Package 'genderBR'

July 13, 2026

Type Package

Title Predict Gender from Brazilian First Names

Version 1.4.0

Description A generalized method to predict and report gender from Brazilian first names using the Brazilian Institute of Geography and Statistics' Census data and neural networks.

License GPL (>= 2)

Depends R (>= 4.1.0)

Imports data.table, jsonlite, httr, purrr

Encoding UTF-8

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

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

RoxygenNote 7.3.3

Suggests testthat (>= 3.1.7), covr, httr2, luz, torch (>= 0.13.0)

Config/testthat/edition 3

NeedsCompilation no

Author Fernando Meireles [aut, cre] (ORCID: <https://orcid.org/0000-0002-7027-2058>)

Maintainer Fernando Meireles <fernando.meireles@iesp.uerj.br>

Repository CRAN

Date/Publication 2026-07-12 23:10:07 UTC

Contents

clear_nn_cache 2
download_gender_model 2
get_gender 3
get_gender_nn 6
get_states 7
map_gender 8

Index **10**

`clear_nn_cache` *Clear the neural network in-memory cache*

Description

Removes the model and vocabulary metadata from the in-memory session cache. The next call to `get_gender_nn` will reload them from the on-disk cache (no re-download needed if the files are already cached).

Usage

```
clear_nn_cache()
```

Value

Invisible `NULL`.

Examples

```
## Not run:
clear_nn_cache()

## End(Not run)
```

`download_gender_model` *Download neural network model files*

Description

Downloads the pre-trained model weights and vocabulary from Hugging Face to a local cache directory. This is required before using `get_gender_nn`.

Usage

```
download_gender_model()
```

Details

Files are stored in `tools::R_user_dir("genderBR", "cache")` and only downloaded if not already present.

Value

Invisible character vector with the paths to the downloaded files.

Examples

```
## Not run:
download_gender_model()

## End(Not run)
```

`get_gender` *Predict gender from Brazilian first names*

Description

`get_gender` uses the IBGE's Census data to predict gender from Brazilian first names (2010 by default, optionally 2022). 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 *female* when that proportion is higher than the female threshold (`proportion > 0.9`, by default), and as *male* when it is lower than one minus the male threshold (`proportion < 0.1`, by default); proportions in between are classified as `Unknown`. The two thresholds can be set independently (see the `threshold` argument). 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 (2010 and 2022), 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",
  year = 2022,
  nn = FALSE,
  nn_size = NULL,
  device = NULL
)
```

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. A single value sets the same threshold for both sexes; a vector with two values sets one threshold per sex, the first for females and the second for males (e.g., c(0.9, 0.8)). The two values can also be named, in any order (e.g., c(Female = 0.9, Male = 0.8) or c(F = 0.9, M = 0.8)). Because a name cannot be female and male at the same time, the two thresholds must sum to at least 1.
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	(Deprecated) Previously used to strip accents via iconv. Accents are now removed with a platform-independent method and this argument is ignored. It will be removed in a future version.
year	Census year used in the prediction. Supported values are 2010 and 2022 (default).
nn	Logical. If TRUE, use a character-level neural network model to predict gender instead of the IBGE Census data. This allows the function to generalise to names not present in the IBGE dataset. When nn = TRUE, the state, internal, and year arguments are ignored. Model files must be downloaded first with download_gender_model. Defaults to FALSE.
nn_size	Batch size for neural network inference, used only when nn = TRUE. When NULL (the default), all names are classified in a single pass; set it to a positive integer to split a large input vector into batches and avoid out-of-memory crashes. See get_gender_nn.
device	Device used for neural network inference, used only when nn = TRUE. When NULL (the default), the CPU is used; set it to "cuda" or "mps" to run on a GPU. See get_gender_nn.

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.

When year = 2022, the function queries the IBGE names API with 2022 data or uses the 2022 internal dataset when internal = TRUE and state is NULL.

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 set a different threshold for each sex (female first)
get_gender('marion', threshold = c(0.7, 0.9))
get_gender('marion', threshold = c(Female = 0.7, Male = 0.9))

# 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_gender_nn *Predict gender from Brazilian first names using a neural network*

Description

get_gender_nn uses a 2-layer bidirectional GRU neural network with attention pooling to predict gender from Brazilian first names. Unlike get_gender, this function can generalise to names not present in the IBGE census dataset.

Usage

```
get_gender_nn(
  names,
  prob = FALSE,
  threshold = 0.9,
  nn_size = NULL,
  device = NULL,
  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_nn is case insensitive.
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. A single value sets the same threshold for both sexes; a vector with two values sets one threshold per sex, the first for females and the second for males (e.g., c(0.9, 0.8)). The two values can also be named, in any order (e.g., c(Female = 0.9, Male = 0.8) or c(F = 0.9, M = 0.8)). Because a name cannot be female and male at the same time, the two thresholds must sum to at least 1.
nn_size	Batch size for neural network inference. When NULL (the default), all names are classified at once. Set it to an integer to split a large input vector of first names into batches of that size, which avoids out-of-memory crashes when analyzing large datasets.
device	Device used for inference. When NULL (the default), the CPU is used. Set it to "cuda" or "mps" to run on a GPU (note that Apple Silicon's MPS shares system memory, so nn_size is what prevents memory crashes there).
encoding	(Deprecated) Previously used to strip accents via iconv. Accents are now removed with a platform-independent method and this argument is ignored. It will be removed in a future version.

Details

Model weights and vocabulary must be downloaded before first use with download_gender_model. If the files are not found in an interactive session, you will be prompted to download them. Subsequent calls within the same session use an in-memory cache.

Value

get_gender_nn may return 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.

See Also

get_gender, download_gender_model

Examples

```
## Not run:
get_gender_nn("Maria")
get_gender_nn(c("Maria", "Joao"), prob = TRUE)
get_gender_nn("Ana Maria de Souza")

## 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	(Deprecated) Previously used to strip accents via iconv. Accents are now removed with a platform-independent method and this argument is ignored. It will be removed in a future version.

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

clear_nn_cache, 2

download_gender_model, 2, *4*, *7*

get_gender, 3, *6*, *7*, *9*
get_gender_nn, *2*, *4*, 6
get_states, 7

map_gender, *5*, 8