NPFL142.C4DHI – tutorial #2

Word frequency with *Migrant stories* and probabilities with *Titanic* data

---

Exercises with *Migrant stories* dataset

Exercise 1.1 – Loading the data set

In RStudio create a blank R script

- Move to Files&Plots desktop
- In Files manager
  - move to the home directory
  - create a new directory New folder > 2
  - move to the directory `2` and create a new directory New folder > `migrants`
  - move to the directory `2/migrants`
  - use More in the menu to run Set as working directory
  - use New Blank File in the menu to create a blank R Script and name it `migrants.t.R`. Then the script is open in the Code editor window (upper-left window) and you can add the `commands` listed below to the script.

We suppose using these packages

```
library(tidyverse)
library(tidytext)
library(stringr)
```

Load the *Migrant stories* dataset into your R environment.

```
dataset <- read_tsv("dataset <- read_tsv("../../1/migrants/migrants.tsv")
names(dataset)   # attributes
```

Exercise 1.2 – Visualize a frequency dictionary of men who moved to Switzerland

Draw a barplot displaying a word frequency. Focus on the words that occur in the stories of men who moved to Switzerland. Display the words with frequency at least 10. The `ggplot2` library (part of `tidyverse` library) is a powerful and popular data visualization package in R

```
dataset %>%
  filter(gender == "male")  %>%              # males only
  filter(country_de == "Switzerland") %>%    # Switzerland as a destination country only
  unnest_tokens(word, "story", to_lower = TRUE)  %>% tokenization,convert tokens to lowercase
  anti_join(stop_words) %>%                  # exclude stop words
  count(word)   %>%                           # count word frequencies
  filter(n > 10) %>%                         # filter out the words used more than 10 times
  mutate(word = reorder(word, n)) %>%        # sort by n
  ggplot(aes(word, n)) +                     # draw a bar plot
  geom_col() +
  xlab("Token") + ylab("Frequency 10+") + coord_flip() +
  ggtitle("Frequency dictionary of men who moved to Switzerland")
```

Save the plot to a file

```
 ggsave(file = 'switzerland.png', height = 5, width = 5)
```

Exercises with *Titanic* dataset

Exercise 2.1 – Loading the data set

In RStudio create a blank R script

- Move to Files&Plots desktop
- In Files manager
    - move to the directory `2` and create a new directory New folder > `titanic`
    - move to the directory `2/titanic`
    - use More in the menu to run Set as working directory
    - use New Blank File in the menu to create a blank R Script and name it `titanic.t.R`. Then the script is open in the Code editor window (upper-left window) and you can add the `commands` listed below to the script.

We suppose using these packages

```
library(tidyverse)
```

Load the *Titanic* dataset into your R environment.

```
dataset <- read_csv("../../1/titanic/titanic.csv")
names(dataset)   # attributes
```

Exercise 2.2 – Calculate probabilities

- If one of the passengers is randomly selected,
    - what is the probability that this passenger was in first class?

```
tmp1 <- dataset %>%
  filter(Pclass == 1)  %>%
  nrow()
round(tmp1/nrow(dataset), 2)
```

    - what is the probability that this passenger survived?
    - what is the probability that this passenger was in first class and survived?
- If one of the passengers is randomly selected from the first class passengers, what is the probability that this passenger survived? (That is, what is the probability that the passenger survived, given that this passenger was in first class?)

```
tmp4 <- dataset %>%
  filter(Survived == 1 & Pclass == 1)  %>%
  nrow()
round(tmp4/tmp1, 2)
```

- If one of the passengers who survived is randomly selected,
    - what is the probability that this passenger was in first class?
    - what is the probability that this passenger was in third class?

Exercise 2.3 – Calculate probabilities

Calculate the conditional probability that a person survives given his/her `Pclass` and `Sex`, i.e.

```
Pr(Survived = 1|Sex = female, Pclass = 2)= ?
Pr(Survived = 1|Sex = female, Pclass = 3) = ?
Pr(Survived = 1|Sex = male,Pclass = 1) = ?
Pr(Survived = 1|Sex = male,Pclass = 2) = ?
Pr(Survived = 1|Sex = male,Pclass = 3)= ?
```