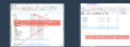
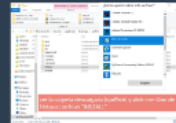


Extracción de texto



1. Guardar carpeta con múltiples pdfs en escritorio
2. Descargar: <https://www.xpdfreader.com/download.html>

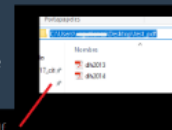


Microsoft Windows [versión 10.0.17134.0]
(c) 2018 Microsoft Corporation. Todos los derechos reservados.
C:\Users\salvo>cd Desktop\test



Escribe cd seguido de la dirección en donde
están tus pdfs, ejemplo:

```
C:\WINDOWS\system32>cd C:\Users\salvo\Desktop\test_pdf
```



Ingresar las palabras mágicas

```
C:\Users\salvo\Desktop\test_pdf>FORFILES /M *.pdf /C "cmd /c pdftotext @file"
```

```
FORFILES /M *.pdf /C "cmd /c pdftotext @file"
```

Extraccción de texto

Silvia Gutiérrez De la Torre



@espejolento



BIBLIOTECA

DANIEL COSÍO VILLEGAS

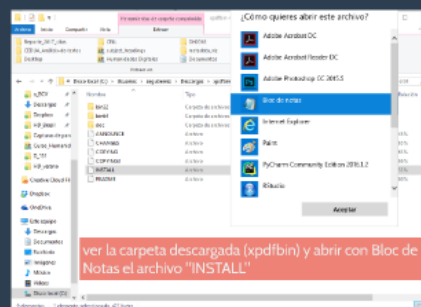
Extracción de texto

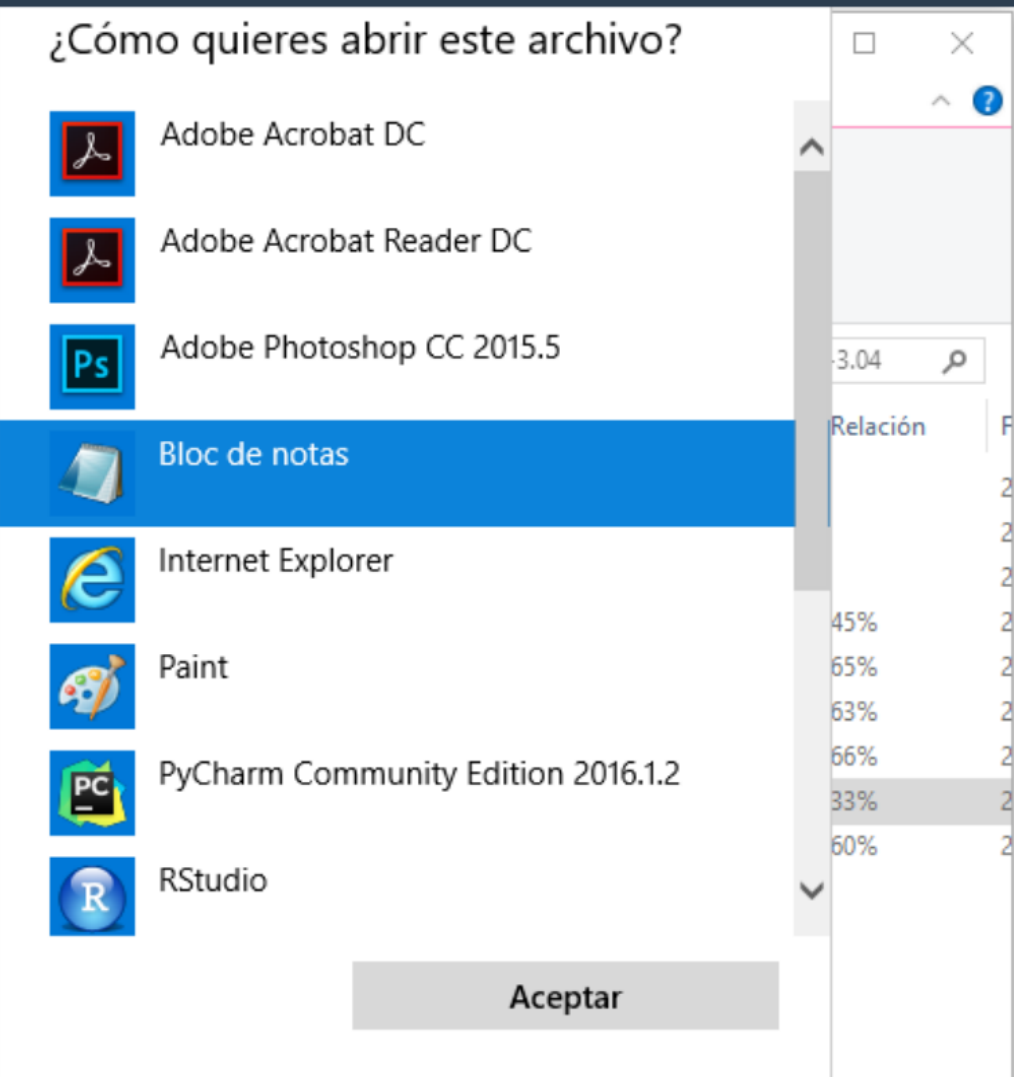
Silvia Gutiérrez De la Torre

 @espejolento



1. Guardar carpeta con múltiples pdfs en escritorio
2. Descargar: <https://www.xpdfreader.com/download.h>





ver la carpeta descargada (xpdfbin) y abrir con Bloc de Notas el archivo "INSTALL"



INSTALL: Bloc de notas

Archivo Edición Formato Ver Ayuda

Xpdf - Win32 binaries

=====

The Xpdf software and documentation are
copyright 1996-2014 Glyph & Cog, LLC.

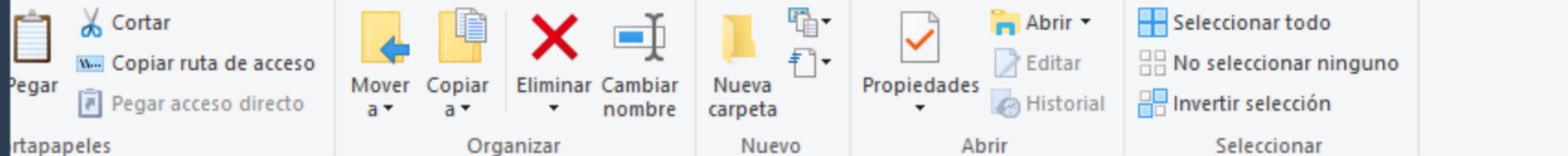
Email: derekn@foolabs.com

WWW: <http://www.foolabs.com/xpdf/>

To install this binary package:

1. Copy everything to an installation directory, e.g.,
C:/Program Files/Xpdf
2. Edit the xpdfrc file (as distributed, everything is commented out)
-- see xpdfrc.txt for details.

seguir instrucciones 1 y 2
(detalles a continuación)



Nombre	Fecha de modifica...	Tipo	Tamaño
Dropbox	18/07/2017 11:09 a...	Carpeta de archivos	
Evernote	06/05/2016 04:36 ...	Carpeta de archivos	
Gephi-0.9.1	02/08/2016 04:02 ...	Carpeta de archivos	
Google	25/02/2016 12:39 ...	Carpeta de archivos	
Harvard University	15/06/2017 05:47 ...	Carpeta de archivos	
Hewlett-Packard	25/02/2016 01:44 ...	Carpeta de archivos	
HTC	06/05/2016 07:05 ...	Carpeta de archivos	
Internet Explorer	15/06/2017 05:47 ...	Carpeta de archivos	
Java	18/01/2017 10:55 a...	Carpeta de archivos	
JetBrains	27/04/2016 03:12 ...	Carpeta de archivos	
McAfee	02/08/2016 05:44 ...	Carpeta de archivos	
Messenger for Desktop	11/04/2016 11:11 a...	Carpeta de archivos	
Microsoft Analysis Services	25/02/2016 12:18 ...	Carpeta de archivos	
Microsoft Office	25/02/2016 12:18 ...	Carpeta de archivos	
Microsoft SQL Server	25/02/2016 12:22 ...	Carpeta de archivos	
Microsoft.NET	19/05/2017 11:28 a...	Carpeta de archivos	
Mozilla Firefox	18/07/2017 11:05 a...	Carpeta de archivos	
Mozilla Maintenance Service	18/07/2017 11:05 a...	Carpeta de archivos	
MSBuild	19/05/2017 11:55 a...	Carpeta de archivos	
Nero	25/02/2016 02:12 ...	Carpeta de archivos	
Notepad++	27/04/2016 03:10 ...	Carpeta de archivos	
XPdf	26/07/2017 12:30 ...	Carpeta de archivos	

instrucción 1 nos dice que en nuestros "Archivos de Programa" debemos crear la carpeta Xpdf

Windows Explorer window titled "bin64" showing the contents of the "Xpdf" folder on the "Disco local (C:)" drive. The address bar shows the path: "Este equipo > Disco local (C:) > Archivos de programa (x86) > Xpdf > bin64". The search bar contains "Buscar en bin64".

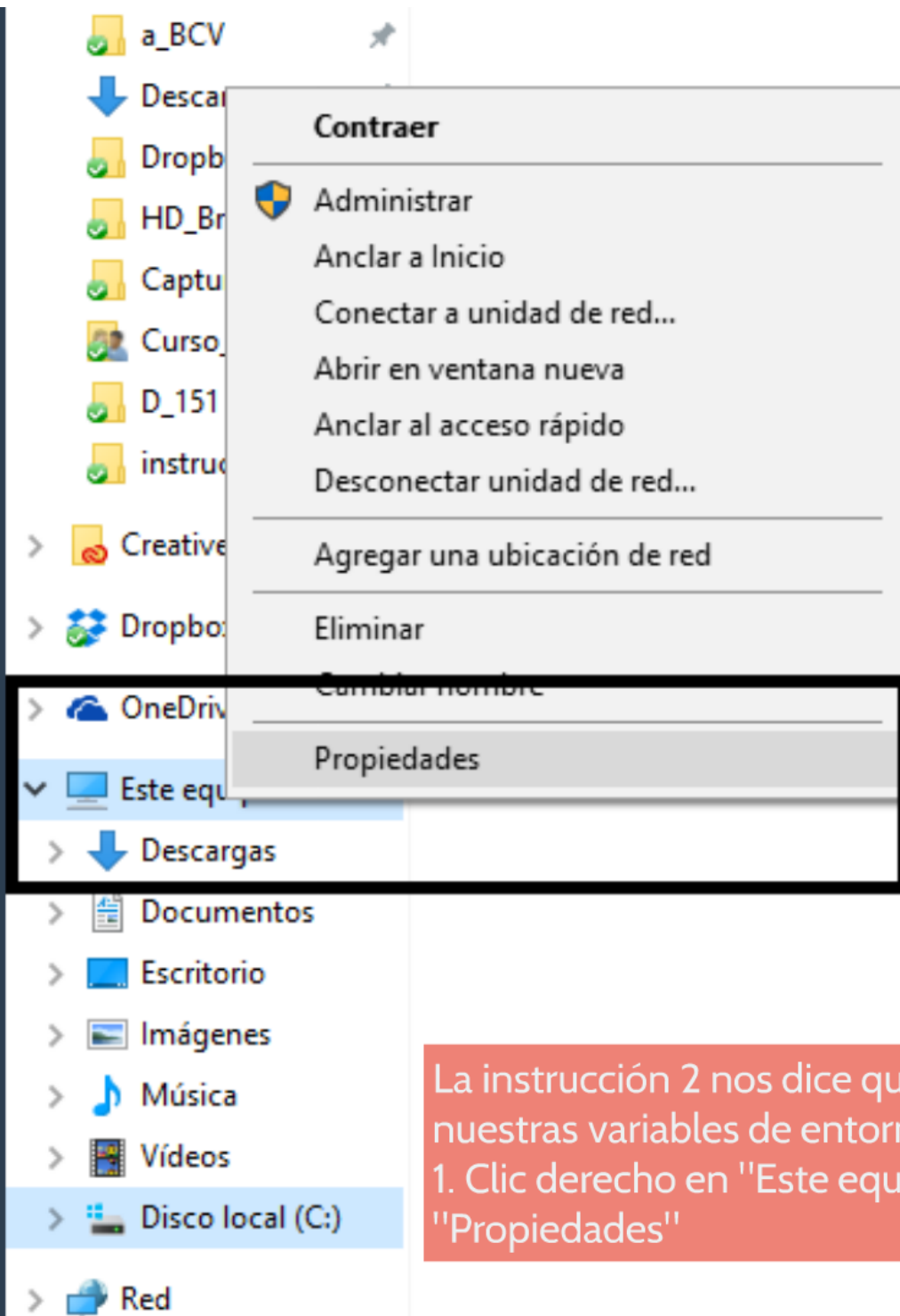
The ribbon includes the following tabs: "Herramientas de aplicación", "Administrar", "Compartir", and "Vista". The "Herramientas de aplicación" tab is active, showing the following groups of icons:

- Organizar: Mover a, Copiar a, Eliminar, Cambiar nombre
- Nuevo: Nueva carpeta
- Abir: Abrir, Editar, Historial, Propiedades
- Seleccionar: Seleccionar todo, No seleccionar ninguno, Invertir selección

The file list shows the following files:

Nombre	Fecha de modifica...	Tipo	Tamaño
pdfdetach	26/07/2017 12:30 ...	Aplicación	1,061 KB
pdffonts	26/07/2017 12:30 ...	Aplicación	1,078 KB
pdfimages	26/07/2017 12:30 ...	Aplicación	1,086 KB
pdfinfo	26/07/2017 12:30 ...	Aplicación	1,076 KB
pdftohtml	26/07/2017 12:30 ...	Aplicación	2,393 KB
pdftopng	26/07/2017 12:30 ...	Aplicación	2,240 KB
pdftoppm	26/07/2017 12:30 ...	Aplicación	2,061 KB
pdftops	26/07/2017 12:30 ...	Aplicación	2,196 KB
pdftotext	26/07/2017 12:30 ...	Aplicación	1,152 KB

asegúrense que en bin64 estén todos estos archivos



La instrucción 2 nos dice que debemos añadir Xpdf a nuestras variables de entorno eso se hace así:
1. Clic derecho en "Este equipo" -> seleccionar "Propiedades"



Panel de control > Sistema y seguridad > Sistema

Ventana principal del Panel de control

- Administrador de dispositivos
- Configuración de Acceso remoto
- Protección del sistema
- Configuración avanzada del sistema

Ver información básica acerca del equipo

Propiedades del sistema

Nombre del equipo Hardware

Opciones avanzadas Protección del sistema Remoto

Para realizar la mayoría de estos cambios, inicie sesión como administrador.

Rendimiento
Efectos visuales, programación del procesador, uso de memoria y memoria virtual
Configuración...

Perfiles de usuario
Configuración del escritorio correspondiente al inicio de sesión
Configuración...

Inicio y recuperación
Inicio del sistema, errores del sistema e información de depuración
Configuración...

Variables de entorno...

Aceptar Cancelar Aplicar

Id. del producto: 00330-50029-39940-AAOEM

Variables de entorno

Variables de usuario para segutierrez

Variable	Valor
Path	C:\Users\segutierrez\AppData\Local\Programs\Python\Python35-32\...
TEMP	C:\Users\segutierrez\AppData\Local\Temp
TMP	C:\Users\segutierrez\AppData\Local\Temp

Nuevo... Editar... Eliminar

Variables del sistema

Variable	Valor
ComSpec	C:\WINDOWS\system32\cmd.exe
ESET_OPTIONS	
MALLET_HOME	C:\mallet-2.0.8\
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\ProgramData\Oracle\Java\javapath;C:\WINDOWS\system32;C:\W...
Path	Q:\entorno\bin;CMD;VBS;VBE;JS;JSE;WSF;WSH;MSC...

Nueva... Editar... Eliminar

Aceptar Cancelar

Editar variable de entorno


C:\Users\segutierrez\AppData\Local\Programs\Python\Python35-32\...
C:\Users\segutierrez\AppData\Local\Programs\Python\Python35-32\
C:\Program Files\DPF Manager
C:\Users\segutierrez\AppData\Local\Microsoft\WindowsApps
C:\Program Files (x86)\Xpdf\

Nuevo Editar Examinar... Eliminar Subir Bajar Editar texto...


Aceptar Cancelar

2. Seguir estos cuatro pasos (OJO: empieza con seleccionar "Configuración avanzada del sistema")

Mejor coincidencia

 **Símbolo del sistema**
Aplicación

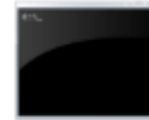
Aplicaciones

 **Anaconda Prompt (Anaconda3)** >

Buscar en Internet


🔍 **cmd** - Ver resultados web >

Configuración (1)




Símbolo del sistema
Aplicación

 **Abrir**

 **Ejecutar como administrador**

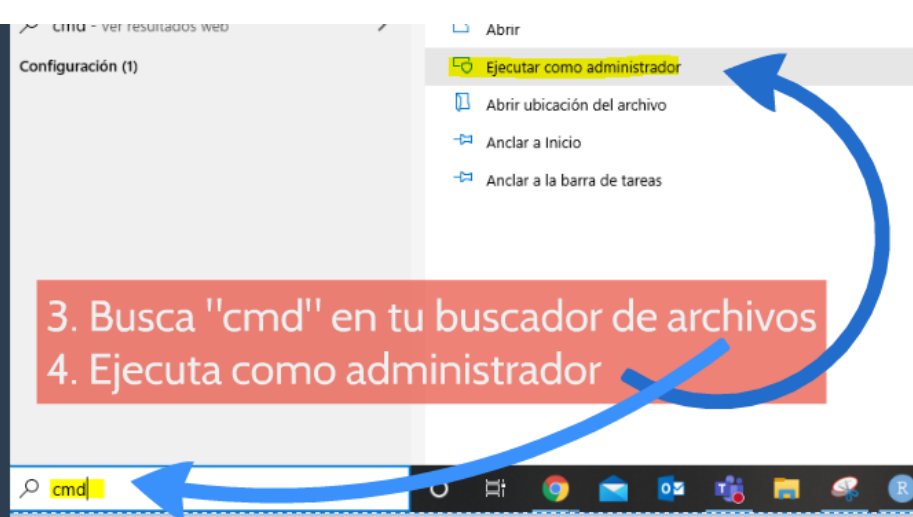
 Abrir ubicación del archivo

 Anclar a Inicio

 Anclar a la barra de tareas

3. Busca "cmd" en tu buscador de archivos
4. Ejecuta como administrador

🔍 **cmd**



```
Microsoft Windows [Versión 10.0.15063]  
(c) 2017 Microsoft Corporation. Todos los derechos reservados.  
C:\WINDOWS\system32>pdftotext
```

5. escribe "pdftotext" y da clic en Enter

```
Microsoft Windows [Versión 10.0.15063]  
(c) 2017 Microsoft Corporation. Todos los derechos reservados.  
C:\WINDOWS\system32>pdftotext  
pdftotext version 3.04  
Copyright 1996-2014 Glyph & Cog, LLC  
Usage: pdftotext [options] <PDF-file> [<text-file>]  
-f <int>                : first page to convert  
-l <int>                : last page to convert  
-layout                 : maintain original physical layout
```

si te aparece esto, los pasos anteriores están bien

Microsoft Windows [Versión 10.0.15063]

(c) 2017 Microsoft Corporation. Todos los derechos reservados.

C:\WINDOWS\system32>pdftotext

pdftotext version 3.04

Copyright 1996-2014 Glyph & Cog, LLC

Usage: pdftotext [options] <PDF-file> [<text-file>]

-f <int> : first page to convert
-l <int> : last page to convert
-layout : maintain original physical layout
-table : similar to -layout, but optimized for tables
-lineprinter : use strict fixed-pitch/height layout
-raw : keep strings in content stream order
-fixed <fp> : assume fixed-pitch (or tabular) text
-linespacing <fp> : fixed line spacing for LinePrinter mode
-clip : separate clipped text
-enc <string> : output text encoding name
-eol <string> : output end-of-line convention (unix, dos, or mac)
-nopgbrk : don't insert page breaks between pages
-opw <string> : owner password (for encrypted files)
-upw <string> : user password (for encrypted files)
-q : don't print any messages or errors
-cfg <string> : configuration file to use in place of .xpdfrc
-v : print copyright and version info
-h : print usage information
-help : print usage information
--help : print usage information
-? : print usage information

si te aparece esto, los pasos anteriores están bien

```

Microsoft Windows [Versión 10.0.15063]
(c) 2017 Microsoft Corporation. Todos los derechos reservados.

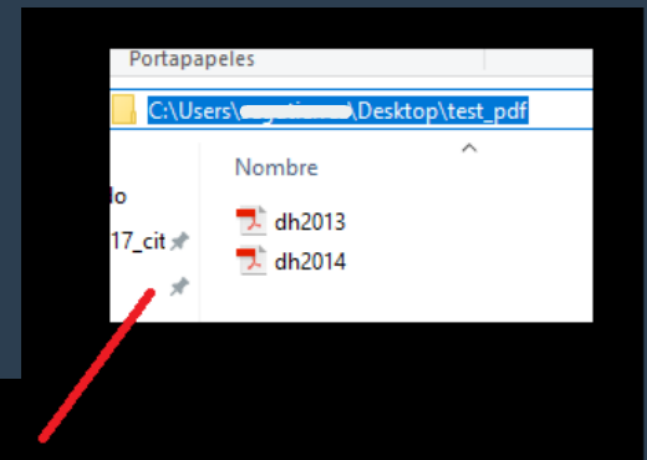
C:\WINDOWS\system32>pdftotext
pdftotext ver
Copyright 1996-2017 by Glyph & Associates
Usage: pdftotext [options] <PDF-file> [<text-file>]
-f <int>          : first page to convert
-l <int>          : last page to convert
-layout          : maintain original physical layout
-table          : similar to -layout, but optimized for tables
-lineprinter     : use strict fixed-pitch/height layout
-raw            : keep strings in content stream order
-fixed <fp>      : assume fixed-pitch (or tabular) text
-linespacing <fp> : fixed line spacing for LinePrinter mode
-clip           : separate clipped text
-enc <string>    : output text encoding name
-eol <string>    : output end-of-line convention (unix, dos, or mac)
-nogbmk         : don't insert page breaks between pages
-opw <string>    : owner password (for encrypted files)
-upw <string>    : user password (for encrypted files)
-q             : don't print any messages or errors
-cfg <string>    : configuration file to use in place of .xpdfrc
-v            : print copyright and version info
-h            : print usage information
-help        : print usage information
--help       : print usage information
-?           : print usage information

```

si te aparece esto, los pasos anteriores están bien

Escribe cd seguido de la dirección en donde están tus pdfs, ejemplo:

```
C:\WINDOWS\system32>cd C:\Users\[usuario]\Desktop\test_pdf
```



Ingresa las palabras mágicas

```
ers\[usuario]\Desktop\test_pdf>FORFILES /M *.pdf /C "cmd /c pdftotext @file"
```

```
C:\WINDOWS\system32>cd C:\Users\[redacted]\Desktop\test_pdf
```

Ingresar las palabras mágicas

```
C:\Users\[redacted]\Desktop\test_pdf>FORFILES /M *.pdf /C "cmd /c pdftotext @file"
```

FORFILES /M *.pdf /C "cmd /c pdftotext @file"