Chapter3_exercises

Anna Wohlmann

2024-02-05

First, you need to install or load *quanteda*.

```
library(quanteda)
```

```
## Warning: Paket 'quanteda' wurde unter R Version 4.2.3 erstellt
```

```
## Package version: 3.3.1
## Unicode version: 13.0
## ICU version: 69.1
```

```
## Parallel computing: 12 of 12 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

1) Create a corpus from a character vector that consists of multiple texts; create the character vector yourself. Hint: You can link character vectors together into one character vector with c().

```
we_love_R <- c("R is fun to learn, and there is so much to do.")
we_love_data <- c("Loading datasets into R and exploring them is great.
                  We can learn so many new things about whatever interests us.")
we_love_QTA <- c("QTA is fun, too.
                 We can analyse so much text at once.
                 We can even create colorful plots!")
what_we_love <- c(we_love_R, we_love_data, we_love_QTA)
```

```
what_we_love_corpus <- corpus(what_we_love)
summary(what_we_love_corpus)
```

```
## Corpus consisting of 3 documents, showing 3 documents:
##
##   Text Types Tokens Sentences
##  text1    12     14         1
##  text2    21     22         2
##  text3    19     22         3
```

2) Go to the polidoc page (shiny.mzes.uni-mannheim.de/polidoc) and download a few party manifestos of your choice (as .txt), read them into R with the *readtext* package, and create a corpus. Now you can play with that data: add document-level variables, tokenize and create plots. Use the examples from chapters 2 and 3 as a guide.

First, you need to load or install the required packages for this exercise. You need *readtext* and *quanteda*.

```
library(readtext)
library(quanteda)
```

Go to the polidoc site, register, and choose countries, years, or parties you are interested in. I chose the national manifestos of the 2010 election in the Czech Republic. Download them into a folder in your R working directory; my subfolder is exercise3. Your data frame needs a name; I chose df_polidoc.

```
datadir <- "./data_exercises"
df_polidoc <- readtext(paste0(datadir,"/exercise3"), encoding="UTF8")
head(df_polidoc)
```

```
## readtext object consisting of 6 documents and 0 docvars.
## # Description: df [6 x 2]
##   doc_id                 text
##   <chr>                  <chr>
## 1 82110.000.2010.1.1.txt "\"Volební pr\"..."
## 2 82220.000.2010.1.1.txt "\"Otevřený v\"..."
## 3 82320.000.2010.1.1.txt "\"PROGRAM ZM\"..."
## 4 82413.000.2010.1.1.txt "\"Podrobný v\"..."
## 5 82414.000.2010.1.1.txt "\"VĚCI VEŘEJ\"..."
## 6 82523.000.2010.1.1.txt "\"KDU-ČSL\nto\"..."
```

The corpus() function applied to the data frame df_polidoc results in a corpus, I called corpus_polidoc.

```
corpus_polidoc <- corpus(df_polidoc)
summary(corpus_polidoc, n=5)
```

```
## Corpus consisting of 7 documents, showing 5 documents:
##
##                    Text Types Tokens Sentences
##  82110.000.2010.1.1.txt  9438  35374      1716
##  82220.000.2010.1.1.txt  2193   4526        34
##  82320.000.2010.1.1.txt  4542  12735       667
##  82413.000.2010.1.1.txt  6002  18875      1252
##  82414.000.2010.1.1.txt  5210  15254       411
```

3) Sign up for a Genius API (or another API of your choice). Check the terms of use to make sure what you want to do is legal. Store your genius token in R. Now load the lyrics to your favourite song in R.

If you choose the Genius API, go to this website: https://docs.genius.com/#/getting-started-h1, which tells you how to get started with the Genius API. First, we register and obtain a client ID, secret, and access token. Back in R, you need to load or install *geniusr*. To get more info on how to use the API we visit https://ewenme.github.io/geniusr/

```
library(geniusr)
```

You can save your access token by entering it manually after running this line:

Now, we wanted to check if loading the lyrics of a song is legal:

```
library(robotstxt)
```

```
paths_allowed("https://genius.com/Nura-fair-lyrics")
```

```
##  genius.com
```

```
## [1] TRUE
```

Additionally, check the documentation on their website.

The geniusr documentation tells us how to find a song. First, we need the song ID, which we get by searching for the song:

```
search_song("fair Nura", n_results = 5)
```

```
## # A tibble: 10 x 5
##     song_id song_name                                    song_~1 artis~2 artis~3
```

```
##       <int> <chr>                                          <chr>     <int> <chr>
##  1  7105453 Fair                                         https:~ 1037433 Nura
##  2 11233892 All In Love Is Fair                          https:~ 2866598 Nora A~
##  3   662385 Leeres Blatt                                 https:~   14316 Bizzy ~
##  4  3541820 Aufgeben (Ft. .fab (DEU))                    https:~   12202 Curse
##  5    93385 Süßholz (Balsam für meine Seele)             https:~   12202 Curse
##  6  6644647 JE T'AIME                                    https:~   20441 Moe Ph~
##  7  6213021 Herşey Yalan                                 https:~   17856 Alpa G~
##  8   111089 Hot Sun, Cool Fire                           https:~   35623 George~
##  9  4254678 Ondi Vil - Please Come Back ft. Thomas Reid~ https:~ 1799569 Genius~
## 10  9250942 Arachnophilic archangel (Ft. Kidd Klaviu & ~ https:~ 3346633 Zyrexo~
## # ... with abbreviated variable names 1: song_lyrics_url, 2: artist_id,
## #   3: artist_name
```

The first result is my favorite song, so we can now access the lyrics through the ID.

This is a data frame with information about the song and the album it is on:

```
get_song_df("7105453")
```

```
## # A tibble: 1 x 13
##   song_id song_name song_lyric~1 song_~2 song_~3 song_~4 song_~5 artis~6 artis~7
##     <int> <chr>     <chr>        <chr>   <chr>     <int>   <int>   <int> <chr>
## 1 7105453 Fair      https://gen~ https:~ 2021-0~   49317       7 1037433 Nura
## # ... with 4 more variables: artist_url <chr>, album_id <int>,
## #   album_name <chr>, album_url <chr>, and abbreviated variable names
## #   1: song_lyrics_url, 2: song_art_image_url, 3: song_release_date,
## #   4: song_pageviews, 5: song_annotation_count, 6: artist_id, 7: artist_name
```

And here is the lyrics:

```
get_lyrics_id(song_id = "7105453")
```

Alternatively, you could get the lyrics through the song URL as well:

```
get_lyrics_url("https://genius.com/Nura-fair-lyrics")
```