

Search Terms Examples

Regex Pattern Description	Matching Group	Example Match
\bl'm (not)?({nationalities_pattern})\b	Nationalities	"I'm Brazilian", "I'm not American"
\bl'm (not)?({nationalities_pattern})\b	Nationalities	"I'm Canadian", "I'm not French"
\b being ?({nationalities_pattern})\b	Nationalities	"being Australian", "being British"
\bl am (not)?({nationalities_pattern})\b	Nationalities	"I am German", "I am not Japanese"
\bl'm (not)?from ({countries_pattern})\b	Countries	"I'm from Brazil", "I'm not from India"
\bl'm (not)?from ({countries_pattern})\b	Countries	"I'm from China", "I'm not from Canada"
\bl am (not)?from ({countries_pattern})\b	Countries	"I am from the US", "I am not from Italy"
\blive in ({countries_pattern})\b	Countries	"live in Germany", "live in Australia"
\bnative language is ({languages_pattern})\b	Languages	"native language is English"
\bnative language isn't ({languages_pattern})\b	Languages	"native language isn't Spanish"
\bnative language isn't ({languages_pattern})\b	Languages	"native language isn't French"
\bnative language is not ({languages_pattern})\b	Languages	"native language is not German"
\b({languages_pattern}) is my native language\b	Languages	"Portuguese is my native language"
\b({languages_pattern}) isn't my native language\b	Languages	"Japanese isn't my native language"
\b({languages_pattern}) isn't my native language\b	Languages	"Mandarin isn't my native language"
\b({languages_pattern}) is not my native language\b	Languages	"Korean is not my native language"
\bspeak ({languages_pattern})\b	Languages	"speak Arabic", "speak Hindi"
\bmain language is ({languages_pattern})\b	Languages	"main language is Tamil"
\bmain language isn't ({languages_pattern})\b	Languages	"main language isn't Russian"
\bmain language isn't ({languages_pattern})\b	Languages	"main language isn't Turkish"
\bmain language is not ({languages_pattern})\b	Languages	"main language is not Swahili"

\b({languages_pattern}) is my main language\b	Languages	"Urdu is my main language"
\b({languages_pattern}) isn't my main language\b	Languages	"Punjabi isn't my main language"
\b({languages_pattern}) isn't my main language\b	Languages	"Bengali isn't my main language"
\b({languages_pattern}) is not my main language\b	Languages	"Zulu is not my main language"

nationalities = [
 "Afghan", "Albanian", "Algerian", "American", "Andorran", "Angolan",
 "Antiguan", "Argentine", "Armenian", "Australian", "Austrian", "Azerbaijani",
 "Bahamian", "Bahraini", "Bangladeshi", "Barbadian", "Barbudan", "Batswana",
 "Belarusian", "Belgian", "Belizean", "Beninese", "Bhutanese", "Bolivian",
 "Bosnian", "Brazilian", "British", "Bruneian", "Bulgarian", "Burkinabé",
 "Burmese", "Burundian", "Cabo Verdean", "Cambodian", "Cameroonian",
 "Canadian",
 "Central African", "Chadian", "Chilean", "Chinese", "Colombian", "Comoran",
 "Congolese", "Costa Rican",
 "Croatian", "Cuban", "Cypriot", "Czech", "Danish", "Djiboutian",
 "Dominican",
 "Dutch", "East Timorese", "Ecuadorean", "Egyptian", "Emirati", "Equatorial
 Guinean",
 "Eritrean", "Estonian", "Ethiopian", "Fijian", "Finnish", "French",
 "Gabonese", "Gambian", "Georgian", "German", "Ghanaian", "Greek",
 "Grenadian", "Guatemalan", "Guinean", "Guinea-Bissauan", "Guyanese",
 "Haitian",
 "Honduran", "Hungarian", "Icelandic", "Indian", "Indonesian", "Iranian",
 "Iraqi", "Irish", "Israeli", "Italian", "Ivorian", "Jamaican",
 "Japanese", "Jordanian", "Kazakh", "Kenyan", "Kiribati", "Kittian and
 Nevisian",
 "Kuwaiti", "Kyrgyz", "Laotian", "Latvian", "Lebanese", "Liberian",
 "Libyan", "Liechtensteiner", "Lithuanian", "Luxembourgish", "Macedonian",
 "Malagasy",
 "Malawian", "Malaysian", "Maldivian", "Malian", "Maltese", "Marshallese",
 "Mauritanian", "Mauritian", "Mexican", "Micronesian", "Moldovan",
 "Monacan",
 "Mongolian", "Montenegrin", "Moroccan", "Mozambican", "Namibian",
 "Nauruan",
 "Nepalese", "New Zealander", "Nicaraguan", "Nigerien", "Nigerian", "North
 Korean",
 "Norwegian", "Omani", "Pakistani", "Palauan", "Palestinian", "Panamanian",
 "Papua New Guinean", "Paraguayan", "Peruvian", "Polish", "Portuguese",
 "Qatari",
 "Romanian", "Russian", "Rwandan", "Saint Lucian", "Salvadoran", "Samoan",
 "San Marinese", "Sao Tomean", "Saudi", "Senegalese", "Serbian",
 "Seychellois",
 "Sierra Leonean", "Singaporean", "Slovak", "Slovenian", "Solomon Islander",
 "Somali",
 "South African", "South Korean", "South Sudanese", "Spanish", "Sri Lankan",
 "Sudanese",

"Surinamese", "Swazi", "Swedish", "Swiss", "Syrian", "Taiwanese",
"Tajik", "Tanzanian", "Thai", "Togolese", "Tongan", "Trinidadian or
Tobagonian",
"Tunisian", "Turkish", "Turkmen", "Tuvaluan", "Ugandan", "Ukrainian",
"Uruguayan", "Uzbek", "Vanuatuan", "Venezuelan", "Vietnamese",
"Vincentian",
"Yemeni", "Zambian", "Zimbabwean"

]

countries = [
"Afghanistan", "Albania", "Algeria", "Andorra", "Angola", "Antigua and
Barbuda",
"Argentina", "Armenia", "Australia", "Austria", "Azerbaijan", "Bahamas",
"Bahrain", "Bangladesh", "Barbados", "Belarus", "Belgium", "Belize",
"Benin", "Bhutan", "Bolivia", "Bosnia and Herzegovina", "Botswana",
"Brazil",
"Brunei", "Bulgaria", "Burkina Faso", "Burundi", "Cabo Verde", "Cambodia",
"Cameroon", "Canada", "Central African Republic", "Chad", "Chile", "China",
"Colombia", "Comoros", "Congo, Democratic Republic of the", "Congo,
Republic of the",
"Costa Rica", "Croatia", "Cuba", "Cyprus", "Czech Republic", "Denmark",
"Djibouti", "Dominica", "Dominican Republic", "East Timor", "Ecuador",
"Egypt",
"El Salvador", "Equatorial Guinea", "Eritrea", "Estonia", "Eswatini",
"Ethiopia",
"Fiji", "Finland", "France", "Gabon", "Gambia", "Georgia",
"Germany", "Ghana", "Greece", "Grenada", "Guatemala", "Guinea",
"Guinea-Bissau", "Guyana", "Haiti", "Honduras", "Hungary", "Iceland",
"India", "Indonesia", "Iran", "Iraq", "Ireland", "Israel",
"Italy", "Ivory Coast", "Jamaica", "Japan", "Jordan", "Kazakhstan",
"Kenya", "Kiribati", "Korea, North", "Korea, South", "Kosovo", "Kuwait",
"Kyrgyzstan", "Laos", "Latvia", "Lebanon", "Lesotho", "Liberia",
"Libya", "Liechtenstein", "Lithuania", "Luxembourg", "Madagascar",
"Malawi",
"Malaysia", "Maldives", "Mali", "Malta", "Marshall Islands", "Mauritania",
"Mauritius", "Mexico", "Micronesia", "Moldova", "Monaco", "Mongolia",
"Montenegro", "Morocco", "Mozambique", "Myanmar (Burma)", "Namibia",
"Nauru",
"Nepal", "Netherlands", "New Zealand", "Nicaragua", "Niger", "Nigeria",
"North Macedonia", "Norway", "Oman", "Pakistan", "Palau", "Palestine",
"Panama", "Papua New Guinea", "Paraguay", "Peru", "Philippines", "Poland",
"Portugal", "Qatar", "Romania", "Russia", "Rwanda", "Saint Kitts and Nevis",
"Saint Lucia", "Saint Vincent and the Grenadines", "Samoa", "San Marino",
"Sao Tome and Principe", "Saudi Arabia",
"Senegal", "Serbia", "Seychelles", "Sierra Leone", "Singapore", "Slovakia",
"Slovenia", "Solomon Islands", "Somalia", "South Africa", "South Sudan",
"Spain",
"Sri Lanka", "Sudan", "Suriname", "Sweden", "Switzerland", "Syria",
"Taiwan", "Tajikistan", "Tanzania", "Thailand", "Togo", "Tonga",
"Trinidad and Tobago", "Tunisia", "Turkey", "Turkmenistan", "Tuvalu",
"Uganda",

```
    "Ukraine", "United Arab Emirates", "United Kingdom", "United States",  
    "Uruguay", "Uzbekistan",  
    "Vanuatu", "Vatican City", "Venezuela", "Vietnam", "Yemen", "Zambia",  
    "Zimbabwe"  
]
```

```
languages = [  
    "Afrikaans", "Akan", "Albanian", "Amharic", "Arabic", "Armenian",  
    "Assamese", "Aymara",  
    "Azerbaijani", "Balochi", "Bamanankan", "Bashkir", "Basque", "Belarusian",  
    "Bengali",  
    "Bhojpuri", "Bislama", "Bosnian", "Brahui", "Bulgarian", "Burmese",  
    "Cantonese",  
    "Catalan", "Cebuano", "Chechen", "Cherokee", "Croatian", "Czech", "Danish",  
    "Dari",  
    "Dholuo", "Dutch", "Dzongkha", "English", "Esperanto", "Estonian", "Ewe",  
    "Finnish",  
    "Flemish", "French", "Fulfulde", "Gaelic", "Galician",  
    "Ganda", "Georgian", "German", "Gikuyu", "Greek", "Guarani", "Gujarati",  
    "Haitian Creole",  
    "Hausa", "Hawaiian", "Hebrew", "Hindi", "Hmong", "Hungarian", "Icelandic",  
    "Igbo",  
    "Ilocano", "Indonesian", "Inuit/Inuktitut", "Italian", "Japanese", "Jarai",  
    "Javanese",  
    "Kiche", "Kannada", "Kashmiri", "Kazakh", "Khmer", "Kinyarwanda",  
    "Kirundi", "Korean",  
    "Kurdish", "Kyrgyz", "Lao", "Latvian", "Lingala", "Lithuanian", "Luganda",  
    "Luxembourgish",  
    "Macedonian", "Maithili", "Malagasy", "Malay", "Malayalam", "Maltese",  
    "Māori",  
    "Marathi", "Mende", "Mongolian", "Nahuatl", "Nauruan", "Navajo", "Nepali",  
    "Norwegian",  
    "Ojibwa", "Oriya", "Oromo", "Pashto", "Persian", "Polish", "Portuguese",  
    "Punjabi", "Quechua", "Romani", "Romanian", "Russian", "Rwanda",  
    "Samoan", "Sango",  
    "Sanskrit", "Serbian", "Shona", "Sindhi", "Sinhala", "Slovak", "Slovene",  
    "Somali",  
    "Spanish", "Swahili", "Swedish", "Tajiki", "Tamil", "Tatar", "Telugu", "Thai",  
    "Tibetic",  
    "Tigrigna", "Tok Pisin", "Turkish", "Turkmen", "Twi", "Uighur", "Ukrainian",  
    "Urdu",  
    "Uzbek", "Vietnamese", "Warlpiri", "Welsh", "Wolof", "Xhosa", "Yoruba",  
    "Yucatec",  
    "Zapotec", "Zulu"]
```

```
nationalities_pattern = '|'.join(nationalities)  
countries_pattern = '|'.join(countries)  
languages_pattern = '|'.join(languages)
```