Dropout Prediction: A Systematic Literature Review

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. For the analysis of the articles selected after the abstract teste screening for the Systematic Review. <https://cran.r-project.org/web/packages/bibtex/bibtex.pdf>

This was developed with quanteda version 1.1.1, installation instructions available [here](https://cran.r-project.org/web/packages/quanteda/index.html), [vignetes](https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html). The pdfs where readed with [readtext](https://cran.r-project.org/web/packages/readtext/index.html) which suport functions for importing and handling text files and formatted text files with additional-data, such: ‘.csv’, ‘.tab’, ‘.json’, ‘.xml’, ‘.html’, ‘.pdf’, ‘.doc’, ‘.docx’, ‘.rtf’, ‘.xls’, ‘.xlsx’, and others.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30), 774. https://doi.org/10.21105/joss.00774

Silge & Robinson (2016) developed [tidytext] to make text mining tasks easier, more effective and consistent with tools already in wide use. This package provides commands that allow you to convert text to and from tidy formats. Allow analysis and visualisation: sentiment analysis, tf-idf statistics, n-grams or topic modelling. Best for visualization of the output

We will be using also [revtools](https://cran.r-project.org/web/packages/revtools/index.html) Screening is usually achieved by manually sorting titles or abstracts one at a time. screen\_topics offers an alternative by allowing the user to group data by any column in the input dataset, and running a topic model on the resulting data. This allows a great deal of flexibility to locate patterns in journals, years, or authors, rather than just articles.  
Data points can be selected or excluded individually, or by topic.

# get rmarkdown directory  
caminho<-getwd()  
# set working directory  
setwd(caminho)

# Import Libraries

# Setup library  
library('quanteda')  
library('readtext')  
library('revtools')

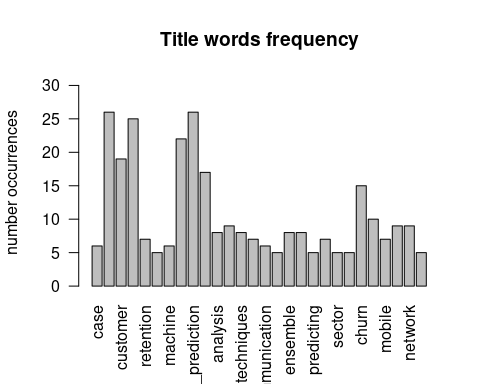
# revtools

data\_revtools<-read\_bibliography("asreview/asreview\_result\_dropout-prediction-a-systematic-literature-review-v3\_FINAL.csv")  
data\_revtools<- data\_revtools[data\_revtools$final\_included==1,]  
#analysis\_revtools<-screen\_topics(data\_revtools)  
#analysis\_revtools

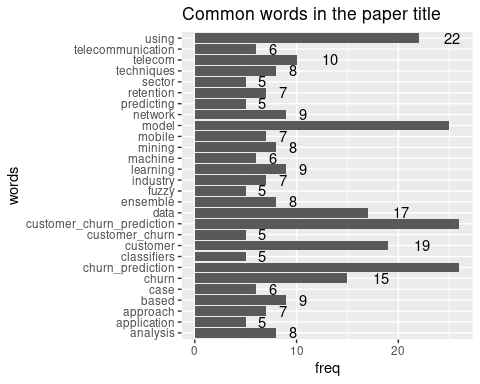
## Creating a document-term matrix for the title

min\_freq: minimum proportion of entries that a term must be found in to be retained in the analysis. Defaults to 0.01. We have adjusted to 0.05

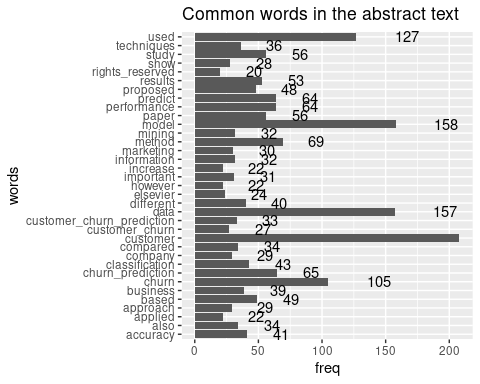
library(lattice)  
library(ggplot2)  
  
x\_dtm<-make\_dtm(data\_revtools$title,min\_freq = 0.05)  
  
#create a matrix of the document term-matrix  
x\_dtm\_matrix <- as.matrix(x\_dtm)  
#apply function to calculate columns totals  
totals<-apply(x\_dtm\_matrix,MARGIN = 2,sum)  
par(las=2)  
barplot(totals,ylim = c(0,max(totals)+5),main = "Title words frequency",ylab = "number occurrences")



#create dataframe  
df\_dtm<-data.frame(totals)  
df\_dtm$words<-row.names(df\_dtm)  
   
ggplot(data = df\_dtm,aes(x=words,y=totals)) +  
 labs(title = "Common words in the paper title",y = "freq")+ #modify legend and plot labels  
 geom\_bar(stat = "identity") + coord\_flip()+  
 geom\_text(aes(label=totals),hjust=-1.6)



x\_dtm<-make\_dtm(data\_revtools$abstract,min\_freq = 0.20)  
#create a matrix of the document term-matrix  
x\_dtm\_matrix <- as.matrix(x\_dtm)  
#apply function to calculate columns totals  
totals<-apply(x\_dtm\_matrix,MARGIN = 2,sum)  
#create dataframe  
df\_dtm<-data.frame(totals)  
df\_dtm$words<-row.names(df\_dtm)  
   
ggplot(data = df\_dtm,aes(x=words,y=totals)) +  
 labs(title = "Common words in the abstract text",y = "freq")+ #modify legend and plot labels  
 geom\_bar(stat = "identity") + coord\_flip()+  
 geom\_text(aes(label=totals),hjust=-1.6)

 #Select only select articles

pdf.files <- list.files(path=caminho,recursive=T,pattern="pdf$",full.names=TRUE)

88 PDFs to analyse

# Reading the PDFs

pdfs<-readtext(pdf.files[1],docvarsfrom = "filenames", sep = "\_", docvarnames = c("author", "year","title"))  
for (file in pdf.files[2:length(pdf.files)]) {  
 pdf<-readtext(file,docvarsfrom = "filenames", sep = "\_", docvarnames = c("author", "year","title"))  
 pdfs <- rbind(pdfs,pdf)  
}  
pdfs

## readtext object consisting of 85 documents and 3 docvars.  
## # Description: df[,5] [85 × 5]  
## doc\_id text author year title   
## <chr> <chr> <chr> <int> <chr>   
## 1 Hung et al\_2006\_Applying … "\"Assoc… Hung et … 2006 Applying data mining to …  
## 2 Dierkes et al\_2011\_Estima… "\" … Dierkes … 2011 Estimating the effect of…  
## 3 Liu et al\_2020\_Micro- and… "\" … Liu et al 2020 Micro- and macro-level c…  
## 4 Kaya et al\_2018\_Behaviora… "\"Kaya … Kaya et … 2018 Behavioral attributes an…  
## 5 Mitrovic et al\_2018\_On th… "\" On … Mitrovic… 2018 On the operational effic…  
## 6 Oskarsdottir et al\_2018\_P… "\" Prof… Oskarsdo… 2018 Profit-Based Model Selec…  
## # … with 79 more rows

# Transform the journal articles into a corpus object  
pdfs\_corpus <- corpus(pdfs)  
summary(pdfs\_corpus, 5)

## Corpus consisting of 85 documents, showing 5 documents:  
##   
## Text  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf  
## Types Tokens Sentences author year  
## 1607 6424 245 Hung et al 2006  
## 2292 12093 586 Dierkes et al 2011  
## 2537 16527 643 Liu et al 2020  
## 2570 12573 491 Kaya et al 2018  
## 2608 14009 496 Mitrovic et al 2018  
## title  
## Applying data mining to telecom churn management  
## Estimating the effect of word of mouth on churn and cross-buying in the mobile  
## Micro- and macro-level churn analysis of large-scale mobile games  
## Behavioral attributes and financial churn prediction  
## On the operational efficiency of different feature types for telco Churn

# Build a document feature matrix

pdfs\_corpus <- corpus(pdfs[1:10,])  
corpus\_DFM <- dfm(pdfs\_corpus, tolower = TRUE, stem = FALSE, remove = c('@',"et", "al", "fig", "table", "ml", "http", stopwords(language = 'en',source = "smart")),  
 remove\_punct = TRUE, remove\_numbers = TRUE)  
  
dfm\_sort(corpus\_DFM)

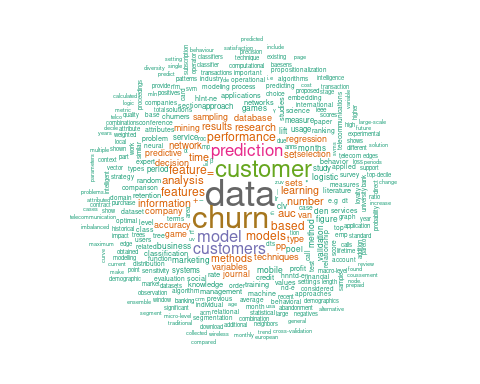
## Document-feature matrix of: 10 documents, 7,830 features (80.0% sparse) and 3 docvars.  
## features  
## docs data  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 67  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 87  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 49  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 140  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 36  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 17  
## features  
## docs churn  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 77  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 121  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 191  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 84  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 26  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 37  
## features  
## docs customer  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 53  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 70  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 3  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 74  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 32  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 83  
## features  
## docs prediction  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 20  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 25  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 80  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 59  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 16  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 18  
## features  
## docs model  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 62  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 36  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 52  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 21  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 23  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 70  
## features  
## docs customers  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 32  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 55  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 0  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 72  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 13  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 33  
## features  
## docs performance  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 24  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 6  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 26  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 17  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 19  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 54  
## features  
## docs =  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 25  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 12  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 95  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 11  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 23  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 49  
## features  
## docs feature  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 1  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 1  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 28  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 36  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 136  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 3  
## features  
## docs models  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 22  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 28  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 17  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 31  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 4  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 37  
## [ reached max\_ndoc ... 4 more documents, reached max\_nfeat ... 7,820 more features ]

stopwords("english")

## [1] "i" "me" "my" "myself" "we"   
## [6] "our" "ours" "ourselves" "you" "your"   
## [11] "yours" "yourself" "yourselves" "he" "him"   
## [16] "his" "himself" "she" "her" "hers"   
## [21] "herself" "it" "its" "itself" "they"   
## [26] "them" "their" "theirs" "themselves" "what"   
## [31] "which" "who" "whom" "this" "that"   
## [36] "these" "those" "am" "is" "are"   
## [41] "was" "were" "be" "been" "being"   
## [46] "have" "has" "had" "having" "do"   
## [51] "does" "did" "doing" "would" "should"   
## [56] "could" "ought" "i'm" "you're" "he's"   
## [61] "she's" "it's" "we're" "they're" "i've"   
## [66] "you've" "we've" "they've" "i'd" "you'd"   
## [71] "he'd" "she'd" "we'd" "they'd" "i'll"   
## [76] "you'll" "he'll" "she'll" "we'll" "they'll"   
## [81] "isn't" "aren't" "wasn't" "weren't" "hasn't"   
## [86] "haven't" "hadn't" "doesn't" "don't" "didn't"   
## [91] "won't" "wouldn't" "shan't" "shouldn't" "can't"   
## [96] "cannot" "couldn't" "mustn't" "let's" "that's"   
## [101] "who's" "what's" "here's" "there's" "when's"   
## [106] "where's" "why's" "how's" "a" "an"   
## [111] "the" "and" "but" "if" "or"   
## [116] "because" "as" "until" "while" "of"   
## [121] "at" "by" "for" "with" "about"   
## [126] "against" "between" "into" "through" "during"   
## [131] "before" "after" "above" "below" "to"   
## [136] "from" "up" "down" "in" "out"   
## [141] "on" "off" "over" "under" "again"   
## [146] "further" "then" "once" "here" "there"   
## [151] "when" "where" "why" "how" "all"   
## [156] "any" "both" "each" "few" "more"   
## [161] "most" "other" "some" "such" "no"   
## [166] "nor" "not" "only" "own" "same"   
## [171] "so" "than" "too" "very" "will"

textplot\_wordcloud(corpus\_DFM, min.freq = 30, random.order=F, rot.per = .10,colors = RColorBrewer::brewer.pal(8,'Dark2'))

## Warning: min.freqcolorsrandom.orderrot.per is deprecated; use  
## min\_countcolorrandom\_orderrotation instead



#summary(corpus\_DFM)  
topfeatures(corpus\_DFM, 100)

## data churn customer prediction model   
## 1018 842 616 390 378   
## customers performance = feature models   
## 357 243 241 240 239   
## features based analysis auc number   
## 238 231 223 215 202   
## learning research time information methods   
## 200 199 198 196 190   
## results set network sets techniques   
## 181 165 162 161 156   
## + database company decision variables   
## 149 145 144 143 141   
## accuracy pp sampling van type   
## 141 141 140 140 139   
## regression mining predictive game selection   
## 137 136 132 132 130   
## journal method mobile marketing business   
## 128 125 124 123 123   
## clv logistic figure lift validation   
## 120 119 118 118 117   
## den poel games months study   
## 116 115 114 113 109   
## period − service usage approach   
## 104 104 103 103 103   
## measure systems ∗ retention classification   
## 101 100 100 98 97   
## rate credit networks knowledge studies   
## 96 95 94 93 93   
## management applications random training profit   
## 92 91 91 91 89   
## literature churners behavior social hlnt-ne   
## 88 86 85 84 84   
## problem values machine relationship section   
## 81 81 80 79 79   
## e.g class services test science   
## 79 78 77 77 77   
## level tree expert process call   
## 76 75 75 74 73   
## attributes comparison ranking dataset order   
## 73 73 73 73 72

library(quanteda.textmodels)

##   
## Attaching package: 'quanteda.textmodels'

## The following object is masked from 'package:quanteda':  
##   
## data\_dfm\_lbgexample

tmod\_wf <- textmodel\_wordfish(corpus\_DFM,c(6,5))  
summary(tmod\_wf)

##   
## Call:  
## textmodel\_wordfish.dfm(x = corpus\_DFM, dir = c(6, 5))  
##   
## Estimated Document Positions:  
## theta  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf -0.25075  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 0.30852  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 1.50768  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 0.07849  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 1.09740  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 0.71278  
## Zhu et al\_2018\_Benchmarking sampling techniques for imbalance learning in churn prediction.pdf 0.28315  
## Garcia et al\_2017\_Intelligent data analysis approaches to churn as a business problem.pdf -1.01442  
## Ballings Van den Poel\_2012\_Customer event history for churn prediction.pdf -1.23073  
## Ballings et al\_2012\_Improving Customer Churn Prediction by Data Augmentation Using Pictorial.pdf -1.49212  
## se  
## Hung et al\_2006\_Applying data mining to telecom churn management.pdf 0.024607  
## Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf 0.017540  
## Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf 0.005849  
## Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf 0.017138  
## Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf 0.010834  
## Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf 0.016474  
## Zhu et al\_2018\_Benchmarking sampling techniques for imbalance learning in churn prediction.pdf 0.018177  
## Garcia et al\_2017\_Intelligent data analysis approaches to churn as a business problem.pdf 0.009288  
## Ballings Van den Poel\_2012\_Customer event history for churn prediction.pdf 0.016191  
## Ballings et al\_2012\_Improving Customer Churn Prediction by Data Augmentation Using Pictorial.pdf 0.009036  
##   
## Estimated Feature Scores:  
## association information systems ais electronic library aisel  
## beta -1.7466 -0.3902 -0.9337 -0.5899 -0.4507 -0.1264 -0.5899  
## psi -0.3103 2.8355 2.0106 -1.3768 -0.2524 -1.0556 -1.3768  
## pacific asia conference pacis proceedings december applying data  
## beta -0.06554 -0.10361 0.5818 -0.6024 1.3765 -1.552 -0.2033 -0.9112  
## psi -1.06221 0.04084 1.5229 -0.8689 0.4712 -1.598 0.6984 4.3412  
## mining telecom churn management shin-yuan hung national  
## beta -0.5269 -1.067 -0.0002762 -0.8703 -0.4011 -1.2979 -0.6291  
## psi 2.4506 1.364 4.2773796 1.9549 -1.0574 -0.4019 -0.4055  
## chung-cheng university hsiu-yu wang follow additional works  
## beta -0.5976 0.3209 -0.4011 -0.5597 -0.2977 -0.02532 1.7240  
## psi -1.0909 1.3024 -1.0574 0.7094 -0.8275 1.30694 -0.3161  
## http://aisel.aisnet.org/pacis2004  
## beta -0.5351  
## psi -2.4636

# Extract data from PDFs

## Images

pdf.files

## [1] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11391/Hung et al\_2006\_Applying data mining to telecom churn management.pdf"   
## [2] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11393/Dierkes et al\_2011\_Estimating the effect of word of mouth on churn and cross-buying in the mobile.pdf"   
## [3] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11397/Liu et al\_2020\_Micro- and macro-level churn analysis of large-scale mobile games.pdf"   
## [4] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11402/Kaya et al\_2018\_Behavioral attributes and financial churn prediction.pdf"   
## [5] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11406/Mitrovic et al\_2018\_On the operational efficiency of different feature types for telco Churn.pdf"   
## [6] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11408/Oskarsdottir et al\_2018\_Profit-Based Model Selection for Customer Retention Using Individual Customer.pdf"   
## [7] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11411/Zhu et al\_2018\_Benchmarking sampling techniques for imbalance learning in churn prediction.pdf"   
## [8] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11415/Garcia et al\_2017\_Intelligent data analysis approaches to churn as a business problem.pdf"   
## [9] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11432/Ballings Van den Poel\_2012\_Customer event history for churn prediction.pdf"   
## [10] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11440/Ballings et al\_2012\_Improving Customer Churn Prediction by Data Augmentation Using Pictorial.pdf"   
## [11] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11453/Burez Van den Poel\_2009\_Handling class imbalance in customer churn prediction.pdf"   
## [12] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11456/Burez Vandenpoel\_2008\_Separating financial from commercial customer churn.pdf"   
## [13] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11459/Burez Van den Poel\_2007\_CRM at a pay-TV company.pdf"   
## [14] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11461/Wai-Ho Au et al\_2003\_A novel evolutionary data mining algorithm with applications to churn prediction.pdf"   
## [15] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11467/De Bock Van den Poel\_2010\_Ensembles of Probability Estimation Trees for Customer Churn Prediction.pdf"   
## [16] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11477/Antipov Pokryshevskaya\_2010\_Applying CHAID for logistic regression diagnostics and classification accuracy.pdf"   
## [17] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11479/Al-Molhem et al\_2019\_Social network analysis in Telecom data.pdf"   
## [18] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11487/Gok et al\_2015\_A Case Study for the Churn Prediction in Turksat Internet Service Subscription.pdf"   
## [19] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11488/Azeem et al\_2017\_A churn prediction model for prepaid customers in telecom using fuzzy.pdf"   
## [20] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11491/Agrawal et al\_2018\_Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis.pdf"   
## [21] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11492/Amornvetchayakul Phumchusri\_2020\_Customer Churn Prediction for a Software-as-a-Service Inventory Management.pdf"  
## [22] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11493/Benedek et al\_2014\_The Importance of Social Embeddedness.pdf"   
## [23] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11494/Esteves-Mendes-Moreira\_2016\_Churn perdiction in the telecom business.pdf"   
## [24] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11495/Farquad et al\_2009\_Data Mining Using Rules Extracted from SVM.pdf"   
## [25] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11496/Jafari-Marandi et al\_2020\_Optimum profit-driven churn decision making.pdf"   
## [26] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11497/Jahromi et al\_2016\_Customer Churn Models.pdf"   
## [27] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11498/Mitrovic et al\_2017\_Scalable RFM-enriched Representation Learning for Churn Prediction.pdf"   
## [28] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11499/Huang et al\_2009\_Customer Churn Prediction for Broadband Internet Services.pdf"   
## [29] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11500/Idris et al\_2013\_Intelligent churn prediction in telecom.pdf"   
## [30] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11501/Mohanty-Rani\_2015\_Application of Computational Intelligence to Predict Churn and Non-Churn of.pdf"   
## [31] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11502/Wang-Xiao\_2011\_Transfer Ensemble Model for Customer Churn Prediction with Imbalanced Class.pdf"   
## [32] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11503/Xiao et al\_2015\_Feature-selection-based dynamic transfer ensemble model for customer churn.pdf"   
## [33] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11504/Vijaya-Sivasankar\_2019\_An efficient system for customer churn prediction through particle swarm.pdf"   
## [34] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11505/Semrl-Matei\_2017\_Churn prediction model for effective gym customer retention.pdf"   
## [35] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11506/Shao et al\_2008\_Construction of Bayesian Classifiers with GA for Predicting Customer Retention.pdf"   
## [36] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11507/Sivasankar-Vijaya\_2019\_Hybrid PPFCM-ANN model.pdf"   
## [37] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11508/Ullah et al\_2019\_Churn Prediction in Banking System using K-Means, LOF, and CBLOF.pdf"   
## [38] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11509/Perianez et al\_2016\_Churn Prediction in Mobile Social Games.pdf"   
## [39] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11510/Prasasti-Ohwada\_2014\_Applicability of machine-learning techniques in predicting customer defection.pdf"   
## [40] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11511/Routh et al\_2020\_Estimating customer churn under competing risks.pdf"   
## [41] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11512/Naik-Reddy\_2017\_An innovative optimized model to anticipate clients about immigration in.pdf"   
## [42] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11514/Zhang et al\_2012\_Predicting customer churn through interpersonal influence.pdf"   
## [43] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11517/Xie et al\_2009\_Customer churn prediction using improved balanced random forests.pdf"   
## [44] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11520/Xiao et al\_2012\_Dynamic classifier ensemble model for customer classification with imbalanced.pdf"   
## [45] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11525/Wei-Chiu\_2002\_Turning telecommunications call details to churn prediction.pdf"   
## [46] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11528/Shirazi-Mohammadi\_2019\_A big data analytics model for customer churn prediction in the retiree segment.pdf"   
## [47] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11531/Coussement et al\_2017\_A comparative analysis of data preparation algorithms for customer churn.pdf"   
## [48] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11534/De Bock-Poel\_2011\_An empirical evaluation of rotation-based ensemble classifiers for customer.pdf"   
## [49] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11537/Verbeke et al\_2011\_Building comprehensible customer churn prediction models with advanced rule.pdf"   
## [50] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11540/Kianmehr-Alhajj\_2009\_Calling communities analysis and identification using machine learning.pdf"   
## [51] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11543/Coussement-Van den Poel\_2008\_Churn prediction in subscription services.pdf"   
## [52] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11546/Farquad et al\_2014\_Churn prediction using comprehensible support vector machine.pdf"   
## [53] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11549/Nie et al\_2011\_Credit card churn forecasting by logistic regression and decision tree.pdf"   
## [54] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11552/Coussement-De Bock\_2013\_Customer churn prediction in the online gambling industry.pdf"   
## [55] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11555/Amin et al\_2017\_Customer churn prediction in the telecommunication sector using a rough set.pdf"   
## [56] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11558/Delen et al\_2020\_Development of a Bayesian Belief Network-based DSS for predicting and.pdf"   
## [57] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11561/Alboukaey et al\_2020\_Dynamic behavior based churn prediction in mobile telecom.pdf"   
## [58] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11564/Gur Ali-Aniturk\_2014\_Dynamic churn prediction framework with more effective use of rare event data.pdf"   
## [59] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11567/Schaeffer-Rodriguez Sanchez\_2020\_Forecasting client retention — A machine-learning approach.pdf"   
## [60] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11570/Keramati et al\_2014\_Improved churn prediction in telecommunication industry using data mining.pdf"   
## [61] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11572/Coussement et al\_2010\_Improved marketing decision making in a customer churn prediction context using.pdf"   
## [62] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11576/Coussement-Poel\_2009\_Improving customer attrition prediction by integrating emotions from.pdf"   
## [63] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11579/Benoit-Van den Poel\_2012\_Improving customer retention in financial services using kinship network.pdf"   
## [64] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11582/Moeyersoms-Martens\_2015\_Including high-cardinality attributes in predictive models.pdf"   
## [65] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11585/Glady et al\_2009\_Modeling churn using customer lifetime value.pdf"   
## [66] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11587/Verbeke et al\_2012\_New insights into churn prediction in the telecommunication sector.pdf"   
## [67] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11594/De Bock-Van den Poel\_2012\_Reconciling performance and interpretability in customer churn prediction using.pdf"   
## [68] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11597/Verbeke et al\_2014\_Social network analysis for customer churn prediction.pdf"   
## [69] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11600/Risselada et al\_2010\_Staying Power of Churn Prediction Models.pdf"   
## [70] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11603/Tsai-Chen\_2010\_Variable selection by association rules for customer churn prediction of.pdf"   
## [71] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11667/Liao-Chueh\_2011\_Applying Fuzzy Data Mining to Telecom Churn Management.pdf"   
## [72] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11668/Lee-Jo\_2010\_Bayesian Network Approach to Predict Mobile Churn Motivations.pdf"   
## [73] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11669/Mohanty-Naga Ratna Sree\_2018\_Churn and Non-churn of Customers in Banking Sector Using Extreme Learning.pdf"   
## [74] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11670/Khodabandehlou-Zivari Rahman\_2017\_Comparison of supervised machine learning techniques for customer churn.pdf"   
## [75] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11671/Vijaya et al\_2019\_Fuzzy Clustering with Ensemble Classification Techniques to Improve the.pdf"   
## [76] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11672/Radosavljevik-van der Putten\_2013\_Preventing Churn in Telecommunications.pdf"   
## [77] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11673/Verbraken et al\_2014\_Profit optimizing customer churn prediction with Bayesian network classifiers.pdf"   
## [78] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11674/Jiang et al\_2014\_Research on Customers Churn Prediction Model Based on Logistic.pdf"   
## [79] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11675/Ascarza\_2018\_Retention Futility.pdf"   
## [80] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11676/Hutchison et al\_2010\_Rule Extraction from Support Vector Machine Using Modified Active Learning.pdf"   
## [81] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11679/Martono et al\_2014\_Utilizing Customers’ Purchase and Contract Renewal Details to Predict Defection.pdf"   
## [82] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11680/Amin et al\_2015\_Churn Prediction in Telecommunication Industry Using Rough Set Approach.pdf"   
## [83] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11681/Saravanan-Vijay Raajaa\_2012\_A Graph-Based Churn Prediction Model for Mobile Telecom Networks.pdf"   
## [84] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11682/Farquad et al\_2012\_Analytical CRM in banking and finance using SVM.pdf"   
## [85] "/mnt/hgfs/nuvem/Dropbox/doutoramento/tese/SLRDropout/analysis/selected\_articles/final\_dataset/files/11683/Kianmehr-Alhajj\_2011\_A fuzzy prediction model for calling communities.pdf"

for(image in pdf.files) {  
   
}  
#imageFiles <-PDF\_extractImages(pdf.files[2])

## Tables

Creating a list of lists. Lists are indexed with double brackets.

library("tabulizer")  
library(utils)  
  
pb = txtProgressBar(min = 0, max = length(pdf.files), initial = 0,)  
tab = list()  
for (i in 1:length(pdf.files)){  
 setTxtProgressBar(pb,i)  
 try(  
 tab[[i]]<-extract\_tables(pdf.files[i]),  
 silent = TRUE  
 )  
}

## ================================================================================

### Identifying elements

Perceber quais são os artigos que abordam isto…. tentar fazer um quadro global Por exemplo x artigos abordam AUC, sensitivy…. blá… blá…

Estão exemplos no Edge AOC

# Where is AUC  
grep("AUC",tab)

## [1] 6 7 8 10 23 27 29 32 38 42 44 47 48 49 51 56 60 61 62 63 64 66 67

grep("AUC",tab[[6]][[4]])

## [1] 56

tab[[6]][[4]]

## [,1] [,2] [,3] [,4] [,5]   
## [1,] "" "" "" "" "Measure"   
## [2,] "Dataset" "Method" "AUC" "H-measure" "Top decile lift EMP"   
## [3,] "" "LR" "0.75" "0.22" "1.29 306.63"   
## [4,] "" "DT" "0.82" "0.36" "1.47 306.60"   
## [5,] "D1" "RF XGB" "0.85 0.85" "0.41 0.41" "1.46 1.51 306.79 306.81"  
## [6,] "" "NN" "0.86" "0.44" "1.53 306.63"   
## [7,] "" "SVM" "0.83" "0.38" "1.51 306.59"   
## [8,] "" "LR" "0.71" "0.21" "1.86 224.25"   
## [9,] "" "DT" "0.72" "0.26" "2.13 224.24"   
## [10,] "D2" "RF XGB" "0.75 0.82" "0.30 0.38" "2.07 2.87 224.26 224.56"  
## [11,] "" "NN" "0.73" "0.23" "1.98 224.13"   
## [12,] "" "SVM" "0.72" "0.23" "2.14 224.17"   
## [13,] "" "LR" "0.58" "0.03" "1.04 389.32"   
## [14,] "" "DT" "0.62" "0.05" "1.05 389.32"   
## [15,] "D3" "RF XGB" "0.64 0.64" "0.07 0.07" "1.05 1.04 389.32 389.32"  
## [16,] "" "NN" "0.63" "0.06" "1.03 389.32"   
## [17,] "" "SVM" "0.58" "0.03" "1.05 389.32"   
## [18,] "" "LR" "0.69" "0.16" "1.26 171.54"   
## [19,] "" "DT" "0.90" "0.55" "2.86 173.46"   
## [20,] "D4" "RF XGB" "0.92 0.95" "0.58 0.66" "2.14 3.39 174.37 174.50"  
## [21,] "" "NN" "0.85" "0.43" "2.55 173.00"   
## [22,] "" "SVM" "0.80" "0.37" "2.02 171.52"   
## [23,] "" "LR" "0.84" "0.40" "2.29 98.10"   
## [24,] "" "DT" "0.88" "0.64" "5.04 97.65"   
## [25,] "D5" "RF XGB" "0.91 0.93" "0.71 0.75" "3.05 5.77 97.90 98.54"   
## [26,] "" "NN" "0.75" "0.26" "2.52 97.54"   
## [27,] "" "SVM" "0.87" "0.48" "3.16 97.98"   
## [,6] [,7]   
## [1,] "" ""   
## [2,] "Mean EMPβ′" "Median EMPβ′"  
## [3,] "281.98" "172.57"   
## [4,] "288.96" "172.22"   
## [5,] "282.77 283.42" "170.95168.68"  
## [6,] "293.89" "171.55"   
## [7,] "297.29" "171.58"   
## [8,] "213.57" "124.91"   
## [9,] "203.94" "124.02"   
## [10,] "203.87 208.62" "124.35123.94"  
## [11,] "203.30" "123.00"   
## [12,] "208.99" "123.45"   
## [13,] "355.01" "220.35"   
## [14,] "354.15" "220.07"   
## [15,] "365.49 356.80" "219.67218.81"  
## [16,] "365.11" "219.39"   
## [17,] "357.10" "218.30"   
## [18,] "157.41" "92.93"   
## [19,] "158.28" "95.06"   
## [20,] "158.78 159.24" "96.1996.10"   
## [21,] "161.57" "95.51"   
## [22,] "165.27" "94.60"   
## [23,] "90.95" "53.03"   
## [24,] "89.66" "51.99"   
## [25,] "89.44 92.41" "52.6753.04"   
## [26,] "90.05" "51.08"   
## [27,] "91.89" "53.25"