R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

Import Libraries

```
# Setup library

#install.packages("bib2df")
#install.packages('journalabbr')
library(journalabbr)
library(bib2df)
library(tibble)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
setwd('C:/nuvem/Dropbox/doutoramento/tese/SLRDropout/sources/allSources')
```

List all the files with bib extension

```
files<-list.files(pattern = '\\.bib$')
files
```

```
## [1] "acm.bib"                              "ieee.bib"
## [3] "scienceDirect_withContractual1.bib" "scienceDirect_withContractual2.bib"
## [5] "scopus.bib"                           "springer_Exported Items.bib"
## [7] "webofscience_Exported Items.bib"
```

Reading bibtex source files

```
acm<-bib2df(file = files[1])
```

```
## Warning: `as_data_frame()` is deprecated as of tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
acm
```

```
## # A tibble: 8 x 36
##   CATEGORY BIBTEXKEY ADDRESS ANNOTE AUTHOR BOOKTITLE CHAPTER CROSSREF EDITION
##   <chr>    <chr>     <chr>   <chr>  <list> <chr>     <chr>   <chr>    <chr>
## 1 INPROCE~ 10.1145/~ New Yo~ <NA>   <chr ~ Proceedi~ <NA>    <NA>     <NA>
## 2 INPROCE~ 10.1145/~ New Yo~ <NA>   <chr ~ Proceedi~ <NA>    <NA>     <NA>
## 3 INPROCE~ 10.1145/~ New Yo~ <NA>   <chr ~ Proceedi~ <NA>    <NA>     <NA>
## 4 ARTICLE  10.1145/~ New Yo~ <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 5 ARTICLE  10.1145/~ New Yo~ <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 6 ARTICLE  10.1145/~ New Yo~ <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 7 PROCEED~ 10.1145/~ New Yo~ <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 8 INPROCE~ 10.1145/~ New Yo~ <NA>   <chr ~ Proceedi~ <NA>    <NA>     <NA>
## # ... with 27 more variables: EDITOR <list>, HOWPUBLISHED <chr>,
## #   INSTITUTION <chr>, JOURNAL <chr>, KEY <chr>, MONTH <chr>, NOTE <chr>,
## #   NUMBER <chr>, ORGANIZATION <chr>, PAGES <chr>, PUBLISHER <chr>,
## #   SCHOOL <chr>, SERIES <chr>, TITLE <chr>, TYPE <chr>, VOLUME <chr>,
## #   YEAR <dbl>, ISBN <chr>, URL <chr>, DOI <chr>, ABSTRACT <chr>,
## #   NUMPAGES <chr>, KEYWORDS <chr>, LOCATION <chr>, ARTICLENO <chr>,
## #   ISSUE_DATE <chr>, ISSN <chr>
```

```
ieee<-bib2df(file = files[2])
```

```
## Column `YEAR` contains character strings.
##               No coercion to numeric applied.
```

```
ieee
```

```
## # A tibble: 20 x 182
##    CATEGORY BIBTEXKEY ADDRESS ANNOTE AUTHOR BOOKTITLE CHAPTER CROSSREF EDITION
##    <chr>    <chr>     <chr>   <chr>  <list> <chr>     <chr>   <chr>    <chr>
##  1 INPROCE~ 8256385   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  2 INPROCE~ 5461170   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  3 INPROCE~ 5351389   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  4 INPROCE~ 7564273   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  5 INPROCE~ 7224392   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  6 INPROCE~ 8428702   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
```

```
##  7 INPROCE~ 6602505   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  8 INPROCE~ 8615575   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  9 ARTICLE  9003617   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 10 INPROCE~ 5488654   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 11 INPROCE~ 6410619   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 12 INPROCE~ 1007098   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 13 INPROCE~ 8725275   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 14 INPROCE~ 8647765   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 15 INPROCE~ 9112988   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 16 ARTICLE  1514768   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 17 INPROCE~ 4378480   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 18 ARTICLE  9057003   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 19 INPROCE~ 5261844   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 20 ARTICLE  1638205   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## # ... with 173 more variables: EDITOR <list>, HOWPUBLISHED <chr>,
## #   INSTITUTION <chr>, JOURNAL <chr>, KEY <chr>, MONTH <chr>, NOTE <chr>,
## #   NUMBER <chr>, ORGANIZATION <chr>, PAGES <chr>, PUBLISHER <chr>,
## #   SCHOOL <chr>, SERIES <chr>, TITLE <chr>, TYPE <chr>, VOLUME <chr>,
## #   YEAR <chr>, AUTHOR..J. <chr>, BOOKTITLE..2017 <chr>, TITLE..CHURN <chr>,
## #   YEAR..2017.. <chr>, VOLUME.... <chr>, NUMBER.... <chr>, PAGES..1.3.. <chr>,
## #   ABSTRACT..IN <chr>, KEYWORDS..CUSTOMER <chr>,
## #   DOI..10.1109.BESC.2017.8256385.. <chr>, ISSN.... <chr>, AUTHOR..Y. <chr>,
## #   BOOKTITLE..2010 <chr>, TITLE..CLUSTERING <chr>, YEAR..2010.. <chr>,
## #   VOLUME..3.. <chr>, PAGES..1284.1287.. <chr>, ABSTRACT..THE <chr>,
## #   KEYWORDS..DATA <chr>, DOI..10.1109.ICLSIM.2010.5461170.. <chr>,
## #   AUTHOR..W. <chr>, BOOKTITLE..2009 <chr>, TITLE..A <chr>,
## #   YEAR..2009.. <chr>, PAGES..1089.1092.. <chr>, ABSTRACT..A <chr>,
## #   KEYWORDS..MIXED <chr>, DOI..10.1109.ASICON.2009.5351389.. <chr>,
## #   ISSN..2162.755X.. <chr>, AUTHOR..F. <chr>, BOOKTITLE..2016 <chr>,
## #   TITLE..FAILURE <chr>, YEAR..2016.. <chr>, PAGES..169.172.. <chr>,
## #   ABSTRACT..THERE <chr>, KEYWORDS..CMOS <chr>,
## #   DOI..10.1109.IPFA.2016.7564273.. <chr>, ISSN..1946.1550.. <chr>,
## #   AUTHOR..R. <chr>, BOOKTITLE..2015 <chr>, YEAR..2015.. <chr>,
## #   PAGES..332.335.. <chr>, KEYWORDS..FAULT <chr>,
## #   DOI..10.1109.IPFA.2015.7224392.. <chr>, AUTHOR..S. <chr>,
## #   BOOKTITLE..2018 <chr>, TITLE..CUSTOMER <chr>, YEAR..2018.. <chr>,
## #   PAGES..211.218.. <chr>, ABSTRACT..DEEP <chr>, KEYWORDS..COMPUTER <chr>,
## #   DOI..10.1109.SYSOSE.2018.8428702.. <chr>, BOOKTITLE..2013 <chr>,
## #   TITLE..MICROBLOG <chr>, YEAR..2013.. <chr>, PAGES..481.486.. <chr>,
## #   ABSTRACT..AS <chr>, KEYWORDS..BEHAVIOURAL <chr>,
## #   DOI..10.1109.ICSSSM.2013.6602505.. <chr>, ISSN..2161.1904.. <chr>,
## #   AUTHOR..N. <chr>, TITLE..ASPECT <chr>, PAGES..215.220.. <chr>,
## #   ABSTRACT..CONVOLUTIONAL <chr>, KEYWORDS..CONSUMER <chr>,
## #   DOI..10.1109.ICTER.2018.8615575.. <chr>, ISSN..2472.7598.. <chr>,
## #   AUTHOR..B. <chr>, JOURNAL..IEEE <chr>, TITLE..GENERATE <chr>,
## #   YEAR..2020.. <chr>, VOLUME..7.. <chr>, NUMBER..1.. <chr>,
## #   PAGES..28.34.. <chr>, ABSTRACT..CRITICAL <chr>, KEYWORDS..AIRCRAFT <chr>,
## #   DOI..10.1109.MPEL.2019.2959099.. <chr>, ISSN..2329.9215.. <chr>,
## #   AUTHOR..M. <chr>, TITLE..IMPLEMENTATION <chr>, PAGES..1467.1471.. <chr>,
## #   ABSTRACT..RECENTLY. <chr>, KEYWORDS..POWER <chr>, ...
```

```
science1<-bib2df(file = files[3])
science1
```

```
## # A tibble: 100 x 32
##    CATEGORY BIBTEXKEY ADDRESS ANNOTE AUTHOR BOOKTITLE CHAPTER CROSSREF EDITION
##    <chr>    <chr>     <chr>   <chr>  <list> <chr>     <chr>   <chr>    <chr>
##  1 ARTICLE  MARTINS2~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  2 ARTICLE  HILAS200~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  3 ARTICLE  WANG2015~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  4 ARTICLE  MOSTAFA2~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  5 ARTICLE  MARTINEZ~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  6 ARTICLE  CASCIO20~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  7 ARTICLE  FARQUAD2~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  8 ARTICLE  SHIRAZI2~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  9 ARTICLE  ROSENTRE~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 10 ARTICLE  PAOLANTI~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## # ... with 90 more rows, and 23 more variables: EDITOR <list>,
## #   HOWPUBLISHED <chr>, INSTITUTION <chr>, JOURNAL <chr>, KEY <chr>,
## #   MONTH <chr>, NOTE <chr>, NUMBER <chr>, ORGANIZATION <chr>, PAGES <chr>,
## #   PUBLISHER <chr>, SCHOOL <chr>, SERIES <chr>, TITLE <chr>, TYPE <chr>,
## #   VOLUME <chr>, YEAR <dbl>, ISSN <chr>, DOI <chr>, URL <chr>, KEYWORDS <chr>,
## #   ABSTRACT <chr>, ISBN <chr>
```

```
science2<-bib2df(file = files[4])
science2
```

```
## # A tibble: 26 x 32
##    CATEGORY BIBTEXKEY ADDRESS ANNOTE AUTHOR BOOKTITLE CHAPTER CROSSREF EDITION
##    <chr>    <chr>     <chr>   <chr>  <list> <chr>     <chr>   <chr>    <chr>
##  1 INCOLLE~ DEPAMPHI~ <NA>    <NA>   <chr ~ Mergers,~ <NA>    <NA>     Tenth ~
##  2 ARTICLE  2020450   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  3 ARTICLE  2013106   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  4 ARTICLE  AHAD20161 <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  5 ARTICLE  2020334   <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  6 INCOLLE~ 2004703   New Yo~ <NA>   <chr ~ Encyclop~ <NA>    <NA>     <NA>
##  7 INCOLLE~ KOTU2015~ Boston  <NA>   <chr ~ Predicti~ <NA>    <NA>     <NA>
##  8 ARTICLE  SHARMA20~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  9 INCOLLE~ KOTU2019~ <NA>    <NA>   <chr ~ Data Sci~ <NA>    <NA>     Second~
## 10 ARTICLE  REHMAN20~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## # ... with 16 more rows, and 23 more variables: EDITOR <list>,
## #   HOWPUBLISHED <chr>, INSTITUTION <chr>, JOURNAL <chr>, KEY <chr>,
## #   MONTH <chr>, NOTE <chr>, NUMBER <chr>, ORGANIZATION <chr>, PAGES <chr>,
## #   PUBLISHER <chr>, SCHOOL <chr>, SERIES <chr>, TITLE <chr>, TYPE <chr>,
## #   VOLUME <chr>, YEAR <dbl>, ISBN <chr>, DOI <chr>, URL <chr>, KEYWORDS <chr>,
## #   ABSTRACT <chr>, ISSN <chr>
```

```
scopus<-read_bib2tib(file = files[5])
scopus
```

```
## # A tibble: 210 x 18
##    sitenum value rawchar keybib typebib AUTHOR TITLE JOURNAL YEAR  VOLUME DOI
##      <int> <lis> <list>  <chr>  <chr>   <chr>  <chr> <chr>   <chr> <chr>  <chr>
##  1       1 <chr~ <chr [~ Albou~ ARTICLE Albou~ Dyna~ Expert~ 2020  162    10.1~
##  2       2 <chr~ <chr [~ Gonzá~ ARTICLE Gonzá~ A pr~ Inform~ 2020  64     10.1~
##  3       3 <chr~ <chr [~ Choi2~ ARTICLE Choi,~ Soci~ Inform~ 2020  57     10.1~
##  4       4 <chr~ <chr [~ Cho20~ ARTICLE Cho, ~ Inst~ Patter~ 2020  23     10.1~
```

```
##  5       5 <chr~ <chr [~ Guan2~ ARTICLE Guan,~ The ~ Intern~ 2020  34     10.1~
##  6       6 <chr~ <chr [~ Shukl~ ARTICLE Shukl~ A bi~ Engine~ 2020  92     10.1~
##  7       7 <chr~ <chr [~ Beggs~ ARTICLE Beggs~ A CU~ Sport,~ 2020  10     10.1~
##  8       8 <chr~ <chr [~ Amorn~ CONFER~ Amorn~ Cust~ 2020 I~ 2020  <NA>   10.1~
##  9       9 <chr~ <chr [~ Liu20~ ARTICLE Liu, ~ Micr~ Knowle~ 2020  62     10.1~
## 10      10 <chr~ <chr [~ Martí~ ARTICLE Martí~ A ma~ Europe~ 2020  281    10.1~
## # ... with 200 more rows, and 7 more variables: ART_NUMBER <chr>, NOTE <chr>,
## #   URL <chr>, DOCUMENT_TYPE <chr>, SOURCE <chr>, PAGES <chr>, NUMBER <chr>
```

```
springer<-bib2df(file = files[6])
springer
```

```
## # A tibble: 79 x 35
##    CATEGORY BIBTEXKEY ADDRESS ANNOTE AUTHOR BOOKTITLE CHAPTER CROSSREF EDITION
##    <chr>    <chr>     <chr>   <chr>  <list> <chr>     <chr>   <chr>    <chr>
##  1 INCOLLE~ kalita_f~ Singap~ <NA>   <chr ~ Recent {~ <NA>    <NA>     <NA>
##  2 INCOLLE~ patnaik_~ Singap~ <NA>   <chr ~ Advances~ <NA>    <NA>     <NA>
##  3 INCOLLE~ casillas~ Berlin~ <NA>   <chr ~ Manageme~ <NA>    <NA>     <NA>
##  4 INCOLLE~ campbell~ Cham    <NA>   <chr ~ Looking ~ <NA>    <NA>     <NA>
##  5 ARTICLE  sivasank~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
##  6 INCOLLE~ garcia-p~ Berlin~ <NA>   <chr ~ Trends i~ <NA>    <NA>     <NA>
##  7 INCOLLE~ washio_c~ Berlin~ <NA>   <chr ~ Advances~ <NA>    <NA>     <NA>
##  8 INCOLLE~ pant_rul~ Singap~ <NA>   <chr ~ Soft {Co~ <NA>    <NA>     <NA>
##  9 INCOLLE~ pedersen~ Berlin~ <NA>   <chr ~ Data {Wa~ <NA>    <NA>     <NA>
## 10 ARTICLE  ascarza_~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## # ... with 69 more rows, and 26 more variables: EDITOR <list>,
## #   HOWPUBLISHED <chr>, INSTITUTION <chr>, JOURNAL <chr>, KEY <chr>,
## #   MONTH <chr>, NOTE <chr>, NUMBER <chr>, ORGANIZATION <chr>, PAGES <chr>,
## #   PUBLISHER <chr>, SCHOOL <chr>, SERIES <chr>, TITLE <chr>, TYPE <chr>,
## #   VOLUME <chr>, YEAR <dbl>, ISBN <chr>, URL <chr>, URLDATE <chr>, DOI <chr>,
## #   LANGUAGE <chr>, FILE <chr>, SHORTTITLE <chr>, ISSN <chr>,
## #   COLLABORATOR <chr>
```

```
wos<-bib2df(file = files[7])
wos
```

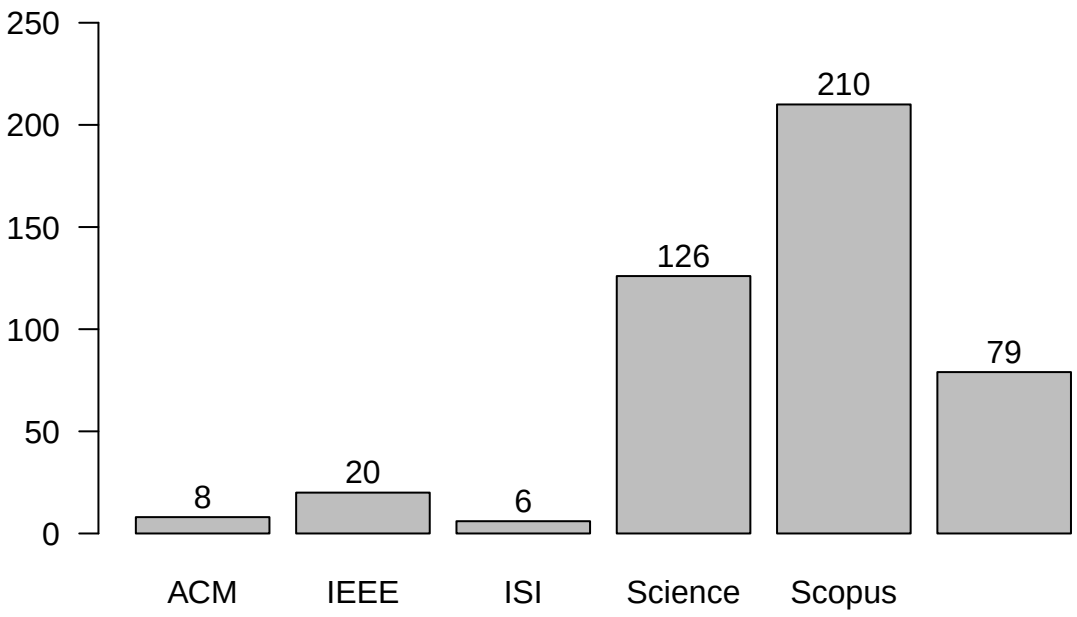
```
## # A tibble: 6 x 35
##   CATEGORY BIBTEXKEY ADDRESS ANNOTE AUTHOR BOOKTITLE CHAPTER CROSSREF EDITION
##   <chr>    <chr>     <chr>   <chr>  <list> <chr>     <chr>   <chr>    <chr>
## 1 ARTICLE  routh_es~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 2 ARTICLE  martinez~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 3 ARTICLE  brmez_ca~ <NA>    <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 4 BOOK     semrl_ch~ New Yo~ <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 5 BOOK     semrl_ch~ New Yo~ <NA>   <chr ~ <NA>      <NA>    <NA>     <NA>
## 6 INPROCE~ buckinx_~ <NA>    <NA>   <chr ~ Data min~ <NA>    <NA>     <NA>
## # ... with 26 more variables: EDITOR <list>, HOWPUBLISHED <chr>,
## #   INSTITUTION <chr>, JOURNAL <chr>, KEY <chr>, MONTH <chr>, NOTE <chr>,
## #   NUMBER <chr>, ORGANIZATION <chr>, PAGES <chr>, PUBLISHER <chr>,
## #   SCHOOL <chr>, SERIES <chr>, TITLE <chr>, TYPE <chr>, VOLUME <chr>,
## #   YEAR <dbl>, ISSN <chr>, URL <chr>, DOI <chr>, LANGUAGE <chr>,
## #   URLDATE <chr>, ABSTRACT <chr>, KEYWORDS <chr>, ISBN <chr>, FILE <chr>
```

Articles by year

```
library(xlsx)
library(tibble)
all_articles<-read.xlsx(file = 'articles_withoutDuplicates.xls',sheetIndex = 1)
all_articles<-as_tibble(all_articles)
#Create contigency table for source column
tb<-table(all_articles$source)
#Change names
names(tb)
```

```
## [1] "ACM Digital Library"  "IEEE Digital Library" "ISI Web of Science"
## [4] "Science@Direct"       "Scopus"               "SpringerLink"
```

```
names(tb)[1]='ACM'
names(tb)[2]='IEEE'
names(tb)[3]='ISI'
names(tb)[4]='Science'
# barplot with articles by source
bb<-barplot(height = tb,ylim = c(0,250),las=1)
text(x = bb,y = tb+10,labels = tb)
```



Remove duplicates

```
print(paste('Articles:',nrow(all_articles)))
```

```
## [1] "Articles: 449"
```

```r
print(unique(all_articles$status))
```

```
## [1] "Unclassified" "Duplicated"
```

```r
all_articles<-all_articles %>% filter(all_articles$status=="Unclassified")
print(paste('Duplicates removed:',nrow(all_articles)))
```

```
## [1] "Duplicates removed: 420"
```

Cleaning some references and preparing to ASReview

The requirements for ASReview: 1. title 2. abstract - author - date - keywords - doi

```r
dim(all_articles)
```

```
## [1] 420  26
```

```r
names(all_articles)
```

```
##  [1] "bibtex_key"         "title"              "author"
##  [4] "journal"            "year"               "source"
##  [7] "pages"              "volume"             "abstract"
## [10] "document_type"      "doi"                "url"
## [13] "affiliation"        "author_keywords"    "keywords"
## [16] "publisher"          "issn"               "language"
## [19] "note"               "selection_criteria" "created_at"
## [22] "updated_at"         "created_by"         "updated_by"
## [25] "status"             "comments"
```

```r
all_articles<-all_articles %>% select(title,abstract,author,year,keywords,doi,source)
summary(all_articles)
```

```
##     title             abstract           author             year
##  Length:420         Length:420         Length:420         Length:420
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##    keywords             doi               source
##  Length:420         Length:420         Length:420
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```r
# sapply with anonymous function
sapply(all_articles, function(x) sum(is.na(x)))
```

```
##    title abstract   author     year keywords      doi   source
##        1      391       15        0      289       13        0
```

Several values missing checking the articles with missing doi

```
articles_wdoi<-all_articles%>%filter(is.na(doi))
articles_wdoi$title
```

```
##  [1] "SIGMOD '15: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data"
##  [2] "A {Case} of {Churn} {Prediction} in {Telecommunications} {Industry}"
##  [3] "Using machine learning techniques to preduct defection of top clients"
##  [4] "Comparing and evaluating machine learning algorithms for predicting customer churn in telecommu
##  [5] "An experimental analysis of churn prediction techniques on real time datasets"
##  [6] "Research on grain storage temperature prediction model based on time series method"
##  [7] "Identifying attrition causing products using RFM analysis"
##  [8] "Feasibility of B2C customer relationship analytics in the B2B industrial context"
##  [9] "Maximize what matters: Predicting customer churn with decisioncentric ensemble selection"
## [10] "Review of data mining techniques for churn prediction in telecom"
## [11] "Agent based modelling and simulation: Toward a new model of customer retention in the mobile ma
## [12] "Mining the telecom marketing information to optimizing the customer retention strategies"
## [13] "Using machine learning techniques to predict defection of top clients"
```

Check if missing are duplicated by name

```
for (x in articles_wdoi$title){
    print(paste(x,'=',dim(filter(articles_wdoi,articles_wdoi$title==x))[1]))
}
```

```
## [1] "SIGMOD '15: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data =
## [1] "A {Case} of {Churn} {Prediction} in {Telecommunications} {Industry} = 1"
## [1] "Using machine learning techniques to preduct defection of top clients = 1"
## [1] "Comparing and evaluating machine learning algorithms for predicting customer churn in telecommu
## [1] "An experimental analysis of churn prediction techniques on real time datasets = 1"
## [1] "Research on grain storage temperature prediction model based on time series method = 1"
## [1] "Identifying attrition causing products using RFM analysis = 1"
## [1] "Feasibility of B2C customer relationship analytics in the B2B industrial context = 1"
## [1] "Maximize what matters: Predicting customer churn with decisioncentric ensemble selection = 1"
## [1] "Review of data mining techniques for churn prediction in telecom = 1"
## [1] "Agent based modelling and simulation: Toward a new model of customer retention in the mobile ma
## [1] "Mining the telecom marketing information to optimizing the customer retention strategies = 1"
## [1] "Using machine learning techniques to predict defection of top clients = 1"
```

Remove errors

Tasks

-SIGMOID -Remove special characters and search again -Try to identify DOI or remove the others