

Homework, Session #5

Visualisation and Statistical Analysis

2018-10-22

Data manipulation and plotting exercise

Format: Submit a markdown report with text explanation and functional R code. The report has two parts: each part should have a couple of paragraphs of text and one or two graphics. Please submit the file in rmarkdown (Rmd) format only: I will run the code myself to produce the graphics etc. Style note: it's good practice to e.g. check every step of a pipeline as you write it, but once it works you don't have to show me the parts. Just give me the full analysis. Likewise, it's good practice to e.g. run `head()` on your data to confirm that it contains what you think it does, but don't put that in your report — I can check it myself if I want to.

Due: Sunday evening (28 Oct). I will go through interesting results in the lecture the next day. You are welcome to work on this together, but ultimately you should carry out and report distinct analyses.

Part 1: Numerals

Please download from the course website the file `numeral-frequency.tsv`, kindly produced by our colleagues Marc Tang and Marie Dubremetz in April 2018. This contains the corpus frequency of Swedish numerals printed in word form (e.g. *två* but not *2*), for numerals from 1 to 50.

The corpora used contain in total 641,404,367 tokens and originate from:

- Swedish Wikipedia available at Wikipedia Monolingual Corpora
- Swedish web news corpora (2001-2013) and Swedish Wikipedia corpus collected by Språkbanken. <https://spraakbanken.gu.se/eng/resources/corpus>

There are three columns:

```
numerals <- read_tsv("numeral-frequency.tsv")
```

Parsed with column specification:

```
cols(  
  freq = col_integer(),  
  numeral = col_integer(),  
  word = col_character()  
)
```

```
head(numerals)
```

```
# A tibble: 6 x 3
```

	freq	numeral	word
	<int>	<int>	<chr>
1	3249009	1	ett
2	729451	2	två
3	427844	3	tre
4	253325	4	fyra
5	183557	5	fem
6	140779	6	sex

The numeral column has a numeric representation of the number, and the word column has the Swedish number word. Please investigate the frequency of Swedish number words and report your analysis and findings.

Part 2: Babynames

Spend some more time exploring the babynames data, and write a report on something interesting that you find. Here are some ideas to get you started:

- Find different spellings of a common name (e.g. Anna, Anne, Ann) and plot the change in proportion of how the popularity of the different spellings has changed over time
- Can you find evidence of e.g. *Game of Thrones*, pop stars, or US presidents having an effect on naming practices?
- Calculate the proportion of vowels in each name and plot the trends in mean vowel proportion over time for girls' and boys' names