Simple Corpus Stats

Chris Bentz

February 28, 2020

Load libraries

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

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

List files

List all the files in the directory "corpus".

```r
file.list <- list.files(path = "/home/chris/Github/StringBase/corpus/original",
                        recursive = T, full.names = T)
#print(file.list)
length(file.list)
```

```
## [1] 191
```

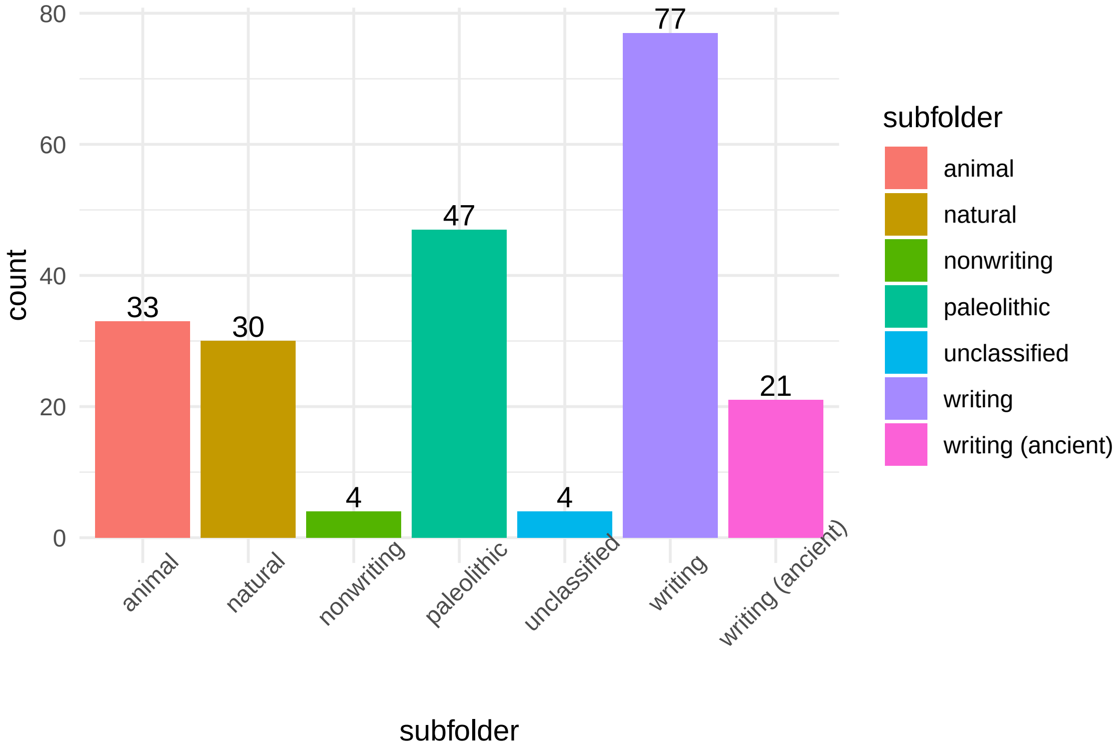
Count number of files in subfolders

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

```r
#number of "animal" files
animal.count <- length(file.list[grepl("animal", file.list)])
#number of "natural" files
natural.count <- length(file.list[grepl("natural", file.list)])
#number of "nonwriting" files
nonwriting.count <- length(file.list[grepl("nonwriting", file.list)])
#number of "unclassified"" files
unclassified.count <- length(file.list[grepl("unclassified", file.list)])
#number of "paleolithic" files
paleolithic.count <- length(file.list[grepl("paleolithic", file.list)])
#number of "writing" files
writing.count <- length(file.list[grepl("writing", file.list)])
#number of "ancient" files
ancient.count <- length(file.list[grepl("ancient", file.list)])

#create data frame
df <- data.frame(subfolder = c("animal", "natural", "nonwriting", "unclassified", "paleolithic",
                               "writing", "writing (ancient)"),
                 count = c(animal.count, natural.count, nonwriting.count,
                           unclassified.count, paleolithic.count, writing.count, ancient.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("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
```

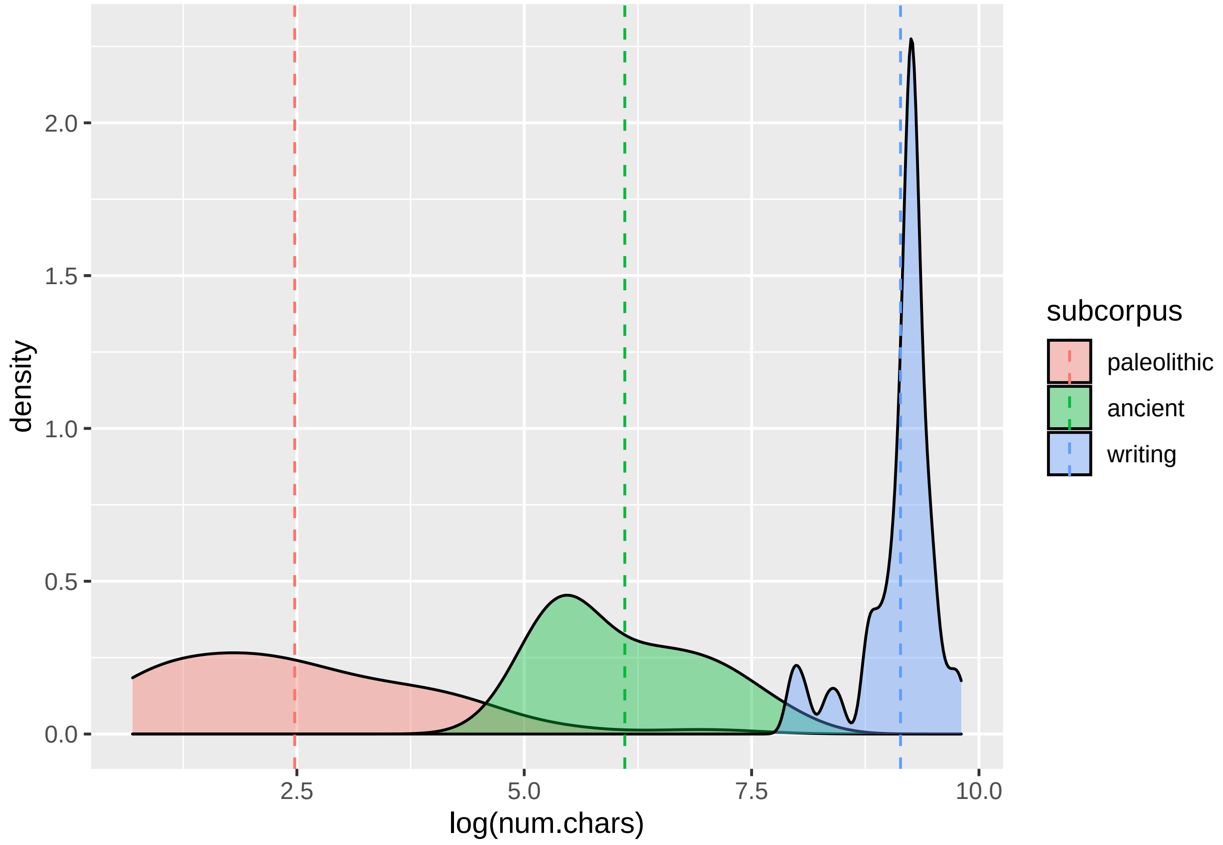
```r
simpleStats.df <- data.frame(filename = character(0), subcorpus = character(0),
                             num.lines = numeric(0), num.chars = numeric(0))

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         animal_bhg_0001.txt       animal         1       765
## 2         animal_bhg_0002.txt       animal         1       272
## 3         animal_bhg_0003.txt       animal         1       734
## 4         animal_bhg_0004.txt       animal         1       341
## 5         animal_bhg_0005.txt       animal         1       182
## 6         animal_bhg_0006.txt       animal         1      1598
## 7         animal_bhg_0007.txt       animal         1      1118
## 8         animal_bhg_0008.txt       animal         1       995
## 9         animal_bhg_0009.txt       animal         1       213
## 10        animal_bhg_0010.txt       animal         1       426
## 11        animal_cad_0001.txt       animal         6        36
## 12        animal_cad_0002.txt       animal        17       151
## 13        animal_cav_0001.txt       animal         1       407
## 14        animal_cav_0002.txt       animal         1       636
## 15        animal_cav_0003.txt       animal         1       176
## 16        animal_cav_0004.txt       animal         1        23
## 17        animal_cav_0005.txt       animal         1       167
## 18        animal_cav_0006.txt       animal         1       116
## 19        animal_cav_0007.txt       animal         1        32
## 20        animal_cav_0008.txt       animal         1       161
## 21        animal_cav_0009.txt       animal         1       116
## 22        animal_cth_0001.txt       animal         1       287
## 23        animal_cth_0002.txt       animal         1       871
## 24        animal_cth_0003.txt       animal         1       391
## 25        animal_cth_0004.txt       animal         1       779
## 26        animal_cth_0005.txt       animal         1      1533
## 27        animal_cth_0006.txt       animal         1       608
## 28        animal_cth_0007.txt       animal         1       560
## 29        animal_cth_0008.txt       animal         1      2288
## 30        animal_cth_0009.txt       animal         1      1440
## 31        animal_cth_0010.txt       animal         1      4492
## 32        animal_cth_0011.txt       animal         1       604
## 33        animal_zfi_0001.txt       animal         3        31
## 34       natural_dna_0001.txt      natural       157     10958
## 35       natural_dna_0002.txt      natural       209     14621
## 36       natural_dna_0003.txt      natural       713     49910
## 37       natural_dna_0004.txt      natural       713     49910
## 38       natural_dna_0005.txt      natural       713     49910
## 39       natural_dna_0006.txt      natural       713     49910
## 40       natural_dna_0007.txt      natural       713     49910
## 41       natural_dna_0008.txt      natural       713     49910
## 42       natural_dna_0009.txt      natural       713     49910
## 43       natural_dna_0010.txt      natural       713     49910
## 44       natural_dna_0011.txt      natural       713     49910
## 45       natural_dna_0012.txt      natural       713     49910
## 46       natural_dna_0013.txt      natural       713     49910
## 47       natural_dna_0014.txt      natural       713     49910
## 48       natural_dna_0015.txt      natural       713     49910
## 49       natural_dna_0016.txt      natural       713     49910
## 50       natural_dna_0017.txt      natural       713     49910
## 51       natural_dna_0018.txt      natural       713     49910
## 52       natural_dna_0019.txt      natural       713     49910
## 53       natural_dna_0020.txt      natural       713     49910
## 54       natural_dna_0021.txt      natural       713     49910
## 55       natural_dna_0022.txt      natural       713     49910
## 56       natural_dna_0023.txt      natural       713     49910
## 57       natural_dna_0024.txt      natural       713     49910
## 58       natural_dna_0025.txt      natural       713     49910
## 59       natural_dna_0026.txt      natural       713     49910
## 60       natural_dna_0027.txt      natural       713     49910
## 61       natural_dna_0028.txt      natural       713     49910
## 62       natural_dna_0029.txt      natural       713     49910
## 63       natural_dna_0030.txt      natural       713     49910
## 64    nonwriting_moc_0001.txt   nonwriting         1      1291
## 65    nonwriting_moc_0002.txt   nonwriting         3      2257
## 66    nonwriting_tsl_0001.txt   nonwriting        14       479
## 67    nonwriting_wsy_0001.txt   nonwriting        26      3987
## 68   paleolithic_aur_0001.txt  paleolithic         2        14
## 69   paleolithic_bla_0001.txt  paleolithic         2        84
## 70   paleolithic_bla_0002.txt  paleolithic         2       110
## 71   paleolithic_bla_0003.txt  paleolithic         2        15
## 72   paleolithic_bla_0004.txt  paleolithic         2        16
## 73   paleolithic_bla_0006.txt  paleolithic         2        44
## 74   paleolithic_bla_0007.txt  paleolithic         2        77
## 75   paleolithic_bla_0008.txt  paleolithic         2         4
## 76   paleolithic_bla_0009.txt  paleolithic         2         3
## 77   paleolithic_bla_0010.txt  paleolithic         2         2
## 78   paleolithic_bla_0011.txt  paleolithic         2         4
## 79   paleolithic_bla_0013.txt  paleolithic         2         2
## 80   paleolithic_bla_0014.txt  paleolithic         2        29
## 81   paleolithic_brs_0003.txt  paleolithic         2         6
## 82   paleolithic_brs_0004.txt  paleolithic         2         9
## 83   paleolithic_cas_0001.txt  paleolithic         2        31
## 84   paleolithic_cas_0003.txt  paleolithic         2         8
## 85   paleolithic_cas_0006.txt  paleolithic         2         5
## 86   paleolithic_cas_0007.txt  paleolithic         2         3
## 87   paleolithic_cas_0008.txt  paleolithic         2         2
## 88   paleolithic_cel_0002.txt  paleolithic         2         6
## 89   paleolithic_cel_0003.txt  paleolithic         2         2
## 90   paleolithic_gdr_0001.txt  paleolithic         2         7
## 91   paleolithic_gdr_0002.txt  paleolithic         2         8
## 92   paleolithic_gdr_0003.txt  paleolithic         2        25
## 93   paleolithic_hfc_0012.txt  paleolithic         2         3
## 94   paleolithic_hss_0002.txt  paleolithic         2         8
## 95   paleolithic_laf_0004.txt  paleolithic         2        13
## 96   paleolithic_laf_0005.txt  paleolithic         2        14
## 97   paleolithic_laf_0007.txt  paleolithic         2        11
## 98   paleolithic_laf_0008.txt  paleolithic         2         5
## 99   paleolithic_laf_0011.txt  paleolithic         2        43
## 100  paleolithic_laf_0012.txt  paleolithic         2         7
## 101  paleolithic_laf_0013.txt  paleolithic         2         2
## 102  paleolithic_laf_0014.txt  paleolithic         2         2
## 103  paleolithic_laf_0015.txt  paleolithic         2         2
## 104  paleolithic_laf_0016.txt  paleolithic         2        27
## 105  paleolithic_laf_0018.txt  paleolithic         2        30
## 106  paleolithic_laf_0019.txt  paleolithic         2        62
## 107  paleolithic_lar_0001.txt  paleolithic         2         4
## 108  paleolithic_lar_0002.txt  paleolithic         2       232
## 109  paleolithic_pal_0001.txt  paleolithic        46      1097
## 110  paleolithic_tdc_0001.txt  paleolithic         2        69
## 111  paleolithic_tri_0001.txt  paleolithic         2         4
## 112  paleolithic_vhc_0090.txt  paleolithic         2        11
## 113  paleolithic_vhc_0161.txt  paleolithic         2        62
## 114  paleolithic_vhc_0165.txt  paleolithic         2        16
## 115 unclassified_pic_0001.txt unclassified       249      5343
## 116 unclassified_rrg_0001.txt unclassified        16      6149
## 117 unclassified_rrg_0002.txt unclassified        22      4478
## 118 unclassified_voy_0001.txt unclassified      5472    223769
## 119      ancient_akk_0001.txt      ancient        10      2698
## 120      ancient_akk_0002.txt      ancient         8      1603
## 121      ancient_cre_0001.txt      ancient        19       270
## 122      ancient_cre_0002.txt      ancient        48      1099
## 123      ancient_ela_0001.txt      ancient        30       578
## 124      ancient_ela_0002.txt      ancient        59       692
## 125      ancient_pra_0001.txt      ancient         4      1101
## 126      ancient_prc_0001.txt      ancient        46       848
## 127      ancient_prc_0002.txt      ancient        24       448
## 128      ancient_prc_0003.txt      ancient        18       210
## 129      ancient_prc_0004.txt      ancient        16       217
## 130      ancient_sum_0001.txt      ancient        70      1560
## 131      ancient_sum_0002.txt      ancient        30       361
## 132      ancient_sum_0003.txt      ancient         9       171
## 133      ancient_sum_0004.txt      ancient        18       237
## 134      ancient_sum_0005.txt      ancient         9       275
## 135      ancient_sum_0006.txt      ancient        11       137
## 136      ancient_sum_0007.txt      ancient        10       227
## 137      ancient_sum_0008.txt      ancient         9       192
## 138      ancient_sum_0009.txt      ancient        19       579
## 139      ancient_sum_0010.txt      ancient        10       228
## 140    heraldics_bla_0001.txt    heraldics       503     44198
## 141    heraldics_bla_0002.txt    heraldics       207     13932
## 142      writing_aii_0001.txt      writing        88      6471
## 143      writing_arb_0001.txt      writing        90      7629
## 144      writing_azj_0001.txt      writing        90     10749
## 145      writing_azj_0002.txt      writing        90     10745
## 146      writing_ben_0001.txt      writing        94      9718
## 147      writing_blt_0001.txt      writing        89      8804
## 148      writing_bod_0001.txt      writing        90     12649
## 149      writing_bos_0001.txt      writing        90      9694
## 150      writing_bos_0002.txt      writing        90      9871
## 151      writing_chr_0001.txt      writing        89      8985
## 152      writing_cmn_0001.txt      writing        89      2960
## 153      writing_cmn_0002.txt      writing        90      2789
## 154      writing_csw_0001.txt      writing        67      6513
## 155      writing_div_0001.txt      writing        88     18129
## 156      writing_ell_0001.txt      writing        90     12392
## 157      writing_eng_0001.txt      writing        91     10606
## 158      writing_epo_0001.txt      writing        91      9884
## 159      writing_eus_0001.txt      writing        93     10907
## 160      writing_gaz_0001.txt      writing        92     10473
## 161      writing_guj_0001.txt      writing        91      9959
## 162      writing_heb_0001.txt      writing        89      7261
## 163      writing_hin_0001.txt      writing        90     10497
## 164      writing_hye_0001.txt      writing        92     11119
## 165      writing_ibb_0001.txt      writing       100     13205
## 166      writing_iii_0001.txt      writing        88      3272
## 167      writing_ike_0001.txt      writing        67      8541
## 168      writing_jav_0001.txt      writing        92     10816
## 169      writing_jav_0002.txt      writing        92     13651
## 170      writing_jpn_0001.txt      writing        89      4157
## 171      writing_kal_0001.txt      writing        90     16786
## 172      writing_kan_0001.txt      writing        89     10534
## 173      writing_kat_0001.txt      writing        90     11828
## 174      writing_khm_0001.txt      writing        90     10652
## 175      writing_kkh_0001.txt      writing        82      9938
## 176      writing_kor_0001.txt      writing        91      4709
## 177      writing_lao_0001.txt      writing        90     10502
## 178      writing_lug_0001.txt      writing        87     10354
## 179      writing_mal_0001.txt      writing        82     10557
## 180      writing_mya_0001.txt      writing        89     15131
## 181      writing_pan_0001.txt      writing        90     10681
## 182      writing_rus_0001.txt      writing        90     11684
## 183      writing_sin_0001.txt      writing        90     10543
## 184      writing_tam_0001.txt      writing        89     13133
## 185      writing_tel_0001.txt      writing        89     11075
## 186      writing_tgl_0001.txt      writing        94     12246
## 187      writing_tha_0001.txt      writing        89      9278
## 188      writing_tir_0001.txt      writing        89      6640
## 189      writing_vai_0001.txt      writing        91      8556
## 190      writing_zgh_0001.txt      writing        89      7770
## 191      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("ancient", "paleolithic", "writing")
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)
```



Safe figure to file

```
ggsave("Figures/simpleStats_density.pdf", density.plot, dpi = 300,
       scale = 1, device = cairo_pdf)
```

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