**Documentation for the OCR Code**

**NOTE: File should be uploaded in the Google drive and using! Wget it should be recalled and authenticated by a code generated using google account**

**Poppler utils** is the package for pdf rendering in linux environment

**Pdf2image** is like a wrapper for poppler

**Easy ocr** is package used to extract data from pdf

Import package convert from path to convert pdf to image

**PIL** is function used to load the image

**Spacy** is used to perform named entity recognition and to identify key entities

**Easyocr.Reader** to give reader object and select English as language or any language can be selected based on the document language

**!Wget** is used to download the pdf from the web to local or can be directly loaded from local computer using open() function

**Ipython.display** is used to show the downloaded image

Convert the PIL image to array by setting parameters and setting decoder and returning different bounding ranges i.e bounds

Draw bounding boxes to highlight the data that is getting extracted

Iterate through bounds and concatenate in a single text using **for loop**

Pass the text to spacy loading default English language

**Displacy** function is used to display the data output in rich format