Estimation Plots

Chris Bentz

10/01/2023

Load libraries

If the libraries are not installed yet, you need to install them using, for example, the command: install.packages("ggplot2"). For the Hrate package this is different, since it comes from github. The devtools library needs to be installed, and then the install_github() function is used.

```r
library(stringr)
library(ggplot2)
library(ggrepel)
library(plyr)
library(ggExtra)
library(ggpubr)
```

```
##
## Attaching package: 'ggpubr'

## The following object is masked from 'package:plyr':
##
##     mutate
```

Load Data

Load data table with values per text file.

```r
# load estimations from stringBase corpus
estimations10.df <- read.csv("~/Github/NaLaFi/results/estimation10chars.csv")
estimations100.df <- read.csv("~/Github/NaLaFi/results/estimation100chars.csv")
estimations1000.df <- read.csv("~/Github/NaLaFi/results/estimation1000chars.csv")
#head(estimations10.df.)
```

Exclude subcorpora (if needed).

```r
selected <- c("shuffled", "random")
estimations10.df <- estimations10.df[!(estimations10.df$subcorpus %in% selected), ]
estimations100.df <- estimations100.df[!(estimations100.df$subcorpus %in% selected), ]
estimations1000.df <- estimations1000.df[!(estimations1000.df$subcorpus %in% selected), ]
```

Scatterplots

10 Characters

Entropy rate vs. unigram entropy for characters

```
huni.hrate.10chars.plot <- ggplot(estimations10.df,
                                  aes(x = huni.chars, y = ttr.chars,
                                      colour = corpus)) +
  geom_point(alpha = 0.3, size  = 1) +
  labs(x = "Unigram entropy for 10 characters", y = "TTR for 10 characters") +
  theme(legend.position = "none")
huni.hrate.10chars.plot <- ggMarginal(huni.hrate.10chars.plot, groupFill = T,
                                      type = "histogram", colour = "white")
#huni.hrate.10chars.plot
```

TTR vs. repetition rate for characters

```
ttr.rm.10chars.plot <- ggplot(estimations10.df,
                    aes(x = hrate.chars, y = rm.chars,
                    colour = corpus)) +
  geom_point(alpha = 0.3, size  = 1) +
  theme(legend.position = "left") +
  labs(x = "Entropy rate for 10 characters", y = "Repetition rate for 10 characters")
ttr.rm.10chars.plot <- ggMarginal(ttr.rm.10chars.plot, groupFill = T,
                                  type = "histogram", colour = "white")
#ttr.rm.10chars.plot
```

100 Characters

Entropy rate vs. unigram entropy for characters

```
huni.hrate.100chars.plot <- ggplot(estimations100.df,
                                   aes(x = huni.chars, y = ttr.chars,
                                       colour = corpus)) +
  geom_point(alpha = 0.3, size  = 1) +
  labs(x = "Unigram entropy for 100 characters", y = "TTR for 100 characters") +
  theme(legend.position = "none")
huni.hrate.100chars.plot <- ggMarginal(huni.hrate.100chars.plot, groupFill = T,
                                       type = "histogram", colour = "white")
#huni.hrate.100chars.plot
```

TTR vs. repetition rate for characters

```
ttr.rm.100chars.plot <- ggplot(estimations100.df,
                    aes(x = hrate.chars, y = rm.chars,
                    colour = corpus)) +
  geom_point(alpha = 0.3, size  = 1) +
```

```
  theme(legend.position = "left") +
  labs(x = "Entropy rate for 100 characters", y = "Repetition rate for 100 characters")
ttr.rm.100chars.plot <- ggMarginal(ttr.rm.100chars.plot, groupFill = T,
                                   type = "histogram", colour = "white")
#ttr.rm.100chars.plot
```

1000 Characters

Entropy rate vs. unigram entropy for characters

```
huni.hrate.1000chars.plot <- ggplot(estimations1000.df,
                                    aes(x = huni.chars, y = ttr.chars,
                                        colour = corpus, shape = subcorpus)) +
  scale_shape_manual(values = 1:nlevels(estimations1000.df$subcorpus)) +
  geom_point(alpha = 0.5, size  = 1.5) +
  labs(x = "Unigram entropy for 1000 characters", y = "TTR for 1000 characters") +
  theme(legend.position = "none")
huni.hrate.1000chars.plot <- ggMarginal(huni.hrate.1000chars.plot, groupFill = T,
                                        type = "histogram", colour = "white")
#huni.hrate.1000chars.plot
```

TTR vs. repetition rate for characters

```
ttr.rm.1000chars.plot <- ggplot(estimations1000.df,
                   aes(x = hrate.chars, y = rm.chars,
                   colour = corpus, shape = subcorpus)) +
  scale_shape_manual(values = 1:nlevels(estimations1000.df$subcorpus)) +
  geom_point(alpha = 0.5, size  = 1.5) +
  theme(legend.position = "left") +
  labs(x = "Entropy rate for 1000 characters", y = "Repetition rate for 1000 characters")
ttr.rm.1000chars.plot <- ggMarginal(ttr.rm.1000chars.plot, groupFill = T,
                                    type = "histogram", colour = "white")
#ttr.rm.1000chars.plot
```
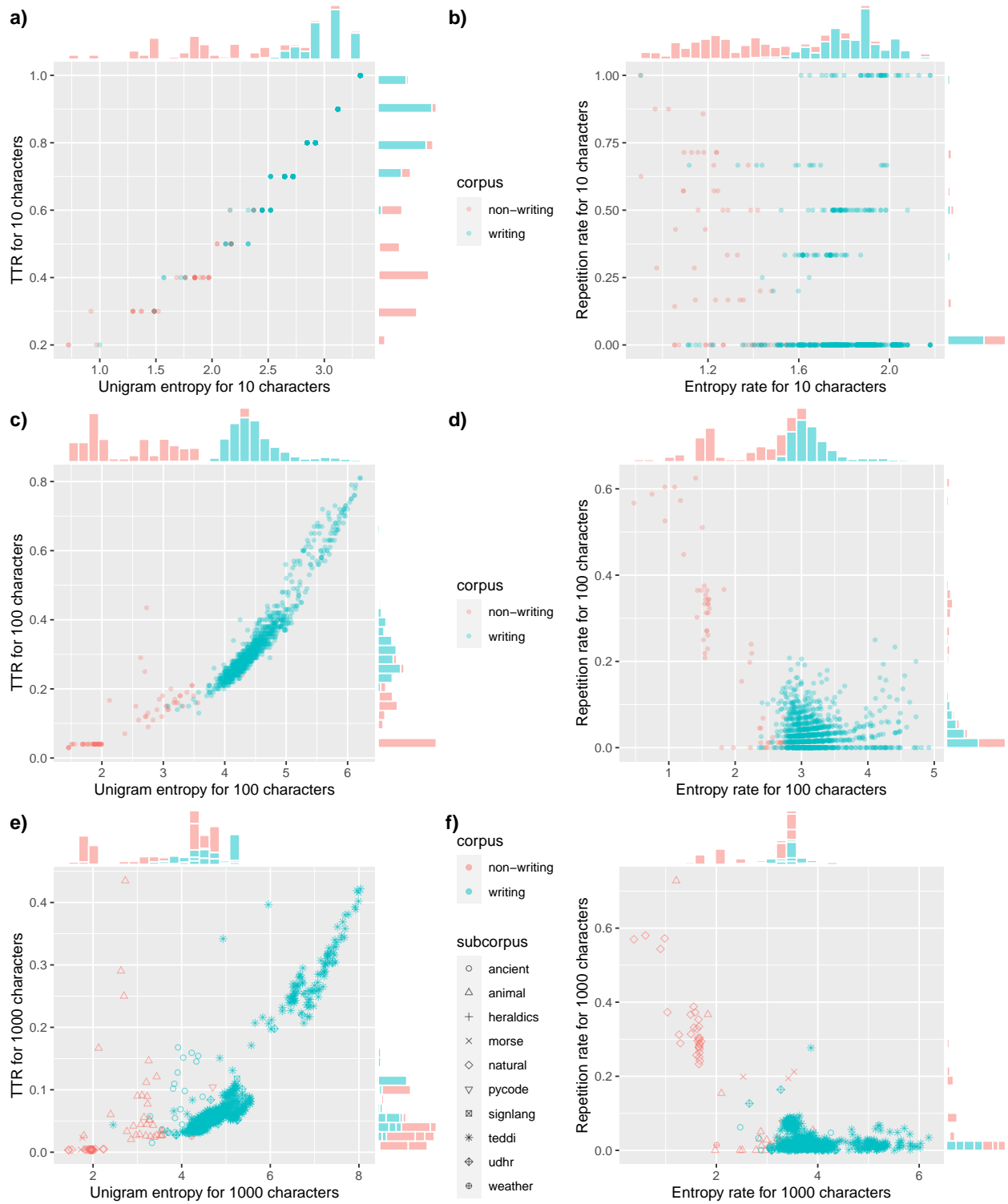
Combined Plots

```
plots.combined <- ggarrange(huni.hrate.10chars.plot, ttr.rm.10chars.plot,
                            huni.hrate.100chars.plot, ttr.rm.100chars.plot,
                            huni.hrate.1000chars.plot, ttr.rm.1000chars.plot,
                   labels = c("a)", "b)", "c)", "d)",
                              "e)", "f)"),
                   ncol = 2, nrow = 3, widths = c(1, 1.3) )
plots.combined
```

Safe complete figure to file

```
ggsave("~/Github/NaLaFi/figures/plots_combined.pdf", plots.combined, width = 10,
        height = 12, dpi = 300, scale = 1, device = cairo_pdf)
```