PDF Extractor Plugin

# Cases:

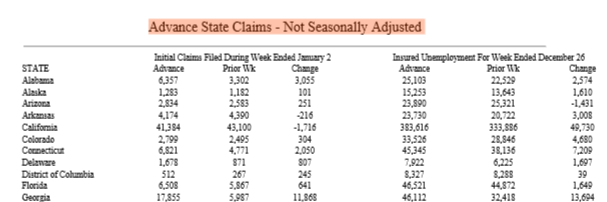
1. PDF Searching for header text is based on the following order till text is not found:
   1. Reading the pdf in the given rotation, removing landscape characters
   2. Reading the pdf in given rotation, removing portrait characters
   3. If text is not found, rotate the pdf 90 degrees, and follow (a.) and (b.)

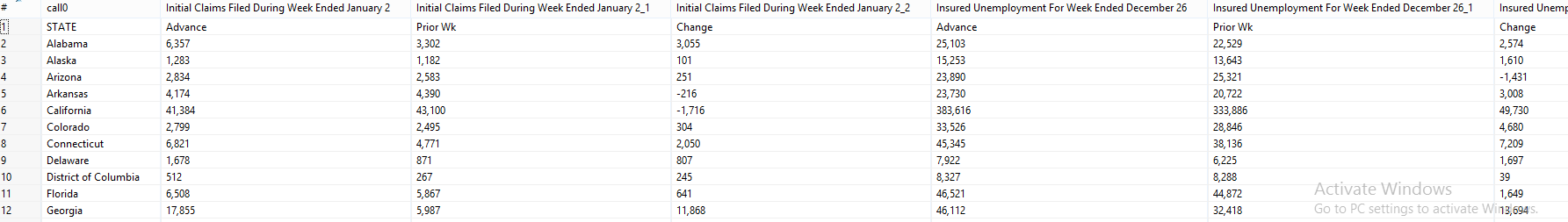
Still if text is not found, no table will be extracted. If text is found, the given rotation and type of reading are saved and footer text is searched on these saved parameters.

1. Header text is mandatory. Table is assumed to start from the next line of given header text. Header text is searched on similar case (uppercase or lowercase characters) basis.
2. Footer text is optional. If given, then the table is assumed to end till the footer. If footer lies on a page different from header text, then the table is extracted from the header text page to the footer text page. (Provided footer text lies after header text).
3. In case there are more than one header text matches in pdf file, table is extracted from first header text only.
4. In case there are more than one footer text matches in pdf file, table is extracted from the header text to the first footer text match found.
5. Since the lines are extracted with coordinates based on portrait mode with origin on left top, so the lines are rotated in case the rotation is 90 and type of reading is while removing portrait characters.
6. Vertical lines are used to extend the cells and make a particular cell lie on more than one column (extension is based on nearest vertical line found on left and right, provided no other cell is before the vertical line found). In case no vertical lines are found, cells are not extended.
7. Horizontal lines are used to find the headers and end of the table.
   1. Header line is assumed to be the one which extends from main row first cell to main row last cell, and is after the line from where the table starts. In case such a line is not found, then table is returned as it is without making any hierarchy in headers.
   2. All the data that is below a horizontal line which extends from last row first cell to last row last cell is assumed to be footer and removed. If in case such a line is not found, then footer lines are assumed to be those which overlap with more than one cell. In this case, lines are removed from the bottom which satisfy the criteria mentioned previously, and stops removing when it finds a row which doesn’t overlap with more than one column.
8. If blank entries are found in header, they are replaced by “call” + column number-1; for example, if first column header is an empty string, “call0” will be made its default header.
9. If text preceding the release date in pdf is mentioned, then we try to extract the date time from the pdf till next two lines from the pre text given. But in case date is not found, then automatically, we save the runtime date time.

# Screenshots of cases

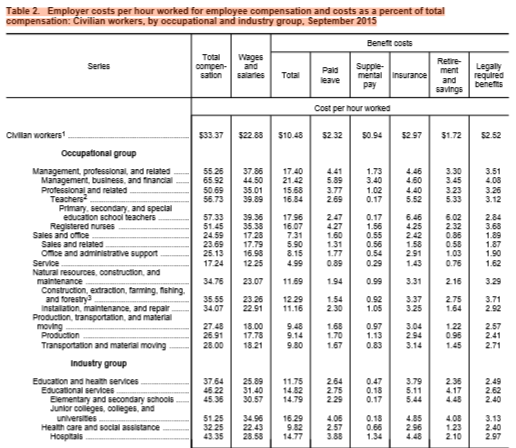
## No horizontal line is found and blank value found in header

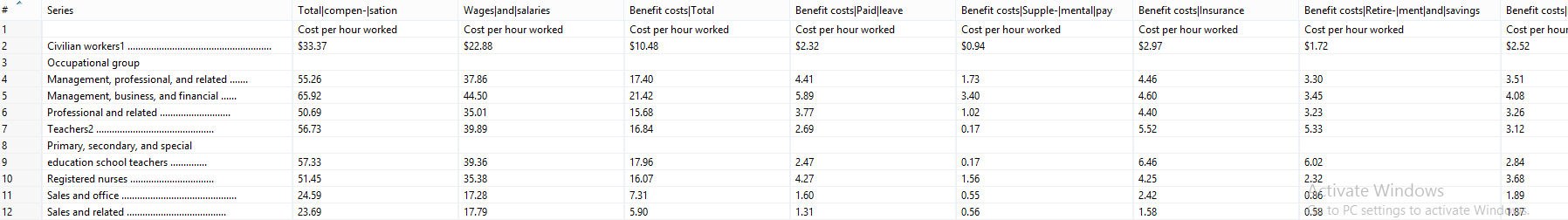
Input:

Output:

## Horizontal and vertical lines found

Input



Output:

Footer text not