Package ‘franc’

October 13, 2022

Title  Detect the Language of Text
Version  1.1.4
Author  Gabor Csardi, Titus Wormer, Maciej Ceglowski, Jacob R. Rideout,
        and Kent S. Johnson
Maintainer  Gábor Csárdi <csardi.gabor@gmail.com>
Description  With no external dependencies and
             support for 335 languages; all languages spoken by
             more than one million speakers. ‘Franc’ is a port
             of the ‘JavaScript’ project of the same name,
             see <https://github.com/wooorm/franc>.
License  MIT + file LICENSE
URL  https://github.com/gaborcsardi/franc#readme
BugReports  https://github.com/gaborcsardi/franc/issues
Suggests  testthat
RoxygenNote  6.1.1
Encoding  UTF-8
Imports  jsonlite
NeedsCompilation  no
Repository  CRAN
Date/Publication  2021-10-18 08:50:17 UTC

R topics documented:

franc .................................................. 2
franc_all .............................................. 3
speakers .............................................. 4

Index  5
franc

Detect the language of a string

Description

Detect the language of a string

Usage

franc(text, min_speakers = 1e+06, whitelist = NULL, blacklist = NULL, min_length = 10, max_length = 2048)

Arguments

text A string constant. Should be at least min_length characters long, this is 10 characters by default. Only the first max_length characters are used (2048 by default), to make the detection reasonably fast.

min_speakers Languages with at least this many speakers are checked. By default this is one million. Set it to zero to include all languages known by franc. See also speakers.

whitelist List of three letter language codes to check against.

blacklist List of three letter language codes not to check against.

min_length Minimum number of characters required in the text.

max_length Maximum number of characters used from the text. By default only the first 2048 characters are used.

Value

A three letter ISO-639-3 language code, the detected language of the text. "und" is returned for too short input.

See Also

franc_all for scores against many languages, speakers.

Examples

## afr
franc("Alle menslike wesens word vry")

## nno
franc("Alle mennesker er født frie og")

## Too short, und
franc("the")

## You can change what's too short (default: 10), sco
franc("the", min_length = 3)
Description

Returns the scores for all languages that use the same script as the input text, in decreasing order of probability. The score is calculated from the distances of the trigram distributions in the input text and in the language model. The closer the languages, the higher the score. Scores are scaled, so that the closest language will have a score of 1.

Usage

\[
\text{franc\_all(text, min\_speakers = 1e+06, whitelist = NULL, blacklist = NULL, min\_length = 10, max\_length = 2048)}
\]

Arguments

- **text**: A string constant. Should be at least \text{min\_length} characters long, this is 10 characters by default. Only the first \text{max\_length} characters are used (2048 by default), to make the detection reasonably fast.
- **min\_speakers**: Languages with at least this many speakers are checked. By default this is one million. Set it to zero to include all languages known by franc. See also \text{speakers}.
- **whitelist**: List of three letter language codes to check against.
- **blacklist**: List of three letter language codes not to check against.
- **min\_length**: Minimum number of characters required in the text.
- **max\_length**: Maximum number of characters used from the text. By default only the first 2048 characters are used.

Value

A data frame with columns \text{language} and \text{score}. The \text{language} column contains the three letter ISO-639-3 language codes. The \text{score} column contains the scores.

See Also

- \text{franc} if you only want the top result, \text{speakers}.

Examples

\[
\text{head(franc\_all("O Brasil caiu 26 posições"))}
\]

## Provide a whitelist:
\[
\text{franc\_all("O Brasil caiu 26 posições",}
\]
\[
\text{whitelist = c("por", "src", "glg", "spa"))}
\]

## Provide a blacklist:
head(franc_all("O Brasil caiu 26 posições",
blacklist = c("src", "glg", "lav")))

<table>
<thead>
<tr>
<th>speakers</th>
<th>Number of speakers for 370 languages</th>
</tr>
</thead>
</table>

**Description**

This is a superset of all languages detected by franc. Numbers were collected by Titus Wormer. To quote him: *Painstakingly crawled by hand from OHCHR, the numbers are (in some cases, very) rough estimates or out-of-date.*

**Usage**

speakers

**Format**

A data frame with columns:

- **language**  Three letter language code.
- **speakers**  Number of speakers.
- **name**  Full name of language.
Index

* datasets
  speakers, 4
franc, 2, 3
franc_all, 2, 3
speakers, 2, 3, 4