This is a bug example to showcase a bug format.

The bug below is based on an actual bug in Tesseract. Note that it deals with parameters that are not covered in the exercise (*"text in the wild", colored backgrounds, 'deu' language etc*), so it is just an example.

Note that we are not attaching the images and files referenced in this bug; it's not important; this is an example how a bug report should be written. Note however that when *you* submit bugs that reference images and files – we certainly do expect you to provide the referenced files!

Background border causes severe OCR failure on low resolution TIFF images

Description:

Trying OCR of a high resolution full scan of book page with colored border (book background) generates an answer almost entirely devoid of text. See the attached image `fullborder_lowres.tif` which OCRs to the blank `fullborder_lowres.txt`.

Steps to reproduce:

1) Perform OCR on image `whiteborder_lowres.tif`. Suggested command:
"`tesseract.exe whiteborder_lowres.tif stdout -l deu`"

2) Compare OCR results with the image contents. Results will be very similar to `whiteborder_lowres.txt`.

3) Perform OCR on image `fullborder_lowres.tif`. Suggested command:
"`tesseract.exe fullborder_lowres.tif stdout -l deu`"

4) Compare OCR results with the image contents.

Expected Result: Results similar to what we got in step (1). See attachment `whiteborder_lowres.txt` for example.

Actual Results: The OCR results are mostly empty. See attached `fullborder_lowres.txt` for example.

Setup and Additional Info:

Version of Tesseract used: **tesseract v5.0.0-alpha.20201127**

No config file or config parameter was provided, all the OCR is based on the defaults.

OS: Windows 10, Version 10.0.18363.1621

Executed on Lenovo T480, i5-8350U CPU

Additional Information

It should be noted that:

- a high resolution image with the same colored border succeeds OCR. See example `fullborder_highres.tif` attached (OCRs to file `fullborder_highres.txt`).

- an image of the same resolution succeeds OCR if the border is recolored. See example `whiteborder_lowres.tif` attached (OCRs to file `whiteborder_lowres.txt`).

This shows that the resolution is enough for character recognition, and that Tesseract can identify and ignore the border, however Tesseract is not able to both ignore the border while under the combination of both parameters.

Suggested Severity: Medium

Rationale: As long as a human can easily read the image, we expect Tesseract to OCR it correctly.

Suggesting a Medium severity due to the specificity of this image format and the combination of the two parameters.