

Manual de Instalacion.

Software de Categorización y Procesamiento de Archivos PDF y Excel

Desarrollador: Stiven Colorado

Cargo: Desarrollador de Software Freelance

Experiencia: Más de 4 años de experiencia en desarrollo de software.

Objetivo del Manual

Este manual tiene como propósito proporcionar a los usuarios una guía clara y concisa sobre cómo **instalar** el **Software de Categorización y Procesamiento de Archivos PDF y Excel**. Incluye las instrucciones detalladas para preparar el entorno, realizar la instalación y configurar las herramientas necesarias para su correcto funcionamiento. Además, busca garantizar una instalación sencilla y libre de complicaciones.

1. Introducción

Bienvenido al **Manual de Instalación** del Software de Categorización y Procesamiento de Archivos PDF y Excel. Este manual tiene como objetivo guiar al usuario a través del proceso de instalación y configuración del programa, permitiendo que su implementación sea eficiente y efectiva.

El software ha sido diseñado para optimizar el procesamiento manual de archivos PDF y Excel, mejorando la eficiencia y reduciendo drásticamente los tiempos operativos en la categorización y organización de datos. Con esta guía, tendrás toda la información necesaria para asegurar que el sistema funcione correctamente desde el inicio.

2. Presentación del Programa

El software proporciona una solución integral que incluye la capacidad de:

- **Importar automáticamente archivos PDF y Excel** para su análisis y categorización.
- **Clasificar información por cliente y subpartida** mediante herramientas avanzadas de análisis de datos.
- Utilizar tecnologías de vanguardia como **OCR con Tesseract**, optimizando la lectura de documentos PDF escaneados.
- **Generar informes automáticos en formatos Excel y PDF**, facilitando la exportación y compatibilidad con otros sistemas.

El programa combina facilidad de uso con funcionalidad avanzada, asegurando resultados precisos y consistentes en menos tiempo. Este manual se enfocará en el proceso de instalación necesario para habilitar estas capacidades en tu entorno de trabajo.

Requisitos Mínimos

Para poder ejecutar correctamente el **Software de Categorización y Procesamiento de Archivos PDF y Excel**, es necesario contar con el siguiente entorno mínimo en tu sistema:

Requisitos del Sistema Operativo

- **Windows 10** o versiones posteriores.
- **MacOS 11.7.10** (Big Sur) o versiones posteriores.
- **Linux** (cualquier distribución compatible con Python 3.7+).

Requisitos de Hardware

- **Procesador:** Intel i3 o superior.

- **Memoria RAM:** 4GB de RAM como mínimo.
- **Espacio en disco:** 1 GB de espacio libre en disco duro para la instalación del software y las dependencias.

Requisitos de Software

- **Python 3.7+** (se recomienda la versión más reciente).
- **Dependencias de Python:**
 - pandas
 - numpy
 - openpyxl
 - fpdf
 - Pillow
 - PyMuPDF
 - python-docx
 - pytesseract
 - pdf2image
 - opencv-python

Software Adicional

- **Tesseract OCR:** Necesario para el procesamiento de documentos PDF mediante OCR. Puedes descargarlo desde [aquí](#).
- **Librerías de Python:** Pueden instalarse fácilmente con el siguiente comando en la terminal o línea de comandos:

```
pip install pandas numpy openpyxl fpdf Pillow PyMuPDF python-docx pytesseract pdf2image opencv-python
```

1. Proceso de instalacion.

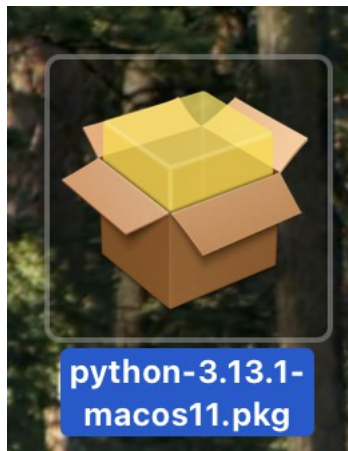
En el sistema operativo Windows:

Primero se debera de instalar Python en el dispositivo para hacerlo deberas de ingresar a la pagina oficial de Python, enlace: <https://www.python.org/>

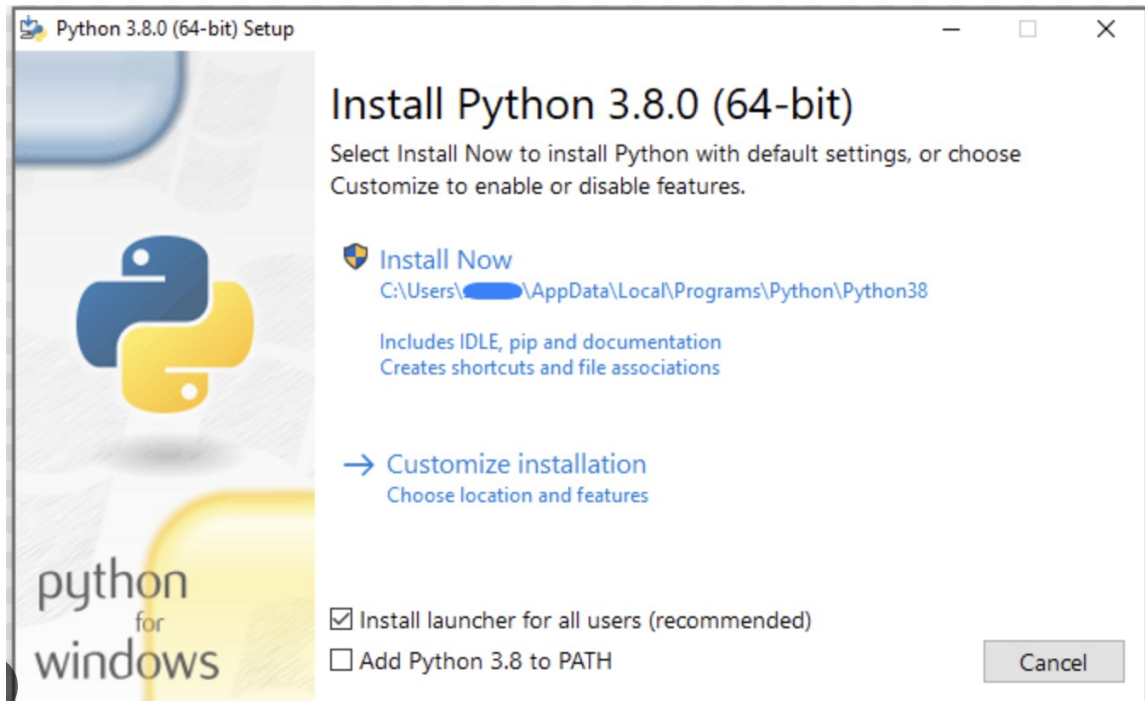
Busca la opcion “Descargar” y luego presiona en el boton “Python 3.13.1”



Deberas de esperar unos segundos a que descargue el instalador, debera de quedar algo como: (en caso de Mac sera un archivo .pkg, en windows podras verlo como .exe)



Deberas de abrir el instalador y presionar en el boton “Siguiente”, es recomendable dejar la configuracion por defecto que tiene el instalador y no cambiar rutas. tendras una ventana como la siguiente: (deberas de activar las opciones install launcher for all users y Add to Python to Path) y presionar en “instalar”.

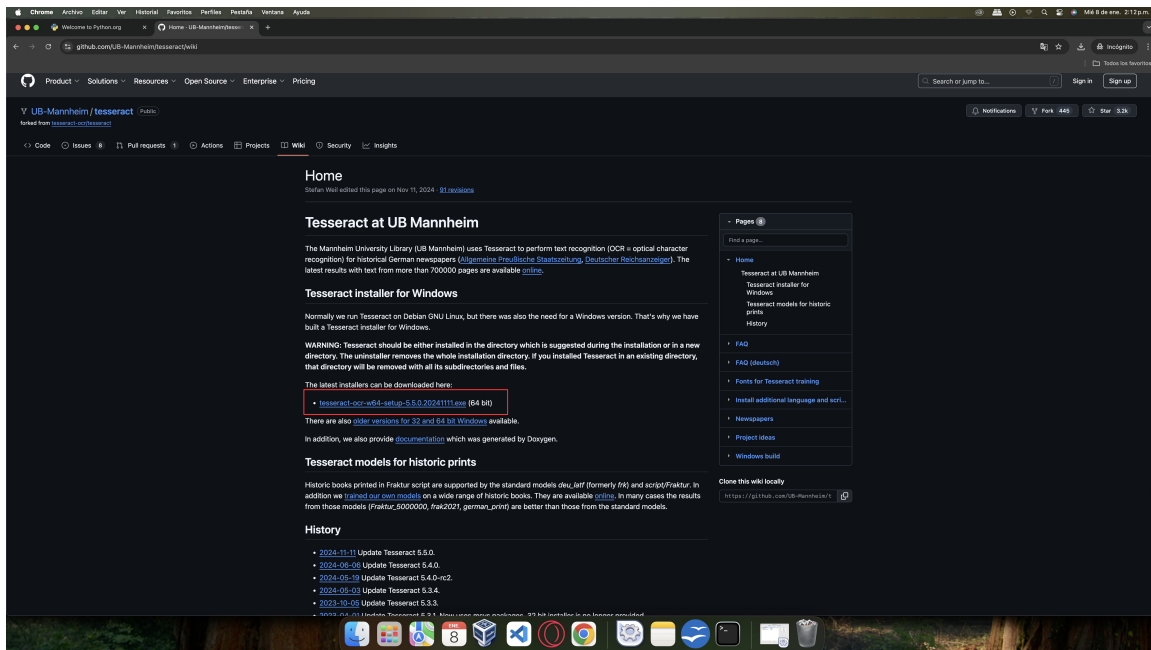


Ahora deberemos de instalar Tesseract OCR:

ingresa al siguiente enlace: <https://github.com/UB-Mannheim/tesseract/wiki>

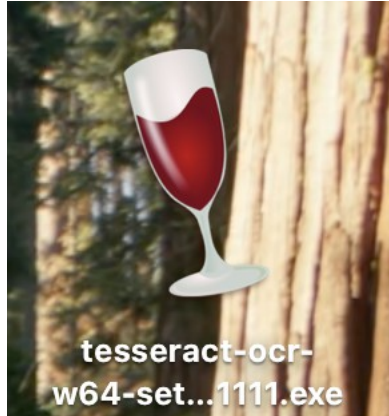
Busca el enlace para la arquitectura x64 (si tu equipo cuenta con 64 bits).

Aqui señalo en el recuadro rojo como seria el enlace.

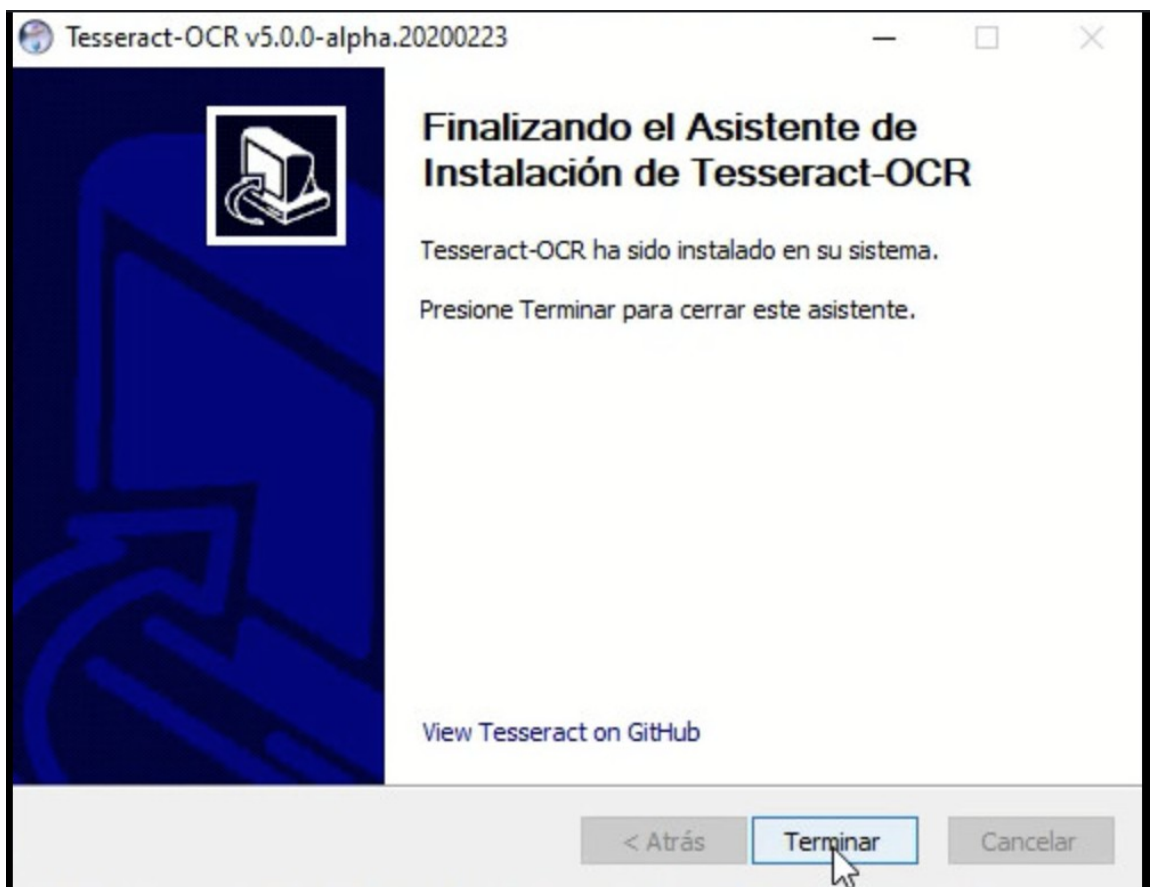


Se debera de comenzar la descarga del instalador de Tesseract OCR.

Cuando se descargue deberas de ejecutar el instalador. Instalarlo es muy sencillo solo consta en aceptar terminos y de nuevo, **es recomendable no cambiar la ruta donde se instalar ya que esto influira mas adelante en la instalacion.**



Al final deberas de ver una ventana como:



Ahora, es importante entender que en el computador o dispositivo en el que se haya realizado la anterior instalacion desconoce globalmente en el sistema al software “Tesseract OCR”, lo cual, deberemos de configurar para que nuestro sistema lo reconozca.

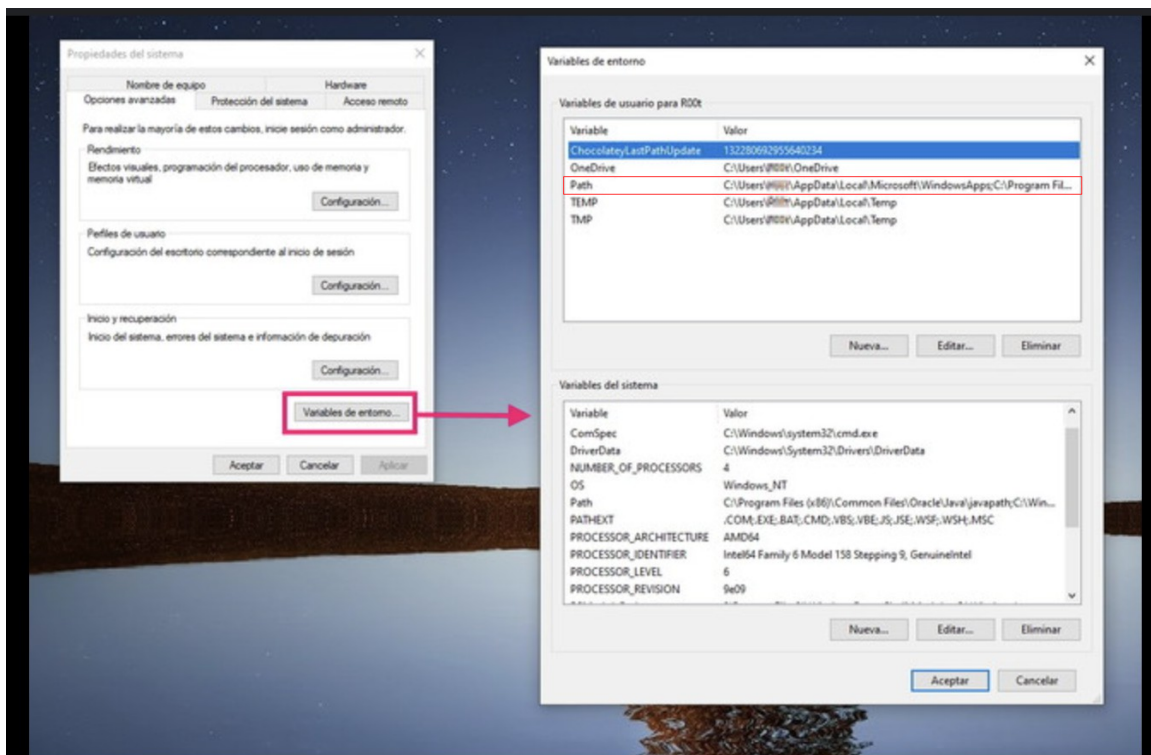
1. Agregar Tesseract a variables de entorno en windows:

Para ver o cambiar las variables de entorno:

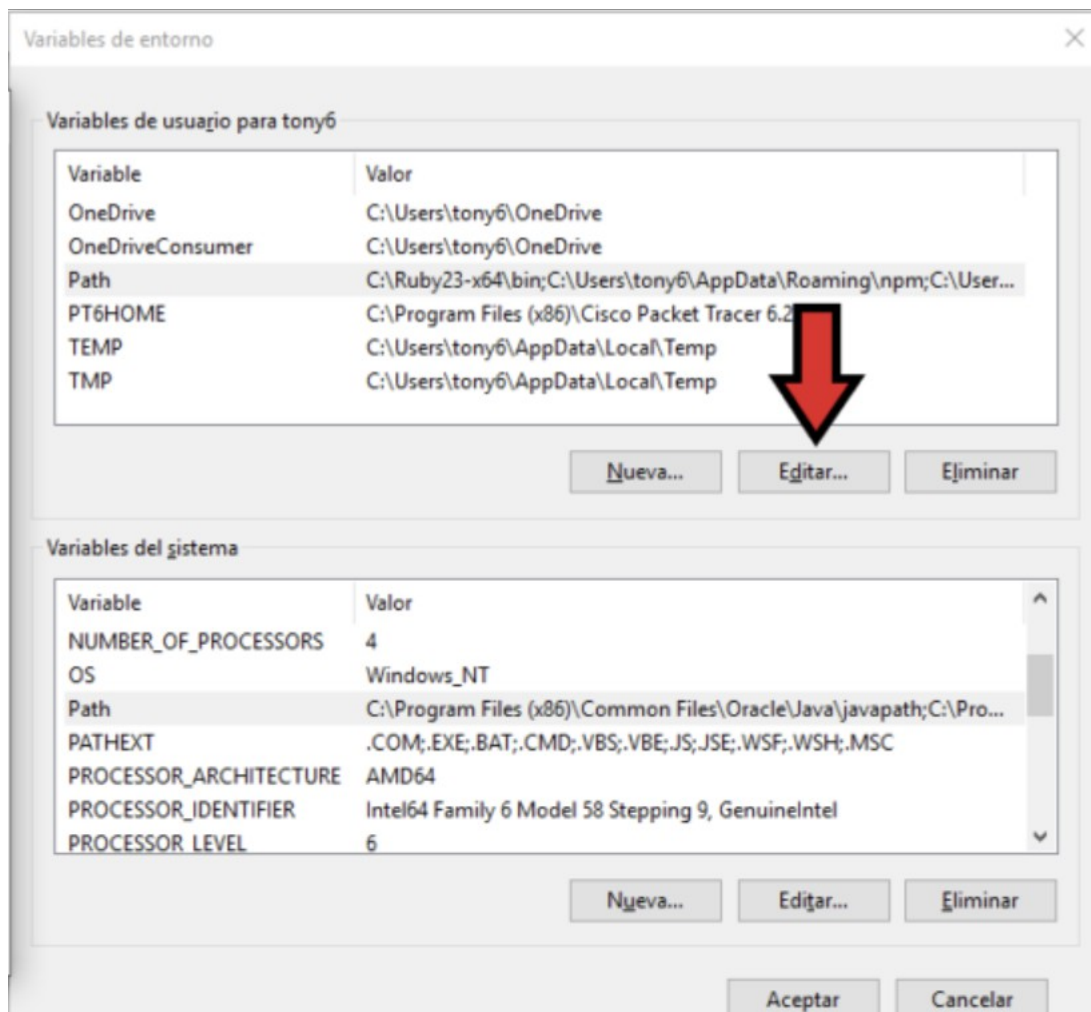
- Haga clic con el botón secundario en Mi PC y, a continuación, haga clic en Propiedades.
- Haga clic en la pestaña Opciones avanzadas.
- Haga clic en Variables de entorno.

Los puntos que mas nos importan son los señalados en los recuadros rojos:

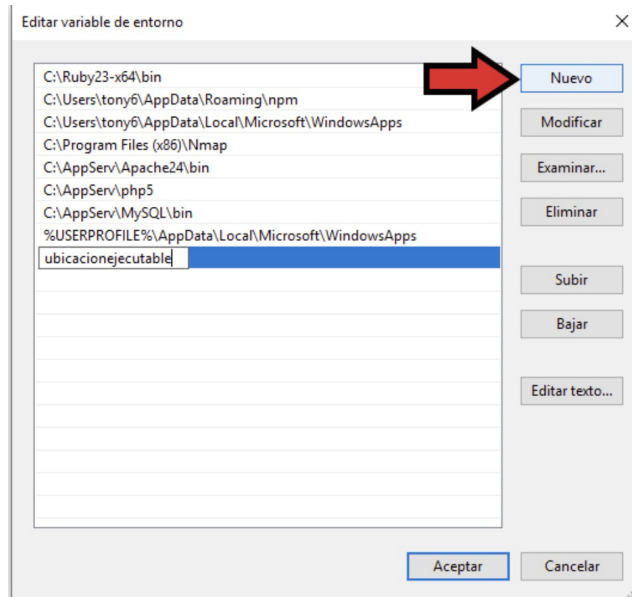
Variables de entorno y Path.



Deberas de presionar en Path “doble click”o en editar como:



Se debera de abrir una ventana llamada “Editar variable de entorno”, luego daremos click en Nuevo:



Copiaremos lo siguiente y agregaremos como “Nuevo” y donde se enmarca la entrada de escritura pegaremos:

C:\Program Files\Tesseract-OCR,

De nuevo presionaremos en “Nuevo” y pegaremos:

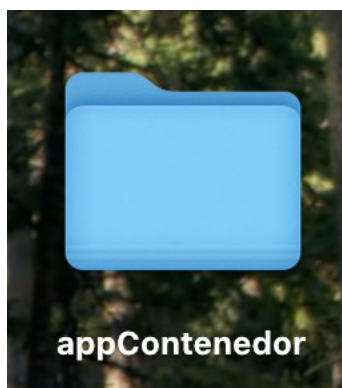
C:\Users\mpc\AppData\Local\Programs\Python\Python313,

Por ultimo presionamos de nuevo en “Nuevo” y pegaremos:

C:\Users\mpc\AppData\Local\Programs\Python\Python313\Scripts

“Estas rutas pueden cambiar si el directorio en donde instalaste Python y Tesseract no han sido las predeterminadas”

Ahora debemos de asegurarnos de tener el codigo fuente del programa, este sera suministrado por el desarrollador. Dentro de la carpeta de la aplicacion debera de llamarse appContenedor.



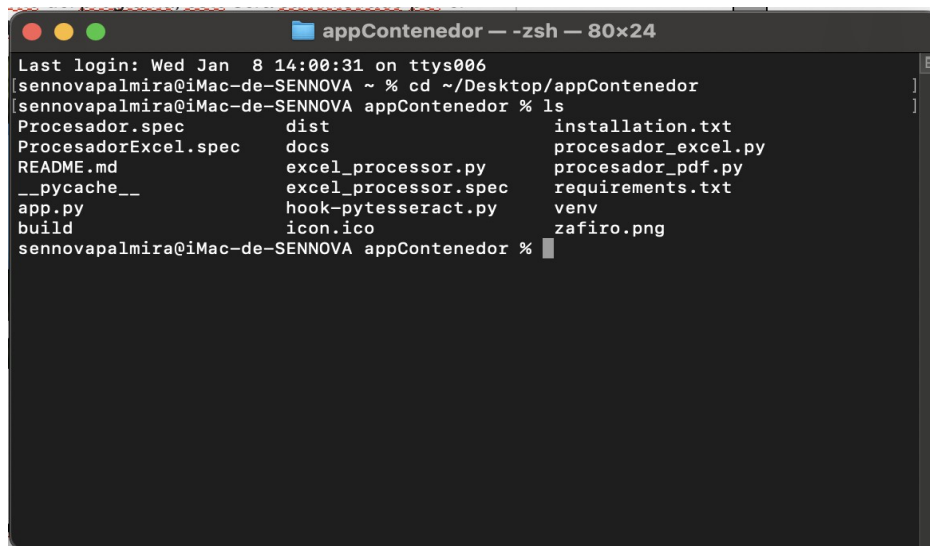
En Windows:

Dentro de esta carpeta presionaremos click “derecho” y presionar “abrir Terminal”.

En Mac:

deberemos de abrir una terminal y movernos al directorio de la aplicacion, recomiendo poner la carpeta en el escritorio.

En la terminal escribir y presionar Enter: `cd ~/Desktop/appContenedor/`

A terminal window titled 'appContenedor - zsh - 80x24' showing the output of the 'ls' command. The output lists files and directories: 'dist', 'docs', 'excel_processor.py', 'excel_processor.spec', 'hook-pytesseract.py', 'icon.ico', 'installation.txt', 'procesador_excel.py', 'procesador_pdf.py', 'requirements.txt', 'venv', and 'zafiro.png'.

```
Last login: Wed Jan  8 14:00:31 on ttys006
sennovapalmira@iMac-de-SENNOVA ~ % cd ~/Desktop/appContenedor
sennovapalmira@iMac-de-SENNOVA appContenedor % ls
Procesador.spec      dist                  installation.txt
ProcesadorExcel.spec docs                  procesador_excel.py
README.md            excel_processor.py   procesador_pdf.py
__pycache__          excel_processor.spec requirements.txt
app.py               hook-pytesseract.py venv
build                icon.ico              zafiro.png
sennovapalmira@iMac-de-SENNOVA appContenedor %
```

posterior a esto verificaremos que python este instalado correctamente con el comando:

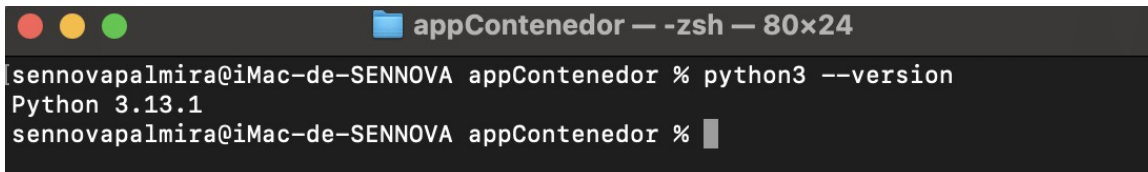
1. En Mac(deberas de probar con los sgtes):

1. `python3 --version`
2. `python --version`

2. En Windows (deberas de probar con los sgtes):

1. `python3 --version`
2. `python --version`
3. `py --version`

al escribir el comando debera de aparecer:

A terminal window titled 'appContenedor - zsh - 80x24' showing the output of the 'python3 --version' command, which is 'Python 3.13.1'.

```
sennovapalmira@iMac-de-SENNOVA appContenedor % python3 --version
Python 3.13.1
sennovapalmira@iMac-de-SENNOVA appContenedor %
```

Ahora deberemos de verificar que Tesseract-OCR este correctamente instalado escribiendo el comando:

`tesseract --version`

Deberas de ver algo como:

```
sennovapalmira@iMac-de-SENNOVA appContenedor % tesseract --version ]
tesseract 5.5.0
leptonica-1.85.0
libgif 5.2.2 : libjpeg 8d (libjpeg-turbo 3.0.4) : libpng 1.6.44 : libtiff 4.7.
0 : zlib 1.2.12 : libwebp 1.5.0 : libopenjp2 2.5.3
Found AVX2
Found AVX
Found FMA
Found SSE4.1
Found libarchive 3.7.7 zlib/1.2.12 liblzma/5.6.3 bz2lib/1.0.8 liblz4/1.10.0 lib
zstd/1.5.6
Found libcurl/8.7.1 SecureTransport (LibreSSL/3.3.6) zlib/1.2.12 nghttp2/1.62.0
sennovapalmira@iMac-de-SENNOVA appContenedor %
```

si no aparece el mensaje anterior deberas de volver al paso de agregar Variables de entorno.

Ahora deberemos de instalar las dependencias para que el proyecto funcione correctamente, corriendo los siguientes comandos:

crear e inciar entorno virtual.

En Windows

```
python -m venv venv
```

```
source venv\Scripts\activate
```

En Mac/Linux

```
python3 -m venv venv
```

```
source venv/bin/activate
```

deberas de ver algo como:

```
appContenedor — zsh — 80x24
sennovapalmira@iMac-de-SENNOVA appContenedor % python3 -m venv venv
source venv/bin/activate
(env) sennovapalmira@iMac-de-SENNOVA appContenedor %
```

Instalar dependencias

escribe el siguiente comando en la terminal, posterior a pegarlo presiona en “Enter”:

```
pip install pandas numpy openpyxl fpdf Pillow PyMuPDF python-docx pyesseract pdf2image opencv-python pyinstaller
```

o con el comando:

```
pip3 install pandas numpy openpyxl fpdf Pillow PyMuPDF python-docx pyesseract pdf2image opencv-python pyinstaller
```

Deberas de ver algo como:

```
appContenedor — zsh — 147x30
sennovapalmira@iMac-de-SENNOVA appContenedor % python3 -m venv venv
source venv/bin/activate
(env) sennovapalmira@iMac-de-SENNOVA appContenedor % pip3 install pandas numpy openpyxl fpdf Pillow PyMuPDF python-docx pyesseract pdf2image open
cv-python pyinstaller
Requirement already satisfied: pandas in ./venv/lib/python3.13/site-packages (2.2.3)
Requirement already satisfied: numpy in ./venv/lib/python3.13/site-packages (2.2.1)
Requirement already satisfied: openpyxl in ./venv/lib/python3.13/site-packages (3.1.5)
Requirement already satisfied: fpdf in ./venv/lib/python3.13/site-packages (1.7.2)
Requirement already satisfied: Pillow in ./venv/lib/python3.13/site-packages (11.0.0)
Requirement already satisfied: PyMuPDF in ./venv/lib/python3.13/site-packages (1.25.1)
Requirement already satisfied: python-docx in ./venv/lib/python3.13/site-packages (1.1.2)
Requirement already satisfied: pyesseract in ./venv/lib/python3.13/site-packages (0.3.13)
Requirement already satisfied: pdf2image in ./venv/lib/python3.13/site-packages (1.17.0)
Requirement already satisfied: opencv-python in ./venv/lib/python3.13/site-packages (4.10.0.84)
Requirement already satisfied: pyinstaller in ./venv/lib/python3.13/site-packages (6.11.1)
Requirement already satisfied: python-dateutil>=2.8.2 in ./venv/lib/python3.13/site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in ./venv/lib/python3.13/site-packages (from pandas) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in ./venv/lib/python3.13/site-packages (from pandas) (2024.2)
Requirement already satisfied: et-xmlfile in ./venv/lib/python3.13/site-packages (from openpyxl) (2.0.0)
Requirement already satisfied: lxml>=3.1.0 in ./venv/lib/python3.13/site-packages (from python-docx) (5.3.0)
Requirement already satisfied: typing-extensions>=4.9.0 in ./venv/lib/python3.13/site-packages (from python-docx) (4.12.2)
Requirement already satisfied: packaging>=21.3 in ./venv/lib/python3.13/site-packages (from pyesseract) (24.2)
Requirement already satisfied: setuptools>=42.0.0 in ./venv/lib/python3.13/site-packages (from pyinstaller) (75.7.0)
Requirement already satisfied: altgraph in ./venv/lib/python3.13/site-packages (from pyinstaller) (0.17.4)
Requirement already satisfied: pyinstaller-hooks-contrib>=2024.9 in ./venv/lib/python3.13/site-packages (from pyinstaller) (2024.11)
Requirement already satisfied: macholib>=1.8 in ./venv/lib/python3.13/site-packages (from pyinstaller) (1.16.3)
Requirement already satisfied: six>=1.5 in ./venv/lib/python3.13/site-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
(env) sennovapalmira@iMac-de-SENNOVA appContenedor %
```

Crear ejecutable de la aplicacion

Para crear el ejecutable de la aplicacion usaremos la misma consola, deberemos de escribir el comando en la terminal:

Nota: debes de tener en cuenta que el archivo icon.ico y la ruta de tesseract-ocr es la correcta.

```
pyinstaller --onefile --windowed --icon=icon.ico --name=Procesador --add-binary  
"/usr/local/bin/tesseract:." app.py
```

Debera de verse como:

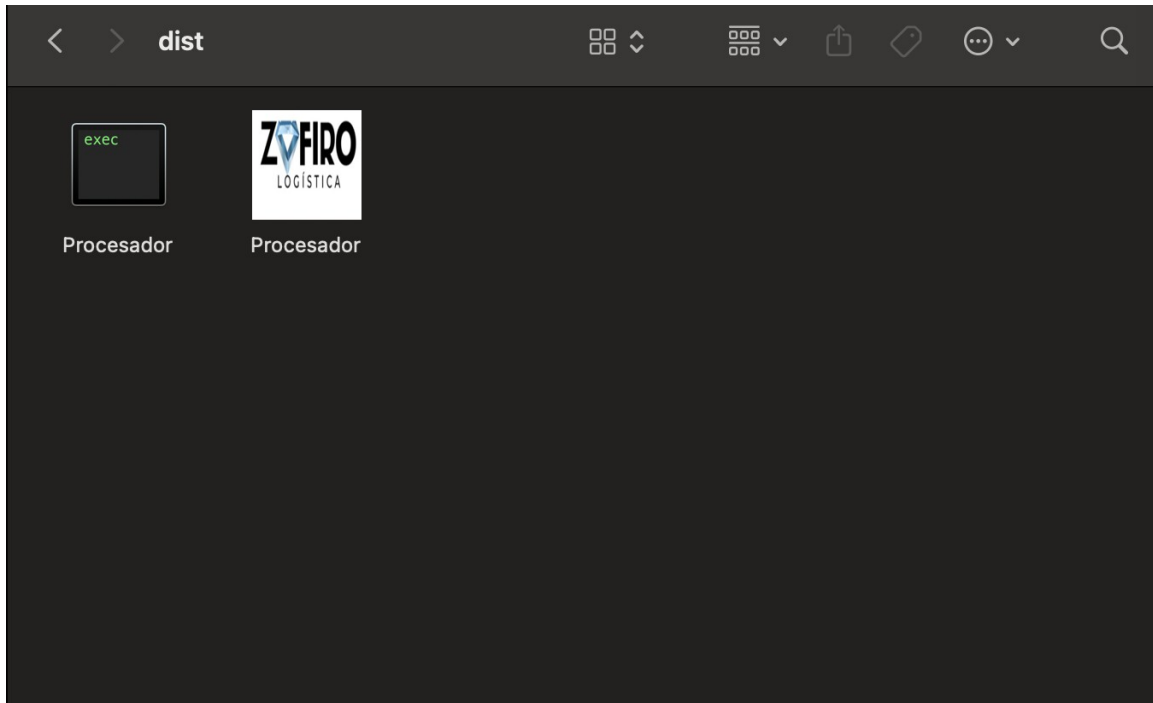
```
appContenedor — Python < pyinstaller --onefile --windowed --icon=icon.ico --name=Procesador --add-binary /usr/l...  
(venv) sennovapalmira@iMac-de-SENNOVA appContenedor % pyinstaller --onefile --windowed --icon=icon.ico --name=Proces  
ador --add-binary "/usr/local/bin/tesseract:." app.py  
  
146 INFO: PyInstaller: 6.11.1, contrib hooks: 2024.11  
146 INFO: Python: 3.13.1  
164 INFO: Platform: macOS-15.1.1-x86_64-i386-64bit-Mach-O  
164 INFO: Python environment: /Users/sennovapalmira/Desktop/appContenedor/venv  
165 INFO: wrote /Users/sennovapalmira/Desktop/appContenedor/Procesador.spec  
169 INFO: Module search paths (PYTHONPATH):  
['/usr/local/Cellar/python@3.13/3.13.1/Frameworks/Python.framework/Versions/3.13/lib/python313.zip',  
 '/usr/local/Cellar/python@3.13/3.13.1/Frameworks/Python.framework/Versions/3.13/lib/python3.13',  
 '/usr/local/Cellar/python@3.13/3.13.1/Frameworks/Python.framework/Versions/3.13/lib/python3.13/lib-dynload',  
 '/Users/sennovapalmira/Desktop/appContenedor/venv/lib/python3.13/site-packages',  
 '/usr/local/opt/python-tk@3.13/libexec',  
 '/Users/sennovapalmira/Desktop/appContenedor/venv/lib/python3.13/site-packages/setuptools/_vendor',  
 '/Users/sennovapalmira/Desktop/appContenedor']  
474 INFO: Appending 'binaries' from .spec  
474 INFO: checking Analysis  
499 INFO: Building because /Users/sennovapalmira/Desktop/appContenedor/procesador_pdf.py changed  
499 INFO: Running Analysis Analysis-00.toc  
499 INFO: Target bytecode optimization level: 0  
499 INFO: Initializing module dependency graph...  
500 INFO: Initializing module graph hook caches...  
511 INFO: Analyzing base_library.zip ...  
1210 INFO: Processing standard module hook 'hook-heapq.py' from '/Users/sennovapalmira/Desktop/appContenedor/venv/li  
b/python3.13/site-packages/PyInstaller/hooks'  
1477 INFO: Processing standard module hook 'hook-encodings.py' from '/Users/sennovapalmira/Desktop/appContenedor/ven  
v/lib/python3.13/site-packages/PyInstaller/hooks'  
2809 INFO: Processing standard module hook 'hook-pickle.py' from '/Users/sennovapalmira/Desktop/appContenedor/venv/li  
b/python3.13/site-packages/PyInstaller/hooks'  
4359 INFO: Caching module dependency graph...  
4420 INFO: Looking for Python shared library...  
4424 INFO: Using Python shared library: /usr/local/Cellar/python@3.13/3.13.1/Frameworks/Python.framework/Versions/3.  
13/Python  
4424 INFO: Analyzing /Users/sennovapalmira/Desktop/appContenedor/app.py  
4433 INFO: Processing pre-find-module-path hook 'hook-tkinter.py' from '/Users/sennovapalmira/Desktop/appContenedor/  
venv/lib/python3.13/site-packages/PyInstaller/hooks/pre_find_module_path'  
4434 INFO: TclTkInfo: initializing cached Tcl/Tk info...  
4664 INFO: Processing standard module hook 'hook-tkinter.py' from '/Users/sennovapalmira/Desktop/appContenedor/venv  
/lib/python3.13/site-packages/PyInstaller/hooks'  
4850 INFO: Processing standard module hook 'hook-pandas.py' from '/Users/sennovapalmira/Desktop/appContenedor/venv/li  
b/python3.13/site-packages/PyInstaller/hooks'
```

el proceso de instalacion dependera de tu conexion a internet y las características de tu dispositivo.

El proceso culmina de forma exitosa si ves el mensaje:

```
On your own risk, you can use the option '--noconfirm' to get rid of this question.  
61935 INFO: Removing dir /Users/sennovapalmira/Desktop/appContenedor/dist/Procesador.app  
61939 INFO: Building BUNDLE BUNDLE-00.toc  
62227 INFO: Signing the BUNDLE...  
62688 INFO: Building BUNDLE BUNDLE-00.toc completed successfully.  
(venv) sennovapalmira@iMac-de-SENNOVA appContenedor %
```

ahora deberas de dirgirte a la carpeta appContenedor > Dist, aqui encontraras el ejecutable de la aplicacion:



Recomendacion: Crear un acceso directo en el escritorio y agregarlo a la barra de tareas.

Utilidades

En la carpeta **appContenedor** se encuentra el **código fuente** del software. Además, dentro del archivo **installation.txt** se proporcionan instrucciones detalladas para instalar, configurar y ejecutar el programa, suministradas por el desarrollador.

El programa está compuesto por **cuatro archivos principales**, cada uno con una función clave:

1. *App.py*

Este es el **archivo principal** que sirve como punto de entrada para el programa. Su función es administrar las diferentes ventanas de la interfaz gráfica, permitiendo la navegación entre módulos como el **Procesador Excel** y el **Procesador PDF**. Este archivo asegura una comunicación eficiente entre los componentes del software.

2. Procesador_excel.py

- **Descripción:**
Código fuente completamente documentado, diseñado con una estructura **orientada a objetos** basada en clases padre e hijas.
 - **Estructura:**
Cada clase contiene métodos específicos para configurar las ventanas gráficas y desarrollar la lógica de procesamiento. Este archivo incluye toda la lógica necesaria para manipular y procesar archivos Excel de manera **modular y escalable**.
 - **Principales funcionalidades:**
 - Categorización automatizada por criterios definidos.
 - Aplicación de fórmulas, eliminación de columnas y generación de informes en formatos Excel y PDF.
 - Eficiencia mejorada al reducir significativamente el tiempo de procesamiento.
-

3. Procesador_pdf.py

- **Descripción:**
Este archivo contiene el código dedicado al manejo y procesamiento de archivos PDF. Siguiendo los principios de diseño modular, integra funcionalidades avanzadas para convertir PDFs escaneados en datos estructurados mediante **OCR (Reconocimiento Óptico de Caracteres)**.
 - **Características clave:**
 - Uso de **Tesseract OCR** para leer y convertir imágenes en texto procesable.
 - Preprocesamiento de imágenes utilizando **OpenCV** para mejorar la precisión de reconocimiento.
 - Clasificación y categorización de datos por cliente o subpartidas.
 - Exportación automatizada en formatos PDF y Excel.
-

4. Otros Archivos Complementarios

El software también incluye scripts adicionales para manejar configuraciones específicas, documentación de soporte y archivos necesarios para el funcionamiento de bibliotecas externas. Como el archivo (hook-pytesseract.py)