Métodos Computacionales para Políticas Públicas - URosario Entrega: viernes 6-nov-2020 11:59 PM **[Nicolás Garcés R]** [nicolas.garces@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: 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]: ob1 = "JEFF BEZOS, the founder of Amazon, has reached a divorce settlement with his wife, Ma cKenzie. Mr Bezos will keep all the shares in the Washington Post and Blue Origin, a space-e xploration firm, as well as 75% of the couple's Amazon stock. Mrs Bezos will retain a 4% sta ke in the tech giant, worth nearly \$36bn, which is likely to make her the third-richest woma n alive when the divorce is finalised." In [8]: import re re.findall("[0-9]+",ob1) Out[8]: ['75', '4', '36'] 2. [1 punto] Usando expresiones regulares ahora extraiga de ob1 sólo los números que correspondan a porcentajes. In [12]: re.findall("([0-9]+)%",ob1) Out[12]: ['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 [1]: def dater(x): import re y = re.findall("([A-Z][a-zA-Z]*)",x)**if** y[0] =="Enero": yy= "01" elif y[0] =="Febrero": yy= "02" elif y[0] =="Marzo": yy= "03" elif y[0] =="Abril": yy= "04" **elif** y[0] =="Mayo": yy= "05" elif y[0] =="Junio": yy= "06" elif y[0] =="Julio": yy= "07" elif y[0] =="Agosto": yy= "08" elif y[0] =="Septiembre": yy= "09" elif y[0] =="Octubre": yy= **"10"** elif y[0] =="Noviembre": yy= **"11"** elif y[0] =="Diciembre": yy= **"12"** else: yy="00" z = re.findall("[0-9]+",x)zz = z[0]**if** len(zz) <= 1: zz= "0"+zz else: zz = zzzzz=z[1]re= zzz + "-" + zz + "-"+ yy **return** re In [3]: x = "Septiembre 25, 2019"dater(x)Out[3]: '2019-25-09' In [4]: x = "Marzo 7, 2019"dater(x) Out[4]: '2019-07-03' 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 [24]: ob2 = "COMPSCI 270 (Spring 2019): Introduction to Artificial Intelligence. COMPSCI 590.2 (Fa ll 2018): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. CO MPSCI 223 (Spring 2018): Computational Microeconomics. COMPSCI 570 (Fall 2017): Artificial I ntelligence. 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 (Spri ng 2016): Computational Microeconomics: Game Theory, Social Choice, and Mechanism Design. CO MPSCI 290.4/590.4 (Spring 2015): Crowdsourcing Societal Tradeoffs. COMPSCI 570 (Fall 2014): Artificial Intelligence. COMPSCI 590.4 (Spring 2014): Computational Microeconomics: Game Th eory, Social Choice, and Mechanism Design. COMPSCI 590.1 (Fall 2012): Linear and Integer Pro gramming. COMPSCI 173 (Spring 2012): Computational Microeconomics. COMPSCI 296.1 (Fall 201 1): 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 2 008): Linear and Integer Programming. COMPSCI 196.2 (Fall 2007): Introduction to Computation al Economics. COMPSCI 296.3 (Spring 2007): Topics in Computational Economics. COMPSCI 296.2 (Fall 2006): Computational Game Theory and Mechanism Design." In [26]: import re y = re.findall("[A-Z] ([0-9.]*) ", ob2)x = re.findall("([0-9.]+)/", ob2)z = re.findall("/([0-9.]+)", ob2)final = x + y + zfinal Out[26]: ['290.4', '196.1', '196', '270', '590.2', '223', 570' '590.3' '590.2' '590.4', '570', 590.4 '590.1', '173', '296.1' '296.1', '173', '170', '270', '196.2', '296.3', '296.2' '590.4' '296.1', '296.2'] 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 [6]: 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 po litical 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, M erhout, 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): eaao5 948-null. Full Text Open Access Copy.\n", "Bail, CA, Brown, TW, and Mann, M. "Channeling Hea rts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversatio n." American Sociological Review 82.6 (December 1, 2017): 1188-1213. Full Text.\n", "Bail, C A. "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, 20 17): 1-14. Full Text.\n", "Bail, CA. "Combining natural language processing and network anal ysis to examine how advocacy organizations stimulate conversation on social media." Proceedi ngs of the National Academy of Sciences of the United States of America 113.42 (October 201 6): 11823-11828. Full Text.\n", "Bail, CA. "Emotional Feedback and the Viral Spread of Socia 1 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, d isclosure, 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 wi th big data." Theory and Society 43.3 (January 1, 2014): 465-524. Full Text.""'' In [7]: $re.findall('''." ([A-Z \setminus s a-z]+) [0-9]*''', ob3)$ Out[7]: ['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 559 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) In [1]: import requests from bs4 import BeautifulSoup html = requests.get('https://archive.ics.uci.edu/ml/datasets.php').text page = BeautifulSoup(html) In [2]: items = page.find_all('p') In [3]: import re data= re.findall('class="normal".+', str(items)) In [4]: nombres = re.findall('''>([\w'' "0-9/?^.+\,;*\$:!@()_+-&-<>/]*)</table >''', str(page)) In [5]: len(nombres) Out[5]: 559 In [11]: other= re.findall('class="normal">([A-Z \s a-z,-]*)', str(data)) other= other[9:] len(other) # de 3 en 3 Out[11]: 3914 In [7]: numbers = re.findall('class="normal">([0-9]*)', str(data)) numbers = numbers[12:] In [8]: data_types =[] iterable = [] **for** i **in** range(3914): **if** i % 7 == 0: iterable.append(i) **for** i **in** iterable: data_types.append(other[i]) data_types=data_types[:559] In [124]: |len(data_types) Out[124]: 559 In [9]: | default_task=[] iterablee = [] **for** i **in** range(3914): **if** i % 7 == 0: iterablee.append(i+1) iterablee.remove(3914) **for** i **in** iterablee: default_task.append(other[i]) In [126]: len(default_task) Out[126]: 559 In [12]: | attribute_types=[] iterableee = [] for i in range(3914): **if** i % 7 == 0: iterableee.append(i+2) iterableee.remove(3915) for i in iterableee: attribute_types.append(other[i]) In [13]: len(attribute_types) Out[13]: 559 In [14]: | num_instances =[] iiterable = [] **for** i **in** range(3911): **if** i % 7 == 0: iiterable.append(i) **for** i **in** iiterable: num_instances.append(numbers[i]) In [15]: len(num_instances) Out[15]: 559 In [16]: | num_attributes=[] iiterablee = [] **for** i **in** range(3911): **if** i % 7 == 0: iiterablee.append(i+1) **for** i **in** iiterablee: num_attributes.append(numbers[i]) In [17]: len(num_attributes) Out[17]: 559 In [18]: | year=[] iiterableee = [] **for** i **in** range(3911): **if** i % 7 == 0: iiterableee.append(i+2) **for** i **in** iiterableee: year.append(numbers[i]) In [19]: len(year) Out[19]: 559 In [20]: $| links = re.findall("<a href=\"([/\%=\w'+():. ,-]*)\"", str(page))$ link=[] for i in links: link.append("https://archive.ics.uci.edu/ml/"+i) In [21]: len(link) Out[21]: 559 In [138]: import pandas as pd dtafrm = {'Data Types':data_types,'Default Task':default_task,'Attribute Types':attribute_ty 'Num Instances':num_instances,'Num Attributes':num_attributes, 'Year':year,'Link': link} halfdf = pd.DataFrame(dtafrm,index =nombres) halfdf Out[138]: Attribute Data Num Num **Default Task** Year Link instances Attributes Categorical, 4177 **Abalone** Multivariate Classification 8 1995 https://archive.ics.uci.edu/ml/datasets/Abalone Integer, Categorical, Classification 48842 14 1996 **Adult** Multivariate https://archive.ics.uci.edu/ml/datasets/Adult Categorical, **Annealing** Multivariate Classification Integer, 798 38 https://archive.ics.uci.edu/ml/datasets/Annealing Real **Anonymous** Recommender-**Microsoft Web** Categorical 37711 294 1998 https://archive.ics.uci.edu/ml/datasets/Anonym... Systems Data Categorical, Arrhythmia Multivariate Classification Integer, 452 279 1998 https://archive.ics.uci.edu/ml/datasets/Arrhyt... Real IIWA14-R820-Gazebohttps://archive.ics.uci.edu/ml/datasets/IIWA14... Regression Integer **Dataset-**10Trajectories **Guitar Chords** Classification 2633 5 2020 https://archive.ics.uci.edu/ml/datasets/Guitar... finger positions Russian Corpus of 200 Text Classification https://archive.ics.uci.edu/ml/datasets/Russia... **Biographical Texts** Classification, 13028 Codon usage Multivariate 69 2020 https://archive.ics.uci.edu/ml/datasets/Codon+... Clustering Intelligent Media **Accelerometer** Time-800 Classification Real 9 2020 https://archive.ics.uci.edu/ml/datasets/Intell.. and Series Gyroscope (IM-AccGyro) **Dataset** 559 rows × 7 columns In [23]: #description import requests from bs4 import BeautifulSoup des= requests.get('https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numA tt=&numIns=&type=&sort=nameUp&view=list').text description= BeautifulSoup(des) In [24]: import requests descrip = re.findall(''': ([\w'' "0-9/?'^.%+@\,;*\$:!@#(){}_+-&-<>/~\±]*)''', st r(description)) descrip.insert(10, 'This dataset comprises information regarding the ADLs performed by two us ers on a daily basis in their own homes.') descrip.insert(15,'The data was created by a medical expert as a data set to test the expert system, which will perform the presumptive diagnosis of two diseases of the urinary system.' descrip.insert(28, 'This is a dataset of classified for apartments for rent in USA.') descrip.insert(46, 'The data set is composed of 60 chorales (5665 events) by J.S. Bach (1675-1750). Each event of each chorale is labelled using 1 among 101 chord labels and described t hrough 14 features.') descrip.insert(61,"It contains fourteen numerical weather prediction (NWP)'s meteorological forecast data, two in-situ observations, and five geographical auxiliary variables over Seo ul, South Korea in the summer.") descrip.insert(66,'In this paper, we look for to recognize the causes of users tend to cyber space in Kohkiloye and Boyer Ahmad Province in Iran') descrip.insert(97, 'ChIP-seq experiments characterize protein modifications or binding at spe cific genomic locations in specific samples. The machine learning problem in these data is s tructured binary classification.') descrip.insert(104, 'This is a data set containing 1080 documents of free text business descr iptions of Brazilian companies categorized into a subset of 9 categories') descrip.insert(133,'The dataset comprises motion sensor data of 19 daily and sports activiti es each performed by 8 subjects in their own style for 5 minutes. Five Xsens MTx units are u sed on the torso, arms, and legs.') descrip.insert(135, "This dataset contains the annotated readings of 3 acceleration sensors a t the hip and leg of Parkinson's disease patients that experience freezing of gait (FoG) dur ing walking tasks.") descrip.insert(150, 'DEXTER is a text classification problem in a bag-of-word representation. This is a two-class classification problem with sparse continuous input variables. This data set is one of five datasets of the NIPS 2003 feature selection challenge.') descrip.insert(153, 'This data has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes.') descrip.insert(155, 'Discrete Tone Images(DTI) are available which needs to be analyzed in det ail. Here, we created this dataset for those who do research in DTI. ') descrip.insert(157, 'Participants completed the "Personal Information Form" and "Divorce Pred ictors Scale"') descrip.insert(165,'The DrivFace contains images sequences of subjects while driving in real scenarios. It is composed of 606 samples of 640×480, acquired over different days from 4 dri vers with several facial features.') descrip.insert(174, 'This dataset contains the sign and symptpom data of newly diabetic or wo uld be diabetic patient.') descrip.insert(186,'This data set contains electricity consumption of 370 points/clients.') descrip.insert(214, 'A chemical detection platform composed of 14 temperature-modulated metal oxide (MOX) gas sensors was exposed during 3 weeks to mixtures of carbon monoxide and humid synthetic air in a gas chamber.') descrip.insert(218,'100 recordings of a sensor array under different conditions in a home se tting: background, wine and banana presentations. The array includes 8 MOX gas sensors, and humidity and temperature sensors.') descrip.insert(223, 'Instances in this dataset contain audio features extracted from 1059 wav e files. The task associated with the data is to predict the geographical origin of music. descrip.insert(225, 'GISETTE is a handwritten digit recognition problem. The problem is to se parate the highly confusible digits 4 and 9. This dataset is one of five datasets of the NIP S 2003 feature selection challenge.') descrip.insert(229, 'The dataset has been feed by Android app called Go!Track. It is availabl e at Goolge Play Store(https://play.google.com/store/apps/details?id=com.go.router).') descrip.insert(231, 'Design an observing network to monitor emissions of a greenhouse gas (GH G) in California given time series of synthetic observations and tracers from weather model simulations.') descrip.insert(265,'The IM-AccGyro dataset is devised to benchmark techniques dealing with h uman activity recognition based on inertial sensors') descrip.insert(292, 'The data set contains 15 classes of 24 instances each. Each class refere nces to a hand movement type in LIBRAS (Portuguese name Lngua BRAsileira de Sinais, oficial brazilian signal language).') descrip.insert(307,'Mesothelioma's disease data set were prepared at Dicle University Facult y of Medicine in Turkey. Three hundred and twenty-four Mesothelioma patient data. In the dat aset, all samples have 34 features.') descrip.insert(311, 'The MEx Multi-modal Exercise dataset contains data of 7 different physio therapy exercises, performed by 30 subjects recorded with 2 accelerometers, a pressure mat a nd a depth camera.') descrip.insert(314, 'MicroblogPCU data is crawled from sina weibo microblog[http://weibo.co m/]. This data can be used to study machine learning methods as well as do some social netwo rk research.') descrip.insert(341, 'Nomao collects data about places (name, phone, localization...) from man y sources. Deduplication consists in detecting what data refer to the same place. Instances in the dataset compare 2 spots.') descrip.insert(349, 'This is a dataset of 8235 online handwritten assamese characters. The "o nline" process involves capturing of data as text is written on a digitizing tablet with an electronic pen.') descrip.insert(353,'0f the 12,330 sessions in the dataset, 84.5% (10,422) were negative clas s samples that did not end with shopping, and the rest (1908) were positive class samples en ding with shopping.') descrip.insert(357, 'This dataset contains sentences extracted from user reviews on a given t opic. Example topics are "performance of Toyota Camry" and "sound quality of ipod nano".') descrip.insert(358, 'This data set contains user reviews of cars and and hotels collected fro m Tripadvisor (~259,000 reviews) and Edmunds (~42,230 reviews).') descrip.insert(364, 'The goal is to model mutant p53 transcriptional activity (active vs inac tive) based on data extracted from biophysical simulations.') descrip.insert(369,'Data collected from car parks in Birmingham that are operated by NCP fro m Birmingham City Council. UK Open Government Licence (OGL). https://data.birmingham.gov.uk/ dataset/birmingham-parking') descrip.insert(393, 'PPG-DaLiA contains data from 15 subjects wearing physiological and motio n sensors, providing a PPG dataset for motion compensation and heart rate estimation in Dail y Life Activities.') descrip.insert(413, 'The "real estate valuation" is a regression problem. The market historic al data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiw descrip.insert(423, 'This dataset is created by reading out 200 files from the 10 largest Reu ters classes and using an Automatic Speech Recognition system to create corresponding transc descrip.insert(436,'The data describe the problem of high energy (higher than 10^4 J) seismi c bumps forecasting in a coal mine. Data come from two of longwalls located in a Polish coal descrip.insert(437, 'The SELFBACK dataset is a Human Activity Recognition Dataset of 9 activi ty classes recorded with two tri-axial accelerometers.') descrip.insert(455, 'Activity recognition data set built from the recordings of 30 subjects p erforming basic activities and postural transitions while carrying a waist-mounted smartphon e with embedded inertial sensors. ') descrip.insert(479,'A dataset of steel plates' faults, classified into 7 different types. Th e goal was to train machine learning for automatic pattern recognition.') descrip.insert(480, 'The dataset is provided by the "Trialto Latvia LTD", the third-party log istics operator. Each observation stands for a distinct type of item for sale.') descrip.insert(482,'Stone flakes are waste products of the stone tool production in the preh istoric era. The variables are means of geometric and stylistic features of the flakes conta ined in different inventories.') descrip.insert(494, 'An accurate dataset describing trajectories performed by all the 442 tax is running in the city of Porto, in Portugal.') descrip.insert(509, 'UbiqLog is the smartphone lifelogging tool that runs on the smartphone o f 35 users for about 2 months.') descrip.insert(524,' The dataset collects data from an Android smartphone positioned in the chest pocket from 22 participants walking in the wild over a predefined path.') descrip.insert(540,'') descrip.insert(547, 'Two datasets are included, related to red and white vinho verde wine sam ples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], http://www3.dsi.uminho.pt/pcortez/wine/).') descrip.insert(554,' The datasets are taken from top 2 Indian cooking channel named Nisha Ma dhulika channel and Kabita's Kitchen channel. The data set is in Hinglish Language.') # no supe como mas hacerlo :<</pre> In [141]: len(descrip) Out[141]: 559 In [25]: |bb= re.findall('''">([\w'' "0-9/?^.f\,;*\$:!@()_+-&-<>/]*)''', str(description)) # names en orden In [26]: len(bb) Out[26]: 559 In [144]: import pandas as pd dframe = {'Description':descrip} ohalfdf = pd.DataFrame(dframe,index =bb) ohalfdf Out[144]: Description 2.4 GHZ Indoor Channel Measurements Measurement of the S21, consists of 10 sweeps, ... 3D Road Network (North Jutland, Denmark) 3D road network with highly accurate elevation... 3W dataset The first realistic and public dataset with ra... : Simulated Data set of Iraqi tourism places Simulated Data set of Iraqi tourism places wit... A study of Asian Religious and Biblical Texts Mainly from Project Gutenberg, we combine Upan... Youtube cookery channels viewers comments in Hinglish The datasets are taken from top 2 Indian cook... YouTube Multiview Video Games Dataset This dataset contains about 120k instances, ea... It is a public set of comments collected for s... YouTube Spam Collection It was collected for CAD diagnosis. Z-Alizadeh Sani Artificial, 7 classes of animals Zoo 559 rows × 1 columns In [152]: final_data= pd.merge(halfdf,ohalfdf,how='inner',left_index=True, right_index=True) final_data Out[152]: Num Data Types Default Task Link Year Types Instances Attributes 2.4 GHZ Indoor Multivariate Classification 5 2018 https://archive.ics.uci.edu/ml/datasets/2.4+GH... Real 7840 Channel 0 Measurements 3D Road Network Sequential, Regression, 434874 4 2013 https://archive.ics.uci.edu/ml/datasets/3D+Roa... (North Real Text Clustering Jutland, Denmark) Multivariate, Classification, Integer, 3W dataset 1984 8 2019 https://archive.ics.uci.edu/ml/datasets/3W+dat... Time-Series Clustering Real : Simulated Data set of Classification, Multivariate 16 2020 https://archive.ics.uci.edu/ml/datasets/%3A+Si... 232 Iraqi tourism places A study of Asian Multivariate. Classification 590 https://archive.ics.uci.edu/ml/datasets/A+stud... Integer

Religious and

Biblical Texts

seismic-

Text

seeds Multivariate

Clustering

Classification.

Clustering

Real

210

7 2012

https://archive.ics.uci.edu/ml/datasets/seeds

Taller 8

