American University of Armenia

Bachelor of Science in Data Science

Student: Astghik Kostanyan

Capstone Project Title: Success Prediction in IT Studies Based on Handwriting Samples

Supervisors: Aleksandr Hayrapetyan, Suren Khachatryan

Date: 22 of March, 2024

**Literature Review**

In the initial phase of my capstone project, the focus was on image preprocessing. "Digital Image Processing" by Wilhelm Burger and Mark J. Burge provided a rich foundation in understanding various image processing techniques, such as median filtering, denoising, eroding, histogram matching, histogram equalization and skeletonization. I also researched relevant scholarly articles that complemented the principles discussed in the book and showed real word usages of those principles. Furthermore, I explored various Python libraries which offered the implementations of the abovementioned techniques.

**Methodology**

The methodology involves a series of steps: data collection and integration, identifying similar patterns among handwritten characters through personal research, preprocessing JPG images to isolate skeletonized handwritten parts, and researching pattern recognition machine learning models to identify the most effective approach for achieving the goal.

**Dataset Constructing Phase**

During the dataset constructing phase, handwritten exam samples from the Object-Oriented Programming (OOP) course between 2015 and 2018 were collected. To maintain anonymity, student names were removed from the exam papers. Following this, each student's exam was scanned and stored as a separate PDF file. The subsequent step involved extracting individual pages from each exam and saving them as JPG images in a specified data directory, facilitating further analysis.

**Data processing phase**

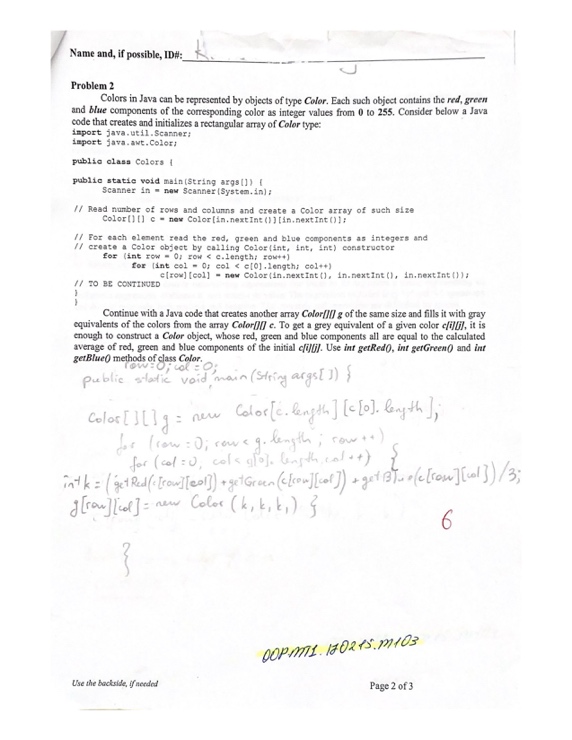
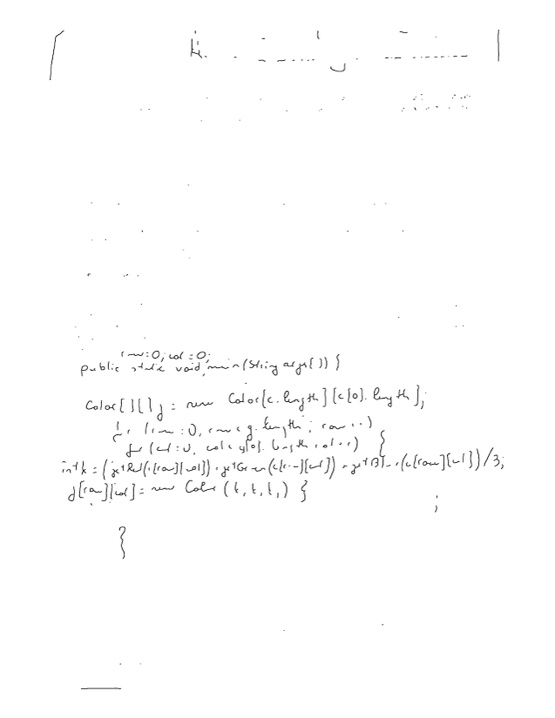
In the stage of data processing, the problem was to separate quite distinctly the handwritings of the students from the printed part of the text in examination papers. To achieve this, the following two hypotheses were developed:

The goal of the first hypothesis is to get the separation margins (height and width) of the handwritten part from the printed part. The hypothesis is to binarize and apply gaussian blur to the grayscale image and as the printed part has darker average color than the handwritten part, the result would be dark blobs representing printed text and white pixels representing handwritten part. Basically, after this step, identifying the printed text margin would be extremely easy. This approach was validated through the ImageJ software.

The second hypothesis was intended for clear skeletonization of handwritten text and disposal of printed parts. The first proposed method was histogram equalization. The original image was split into handwritten and printed parts. Histogram equalization was applied on both images to make the handwritten part darker. In the end, subtracting the result from the original image would output handwritten part only. This hypothesis was tested using ImageJ and Python code, both gave the same unexpected output. Turned out, the subtraction does not result in complete removal of the printed part, but the edges remain. However, this was easy to remove from the final output by applying median filtering technique.

**Initial results**

The first image is the input data point. The second image is obtained after performing all the preprocessing steps.

**OCR Model Constructing Phase**

After doing some research I have decided to use following OCR models on the dataset, compare their performances and choose the one which works the best.

* TrOCR: Transformer-based Optical Character Recognition
* Tesseract
* EasyOCR