National University of Sciences & Technology (NUST)
School of Electrical Engineering and Computer Science (SEECS)

Assignment 02
Reading Order Determination by Topological Sort

Deadline: April 14, 2023 (23:59)

Q1. Provide a list of five commonly used OCR systems (either commercial or open source).

Q2. Implement a Directed Acyclic Graph for bounding boxes at the line level, using the information/algorithm provided in Section 6 of the research paper given below:
https://www.researchgate.net/publication/2564797_High_Performance_Document_Layout_Analysis
Please note that OCR engines such as Tesseract and EasyOCR can be used to extract bounding boxes from document images.

Q3. Sort the bounding boxes based on their width and then draw them on given image sample: Assignment02-Test-Page.jpg , along with their order numbers.

Q4. Implement Topological sort to sort the DAG created in Section B, and then draw the bounding boxes on given image sample: Assignment02-Test-Page.jpg with their order given by the topological sort.

Q5. Find three sample test pages from the Internet, having varying page layouts, and show your results on these test pages.

Deliverables

1) List of OCR Systems along with their attributes
2) Notebook (.ipynb file) of your code corresponding to Q2 – Q4
3) Result of topological sort on the sample image: Assignment02-Test-Page.jpg (uploaded on LMS)
4) Results of your implementation on three sample test pages of your choice