Simple Corpus Stats

Chris Bentz

20/12/2022

Load libraries

If the libraries are not installed yet, you need to install them using, for example, the command: install.packages("ggplot2").

```
library(stringr)
library(ggplot2)
library(plyr)
```

List files

List all the files in the directory "corpus".

```
file.list <- list.files(path = "~/Github/NaLaFi/data/",
                        recursive = T, full.names = T)
#print(file.list)
length(file.list)
```

```
## [1] 329
```

Count number of files in subfolders

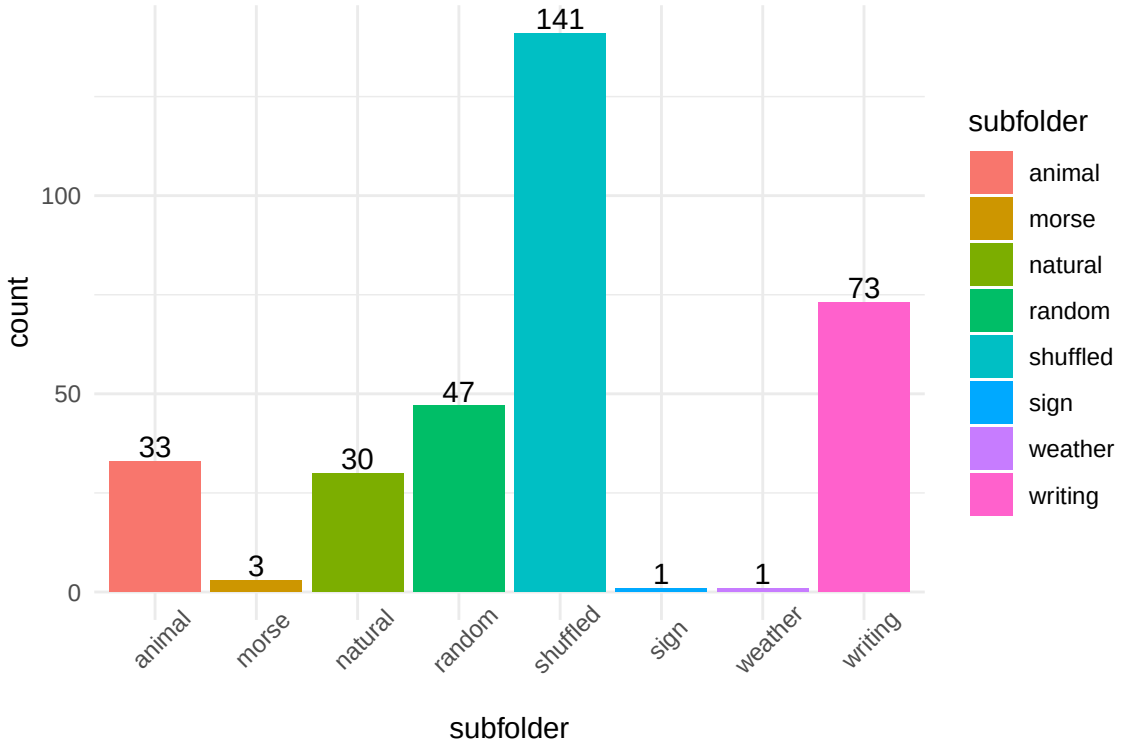
Count how many files are in each of the highest level subfolders of "data", and create a dataframe with counts.

```
#number of "animal" files
animal.count <- length(file.list[grepl("animal", file.list)])
#number of "morse" files
morse.count <- length(file.list[grepl("morse", file.list)])
#number of "natural" files
natural.count <- length(file.list[grepl("natural", file.list)])
#number of "sign" files
sign.count <- length(file.list[grepl("sign", file.list)])
#number of "weather" files
weather.count <- length(file.list[grepl("weather", file.list)])
#number of "writing" files
writing.count <- length(file.list[grepl("writing", file.list)])
#number of "random" files
random.count <- length(file.list[grepl("random", file.list)])
#number of "shuffled" files
shuffled.count <- length(file.list[grepl("shuffled", file.list)])
```

```
#create data frame
df <- data.frame(subfolder = c("animal", "morse", "natural", "sign",
                               "weather", "writing", "random", "shuffled"),
                 count = c(animal.count, morse.count, natural.count, sign.count,
                           weather.count, writing.count, random.count, shuffled.count))
```

Create a bar plot with counts.

```
counts.plot <- ggplot(data = df, aes(x = subfolder, y = count, fill = subfolder)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = count), vjust = -0.2, color = "black", size = 4) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45))
counts.plot
```



Safe figure to file

```
ggsave("~/Github/NaLaFi/figures/simpleStats_counts.pdf", counts.plot, dpi = 300,
       scale = 1, device = cairo_pdf)
```

```
## Saving 6 x 4.5 in image
```

Lengths of files in characters

Read files and count characters

```
# set counter
counter = 0
# initialize dataframe to append results to
simpleStats.df <- data.frame(filename = character(0), subcorpus = character(0),
                             num.lines = numeric(0), num.chars = numeric(0))

for (file in file.list)
{
  # loading textfile
  textfile <- scan(file, what = "char", quote = "", comment.char = "",
                   encoding = "UTF-8", sep = "\n", skip = 7)
  textfile <- gsub("\t","",textfile) # remove tabs
  textfile <- gsub("<.*>","",textfile) # remove annotations marked by '<>'
  #print(head(textfile))
  # get filename
  filename <- basename(file)
  #print(filename) # for visual inspection
  # get subcorpus category
  subcorpus <- sub("_.*", "", filename)
  #print(subcorpus) # for visual inspection
  # count number of lines in text file
  num.lines <- length(textfile)
  # count the number of utf-8 characters in text file (note that this includes white
  # spaces)
  num.chars <- sum(nchar(textfile, type ="chars"))
  #print(num.chars) # for visual inspection
  # append results to dataframe
  local.df <- data.frame(filename, subcorpus, num.lines, num.chars)
  simpleStats.df <- rbind(simpleStats.df, local.df)
  # counter
  counter <- counter + 1
  #print(counter)
}
simpleStats.df
```

```
##                  filename subcorpus num.lines num.chars
## 1            random_ran_10    random         1      1000
## 2            random_ran_11    random         1      1000
## 3            random_ran_12    random         1      1000
## 4            random_ran_13    random         1      1000
## 5            random_ran_14    random         1      1000
## 6            random_ran_15    random         1      1000
## 7            random_ran_16    random         1      1000
## 8            random_ran_17    random         1      1000
## 9            random_ran_18    random         1      1000
## 10           random_ran_19    random         1      1000
## 11            random_ran_2    random         1      1000
## 12           random_ran_20    random         1      1000
## 13           random_ran_21    random         1      1000
```

```
## 14          random_ran_22    random         1      1000
## 15          random_ran_23    random         1      1000
## 16          random_ran_24    random         1      1000
## 17          random_ran_25    random         1      1000
## 18          random_ran_26    random         1      1000
## 19          random_ran_27    random         1      1000
## 20          random_ran_28    random         1      1000
## 21          random_ran_29    random         1      1000
## 22           random_ran_3    random         1      1000
## 23          random_ran_30    random         1      1000
## 24          random_ran_31    random         1      1000
## 25          random_ran_32    random         1      1000
## 26          random_ran_33    random         1      1000
## 27          random_ran_34    random         1      1000
## 28          random_ran_35    random         1      1000
## 29          random_ran_36    random         1      1000
## 30          random_ran_37    random         1      1000
## 31          random_ran_38    random         1      1000
## 32          random_ran_39    random         1      1000
## 33           random_ran_4    random         1      1000
## 34          random_ran_40    random         1      1000
## 35          random_ran_41    random         1      1000
## 36          random_ran_42    random         1      1000
## 37          random_ran_43    random         1      1000
## 38          random_ran_44    random         1      1000
## 39          random_ran_45    random         1      1000
## 40          random_ran_46    random         1      1000
## 41          random_ran_47    random         1      1000
## 42          random_ran_48    random         1      1000
## 43           random_ran_5    random         1      1000
## 44           random_ran_6    random         1      1000
## 45           random_ran_7    random         1      1000
## 46           random_ran_8    random         1      1000
## 47           random_ran_9    random         1      1000
## 48      shuffled_aii_0001  shuffled         1      1000
## 49      shuffled_akk_0001  shuffled         1      1000
## 50      shuffled_akk_0002  shuffled         1      1000
## 51      shuffled_arb_0001  shuffled         1      1000
## 52      shuffled_azj_0001  shuffled         1      1000
## 53      shuffled_azj_0002  shuffled         1      1000
## 54      shuffled_ben_0001  shuffled         1      1000
## 55      shuffled_bhg_0001  shuffled         1       765
## 56      shuffled_bhg_0002  shuffled         1       272
## 57      shuffled_bhg_0003  shuffled         1       734
## 58      shuffled_bhg_0004  shuffled         1       341
## 59      shuffled_bhg_0005  shuffled         1       182
## 60      shuffled_bhg_0006  shuffled         1      1000
## 61      shuffled_bhg_0007  shuffled         1      1000
## 62      shuffled_bhg_0008  shuffled         1       995
## 63      shuffled_bhg_0009  shuffled         1       213
## 64      shuffled_bhg_0010  shuffled         1       426
## 65      shuffled_bla_0001  shuffled         1      1000
## 66      shuffled_bla_0002  shuffled         1      1000
## 67      shuffled_blt_0001  shuffled         1      1000
```

```
## 68       shuffled_bod_0001  shuffled         1      1000
## 69       shuffled_bos_0001  shuffled         1      1000
## 70       shuffled_bos_0002  shuffled         1      1000
## 71       shuffled_cad_0001  shuffled         1        36
## 72       shuffled_cad_0002  shuffled         1       151
## 73       shuffled_cav_0001  shuffled         1       407
## 74       shuffled_cav_0002  shuffled         1       636
## 75       shuffled_cav_0003  shuffled         1       176
## 76       shuffled_cav_0004  shuffled         1        23
## 77       shuffled_cav_0005  shuffled         1       167
## 78       shuffled_cav_0006  shuffled         1       116
## 79       shuffled_cav_0007  shuffled         1        32
## 80       shuffled_cav_0008  shuffled         1       161
## 81       shuffled_cav_0009  shuffled         1       116
## 82       shuffled_chr_0001  shuffled         1      1000
## 83       shuffled_cmn_0001  shuffled         1      1000
## 84       shuffled_cmn_0002  shuffled         1      1000
## 85       shuffled_cre_0001  shuffled         1       261
## 86       shuffled_cre_0002  shuffled         1      1000
## 87       shuffled_csw_0001  shuffled         1      1000
## 88       shuffled_cth_0001  shuffled         1       287
## 89       shuffled_cth_0002  shuffled         1       871
## 90       shuffled_cth_0003  shuffled         1       391
## 91       shuffled_cth_0004  shuffled         1       779
## 92       shuffled_cth_0005  shuffled         1      1000
## 93       shuffled_cth_0006  shuffled         1       608
## 94       shuffled_cth_0007  shuffled         1       560
## 95       shuffled_cth_0008  shuffled         1      1000
## 96       shuffled_cth_0009  shuffled         1      1000
## 97       shuffled_cth_0010  shuffled         1      1000
## 98       shuffled_cth_0011  shuffled         1       604
## 99       shuffled_div_0001  shuffled         1      1000
## 100      shuffled_dna_0001  shuffled         1      1000
## 101      shuffled_dna_0002  shuffled         1      1000
## 102      shuffled_dna_0003  shuffled         1      1000
## 103      shuffled_dna_0004  shuffled         1      1000
## 104      shuffled_dna_0005  shuffled         1      1000
## 105      shuffled_dna_0006  shuffled         1      1000
## 106      shuffled_dna_0007  shuffled         1      1000
## 107      shuffled_dna_0008  shuffled         1      1000
## 108      shuffled_dna_0009  shuffled         1      1000
## 109      shuffled_dna_0010  shuffled         1      1000
## 110      shuffled_dna_0011  shuffled         1      1000
## 111      shuffled_dna_0012  shuffled         1      1000
## 112      shuffled_dna_0013  shuffled         1      1000
## 113      shuffled_dna_0014  shuffled         1      1000
## 114      shuffled_dna_0015  shuffled         1      1000
## 115      shuffled_dna_0016  shuffled         1      1000
## 116      shuffled_dna_0017  shuffled         1      1000
## 117      shuffled_dna_0018  shuffled         1      1000
## 118      shuffled_dna_0019  shuffled         1      1000
## 119      shuffled_dna_0020  shuffled         1      1000
## 120      shuffled_dna_0021  shuffled         1      1000
## 121      shuffled_dna_0022  shuffled         1      1000
```

```
## 122      shuffled_dna_0023  shuffled         1      1000
## 123      shuffled_dna_0024  shuffled         1      1000
## 124      shuffled_dna_0025  shuffled         1      1000
## 125      shuffled_dna_0026  shuffled         1      1000
## 126      shuffled_dna_0027  shuffled         1      1000
## 127      shuffled_dna_0028  shuffled         1      1000
## 128      shuffled_dna_0029  shuffled         1      1000
## 129      shuffled_dna_0030  shuffled         1      1000
## 130      shuffled_ela_0001  shuffled         1       556
## 131      shuffled_ela_0002  shuffled         1       692
## 132      shuffled_ell_0001  shuffled         1      1000
## 133      shuffled_eng_0001  shuffled         1      1000
## 134      shuffled_epo_0001  shuffled         1      1000
## 135      shuffled_eus_0001  shuffled         1      1000
## 136      shuffled_gaz_0001  shuffled         1      1000
## 137      shuffled_guj_0001  shuffled         1      1000
## 138      shuffled_heb_0001  shuffled         1      1000
## 139      shuffled_hin_0001  shuffled         1      1000
## 140      shuffled_hye_0001  shuffled         1      1000
## 141      shuffled_ibb_0001  shuffled         1      1000
## 142      shuffled_iii_0001  shuffled         1      1000
## 143      shuffled_ike_0001  shuffled         1      1000
## 144      shuffled_jav_0001  shuffled         1      1000
## 145      shuffled_jav_0002  shuffled         1      1000
## 146      shuffled_jpn_0001  shuffled         1      1000
## 147      shuffled_kal_0001  shuffled         1      1000
## 148      shuffled_kan_0001  shuffled         1      1000
## 149      shuffled_kat_0001  shuffled         1      1000
## 150      shuffled_khm_0001  shuffled         1      1000
## 151      shuffled_kkh_0001  shuffled         1      1000
## 152      shuffled_kor_0001  shuffled         1      1000
## 153      shuffled_lao_0001  shuffled         1      1000
## 154      shuffled_lug_0001  shuffled         1      1000
## 155      shuffled_mal_0001  shuffled         1      1000
## 156      shuffled_moc_0001  shuffled         1      1000
## 157      shuffled_moc_0002  shuffled         1      1000
## 158      shuffled_moc_0003  shuffled         1       175
## 159      shuffled_mya_0001  shuffled         1      1000
## 160      shuffled_pan_0001  shuffled         1      1000
## 161      shuffled_pra_0001  shuffled         1      1000
## 162      shuffled_prc_0001  shuffled         1       716
## 163      shuffled_prc_0002  shuffled         1       394
## 164      shuffled_prc_0003  shuffled         1       182
## 165      shuffled_prc_0004  shuffled         1       191
## 166      shuffled_rus_0001  shuffled         1      1000
## 167      shuffled_sin_0001  shuffled         1      1000
## 168      shuffled_sum_0001  shuffled         1      1000
## 169      shuffled_sum_0002  shuffled         1       349
## 170      shuffled_sum_0003  shuffled         1       165
## 171      shuffled_sum_0004  shuffled         1       227
## 172      shuffled_sum_0005  shuffled         1       255
## 173      shuffled_sum_0006  shuffled         1       137
## 174      shuffled_sum_0007  shuffled         1       219
## 175      shuffled_sum_0008  shuffled         1       188
```

```
## 176      shuffled_sum_0009  shuffled         1       567
## 177      shuffled_sum_0010  shuffled         1       226
## 178      shuffled_tam_0001  shuffled         1      1000
## 179      shuffled_tel_0001  shuffled         1      1000
## 180      shuffled_tgl_0001  shuffled         1      1000
## 181      shuffled_tha_0001  shuffled         1      1000
## 182      shuffled_tir_0001  shuffled         1      1000
## 183      shuffled_tsl_0001  shuffled         1       451
## 184      shuffled_vai_0001  shuffled         1      1000
## 185      shuffled_wsy_0001  shuffled         1      1000
## 186      shuffled_zfi_0001  shuffled         1        31
## 187      shuffled_zgh_0001  shuffled         1      1000
## 188      shuffled_zul_0001  shuffled         1      1000
## 189    animal_bhg_0001.txt    animal         1       765
## 190    animal_bhg_0002.txt    animal         1       272
## 191    animal_bhg_0003.txt    animal         1       734
## 192    animal_bhg_0004.txt    animal         1       341
## 193    animal_bhg_0005.txt    animal         1       182
## 194    animal_bhg_0006.txt    animal         1      1598
## 195    animal_bhg_0007.txt    animal         1      1118
## 196    animal_bhg_0008.txt    animal         1       995
## 197    animal_bhg_0009.txt    animal         1       213
## 198    animal_bhg_0010.txt    animal         1       426
## 199    animal_cad_0001.txt    animal         6        36
## 200    animal_cad_0002.txt    animal        17       151
## 201    animal_cav_0001.txt    animal         1       407
## 202    animal_cav_0002.txt    animal         1       636
## 203    animal_cav_0003.txt    animal         1       176
## 204    animal_cav_0004.txt    animal         1        23
## 205    animal_cav_0005.txt    animal         1       167
## 206    animal_cav_0006.txt    animal         1       116
## 207    animal_cav_0007.txt    animal         1        32
## 208    animal_cav_0008.txt    animal         1       161
## 209    animal_cav_0009.txt    animal         1       116
## 210    animal_cth_0001.txt    animal         1       287
## 211    animal_cth_0002.txt    animal         1       871
## 212    animal_cth_0003.txt    animal         1       391
## 213    animal_cth_0004.txt    animal         1       779
## 214    animal_cth_0005.txt    animal         1      1533
## 215    animal_cth_0006.txt    animal         1       608
## 216    animal_cth_0007.txt    animal         1       560
## 217    animal_cth_0008.txt    animal         1      2288
## 218    animal_cth_0009.txt    animal         1      1440
## 219    animal_cth_0010.txt    animal         1      4492
## 220    animal_cth_0011.txt    animal         1       604
## 221    animal_zfi_0001.txt    animal         3        31
## 222     morse_moc_0001.txt     morse         1      1291
## 223     morse_moc_0002.txt     morse         3      2257
## 224     morse_moc_0003.txt     morse         1       175
## 225   natural_dna_0001.txt   natural       157     10958
## 226   natural_dna_0002.txt   natural       209     14621
## 227   natural_dna_0003.txt   natural       713     49910
## 228   natural_dna_0004.txt   natural       713     49910
## 229   natural_dna_0005.txt   natural       713     49910
```

```
## 230   natural_dna_0006.txt   natural       713     49910
## 231   natural_dna_0007.txt   natural       713     49910
## 232   natural_dna_0008.txt   natural       713     49910
## 233   natural_dna_0009.txt   natural       713     49910
## 234   natural_dna_0010.txt   natural       713     49910
## 235   natural_dna_0011.txt   natural       713     49910
## 236   natural_dna_0012.txt   natural       713     49910
## 237   natural_dna_0013.txt   natural       713     49910
## 238   natural_dna_0014.txt   natural       713     49910
## 239   natural_dna_0015.txt   natural       713     49910
## 240   natural_dna_0016.txt   natural       713     49910
## 241   natural_dna_0017.txt   natural       713     49910
## 242   natural_dna_0018.txt   natural       713     49910
## 243   natural_dna_0019.txt   natural       713     49910
## 244   natural_dna_0020.txt   natural       713     49910
## 245   natural_dna_0021.txt   natural       713     49910
## 246   natural_dna_0022.txt   natural       713     49910
## 247   natural_dna_0023.txt   natural       713     49910
## 248   natural_dna_0024.txt   natural       713     49910
## 249   natural_dna_0025.txt   natural       713     49910
## 250   natural_dna_0026.txt   natural       713     49910
## 251   natural_dna_0027.txt   natural       713     49910
## 252   natural_dna_0028.txt   natural       713     49910
## 253   natural_dna_0029.txt   natural       713     49910
## 254   natural_dna_0030.txt   natural       713     49910
## 255  signlang_tsl_0001.txt  signlang        14       479
## 256   weather_wsy_0001.txt   weather        26      3987
## 257   ancient_akk_0001.txt   ancient        10      2698
## 258   ancient_akk_0002.txt   ancient         8      1603
## 259   ancient_cre_0001.txt   ancient        19       270
## 260   ancient_cre_0002.txt   ancient        48      1099
## 261   ancient_ela_0001.txt   ancient        30       578
## 262   ancient_ela_0002.txt   ancient        59       692
## 263   ancient_pra_0001.txt   ancient         4      1101
## 264   ancient_prc_0001.txt   ancient        46       848
## 265   ancient_prc_0002.txt   ancient        24       448
## 266   ancient_prc_0003.txt   ancient        18       210
## 267   ancient_prc_0004.txt   ancient        16       217
## 268   ancient_sum_0001.txt   ancient        70      1560
## 269   ancient_sum_0002.txt   ancient        30       361
## 270   ancient_sum_0003.txt   ancient         9       171
## 271   ancient_sum_0004.txt   ancient        18       237
## 272   ancient_sum_0005.txt   ancient         9       275
## 273   ancient_sum_0006.txt   ancient        11       137
## 274   ancient_sum_0007.txt   ancient        10       227
## 275   ancient_sum_0008.txt   ancient         9       192
## 276   ancient_sum_0009.txt   ancient        19       579
## 277   ancient_sum_0010.txt   ancient        10       228
## 278 heraldics_bla_0001.txt heraldics       503     44198
## 279 heraldics_bla_0002.txt heraldics       207     13932
## 280   writing_aii_0001.txt   writing        88      6471
## 281   writing_arb_0001.txt   writing        90      7629
## 282   writing_azj_0001.txt   writing        90     10749
## 283   writing_azj_0002.txt   writing        90     10745
```

```
## 284   writing_ben_0001.txt   writing        94      9718
## 285   writing_blt_0001.txt   writing        89      8804
## 286   writing_bod_0001.txt   writing        90     12649
## 287   writing_bos_0001.txt   writing        90      9694
## 288   writing_bos_0002.txt   writing        90      9871
## 289   writing_chr_0001.txt   writing        89      8985
## 290   writing_cmn_0001.txt   writing        89      2960
## 291   writing_cmn_0002.txt   writing        90      2789
## 292   writing_csw_0001.txt   writing        67      6513
## 293   writing_div_0001.txt   writing        88     18129
## 294   writing_ell_0001.txt   writing        90     12392
## 295   writing_eng_0001.txt   writing        91     10606
## 296   writing_epo_0001.txt   writing        91      9884
## 297   writing_eus_0001.txt   writing        93     10907
## 298   writing_gaz_0001.txt   writing        92     10473
## 299   writing_guj_0001.txt   writing        91      9959
## 300   writing_heb_0001.txt   writing        89      7261
## 301   writing_hin_0001.txt   writing        90     10497
## 302   writing_hye_0001.txt   writing        92     11119
## 303   writing_ibb_0001.txt   writing       100     13205
## 304   writing_iii_0001.txt   writing        88      3272
## 305   writing_ike_0001.txt   writing        67      8541
## 306   writing_jav_0001.txt   writing        92     10816
## 307   writing_jav_0002.txt   writing        92     13651
## 308   writing_jpn_0001.txt   writing        89      4157
## 309   writing_kal_0001.txt   writing        90     16786
## 310   writing_kan_0001.txt   writing        89     10534
## 311   writing_kat_0001.txt   writing        90     11828
## 312   writing_khm_0001.txt   writing        90     10652
## 313   writing_kkh_0001.txt   writing        82      9938
## 314   writing_kor_0001.txt   writing        91      4709
## 315   writing_lao_0001.txt   writing        90     10502
## 316   writing_lug_0001.txt   writing        87     10354
## 317   writing_mal_0001.txt   writing        82     10557
## 318   writing_mya_0001.txt   writing        89     15131
## 319   writing_pan_0001.txt   writing        90     10681
## 320   writing_rus_0001.txt   writing        90     11684
## 321   writing_sin_0001.txt   writing        90     10543
## 322   writing_tam_0001.txt   writing        89     13133
## 323   writing_tel_0001.txt   writing        89     11075
## 324   writing_tgl_0001.txt   writing        94     12246
## 325   writing_tha_0001.txt   writing        89      9278
## 326   writing_tir_0001.txt   writing        89      6640
## 327   writing_vai_0001.txt   writing        91      8556
## 328   writing_zgh_0001.txt   writing        89      7770
## 329   writing_zul_0001.txt   writing        91     10217
```

Density plot

Density plot for numbers of characters per data file and subcorpus.

```
# select subcorpora (if applicable)
# selection <- c("writing", "shuffled")
```

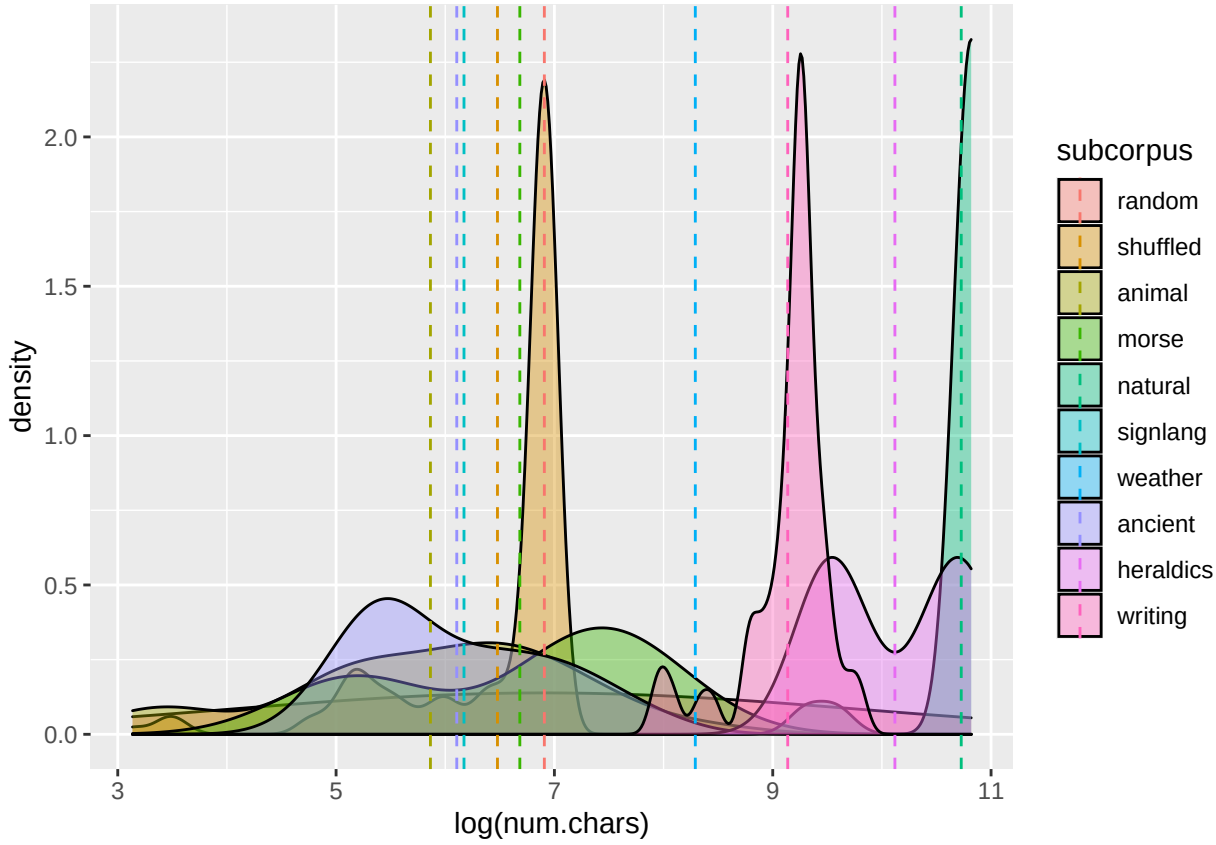
```
#simpleStats.df <- simpleStats.df[simpleStats.df$subcorpus %in% selection, ]
# get means per subcorpus
mu <- ddply(simpleStats.df, "subcorpus", summarise, grp.mean = mean(log(num.chars)))
# plot densities with mean values as vertical lines
density.plot <- ggplot(simpleStats.df, aes(x = log(num.chars), fill = subcorpus)) +
  geom_density(alpha = 0.4) +
  geom_vline(data = mu, aes(xintercept = grp.mean, color = subcorpus),
             linetype = "dashed")
print(density.plot)
```

```
## Warning: Groups with fewer than two data points have been dropped.

## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf
```



Safe figure to file

```
ggsave("~/Github/NaLaFi/figures/simpleStats_density.pdf", density.plot, dpi = 300,
       scale = 1, device = cairo_pdf)
```

```
## Saving 8 x 4 in image
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf
```