Qns2:Why is a high quality conversion from PDF to plain text hard? Your answer does not need to be exhaustive but should outline some of the most important reasons.

A PDF page is drawn based on instructions contained in the content stream. Following are the procedures which are difficult to compute while converting from pdf to plain text:

1. It is also possible to draw text in areas where text extractors do not look: some tools hide text by placing it in a pattern and filling the page area with that pattern; similarly, type 3 fonts are available; each character in a type 3 font has its own content stream; therefore, a tool can draw all the text within a single glyph of a type 3 font and then draw it on the page.
2. The process of reading through PDFs line by line may seem straightforward. There can be differences in the way documents are structured, and data can be arranged ad hoc. Also, documents can be skewed in various ways, making them difficult to read.
3. There are many PDFs that follow different document structures, so the data front, the size, and the angle can differ from one PDF to another. Therefore, it will be more difficult to process data is challenging.

Qns 4: Apply your solution from task4 (i)tothefilesinthefolderscans/whichconsistsofpagesscannedfrom a phone book. Analyse how well your solution performs by giving examples.

This mupypdf method as considered scanning this pdf column-vise, Based on ads they are 4 columns in the original scanned OCR.

Consider an ad as shown below:

A picture containing text, newspaper, receipt

Description automatically generated

After using Mupypdf, below is the text format conversion:

'Alt Peter \n',

'Rechtsanwalt \n',

'Fachanwalt für Arbeitsrecht \n',

'Kirschgartenstr. 19 \n',

'33836 70 \n',

'www.kanzlei-buehler-alt.de \n',

'm) \n',

'Gratis anrufen | mit dastelefonbuch.de \n',

' \n',

'Bau, Becky & Collegen \n',

'Poststr. 2 \n',

'www.bau-coll.de \n',

'ed al Gratis anrufen \n',

'| mit dastelefonbuch.de \n',

Most of the text from the sample has been detected by the method except in this case the phone from Bau,Becky & Collegen has not been detected.

Example2:

Text, letter

Description automatically generated

Here, the telephone symbol has not been detected properly

'Jakob & Kollegen \n',

'Jakob & Kollegen \n',

'Rechtsanwälte \n',

'Bergheimer Straße 49 \n',

'| 69115 Heidelberg \n',

'2 (0 62 21) 14 07 14 \n',

'www.jakob-ra.de \n',

Now considering the single OCR,

A close-up of a newspaper

Description automatically generated with low confidence

The method has run from left to right but not in columns as doubleOCR

'Altlußheim (0.62.05) \n',

'| \n',

'ne. \n',

'www.dastelefonbuch.de \n',

'Gottfried Werner Tulla-20 ........020...0.... 32030 | Heid Dagmar Rosemarie.................. 364 1733 | Hoffstätter Dieter u. Fischer Beate......... 31973 \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

' \n',

'AESE5-APOTHEKE| \n',

'cottmam Eisabeth Hause \n',

'31942 | \n',

'Erich-Kästner-Weg 1A \n',

'Kurpfalz-9 \n',

'Gottwald G. Fliederweg 6 .................... 834139 | - Dagmar Rosemarie.............. 01791173263 | - Elisabeth Kurpfalz-9 ..................... 34747 \n',

'Tel. 062 21-983 13 31 \n',

'Gottwald-Münch Uta Fliederweg 6 ......... 32648 | \n',

'Lußhardt-5 \n',

'| \n',

'Hofmann Walter Rheinhäuser-S4 .......... 32253 \n',

'. \n',

'ei \n',

'Grabowski Wäscherei Schul- 2: aan 34341 | - Kurtund Dagmar... \n',

'33023 | Hohn Elisabeth Haupt-53 .................... 34975 \n',

'www.atos-apotheke.de \n',

'Graf Madlen Haupt-9 ..........u....... 953159 | \n',

'Erich-Kästner-Weg 1a \n',

'— Josef Hölderlin-9......................... 322 60 \n',

'Grafe Rolf Kurpfalz-S1 ......................... 32499 | Heidemann Rainer u. Alexandra............ 32304 | - Robert Hölderlin-9.............. 0178895 72 64 \n',

'Apotheker M. Schuol \n',

Here the spelling from the ad it is AtoS APOTHEKE but it has read as AESE5 APOTHEKE.Its not much readable as the previous double OCR.