Estimation Plots

Chris Bentz

14/10/2023

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.

```
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.

```
# load estimations from stringBase corpus
estimations.df <- read.csv("~/Github/NaLaFi/results/features.csv")
#head(estimations10.df.)
```

Exclude subcorpora (if needed).

```
#selected <- c("shuffled")
#estimations.df <- estimations.df[!(estimations.df$subcorpus %in% selected), ]
```

Split into separate files by length of chunks in characters.

```
estimations10.df <- estimations.df[estimations.df$num.char == 10, ]
estimations100.df <- estimations.df[estimations.df$num.char == 100, ]
estimations1000.df <- estimations.df[estimations.df$num.char == 1000,]
```

Scatterplots

10 Characters

Unigram entropy vs. TTR for characters

```
huni.ttr.10chars.plot <- ggplot(estimations10.df,
                                aes(x = huni.chars, y = ttr.chars,
                                    shape = subcorpus, colour = corpus)) +
  scale_shape_manual(values = 1:length(unique(estimations10.df$subcorpus))) +
  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.ttr.10chars.plot,
                                      groupFill = T, groupColour = T,
                                      type = "density")
```

Entropy rate vs. repetition rate for characters

```
hrate.rm.10chars.plot <- ggplot(estimations10.df,
                   aes(x = hrate.chars, y = rm.chars,
                   colour = corpus, shape = subcorpus)) +
  scale_shape_manual(values = 1:length(unique(estimations10.df$subcorpus))) +
  geom_point(alpha = 0.3, size  = 1) +
  theme(legend.position = "left") +
  labs(x = "Entropy rate for 10 characters", y = "Repetition rate for 10 characters")
hrate.rm.10chars.plot <- ggMarginal(hrate.rm.10chars.plot,
                                  groupFill = T, groupColour = T,
                                  type = "density")
```

100 Characters

Unigram entropy vs. TTR for characters

```
huni.ttr.100chars.plot <- ggplot(estimations100.df,
                                aes(x = huni.chars, y = ttr.chars,
                                    colour = corpus, shape = subcorpus)) +
  scale_shape_manual(values = 1:length(unique(estimations100.df$subcorpus))) +
  geom_point(aes(fill = corpus), alpha = 0.3, size  = 1) +
  labs(x = "Unigram entropy for 100 characters", y = "TTR for 100 characters") +
  theme(legend.position = "none")
huni.ttr.100chars.plot <- ggMarginal(huni.ttr.100chars.plot,
                                       groupFill = T, groupColour = T,
                                       type = "density")
```

Entropy rate vs. repetition rate for characters

```
hrate.rm.100chars.plot <- ggplot(estimations100.df,
                   aes(x = hrate.chars, y = rm.chars,
                   colour = corpus, shape = subcorpus)) +
  scale_shape_manual(values = 1:length(unique(estimations100.df$subcorpus))) +
  geom_point(alpha = 0.3, size  = 1) +
  theme(legend.position = "left") +
  labs(x = "Entropy rate for 100 characters", y = "Repetition rate for 100 characters")
hrate.rm.100chars.plot <- ggMarginal(hrate.rm.100chars.plot,
                                   groupFill = T, groupColour = T,
                                   type = "density")
```

1000 Characters

Unigram entropy vs. TTR for characters

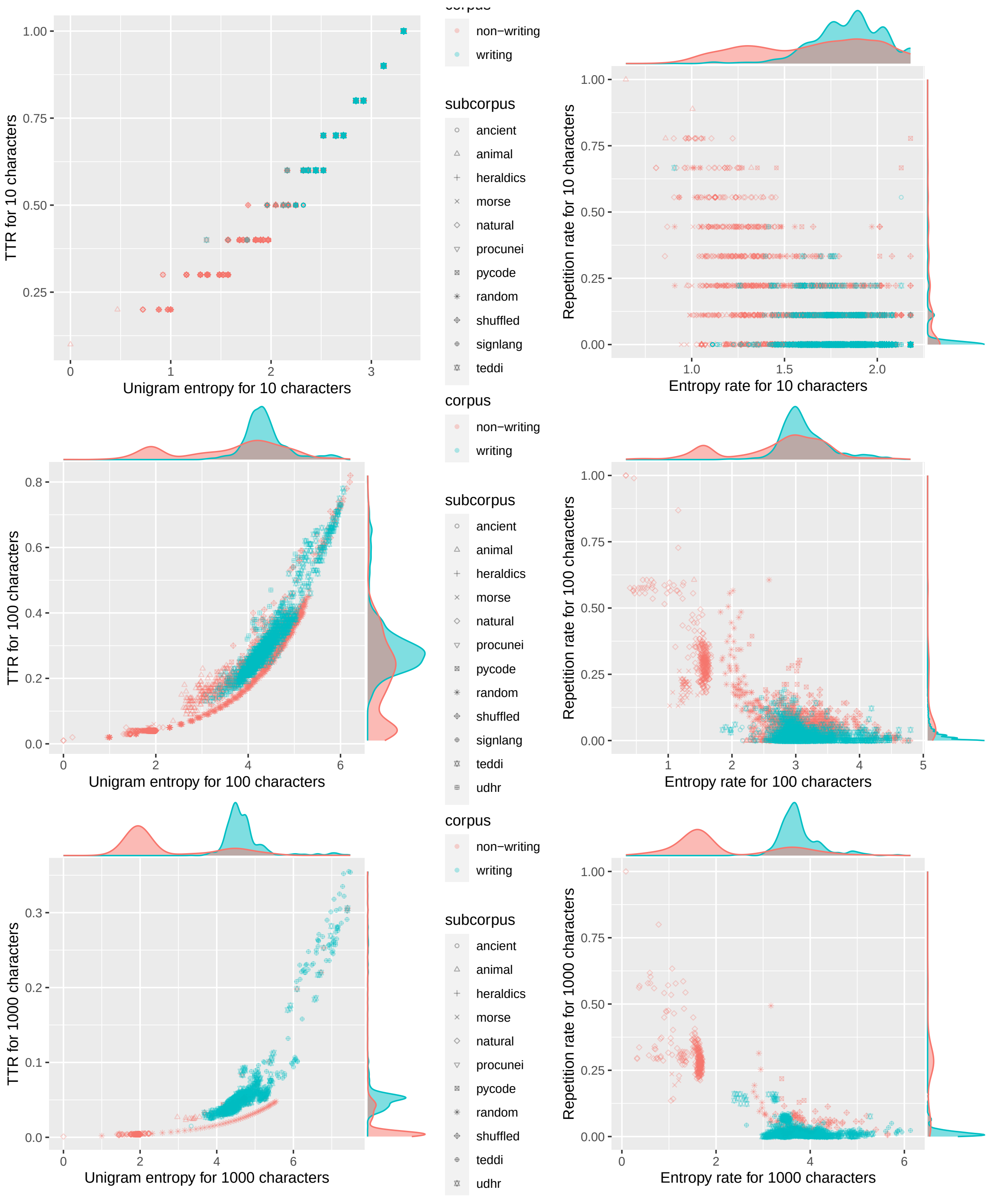
```
huni.ttr.1000chars.plot <- ggplot(estimations1000.df,
                                aes(x = huni.chars, y = ttr.chars,
                                    colour = corpus, shape = subcorpus)) +
  scale_shape_manual(values = 1:length(unique(estimations1000.df$subcorpus))) +
  geom_point(alpha = 0.3, size  = 1) +
  labs(x = "Unigram entropy for 1000 characters", y = "TTR for 1000 characters") +
  theme(legend.position = "none")
huni.ttr.1000chars.plot <- ggMarginal(huni.ttr.1000chars.plot,
                                   groupFill = T, groupColour = T,
                                   type = "density")
```

Entropy rate vs. repetition rate for characters

```
hrate.rm.1000chars.plot <- ggplot(estimations1000.df,
                   aes(x = hrate.chars, y = rm.chars,
                   colour = corpus, shape = subcorpus)) +
  scale_shape_manual(values = 1:length(unique(estimations1000.df$subcorpus))) +
  geom_point(alpha = 0.3, size  = 1) +
  theme(legend.position = "left") +
  labs(x = "Entropy rate for 1000 characters", y = "Repetition rate for 1000 characters")
hrate.rm.1000chars.plot <- ggMarginal(hrate.rm.1000chars.plot,
                                   groupFill = T, groupColour = T,
                                   type = "density")
```

Combined Plots

```
plots.combined <- ggarrange(huni.ttr.10chars.plot, hrate.rm.10chars.plot,
                            huni.ttr.100chars.plot, hrate.rm.100chars.plot,
                            huni.ttr.1000chars.plot, hrate.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)
```