**How to extract information and digitizing data from plots in the literature?**

1. First, using suitable keyword on google scholar download and collect all papers as pdf format inside of a folder.
2. After collecting all papers, then Run “Extract\_PDF2Images\_Ali.py” or “Extract\_Images\_PDF.py” inside of that folder

To run ” Extract\_Images\_PDF.py”” code you need to have “pdfimages.exe” code in the folder. Sometimes the format of images for different journals are different, if there is an issue, you can use another python code.

This code will extract all images from a pdf as “jpg” or “png” and will copy it in a folder (pdf\_name\_IMAGES) and will copy the pdf file into this folder and a generate two txt files:

* “**Pdf\_name.txt**” : This should be used for extracting all datapoint from images by user
* **“Pdf\_name\_REF.txt”**: includes reference information like Author, Title, Subject. This would be extracted automatically

1. Go to each folder and open those images for data extraction (sometimes you need to crop images) and use **webplotdigitzer (**<https://automeris.io/WebPlotDigitizer/>**)** to extract datapoints from those plots.

How you can use webplotdigitizer see (<https://www.youtube.com/watch?v=BfPyWgWWL_8>)

1. After digitizing specific plot, then you need to copy, paste the datapoints should in pdf\_name.txt with similar path as follows

x= ??, y= ???, keyword\_1=value1 , keyword\_2=value2, keyword\_3=value3 , …

X1, Y1

X2, Y2  
..,….

Xn, Xn

Blank space

x= ??, y= ???, keyword\_1=value1 , keyword\_2=value2, keyword\_3=value3 , …

X1, Y1

X2, Y2  
..,….

Xn, Xn

Blank space

…

…

..

1. After completing all data extraction from those papers, then collect all .txt files in a folder by running the below python code

You can run the following code “python Collect\_all\_txt.py” to collect all txt files and REF files in a folder. You need to give the path of destination folder you want to collect txt files there

# =============================================================================

import glob, os, shutil

path = os.getcwd()

# Give the path of the folder you want to collect all data

dest = "C:\\Users\\alima\\OneDrive\\Desktop\\Textfiles\\All\_TXT\_FILES"

# list all the directories in current directory

dirs=[ x for x in os.listdir('.') if os.path.isdir(x) ]

for d in dirs:

# list all files in A/\*, B/\*, C/\*...

files\_to\_copy = os.listdir(os.path.join(d))

for f in files\_to\_copy:

if f.endswith(".txt"): ## copy the relevant files to dest

shutil.copy(os.path.join(d, f), dest)

=============================================================================

1. Give all keywords in txt files in “keywords.xlsx”
2. Then in this folder, run the below python code to extract all datapoints and information as csv format with their corresponding references.

“python main.py”

To run this, you need to have “ExtractDataFromText.py” inside of folder