MINI PROBLEM STATEMENT FOR NLP OCR

Computer Vision and Intelligence Group, CFI, IIT Madras 30/03/2020

Pre-processing for OCR

In this mini problem statement, we shall look at pre-processing for OCR. A certain level of pre-processing the input image is required before feeding it to the OCR engine. This includes perspective warping, noise removal and binarization. These steps help the OCR engine to better comprehend the text in the image.

Steps for preprocessing

1. Firstly, the piece of document in the image may not look exactly as seen from the top. This is a major issue because, if not seen from above, text size will not be uniform throughout the document. This is where perspective warp comes to the rescue. This requires the orientation of the document in the image. For this, convert the image to grayscale and denoise using a gaussian filter. Use `cv2.Canny()` to extract the edges and proceed with morphological transformations to further denoise the mask.

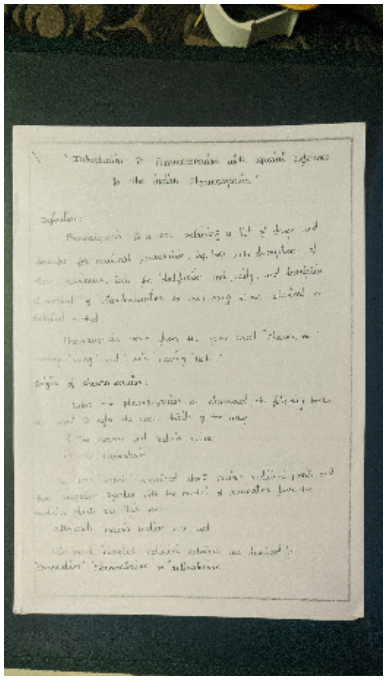
2. The next step is to look for rectangular objects in the image. Use `cv2.findContours()` to find all the contours (closed shapes) in the processed mask. Look for the rectangle with the largest area. This should be the document's boundaries. Look up the internet for how to find rectangular contours.

3. Now, we have the orientation of our document. To flatten the document, we use the methods `cv2.getPerspectiveTransform()` and `cv2.warpPerspective()`. Check out their documentation for parameters and return values. You will end up with a view from the top.

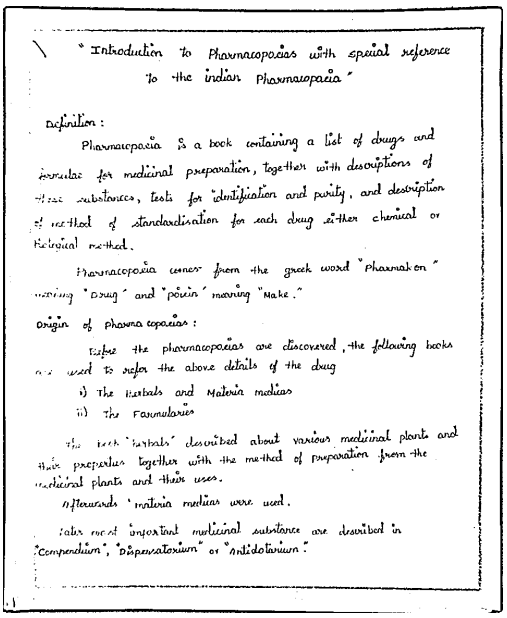
4. Now, it's time to threshold. Binarization means to convert the image to purely black and white. There can be no shades gray. This step immensely helps the ocr engine to process text. Binarization is best achieved by thresholding. Discover how the method `cv2.adaptiveThreshold()` can be applied to ill illuminated images to achieve good results. The warped image, when binarized, can be fed to the ocr engine.

Goal

The results that I got are given below. Try and beat me.



(a) input



"Introduction to Pharmacopoeias with special reference to the Indian Pharmacopoeia"

Definition:

Pharmacopoeia is a book containing a list of drugs and formulae for medicinal preparation, together with descriptions of these substances, tests for identification and purity, and description of method of standardisation for each drug either chemical or Biological method.

Pharmacopoeia comes from the greek word "Pharmakon" meaning "Drug" and "poiein" meaning "Make."

Origin of Pharmacopoeias:

Before the pharmacopoeias are discovered, the following books are used to refer the above details of the drug

i) The Herbals and Materia medicas
ii) The Formularies

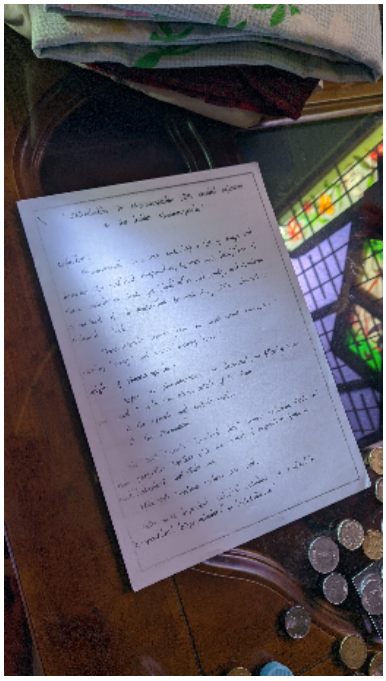
The book 'herbals' described about various medicinal plants and their properties together with the method of preparation from the medicinal plants and their uses.

Afterwards 'materia medicas were used.

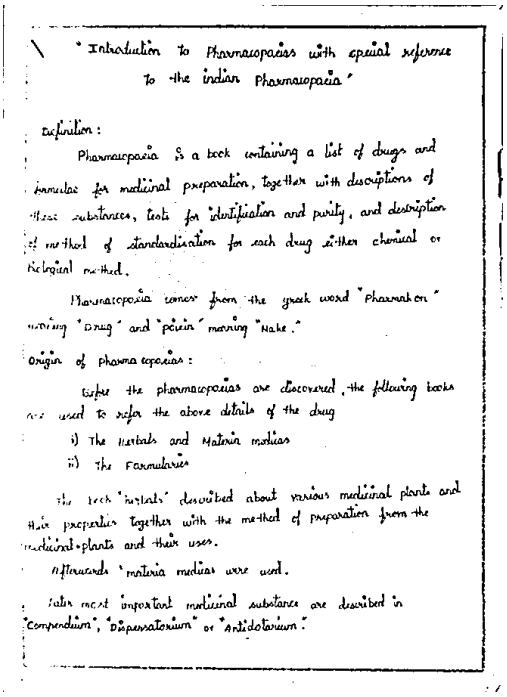
Later most important medicinal substances are described in "Compendium", "Dispensatorium" or "Antidotarium."

(b) result

Well lit text with defined boundaries



(a) input



"Introduction to Pharmacopoeias with special reference to the Indian Pharmacopoeia"

Definition:

Pharmacopoeia is a book containing a list of drugs and formulae for medicinal preparation, together with descriptions of these substances, tests for identification and purity, and description of method of standardisation for each drug either chemical or Biological method.

Pharmacopoeia comes from the greek word "Pharmakon" meaning "Drug" and "poiein" meaning "Make."

Origin of Pharmacopoeias:

Before the pharmacopoeias are discovered, the following books are used to refer the above details of the drug

i) The Herbals and Materia medicas
ii) The Formularies

The book 'herbals' described about various medicinal plants and their properties together with the method of preparation from the medicinal plants and their uses.

Afterwards 'materia medicas were used.

Later most important medicinal substances are described in "Compendium", "Dispensatorium" or "Antidotarium."

(b) result

Ill lit text with noisy boundaries

Remarks

1. I highly encourage you to look up openCV documentation whenever possible. Apart from clearing your doubts, it gives you better intuition for the concept and varied usage of the functionality. Do not hesitate to raise doubts in the group / PM me in this regard.
2. Note that the same code has to be used for both the images given. Parameters cannot be changed between different inputs.
3. Drive link for resource images: `https://drive.google.com/drive/folders/1f67M-6QteEqmN3?usp=sharing`
4. Some useful links are given below:
    - Canny egde detection: `https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_canny/py_canny.html`
    - Morphological transformations: `https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_morphological_ops/py_morphological_ops.html`
    - Contours: `https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_contours/py_contours_begin/py_contours_begin.html`
    - Rectangle detection: `https://theailearner.com/2019/11/22/simple-shape-detectio`
    - Perspective warp: `https://www.pyimagesearch.com/2014/08/25/4-point-opencv-get`
    - Thresholding: `https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_thresholding/py_thresholding.html`

– Good luck! –