**Financial Data Parser – Known Issues**

The following are exceptions of Audits that may not work with the program:

* When the tables use spaces between numbers instead of commas.
  + Ex: 1 288 360 instead of 1,288,360 (See Breen Township’s 2020 Financial Statement)

This messes up the parsing of the row, since the program will not be able to differentiate between spaces that are separating two different numbers and spaces that act as commas.

**Possible fix:** Configure PDFPlumber’s extract\_table() function to detect the different rows/columns in the pages. This eliminates the need for character-by-character comparison and allows the program to read all cell contents as one number.

*Note:* It seems like most PDFs that have this spacing issue were made by the independent auditor CAMPBELL AUDITING CPA, PLC.

* PDF pages are rotated (facing sideways).
  + Ex:

|  |  |  |
| --- | --- | --- |
| Column | 1254125 | 12 |
| Column | 12354 | 5325 |
|  | Row | Row |

PDFPlumber uses PDFMiner that reads text from left to right. The program is unable to detect rotated text.

**Possible Fix:** Find a library or write a function that rotates the page as a whole, then try parsing the text.

* The Font used in the PDF is a CID font.
  + Ex: After extraction from the PDF, the letter A would be read as (CID12) instead of just A. (See City of Hastings 2020 Financial Audit, extracting the text from it will result in CID characters)

This means that the extracted characters will need a map to convert from CID back to normal letters. Since maps can differ between each set of CID characters, there is no way of 100% getting the correct results for all such PDFs. This issue was only detected in one PDF out of 100+ tested PDFs, so it is most likely very rare to happen.