

Wordcloud in R

Benjamin Schlegel

13. Juni 2016

Um eine Wordcloud zu erstellen, wird zuerst ein Text benötigt, von dem die Wordcloud erstellt werden soll. In diesem Beispiel wird dazu der Bericht des Bundesrates zur [Modernisierung des Familienrechts](#) verwendet.

Mithilfe von Word 2016 kann das PDF in ein Word Dokument verwandelt werden. Anschliessend kann der Text in einen Editor kopiert und als Text-Datei gespeichert werden. Zum Schluss werden § und Umbrüche entfernt, bevor der Text in R eingelesen wird.

```
data = read.csv("bericht_br_familie.txt", header=F, sep="§", encoding = "UTF-8")
text = as.character(data$V1)
```

Nun kann mit dem eingelesenen Text gearbeitet werden. Um eine Wordcloud zu erstellen werden die Packages [wordcloud](#), [tm](#) und [RColorBrewer](#) benötigt.

```
library(tm)
library(wordcloud)
library(RColorBrewer)
```

Zuerst wird nun ein Corpus erstellt. Anschliessend werden die Satzzeichen entfernt und alles in die Kleinschreibung umgewandelt. Zum Schluss werden häufige deutsche Wörter entfernt, die irrelevant sind, wie "der, die, das, etc."

```
ap.corpus = Corpus(DataframeSource(data.frame(text)))
ap.corpus = tm_map(ap.corpus, removePunctuation)
ap.corpus = tm_map(ap.corpus, content_transformer(tolower))
ap.corpus = tm_map(ap.corpus, function(x) removeWords(x,
c(stopwords("german"), "sowie", "dass", "bzw")))
```

Nun wird eine Häufigkeitsübersicht der Wörter erstellt. Die Wörter werden zum Schluss in Grossbuchstaben umgewandelt.

```
ap.tdm = TermDocumentMatrix(ap.corpus)
ap.m = as.matrix(ap.tdm)
ap.v = sort(rowSums(ap.m), decreasing=TRUE)
ap.d = data.frame(word = names(ap.v), freq=ap.v)
ap.d$word = toupper(ap.d$word)
```

Jetzt sind wir soweit und können die Wordcloud zeichnen. Wir speichern Sie in die Datei `wordcloud_family.png` mit 1000px Breite und 664px Höhe. Mit `max.words` wird angegeben, wie viele Wörter die Wolke enthalten soll.

```
pal2 = brewer.pal(8, "Dark2")
png("wordcloud_family.png", width=1000, height=664)
```

```
wordcloud(ap.d$word,ap.d$freq, scale=c(8,.2),min.freq=3,  
max.words=30, random.order=FALSE, rot.per=.15, colors=pal2)  
dev.off()
```

