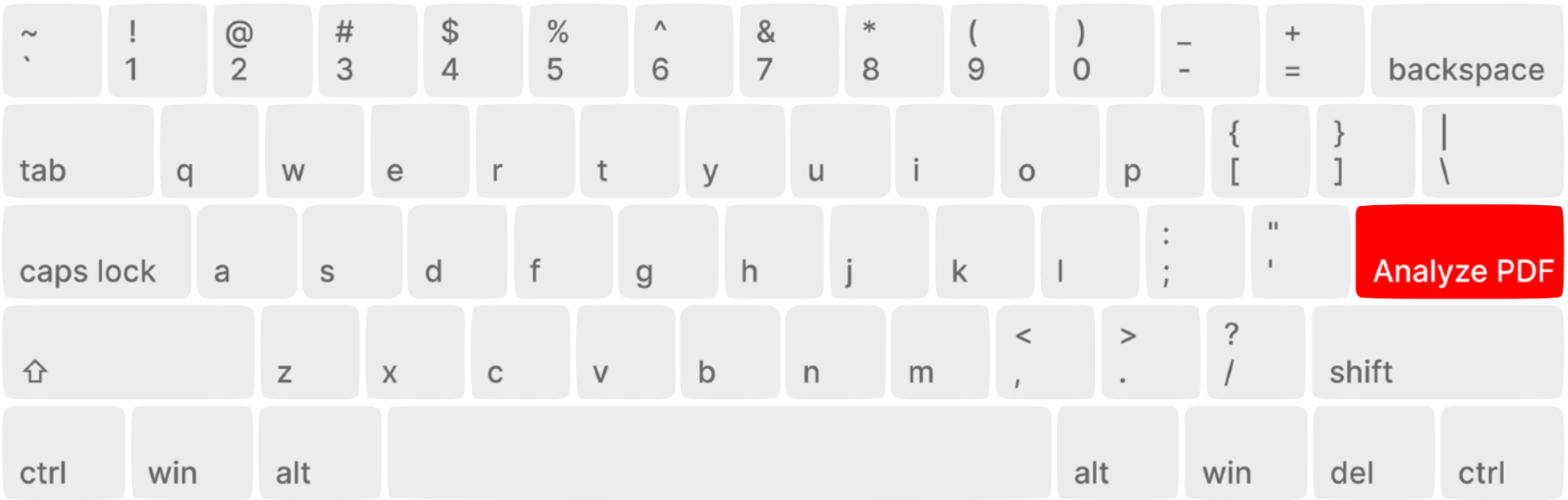
PDF Data Extraction

Techniken zum Lesen und Analysieren von PDF-Dateien

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 Dateien

.. enthalten Textinformationen

Es gibt eine JavaScript Resource und eine Omnis Library, mit denen Sie den Inhalt einer beliebigen PDF-Datei analysieren können.

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

PDF2Json

.. so geht's

- Downloaden und entpacken Sie die Zip-Datei
- Kopieren Sie den Ordner "pdf2json" in den Ordner "jsworker" in Ihrer Omnis-Installation innerhalb von Application Support oder AppData.
- Öffnen Sie ein Terminalfenster oder eine DOS Eingabeaufforderung, navigieren Sie zu diesem Ordner und geben Sie folgendes ein: npm i
- Dadurch werden zusätzliche Node-Module installiert.
- Nun können Sie die Bibliothek "pdf_content_reader.lbs" ausführen

PDF content reader

.. einige Hinweise

- Der JS-Worker wird nur einmal instanziiert, indem eine Getter-Methode zum Abrufen der Referenz auf den Worker erstellt wird. (Siehe Startup_Task)
- Wir werden das JavaScript-Worker-Objekt verwenden. Für weitere Informationen siehe: https://www.omnis.net/developers/resources/onlinedocs/ExtendingOmnis/07webcomms.html#javascript-worker-object
- oPDFDetails ist eine Unterklasse des JavaScript Workers.
- Die Rückgabemethode von oPDFDetails gibt eine nach Y- und X-Koordinaten sortierte Liste von Textobjekten aus der PDF-Datei zurück.

Beispiel