical, Integer\xa0',
 '48842\xa0',
 '14\xa0',
 '1996\xa0',
 'Multivariate\xa0',
 'Classification\xa0',
 'Categorical, Integer, Real\xa0',
 '798\xa0',
 '38\xa0',
 '\xa0',
```

In [77]:
```
atributos = re.findall('class="normal">([\w\s,-]+)<', str(datos))
```

In [78]:
```
atributos.pop(0)
```

Out[78]:
```
'\n'
```

In [85]:
```
len(atributos)/6
# 1.Tipo de dato, tarea, atributos, variables, obs, año
```

Out[85]:
```
497.0
```

In [81]:
```
atributos
```

Out[81]:
```
['Multivariate\xa0',
 'Classification\xa0',
 'Categorical, Integer, Real\xa0',
 '4177\xa0',
 '8\xa0',
 '1995\xa0',
 'Multivariate\xa0',
 'Classification\xa0',
 'Categorical, Integer\xa0',
 '48842\xa0',
 '14\xa0',
 '1996\xa0',
 'Multivariate\xa0',
 'Classification\xa0',
 'Categorical, Integer, Real\xa0',
 '798\xa0',
 '38\xa0',
 '\xa0',
 '\xa0',
```

In [83]:
```
nuevo_atributos = []
for x in atributos:
    nuevo_atributos.append(x.replace("\xa0",""))
nuevo_atributos
```

```
 '2008',
 'Multivariate',
 'Regression',
 'Real',
 '1030',
 '9',
 '2007',
 'Sequential',
 'Classification',
 'Real',
 '606',
 '101',
 '2008',
 'Multivariate',
 'Classification',
 'Real',
 '900',
 '10000',
 '2008',
 'Multivariate'
```

In [103]:
```
elementos = list(range(0,2982,6))
data_types = []
for i in elementos:
    data_types.append(nuevo_atributos[i])
data_types
```

Out[103]:
```
['Multivariate',
 'Multivariate',
 'Multivariate',
 '',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Univariate, Text',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Multivariate',
 'Multivariate',
```

In [97]:
```
elementos = list(range(1,2982,6))
default_task = []
for i in elementos:
    default_task.append(nuevo_atributos[i])
default_task
```

```
 'Regression, Clustering',
 'Regression',
 'Regression',
 'Classification, Regression, Clustering',
 'Classification',
 'Regression, Clustering',
 'Classification',
 'Classification, Clustering',
 'Regression',
 'Classification, Regression, Clustering',
 'Classification',
 'Regression',
 'Regression',
 'Classification',
 'Classification',
 'Classification',
 'Classification',
 'Classification, Regression',
 'Regression',
 'Classification, Clustering',
```

In [98]:
```
elementos = list(range(2,2982,6))
attribute_types = []
for i in elementos:
    attribute_types.append(nuevo_atributos[i])
attribute_types
```

```
 'Integer, Real',
 'Real',
 'Real',
 'Real',
 'Real',
 'Integer',
 'Integer',
 'Integer',
 'Real',
 'Integer, Real',
 'Integer, Real',
 'Integer, Real',
 'Real',
 'Real',
 'Integer, Real',
 'Real',
 '',
 '',
 '',
```

In [99]:
```
elementos = list(range(3,2982,6))
obs = []
for i in elementos:
    obs.append(nuevo_atributos[i])
obs
```

```
 '260000',
 '288000',
 '8300000',
 '125',
 '170',
 '141712',
 '3916',
 '6262',
 '420768',
 '1067371',
 '1385',
 '908',
 '546',
 '13956534',
 '15630426',
 '8992',
 '1687',
 '779',
 '1056',
```

In [100]:
```
elementos = list(range(4,2982,6))
cols = []
for i in elementos:
    cols.append(nuevo_atributos[i])
cols
```

```
 '8',
 '49',
 '11',
 '8',
 '54',
 '36',
 '3916',
 '710',
 '18',
 '8',
 '29',
 '7',
 '9',
 '37',
 '6',
 '1024',
 '1024',
 '14',
 '7',
```

In [101]:
```
elementos = list(range(5,2982,6))
year = []
for i in elementos:
    year.append(nuevo_atributos[i])
year
```

```
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
 '2019',
```

In [104]:
```
table = {"Name": names,
        "Data types": data_types,
        "Default task": default_task,
        "Attributes types": attribute_types,
        "# Instances": obs,
        "# Attributes": cols,
        "Year": year,
        "Links": url_completa}
base_de_datos = pd.DataFrame(table)
```

In [105]:
```
base_de_datos
```

Out[105]:

	Name	Data types	Default task	Attributes types	# Instances	# Attributes	Year	
0	Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995	https://arc
1	Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996	https://
2	Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38		https://archi
3	Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998	https://archi
4	Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998	https://arc

In [112]:
```
base_de_datos["Name"] = base_de_datos["Name"].str.upper()
base_de_datos.sort_values(["Name"], axis=0,
                 ascending=[True], inplace=True)
base_de_datos
```

Out[112]:

	Name	Data types	Default task	Attributes types	# Instances	# Attributes	Year	
462	2.4 GHZ INDOOR CHANNEL MEASUREMENTS	Multivariate	Classification	Real	7840	5	2018	https://ar
237	3D ROAD NETWORK (NORTH JUTLAND, DENMARK)	Sequential, Text	Regression, Clustering	Real	434874	4	2013	https://ar
492	A STUDY OF ASIAN RELIGIOUS AND BIBLICAL TEXTS	Multivariate, Text	Classification, Clustering	Integer	590	8265	2019	https://a
301	AAAI 2013 ACCEPTED PAPERS	Multivariate	Clustering		150	5	2014	https://ar
294	AAAI 2014 ACCEPTED PAPERS	Multivariate	Clustering		399	6	2014	https://ar
...	...	...	...	...	...	...	...	
215	YOUTUBE COMEDY SLAM PREFERENCE DATA	Text	Classification		1138562	3	2012	https://ar
258	YOUTUBE MULTIVIEW VIDEO GAMES DATASET	Multivariate, Text	Classification, Clustering	Integer, Real	120000	1000000	2013	https://ar
364	YOUTUBE SPAM COLLECTION	Text	Classification		1956	5	2017	https://ar
396	Z-ALIZADEH SANI		Classification	Integer, Real	303	56	2017	https://
108	ZOO	Multivariate	Classification	Categorical, Integer	101	17	1990	htt

497 rows × 8 columns

Añadir resumen

In [109]:
```
link = requests.get('https://archive.ics.uci.edu/ml/datasets.php?format=&task=&at
s_link = BeautifulSoup(link, "lxml")
```

In [114]:
```
des = re.findall('<b><a href="datasets[\s\S]*?</p>',str(s_link))
des
:\s([\w\s:,.¿\'%()-]+)
```

```
of Wisdom, Book of Proverbs, Book of Ecclesiastes and Book of Ecclesiasticus
</p>',
 '<b><a href="datasets/AAAI+2013+Accepted+Papers">AAAI 2013 Accepted Papers</
a></b>: This data set compromises the metadata for the 2013 AAAI conference
\'s accepted papers (main track only), including paper titles, abstracts, and
keywords of varying granularity.</p>',
 '<b><a href="datasets/AAAI+2014+Accepted+Papers">AAAI 2014 Accepted Papers</
a></b>: This data set compromises the metadata for the 2014 AAAI conference
\'s accepted papers, including paper titles, authors, abstracts, and keywords
of varying granularity.</p>',
 '<b><a href="datasets/Abalone">Abalone</a></b>: Predict the age of abalone f
rom physical measurements</p>',
 '<b><a href="datasets/Abscisic+Acid+Signaling+Network">Abscisic Acid Signali
ng Network</a></b>: The objective is to determine the set of boolean rules th
at describe the interactions of the nodes within this plant signaling networ
k.  The dataset includes 300 separate boolean pseudodynamic simulations using
an asynchronous update scheme. </p>',
 '<b><a href="datasets/Absenteeism+at+work">Absenteeism at work</a></b>: The
database was created with records of absenteeism at work from July 2007 to Ju
ly 2010 at a courier company in Brazil.</p>',
```

In [115]:
```
abstract = []
for i in des:
    abstract.append(re.findall(":\s([\w\s:,.¿\'%()-]+)", i))
abstract
```

```
two and three-dimensional airfoil blade sections conducted in an anechoic win
d tunnel.'],
 ['Five different QCM gas sensors are used, and five different gas measuremen
ts (1-octanol, 1-propanol, 2-butanol, 2-propanol and 1-isobutanol) are conduc
ted in each of these sensors.'],
 ["Amazon's InfoSec is getting smarter about the way Access data is leverage
d. This is an anonymized sample of access provisioned within the company."],
 ['The dataset is used for authorship identification in online Writeprint whi
ch is a new research field of pattern recognition. '],
 ['Steel annealing data'],
 ['Log of anonymous users of www.microsoft.com'],
 ['Acoustic features extracted from syllables of anuran (frogs) calls, includ
ing the family, the genus, and the species labels (multilabel). '],
 ['Experimental data used to create regression models of appliances energy us
e in a low energy building.'],
 ["The datasets' positive class consists of component failures for a specific
component of the APS system. The negative class consists of trucks with failu
res for components not related to the APS."],
 ["ARCENE's task is to distinguish cancer versus normal patterns from mass-sp
ectrometric data. This is a two-class classification problem with continuous
```

In [116]:
```
len(abstract)
```

Out[116]: 497

In [118]:
```
resumen = []
for i in range(len(abstract)):
    resumen.append(abstract[i][0])
len(resumen)
```

Out[118]: 497

In [119]:
```
base_de_datos["Abstract"] = resumen
```

In [120]:
```
base_de_datos
```

Out[120]:

	Name	Data types	Default task	Attributes types	# Instances	# Attributes	Year	
462	2.4 GHZ INDOOR CHANNEL MEASUREMENTS	Multivariate	Classification	Real	7840	5	2018	https://ar
237	3D ROAD NETWORK (NORTH JUTLAND, DENMARK)	Sequential, Text	Regression, Clustering	Real	434874	4	2013	https://ar
492	A STUDY OF ASIAN RELIGIOUS AND BIBLICAL TEXTS	Multivariate, Text	Classification, Clustering	Integer	590	8265	2019	https://a
301	AAAI 2013 ACCEPTED PAPERS	Multivariate	Clustering		150	5	2014	https://ar
294	AAAI 2014 ACCEPTED PAPERS	Multivariate	Clustering		399	6	2014	https://ar
...	...	...	...	...	...	...	...	
215	YOUTUBE COMEDY SLAM PREFERENCE DATA	Text	Classification		1138562	3	2012	https://ar
258	YOUTUBE MULTIVIEW VIDEO GAMES DATASET	Multivariate, Text	Classification, Clustering	Integer, Real	120000	1000000	2013	https://ar
364	YOUTUBE SPAM COLLECTION	Text	Classification		1956	5	2017	https://ar
396	Z-ALIZADEH SANI		Classification	Integer, Real	303	56	2017	https://
108	ZOO	Multivariate	Classification	Categorical, Integer	101	17	1990	htt

497 rows × 9 columns