Kønsfordeling i A History of World Societies 11. udgave

Jeg downloader først og fremmest tidyverse pakken, ved at skrive library(tidyverse). Dernæst downloader jeg here pakken, som gør det muligt for at R at lokalisere og registrere min data. Samt en række andre pakker, man skal bruge når man arbejder med materialet og text mining.

For at kunne arbejde med materialet, bliver man først nødt til at få registrere hvor dataen er placeret. Det gør jeg ved at bruge here() funktionen, og skrive; here("data","ahows.pdf"). På den måde har jeg fortalt R, at tekstfilen ahows.pdf, som er denne opgaves data, ligger i mappen, ved navn data. Jeg kalder denne variabel ahows_path.

```
ahows_path <- here("data","ahows.pdf")
ahows_text <- tibble(text=pdf_text(ahows_path))
```

Nu når der er blevet skabt en "vej" hen til materialet, bliver vi nødt til at putte dataen ind i en data frame, for at vi kan lave den om til tidy text. Dette gør jeg ved at bruge funktionen: tibble(text=pdf_text(ahows_path)), og på den måde skabe variablen ahows_text. Funktionen tibble er en data frame i R. Nu har jeg konstrueret en variabel, hvor hver enkelt side i pdf'en, er på hver sin række.

For at se hvordan R har registeret mit materiale bruger jeg glimpse()-funktionen.

```
ahows_text_tidy <- ahows_text %>%
  unnest_tokens(word,text)

glimpse(ahows_text_tidy)
```

```
## Rows: 731,326
## Columns: 1
## $ word <chr> "23.04.2025", "12.25", "full", "text", "of", "a", "history", "of"…
```

Det kan være vigtigt at gøre opmærksom på, at det er alle ord som pdf'en indeholder der er med i denne variabel. Dette gælder også det sidehovede og den sidefod der blev konstrueret, da jeg downloadet bogens tekst ned fra hjemmesiden Internet Archive. Jeg har ikke valgt at fjerne dette, da jeg ikke ser det relevant for denne opgave. Hvis jeg havde arbejdet med materialet på en anden måde, kunne det dog have været nødvendigt.

Nu er dataen blevet omstruktureret så det nu er tidy text. Nu kan vi isolere og filtre udvalgte ord.

Jeg bruger funktionen filter(word %in% c()) til filtrer de ord fra, jeg ønsker at isolere. I dette tilfælde de engelske pronominer, samt count(word,sort=TRUE) som er den del af funktionen, der kommer til at fortælle og sortere efter hyppigheden . Jeg giver denne variabel navnet pronominer

```
pronominer <-ahows_text_tidy %>%
  filter(word %in%  c("he","him","his", "she","her","hers")) %>%
  count(word, sort=TRUE)

glimpse(pronominer)
```

```
## Rows: 6
## Columns: 2
## $ word <chr> "his", "he", "her", "him", "she", "hers"
## $ n    <int> 2145, 1705, 683, 409, 408, 4
```

Nu kan vi se hvor mange gange de enkelte pronominer er blevet brugt. N står for antal. Her kan man se at det pronomen der fremgår hyppigst i bogen, er his, og det der fremgår mindst, er hers.

Jeg ønsker at lave et søjlediagram som visualisering, hvor henholdsvis pronominerne er på den ene akse, og antal gange er på den anden.

Jeg konstruerer derfor en variabel ved navnet data_pronominer, hvor jeg putter min nye data om pronominer ind i en dataframe, fordelt på ord (word), og antal, som er fordelt på hver sin kolonne.

```
data_pronominer <- data.frame(
  word = c("his", "he", "her", "him", "she", "hers"),
  antal = c(2145, 1705, 683, 409, 408, 4)
)
```

Herefter laver jeg en variabel ved navnet df, hvor jeg sorteret pronominerne i valgt rækkefølge. Da jeg var i tvivl om konstruktionen af en sådan funktion spurgte jeg ChatGPT om sprøgmål:

if i have a list of values such as (he, she, hers, his, her, him) and their frequency in R vector, how can i show them in a ggplot in deliberate order? show me a soultion for tidyverse.

```
df <- tibble(
  word = factor(data_pronominer$word, levels = c("he", "she", "him", "her", "his", "hers")),
  frequency = data_pronominer$antal
  )
```

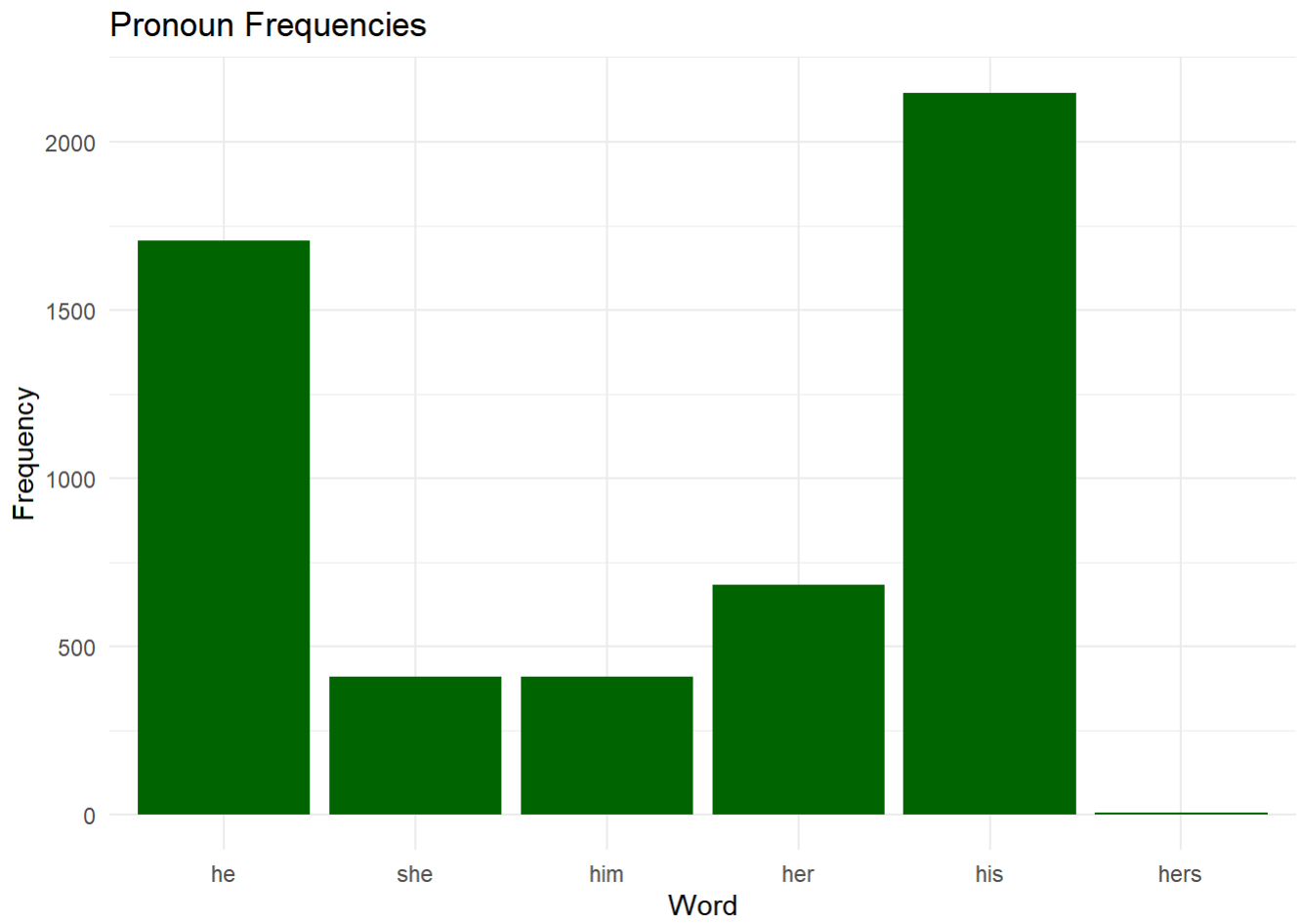
Da jeg nu har laver en variabel med den ønskede række følge, er mit data klar til at blive lavet om til en visualisering.

Laver visualisering ved brug af ggplot. Her indsætter jeg min df variabel, som konstrueret således at den har pronominerne den ønskede rækkefølge. ord kommer på x aksen, og antal (frequence) kommer på y aksen. Samt gør søjlerne grønne.

Jeg bruger ggplot koden for det udleverede script ‘Text mining, sentiment analysis, and visualization’.

```
Profreq <- ggplot(df, aes(x = word, y = frequency)) +
  geom_col(fill = "darkgreen") +
  labs(title = "Pronoun Frequencies", x = "Word", y = "Frequency") +
  theme_minimal()

Profreq
```



Gemmer min visualisering i mappen figures

```
ggsave(plot = Profreq,
       here("figures","profreq.pdf"),
       height = 6,
       width = 8)
```