Using Tesseract OCR with Java

# Introduction

[Tesseract](https://github.com/tesseract-ocr/tesseract) is one of the most popular open source Optical Character Recognition systems around. It supports many languages. It is written in C++ and needs a lot of other libraries as well to work. This blog assumes that you are already familiar with Tesseract and how it works.

# Implementation Details

We would use the [bytedeco javacpp-presets](https://github.com/bytedeco/javacpp-presets/tree/master/tesseract) to call Tesseract API from Java. This library comes with the needed binaries for the given platform. So, we would just declare the Maven dependency and pretty much done.

## pom.xml

|  |
| --- |
| <dependency>  <groupId>org.bytedeco</groupId>  <artifactId>tesseract-platform</artifactId>  <version>4.1.1-1.5.3</version>  </dependency> |

Tesseract can be run in many modes. We will first see how we can detect lines in a given image.

## Detecting lines in an image

|  |
| --- |
| try (TessBaseAPI api = new TessBaseAPI();) {  int returnCode = api.Init(tessDataDirectory, language);  if (returnCode != 0) {  throw new RuntimeException("could not initialize tesseract, error code: " + returnCode);  }   PIX image = pixRead(imagePath.toFile().getAbsolutePath());   LOGGER.info("The image has a width of {} and height of {}", image.w(), image.h());   api.SetImage(image);   BOXA boxes = api.GetComponentImages(tesseract.RIL\_TEXTLINE, true, (PIXA) null, (IntBuffer) null);   LOGGER.info("Found {} textline image components.", boxes.n());   lines = IntStream.range(0, boxes.n()).mapToObj((int lineSequenceNumber) -> {  BOX box = boxaGetBox(boxes, lineSequenceNumber, L\_CLONE);  api.SetRectangle(box.x(), box.y(), box.w(), box.h());  BytePointer ocrResult = api.GetUTF8Text();  String ocrLineText = ocrResult.getString().trim();  ocrResult.deallocate();  int confidence = api.MeanTextConf();  int x1 = box.x();  int y1 = box.y();  int width = box.w();  int height = box.h();   OcrWord lineTextBox = new OcrWord(x1, y1, x1 + width, y1 + height, confidence, ocrLineText, lineSequenceNumber + 1);  LOGGER.debug("lineTextBox: {}", lineTextBox);  return lineTextBox;  }).collect(Collectors.toList());   api.End();  api.close();  pixDestroy(image);  } |

Next, we would see how to detect individual words with Tesseract.

## Detecting words in an image

|  |
| --- |
| try (TessBaseAPI api = new TessBaseAPI();) {  int returnCode = api.Init(tessDataDirectory, language);  if (returnCode != 0) {  throw new RuntimeException("could not initialize tesseract, error code: " + returnCode);  }   PIX image = pixRead(imagePath.toFile().getAbsolutePath());   api.SetImage(image);  int code = api.Recognize(new ETEXT\_DESC());   if (code != 0) {  throw new IllegalArgumentException("could not recognize text");  }   try (ResultIterator ri = api.GetIterator();) {  int level = tesseract.RIL\_WORD;  int wordSequenceNumber = 1;  Supplier<IntPointer> intPointerSupplier = () -> new IntPointer(new int[1]);  do {  BytePointer ocrResult = ri.GetUTF8Text(level);  String ocrText = ocrResult.getString().trim();   float confidence = ri.Confidence(level);  IntPointer x1 = intPointerSupplier.get();  IntPointer y1 = intPointerSupplier.get();  IntPointer x2 = intPointerSupplier.get();  IntPointer y2 = intPointerSupplier.get();  boolean foundRectangle = ri.BoundingBox(level, x1, y1, x2, y2);   if (!foundRectangle) {  throw new IllegalArgumentException("Could not find any rectangle here");  }   OcrWord wordTextBox = new OcrWord(x1.get(), y1.get(), x2.get(), y2.get(), confidence, ocrText, wordSequenceNumber++);  LOGGER.info("wordTextBox: {}", wordTextBox);   words.add(wordTextBox);   x1.deallocate();  y1.deallocate();  x2.deallocate();  y2.deallocate();  ocrResult.deallocate();   } while (ri.Next(level));   ri.deallocate();  }  api.End();  api.deallocate();  pixDestroy(image);  } |

# Source Code

<https://github.com/paawak/blog/tree/master/code/tesseract-ocr/tesseract-java-demo>

# Reference

<https://tesseract-ocr.github.io/tessdoc/APIExample>

<https://github.com/tesseract-ocr/tesseract>

<https://github.com/tesseract-ocr/tessdata>

<https://github.com/bytedeco/javacpp-presets>

<https://github.com/bytedeco/javacpp-presets/tree/master/tesseract>