Automatic Sampling and Analysis of YouTube Data

Processing and Cleaning User Comments

Julian Kohne
Johannes Breuer
M. Rohangis Mohseni

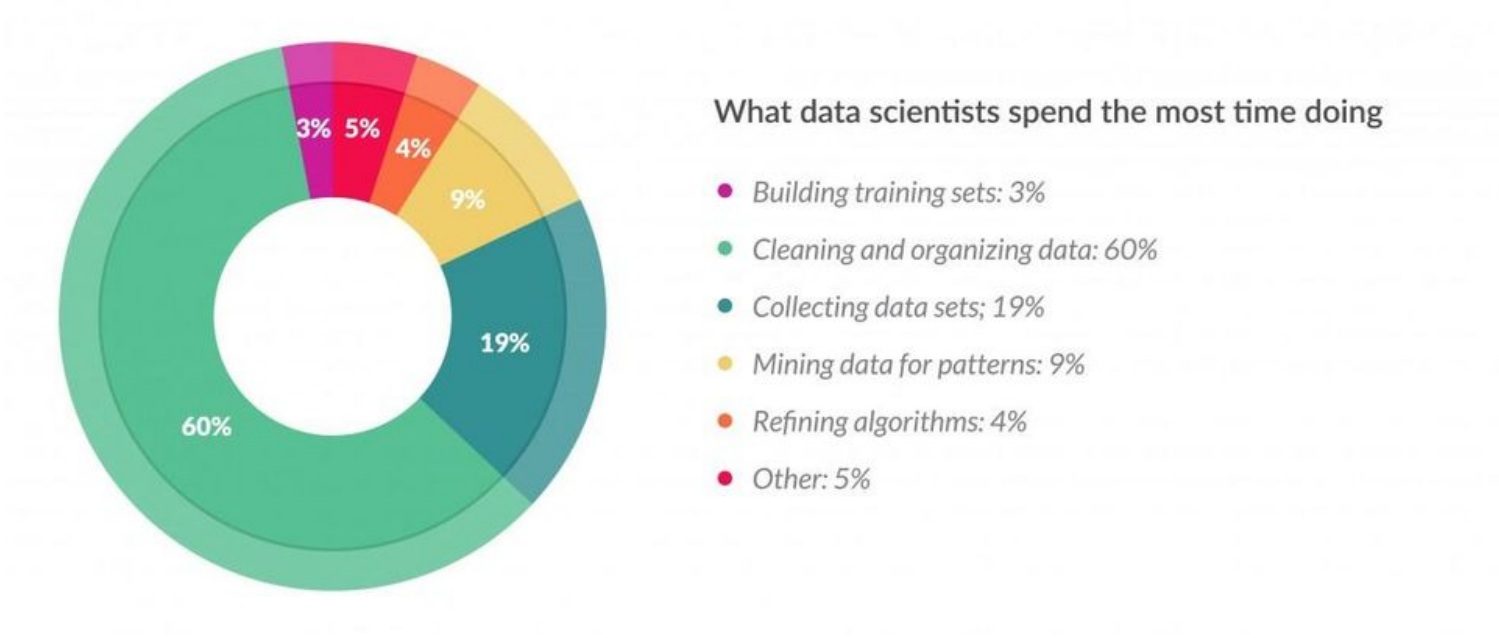
2022-02-21

Preprocessing

- Preprocessing refers to all steps that need to be taken to make the data suitable for the actual analysis
- For webscraping data, this is often more tedious and time-consuming than for survey data because:
    - the data are not designed with your analysis in mind
    - the data are typically less structured
    - the data are typically more complex
    - the data are typically more heterogenous
    - the data are typically larger
- *Note*: In addition, with large amounts of data it is often necessary to work on servers or clusters instead of regular desktop or laptop computers
    - Even then, restructuring or transforming data can take days, so mistakes hurt more

Preprocessing

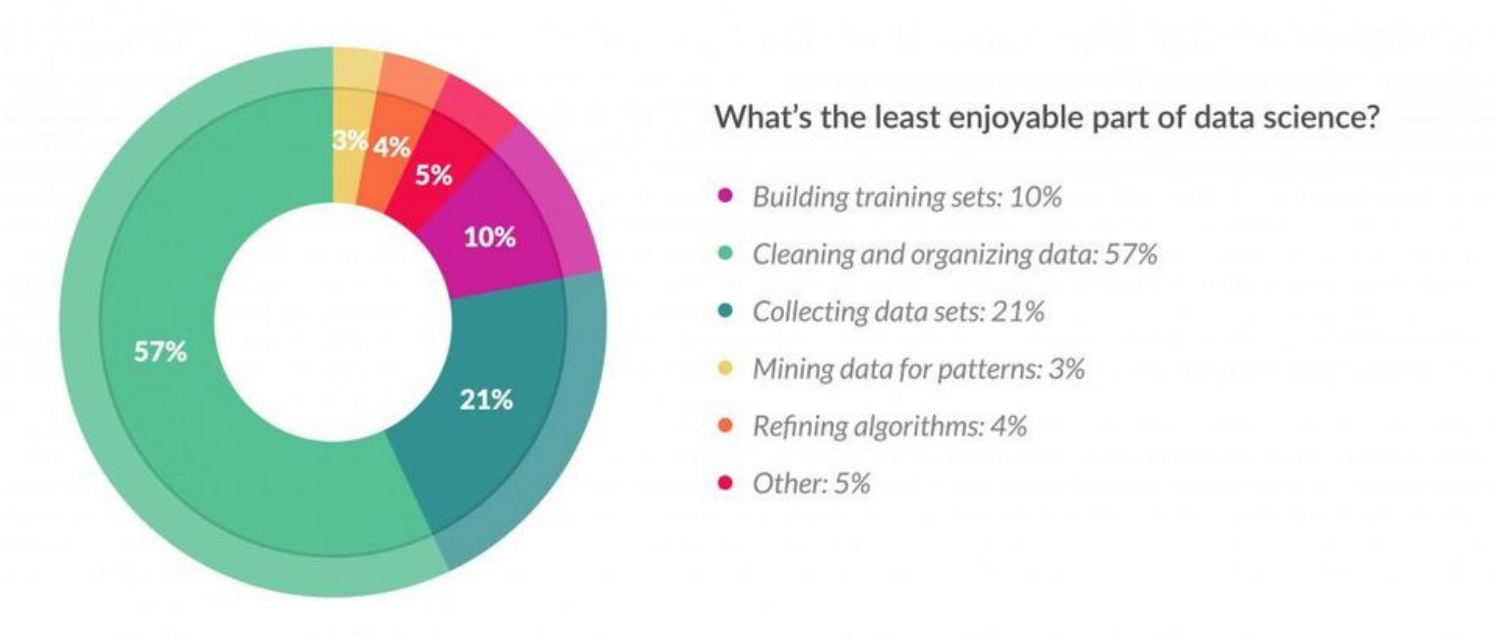
- In *Data Science*, most time is typically spent on the preprocessing rather than the actual analysis



Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#157890a96f63

Preprocessing

- Also, it is perceived as the least enjoyable part of the process



Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#157890a96f63

Preprocessing *YouTube* comments

- The `tuber` package returns an `R` dataframe instead of a JSON
- We can select which data we need by using the API through `tuber`
- For single videos, the data are small enough to be processed on a regular desktop/laptop computer
- However, this doesn't mean that the data are already usable for all intents and purposes
- We still need to:
    - select
    - format
    - extract
    - link

the information that is relevant to us

Preprocessing *YouTube* Comments

For this session, we will use comments from the Emoji Movie Trailer (https://www.youtube.com/watch?v=r8pJt4dK_s4)

Understanding Your Data (1)

The first step is always to explore your data. This is especially crucial for so-called *found data* because they were not designed with your analysis in mind.

```
# load raw data
comments <- readRDS("../../data/RawEmojiComments.rds")

# list all column names
colnames(comments)
```

```
##  [1] "videoId"               "textDisplay"           "textOriginal"
##  [4] "authorDisplayName"     "authorProfileImageUrl" "authorChannelUrl"
##  [7] "authorChannelId.value" "canRate"               "viewerRating"
## [10] "likeCount"             "publishedAt"           "updatedAt"
## [13] "id"                    "parentId"              "moderationStatus"
```

Luckily, the *YouTube* API is very well documented and provides brief explanations for all the variables you can extract from it

Understanding Your Data (2)

This information is valuable for understanding what type of comments the dataframe contains

```
table(is.na(comments$parentId))
```

```
##
## FALSE  TRUE
## 15734 22600
```

A quick look at the documentation reveals:

parentID: *The unique ID of the parent comment. This property is only set if the comment was submitted as a reply to another comment.*

Understanding Your Data (3)

...or for knowing how specific data types are formatted

```
head(comments$publishedAt)
```

```
## [1] "2022-02-10T06:38:33Z" "2022-02-08T04:05:05Z" "2022-02-06T16:43:18Z"
## [4] "2022-02-06T12:42:39Z" "2022-02-06T01:10:24Z" "2022-02-05T23:23:26Z"
```

```
class(comments$publishedAt)
```

```
## [1] "character"
```

A quick look at the documentation reveals:

publishedAt: *The date and time when the comment was originally published. The value is specified in ISO 8601 (YYYY-MM-DDThh:mm:ss.sZ) format.*

Understanding Your Data (4)

...or how similarly named variables differ from each other

```
comments$textOriginal[6]
```

```
## [1] "The best part 2:38"
```

```
comments$textDisplay[6]
```

```
## [1] "The best part <a href=\"https://www.youtube.com/watch?"
## [2] "v=r8pJt4dK_s4&amp;t=2m38s\">2:38</a>"
```

textOriginal: *The original, raw text of the comment as it was initially posted or last updated. The original text is only returned if it is accessible to the authenticated user, which is only guaranteed if the user is the comment's author.*

textDisplay: *The comment's text. The text can be retrieved in either plain text or HTML. (The comments.list and commentThreads.list methods both support a textFormat parameter, which specifies the desired text format). Note that even the plain text may differ from the original comment text. For example, it may replace video links with video titles.*

Selecting What You (Don't) Need

Now we can decide on what we need for our analysis

```
Selection <- subset(comments,select = -c(authorProfileImageUrl,
                                         authorChannelUrl,
                                         authorChannelId.value,
                                         videoId,
                                         canRate,
                                         viewerRating,
                                         moderationStatus))
colnames(Selection)
```

```
## [1] "textDisplay"       "textOriginal"      "authorDisplayName"
## [4] "likeCount"         "publishedAt"       "updatedAt"
## [7] "id"                "parentId"
```

Word of advice: Always keep an unaltered copy of your raw data and don't overwrite it. You never know what kinds of mistakes/oversights you might notice down the line and you don't want to have to recollect everything. Save your parsed data in a separate file (or in multiple steps and versions if your preprocessing pipeline is complex).

Formatting your Data

By default, the data you get out of `tuber` is most likely not in the right format for your analyses

```
sapply(Selection, class)
```

```
##       textDisplay      textOriginal authorDisplayName         likeCount
##       "character"       "character"       "character"       "character"
##       publishedAt         updatedAt                id          parentId
##       "character"       "character"       "character"       "character"
```

```
# summary statistics for like counts
summary(Selection$likeCount)
```

```
##    Length     Class      Mode
##     38334 character character
```

```
# time difference between first comment and now
Sys.time() - Selection$publishedAt[1]
```

```
## Error in unclass(e1) - e2: non-numeric argument to binary operator
```

Formatting the `likeCount`

We want the `likeCount` to be a numeric variable and the timestamps to be datetime objects

```
# transform likeCount to numeric
# (NB: this overwrites the original column)
Selection$likeCount <- as.numeric(Selection$likeCount)

# check
summary(Selection$likeCount)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    2.000    4.829    5.000 4344.000
```

We can now work with the number of likes as a numeric variable

Formatting Timestamps (1)

Timestamps are extremely complex objects due to:

- Different calendars
- Different formattings
- Different origins
- Different time zones
- Historical anomalies
- Different resolutions
- Summer vs. Wintertime (different for each country and depending on hemisphere!)
- Leap years
- etc.

For these reasons, you should **never** try to code your own time stamp translations from scratch. Fortunately, `R` has several build in methods for dealing with this madness. The most basic one is the `as.POSIXct()` function, the most convenient one is the `anytime()` function from the `anytime` package (another powerful option for dealing with times and dates in `R` is the `lubridate` package from the `Tidyverse`).

Formatting Timestamps (2)

```
# transform timestamps to datetime objects
Selection$publishedAt[1]
```

```
## [1] "2022-02-10T06:38:33Z"
```

```
testtime <- as.POSIXct(Selection$publishedAt[1],
                       format = "%Y-%m-%dT%H:%M:%OSZ",
                       tz = "UTC")
testtime
```

```
## [1] "2022-02-10 06:38:33 UTC"
```

```
# test whether we can compute a difference
# with the datetime object
Sys.time() - testtime
```

```
## Time difference of 10.65124 days
```

This internal representation of time objects will be extremely important for plotting trends over time and calculating time differences. You can find an overview of formatting date strings here.

Formatting Timestamps (3)

A more convenient way of transforming datetime variables is the `anytime` package. It automatically tries to guess the format from the character string, so you don't have to. This is especially handy for vectors of datetimes in multiple formats.

```
# transform datetimes using anytime()
library(anytime)
Selection$publishedAt <- anytime(Selection$publishedAt,
                                 asUTC = TRUE)
Selection$updatedAt <- anytime(Selection$updatedAt,
                               asUTC = TRUE)
sapply(list(Selection$publishedAt,Selection$updatedAt),class)
```

```
##      [,1]      [,2]
## [1,] "POSIXct" "POSIXct"
## [2,] "POSIXt"  "POSIXt"
```

Word of Advice: For datetime conversions, always do some sanity checks, especially if you are using methods that automatically detect the format. Pay special attention to the *timezone* in which your data are saved and compare it to the documentation of the standard.

Formatting Timestamps (4)

Be aware of how to interpret your timestamps. Note that the date was interpreted as UTC but converted to our local CET timezone which is 1 hour ahead of UTC. This comment was made at 07:38:33 in *our timezone*, but we have no idea about the time at the location of the user.

```
Selection$publishedAt[1]
```

```
## [1] "2022-02-10 07:38:33 CET"
```

Extracting Information

After having formatted all our selected columns, we usually also want to create some new columns with information that is not directly available in the raw data. For example, consider these comments:

```
# Example comments with extractable information
strwrap(Selection$textOriginal[37445],79)
```

```
## [1] "Watch new Emoji movie [2017] Here: New Emoji movie 2017"
## [2] "https://www.clorox.com/"
```

```
Selection$textOriginal[26]
```

```
## [1] "Here him 🙄🙄🙄🙄🙄🙄"
```

There are two issues exemplified by these comments:

1) Comments contain emojis and hyperlinks that might distort our text analysis later

2) These are features that we'd like to have in a separate column for our analysis

Extracting Hyperlinks (1)

We will start with deleting hyperlinks from our text and saving them in an additional column. We will use the text mining package qdapRegex for this as it has predefined routines for handling large text vectors and regular expressions.

```
# Note that we are using the original text so we don't have
# to deal with the HTML formatting of the links
library(qdapRegex)
Links <- rm_url(Selection$textOriginal, extract = TRUE)
LinkDel <- rm_url(Selection$textOriginal)
head(Links[!is.na(Links)],3)
```

```
## [[1]]
## [1] "https://youtu.be/59Tr9NDr5N4\nI"
##
## [[2]]
## [1] "https://youtu.be/SgX3ggJv1Rw"
##
## [[3]]
## [1] "https://m.youtube.com/watch?v=BLUkgRAy_Vo"
```

Extracting Hyperlinks (2)

We get back a list where each element corresponds to one row in the Selection dataframe and contains a vector of links that were contained in the textOriginal column. At the same time, the link was removed from the Selection dataframe.

```
strwrap(Selection$textOriginal[37445],79)
```

```
## [1] "Watch new Emoji movie [2017] Here: New Emoji movie 2017"
## [2] "https://www.clorox.com/"
```

```
LinkDel[37445]
```

```
## [1] "Watch new Emoji movie [2017] Here: New Emoji movie 2017"
```

```
Links[[37445]]
```

```
## [1] "https://www.clorox.com/"
```

Extracting Emojis (1)

The `qdapRegex` package has a lot of other different predefined functions for extracting or removing certain kinds of strings:

- `rm_citation()`
- `rm_date()`
- `rm_phone()`
- `rm_postal_code()`
- `rm_email()`
- `rm_dollar()`
- `rm_emoticon()`

Unfortunately, it does **not** contain a predefined method for emojis, so we will have to use the `emo` package for removing the emojis and come up with our own method for extracting them.

Extracting Emojis (2)

First we want to replace the emojis with a textual description, so that we can treat it just like any other token in text mining. This is no trivial task, as we have to go through each comment and replace each emoji with its respective textual description. Unfortunately, we did not find a working, easy-to-use out-of-the-box solution for this. But we can always make our own!

Essentially, we want to replace this:

```
## 😄
```

with this

```
## [1] "EMOJI_GrinningFaceWithSmilingEyes"
```

Extracting Emojis (3)

First of all, we need a dataframe that contains the emojis as they are internally represented by R (this means dealing with character encoding which can be quite the hassle). Luckily, this information is contained in the emo package.

```
library(emo)
EmojiList <- jis
EmojiList[1:3,c(1,3,4)]
```

```
## # A tibble: 3 × 3
##   runes emoji name
##   <chr> <chr> <chr>
## 1 1F600 😀    grinning face
## 2 1F601 😁    beaming face with smiling eyes
## 3 1F602 😂    face with tears of joy
```

Extracting Emojis (4)

Next, we need to paste the names of the emojis together while capitalizing the first letter of every word for better readability

```
# Define a function for capitalizing and pasting names together
simpleCap <- function(x) {

  # Split the string
  splitted <- strsplit(x, " ")[[1]]

  # Paste it back together with capital letters
  paste(toupper(substring(splitted, 1,1)),
        substring(splitted, 2),
        sep = "",
        collapse = " ")
}
```

Extracting Emojis (5)

```
# Apply the function to all the names
CamelCaseEmojis <- lapply(jis$name, simpleCap)
CollapsedEmojis <- lapply(CamelCaseEmojis,
                          function(x){gsub(" ",
                                           "",
                                           x,
                                           fixed = TRUE)})
EmojiList[,4] <- unlist(CollapsedEmojis)
EmojiList[1:3,c(1,3,4)]
```

```
## # A tibble: 3 × 3
##   runes emoji name
##   <chr> <chr> <chr>
## 1 1F600 😀    GrinningFace
## 2 1F601 😁    BeamingFaceWithSmilingEyes
## 3 1F602 😂    FaceWithTearsOfJoy
```

Extracting Emojis (6)

After that, we need to order our dictionary from the longest to shortest string, so that we can prevent partial matching of shorter strings later.

```
EmojiList <- EmojiList[rev(order(nchar(jis$emoji))),]
head(EmojiList[,c(1,3,4)],5)
```

```
## # A tibble: 5 × 3
##   runes                                          emoji name
##   <chr>                                          <chr> <chr>
## 1 1F469 200D 2764 FE0F 200D 1F48B 200D 1F469 👩‍❤️‍💋‍👩    Kiss:Woman,Woman
## 2 1F468 200D 2764 FE0F 200D 1F48B 200D 1F468 👨‍❤️‍💋‍👨    Kiss:Man,Man
## 3 1F469 200D 2764 FE0F 200D 1F48B 200D 1F468 👩‍❤️‍💋‍👨    Kiss:Woman,Man
## 4 1F3F4 E0067 E0062 E0077 E006C E0073 E007F  🏴    Wales
## 5 1F3F4 E0067 E0062 E0073 E0063 E0074 E007F  🏴    Scotland
```

Note that what we are ordering by the `emoji` column, not the `text` or `runes` columns.

Extracting Emojis (7)

Now we can `loop` through all of our emojis and replace them consecutively in each comment (*note*: this may take a while)

```
# Assign the column to a an object named TextEmoRep
TextEmoRep <- LinkDel

# Loop over all emojis for all comments in LinkDel
for (i in 1:dim(EmojiList)[1]) {

  TextEmoRep <- rm_default(TextEmoRep,
                  pattern = EmojiList[i,3],
                  replacement = paste0("EMOJI_",
                                       EmojiList[i,4],
                                       " "),
                  fixed = TRUE,
                  clean = FALSE,
                  trim = FALSE)
}
```

Extracting Emojis (8)

As output, we get a large character vector with emojis replaced by textual descriptions.

```
Selection$textOriginal[233]
```

```
## [1] "Current like to dislike ratio:\n👍47K  👎167K"
```

```
TextEmoRep[233]
```

```
## [1] "Current like to dislike ratio"
## [2] " EMOJI_ThumbsUp 47K EMOJI_ThumbsDown 167K"
```

Extracting Emojis Function

```
ExtractEmoji <- function(x){

  SpacerInsert <- gsub(" ","[{[SpAC0R]}]", x)
  ExtractEmoji <- rm_between(SpacerInsert,
                             "EMOJI_","[{[SpAC0R]}]",
                             fixed = TRUE,
                             extract = TRUE,
                             clean = FALSE,
                             trim = FALSE,
                             include.markers = TRUE)

  UnlistEmoji <- unlist(ExtractEmoji)
  DeleteSpacer <- sapply(UnlistEmoji,
                         function(x){gsub("[{[SpAC0R]}]",
                                          " ",
                                          x,
                                          fixed = TRUE)})

  names(DeleteSpacer) <- NULL
  Emoji <- paste0(DeleteSpacer, collapse = "")
  return(Emoji)
}
```

Extracting Emojis Function

We can apply the function to get one vector containing only the emojis as textual descriptions.

```
Emoji <- sapply(TextEmoRep,ExtractEmoji)
names(Emoji) <- NULL
LinkDel[233]
```

```
## [1] "Current like to dislike ratio: 👍47K 👎167K"
```

```
Emoji[233]
```

```
## [1] "EMOJI_ThumbsUp EMOJI_ThumbsDown "
```

Removing Emojis

In addition, we remove the emojis from our `LinkDel` variable to have one *clean* column that we can use for text mining later. This column will not contain hyperlinks or emojis.

```
# We take the LinkDel column and also delete the emojis from it
library(emo)
LinkDel[233]
```

```
## [1] "Current like to dislike ratio: 👍47K 👎167K"
```

```
TextEmoDel <- ji_replace_all(LinkDel,"")
TextEmoDel[233]
```

```
## [1] "Current like to dislike ratio: 47K 167K"
```

Summary: Extracting Information

We now have different versions of our text column

1) The original one, with hyperlinks and emojis (`Selection$textOriginal`)

2) One with only plain text and without hyperlinks and emojis (`TextEmoDel`)

3) One with only hyperlinks (`Links`)

4) One with only emojis (`Emoji`)

We want to integrate all of them into our dataframe.

Linking Everything Back Together

We can now combine our dataframe with the additional columns we created to have the perfect starting point for our analysis! However, because we sometimes have more than two links or two emojis per comment, we need to use the `I()` function so we can put them in the dataframe `as is`. Later, we will have to unlist these columns rowwise if we want to use them.

```
df <- cbind.data.frame(Selection$authorDisplayName,
                       Selection$textOriginal,
                       TextEmoRep,
                       TextEmoDel,
                       Emoji = I(Emoji),
                       Selection$likeCount,
                       Links = I(Links),
                       Selection$publishedAt,
                       Selection$updatedAt,
                       Selection$parentId,
                       Selection$id,
                       stringsAsFactors = FALSE)
```

Linking Everything Back Together

As a final step, we can give the columns appropriate names and save the dataframe for later use

```
# set column names
names(df) <- c("Author",
               "Text",
               "TextEmojiReplaced",
               "TextEmojiDeleted",
               "Emoji",
               "LikeCount",
               "URL",
               "Published",
               "Updated",
               "ParentId",
               "CommentID")

saveRDS(df, file = "../../data/ParsedEmojiComments.rds")
```

Exercise time 🏋️ 💪 🏃 🚴

Solutions