Column Modification - "Country"

Charles J. Gomez, Harshvardhan Singh

2023-06-17

The country column in our dataset contains a more detailed information about the authors' affiliations, including first, middle, and last authors. However, this brings a level of complexity and potential confusion for machine learning models. Therefore, we propose a method to simplify the country column by retaining only the first and last author affiliations, ensuring a more straightforward representation of authorship.

```
library(stringr)

data = read.csv("INPUT_SQL_Text_Data_Astronomy_and_Astrophysics.csv")
data_CountryColumn <- data$country

#We removed all instances of middle authors' country codes. This step involved replacing the pattern "\

data$country <- gsub("\\w+\\+middle", "", data$country)
data$country <- trimws(data$country)
print(head(data$country, 50))
```

```
##  [1] "US+last US+first US+last US+first"
##  [2] "US+first"
##  [3] "US+first"
##  [4] "IN+first US+last"
##  [5] "US+last       US+first                US+last          US+first"
##  [6] "US+first"
##  [7] "US+first  RS+last"
##  [8] "US+last DE+first"
##  [9] "CH+first CH+first"
## [10] "US+first  US+first  US+last US+last"
## [11] "GB+first   GB+first US+last  GB+first  US+last GB+first"
## [12] "CN+first   CN+last CN+first CN+last"
## [13] "US+first"
## [14] "US+first US+last"
## [15] "US+last US+last  US+first    US+first    US+first  US+last US+last US+first"
## [16] "FR+last US+first"
## [17] "IT+first IT+first IT+first     IT+first"
## [18] "US+first US+first  US+last US+last"
## [19] "RU+last RU+first GB+first"
## [20] "US+first US+last US+first US+last"
## [21] "US+last US+first CA+first US+first US+last CA+first"
## [22] "US+last    US+first US+first        US+last"
## [23] "US+first US+first"
## [24] "US+first    US+last"
## [25] "US+last  GB+first                                 US+last                                 GB+first"
## [26] "GB+first NL+last GB+first NL+last"
## [27] "AU+last AU+last   AU+first  AU+first"
```

```
## [28] "US+last GB+first US+last GB+first"
## [29] "US+first US+last   US+last US+first"
## [30] "GB+last GB+first US+first GB+first  US+first GB+last"
## [31] "GB+first GB+last  IT+first  GB+last   GB+first IT+first"
## [32] "GB+last GB+last NL+first  NL+first"
## [33] "US+first"
## [34] "AU+first AU+first  AU+last  AU+last"
## [35] "GB+last"
## [36] "DK+last            US+first US+first     DK+last"
## [37] "CA+first SE+last"
## [38] "US+first US+last    US+first   US+last"
## [39] "US+last   US+last US+first    US+first"
## [40] "FR+last US+first"
## [41] "AU+first AU+first AU+first AU+first"
## [42] "US+last US+first"
## [43] "CL+first US+last"
## [44] "BR+first  BR+last BR+first BR+last"
## [45] "US+first US+first"
## [46] "AU+last   US+first    AU+last   US+first"
## [47] "IT+first  US+first US+last     US+last US+first   IT+first"
## [48] "US+last"
## [49] "US+first   US+last"
## [50] "SE+first  SE+last"
```

```r
#We removed the "+first" and "+last" strings from the remaining country codes using another regular exp

data$country <- gsub("\\+last|\\+first", "", data$country)
# Remove any extra spaces resulting from the removal
data$country <- sapply(strsplit(data$country, "\\s+"), function(x) paste(x[x != ""], collapse = " "))

print(head(data$country, 50))
```

```
##  [1] "US US US US"             "US"
##  [3] "US"                      "IN US"
##  [5] "US US US US"             "US"
##  [7] "US RS"                   "US DE"
##  [9] "CH CH"                   "US US US US"
## [11] "GB GB US GB US GB"       "CN CN CN CN"
## [13] "US"                      "US US"
## [15] "US US US US US US US US" "FR US"
## [17] "IT IT IT IT"             "US US US US"
## [19] "RU RU GB"                "US US US US"
## [21] "US US CA US US CA"       "US US US US"
## [23] "US US"                   "US US"
## [25] "US GB US GB"             "GB NL GB NL"
## [27] "AU AU AU AU"             "US GB US GB"
## [29] "US US US US"             "GB GB US GB US GB"
## [31] "GB GB IT GB GB IT"       "GB GB NL NL"
## [33] "US"                      "AU AU AU AU"
## [35] "GB"                      "DK US US DK"
## [37] "CA SE"                   "US US US US"
## [39] "US US US US"             "FR US"
## [41] "AU AU AU AU"             "US US"
## [43] "CL US"                   "BR BR BR BR"
## [45] "US US"                   "AU US AU US"
```

```
## [47] "IT US US US US IT"       "US"
## [49] "US US"                   "SE SE"
```

```r
# Function to calculate the percentage count of each country code in a vector.
calculate_percentage <- function(vec) {
  counts <- table(vec)
  percentages <- prop.table(counts) * 100
  formatted <- paste0(round(percentages, 1), "%", names(percentages))
  paste(formatted, collapse = " ")
}

# Apply the calculate_percentage function to each row in the 'country' column
data$country <- sapply(strsplit(data$country, "\\s+"), calculate_percentage)

# Print the modified 'country' column (first 20)
print(head(data$country, 50))
```

```
##  [1] "100%US"          "100%US"          "100%US"          "50%IN 50%US"
##  [5] "100%US"          "100%US"          "50%RS 50%US"     "50%DE 50%US"
##  [9] "100%CH"          "100%US"          "66.7%GB 33.3%US" "100%CN"
## [13] "100%US"          "100%US"          "100%US"          "50%FR 50%US"
## [17] "100%IT"          "100%US"          "33.3%GB 66.7%RU" "100%US"
## [21] "33.3%CA 66.7%US" "100%US"          "100%US"          "100%US"
## [25] "50%GB 50%US"     "50%GB 50%NL"     "100%AU"          "50%GB 50%US"
## [29] "100%US"          "66.7%GB 33.3%US" "66.7%GB 33.3%IT" "50%GB 50%NL"
## [33] "100%US"          "100%AU"          "100%GB"          "50%DK 50%US"
## [37] "50%CA 50%SE"     "100%US"          "100%US"          "50%FR 50%US"
## [41] "100%AU"          "100%US"          "50%CL 50%US"     "100%BR"
## [45] "100%US"          "50%AU 50%US"     "33.3%IT 66.7%US" "100%US"
## [49] "100%US"          "100%SE"
```

```r
#print first 50 rows of the dataset
#print(head(data, 50))
```