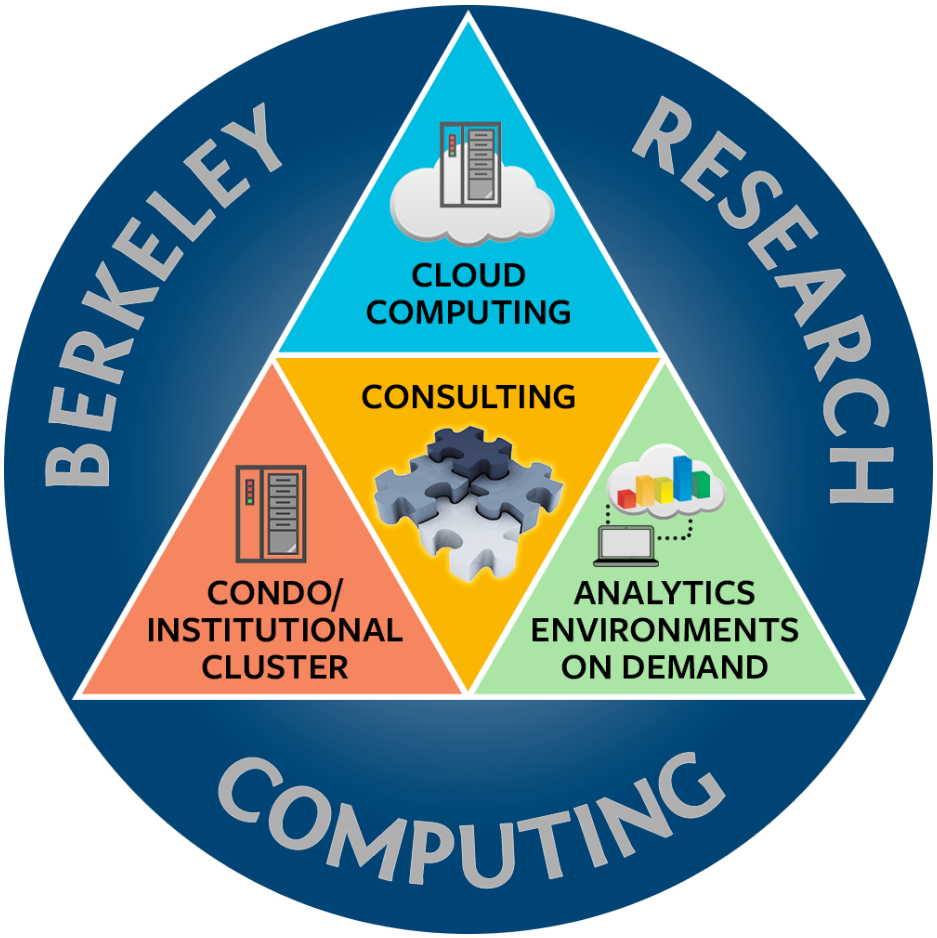
Introduction to Optical Character Recognition (OCR) Software

D-Lab workshop, February 28, 2017

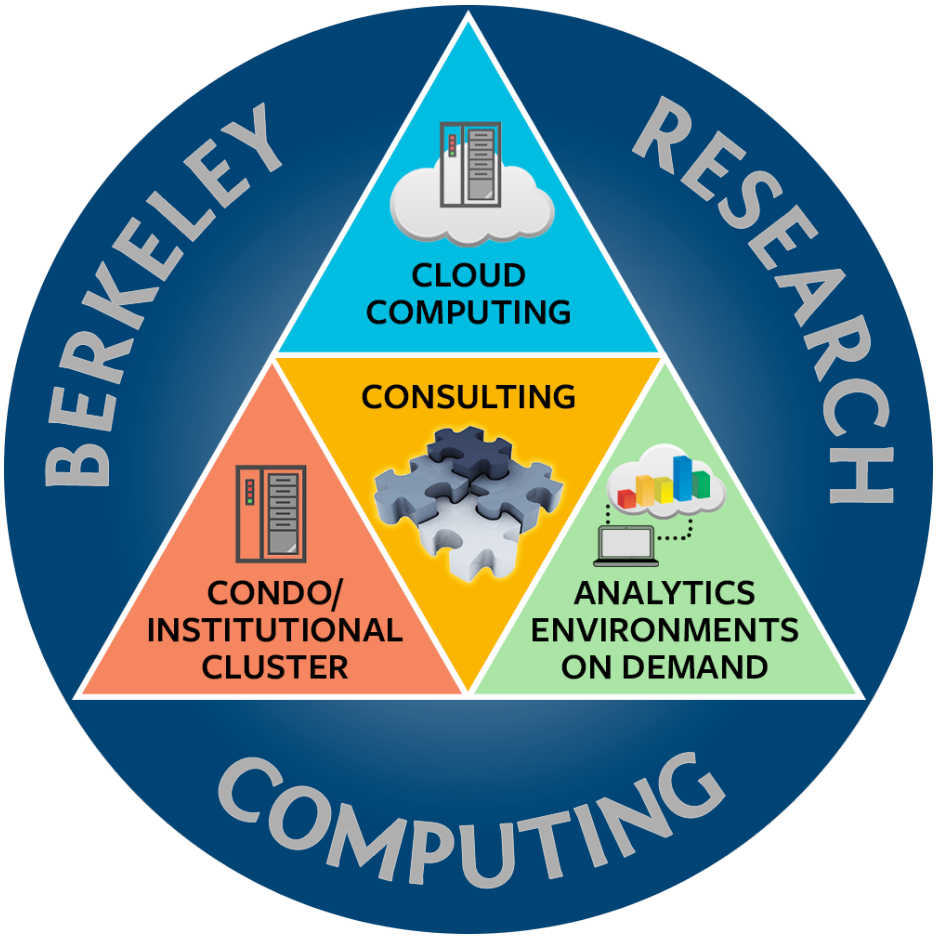
Quinn Dombrowski, DH Coordinator, Research IT, quinnd@berkeley.edu

Research IT
Advancing Research@Berkeley

Berkeley Research Computing

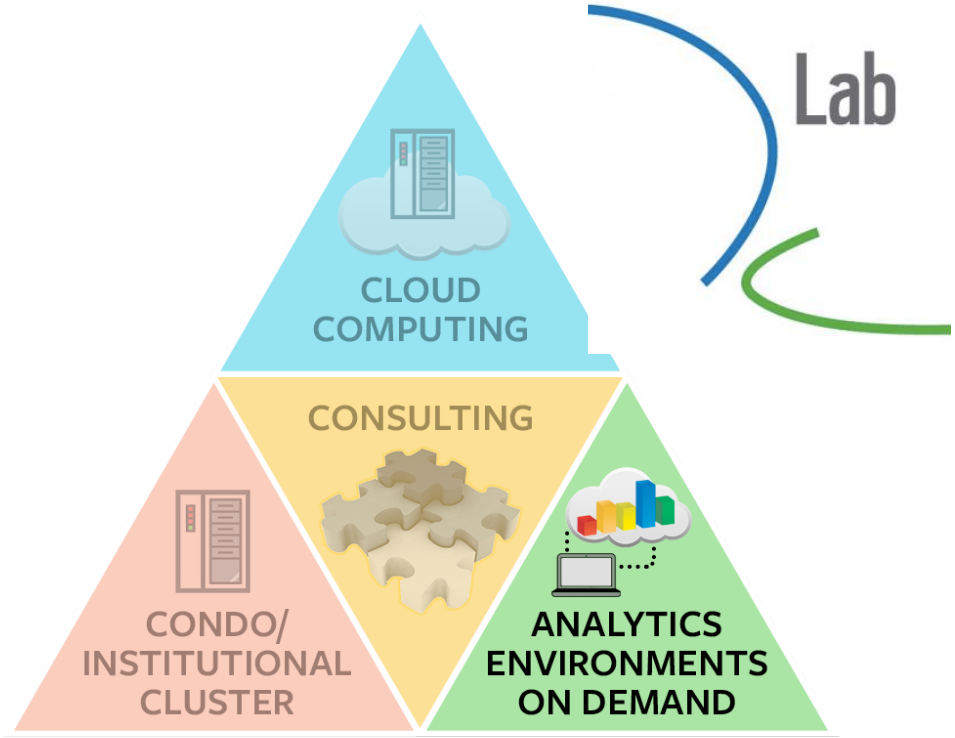


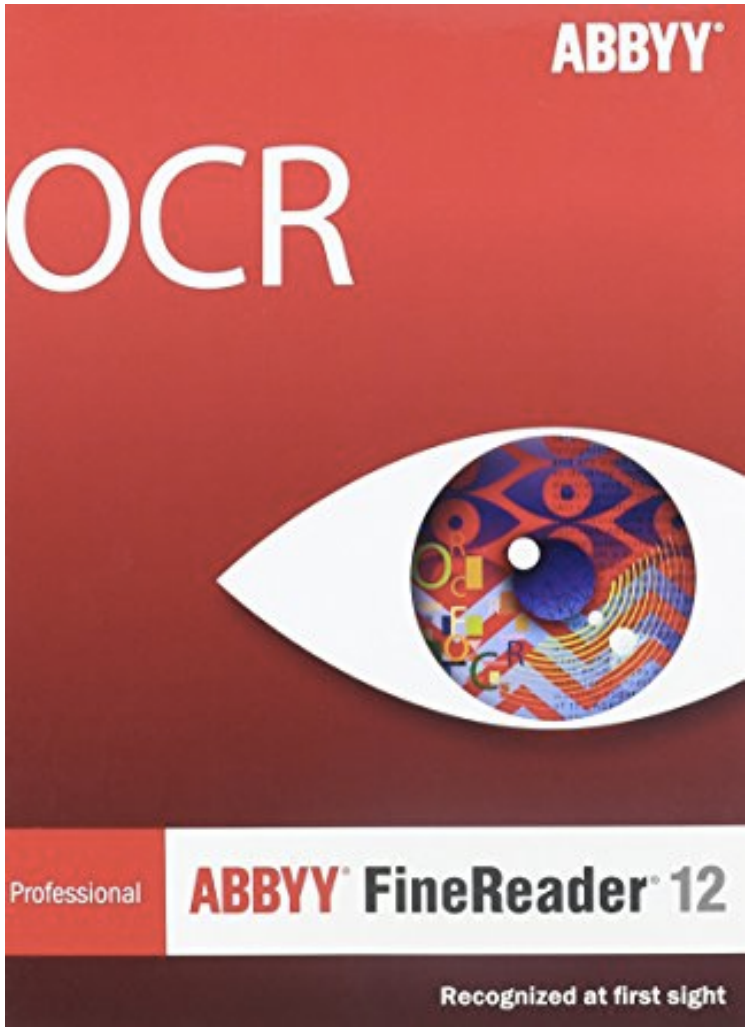
Berkeley Research Computing + DH



DIGITAL HUMANITIES AT BERKELEY

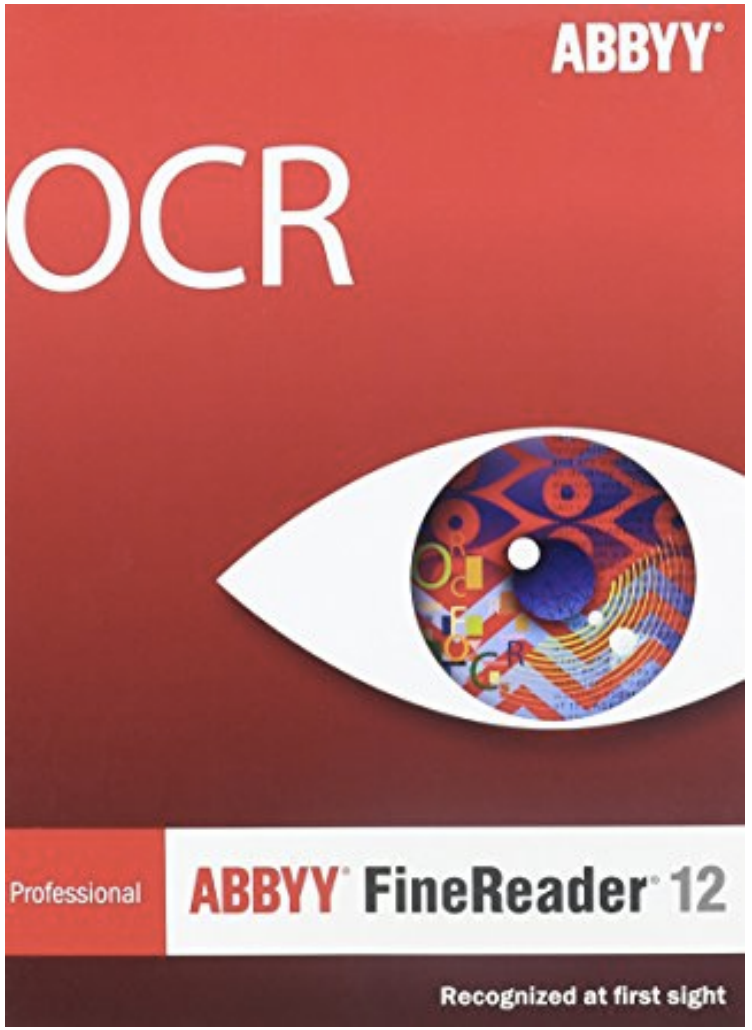
Optical Character Recognition (OCR)



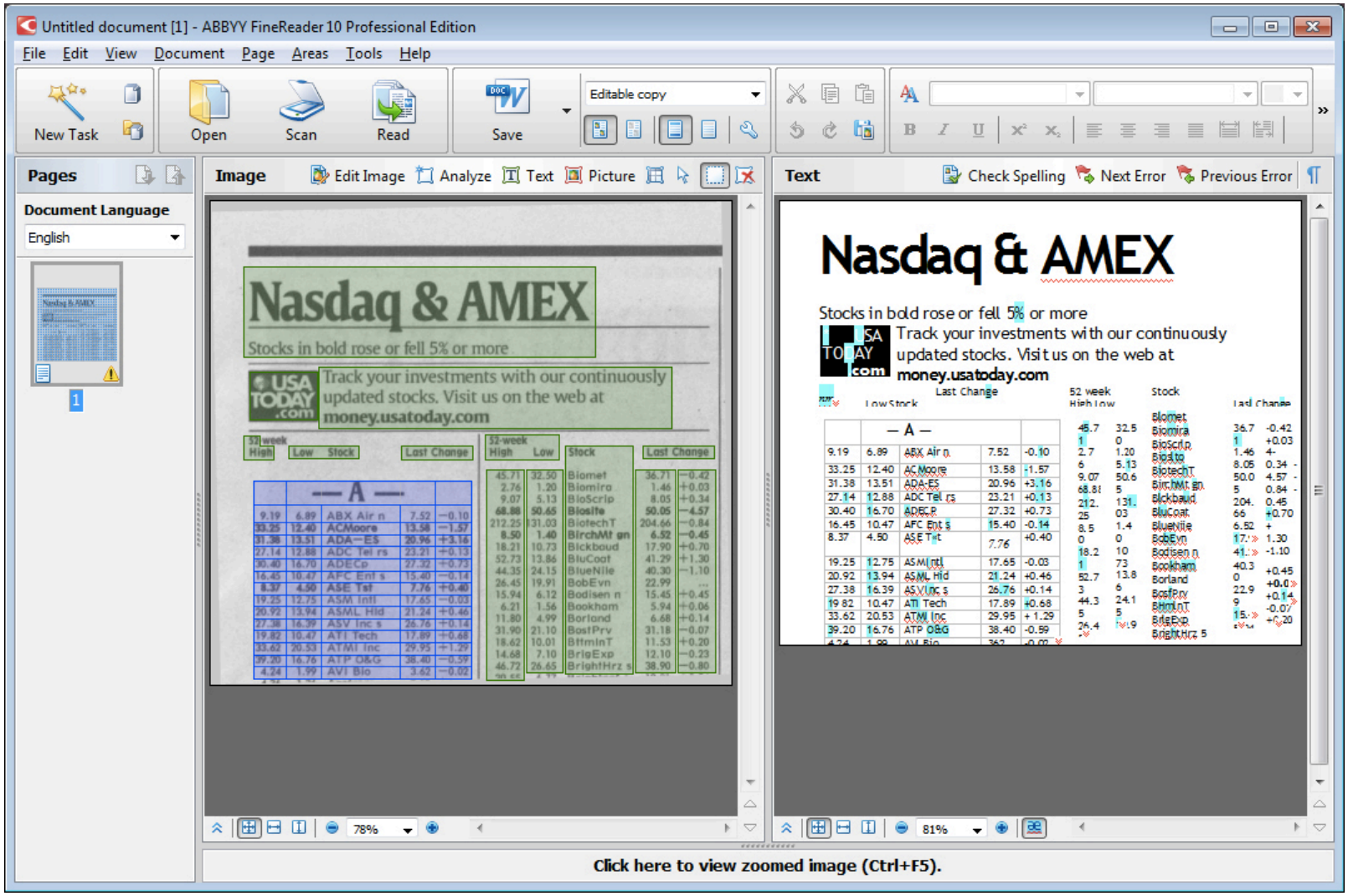


Optical Character Recognition (OCR)

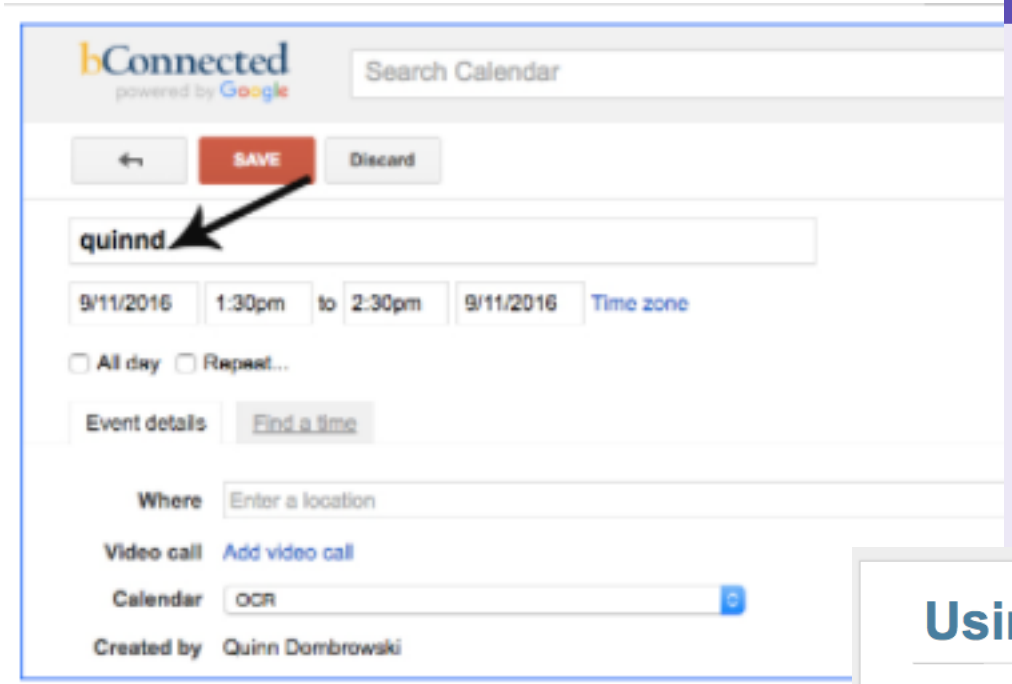
- Complex layout
- Non-English (includes Chinese, Japanese, Korean, Cyrillic, Greek, Arabic, Hebrew)
- Multilingual
- **Precision**

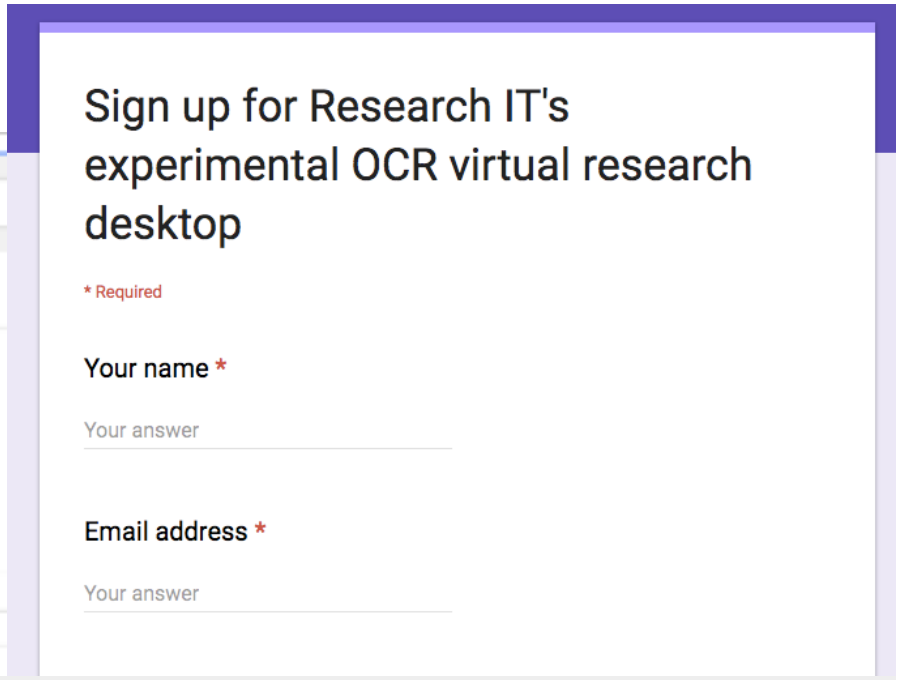


When precision matters



Google Calendar + Docs + Forms





Using FineReader via Citrix

Double-click on **OCR Desktop** to connect to a remote Microsoft Windows computer that has **ABBYY FineReader** installed.

(You will probably only have access to one icon).

Research IT
Advancing Research@Berkeley

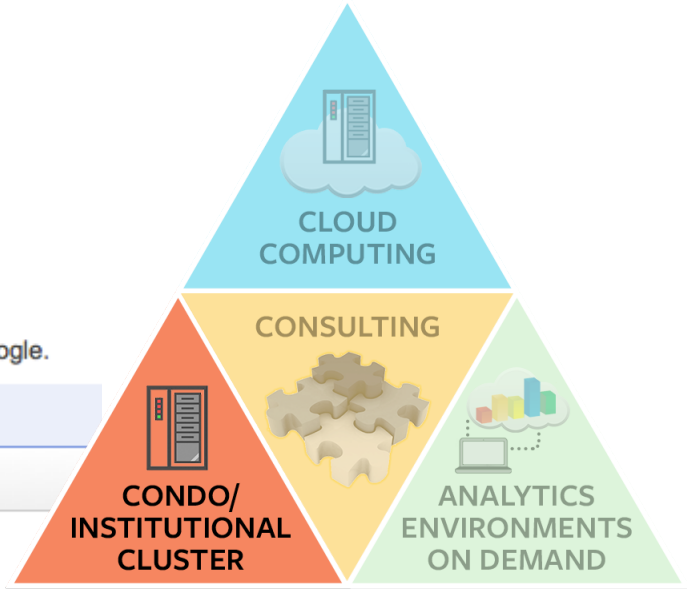
OCR at scale

Postdoc Adam Anderson

+

OCR at scale


tesseract-ocr
An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google.
Project Home
Downloads
Wiki
Issues
Source
Summary
Updates
People
Project Information
Starred by 1864 users
Activity
High
Project feeds
Code license
Apache License 2.0
Labels
OCR, Utility, CPlusPlus, Google
Background
The Tesseract OCR engine was one of the top 3 engines in the 19
but it is probably one of the most accurate open source OCR engi
output text. Image input is managed by the Leptonica Image Proc
Important Download Information:
The language data files are separate from th


CLOUD
COMPUTING
CONSULTING
CONDO/
INSTITUTIONAL
CLUSTER
ANALYTICS
ENVIRONMENTS
ON DEMAND

OCR at scale

Convert all pdf files in working directory to images

In [21]:

```
def runGhostscript(pdfFile):
    print("filename: ", pdfFile)
    if filename.endswith(".pdf") :
        name, extension = os.path.splitext(pdfFile)

        # Ghostscript executable is inside the container.
        # TEMPLATE: gs -dBATCH -dNOPAUSE -dQUIET -sDEVICE=p
        GHOSTSCRIPTCMD = 'gs -dBATCH -dNOPAUSE -dQUIET -sDE
        gcmd = GHOSTSCRIPTCMD.format(tesseractScratchDataDi

        #
        # convert pdf to png
        #
        print("singularity cmd: ", SINGULARITYCMD)
        print("gs cmd: ", gcmd)
        #result = subprocess.call(GHOSTSCRIPTCMD)
        result = !$SINGULARITYCMD $gcmd
        print("gs result: ", result)
```

In [22]:

```
from multiprocessing import Pool

pdffileList = []
for filename in os.listdir(scratchDataDirectory):
    print("filename: ", filename)
    if filename.endswith(".pdf") :
        pdffileList.append(filename)

        name, extension = os.path.splitext(filename)
        pdfnamelist.append(name)

print("pdffileList: ", pdffileList)
print("filenameList: ", pdfnamelist )


#
# multiprocess the pdf to png work work
#
pool0 = Pool(20)
pool0.map(runGhostscript, pdffileList)
pool0.close()
pool0.join()
```

```
filename:  OAAS_1_-_OA_Bib__2003.pdf
filename:  OACC__1976.pdf
pdffileList:  ['OAAS_1_-_OA_Bib__2003.pdf', 'OACC__1976.pdf']
```

Run tesseract on all image files in the working directory

```
def runTesseract(imagefile):
    print("imagefile : ", imagefile)

    # template: tesseract --tessdata-dir /opt/tessdata /scratch/germanocr_Page_01.png  germanc
    TCMD = 'tesseract --tessdata-dir /opt/tessdata {}{} {}{}  -l eng'
    #
    # ocr the png
    #
    basename, ext = os.path.splitext(imagefile)
    tcmd = TCMD.format(tesseractScratchDataDirectory, imagefile, tesseractScratchDataDirectory
    print("tesseract cmd: ", tcmd)
    #print("singularity cmd: ", SINGULARITYCMD)

    result = !$SINGULARITYCMD $tcmd
    print("tesseract result: ", result)
    TCMD = 'tesseract --tessdata-dir /opt/tessdata {}{} {}{}  -l eng'
```

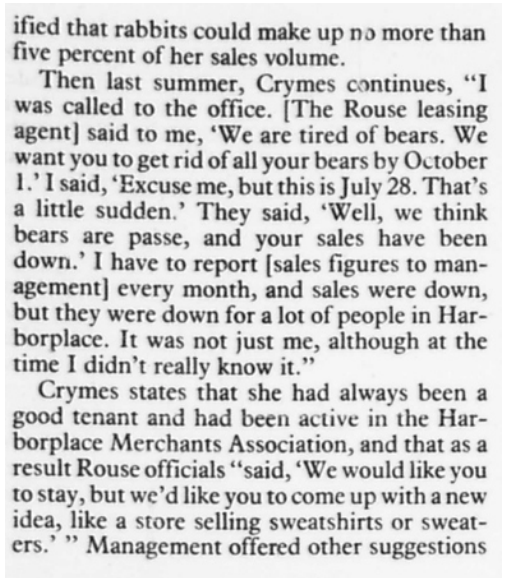
```
from multiprocessing import Pool

imageList = []
for imagename in os.listdir(scratchDataDirectory):
    if imagename.endswith(".png"):
        imageList.append(imagename)

#
# multiprocess the ocr work
#
pool = Pool()
pool.map(runTesseract, imageList)
pool.close()
pool.join()
```

```
imagefile :  OACC__1976-175.png
imagefile :  OACC__1976-112.png
imagefile :  OAAS_1_-_OA_Bib__2003-94.png
imagefile :  OAAS_1_-_OA_Bib__2003-78.png
imagefile :  OACC__1976-68.png
imagefile :  OAAS_1_-_OA_Bib__2003-67.png
imagefile :  OAAS_1_-_OA_Bib__2003-85.png
imagefile :  OACC__1976-4.png
imagefile :  OACC__1976-129.png
imagefile :  OACC__1976-35.png
tesseract cmd:  tesseract --tessdata-dir /opt/tessdata /scratch/OACC__1976-175.png /scratc
h/OACC__1976-175  -l eng
imagefile :  OAAS_1_-_OA_Bib__2003-9.png
```

Tesseract with complex layouts



ified that rabbits could make up no more than five percent of her sales volume.

Then last summer, Crymes continues, "I was called to the office. [The Rouse leasing agent] said to me, 'We are tired of bears. We want you to get rid of all your bears by October 1.' I said, 'Excuse me, but this is July 28. That's a little sudden.' They said, 'Well, we think bears are passe, and your sales have been down.' I have to report [sales figures to management] every month, and sales were down, but they were down for a lot of people in Harborplace. It was not just me, although at the time I didn't really know it."

Crymes states that she had always been a good tenant and had been active in the Harborplace Merchants Association, and that as a result Rouse officials "said, 'We would like you to stay, but we'd like you to come up with a new idea, like a store selling sweatshirts or sweaters.' " Management offered other suggestions

ified that rabbits could make up no more than five percent of her sales volume.

Then last summer, Crymes continues, "I was called to the office. [The Rouse leasing agent] said to me, 'We are tired of bears. We want you to get rid of all your bears by October l.' Isaid, 'Excuse me, but this isJulv 28. That's a little sudden.' They said, We," we think bears are passe, and your sales have been down.' I have to report [sales figures to management] every month, and sales were down, but they were down for a lot of people tn Harborplace. It was not just me, although at the time I didn't really know it. "

Tut-Tut, It's Putt-Putt

THE BROKERAGE BUILDS A MINI-GOLF COURSE

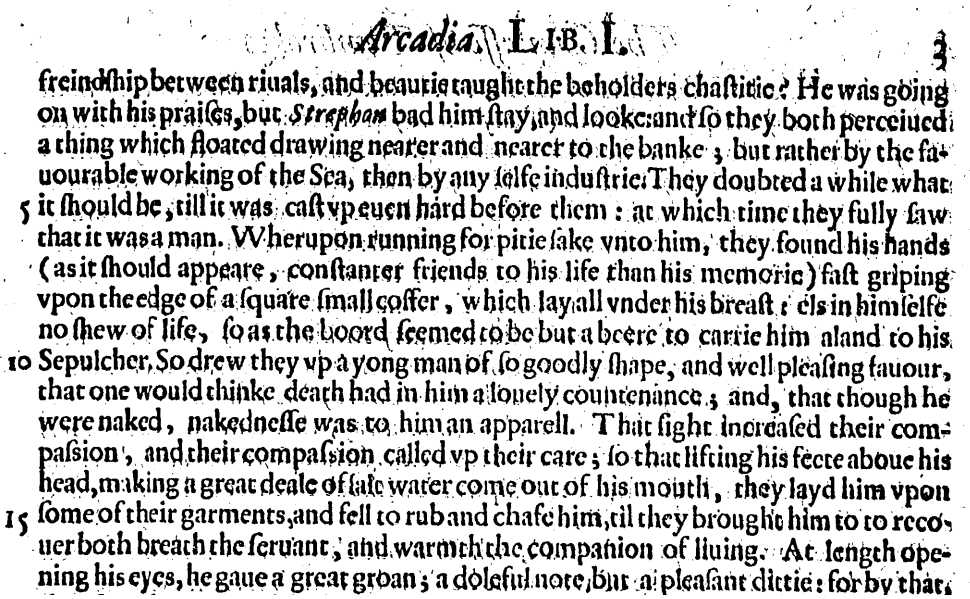
IT WAS INEVITABLE THAT THE ALmost sublime campiness of miniature golf—the flamingos, the windmills, the lighthouses—would one day metamorphose into high art. Yesterday's tacky is today's tacky in a socioeconomic context—a real or hyper-real reflection of the sensibility of the time. Ergo art. Right? Not necessarily bearing this in mind, the people at the Brokerage nonetheless decided to leap on the miniature golf bandwagon (who can blame them?) and open Baltimore's first indoor, year-round miniature golf course. To stir some excitement, they dreamt up a design-and-construct-your-own-golf-hole competition.

a; mi
"' U;
,_ wt «'0'
Ia'mfamwfi-ww"
I.
4!;
i.
'1
W1», 1,
3
i
.3.
'

Early modern OCR

EMOP
EARLY MODERN OCR PROJECT

http://emop.tamu.edu/



Arcadia. LIB. I. 3

freindſhip between riuals, and beautie taught the beholders chaſtitie? He was going on with his praiſes, but *Strephon* bad him ſtay, and looke: and ſo they both perceiued a thing which floated drawing nearer and nearer to the banke; but rather by the fauourable working of the Sea, then by any ſelfe induſtrie. They doubted a while what it ſhould be, till it was caſt vp euen hard before them: at which time they fully ſaw that it was a man. Wherupon running for pitie ſake vnto him, they found his hands (as it ſhould appeare, conſtanter friends to his life than his memorie) faſt griping vpon the edge of a ſquare ſmall coffer, which lay all vnder his breaſt: els in himſelfe no ſhew of life, ſo as the boord ſeemed to be but a beere to carrie him aland to his Sepulcher. So drew they vp a yong man of ſo goodly ſhape, and well pleaſing fauour, that one would thinke death had in him a louely countenance; and, that though he were naked, nakednesſe was to him an apparell. That ſight inordaſed their compaſion, and their compaſion called vp their care; ſo that lifting his feete aboue his head, making a great deale of fale water come out of his mouth, they layd him vpon ſome of their garments, and fell to rub and chafe him, til they brought him to to recouer both breath the ſeruant, and warmth the companion of liuing. At length opening his eyes, he gaue a great groan; a doleful note, but a pleaſant dittie: for by that,

ïll) h mgm ne m
adii:iïot bigg in meGXWa-'-
owne a monſter' kind more '' z
ti be aeth n
hischrefl':ſ ety ;ſh lm lil

twal m abi? id an
' which if muchm dw- ldo not deceiue meas -
i -'Worth)f to bc a ſan[W]?1ngl:'?a gallater offendpr T his ay
p fgx:!?tjhëcauſe!tt

may hc enartſo;ontoſaysh?mn becaulëitw- tibet-uerſo.-
Readëz:tthell atjymmdlahm?sesstrlheffllllgës your good
- melaminst l lind-hamamde mt,but ap e at.Apd-
lea solon saphetterrſtuli,thanlasln a, a e
r ſtmp gia a'm fsathars'eymu willen-onti- ue tolouethe
V and moſtë


Thanks. Questions?
For more information visit
research-it.berkeley.edu/ocr
Email:
quinnd@berkeley.edu
Research IT
Advancing Research@Berkeley