Häufigkeits-Analyse von Bild der Wissenschaft Artikel Themengebieten

Einleitung

Dies ist eine Analyse der Häufigkeitsverteilung von Themengebieten der Artikel in Bild der Wissenschaft (BdW). Ausgewertet wurden die Jahrgänge 2015, 2016 und 2017.

Motivation

Ich bin nun schon viele Jahre Leser, und im wesentlichen doch sehr zufrieden mit der Auswahl der Themen. Auch das die Artikel nicht nur die wissenschaftlichen Ergebnisse darstellen, sondern auch ihre gesellschaftliche Auswirkung diskutieren, gefällt mir sehr gut. Allerdings hat sich bei mir ein Bauchgefühl eingestellt, dass bestimmte Themengebiete (v.a. astronomische Themen) überdurchschnittlich häufig vorgestellt werden.

Methoden

Um diesen Eindruck zu überprüfen habe ich für die Jahrgänge 2015-17 von BdW die Artikel einezelnen Themengebieten zugeordnet. Themengebiete waren Technik, Biologie, Medizin, Physik, Astronomie, Psychologie, Soziologie etc. Die Zuordnung erfolgte manuell in eine OpenOffice Tabelle. Sie war manchmal nicht ganz eindeutig und wurde von mir nach besten Wissen und Gewissen durchgeführt. BdW führt einige Artikel als 'Titelthemen' oder 'Schwerpunktthemen' auf. Diese wurden mit einem boolschen Flag "Titelthema=TRUE" für eine zusätzlichen Auswertung gekennzeichnet. Die Themen-Einteilungen der BdW Redaktion habe ich hier ignoriert.

Nicht einbezogen wurden die Neuigkeiten, Artikel zu den Leserreisen und andere ständige Rubriken. Die Auswertungen erfolgten mit der Programmiersprache 'R', genauer mit Bibliotheken aus dem Tidyverse-Umfeld. An dieser Stelle vielen Dank an Hadley Wickham und den vielen Kontributoren zum Tidyverse. https://www.tidyverse.org Die multiplot Methode die hier verwendet wurde habe ich kopiert von folgender URL: http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/ Sie wurde freundlicherweise von Winston Chang der Allgemeinheit zur Verfügung gestellt.

```
#load libraries
library(tidyverse)
library(ggplot2)
library(utf8)

#http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/
source("multiplot.R")

#import csv
artikel_list <-  read_csv2("bdw_artikel.csv")
```

Rohdaten-Tabelle

```
head(artikel_list)
```

```
## # A tibble: 6 x 4
##   Ausgabe    Bereich                                Titel Titelthema
```

```
##     <chr>      <chr>                              <chr>      <lgl>
## 1 2017-01     Umwelt   Aluminium wie gefährlich ist es      FALSE
## 2 2017-01     Umwelt Aluminium ist kein wegwerfprodukt      FALSE
## 3 2017-01   Biologie              Die Wirkstofffahnder      FALSE
## 4 2017-01    Medizin               HIV unter Kontrolle      FALSE
## 5 2017-01 Astronomie                    Kosmos im Kopf       TRUE
## 6 2017-01 Astronomie                    Zündende Ideen       TRUE
```

Erste Auswertung

```
themen_list <- group_by(artikel_list,Bereich, Titelthema)
alle_themen <- summarise(themen_list, count = n()) %>% arrange(desc(count))
titelthemen <- summarise(themen_list, count = n()) %>% filter(Titelthema == TRUE) %>% arrange(desc(coun
```

Alle Artikel

```
alle_themen
```

```
## # A tibble: 35 x 3
## # Groups:   Bereich [20]
##        Bereich Titelthema count
##          <chr>      <lgl> <int>
##  1     Technik      FALSE    66
##  2  Astronomie      FALSE    54
##  3     Medizin      FALSE    41
##  4    Biologie      FALSE    40
##  5  Archälogie      FALSE    26
##  6   Allgemein      FALSE    25
##  7  Soziologie      FALSE    24
##  8  Astronomie       TRUE    21
##  9 Psychologie      FALSE    20
## 10     Technik       TRUE    20
## # ... with 25 more rows
```

Titel-Themen

```
titelthemen
```

```
## # A tibble: 16 x 3
## # Groups:   Bereich [16]
##          Bereich Titelthema count
##            <chr>      <lgl> <int>
##  1    Astronomie       TRUE    21
##  2       Technik       TRUE    20
##  3    Soziologie       TRUE    18
##  4        Physik       TRUE    16
##  5       Medizin       TRUE    14
##  6 Anthropologie       TRUE    11
##  7    Archälogie       TRUE     9
##  8      Biologie       TRUE     7
##  9   Psychologie       TRUE     6
```

```
## 10     Allgemein       TRUE     5
## 11         Klima       TRUE     4
## 12    Informatik       TRUE     2
## 13        Umwelt       TRUE     2
## 14        Chemie       TRUE     1
## 15      Geologie       TRUE     1
## 16   Philosophie       TRUE     1
```

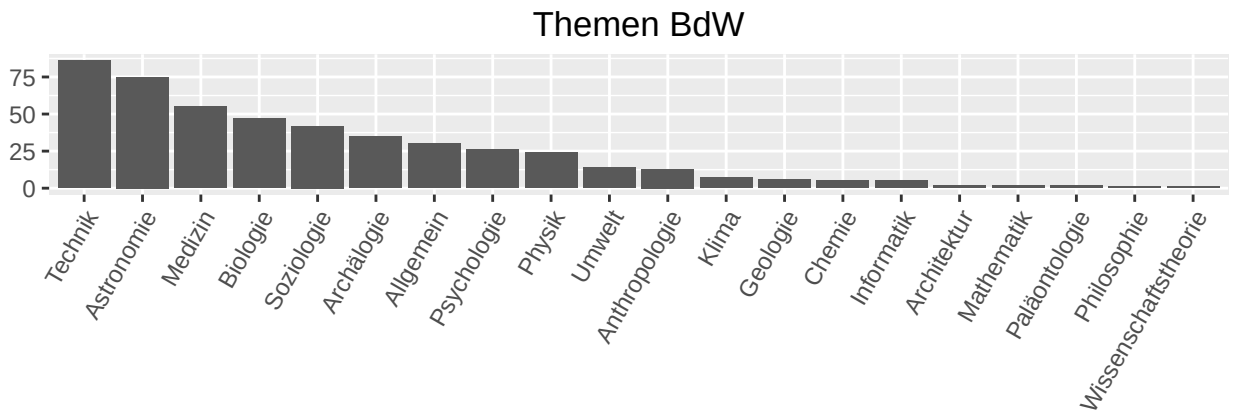
Plots

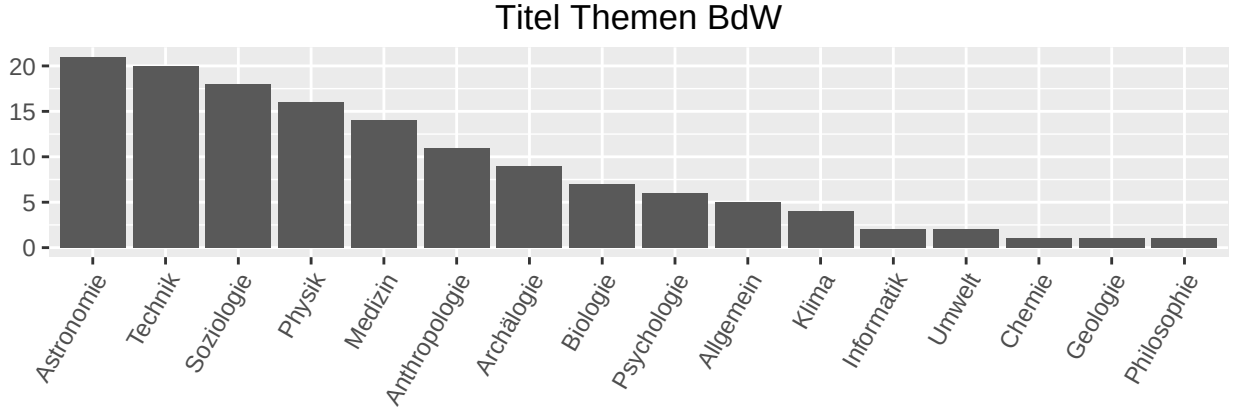
Alle Jahrgänge kombiniert

```
#for all
p1 <- artikel_list %>%
  count(Bereich) %>%
  mutate(Bereich = fct_reorder(Bereich, n, .desc = TRUE)) %>%
  ggplot(aes(x = Bereich, y = n)) + geom_bar(stat = 'identity') + theme(axis.text.x = element_text(angle

p2 <- artikel_list %>%
  filter(Titelthema == TRUE) %>%
  count(Bereich) %>%
  mutate(Bereich = fct_reorder(Bereich, n, .desc = TRUE)) %>%
  ggplot(aes(x = Bereich, y = n)) + geom_bar(stat = 'identity') + theme(axis.text.x = element_text(angle

multiplot(p1,p2, cols=1)
```



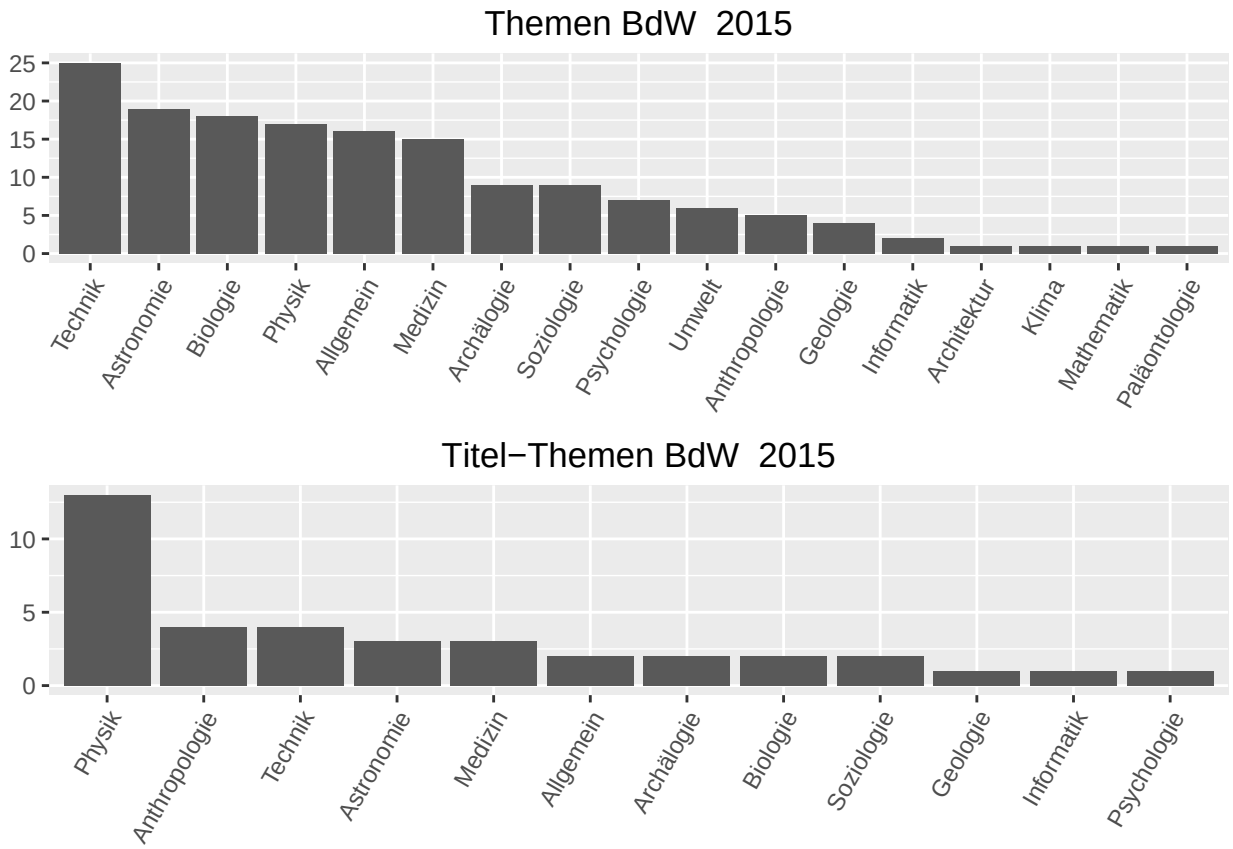


Jahr 2015

```
jahrgang <- "2015"
plottitel = paste('Themen BdW ',jahrgang, sep = " ")
plottitelthemen = paste('Titel-Themen BdW ',jahrgang, sep = " ")
p3 <- artikel_list %>%
  filter(str_detect(Ausgabe, jahrgang)) %>%
  count(Bereich) %>%
  mutate(Bereich = fct_reorder(Bereich, n, .desc = TRUE)) %>%
  ggplot(aes(x = Bereich, y = n)) + geom_bar(stat = 'identity') + theme(axis.text.x = element_text(angl

p4 <- artikel_list %>%
  filter(str_detect(Ausgabe, jahrgang)) %>%
  filter(Titelthema == TRUE) %>%
  count(Bereich) %>%
  mutate(Bereich = fct_reorder(Bereich, n, .desc = TRUE)) %>%
  ggplot(aes(x = Bereich, y = n)) + geom_bar(stat = 'identity') + theme(axis.text.x = element_text(angl

multiplot(p3,p4, cols=1)
```

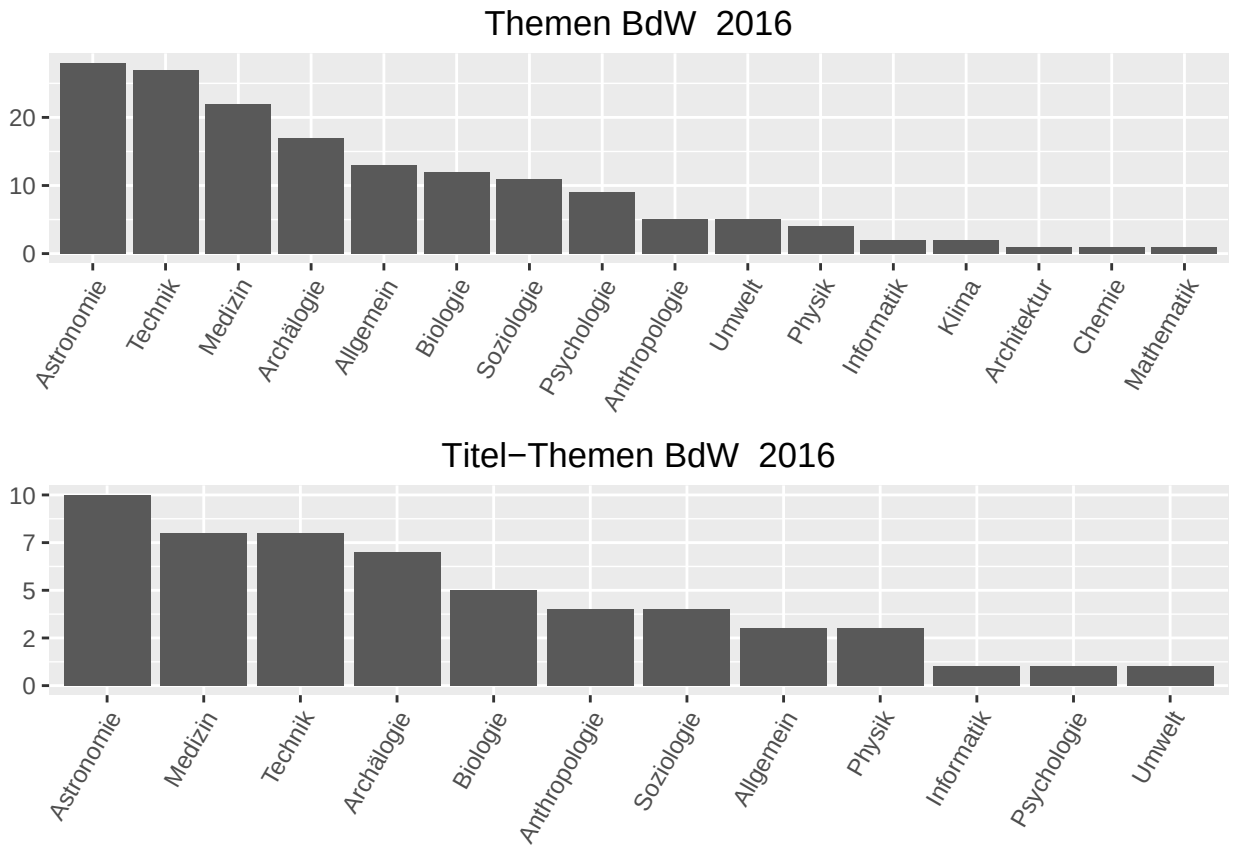


Jahr 2016

```
jahrgang <- "2016"
plottitel = paste('Themen BdW ',jahrgang, sep = " ")
p5 <- artikel_list %>%
```

```
  filter(str_detect(Ausgabe, jahrgang)) %>%
  count(Bereich) %>%
  mutate(Bereich = fct_reorder(Bereich, n, .desc = TRUE)) %>%
  ggplot(aes(x = Bereich, y = n)) + geom_bar(stat = 'identity') + theme(axis.text.x = element_text(angle

plottitelthemen = paste('Titel-Themen BdW ',jahrgang, sep = " ")
p6 <- artikel_list %>%
  filter(str_detect(Ausgabe, jahrgang)) %>%
  filter(Titelthema == TRUE) %>%
  count(Bereich) %>%
  mutate(Bereich = fct_reorder(Bereich, n, .desc = TRUE)) %>%
  ggplot(aes(x = Bereich, y = n)) + geom_bar(stat = 'identity') + theme(axis.text.x = element_text(angle
multiplot(p5,p6, cols=1)
```

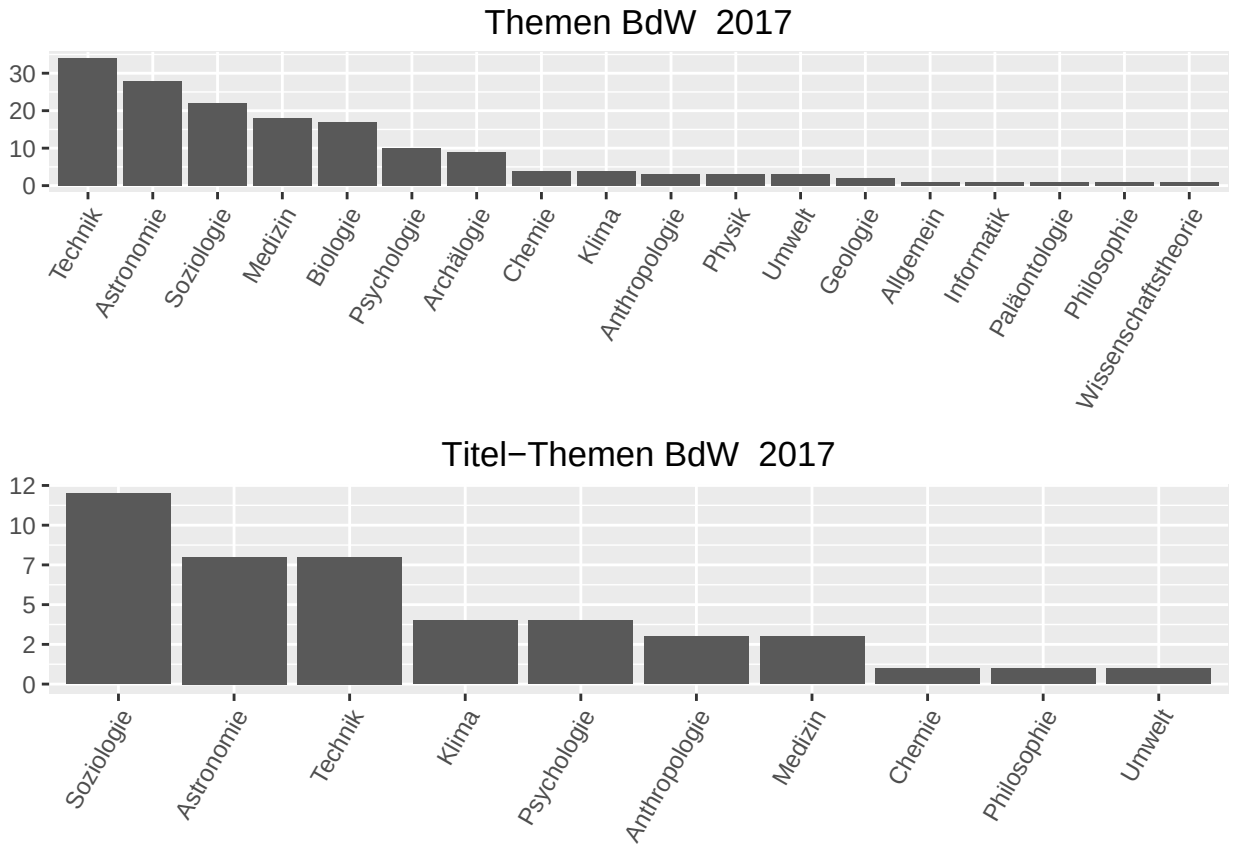


Jahr 2017

```
jahrgang <- "2017"
plottitel = paste('Themen BdW ',jahrgang, sep = " ")
p7 <- artikel_list %>%
  filter(str_detect(Ausgabe, jahrgang)) %>%
  count(Bereich) %>%
  mutate(Bereich = fct_reorder(Bereich, n, .desc = TRUE)) %>%
  ggplot(aes(x = Bereich, y = n)) + geom_bar(stat = 'identity') + theme(axis.text.x = element_text(angle

plottitelthemen = paste('Titel-Themen BdW ',jahrgang, sep = " ")
```

```
p8 <- artikel_list %>%
  filter(str_detect(Ausgabe, jahrgang)) %>%
  filter(Titelthema == TRUE) %>%
  count(Bereich) %>%
  mutate(Bereich = fct_reorder(Bereich, n, .desc = TRUE)) %>%
  ggplot(aes(x = Bereich, y = n)) + geom_bar(stat = 'identity') + theme(axis.text.x = element_text(angl
multiplot(p7,p8, cols=1)
```



Diskussion

- Das Bauchgefühl war nur teilweise richtig. Astronomie war insgesamt betrachtet das häufigste vertretene Themengebiet, aber 2015 war es Physik. Bei den Titelthemen waren in 2017 soziologische Themen die am häufigsten vorkamen und bei Artikeln insgesamt waren es technische Themen. Astronomie immer vorne dabei.
- Eine Erklärung für überdurchschnittliche Anzahl von Artikeln zu astronomischen Themen mögen auch Raumsonden wie Rosetta, Cassini und das Einsteinjahr 2015/16 sein.
- Themen wie Informatik scheinen unterrepräsentiert zu sein, doch habe ich Themn wie autonomes Fahren oder Robotik eher unter 'Technik' subsummiert.

Textanalyse

Zum Abschluss noch eine kleine Darstellung der Worthäufigkeiten in den Titeln der BdW Artikel, enjoy

```
library(tm)
library(ggplot2)
```

```r
library(SnowballC)
library(wordcloud)
library(RColorBrewer)

df_title <- data.frame(doc_id=row.names(artikel_list),
                       text=artikel_list$Titel)
mycorpus <- Corpus(DataframeSource(df_title))

mycorpus <- tm_map(mycorpus, removePunctuation)
mycorpus <- tm_map(mycorpus, content_transformer(tolower))
mycorpus <- tm_map(mycorpus, stripWhitespace)
mycorpus <- tm_map(mycorpus, removeWords, c(stopwords("german")))


tdm <- DocumentTermMatrix(mycorpus)
m <- as.matrix(tdm)
v <- sort(colSums(m),decreasing=TRUE)

words <- names(v)
d <- data.frame(word=words, freq=v)
#display.brewer.all()
pal2 <- brewer.pal(3,"Dark2")
wordcloud(d$word,d$freq, scale=c(3,.2), max.words=150, random.color=TRUE,colors = pal2)
```

```
## Warning in strwidth(words[i], cex = size[i], ...): Konvertierungsfehler für
## '-' in 'mbcsToSbcs': Punkt ersetzt <e2>
```

```
## Warning in strwidth(words[i], cex = size[i], ...): Konvertierungsfehler für
## '-' in 'mbcsToSbcs': Punkt ersetzt <80>
```

```
## Warning in strwidth(words[i], cex = size[i], ...): Konvertierungsfehler für
## '-' in 'mbcsToSbcs': Punkt ersetzt <93>
```

```
## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt
## = rotWord * : Konvertierungsfehler für '-' in 'mbcsToSbcs': Punkt ersetzt
## <e2>
```

```
## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt
## = rotWord * : Konvertierungsfehler für '-' in 'mbcsToSbcs': Punkt ersetzt
## <80>
```

```
## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt
## = rotWord * : Konvertierungsfehler für '-' in 'mbcsToSbcs': Punkt ersetzt
## <93>
```

```
## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : Fontmetrik ist für das Unicode-Zeichen U+2013 unbekannt
```


licht
schatz
leben
zukunft
eigentlich
welt
müll
mann
deutschland
großes
herz
warum
energie
angst
ganz
schön
tiere
große
jahre
millionen
neuer
zeit
löcher
...
strom
100
geburt
groß
einstein
geht
mensch
kreuz
früh
besser
kleine
kopf
immer
menschen
neue
roboter
maschinen
zwei
ende
gefahr
schwarzen
macht
kampf