

UNIVERSIDAD NACIONAL DE INGENIERIA

Facultad de ingeniería industrial y de sistemas



**MANUAL DE FUNCIONAMIENTO Y EJECUCIÓN DE BOT
EN PYTHON CON LA LIBRERÍA SELENIUM**

INTEGRANTES:

LIMAQUISPE HUAMAN MIGUEL ANGEL

MILLER CONYAS JORGE

ORELLANA UGAZ JULIO

CURSO:

ADMINISTRACIÓN DEL CONOCIMIENTO

DOCENTE:

VILCAPOMA ESCURRA, EDGAR

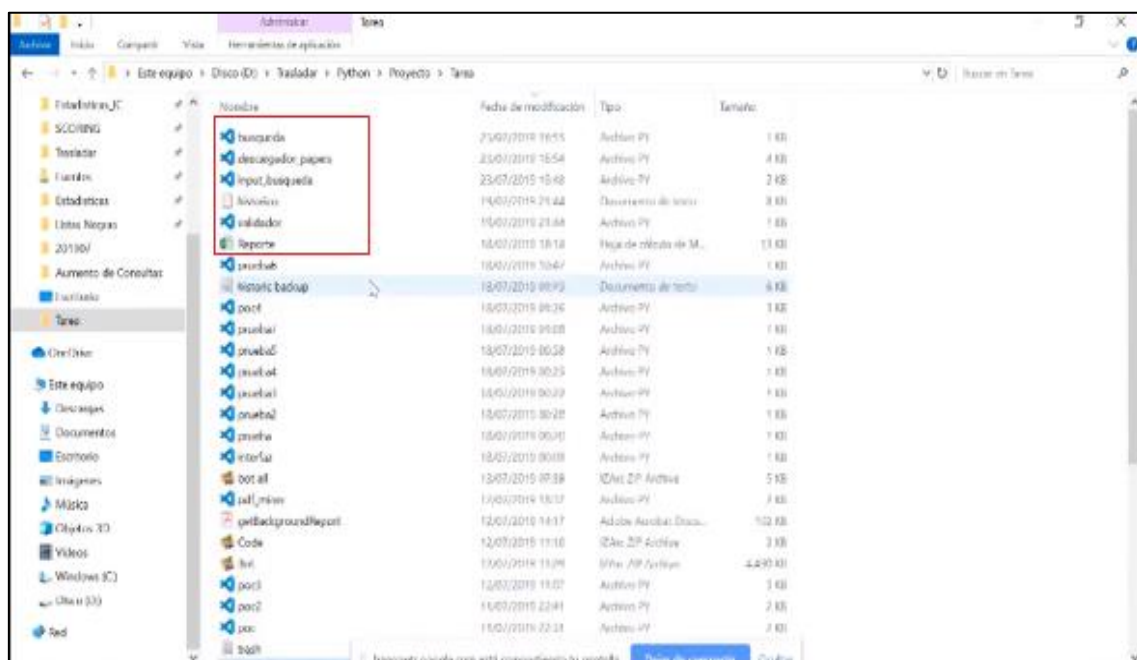
2019-I

MANUAL DE FUNCIONAMIENTO Y EJECUCIÓN DE BOT EN PYTHON

Funcionamiento

El programa ha sido creado en Python, para esto se utilizó principalmente las librerías SELENIUM para la automatización y PYPDF2 para la captura de metadatos de los archivos pdf descargados.

El programa se creó y codificó en distintos módulos que veremos a continuación.




Acá se muestra los archivos .py utilizados para el funcionamiento del programa:

- **Input_busqueda**; este archivo es la interfaz que captura los parámetros de búsqueda. Es modificable según los parámetros que sean conveniente en la búsqueda de los repositorios.
- **Búsqueda y validador**; estos archivos son necesarios para almacenar los parámetros de búsqueda e iniciar la búsqueda.
- **Descargador_papers**; tal y como lo dice su nombre contiene el código para la descarga de los archivos.

Otros:

- **Historico**; archivo que registra historial y datos de los archivos descargados.
- **Reporte**; archivo Excel que se actualiza a través del histórico, en el cual se ordena y almacena los registros deseados de los archivos descargados.

Nota: Es necesario tener descargado el siguiente ejecutable.

 chromedriver

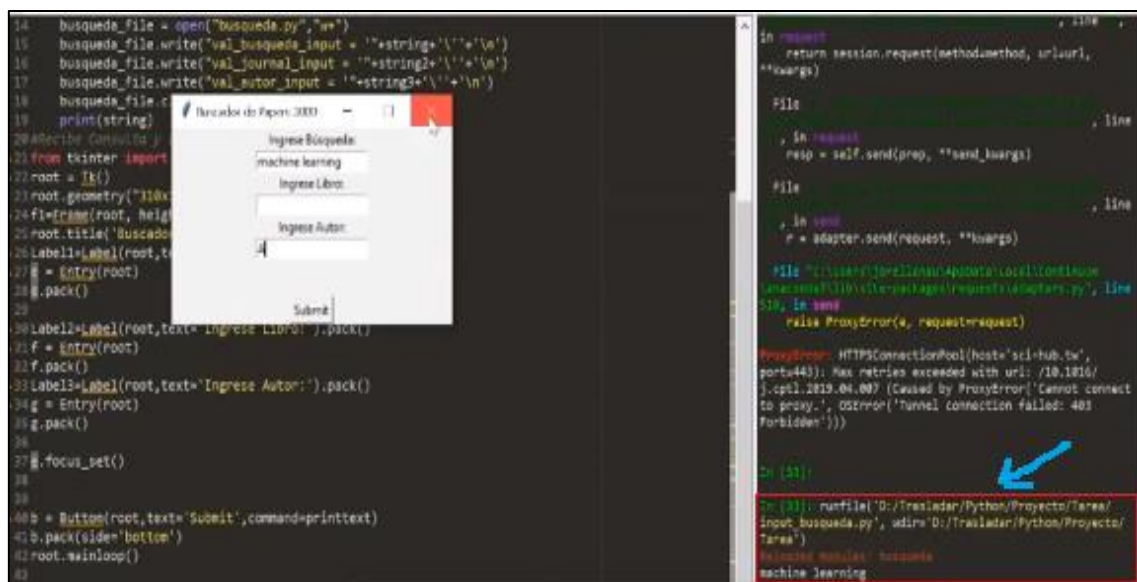
9/05/2019 02:32

Link:

http://chromedriver.chromium.org/downloads?fbclid=IwAR3T1DQNOgVnpJlMGEr6u_CGZr-0CgMm8TVdegZA-rlZUBruCONWg9gaEXY

Ejecución

Se explicará brevemente como se ejecuta la búsqueda de acorde a los archivos utilizados y los resultados de la búsqueda.

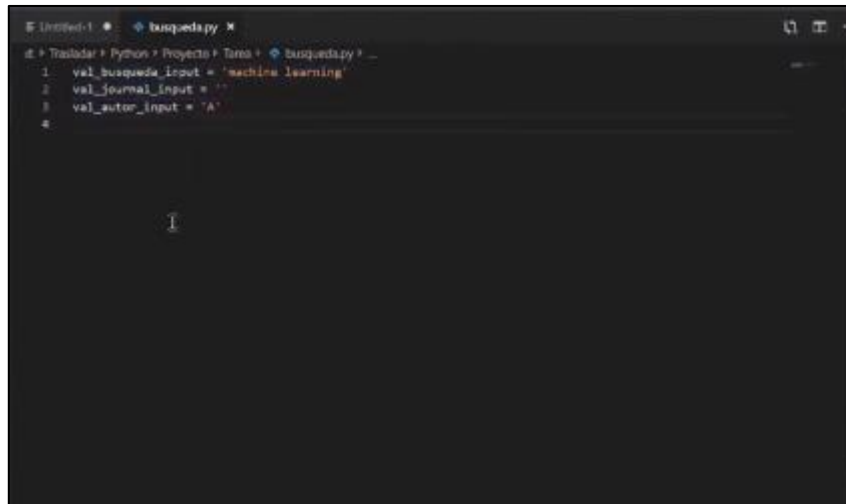


```
14 busqueda_file = open("busqueda.py", "w")
15 busqueda_file.write("val_busqueda_input = '"+string+"'\n")
16 busqueda_file.write("val_journal_input = '"+string2+"'\n")
17 busqueda_file.write("val_autor_input = '"+string3+"'\n")
18 busqueda_file.close()
19 print(string)
20 #Recibe Consulta
21 from tkinter import *
22 root = Tk()
23 root.geometry("1100x600")
24 fl=Frame(root, height=500)
25 root.title('Buscador de Papers 3000')
26 Label1=Label(root, text='Ingrese Búsqueda:')
27 e1 = Entry(root)
28 e1.pack()
29 Label2=Label(root, text='Ingrese Libro:')
30 e2 = Entry(root)
31 e2.pack()
32 Label3=Label(root, text='Ingrese Autor:')
33 e3 = Entry(root)
34 e3.pack()
35 g.pack()
36
37 g.focus_set()
38
39 b = Button(root, text='Submit', command=printtext)
40 b.pack(side='bottom')
41 root.mainloop()
42
```

```
In request
return session.request(method=method, url=url,
**kwargs)
File
, in request
resp = self.send(prepared_request, **kwargs)
File
, in send
r = adapter.send(request, **kwargs)
File "C:\Users\jovellana\AppData\Local\Continuum\anaconda\lib\site-packages\requests\adapters.py", line
510, in send
raise ProxyError(e, request=request)
ProxyError: HTTPConnectionPool(host='sci-hub.tw',
port=443): Max retries exceeded with url: /10.1016/
j.cpl.2019.04.007 (caused by ProxyError[Cannot connect
to proxy.', OSError('Tunnel connection failed: 403
Forbidden')])
In [33]:
python input_busqueda.py, where D:/Traslador/Python/Proyecto/
Tarea1
Welcome Modules: busqueda
machine learning
```

En la imagen superior vemos una pequeña interfaz para la captura de los parámetros de búsqueda; en el ejemplo vemos que se llena los datos de búsqueda de tema y autor (inicial o parte de una cadena de referencia para

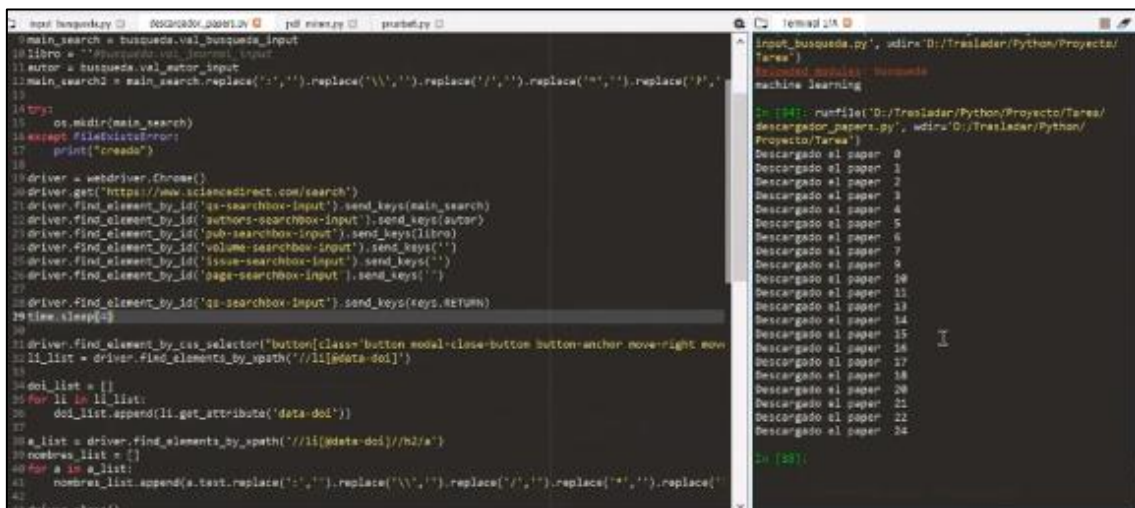
filtrar por nombre), se ingresa “machine learning” y la inicial ‘A’ para luego presionar Submit. Inmediatamente en la parte inferior derecha vemos que el primer paso es almacenar los parámetros de búsqueda.



```
1 val_busqueda_input = 'machine learning'
2 val_journal_input = ''
3 val_autor_input = 'A'
4
```

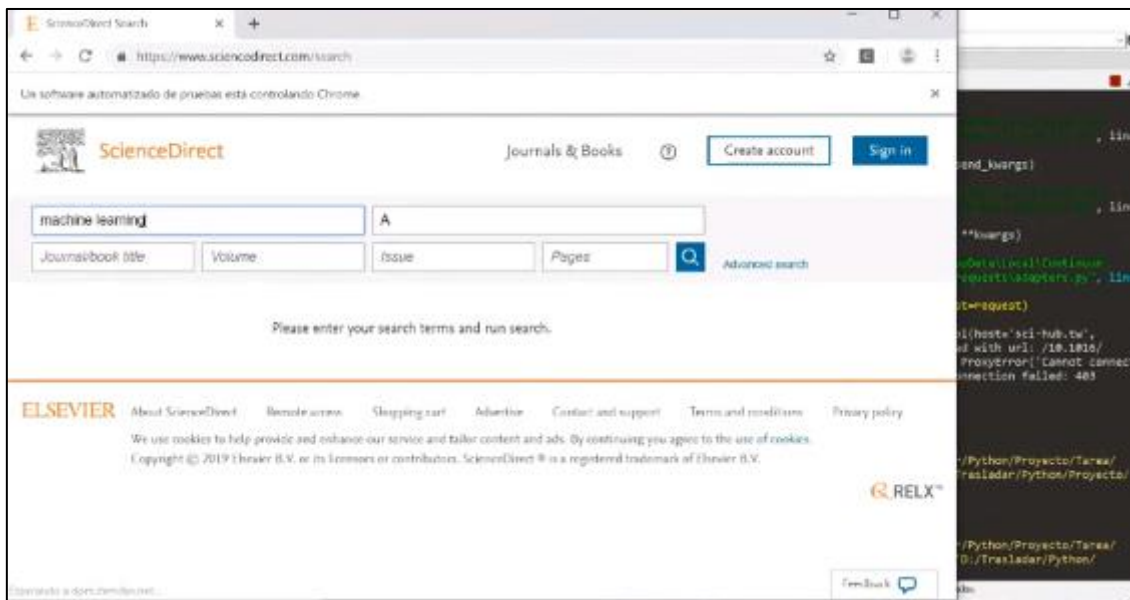
En la imagen superior vemos que el archivo “busqueda.py” ha capturado los datos ingresados.

A continuación, en la imagen inferior vemos el archivo “descargador_papers.py” el cual primero valida los datos de búsqueda según los datos capturados por el archivo “busqueda.py”. Luego inicia la búsqueda en el repositorio web automáticamente, para finalmente iniciar el proceso de descarga el cual vemos en la parte derecha de la imagen.



```
1 main_search = busqueda.val_busqueda_input
2 libro = busqueda.val_journal_input
3 autor = busqueda.val_autor_input
4 main_search2 = main_search.replace(' ', '').replace('\\', '').replace('/', '').replace(':', '').replace('.', '')
5
6 try:
7     os.mkdir(main_search)
8 except FileExistsError:
9     print("creada")
10
11 driver = webdriver.Chrome()
12 driver.get("https://www.sciencedirect.com/search")
13 driver.find_element_by_id('qs-searchbox-input').send_keys(main_search)
14 driver.find_element_by_id('authors-searchbox-input').send_keys(autor)
15 driver.find_element_by_id('pub-searchbox-input').send_keys(libro)
16 driver.find_element_by_id('volume-searchbox-input').send_keys('')
17 driver.find_element_by_id('issue-searchbox-input').send_keys('')
18 driver.find_element_by_id('page-searchbox-input').send_keys('')
19
20 driver.find_element_by_id('qs-searchbox-input').send_keys(keys.RETURN)
21 time.sleep(2)
22 driver.find_element_by_css_selector("button[class='button modal-close-button button-anchor move-right now']").click()
23 li_list = driver.find_elements_by_xpath("//li[@data-doi]")
24
25 doi_list = []
26 for li in li_list:
27     doi_list.append(li.get_attribute('data-doi'))
28
29 a_list = driver.find_elements_by_xpath("//li[@data-doi]//h2/a")
30 nombres_list = []
31 for a in a_list:
32     nombres_list.append(a.text.replace(':', '').replace('\\', '').replace('/', '').replace(':', '').replace('.', ''))
33
34 driver.close()
```

```
1 import busqueda.py, os
2
3 def main():
4     runfile('D:/Traslador/Python/Proyecto/Tareas/
5     descargador_papers.py', wdir='D:/Traslador/Python/
6     Proyecto/Tareas')
7
8 if __name__ == '__main__':
9     main()
10
11 Descargado el paper 0
12 Descargado el paper 1
13 Descargado el paper 2
14 Descargado el paper 3
15 Descargado el paper 4
16 Descargado el paper 5
17 Descargado el paper 6
18 Descargado el paper 7
19 Descargado el paper 8
20 Descargado el paper 9
21 Descargado el paper 10
22 Descargado el paper 11
23 Descargado el paper 12
24 Descargado el paper 13
25 Descargado el paper 14
26 Descargado el paper 15
27 Descargado el paper 16
28 Descargado el paper 17
29 Descargado el paper 18
30 Descargado el paper 19
31 Descargado el paper 20
32 Descargado el paper 21
33 Descargado el paper 22
34 Descargado el paper 23
35 Descargado el paper 24
```



En caso error de descarga en algún archivo el “descargador_paper.py” lo captura para luego saltarse ese error y continuar con la descarga. Ver imagen inferior.

```

44
45
46 num=0
47 flag_descargado = []
48 for i, doi in enumerate(doi_list):
49     r = requests.get('https://sci-hub.tw/' + doi)
50     html_soup = BeautifulSoup(r.text, 'html.parser')
51     try:
52         pdf_link = html_soup.find(id='article') iframe.get('src').split('view')[0]
53         pdf_link = 'http://'+ pdf_link
54         pdf_link = 'http://'+ pdf_link
55         with open(main_search2+'nombres_lista[i]+'.pdf', 'wb') as f:
56             f.write(requests.get(pdf_link).content)
57             f.close()
58         # write(requests.get(pdf_link).content)
59         print("Descargado el paper ", str(i))
60         flag_descargado.append('downloaded')
61         num=num+1
62     except (AttributeError, ConnectionError):
63         flag_descargado.append('failed')
64     except UnicodeDecodeError:
65         flag_descargado.append('failed')
66
67
68 historic_file = open("historico.txt", "a")
69 historic_file.write(main_search2+" LISTA DE PAPIER: "+'\t'+STATUS+"\n")
70 for i, k in enumerate(nombres_lista):
71     historic_file.write(k+"\t"+flag_descargado[i]+"\n")
72
73 historic_file.write("FIN DE BUSQUEDA "+main_search2+"\t"+str(num)+" PAPIER DESCARGADOS"+'\n\n')
74 historic_file.close()
75
76 validation_file = open("validador.py", "w")
77 validation_file.write("val_busqueda = '"+main_search2+"'")
78 validation_file.close()

```

Resultados

Primero vemos el archivo histórico que registra los datos de los papers descargados para luego generar un Excel con los registros almacenados.

historico: Bloc de notas					
Archivo Edición Formato Ver Ayuda					
machine learning LISTA DE PAPERS: STATUS FECHA DOI					
Evaluating machine learning performance in predicting injury severity in agribusiness industries	downloaded	2019/07/25, 11:24:10	10.1016/j.ssci.2019.04.		
Predictors of in-hospital length of stay among cardiac patients A machine learning approach	downloaded	2019/07/25, 11:24:10	10.1016/j.ijcard.2019.01.046		
Machine learning algorithms for predicting scapular kinematics	downloaded	2019/07/25, 11:24:10	10.1016/j.medengphy.2019.01.005		
Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches	downloaded	2019/07/25, 11:24:10	10.1016/j.iot.2019.100059		
Machine learning for ecosystem services	downloaded	2019/07/25, 11:24:10	10.1016/j.ecoser.2018.04.004		
Machine Learning for Sustainable Structures A Call for Data	downloaded	2019/07/25, 11:24:10	10.1016/j.istruc.2018.11.013		
A comparison of machine learning techniques for file system forensics analysis	downloaded	2019/07/25, 11:24:10	10.1016/j.jisa.2019.02.009		
Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty A Review	downloaded	2019/07/25, 11:24:10	10.1016/j.arth.2019.05.055		
Machine learning-assisted early ignition prediction in a complex flow	downloaded	2019/07/25, 11:24:10	10.1016/j.combustflame.2019.05.014		
Development of machine learning algorithms for prediction of mortality in spinal epidural abscess	downloaded	2019/07/25, 11:24:10	10.1016/j.spinee.2019.0		
Machine learning core inflation	downloaded	2019/07/25, 11:24:10	10.1016/j.econlet.2018.05.001		
Identifying psychosis spectrum disorder from experience sampling data using machine learning approaches	downloaded	2019/07/25, 11:24:10	10.1016/j.schres.2019.0		
Machine Learning for Perovskites' Reap-Rest-Recovery Cycle	downloaded	2019/07/25, 11:24:10	10.1016/j.joule.2018.11.010		
Can machine learning predict resectability of a peritoneal carcinomatosis	downloaded	2019/07/25, 11:24:10	10.1016/j.suronc.2019.04.008		
A machine learning classifier for microlensing in wide-field surveys	downloaded	2019/07/25, 11:24:10	10.1016/j.ascom.2019.100298		
A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action	downloaded	2019/07/25, 11:24:10	10.1016/j.cell.2019.04.016		
Statistical downscaling of precipitation using machine learning techniques	downloaded	2019/07/25, 11:24:10	10.1016/j.atmosres.2018.05.022		
Internet of Things A survey on machine learning-based intrusion detection approaches	downloaded	2019/07/25, 11:24:10	10.1016/j.comnet.2019.01.023		
Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning	downloaded	2019/07/25, 11:24:10	10.1016/j.matt.2019.03.002		
Big Data Analysis and Machine Learning in Intensive Care Units	downloaded	2019/07/25, 11:24:10	10.1016/j.medic.2019.06.012		
Finding the right fuel for the analytical engine Expanding the leader trait paradigm through machine learning	downloaded	2019/07/25, 11:24:10	10.1016/j.leaqua.2019.05.005		
Controlling of optical fiber bending losses through 'WARN' parameter and machine learning direction at three communication windows	downloaded	2019/07/25, 11:24:10			
Prediction of tidal currents using Bayesian machine learning	downloaded	2019/07/25, 11:24:10	10.1016/j.oceaneng.2018.03.007		
A Machine Learning Approach for Predicting Execution Time of Spark Jobs	downloaded	2019/07/25, 11:24:10	10.1016/j.aej.2018.03.006		
Sharing the Right Data Right A Symbiosis with Machine Learning	downloaded	2019/07/25, 11:24:10	10.1016/j.tplants.2018.10.016		
FIN DE BUSQUEDA machine learning 25 PAPERS DESCARGADOS 0 PAPERS NO DISPONIBLES TEMA: machine learning					

A continuación, veremos archivo “Reporte”, que es el Excel que se ha generado con los datos más relevantes de los papers descargados.

B2				
downloaded				
A	B	C	D	
machine learning LISTA DE PAPERS:	STATUS	FECHA	DOI	
Evaluating machine learning performance in predicting injury severity in agribusin	downloaded	2019/07/25, 11:24:10	10.1016/j.ssci.2019.04.026	
Predictors of in-hospital length of stay among cardiac patients A machine learning	downloaded	2019/07/25, 11:24:10	10.1016/j.ijcard.2019.01.046	
Machine learning algorithms for predicting scapular kinematics	downloaded	2019/07/25, 11:24:10	10.1016/j.medengphy.2019.01.005	
Attack and anomaly detection in IoT sensors in IoT sites using machine learning a	downloaded	2019/07/25, 11:24:10	10.1016/j.iot.2019.100059	
Machine learning for ecosystem services	downloaded	2019/07/25, 11:24:10	10.1016/j.ecoser.2018.04.004	
Machine Learning for Sustainable Structures A Call for Data	downloaded	2019/07/25, 11:24:10	10.1016/j.istruc.2018.11.013	
A comparison of machine learning techniques for file system forensics analysis	downloaded	2019/07/25, 11:24:10	10.1016/j.jisa.2019.02.009	
Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty A Revi	downloaded	2019/07/25, 11:24:10	10.1016/j.arth.2019.05.055	
Machine learning-assisted early ignition prediction in a complex flow	downloaded	2019/07/25, 11:24:10	10.1016/j.combustflame.2019.05.014	
Development of machine learning algorithms for prediction of mortality in spinal e	downloaded	2019/07/25, 11:24:10	10.1016/j.spinee.2019.06.024	
Machine learning core inflation	downloaded	2019/07/25, 11:24:10	10.1016/j.econlet.2018.05.001	
Identifying psychosis spectrum disorder from experience sampling data using mac	downloaded	2019/07/25, 11:24:10	10.1016/j.schres.2019.04.028	
Machine Learning for Perovskites' Reap-Rest-Recovery Cycle	downloaded	2019/07/25, 11:24:10	10.1016/j.joule.2018.11.010	
Can machine learning predict resectability of a peritoneal carcinomatosis	downloaded	2019/07/25, 11:24:10	10.1016/j.suronc.2019.04.008	
A machine learning classifier for microlensing in wide-field surveys	downloaded	2019/07/25, 11:24:10	10.1016/j.ascom.2019.100298	
A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of A	downloaded	2019/07/25, 11:24:10	10.1016/j.cell.2019.04.016	
Statistical downscaling of precipitation using machine learning techniques	downloaded	2019/07/25, 11:24:10	10.1016/j.atmosres.2018.05.022	
Internet of Things A survey on machine learning-based intrusion detection approa	downloaded	2019/07/25, 11:24:10	10.1016/j.comnet.2019.01.023	
Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine	downloaded	2019/07/25, 11:24:10	10.1016/j.matt.2019.03.002	
Big Data Analysis and Machine Learning in Intensive Care Units	downloaded	2019/07/25, 11:24:10	10.1016/j.medic.2019.06.012	
Finding the right fuel for the analytical engine Expanding the leader trait paradig	downloaded	2019/07/25, 11:24:10	10.1016/j.leaqua.2019.05.005	
Controlling of optical fiber bending losses through 'WARN' parameter and machine	downloaded	2019/07/25, 11:24:10	10.1016/j.ijleo.2019.163054	
Prediction of tidal currents using Bayesian machine learning	downloaded	2019/07/25, 11:24:10	10.1016/j.oceaneng.2018.03.007	
A Machine Learning Approach for Predicting Execution Time of Spark Jobs	downloaded	2019/07/25, 11:24:10	10.1016/j.aej.2018.03.006	
Sharing the Right Data Right A Symbiosis with Machine Learning	downloaded	2019/07/25, 11:24:10	10.1016/j.tplants.2018.10.016	
FIN DE BUSQUEDA machine learning	25 PAPERS DESCARGADOS	0 PAPERS NO DISPONIBLE	TEMA: machine learning	

Finalmente visualizamos que se crea una carpeta con el nombre correspondiente “machine learning” al tema de descarga que contiene los archivos descargados.

Nombre	Fecha de modificación	Tipo	Tamaño
perpetua	18/07/2019 00:08	Archivo PY	1 KB
interior	18/07/2019 00:08	Archivo PY	1 KB
not all	13/07/2019 00:38	Stata ZIP Archive	1 KB
pdf miner	12/07/2019 18:23	Archivo PY	2 KB
getbackgroundreport	12/07/2019 14:17	Adobe Acrobat Docu...	102 KB
Circle	12/07/2019 11:16	Stata ZIP Archive	1 KB
lot	12/07/2019 11:08	Stata ZIP Archive	6,830 KB
pdf	12/07/2019 11:07	Archivo PY	3 KB
pen2	11/07/2019 22:41	Archivo PY	2 KB
pac	11/07/2019 22:31	Archivo PY	2 KB
task	11/07/2019 21:37	Documento de texto	0 KB
chronologies	05/07/2019 00:42	Aplicación	0,471 KB
pycache	24/07/2019 00:29	Carpeta de archivos	
machine learning	24/07/2019 00:19	Carpeta de archivos	

Nombre	Fecha de modificación	Tipo	Tamaño
A comparison of machine learning techniques for the syste...	24/07/2019 00:34	Adobe Acrobat Docu...	631 KB
A Machine learning Approach for Predicting Execution Tim...	24/07/2019 00:31	Adobe Acrobat Docu...	2,122 KB
A machine learning classifier for microlearning in wide field...	24/07/2019 00:30	Adobe Acrobat Docu...	4,126 KB
A Wide-area Machine Learning Approach for Forecasting Acc...	24/07/2019 00:30	Adobe Acrobat Docu...	1,713 KB
Artificial Intelligence and Machine Learning in Lower Urban...	24/07/2019 00:29	Adobe Acrobat Docu...	782 KB
Attacks and anomaly detection in IoT sensors in IoT sdn use...	24/07/2019 00:29	Adobe Acrobat Docu...	2,238 KB
Big Data Analysis and Machine Learning in Intensive Care U...	24/07/2019 00:29	Adobe Acrobat Docu...	1,966 KB
Controlling of optical fiber bending losses through WDM...	24/07/2019 00:29	Adobe Acrobat Docu...	1,520 KB
Development of machine learning algorithms for predictio...	24/07/2019 00:30	Adobe Acrobat Docu...	626 KB
Evaluating machine learning performance in predicting inju...	24/07/2019 00:28	Adobe Acrobat Docu...	642 KB
Internet of Things A survey on machine learning based inte...	24/07/2019 00:30	Adobe Acrobat Docu...	648 KB
Machine learning algorithms for predicting scapular kinem...	24/07/2019 00:29	Adobe Acrobat Docu...	2,270 KB
Machine learning case inference	24/07/2019 00:30	Adobe Acrobat Docu...	1,140 KB
Machine learning for ecosystem services	24/07/2019 00:29	Adobe Acrobat Docu...	1,871 KB
Machine Learning for Protonic/ Super-Rest-Recovery Cycle	24/07/2019 00:30	Adobe Acrobat Docu...	1,239 KB
Machine Learning for Sustainable Structures A Call for Data	24/07/2019 00:29	Adobe Acrobat Docu...	1,240 KB
Prediction of tidal currents using Bayesian machine learning	24/07/2019 00:31	Adobe Acrobat Docu...	4,164 KB
Prediction of in-hospital length of stay among cardiac pati...	24/07/2019 00:29	Adobe Acrobat Docu...	1,210 KB
Sharing the Right Data Right A Symbolic with Machine Lea...	24/07/2019 00:31	Adobe Acrobat Docu...	708 KB
Statistical downscaling of precipitation using machine learn...	24/07/2019 00:30	Adobe Acrobat Docu...	25,097 KB
Structure-Functional Validity Indices of Meta-Organisms i...	24/07/2019 00:30	Adobe Acrobat Docu...	1,251 KB