Project 2 Part 3

Kaushal Khatiwada

2024-03-18

Word Cloud

```
library(pdftools)
```

```
## Using poppler version 23.08.0
```

```
library(tm)
```

```
## Loading required package: NLP
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(RColorBrewer)
```

lapply function is used to apply pdf_text function to all the element of list “pdfs”

```
setwd("~/MDS503P2")
pdfs <- list.files(pattern="*.pdf")
alltext <- lapply(pdfs, pdf_text)
```

```
## PDF error: Expected the optional content group list, but wasn’t able to find it, or it isn’t an Arra
```

```
alltext <- unlist(alltext)
head(alltext,n=4)
```

```
## [1] "Big Data,\nMining, and\nAnalytics\n               Components\n               of Strategic\n
## [2] " Big Data,\nMining, and\n Analytics\n       Components of\n Strategic Decision Making\n"
## [3] ""
## [4] " Big Data,\nMining, and\n Analytics\n       Components of\n Strategic Decision Making\n\n\n\n\n
```

create a single Corpus object containing all the text data, rather than separate Corpus objects for each PDF.

```
myCorpus <- Corpus(VectorSource(alltext))

myCorpus <- tm_map(myCorpus, content_transformer(tolower))
myCorpus <- tm_map(myCorpus,removePunctuation)
myCorpus <- tm_map(myCorpus,removeNumbers)
removeURL <- function(X) gsub("http[^[:space:]]*","",X)
myCorpus <- tm_map(myCorpus, content_transformer(function(x) gsub("\\\n", " ", x)))
myCorpus <- tm_map(myCorpus,removeURL)
myCorpus <- tm_map(myCorpus,removeWords,stopwords("english"))
myCorpus <- tm_map(myCorpus,removeWords,c("can","may","eg","ie","h","b","p","k","g","q","set","used"))
myCorpus <- tm_map(myCorpus, content_transformer(function(x) gsub("\n*","", x)))
myCorpus <- tm_map(myCorpus,stripWhitespace)

inspect(myCorpus[1:2])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 2
##
## [1] big data mining analytics components strategic decision making stephan kudyba foreword thomas da
## [2]  big data mining analytics components strategic decision making
```

```
myTdm <- TermDocumentMatrix(myCorpus,control = list(wordLengths=c(2,Inf)))
m <- as.matrix(myTdm)
freq <- sort(rowSums(m),decreasing = T)
```

Word Cloud

```
set.seed(1234)
wordcloud(words=names(freq),freq=freq, min.freq = 150,random.order = F,
          max.words = 120, colors = brewer.pal(8,"Dark2"),
          rot.per = 0.35,scale = c(5, 0.3),)
```

data
mining
users
general
statistical
threshold
processing
algorithms
time
clusters
discovery
systems
multiple
tuple
chapter
pattern
based
two
will
concepts
system
well
han
dimension
conf
knowledge
clustering
rd
query
itemsets
outliers
training
cube
attributes
class
methods
given
applications
classification
also
classi
large
attribute
learning
edition
example
analysis
rule
different
section
decision
number
networks
multidimensional
order
use
measures
tuples
model
concept
models
cluster
new
object
minimum
table
level
jiawei
using
items
dimensions
databases
process
int
rules
network
value
algorithm
pp
cation
tree
outlier
techniques
values
figure
objects
computer
item
frequent
information
sets
olap
however
computation
hierarchy
support
patterns
one
search
web
user
proc
database
many
form
sales
include
association
method
warehouse
space
measure
suppose
customer
detection
approach
therefore
distance