

Manual de instrucciones de los programas CvLACpyExtract y GrupLACpyExtract

Por: Camilo Eduardo Echeverry Naranjo

Requisitos

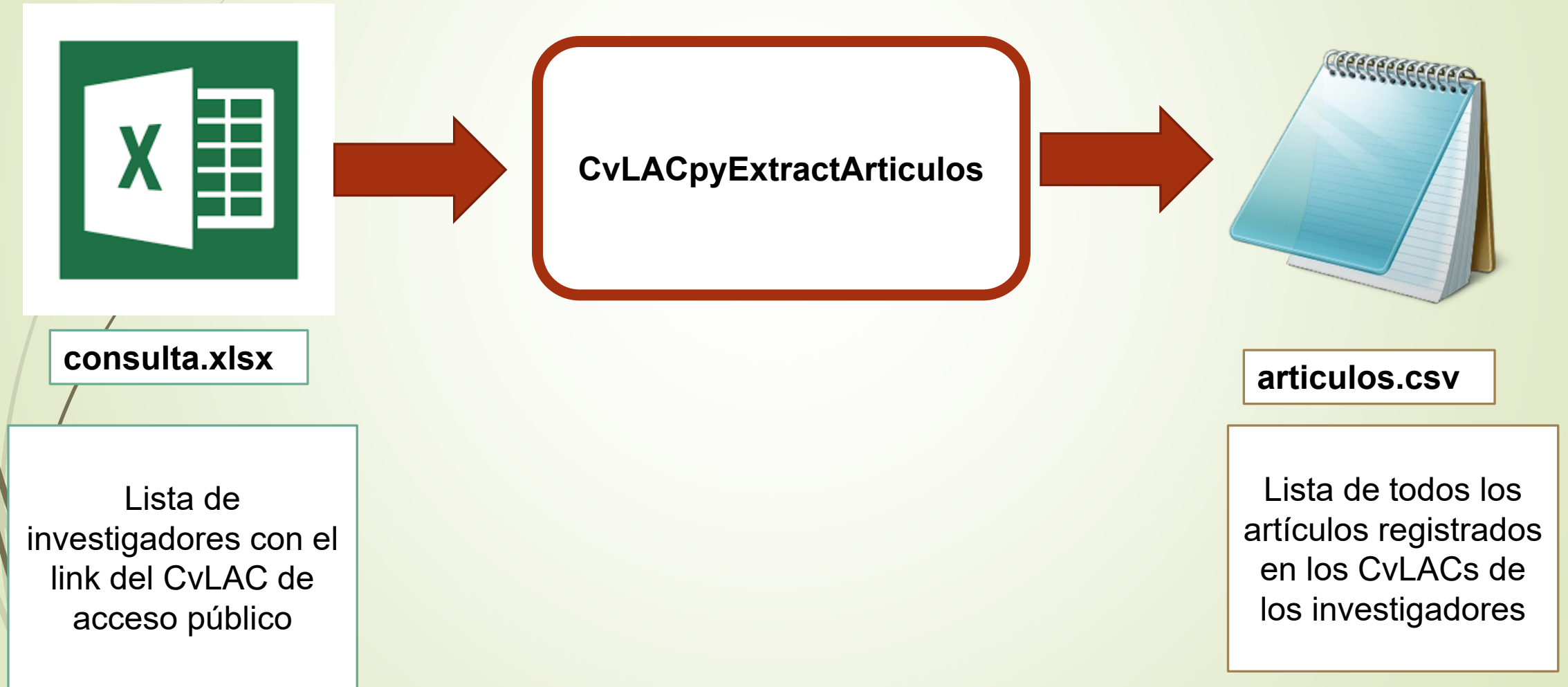
- Tener instalado Visual Studio Code con las opciones de abrir carpetas con VS Code (<https://code.visualstudio.com/>)
- Tener instalado Excel y LibreOffice (Opcional)
- Tener instalado Python (versión>3.7) y las librerías bs4 (Beautiful soup) y openpyxl. Si tiene instalado pip “pip install bs4” y “pip install openpyxl” en la CMD basta para instalarlos.
- Este manual no cubre la instalación de estos programas. Esta información se encuentra fácilmente en internet
- Para la instalación de Python, se recomienda buscar en google de “install python” o accediendo al link <https://www.python.org/downloads/>
- En la instalación asegúrese de añadir python a la variable PATH



Introducción











- Los programas CvLACpyExtract y GrupLACpyExtract son programas hechos en python utilizando las librerías BeautifulSoup (BS4) y openpyxl para “descargar” (el término correcto es Scrapear) la información que los investigadores registran en su CvLAC.
- Para ello, los programas reciben como entrada un archivo en Excel con la lista de los links (hipervínculos) de los CvLACs de varios investigadores y entrega como salida un archivo separado por coma (.CSV) con la lista de productos de cada investigador
- A continuación se muestra un ejemplo de como funciona el programa CvLACpyExtractArticulos

Diagrama simple del funcionamiento del programa CvLACpyExtractArticulos



Tutorial: CvLACpyExtractArticulos











En la carpeta Git CvLAC Scraper se encuentra una lista de carpetas como se ve en la figura:

Nombre	Fecha de modifica...	Tipo	Tamaño
 CvLACpyExtractArticulos	03/08/2022 12:57 ...	Carpeta de archivos	
 CvLACpyExtractCapDeLibros	16/06/2022 02:02 ...	Carpeta de archivos	
 CvLACpyExtractDisenosIndustriales	23/08/2022 09:02 a...	Carpeta de archivos	
 CvLACpyExtractDocumentosDeTrabajo	04/08/2022 03:40 ...	Carpeta de archivos	
 CvLACpyExtractEstrFomentoCTI	28/06/2022 04:56 ...	Carpeta de archivos	
 CvLACpyExtractEventos	28/06/2022 09:57 a...	Carpeta de archivos	
 CvLACpyExtractFormacionAcademica	04/08/2022 03:29 ...	Carpeta de archivos	
 CvLACpyExtractIformesDeInvestigacion	29/06/2022 11:49 a...	Carpeta de archivos	
 CvLACpyExtractJurados	15/07/2022 01:54 ...	Carpeta de archivos	
 CvLACpyExtractLibros	16/06/2022 11:14 a...	Carpeta de archivos	

Cada carpeta es un programa independiente (Se puede correr por separado) destinado a extraer la información descrita en el nombre de la carpeta.

Tutorial: CvLACpyExtractArticulos

Acceda a la carpeta CvLACpyExtractArticulos

Nombre	Fecha de modifica...	Tipo	Tamaño
 CvLACpyExtractArticulos	03/08/2022 12:57 ...	Carpeta de archivos	
 CvLACpyExtractCapDeLibros	16/06/2022 02:02 ...	Carpeta de archivos	
 CvLACpyExtractDisenosIndustriales	23/08/2022 09:02 a...	Carpeta de archivos	
 CvLACpyExtractDocumentosDeTrabajo	04/08/2022 03:40 ...	Carpeta de archivos	
 CvLACpyExtractEstrFomentoCTI	28/06/2022 04:56 ...	Carpeta de archivos	
 CvLACpyExtractEventos	28/06/2022 09:57 a...	Carpeta de archivos	
 CvLACpyExtractFormacionAcademica	04/08/2022 03:29 ...	Carpeta de archivos	
 CvLACpyExtractIformesDeInvestigacion	29/06/2022 11:49 a...	Carpeta de archivos	
 CvLACpyExtractJurados	15/07/2022 01:54 ...	Carpeta de archivos	
 CvLACpyExtractLibros	16/06/2022 11:14 a...	Carpeta de archivos	

Tutorial: CvLACpyExtractArticulos

Dentro de la carpeta CvLACpyExtractArticulos hay dos carpetas de interés y un archivo llamado main.py

Nombre	Fecha de modifica...	Tipo	Tamaño
__pycache__	03/08/2022 12:57 ...	Carpeta de archivos	
cvlacpy	20/04/2022 09:32 a...	Carpeta de archivos	
Input	27/09/2022 08:19 a...	Carpeta de archivos	
Logs	07/04/2022 11:56 a...	Carpeta de archivos	
Output	09/09/2022 05:55 ...	Carpeta de archivos	
main.py	28/04/2022 10:13 a...	Python File	2 KB

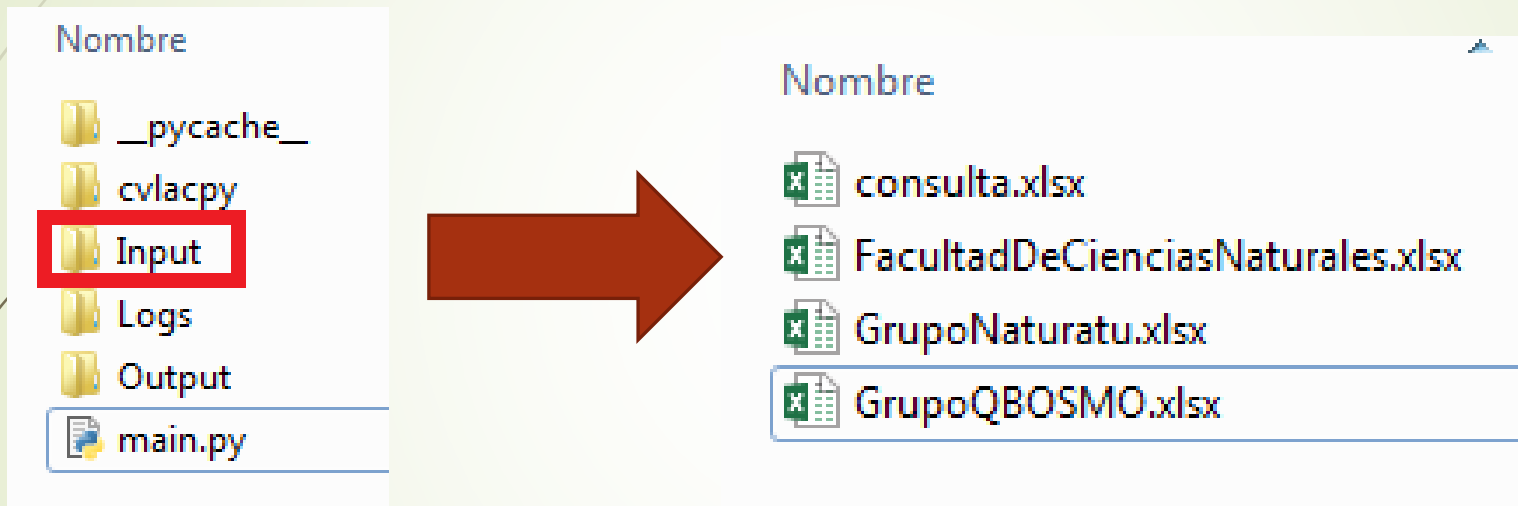
Input: es la carpeta que contiene el archivo de entrada, por ejemplo, consulta.xlsx

Output: es la carpeta que contiene el archivo de salida, por ejemplo, artículos.csv

main.py: Es un script que se corre desde una terminal. Este es el archivo que se debe ejecutar para correr el programa.

Tutorial: CvLACpyExtractArticulos

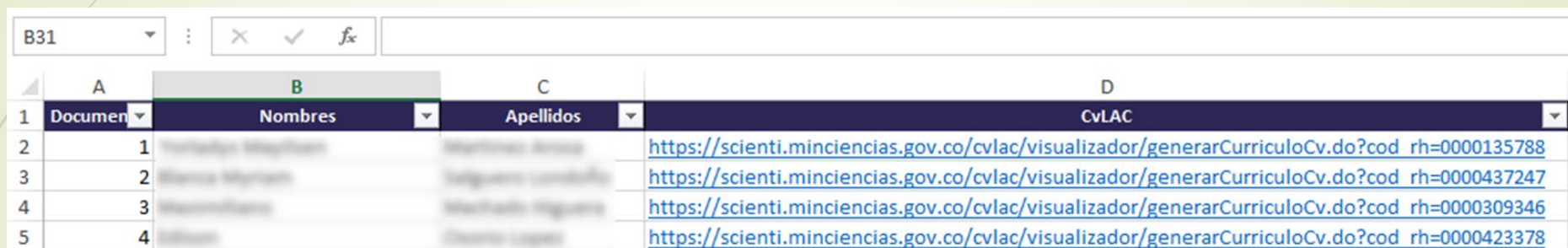
Revisemos el contenido de la carpeta Input



En la carpeta Input hay varios archivos de Excel. El programa es capaz de tomar cada uno de ellos como entrada si se especifica, de lo contrario, el programa toma por defecto el archivo `consulta.xlsx` como entrada

Tutorial: CvLACpyExtractArticulos

Abriendo el archivo consulta.xlsx encontramos lo siguiente:



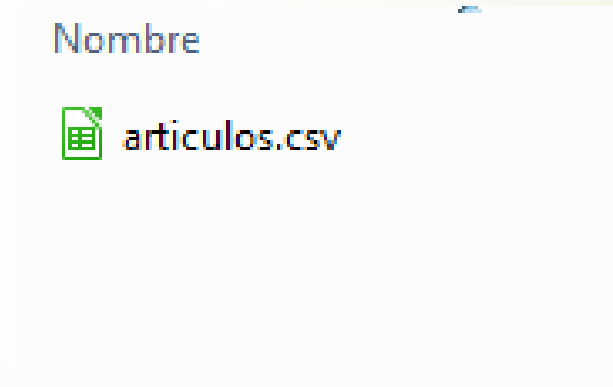
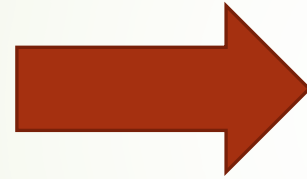
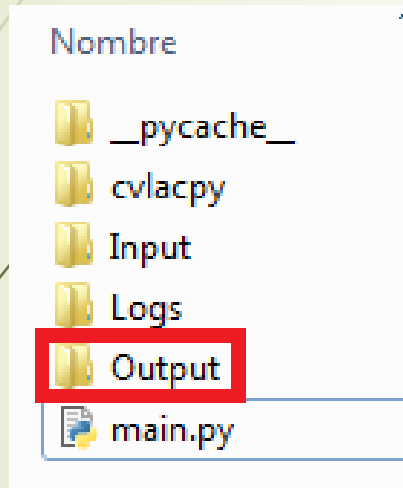
	A	B	C	D
1	Document	Nombres	Apellidos	CvLAC
2	1	[Faded Name]	[Faded Surname]	https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000135788
3	2	[Faded Name]	[Faded Surname]	https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000437247
4	3	[Faded Name]	[Faded Surname]	https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000309346
5	4	[Faded Name]	[Faded Surname]	https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000423378

De este archivo solo hay que entender lo siguiente:

- Todo archivo de entrada debe tener exactamente la misma fila 1. Esta fila es la única que no se puede modificar.
- El contenido de las columnas A, B y C No es de interés para el programa y se puede modificar como se desee (el nombre de los investigadores se toma de los CvLACs, no de esta lista).
- Se recomienda no escribir la cédula de los investigadores en la columna documentos. En su lugar escriba un número que le permita contar el número de integrantes.
- El programa solamente lee la columna D con los links de los respectivos CvLACs

Tutorial: CvLACpyExtractArticulos

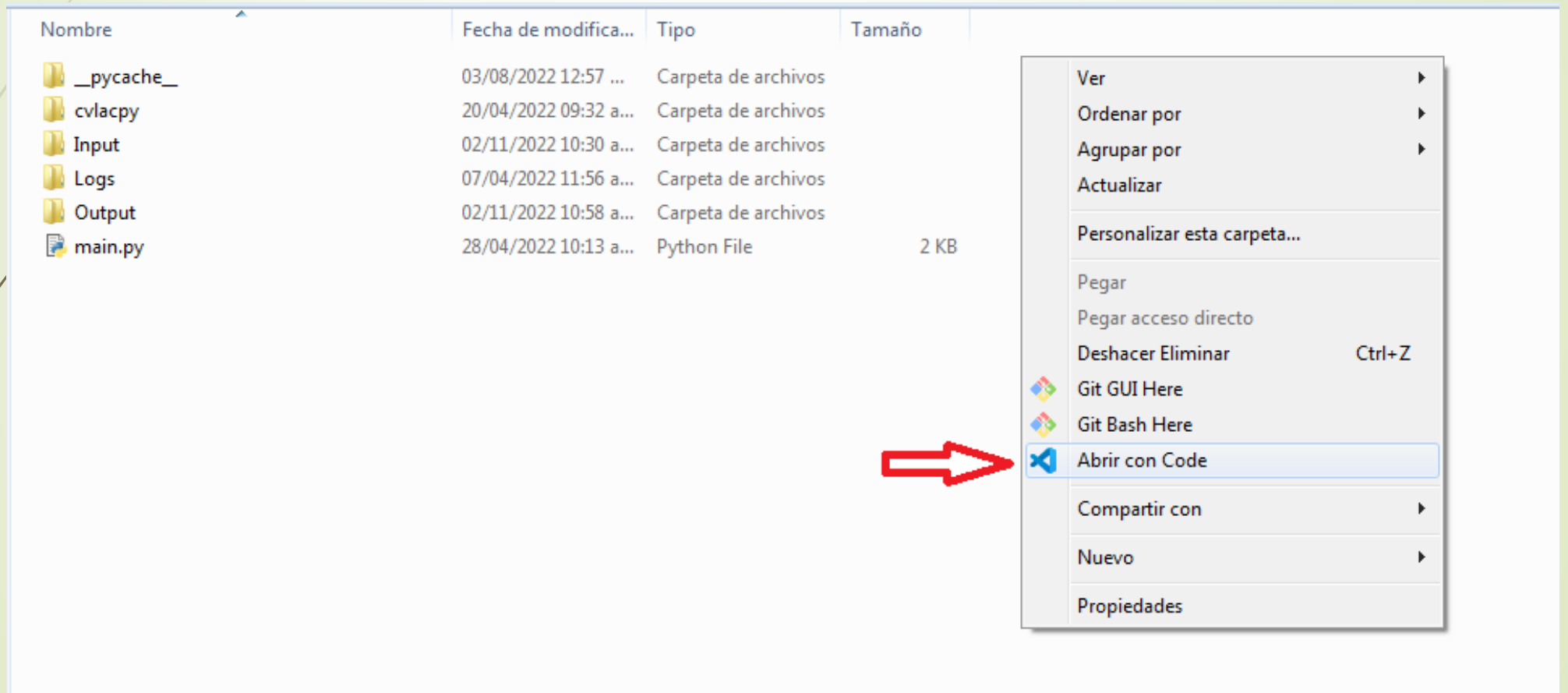
Revisemos el contenido de la carpeta Output



En la carpeta Output se encuentra el archivo de salida artículos.csv. Si bien este archivo se puede abrir con Excel, se recomienda usar el software libre LibreOfficeCalc para abrirlo y salvar como archivo de Excel (.xlsx)

Tutorial: CvLACpyExtractArticulos

Para correr el programa abrimos la carpeta con Visual Studio Code

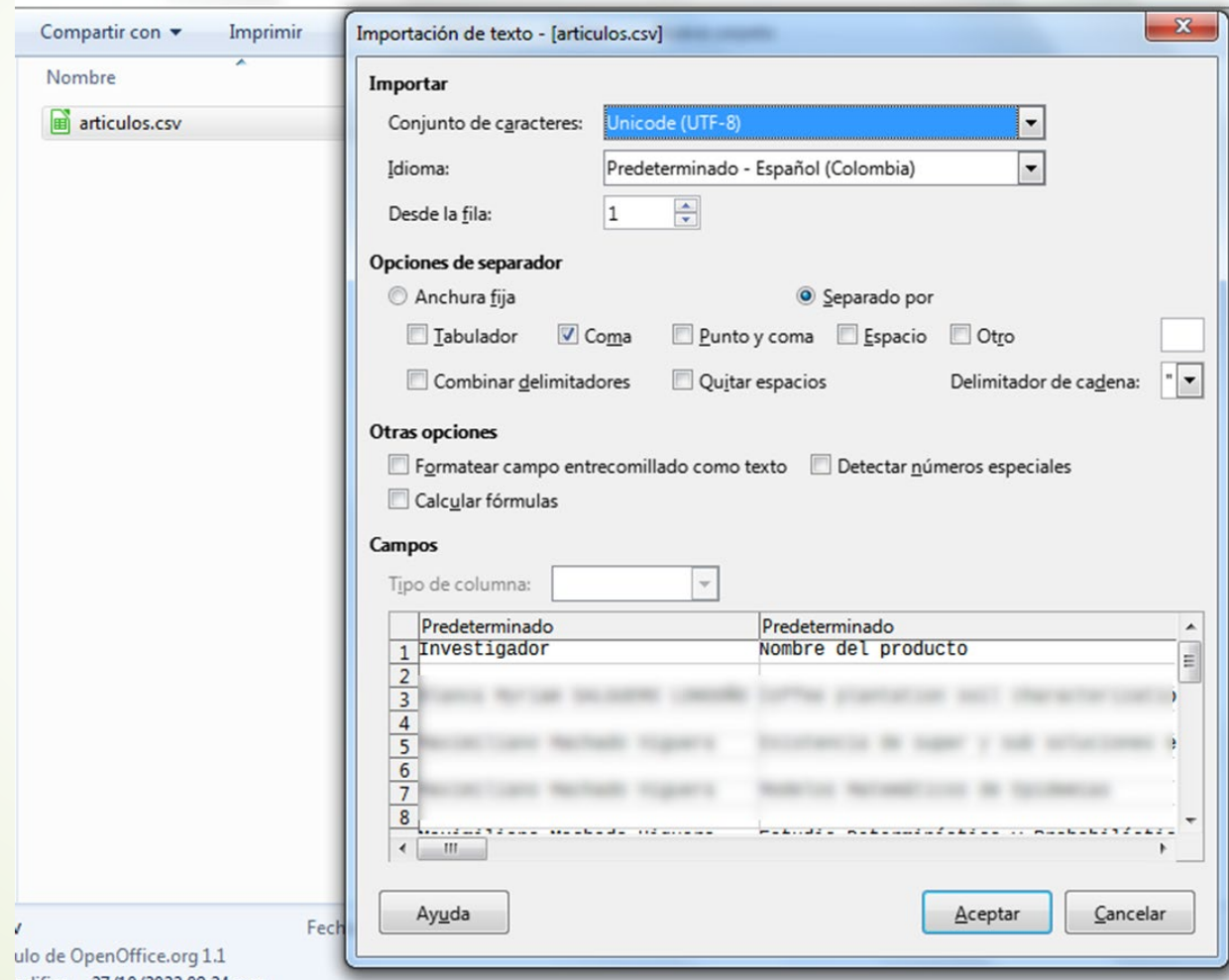


Tutorial: CvLACpyExtractArticulos

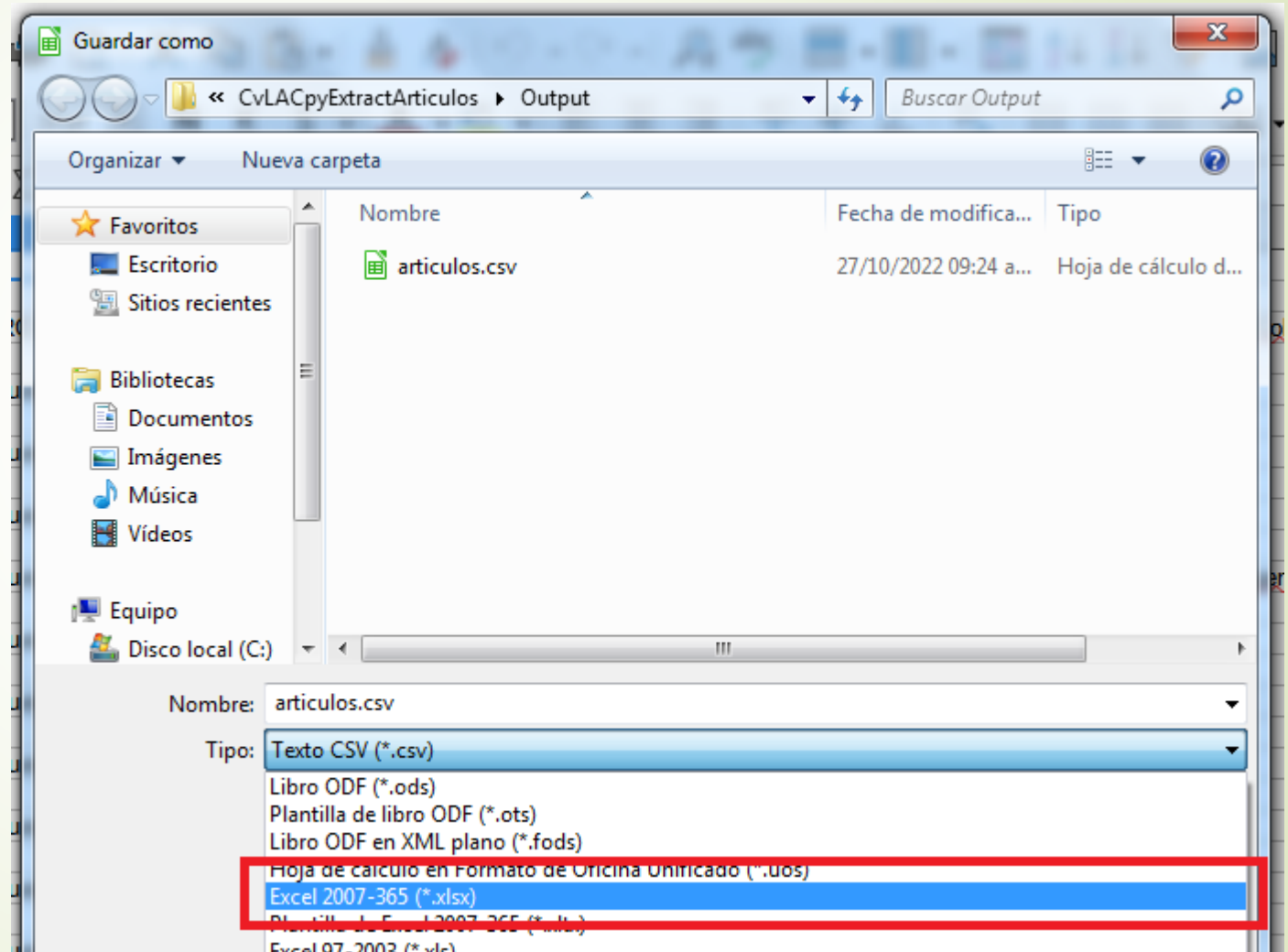
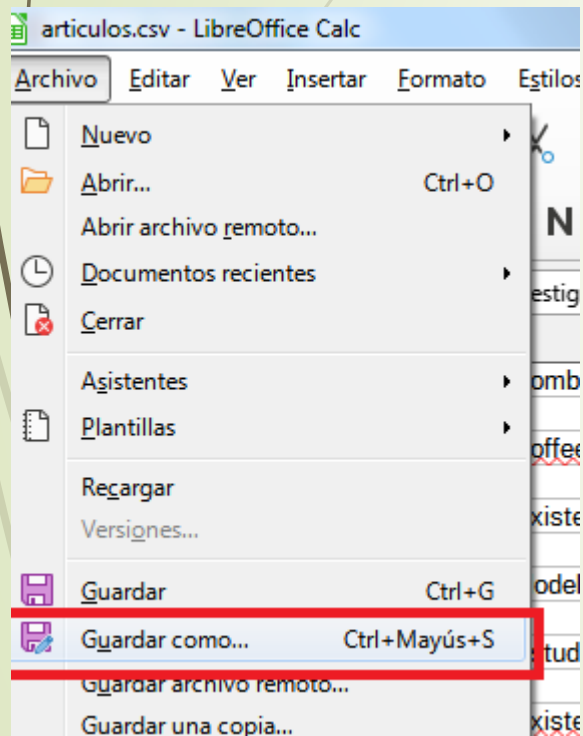
LibreOffice permite dos cosas:

1. Al abrir el archivo permite elegir como leer el archivo.
2. Salvar el archivo en formato Excel sin problemas de codificación de caracteres

Se deben escoger únicamente las opciones que aparecen en la imagen

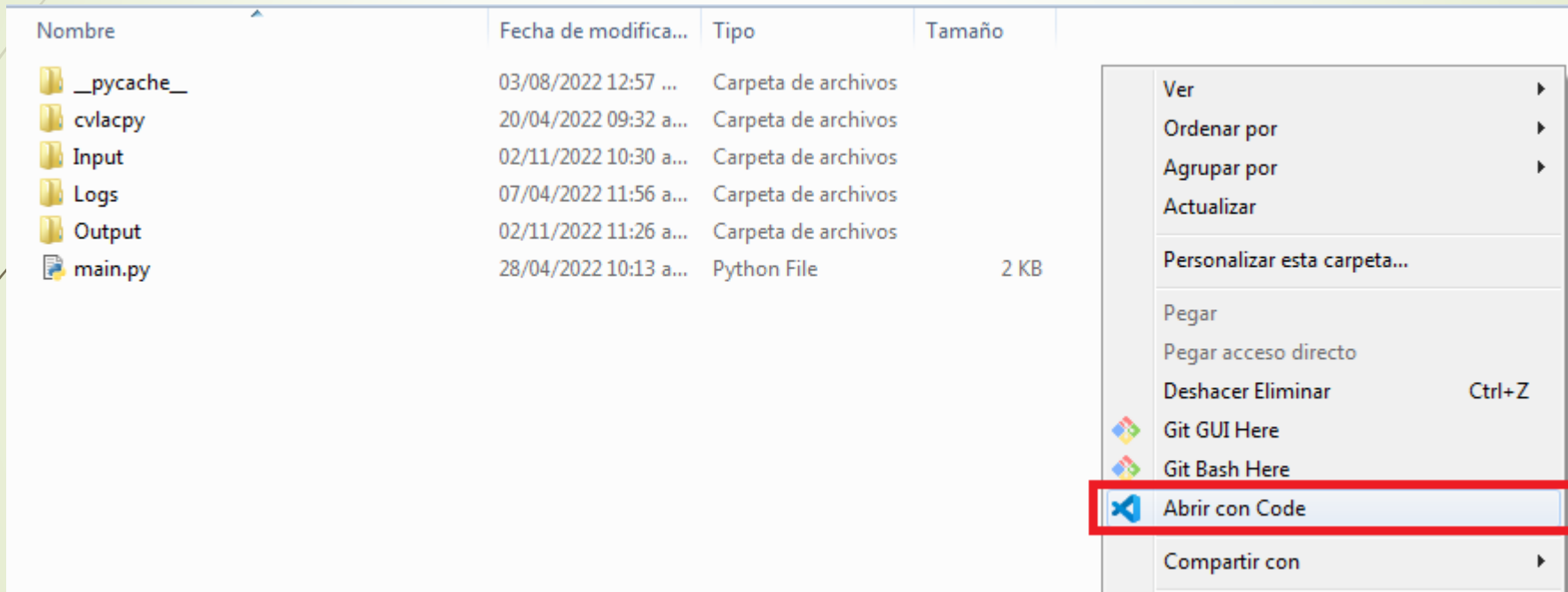


Tutorial: CvLACpyExtractArticulos



Tutorial: CvLACpyExtractArticulos

Uso del programa

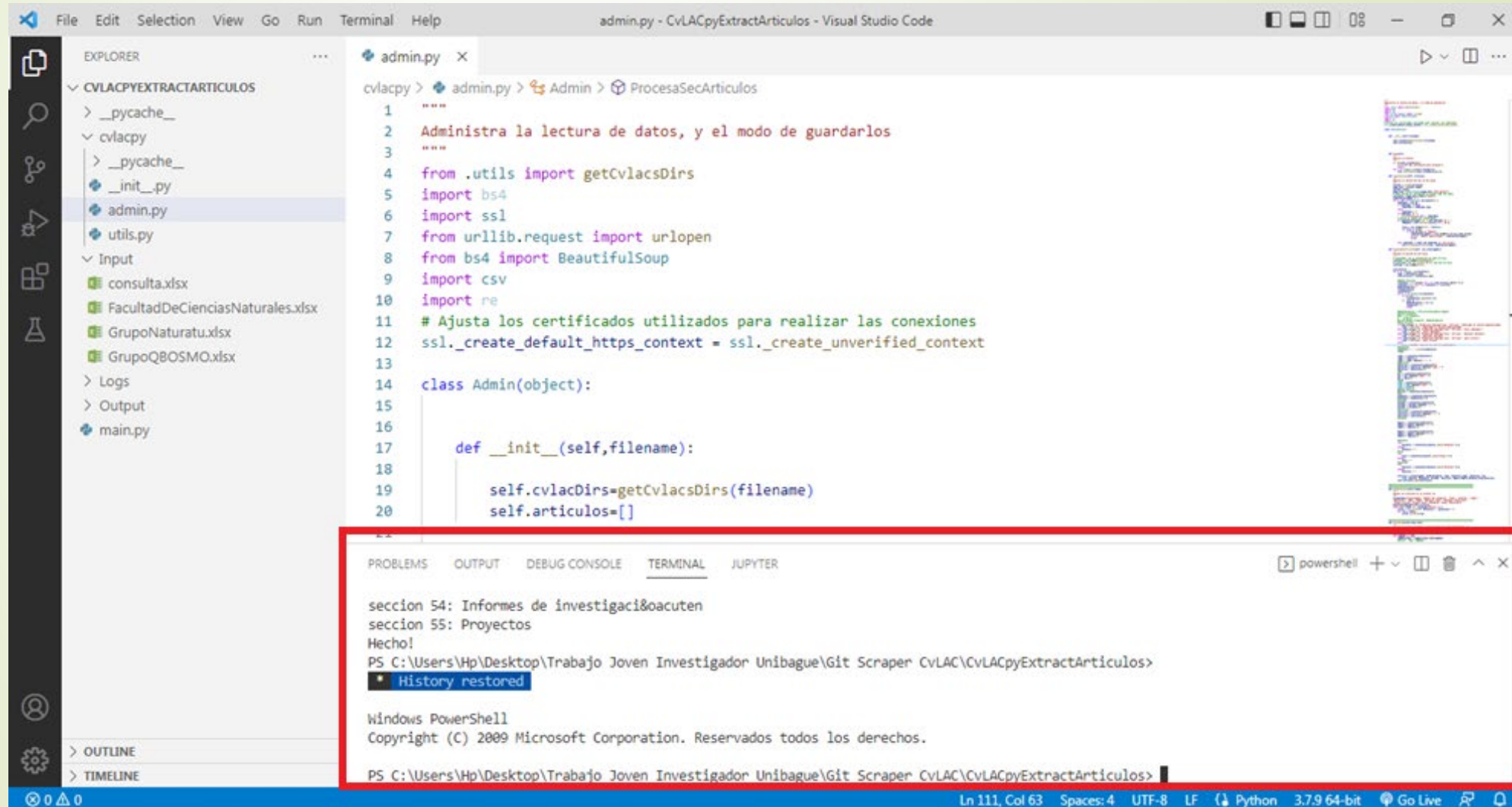


Para correr el programa, primero se abre la carpeta CvLACpyExtractArticulos con VS Code

Tutorial: CvLACpyExtractArticulos

Una vez abierto nos centramos en la terminal (parte resaltada en rojo)

Nota: VS Code no es estrictamente necesario. Los pasos a seguir se pueden ejecutar en cualquier terminal (CMD o PowerShell)



The screenshot shows the Visual Studio Code interface with the project 'CvLACpyExtractArticulos' open. The Explorer panel on the left shows the file structure, with 'admin.py' highlighted. The main editor displays the code in 'admin.py', which includes imports for 'urllib.request', 'BeautifulSoup', 'csv', and 're', and a class 'Admin' with an '__init__' method. The terminal at the bottom, which is highlighted with a red border, shows the command prompt output. The terminal text includes 'seccion 54: Informes de investigación', 'seccion 55: Proyectos', 'Hecho!', and the PowerShell prompt 'PS C:\Users\Hp\Desktop\Trabajo Joven Investigador Unibague\Git Scraper CvLAC\CvLACpyExtractArticulos>'. A blue status bar at the bottom indicates 'Ln 111, Col 63', 'Spaces: 4', 'UTF-8', 'LF', 'Python', '3.7.9 64-bit', and 'Go Live'.

```
File Edit Selection View Go Run Terminal Help
admin.py - CvLACpyExtractArticulos - Visual Studio Code

EXPLORER
CvLACPYEXTRACTARTICULOS
  > __pycache__
  > cvlacpy
    > __pycache__
    > __init__.py
    > admin.py
    > utils.py
  > Input
    > consulta.xlsx
    > FacultadDeCienciasNaturales.xlsx
    > GrupoNaturatu.xlsx
    > GrupoQBOSMO.xlsx
  > Logs
  > Output
  > main.py

admin.py
1  """
2  Administra la lectura de datos, y el modo de guardarlos
3  """
4  from .utils import getCvlacsDirs
5  import bs4
6  import ssl
7  from urllib.request import urlopen
8  from bs4 import BeautifulSoup
9  import csv
10 import re
11 # Ajusta los certificados utilizados para realizar las conexiones
12 ssl._create_default_https_context = ssl._create_unverified_context
13
14 class Admin(object):
15
16     def __init__(self, filename):
17
18         self.cvlacDirs=getCvlacsDirs(filename)
19         self.articulos=[]
```

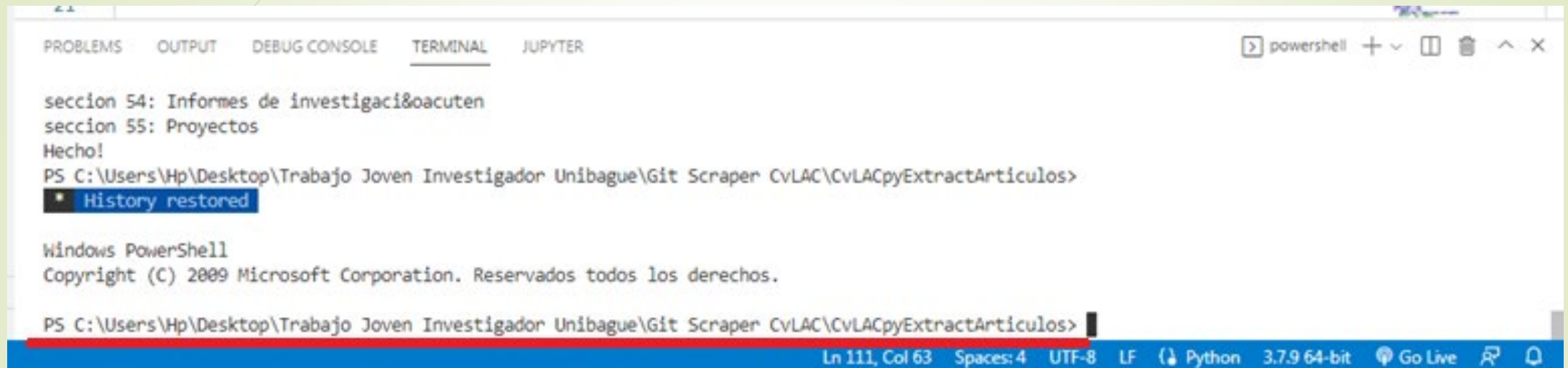
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

seccion 54: Informes de investigación
seccion 55: Proyectos
Hecho!
PS C:\Users\Hp\Desktop\Trabajo Joven Investigador Unibague\Git Scraper CvLAC\CvLACpyExtractArticulos>
History restored

Windows PowerShell
Copyright (C) 2009 Microsoft Corporation. Reservados todos los derechos.
PS C:\Users\Hp\Desktop\Trabajo Joven Investigador Unibague\Git Scraper CvLAC\CvLACpyExtractArticulos>

Ln 111, Col 63 Spaces: 4 UTF-8 LF Python 3.7.9 64-bit Go Live

Tutorial: CvLACpyExtractArticulos



The screenshot shows a JupyterLab interface with a terminal window open. The terminal displays the following text:

```
seccion 54: Informes de investigaci&oacuten
seccion 55: Proyectos
Hecho!
PS C:\Users\Hp\Desktop\Trabajo Joven Investigador Unibague\Git Scraper CvLAC\CvLACpyExtractArticulos>
• History restored

Windows PowerShell
Copyright (C) 2009 Microsoft Corporation. Reservados todos los derechos.

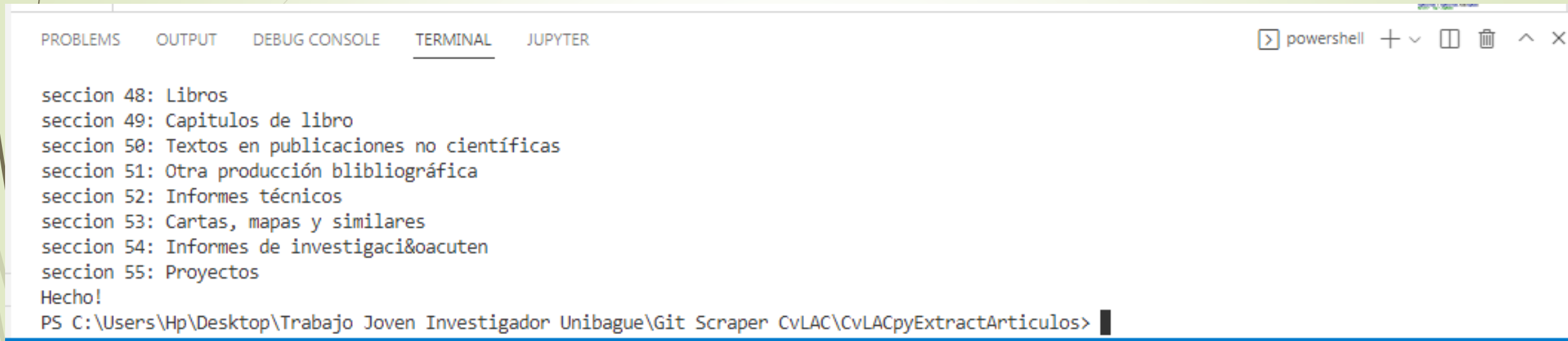
PS C:\Users\Hp\Desktop\Trabajo Joven Investigador Unibague\Git Scraper CvLAC\CvLACpyExtractArticulos>
```

The terminal window has a blue title bar with the text "powershell" and standard window controls. The status bar at the bottom of the terminal shows "Ln 111, Col 63 Spaces: 4 UTF-8 LF Python 3.7.9 64-bit Go Live".

Asegurándose de que la terminal este abierta en la carpeta
CvLACpyExtractArticulos, se procede a digitar lo siguiente:
python main.py
Acto seguido Enter
(Si el comando anterior no funciona prueba con py main.py)

```
PS C:\Users\Hp\Desktop\Trabajo Joven Investigador Unibague\Git Scraper CvLAC\CvLACpyExtractArticulos> python main.py
```

Tutorial: CvLACpyExtractArticulos



The screenshot shows a Jupyter Notebook interface with a terminal window open. The terminal window has tabs for PROBLEMS, OUTPUT, DEBUG CONSOLE, TERMINAL, and JUPYTER. The TERMINAL tab is active, showing the output of a PowerShell command. The output lists several sections of a document, followed by the word 'Hecho!' and the PowerShell prompt 'PS C:\Users\Hp\Desktop\Trabajo Joven Investigador Unibague\Git Scraper CvLAC\CvLACpyExtractArticulos>'. The sections listed are: seccion 48: Libros, seccion 49: Capítulos de libro, seccion 50: Textos en publicaciones no científicas, seccion 51: Otra producción bibliográfica, seccion 52: Informes técnicos, seccion 53: Cartas, mapas y similares, seccion 54: Informes de investigación, and seccion 55: Proyectos.

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER powershell + - [ ] [X] ^ X

seccion 48: Libros
seccion 49: Capítulos de libro
seccion 50: Textos en publicaciones no científicas
seccion 51: Otra producción bibliográfica
seccion 52: Informes técnicos
seccion 53: Cartas, mapas y similares
seccion 54: Informes de investigación
seccion 55: Proyectos
Hecho!
PS C:\Users\Hp\Desktop\Trabajo Joven Investigador Unibague\Git Scraper CvLAC\CvLACpyExtractArticulos>
```

Si el programa se ha ejecutado exitosamente aparecerá la palabra Hecho! Antes de PS C:\

Una vez ejecutado los resultados se pueden encontrar en la carpeta Output en el archivo artículos.csv

Tutorial: Extraer toda la información

En la carpeta junto con todos los programas se encuentra un archivo llamado `ActualizarInfoCvLACs.py`. Este archivo se puede ejecutar de la misma forma que se ejecutó `main.py` y hace lo siguiente:

1. Ejecuta todos los programas `CvLACpyExtract` de forma secuencial
2. Copia los archivos.csv (resultado de ejecutar los programas) a la carpeta `CvLACResults`.

Nombre	Fecha de modifica...	Tipo	Tamaño
CvLACpyExtractNormasRegulaciones	23/08/2022 10:50 a...	Carpeta de archivos	
CvLACpyExtractPlantaPiloto	23/08/2022 09:18 a...	Carpeta de archivos	
CvLACpyExtractPrototipos	23/08/2022 09:48 a...	Carpeta de archivos	
CvLACpyExtractProyectos	29/06/2022 02:33 ...	Carpeta de archivos	
CvLACpyExtractPublicacionesNoCientificas	16/06/2022 02:59 ...	Carpeta de archivos	
CvLACpyExtractSignosDistintivos	23/08/2022 10:06 a...	Carpeta de archivos	
CvLACpyExtractSoftwares	23/08/2022 10:33 a...	Carpeta de archivos	
CvLACpyExtractTrabajosDirigidos	28/06/2022 08:36 a...	Carpeta de archivos	
CvLACResults	02/11/2022 09:21 a...	Carpeta de archivos	
GrupLACpyExtractArticulos	21/07/2022 03:41 ...	Carpeta de archivos	
GrupLACpyExtractCapDeLibros	22/07/2022 09:22 a...	Carpeta de archivos	
GrupLACpyExtractEventos	26/07/2022 10:25 a...	Carpeta de archivos	
GrupLACpyExtractIntegrantes	05/08/2022 03:32 ...	Carpeta de archivos	
GrupLACpyExtractLibros	22/07/2022 08:54 a...	Carpeta de archivos	
GrupLACpyExtractProyectos	22/07/2022 11:11 a...	Carpeta de archivos	
GrupLACpyExtractTrabDirigidos	22/07/2022 10:00 a...	Carpeta de archivos	
GrupLACResults	10/08/2022 05:32 ...	Carpeta de archivos	
Obsoleto	28/06/2022 05:28 ...	Carpeta de archivos	
Scopus_pyExtractArticulos	08/08/2022 10:17 a...	Carpeta de archivos	
18A6B993.tmp	02/11/2022 10:06 a...	Archivo TMP	191 KB
ActualizarInfoCvLACs.py	31/08/2022 04:04 ...	Python File	8 KB
ActualizarInfoGrupLACs.py	05/08/2022 03:37 ...	Python File	4 KB
Git Scraper CvLAC.7z	30/09/2022 08:26 a...	7z Archive	1.653 KB
GrupoNaturatu.xlsx	29/06/2022 04:23 ...	Hoja de cálculo d...	13 KB
Manual de Instrucciones.pptx	02/11/2022 02:20 ...	Presentación de ...	1.051 KB



Extraer información del GrupLAC

- Para extraer la información del GrupLAC se ejecutan los mismos pasos ejecutados con el programa CvLACpyExtractArticulos.
- La única diferencia es que GrupLACpyExtractArticulos recibe una lista de links de GrupLACs en lugar de CvLACs