CSL305 | Assignment-3 | Due 15/Apr/2022 11:59 PM | 100 points

- Important instructions for code submissions are here: https://goo.gl/IMWvdF
- Grading scheme to be followed is available here: https://goo.gl/52D82g
- Assignment description may be underspecified to allow some room for exploration and creativity.
- Assignment description may be underspecified to allow some room for exploration and creativity.
- Your submission should be packaged as a zip file named **exactly** in this format: CS305-[your entry no.]-[assignment no.].zip.
- Submit the ZIP file on Google Classroom (not on GitHub).

We need to design and develop an application which extracts the metadata from a collection of book cover pages. The cover pages of an book are scanned images (taken as a single image file) of the first few pages of the book. An example is shown at the end of this document. The cover page normally contains the following information (in addition to other stuff):

1. Title of the book
2. Names of the authors
3. Publishers
4. ISBN numbers

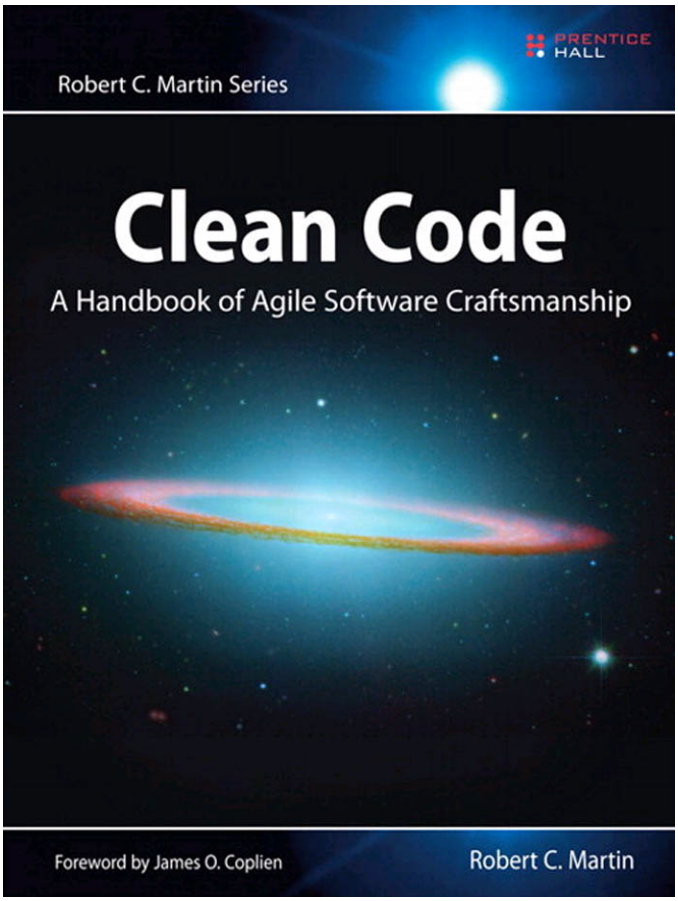
The scanned images will be in the JPEG or PNG format as of now. The application will be a command-line tool which will support the following arguments:

1. A flag that tells whether to process a single input file or a directory containing many files.
2. Path of the input file or the directory.

The output will be a spreadsheet (in .xlsx format) which will have the columns corresponding to the items extracted from the cover page as mentioned above (i.e, title, authors, etc.). The first row of the output .xlsx file will contain the header information.

The crucial thing that will be checked in your solutions is the quality of the software, its adherence to SOLID principles and appropriate use of design techniques. Your tool should be easily extensible to support data extraction from other formats of input files (e.g., PDF, epub, HTML, etc.). You must provide proper unit tests to achieve at least 90% code coverage.

The accuracy of the output data is very important. You can develop the tool in Python 3.x or Java. NOTE: we are not looking for a new ML solution built from ground up for “image recognition”. You can make use of an existing OCR library to do part of the task.


Robert C. Martin Series
PRENTICE HALL
Clean Code
A Handbook of Agile Software Craftsmanship
Foreword by James O. Coplien
Robert C. Martin

Clean Code

A Handbook of Agile Software Craftsmanship

The Object Mentors:

Robert C. Martin

Michael C. Feathers Timothy R. Ottinger
Jeffrey J. Langr Brett L. Schuchert
James W. Grenning Kevin Dean Wampler
Object Mentor Inc.

Writing clean code is what you must do in order to call yourself a professional. There is no reasonable excuse for doing anything less than your best.

Upper Saddle River, NJ • Boston • Indianapolis • San Francisco
New York • Toronto • Montreal • London • Munich • Paris • Madrid
Capetown • Sydney • Tokyo • Singapore • Mexico City