Reading in the Princeton Corpus of Political Emails V1.0 with R

Carsten Schwemmer

5/12/2021

In this document we share R code to read in the Princeton Corpus of Political Emails v1.0 via three different frameworks: `Base R`, `data.table` and `readr` (from the Tidyverse).

We also measure the time it takes to read in the data in order to compare performance across frameworks.

Please make sure to set your working directory to the location of the unzipped dataset before proceeding.

Libraries

Loading the libraries we'll use:

```
library(readr)
library(dplyr)
library(data.table)
```

Base R

```
ptm <- proc.time()

df_base <- read.csv("corpus_v1.0.csv",
                    encoding = "UTF-8")

rtm <- proc.time() - ptm
rtm
```

```
##    user  system elapsed
##    9.28    0.46    9.75
```

```
glimpse(df_base)
```

```
## Rows: 317,366
## Columns: 21
## $ from_name         <chr> "Kathleen Williams", "Goal Update -- via Team Kathle~
## $ from_address      <chr> "info@kathleenformontana.com", "info@kathleenformont~
## $ subject           <chr> "Welcome to the team!", "Quickly closing in", "Welco~
## $ body_text         <chr> "Thanks for joining the team! My name is Kathleen Wi~
## $ name              <chr> "Kathleen Williams", "Kathleen Williams", "Kathleen ~
## $ office_sought     <chr> "U.S. House Montana At-large District", "U.S. House ~
```

```
## $ party_affiliation <chr> "Democratic Party", "Democratic Party", "Democratic ~
## $ office_level      <chr> "Federal", "Federal", "Federal", "Federal", "Federal~
## $ district_type     <chr> "Congress", "Congress", "Congress", "Congress", "Con~
## $ final_website     <chr> "https://kathleenformontana.com/", "https://kathleen~
## $ crawl_date        <chr> "12-02-2019", "12-02-2019", "12-02-2019", "12-02-201~
## $ source            <chr> "ballotpedia-campaign", "ballotpedia-campaign", "bal~
## $ state             <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ~
## $ type              <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ~
## $ subtype           <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ~
## $ date              <chr> "2019-12-05", "2020-06-25", "2019-12-03", "2020-01-2~
## $ hour              <int> 19, 14, 19, 12, 17, 17, 14, 11, 17, 22, 18, 17, 17, ~
## $ day               <chr> "Thu", "Thu", "Tue", "Mon", "Mon", "Fri", "Sat", "Fr~
## $ uid_email         <chr> "7182e4e604717330ecaf2699be61b200", "00768081c0a2487~
## $ uid_inbox         <chr> "08f13962c5b2090e32c902552a4ff634", "08f13962c5b2090~
## $ incumbent         <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No"~
```

data.table

```
ptm <- proc.time()

df_dt <- fread("corpus_v1.0.csv",
                     encoding = "UTF-8")

rtm <- proc.time() - ptm
rtm
```

```
##    user  system elapsed
##    2.33    0.14    2.47
```

```
glimpse(df_dt)
```

```
## Rows: 317,366
## Columns: 21
## $ from_name         <chr> "Kathleen Williams", "Goal Update -- via Team Kathle~
## $ from_address      <chr> "info@kathleenformontana.com", "info@kathleenformont~
## $ subject           <chr> "Welcome to the team!", "Quickly closing in", "Welco~
## $ body_text         <chr> "Thanks for joining the team! My name is Kathleen Wi~
## $ name              <chr> "Kathleen Williams", "Kathleen Williams", "Kathleen ~
## $ office_sought     <chr> "U.S. House Montana At-large District", "U.S. House ~
## $ party_affiliation <chr> "Democratic Party", "Democratic Party", "Democratic ~
## $ office_level      <chr> "Federal", "Federal", "Federal", "Federal", "Federal~
## $ district_type     <chr> "Congress", "Congress", "Congress", "Congress", "Con~
## $ final_website     <chr> "https://kathleenformontana.com/", "https://kathleen~
## $ crawl_date        <chr> "12-02-2019", "12-02-2019", "12-02-2019", "12-02-201~
## $ source            <chr> "ballotpedia-campaign", "ballotpedia-campaign", "bal~
## $ state             <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ~
## $ type              <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ~
## $ subtype           <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ~
## $ date              <date> 2019-12-05, 2020-06-25, 2019-12-03, 2020-01-20, 202~
## $ hour              <int> 19, 14, 19, 12, 17, 17, 14, 11, 17, 22, 18, 17, 17, ~
## $ day               <chr> "Thu", "Thu", "Tue", "Mon", "Mon", "Fri", "Sat", "Fr~
```

```
## $ uid_email         <chr> "7182e4e604717330ecaf2699be61b200", "00768081c0a2487~
## $ uid_inbox         <chr> "08f13962c5b2090e32c902552a4ff634", "08f13962c5b2090~
## $ incumbent         <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No"~
```

Readr (Tidyverse)

```
ptm <- proc.time()

df_tidy <- read_csv("corpus_v1.0.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   state = col_logical(),
##   type = col_logical(),
##   subtype = col_logical(),
##   date = col_date(format = "")
## )
## i Use `spec()` for the full column specifications.

## Warning: 197010 parsing failures.
##    row     col           expected                       actual              file
## 156623 type    1/0/T/F/TRUE/FALSE Hybrid PAC                   'corpus_v1.0.csv'
## 156623 subtype 1/0/T/F/TRUE/FALSE Republican/Conservative PAC 'corpus_v1.0.csv'
## 156624 type    1/0/T/F/TRUE/FALSE Hybrid PAC                   'corpus_v1.0.csv'
## 156624 subtype 1/0/T/F/TRUE/FALSE Republican/Conservative PAC 'corpus_v1.0.csv'
## 156625 type    1/0/T/F/TRUE/FALSE Hybrid PAC                   'corpus_v1.0.csv'
## ...... ....... .................. ............................ .................
## See problems(...) for more details.
```

```
rtm <- proc.time() - ptm
rtm
```

```
##    user  system elapsed
##    3.66    0.16    3.92
```

We have a problem here. The `read_csv()` guess for 1000 rows of missings is "logical": https://github.com/tidyverse/readr/issues/839

Solving the problem by increasing the number of cases to be used for inferring column types:

```
df_tidy <- read_csv("corpus_v1.0.csv",
                    guess_max = 200000) # increase n for guessing types
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   date = col_date(format = "")
## )
## i Use `spec()` for the full column specifications.
```

```
glimpse(df_tidy)
```

```
## Rows: 317,366
## Columns: 21
## $ from_name         <chr> "Kathleen Williams", "Goal Update -- via Team Kathle~
## $ from_address      <chr> "info@kathleenformontana.com", "info@kathleenformont~
## $ subject           <chr> "Welcome to the team!", "Quickly closing in", "Welco~
## $ body_text         <chr> "Thanks for joining the team! My name is Kathleen Wi~
## $ name              <chr> "Kathleen Williams", "Kathleen Williams", "Kathleen ~
## $ office_sought     <chr> "U.S. House Montana At-large District", "U.S. House ~
## $ party_affiliation <chr> "Democratic Party", "Democratic Party", "Democratic ~
## $ office_level      <chr> "Federal", "Federal", "Federal", "Federal", "Federal~
## $ district_type     <chr> "Congress", "Congress", "Congress", "Congress", "Con~
## $ final_website     <chr> "https://kathleenformontana.com/", "https://kathleen~
## $ crawl_date        <chr> "12-02-2019", "12-02-2019", "12-02-2019", "12-02-201~
## $ source            <chr> "ballotpedia-campaign", "ballotpedia-campaign", "bal~
## $ state             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ type              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ subtype           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ date              <date> 2019-12-05, 2020-06-25, 2019-12-03, 2020-01-20, 202~
## $ hour              <chr> "19", "14", "19", "12", "17", "17", "14", "11", "17"~
## $ day               <chr> "Thu", "Thu", "Tue", "Mon", "Mon", "Fri", "Sat", "Fr~
## $ uid_email         <chr> "7182e4e604717330ecaf2699be61b200", "00768081c0a2487~
## $ uid_inbox         <chr> "08f13962c5b2090e32c902552a4ff634", "08f13962c5b2090~
## $ incumbent         <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No"~
```

For reference, this is the R setup used to create this document:

```
sessionInfo()
```

```
## R version 4.0.3 (2020-10-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] data.table_1.14.0 dplyr_1.0.4       readr_1.4.0
##
## loaded via a namespace (and not attached):
##  [1] rstudioapi_0.13   knitr_1.31        magrittr_2.0.1    hms_1.0.0
##  [5] tidyselect_1.1.0  R6_2.5.0          rlang_0.4.10      fansi_0.4.2
```

```
##  [9] stringr_1.4.0     tools_4.0.3       xfun_0.21         utf8_1.1.4
## [13] cli_2.3.1         DBI_1.1.1         htmltools_0.5.1.1 ellipsis_0.3.1
## [17] assertthat_0.2.1  yaml_2.2.1        digest_0.6.27     tibble_3.0.6
## [21] lifecycle_1.0.0   crayon_1.4.1      purrr_0.3.4       vctrs_0.3.6
## [25] glue_1.4.2        evaluate_0.14     rmarkdown_2.7     stringi_1.5.3
## [29] compiler_4.0.3    pillar_1.5.0      generics_0.1.0    pkgconfig_2.0.3
```