

Textanalys av Sveriges regeringsförklaring med R och tidytext

Filip Wästberg, Ferrologic

2018-03-06

Vad är en regeringsförklaring?

Var finns regeringsförklaringen?

The screenshot shows the homepage of the Regeringskansliet website. At the top, there is a navigation bar with links to 'Lyssna', 'English website', 'Lättläst', 'Teckenspråk', 'Other languages', 'Prenumerera via e-post', and 'Kontakt'. Below the navigation bar is a search bar containing the text 'regeringsförklaring' and a 'Sök' button. Underneath the search bar are three main menu items: 'Sveriges regering' (Statsråden och departementen), 'Regeringens politik' (Detta görs inom olika områden), and 'Så styrs Sverige' (Om regeringen, Regeringskansliet och EU). The main content area is titled 'Filtrera ditt sökresultat' (Filter your search result) and displays a list of search filters: 'Innehållstyper', 'Områden', 'Statsråd', 'Departement/Övriga avsändare', and 'Datum'. To the right of these filters, it says 'Din sökning på **regeringsförklaring** gav 171 träffar.' (Your search for **regeringsförklaring** resulted in 171 hits.) Below this, there are sorting options ('Sortera på Relevans' or 'Datum') and inclusion settings ('Inkludera dokument i sökresultat' with 'Ja' or 'Nej'). The search results list includes two entries: 'Regeringsförklaringen den 12 september 2017' (with a link to '12 september 2017 · Tal från Regeringen, Statsrådsberedningen, Stefan Löfven') and 'Regeringsförklaring'.

Efter en del kodande och mycket copy paste

```
library(tidyverse)
regf <- read_csv("data/regf.csv")
```

En data.frame med stycke, datum och statsminister

```
## # A tibble: 4,283 x 3
##   text                               datum    statsminister
##   <chr>                             <date>   <chr>
## 1 Regeringsförklaring, 19761008, Torbjörn Fäll... 1976-10-08 Torbjörn Fäll...
## 2 Regeringspartierna är ense om att finansdepar... 1976-10-08 Torbjörn Fäll...
## 3 I anslutning till denna anmälan vill jag ge r... 1976-10-08 Torbjörn Fäll...
## 4 Enighet har alltså nåtts mellan Centerpartiet... 1976-10-08 Torbjörn Fäll...
## 5 Med fasthet och ansvar skall vi föra en polit... 1976-10-08 Torbjörn Fäll...
## 6 Regeringen skall sträva efter att bryta tende... 1976-10-08 Torbjörn Fäll...
## 7 Tryggheten för alla generationer och grupper ... 1976-10-08 Torbjörn Fäll...
## 8 Den sociala marknadshushållningen förstärks g... 1976-10-08 Torbjörn Fäll...
## 9 Regeringen skall föra en politik för sysselsä... 1976-10-08 Torbjörn Fäll...
## 10 Sträng hushållning måste ske med naturtillgån... 1976-10-08 Torbjörn Fäll...
## # ... with 4,273 more rows
```

Vad är vi intresserade av?

- Vilka ord som nämns
- Hur regeringsförklaringen ändrats över tid
- Vilka ord som är viktigast

För att analysera text behöver den vara tidy

- Varje variabel är en kolumn
- Varje observation är en rad
- I vårt fall handlar det om varje ord ska vara uppdelat per statsminister och regeringsförklaring
- Data är alltså inte tidy

Vansinnigt enkelt att göra text tidy med tidytext

```
library(tidytext)
tidy_regf <- regf %>%
  filter(!str_detect(text, "Regeringsförklaring")) %>%
  unnest_tokens(ord, text)
```

Tidy text data

```
## # A tibble: 129,341 x 3
##   datum   statsminister   ord
##   <date>   <chr>        <chr>
## 1 1976-10-08 Torbjörn Fälldin regeringspartierna
## 2 1976-10-08 Torbjörn Fälldin är
## 3 1976-10-08 Torbjörn Fälldin ense
## 4 1976-10-08 Torbjörn Fälldin om
## 5 1976-10-08 Torbjörn Fälldin att
## 6 1976-10-08 Torbjörn Fälldin finansdepartementet
## 7 1976-10-08 Torbjörn Fälldin så
## 8 1976-10-08 Torbjörn Fälldin snart
## 9 1976-10-08 Torbjörn Fälldin som
## 10 1976-10-08 Torbjörn Fälldin möjligt
## # ... with 129,331 more rows
```

Stoppord

- Ord som *och*, *så*, *att* och *så* vidare
- Finns lista i :

```
get_stopwords(language = "sv", source = "snowball")
```

```
## # A tibble: 114 x 2
##   word  lexicon
##   <chr> <chr>
## 1 och   snowball
## 2 det   snowball
## 3 att   snowball
## 4 i     snowball
## 5 en    snowball
## 6 jag   snowball
## 7 hon   snowball
## 8 som   snowball
## 9 han   snowball
## 10 på   snowball
## # ... with 104 more rows
```

Stoppord 😺

- Kompletterar med lista från forskaren Peter Dahlgren från Göteborgs universitet

```
sv_stoppord <- read_csv("https://gist.githubusercontent.com/peterdalle/8865eb918a82  
  rename(stoppord = X1)  
  
## Parsed with column specification:  
## cols(  
##   X1 = col_character()  
## )
```

Tvätta data

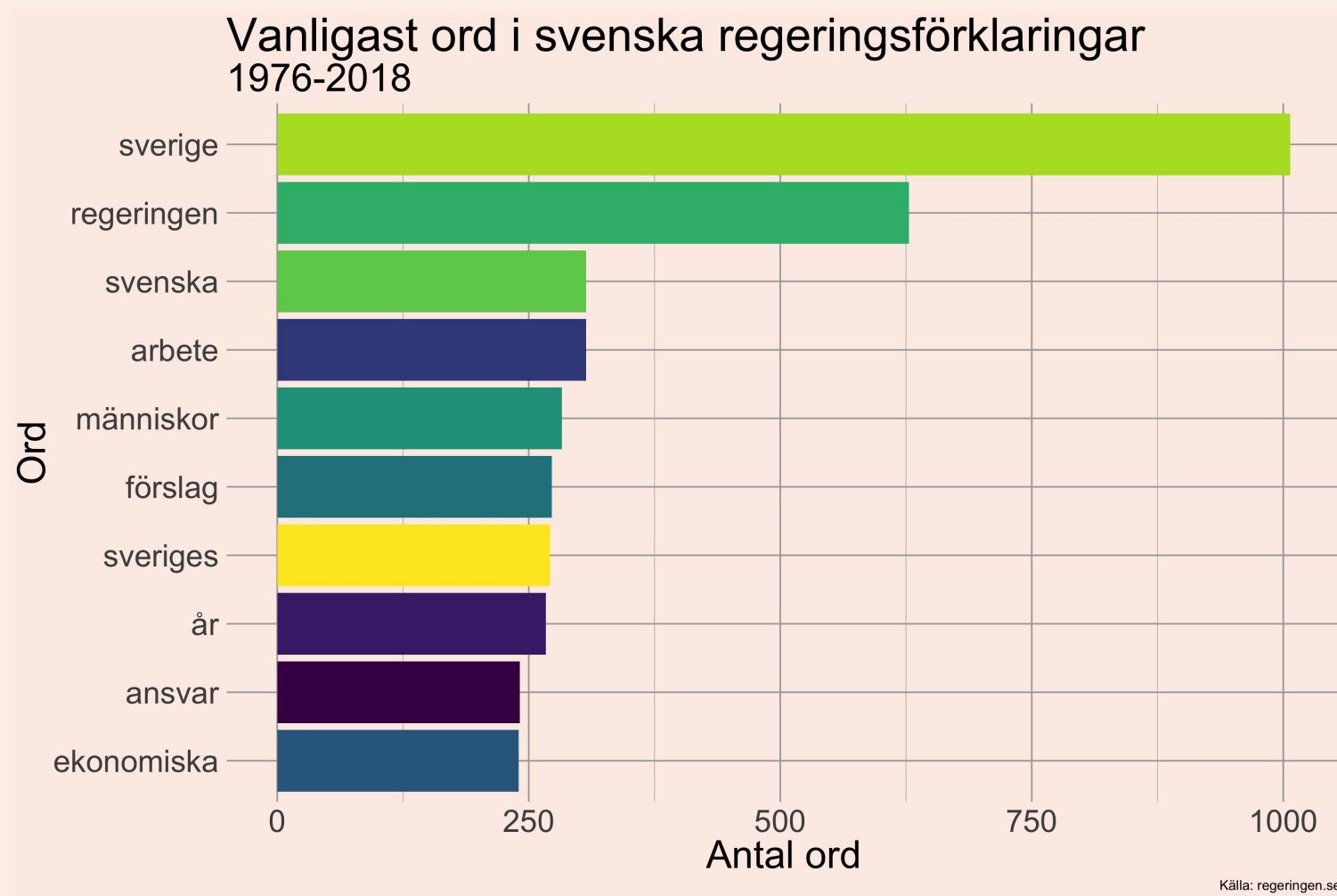
```
tvättade_regf <- tidy_regf %>%
  filter(!str_detect(ord, "[[:digit:]]")) %>%
  anti_join(get_stopwords(language = "sv"), by = c("ord" = "word")) %>%
  anti_join(sv_stoppord, by = c("ord" = "stoppord"))
```

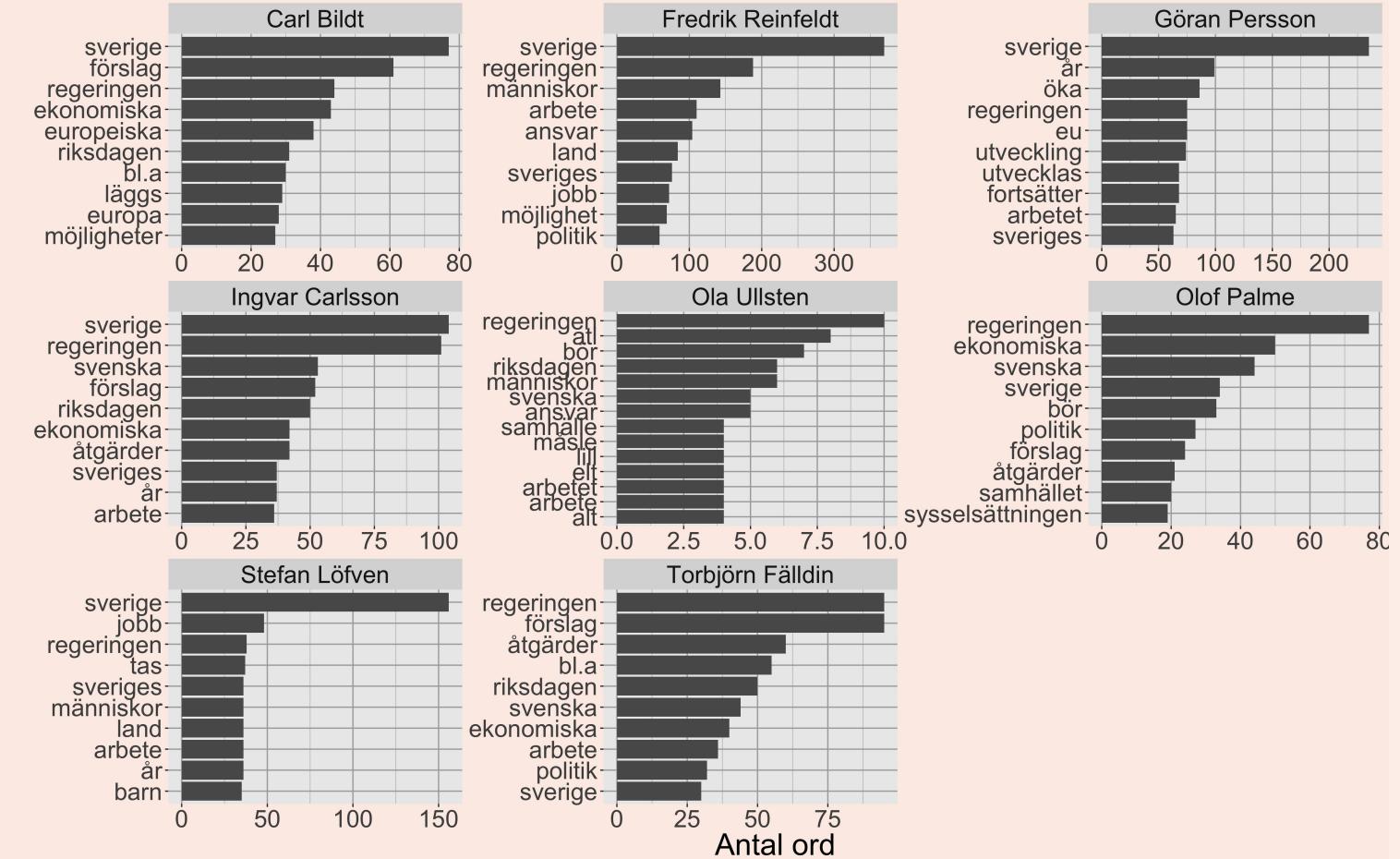
Tvätta data

```
tvättade_regf <- tidy_regf %>%
  filter(!str_detect(ord, "[[:digit:]]")) %>%
  anti_join(get_stopwords(language = "sv"), by = c("ord" = "word")) %>%
  anti_join(sv_stoppord, by = c("ord" = "stoppord"))
```

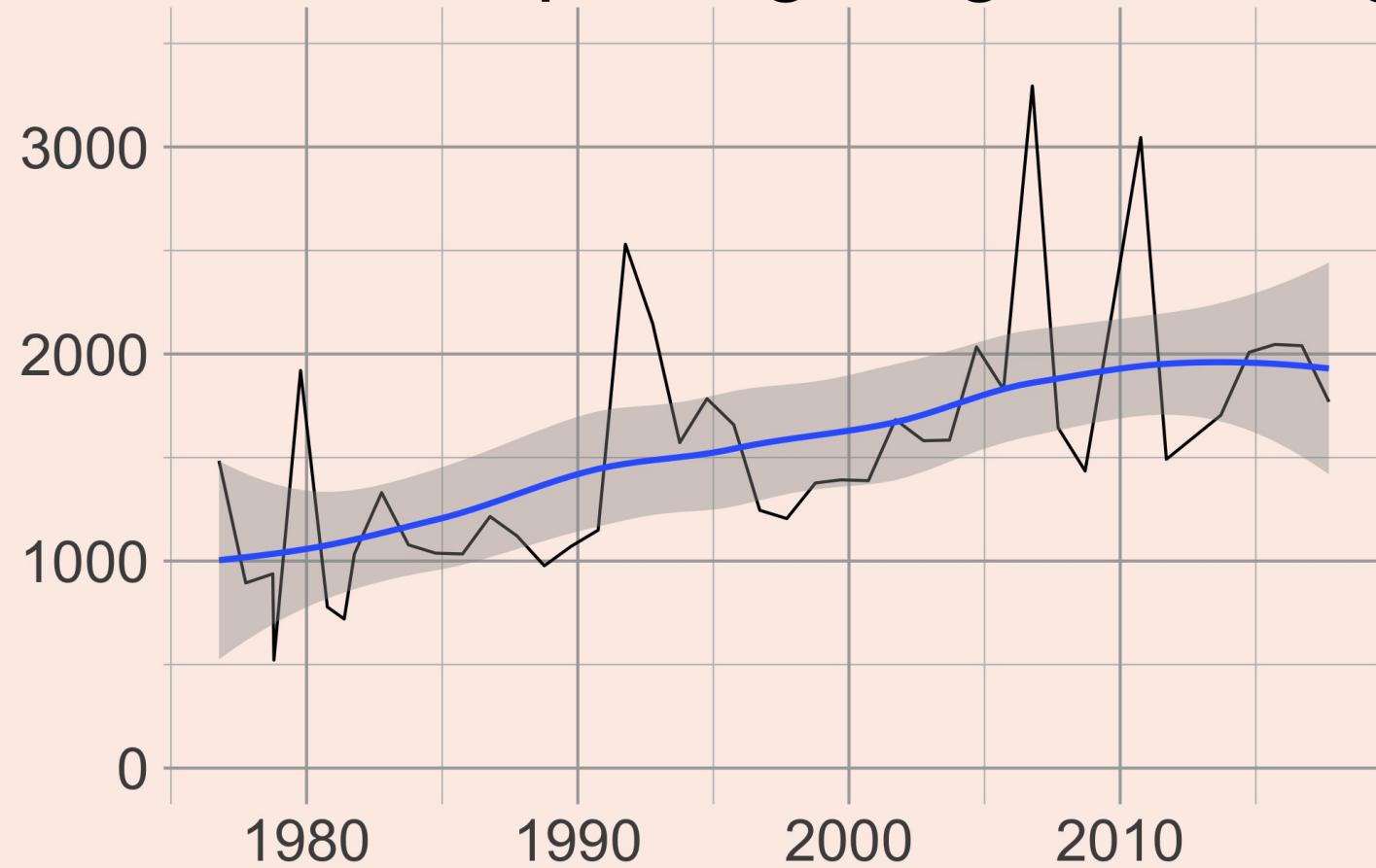
Tvätta data

```
tvättade_regf <- tidy_regf %>%
  filter(!str_detect(ord, "[[:digit:]]")) %>%
  anti_join(get_stopwords(language = "sv"), by = c("ord" = "word")) %>%
  anti_join(sv_stoppord, by = c("ord" = "stoppord"))
```

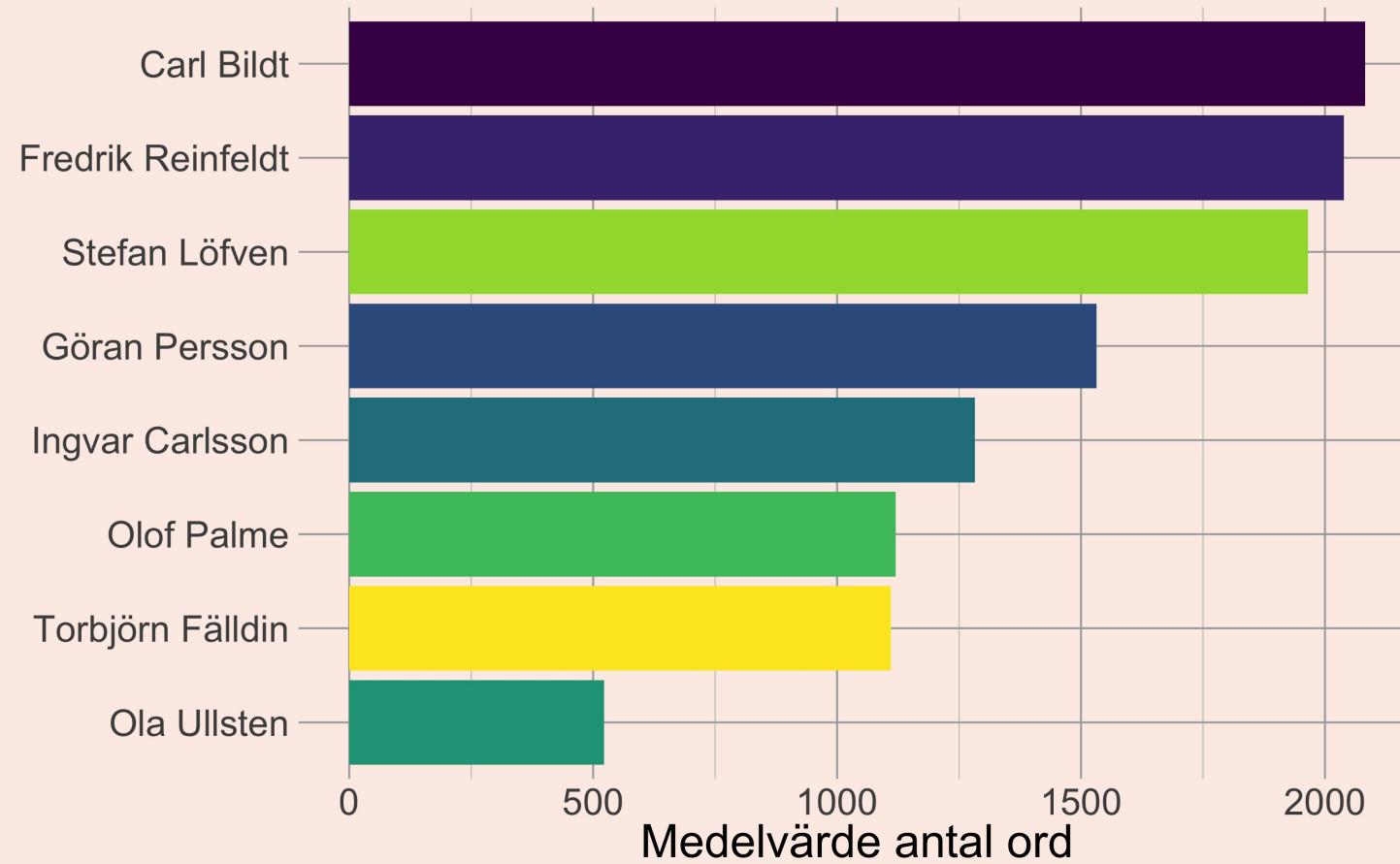


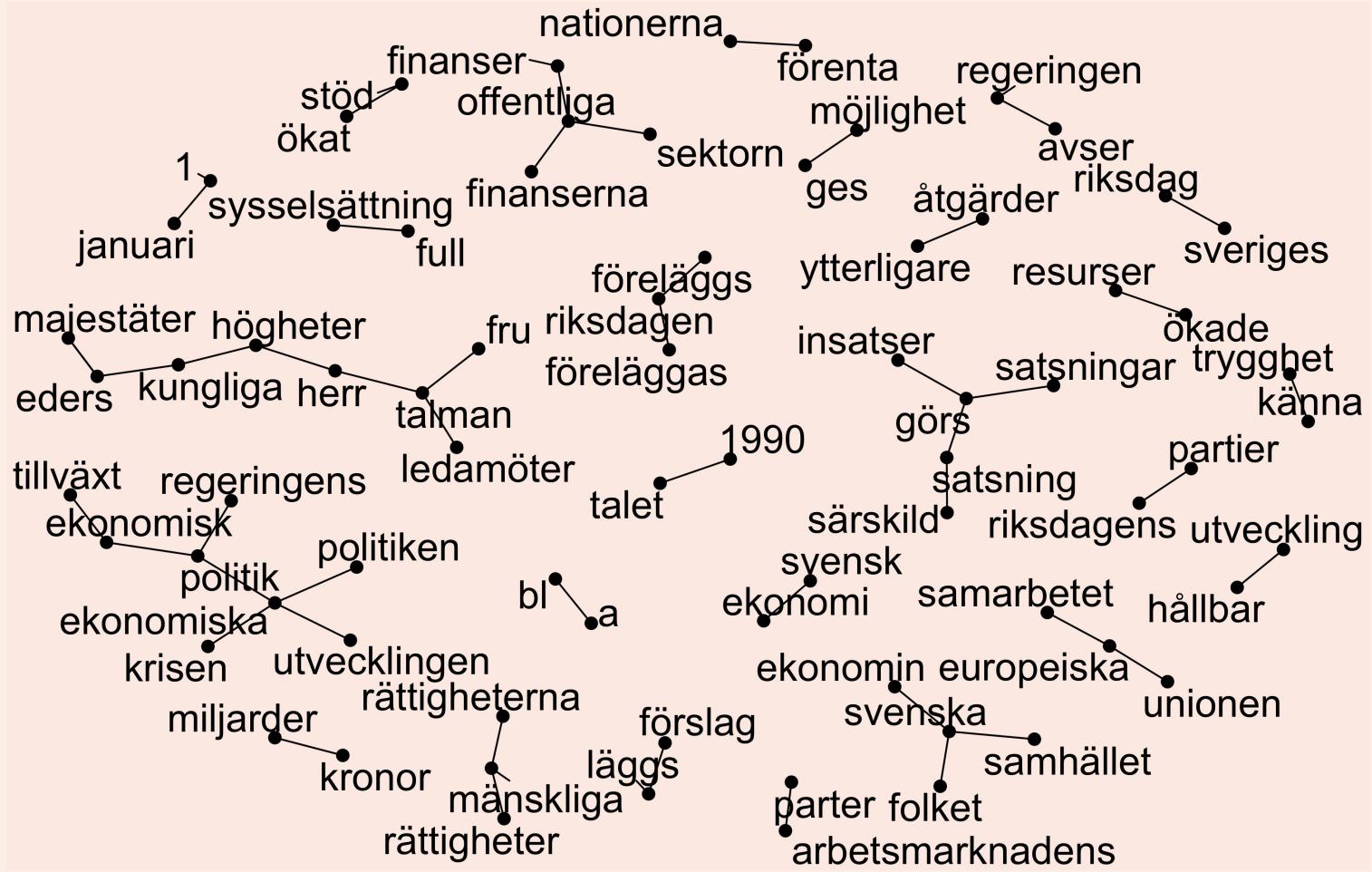


Antal ord per regeringsförklaring



Vem snackar mest?



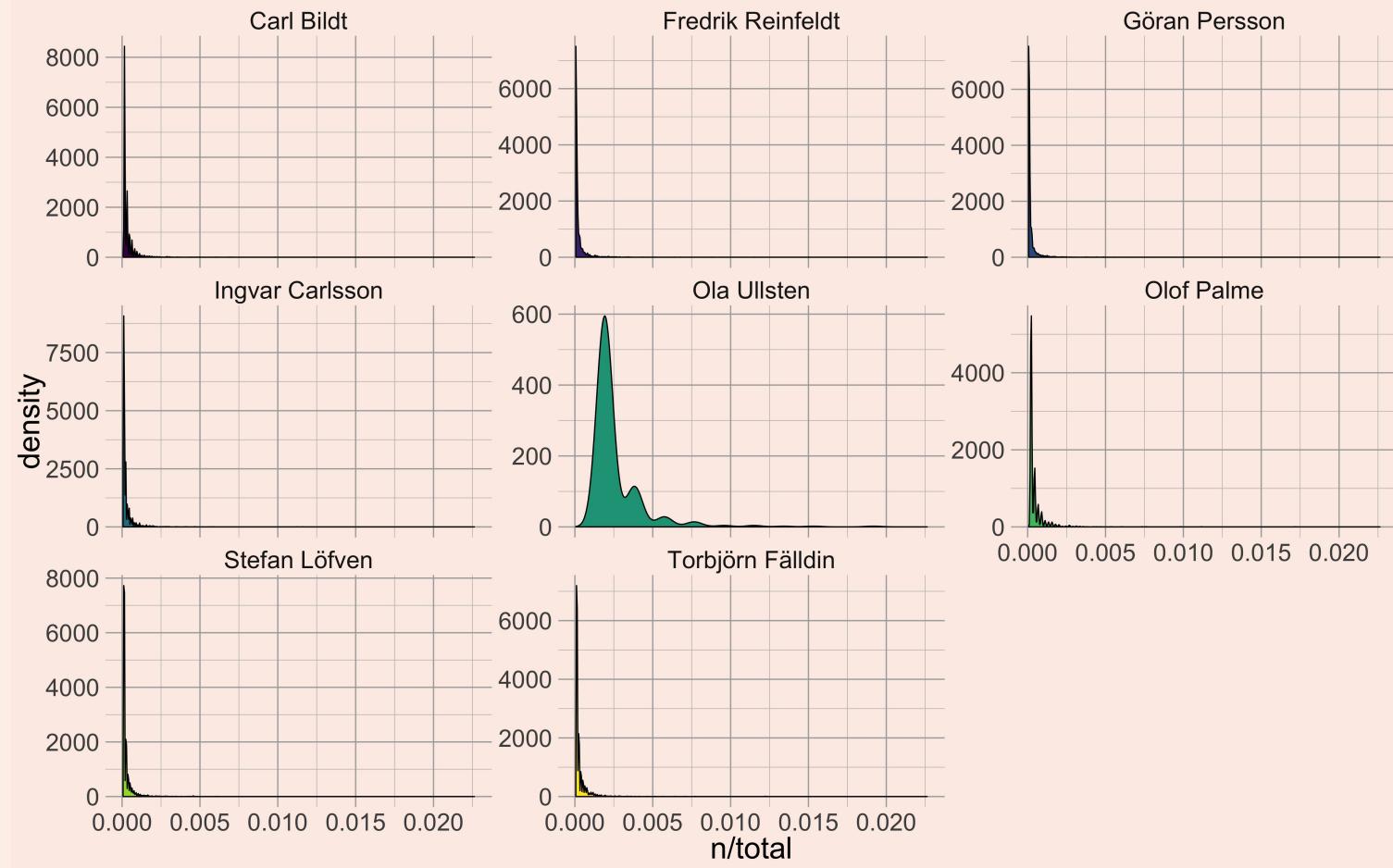


Vilka är de viktigaste orden för respektive statsminister?

Texten är skev!

- Zipfs law





Karen Spärck Jones



Foto Från: University of Cambridge, CC BY 2.5

Term frequency–inverse document frequency

$$tf - idf(term) = tf(term) * idf(term)$$

$$tf(term) = f_t$$

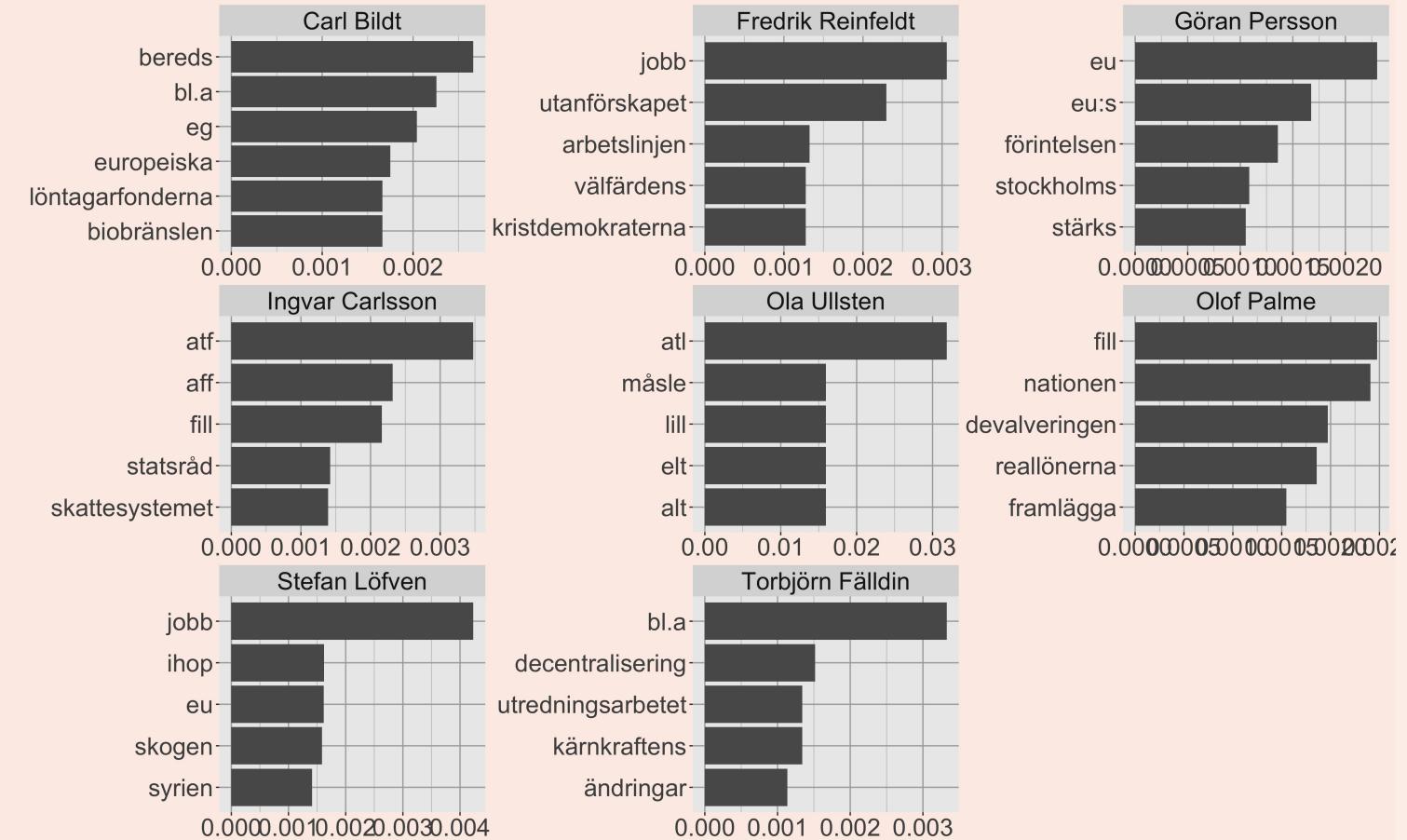
$$idf(term) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right)$$

Term frequency–inverse document frequency

- Skapar en vikt för ord baserat på hur ofta de förekommer i varje regeringsförklaring
- Kombinera med *term frequency* och vi kan ta fram de "viktigaste" orden för repektive statsminister

Term frequency–inverse document frequency

- Exempel:
- Sveriges regering sätter kampen mot **arbetslösheten** främst
- Sverige regering sätter kampen mot **terrorism** främst



- filip.wastberg@ferrologic.se
- All kod för presentationen:
<https://github.com/filipwastberg/regeringsforklaring>