Keyword Searching

How to read pdf documents and extract information based on particular keywords?

Sometimes pdftk not handy in case of reading scanned pdf documents. pdftools will resolve these kinds of issues.

The objective is to find out particular keywords from the list of pdf files.

Suppose we have 1000 pdf files and we want to search specific keywords and extract the pieces of information like page number and pdf file names etc…

Discriminant analysis in R

Data analysis in R pdf tools

The below-mentioned script will be useful for the same.

```
library(pdftools)
library(stringr)
library(gtools)
setwd("/data/common/")
specificwords<-c("Tablet ", "Medicine")
files<-list.files(pattern= ".pdf$")
Final<-NULL
for(k in 1:length(files)
{file<-unlist(pdftools::pdf_text(pdf= paste("/data/common/",file[k],
sep= " "))
file<-matrix(tolower(file)
res<-data.frame(str_detect(specificwords,file))
colnames(res)<- "Result"
res1<-droplevels(subset(res,res$Result==TRUE))
if(dim(res1)>0)
{PageNumber<-data.frame(rownames(res1))
}else
PageNumber<-"Not Found"
out<-data.frame(files[k], PageNumber, specificwords)
colnames(out)<-c("FileName","Page Number","Specific Word")
Final<-rbind(out,Final)}
Final
```

pdftools also can be used for splitting, merging etc…

Here we are using pdftk for splitting, merging, attaching & unpacking.

Merge pdf files

How to merge pdf files in R?

Suppose if you want to merge n number of documents use below mentioned script.

```
as<-list.files(Inputfile,pattern=".pdf")
as<-mixedsort(as)
```

```
BC<-NULL
for(h in 1:length(as))
{AB<-paste(Inputfile,as[h],sep="")
BC<-paste(BC,AB,sep=" ")}
pdf<-paste("pdftk",BC,"cat","output","MergedFile.pdf")
system(pdf)
```

Suppose if you want the merged files with a particular sequence then name the original files accordingly (alphabetically or numbering).

How to calculate sample size?

Split pdf files

How to split the pdf document in R?

Sometimes if you want to split the document, can use the “burst” option.

Refer to the mentioned script for splitting pdf files.

```
as<-list.files(Inputfile,pattern=".pdf")
for(h in 1:length(as))
{ AB<-paste(Inputfile,as[h],sep="")
CD<-paste(OutputDirectory,as[h],"%02d.pdf",sep="")
pdf<-paste("pdftk",AB,"burst","output",CD)
system(pdf)}
```

Unpack pdf files

How to unpack pdf files?

In most cases, pdf files contain some types of attachments. Suppose if you want to extract these attached files use the “unpack_files” option.

```
as<-list.files(Inputfile,pattern=".pdf")
for(h in 1:length(as))
{ AB<-paste(Inputfile,as[h],sep="")
MD<-gsub(".pdf","",as[h])
CD<-paste(OutputDirectory,MD,sep="")
dir.create(CD)
pdf<-paste("pdftk",AB,"unpack_files ","output",CD)
system(pdf) }
```

pdf Attachment

How to attach documents into pdf files?

This method will be very helpful in most situations. You can attach the word, excel, ppt, pdf files, etc… into pdf document.

Use the below-mentioned script for attaching documents into pdf file.

```
Filename<-"MainFile.pdf"
MainFile<-paste(InputFolderName,"/",Filename,sep="")
SaveTo<-paste(OutputFolderName,"/","Attached.",Filename,sep="")
```

```
Attachment<-paste(OutputFolderName,"/*",sep="")
pdf<-paste("pdftk ",MainFile," attach_file ",Attachment," output
",SaveTo)
system(pdf)
```

Compress pdf

pdf files can compress based on below mentioned command

```
BC<-NULL
for(h in 1:length(as))
{AB<-paste("/data/common/",as[h],sep="")
BC<-paste(OutputDirectory,"Compressed-",as[h],sep="")}
pdf<-paste("pdftk",AB,"output",BC,"compress")
system(pdf)
```