Scraping data from .xlsx and .pdf files for the Hawaii-Pacific Weed Risk Assessment

In [32]:

```
from xlrd import open_workbook, XLRDError

# Load the input xls data file
filename   = "Leucaena leucocephala tarramba.xls"
workbook   = open_workbook(filename)
firstsheet = workbook.sheet_by_index(0)
```

In [33]:

```
data_points = []
for row in range(1, 52):
    # Note: array index starts from 0
    data_points.append(firstsheet.cell(row, 3).value)
print(data_points)
```

```
['n', '', '', 2.0, 2.0, 'y', 'y', 'n', 'n', '', '', '', '', '', 'n', 'y', 'n', 'n', 'y', 'y', 'n',
'n', '', 'y', 'n', '', 'n', 'n', 'y', 'n', 'n', 'y', '', 'y', 'n', 'n', '', 'n', 'y', 'n', 'y', 'y
', 'n', 'n', 'y', '', 'y', '', 'y', '', '']
```

In [34]:

```
import xlwt
wb = xlwt.Workbook(encoding='utf-8')
ws = wb.add_sheet('Data Sheet')
```

In [35]:

```
header =
["1.01","1.02","1.03","2.01","2.02","2.03","2.04","2.05","3.01","3.02","3.03","3.04","3.05"," ", "4
.01","4.02","4.03","4.04","4.05","4.06","4.07","4.08","4.09","4.1","4.11","4.12","5.01","5.02","5.0
3","5.04","6.01","6.02","6.03","6.04","6.05","6.06","6.07","7.01","7.02","7.03","7.04","7.05","7.06
","7.07","7.08","8.01","8.02","8.03","8.04","8.05", "total score"]

# write data header on row 0
for index, colname in enumerate(header):
    ws.write(0, index, colname)
```

In [36]:

```
for index, data_item in enumerate(data_points):
    ws.write(1, index, data_item)

wb.save('C:/Users/Kelsey/ParsedExcel.xls')
print ("Done exporting the xls file !!")
```

```
Done exporting the xls file !!
```

In [45]:

```
import tabula

pdf_filename = "Sauvagesia erecta.pdf"
pages_scrapped = tabula.read_pdf(pdf_filename, output_format="json", pages=[1,2])
```

In [46]:

```
for page in pages_scrapped:
    for row in page['data']:
        for column in row:
            print (column['text'][:30], end="\t")
            print (column['text'][:30], end='\t')
        print()
```

```
TAXON: Sauvagesia erecta L.SCO
Fa
Assessor: Chuck ChimeraStatus:
High Risk
Assessor: Chuck ChimeraStatus:
High Risk
Keywords: Naturalized, Pantrop

Qsn # Question Answer Option Answer
101 Is the species highly domestic y=-3, n=0 n
102 Has the species become natural
103 Does the species have weedy ra
201 Species suited to tropical or  (0-low; 1-intermediate; 2-high High
202 Quality of climate match data (0-low; 1-intermediate; 2-high High
203 Broad climate suitability (env y=1, n=0 y
204 Native or naturalized in regio y=1, n=0 y
205 Does the species have a histor y=-2, ?=-1, n=0 y
301 Naturalized beyond native rang y = 1*multiplier (see Appendix y
302 Garden/amenity/disturbance wee n=0, y = 1*multiplier (see App y
303 Agricultural/forestry/horticul
304 Environmental weed n=0, y = 2*multiplier (see App n
305 Congeneric weed
401 Produces spines, thorns or bur y=1, n=0 n
402 Allelopathic
403 Parasitic y=1, n=0 n
404 Unpalatable to grazing animals
405 Toxic to animals y=1, n=0 n
406 Host for recognized pests and
407 Causes allergies or is otherwi y=1, n=0 n
408 Creates a fire hazard in natur y=1, n=0 n
409 Is a shade tolerant plant at s
Creation Date: 10 Dec 2018(Sau
TAXON: Sauvagesia erecta L.SCO

Qsn # Question Answer Option Answer
410 Tolerates a wide range of soil
411 Climbing or smothering growth  y=1, n=0 n
412 Forms dense thickets
501 Aquatic y=5, n=0 n
502 Grass y=1, n=0 n
503 Nitrogen fixing woody plant y=1, n=0 n
504 Geophyte (herbaceous with unde y=1, n=0 n
601 Evidence of substantial reprod y=1, n=0 n
602 Produces viable seed y=1, n=-1 y
603 Hybridizes naturally
604 Self-compatible or apomictic y=1, n=-1 y
605 Requires specialist pollinator y=-1, n=0 n
606 Reproduction by vegetative fra y=1, n=-1 n
607 Minimum generative time (years 1 year = 1, 2 or 3 years = 0,  1
701 Propagules likely to be disper y=1, n=-1 y
702 Propagules dispersed intention
703 Propagules likely to disperse
704 Propagules adapted to wind dis y=1, n=-1 n
705 Propagules water dispersed y=1, n=-1 y
706 Propagules bird dispersed y=1, n=-1 n
707 Propagules dispersed by other
708 Propagules survive passage thr y=1, n=-1 n
801 Prolific seed production (>100
802 Evidence that a persistent pro
803 Well controlled by herbicides
804 Tolerates, or benefits from, m
805 Effective natural enemies pres
Creation Date: 10 Dec 2018(Sau
```

In [47]:

```
left = 20
top = 200
width = 560
height = 520

page1_initial_coords = [(top, left, top + height, top + width)]
scrapped_page1 = tabula.read_pdf(pdf_filename, output_format="json", pages=[1], area=page1_initial
```

```python
coords)

left = 20.07
top = 48.09
width = 565.37
height = 674.04

page2_initial_coords = [(top, left, top + height, top + width)]
scrapped_page2 = tabula.read_pdf(pdf_filename, output_format="json", pages=[2], area=page2_initial_
coords)

pages_json = [scrapped_page1[0], scrapped_page2[0]]

for page in pages_json:
    for row in page['data']:
        for column in row:
            print (column['text'][:30], end="\t")
        print()
```

```
Keywords: Naturalized, Pantrop


Qsn # Question Answer Option Answer
101 Is the species highly domestic y=-3, n=0 n
102 Has the species become natural
103 Does the species have weedy ra
201 Species suited to tropical or  (0-low; 1-intermediate; 2-high High
202 Quality of climate match data (0-low; 1-intermediate; 2-high High
203 Broad climate suitability (env y=1, n=0 y
204 Native or naturalized in regio y=1, n=0 y
205 Does the species have a histor y=-2, ?=-1, n=0 y
301 Naturalized beyond native rang y = 1*multiplier (see Appendix y
302 Garden/amenity/disturbance wee n=0, y = 1*multiplier (see App y
303 Agricultural/forestry/horticul
304 Environmental weed n=0, y = 2*multiplier (see App n
305 Congeneric weed
401 Produces spines, thorns or bur y=1, n=0 n
402 Allelopathic
403 Parasitic y=1, n=0 n
404 Unpalatable to grazing animals
405 Toxic to animals y=1, n=0 n
406 Host for recognized pests and
407 Causes allergies or is otherwi y=1, n=0 n
Creates a fire hazard in natur y=1, n=0 n

Qsn # Question Answer Option Answer
410 Tolerates a wide range of soil
411 Climbing or smothering growth  y=1, n=0 n
412 Forms dense thickets
501 Aquatic y=5, n=0 n
502 Grass y=1, n=0 n
503 Nitrogen fixing woody plant y=1, n=0 n
504 Geophyte (herbaceous with unde y=1, n=0 n
601 Evidence of substantial reprod y=1, n=0 n
602 Produces viable seed y=1, n=-1 y
603 Hybridizes naturally
604 Self-compatible or apomictic y=1, n=-1 y
605 Requires specialist pollinator y=-1, n=0 n
606 Reproduction by vegetative fra y=1, n=-1 n
607 Minimum generative time (years 1 year = 1, 2 or 3 years = 0,  1
701 Propagules likely to be disper y=1, n=-1 y
702 Propagules dispersed intention
703 Propagules likely to disperse
704 Propagules adapted to wind dis y=1, n=-1 n
705 Propagules water dispersed y=1, n=-1 y
706 Propagules bird dispersed y=1, n=-1 n
707 Propagules dispersed by other
708 Propagules survive passage thr y=1, n=-1 n
801 Prolific seed production (>100
802 Evidence that a persistent pro
803 Well controlled by herbicides
804 Tolerates, or benefits from, m
805 Effective natural enemies pres
```

In [48]:

```
type(pages_json)
```

Out[48]:

```
list
```

In [49]:

```
import xlwt
wb_pdf = xlwt.Workbook(encoding='utf-8')
ws_pdf = wb_pdf.add_sheet('Data Sheet')

# write data header on row 0
header =
["1.01","1.02","1.03","2.01","2.02","2.03","2.04","2.05","3.01","3.02","3.03","3.04","3.05", "4.01"
,"4.02","4.03","4.04","4.05","4.06","4.07","4.08","4.09","4.1","4.11","4.12","5.01","5.02","5.03","
5.04","6.01","6.02","6.03","6.04","6.05","6.06","6.07","7.01","7.02","7.03","7.04","7.05","7.06","7
.07","7.08","8.01","8.02","8.03","8.04","8.05", "total score"]
for index, colname in enumerate(header):
    ws_pdf.write(0, index, colname)

# write data on row 1
column_index = 0
for page in pages_json:
    for row in page['data']:

        if (row[0]['text'] == ''):
            # discard the empty row
            continue
        ws_pdf.write(1, column_index, row[3]['text'])
        column_index += 1

ws_pdf.write(1, column_index, row[2]['text'])

wb_pdf.save('ParsedPdfFile.xls')
print ('Done exporting the xls file !!')
```

```
Done exporting the xls file !!
```

In [42]:

```
type(ws_pdf)
```

Out[42]:

```
xlwt.Worksheet.Worksheet
```