Comparing Datasets

Sofia Bliss-Carrascosa

9/12/2022

```
library(tidyverse)
library(knitr)
```

V2: TESTING THE R-VERSION COMPARED TO ORIGINAL VERSION Loading R-version

```
v0_all_us_unsorted <- read_csv("allUS_unsorted.csv")
```

```
v1_all_us_polclaims <- read_csv("all_us_pol_claims_updated_sep15.csv")
```

```
v2_5parties <- v1_all_us_polclaims %>%
  filter(claimant_party == "Republican"|
           claimant_party == "Democratic"|
         claimant_party == "Independent"|
         claimant_party == "Libertarian"|
         claimant_party == "unknown_affiliation")

v2_5parties %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Publisher")
```

Table 1: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	167
checkyourfact.com	10
factcheck.org	1290
factcheck.thedispatch.com	58
newsweek.com	66
nytimes.com	464
politifact.com	4275
polygraph.info	3
poynter.org	10
thegazette.com	7
usatoday.com	19
vox.com	2
washingtonpost.com	1252

```
dim(v2_5parties)
```

```
## [1] 7623   13
```

Loading Grace Original Version

```
grace_v2 <- read_csv("grace_v2_CSVversion.csv")

grace_v2 %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Publisher")
```

Table 2: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	167
checkyourfact.com	10
factcheck.org	1290
factcheck.thedispatch.com	58
newsweek.com	66
nytimes.com	464
politifact.com	4275
polygraph.info	3
poynter.org	10
thegazette.com	7
usatoday.com	19
vox.com	2
washingtonpost.com	1252

```
dim(grace_v2)
```

```
## [1] 7623   13
```

Data counts work out!

```
anti_join(v2_5parties, grace_v2)
```

```
## # A tibble: 0 x 13
## # ... with 13 variables: ...1 <dbl>, url <chr>, title <chr>,
## #   textualRating <chr>, languageCode <chr>, publisher.name <chr>,
## #   publisher.site <chr>, reviewDate <dttm>, text <chr>, claimant <chr>,
## #   claimDate <dttm>, claimant_party <chr>, reason <chr>
```

ALL MATCH!! WOOT!!

V3: TESTING THE R-VERSION COMPARED TO ORIGINAL VERSION Loading R-version

```
v3_deduped <- v2_5parties %>%
  distinct(url, text, .keep_all = TRUE)

v3_deduped %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Publisher")
```

Table 3: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	157
checkyourfact.com	10
factcheck.org	1237
factcheck.thedispatch.com	58
newsweek.com	63
nytimes.com	444
politifact.com	4176
polygraph.info	2
poynter.org	9
thegazette.com	7
usatoday.com	17
vox.com	1
washingtonpost.com	1179

```
dim(v3_deduped)
```

```
## [1] 7360   13
```

Loading Grace version

```
grace_v3 <- read_csv("grace_v3_deduped_CSVversion.csv")

grace_v3 %>%
  count(publisher.site) %>%
  kable(caption = "All Claims in 5 Parties
        Sorted by Publisher")
```

Table 4: All Claims in 5 Parties Sorted by Publisher

publisher.site	n
cbsnews.com	157
checkyourfact.com	10
factcheck.org	1237
factcheck.thedispatch.com	58
newsweek.com	63
nytimes.com	444
politifact.com	4176
polygraph.info	2

publisher.site	n
poynter.org	9
thegazette.com	7
usatoday.com	17
vox.com	1
washingtonpost.com	1179

```
dim(grace_v3)
```

```
## [1] 7360   13
```

Data Counts are not the same: why?

```
anti_join(v3_deduped, grace_v3)
```

```
## # A tibble: 0 x 13
## # ... with 13 variables: ...1 <dbl>, url <chr>, title <chr>,
## #   textualRating <chr>, languageCode <chr>, publisher.name <chr>,
## #   publisher.site <chr>, reviewDate <dttm>, text <chr>, claimant <chr>,
## #   claimDate <dttm>, claimant_party <chr>, reason <chr>
```

```
identified_dupes <- anti_join(v2_5parties, v3_deduped)

dim(identified_dupes)
```

```
## [1] 263  13
```

LOADING V4 – Manual removal by ASA

```
grace_v4 <- read_csv("grace_V4_dejunkedCSVversion.csv")
```

```
manual_removes <- anti_join(v3_deduped, grace_v4)
print(manual_removes)
```

```
## # A tibble: 23 x 13
##      ...1 url     title  textualRating languageCode publisher.name publisher.site
##     <dbl> <chr>   <chr>  <chr>         <chr>        <chr>          <chr>
##  1 11803 http:/~ FACT ~ FALSE         en           Check Your Fa~ checkyourfact~
##  2 15634 https:~ Fact ~ Mostly true   en           Newsweek       newsweek.com
##  3  1096 http:/~ Did G~ Half True     en           PolitiFact     politifact.com
##  4  1941 http:/~ Amend~ FALSE         en           PolitiFact     politifact.com
##  5  6863 http:/~ Does ~ Half True     en           PolitiFact     politifact.com
##  6  2515 http:/~ Fact-~ No evidence   en           PolitiFact     politifact.com
##  7  5758 http:/~ ’Song~ Pants on Fire en           PolitiFact     politifact.com
##  8  1445 http:/~ Was i~ Mostly False  en           PolitiFact     politifact.com
##  9  3473 https:~ Wisco~ Half True     en           PolitiFact     politifact.com
## 10  7065 https:~ No; W~ Mostly False  en           PolitiFact     politifact.com
## # ... with 13 more rows, and 6 more variables: reviewDate <dttm>, text <chr>,
## #   claimant <chr>, claimDate <dttm>, claimant_party <chr>, reason <chr>
```

MANUALLY CREATING V5

```r
v4_dejunked <- v3_deduped %>%
  filter(claimant != "Donald Trump For Prison") %>%
  filter(claimant !=  "Lauren Boebert; Rudy Giuliani") %>%
  filter(claimant != "Rick Scott's Starbuck's heckler") %>%
  filter(claimant != "Americans United for Change") %>%
  filter(claimant !=  "Consumers for Smart Solar") %>%
  filter(claimant !=  "Greg Gianforte's campaign") %>%
  filter(claimant !=  "Vietnam Veterans Against John McCain") %>%
  filter(claimant != "President Trump's lawyers") %>%
  filter(claimant !=    "Robin Vos and Scott Fitzgerald") %>%
  filter(claimant !=  "Robin Vos; Scott Fitzgerald") %>%
  filter(claimant !=  "Donald Trump 2020 Voters") %>%
  filter(claimant !=  "Michael Bloomberg; Joe Biden; Hillary Clinton; Adam Schiff") %>%
  filter(claimant !=  "Donald Trump and Mike Pence") %>%
  filter(claimant !=  "Bill DeBlasio and Brian Kemp") %>%
  filter(claimant !=  "John Roberts; Fox News correspondent") %>%
  filter(claimant !=  "The Trump campaign") %>%
  filter(claimant !=  "Sen. Ted Cruz (R-Texas) and Rep. Mark Meadows (R-N.C.)") %>%
  filter(claimant !=  "Keith Ellison spokesman") %>%
  filter(claimant !=  "Donald Trump ad") %>%
  filter(claimant !=  "Kamala Harris for the People") %>%
  filter(claimant !=  "Don Bolduc campaign") %>%
  filter(claimant !=  "Donald Trump campaign") %>%
  filter(claimant !=  "Joe Biden campaign")
```

CREATING V5 code

```r
v5_pf <- v4_dejunked %>%
  filter(publisher.site == "politifact.com")

v5_wapo <- v4_dejunked %>%
  filter(publisher.site == "washingtonpost.com")

v5_fc <- v4_dejunked %>%
  filter(publisher.site == "factcheck.org")

v5_nyt <- v4_dejunked %>%
  filter(publisher.site == "nytimes.com")
```

V6: step 1 Testing for Trumps

```r
elim_trump <- v4_dejunked %>%
  filter(publisher.site == "politifact.com" |
           publisher.site ==  "washingtonpost.com" |
           publisher.site == "factcheck.org" |
           publisher.site == "nytimes.com")

elim_trump %>%
  filter(grepl('Trump|trump', claimant)) %>%
  group_by(claimant) %>%
  count()
```

```
## # A tibble: 5 x 2
## # Groups:   claimant [5]
##   claimant                      n
##   <chr>                     <int>
## 1 Donald J. Trump             205
## 2 Donald trump                  1
## 3 Donald Trump               2587
## 4 Ivanka Trump                 10
## 5 President Donald J. Trump     2
```