Übungsblatt_2

Christian Kahmann

9 10 2020

Vorbereitung

Viele Funktionen, die R bietet, sind in sogenannte *Pakete* organisiert. Zum Beispiel werden für die Aufgabe 1 die Pakete `RCurl` und `XML` benötigt. Sie können mit folgendem Befehl installiert werden:

```
install.packages(c('RCurl', 'XML'))
```

Hinweis: Für die korrekte Installation von `RCurl` brauchen Sie weitere Dependenzen (z.B. `libcurl4-openssl-dev` auf einem Debian/Ubuntu System). Beachten Sie die Hinweise in der R-Konsole.

Textgewinnung aus RSS-Feeds

Aufgabe 1

Schreiben Sie die Funktion `parse.rss(url)`, die ein RSS-Feed herunterlädt, parst und als `data.frame` darstellt. Benutzen Sie die Funktion, um das RSS-feed der "Tagesschau" herunterzuladen.

Hinweis: Benutzen Sie die Funktion `getURL` aus dem Paket `RCurl`, um das Feed herunterzuladen, sowie `xmlParse` und `xmlToDataFrame` aus `XML`, um die Daten zu extrahieren.

```
#rss <- parse.rss("http://www.tagesschau.de/xml/rss2")
rss <- parse.rss("https://www.tagesschau.de/newsticker.rdf")
rss <- rss[grep("tagesschau.de", rss$link),]
names(rss)
```

```
## [1] "title"       "link"        "description" "guid"        "category"
```

```
head(rss$title, 5)
```

```
## [1] "Vor Corona-Gipfel mit Merkel: Was die Länder planen"
## [2] "Liveblog: ++ Mehr als 142.000 Neuinfektionen in den USA ++"
## [3] "Corona-Fälle in Pflegeheimen häufen sich"
## [4] "Impfstoff von AstraZeneca und Uni Oxford wirkt zu 70 Prozent"
## [5] "Spahn erwartet erste Corona-Impfungen noch in diesem Jahr"
```

```
head(rss$link, 5)
```

```
## [1] "https://www.tagesschau.de/inland/corona-plan-bundeslaender-101.html"
## [2] "https://www.tagesschau.de/newsticker/liveblog-coronavirus-montag-157.html"
## [3] "https://www.tagesschau.de/inland/corona-pflegeheime-103.html"
## [4] "https://www.tagesschau.de/ausland/impfstoff-astrazeneca-103.html"
## [5] "https://www.tagesschau.de/inland/coronavirus-spahn-impfungen-101.html"
```

Aufgabe 2

Schreiben Sie die Funktion `get.page(url)`, die eine HTML-Seite herunterlädt und den Text aus den `<p>`-Tags extrahiert.

Hinweis: Benutzen Sie die Funktion `htmlParse` aus dem Paket `XML` ähnlich, wie in der vorigen Aufgabe.

Korpus-Erstellung

Aufgabe 3

Wenden Sie die Funktionen von Aufgaben 1-2 an, um aus den heutigen Nachrichtenartikeln ein Korpus zu erstellen:

- extrahieren Sie die URLs der Artikel aus dem RSS-Feed,
- laden Sie die Artikeltexte herunter,
- erstellen Sie aus den heruntergeladenen Texten ein Objekt der Klasse `corpus` aus dem Paket `quanteda`.

Aufgabe 4

Wenden Sie die vom `quanteda`-Paket angebotenen Vorverarbeitungsschritte auf das Korpus an: `remove_punct`, `remove_numbers` und `remove_symbols`. Benutzen Sie die Funktion `tokens`, um die Operationen auf das ganze Korpus anzuwenden. Erstellen Sie aus dem Korpus eine Dokument-Term-Matrix.

Worthäufigkeiten

Aufgabe 5

Visualisieren Sie das Zipfsche Gesetz für das Korpus.

Hinweis: Benutzen Sie die Funktion `Zipf_plot` aus dem Paket `tm`.

Aufgabe 6

Visualisieren Sie eins von den Dokumenten als eine Wortwolke.

Hinweis: Benutzen Sie die Funktion `textplot_wordcloud` aus dem Paket `quanteda`.