Package 'genderBR'

October 13, 2022

Type Package

Title Predict Gender from Brazilian First Names

Version 1.1.2

Description A method to predict and report gender from Brazilian first names using the Brazilian Institute of Geography and Statistics' Census data (<https://censo2010.ibge.gov.br/nomes/>).

License GPL (>= 2)

Depends R (>= 3.1.2)

Imports dplyr (>= 0.5.0), jsonlite, httr, magrittr, purrr, tibble

Encoding UTF-8

URL https://github.com/meirelesff/genderBR

BugReports https://github.com/meirelesff/genderBR/issues

RoxygenNote 7.1.1

Suggests testthat (>= 3.0.0), covr

Config/testthat/edition 3

NeedsCompilation no

Author Fernando Meireles [aut, cre]

Maintainer Fernando Meireles <fmeireles@ufmg.br>

Repository CRAN

Date/Publication 2021-05-02 14:20:07 UTC

R topics documented:

get_gender *Predict gender from Brazilian first names*

Description

get_gender uses the IBGE's 2010 Census data to predict gender from Brazilian first names. In particular, the function exploits data on the number of females and males with the same name in Brazil, or in a given Brazilian state, to calculate the proportion of females using it.

The function classifies a name as *male* or *female* only when that proportion is higher than a given threshold (e.g., `female if proportion > 0.9, the default`, or `male if proportion < 0.1`); proportions below this threshold are classified as missings (NA). The method is based on the gender functionality developed by Lincon Mullen in: Mullen (2016). gender: Predict Gender from Names Using Historical Data.

Multiple names can be passed to the function call. To speed the calculations, the package aggregates equal first names to make fewer requests to the IBGE's API. Also, the package contains an internal dataset with all the names reported by the IBGE to make faster classifications – although this option does not support getting results by State.

Usage

```
get_gender(
  names,
  state = NULL,
  prob = FALSE,
  threshold = 0.9,
  internal = TRUE,
  encoding = "ASCII//TRANSLIT"
)
```

Arguments

names	A character vector specifying a person's first name. Names can also be passed to the function as a full name (e.g., Ana Maria de Souza). get_gender is case insensitive. In addition, multiple names can be passed in the same function call.
state	A string with the state of federation abbreviation (e.g., RJ for Rio de Janeiro). If state is set to a value different from NULL, the internal argument is ignored.
prob	Report the proportion of female uses of the name? Defaults to FALSE.
threshold	Numeric indicating the threshold used in predictions. Defaults to 0.9.
internal	Use internal data to predict gender? Allowing this option makes the function faster, but it does not support getting results by State. Defaults to TRUE.
encoding	Encoding used to read Brazilian names and stip accents. Defaults to ASCII//TRANSLIT.

Value

`get_gender` may returns three different values: `Female`, if the name provided is female; `Male`, if the name provided is male; or `NA`, if we can not predict gender from the name given the chosen threshold.

If the `prob` argument is set to `TRUE`, then the function returns the proportion of females uses of the provided name.

Data

Information on the Brazilian first names uses by gender was collect in the 2010 Census (Censo Demografico de 2010, in Portuguese), in July of that year, by the Instituto Brasileiro de Demografia e Estatistica (IBGE). The surveyed population includes 190,8 million Brazilians living in all 27 states. According to the IBGE, there are more than 130,000 unique first names in this population.

Note

Names with different spell (e.g., Ana and Anna, or Marcos and Markos) are considered different names. In addition, only names with more than 20 occurrences, or more than 15 occurrences in a given state, are included in the IBGE's data.

Also note that UTF-8 special characters, common in Portuguese words and names, are not supported by the IBGE's API. Users are encouraged to convert strings to ASCII (it is also possible to set the `encoding` argument to a different value).

References

For more information on the IBGE's data, please check (in Portuguese): https://censo2010.ibge.gov.br/nomes/

See Also

`map_gender`

Examples

```
#' # Use get_gender to predict the gender of a person based on her/his first name
get_gender('MARIA DA SILVA SANTOS')
get_gender('joao')

# To change the employed threshold
get_gender('ariel', threshold = 0.8)

# Or to get the proportion of females
# with the name provided
get_gender('iris', prob = TRUE)

# Multiple names can be predict at the same time
get_gender(c('joao', 'ana', 'benedita', 'rafael'))

## Not run:
```

```
# In different states (using API data, must have internet connection)
get_gender(rep('Ana', 3), c('sp', 'am', 'rs'))

## End(Not run)
```

`get_states` *State's abbreviations*

Description

Use this function to get a `data.frame` with the full names, abbreviations (acronym), and IBGE codes of all Brazilian states.

Usage

```
get_states()
```

Value

A `tbl_df`, `tbl`, `data.frame` with two variables: `state`, `abb`, and `code`.

`map_gender` *Map the use of Brazilian first names by gender and by state*

Description

`map_gender` retrieves data on the number of male or female uses of a given first name by state from the Instituto Brasileiro de Geografia e Estatistica's 2010 Census API.

Usage

```
map_gender(name, gender = NULL, encoding = "ASCII//TRANSLIT")
```

Arguments

`name`	A string with a Brazilian first name. The name can also be passed to the function as a full name (e.g., Ana Maria de Souza). `get_gender` is case insensitive.
`gender`	A string with the gender to look for. Valid inputs are `m`, for males, `f`, for females, and `NULL`, in which case the function returns results for all persons with a given name.
`encoding`	Encoding used to read Brazilian names and stip accents. Defaults to `ASCII//TRANSLIT`.

Details

Information on the gender associated with Brazilian first names was collect in the 2010 Census (Censo Demografico de 2010, in Portuguese), in July of that year, by the Instituto Brasileiro de Demografia e Estatistica (IBGE). The surveyed population includes 190,8 million Brazilians living in all 27 states. According to the IBGE, there are more than 130,000 unique first names in this population.

Value

`get_gender` returns a `tbl_df, tbl, data.frame` with the following variables:

- `nome` State's name.
- `uf` State's abbreviation.
- `freq` Total number of persons with the name provided.
- `populacao` State's total population.
- `sexo` Same as the `sexo` argument provided.
- `prop` Persons with the name and gender provided per 100,000 inhabitants.

Note

Names with different spell (e.g., Ana and Anna, or Marcos and Markos) are considered different names. Additionally, only names with more than 20 occurrences, or more than 15 occurrences in a given state, are considered.

References

For more information on the IBGE's data, please check (in Portuguese): https://censo2010.ibge.gov.br/nomes/

See Also

`get_gender`

Examples

```
## Not run:
# Map the use of the name 'Maria'
map_gender('maria')

# The function accepts full names
map_gender('Maria da Silva Santos')

# Or names in uppercase
map_gender('MARIA DA SILVA SANTOS')

# Select desired gender
map_gender('AUGUSTO ROBERTO', gender = 'm')
map_gender('John da Silva', gender = 'm')

## End(Not run)
```

Index