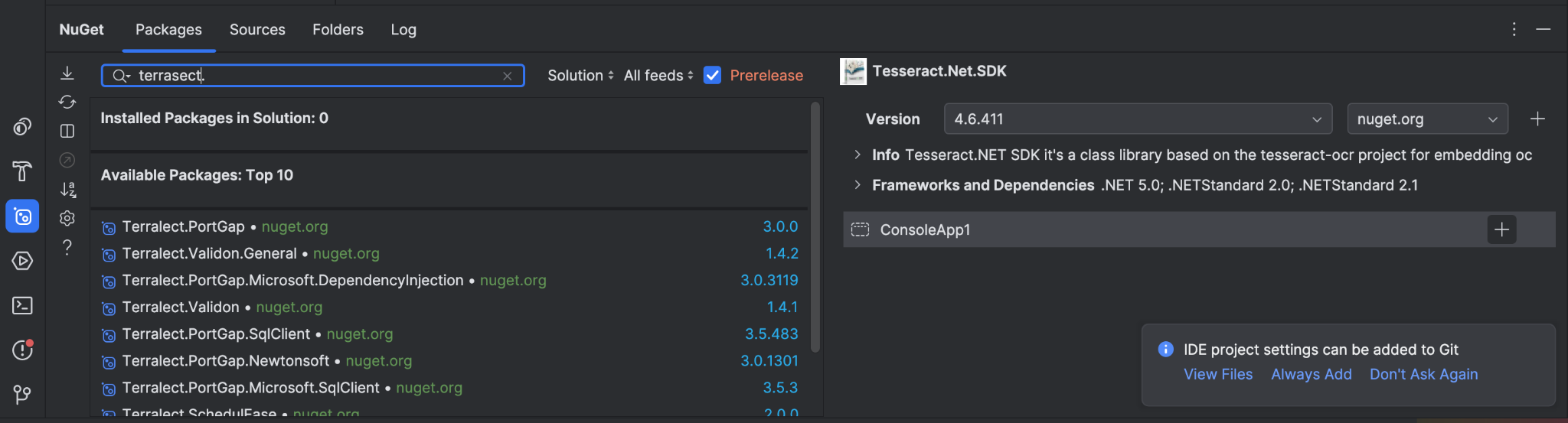
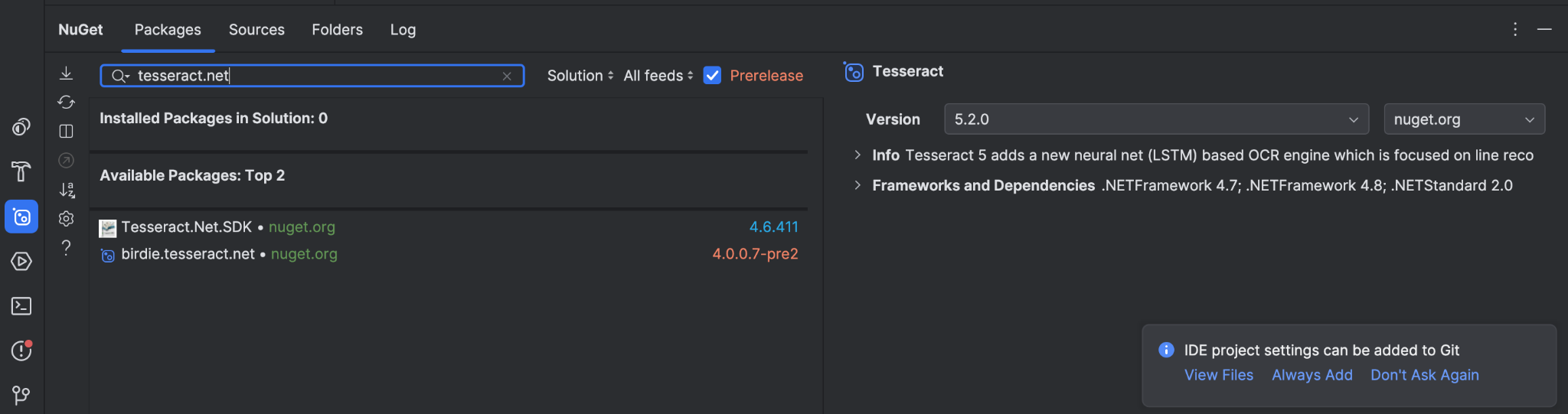


For our Project topic Creating Text from images with OCR API, it has been mentioned that we should use Terrasect. But upon research we saw that Tesseract is free and terrasect is paid and professor Dobric has told to only use free version.





Nuget is present is present only for Tesseract not Terrasect

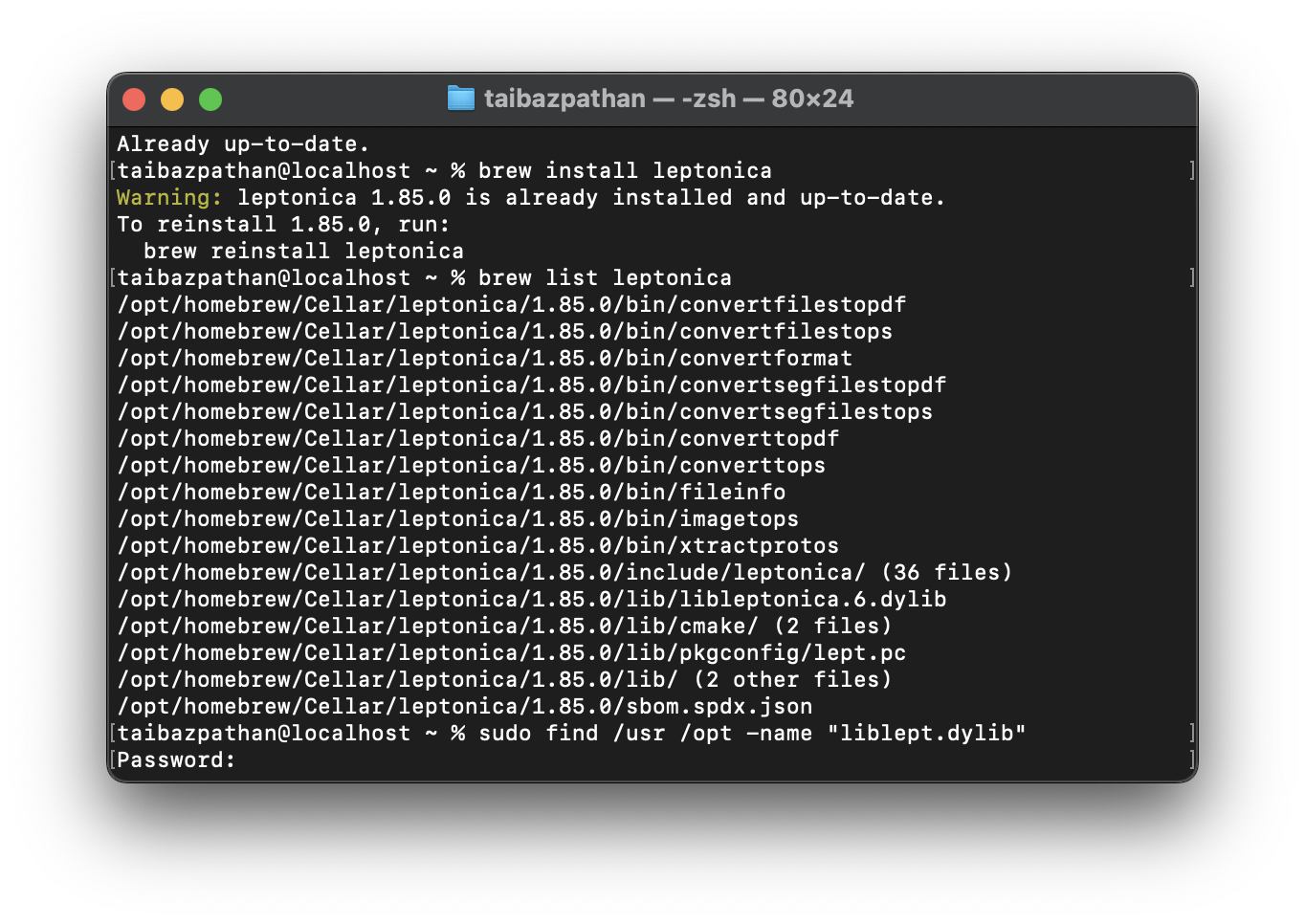
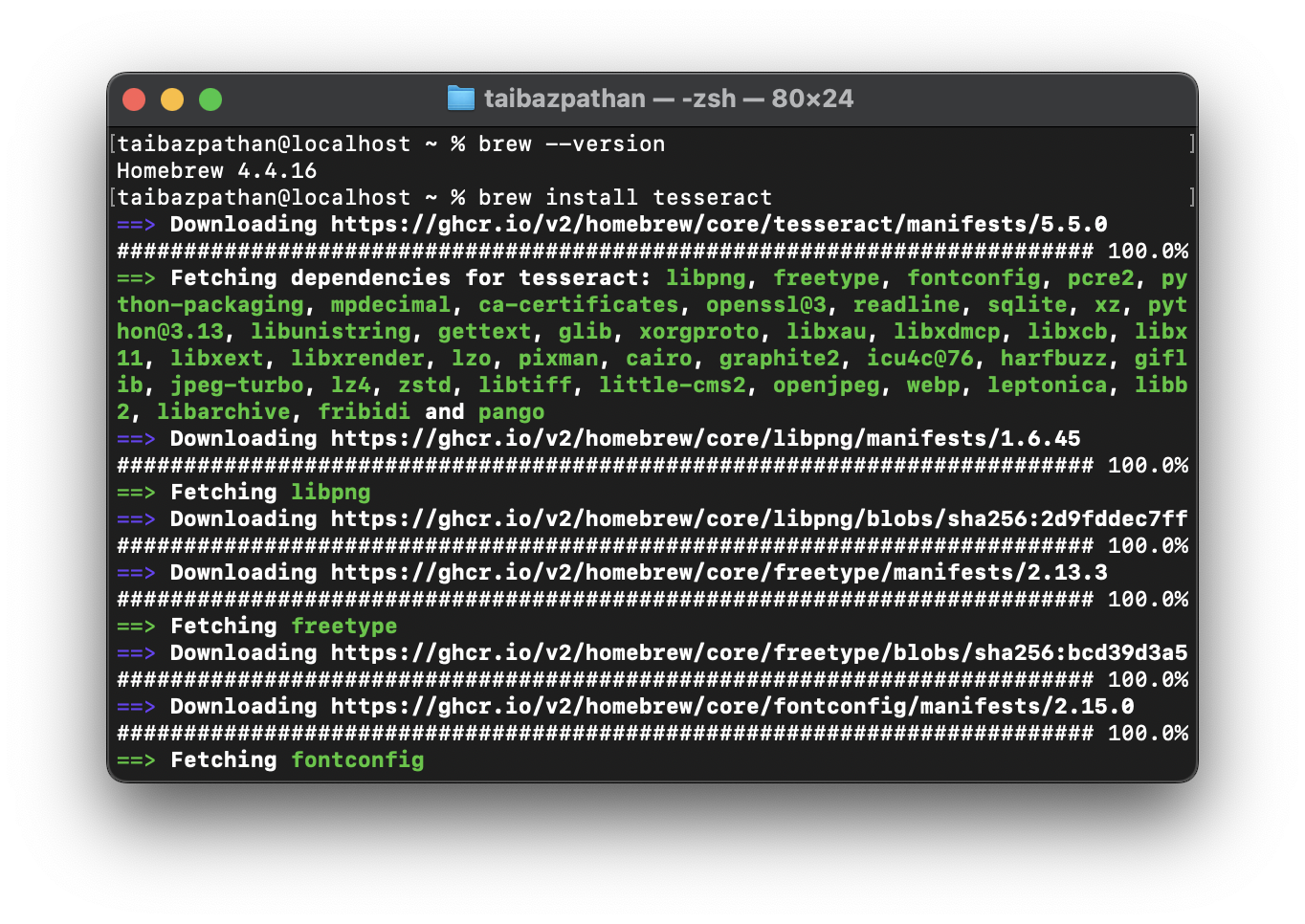


In the Document where project are described there is a potential mistake that Tesseract is replaced with Terrasect.

Installed Tesseract and Leptonica and Brew :

A screenshot of a computer

Description automatically generated



Text Extraction is working in console but not working in IDE Getting Error :

System.Reflection.TargetInvocationException: Exception has been thrown by the target of an invocation.

---> System.DllNotFoundException: Failed to find library "libleptonica-1.82.0.dylib" for platform x64.

at InteropDotNet.LibraryLoader.LoadLibrary(String fileName, String platformName)

at InteropRuntimeImplementer.LeptonicaApiSignaturesInstance.LeptonicaApiSignaturesImplementation..ctor(LibraryLoader loader)

at System.RuntimeMethodHandle.InvokeMethod(Object target, Void\*\* arguments, Signature sig, Boolean isConstructor)

at System.Reflection.MethodBaseInvoker.InvokeDirectByRefWithFewArgs(Object obj, Span`1 copyOfArgs, BindingFlags invokeAttr)

--- End of inner exception stack trace ---

at System.Reflection.MethodBaseInvoker.InvokeDirectByRefWithFewArgs(Object obj, Span`1 copyOfArgs, BindingFlags invokeAttr)

at System.Reflection.MethodBaseInvoker.InvokeWithOneArg(Object obj, BindingFlags invokeAttr, Binder binder, Object[] parameters, CultureInfo culture)

at System.Reflection.RuntimeConstructorInfo.Invoke(BindingFlags invokeAttr, Binder binder, Object[] parameters, CultureInfo culture)

at System.RuntimeType.CreateInstanceImpl(BindingFlags bindingAttr, Binder binder, Object[] args, CultureInfo culture)

at InteropDotNet.InteropRuntimeImplementer.CreateInstance[T]()

at Tesseract.Interop.LeptonicaApi.Initialize()

at Tesseract.Interop.TessApi.Initialize()

at Tesseract.Interop.TessApi.get\_Native()

at Tesseract.TesseractEngine..ctor(String datapath, String language, EngineMode engineMode, IEnumerable`1 configFiles, IDictionary`2 initialOptions, Boolean setOnlyNonDebugVariables)

at Tesseract.TesseractEngine..ctor(String datapath, String language, EngineMode engineMode)

at OCRProcessor.Program.Main() in /Users/taibazpathan/Desktop/Git/test/ConsoleAppTesst/ConsoleAppTesst/Program.cs:line 13

Process finished with exit code 0.

Preprocessing Steps to Improve OCR Accuracy:

• Binarization: Converting the image to black and white to enhance text visibility.

• Noise Removal: Eliminating unwanted pixels or artifacts that may interfere with text recognition.

• Thinning and Skeletonization: Reducing the width of text strokes to a single pixel to simplify character shapes.

**Trying to run ocr but getting some errors :**

The code I have used to implement :

using System;

using Tesseract;

class Program

{

static void Main()

{

string imagePath = "test-image.png"; // Ensure this file is present in the project folder

string tessDataPath = @"./tessdata"; // Download and place the tessdata folder

var ocrEngine = new TesseractEngine(tessDataPath, "eng", EngineMode.Default);

using (var img = Pix.LoadFromFile(imagePath))

{

using (var page = ocrEngine.Process(img))

{

Console.WriteLine("Extracted Text:");

Console.WriteLine(page.GetText());

}

}

}

}

Research about datasets I can use for OCR project:

1. **IIIT 5K-Word Dataset (Scene Text)** - Recognizing text from **natural scene images**: <https://cvit.iiit.ac.in/research/projects/cvit-projects/the-iiit-5k-word-dataset>
2. **ICDAR (International Conference on Document Analysis and Recognition) -**Handwritten & printed text OCR, scanned documents, noisy text.[**https://rrc.cvc.uab.es/**](https://rrc.cvc.uab.es/)
3. **Synthetic Word Dataset (SynthText) - Best for:** Training OCR models on large amounts of synthetic text. https://www.robots.ox.ac.uk/~vgg/data/text/