Automatic Sampling and Analysis of YouTube Data

Excursus: Retrieving Video Subtitles

Julian Kohne
Johannes Breuer
M. Rohangis Mohseni

2020-02-11

Retrieving Video Subtitles (1)

- Instead of transscribing a video, you can retrieve its subtitles.

- What research would you conduct with video subtitles?

Retrieving Video Subtitles (2)

- Retrieve a list of subtitles with

    - `tuber::list_caption_tracks` (costs around 52 quota)

- Videos with automatically created subtitle (*ASR*)

    - Always in English, even if video language is non-English

    - Can be downloaded, but text quality can be bad (especially if translated)

- Videos without any subtitles

    - Not sure if even possible because there always seems to be an *ASR*

- Videos with more than one set of subtitles

    - Examples: *ASR* and regular subtitle, more than one language, more than one subtitle for the same language

    - Can be downloaded, but subtitle for analysis must be selected

- Videos with deactivated 3rd party contribution return Error 403

Retrieving Video Subtitles (3)

- First, we need to get the list of subtitles for a video

```
caption_list <- list_caption_tracks(part = "snippet",
video_id = "nI_0fkQOG6Q")
```

- Next, we need to get the ID of the subtitle and convert it to a string

```
ID <- toString(caption_list[1,"id"])
```

- Adapt the number to select the subtitle that you want (ASR = automatic sub)

- After that, we need to retrieve the subtitles and convert them from raw to char

```
text <- rawToChar(get_captions(id = ID, format = "sbv"))
```

- Now we can save the subtitles to a subtitle file

```
write(Text, file = "Captions.sbv", sep="\n")
```

Converting Subtitles

- Subtitles come in a special format called SBV

- The format contains time stamps etc. that we do not need for text analysis

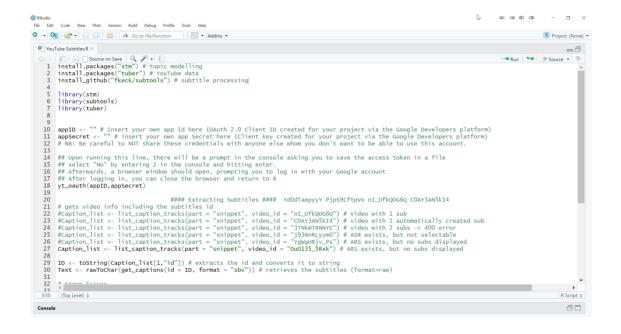- We can read the format with the package `subtools`

  ```
  subs <- read_subtitles("Captions.sbv", format =
  "subviewer")
  ```

- With subtools, we can also retrieve the text from the subtitles

  ```
  subtext <- get_raw_text(Subs)
  ```

- Now the text is ready for text analysis

Time for a Live Demonstration

Any questions?