PDF Data Extraction

Techniques for Reading and Analyzing PDF files

Andreas H. Pfeiffer


~ `
! 1
@ 2
# 3
$ 4
% 5
^ 6
& 7
* 8
(9
) 0
_ -
+ =
backspace
tab
q
w
e
r
t
y
u
i
o
p
{ [
}]
| \
caps lock
a
s
d
f
g
h
j
k
l
: ;
" '
Analyze PDF
shift
z
x
c
v
b
n
m
< ,
> .
? /
ctrl
win
alt
alt
win
del
ctrl

PDF files

.. contain text information

There is some JavaScript and an Omnis library that you can use to analyse the content of any PDF file.

https://omnis-software.de/conf24/pdfcontentreader.zip

PDF2Json

.. how to

- Download and unpack the zip file
- Copy the folder "pdf2json" into the "jsworker" folder in your Omnis installation inside Application Support or AppData.
- Open a terminal window or command prompt, navigate to this folder and run: npm i
- This will install additional Node modules.
- Now you can run the library "pdf_content_reader.lbs".

PDF content reader
.. some remarks

- Only instantiate the JS worker once by making a getter method to retrieve the reference to the worker. (See Startup_Task)
- We will use the JavaScript worker object. For more information please see https://www.omnis.net/developers/resources/onlinedocs/ExtendingOmnis/07webcomms.html#javascript-worker-object
- oPDFDetails is subclassed from the JavaScript worker.
- The return method of oPDFDetails will return a list of text objects from the PDF sorted by Y and X coordinates.

Example