

Package ‘rebus.unicode’

October 14, 2022

Type Package

Title Unicode Extensions for the 'rebus' Package

Version 0.0-2

Date 2017-01-02

Author Richard Cotton [aut, cre]

Maintainer Richard Cotton <richierocks@gmail.com>

Description Build regular expressions piece by piece using human readable code.
This package contains Unicode functionality, and is primarily intended to be used by package developers.

Depends R (>= 3.1.0)

Imports rebus.base (>= 0.0-2)

Suggests stringi

License Unlimited

LazyLoad yes

LazyData yes

Acknowledgments Development of this package was partially funded by the Proteomics Core at Weill Cornell Medical College in Qatar <<http://qatar-weill.cornell.edu>>. The Core is supported by 'Biomedical Research Program' funds, a program funded by Qatar Foundation.

Collate 'imports.R' 'unicode-classes.R' 'unicode-constants.R'
'unicode-general-category-classes.R'
'unicode-general-category-constants.R' 'unicode-operators.R'

RoxygenNote 5.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2017-01-03 07:42:23

R topics documented:

ugc_cased_letter	2
Unicode	6
UnicodeOperators	29
up_alphabetic	31

Index	36
--------------	-----------

ugc_cased_letter	<i>Unicode General Categories</i>
------------------	-----------------------------------

Description

Match a Unicode General Category.

Usage

```
ugc_cased_letter(lo, hi, char_class = TRUE)
ugc_close_punctuation(lo, hi, char_class = TRUE)
ugc_connector_punctuation(lo, hi, char_class = TRUE)
ugc_control(lo, hi, char_class = TRUE)
ugc_currency_symbol(lo, hi, char_class = TRUE)
ugc_dash_punctuation(lo, hi, char_class = TRUE)
ugc_decimal_number(lo, hi, char_class = TRUE)
ugc_enclosing_mark(lo, hi, char_class = TRUE)
ugc_final_punctuation(lo, hi, char_class = TRUE)
ugc_format_control(lo, hi, char_class = TRUE)
ugc_initial_punctuation(lo, hi, char_class = TRUE)
ugc_letter(lo, hi, char_class = TRUE)
ugc_letter_number(lo, hi, char_class = TRUE)
ugc_line_separator(lo, hi, char_class = TRUE)
ugc_lowercase_letter(lo, hi, char_class = TRUE)
```

ugc_mark(lo, hi, char_class = TRUE)
ugc_math_symbol(lo, hi, char_class = TRUE)
ugc_modifier_letter(lo, hi, char_class = TRUE)
ugc_modifier_symbol(lo, hi, char_class = TRUE)
ugc_nonspacing_mark(lo, hi, char_class = TRUE)
ugc_number(lo, hi, char_class = TRUE)
ugc_open_punctuation(lo, hi, char_class = TRUE)
ugc_other(lo, hi, char_class = TRUE)
ugc_other_letter(lo, hi, char_class = TRUE)
ugc_other_number(lo, hi, char_class = TRUE)
ugc_other_punctuation(lo, hi, char_class = TRUE)
ugc_other_symbol(lo, hi, char_class = TRUE)
ugc_paragraph_separator(lo, hi, char_class = TRUE)
ugc_private_use_control(lo, hi, char_class = TRUE)
ugc_punctuation(lo, hi, char_class = TRUE)
ugc_separator(lo, hi, char_class = TRUE)
ugc_space_separator(lo, hi, char_class = TRUE)
ugc_spacing_mark(lo, hi, char_class = TRUE)
ugc_surrogate_control(lo, hi, char_class = TRUE)
ugc_symbol(lo, hi, char_class = TRUE)
ugc_titlecase_letter(lo, hi, char_class = TRUE)
ugc_unassigned_control(lo, hi, char_class = TRUE)
ugc_uppercase_letter(lo, hi, char_class = TRUE)
UGC_UPPERCASE_LETTER

UGC_LOWERCASE_LETTER

UGC_TITLECASE_LETTER

UGC_CASED_LETTER

UGC_MODIFIER_LETTER

UGC_OTHER_LETTER

UGC_LETTER

UGC_NONSPACING_MARK

UGC_SPACING_MARK

UGC_ENCLOSING_MARK

UGC_MARK

UGC_DECIMAL_NUMBER

UGC_LETTER_NUMBER

UGC_OTHER_NUMBER

UGC_NUMBER

UGC_CONNECTOR_PUNCTUATION

UGC_DASH_PUNCTUATION

UGC_OPEN_PUNCTUATION

UGC_CLOSE_PUNCTUATION

UGC_INITIAL_PUNCTUATION

UGC_FINAL_PUNCTUATION

UGC_OTHER_PUNCTUATION

UGC_PUNCTUATION

UGC_MATH_SYMBOL

UGC_CURRENCY_SYMBOL

UGC_MODIFIER_SYMBOL
UGC_OTHER_SYMBOL
UGC_SYMBOL
UGC_SPACE_SEPARATOR
UGC_LINE_SEPARATOR
UGC_PARAGRAPH_SEPARATOR
UGC_SEPARATOR
UGC_CONTROL
UGC_FORMAT_CONTROL
UGC_SURROGATE_CONTROL
UGC_PRIVATE_USE_CONTROL
UGC_UNASSIGNED_CONTROL
UGC_OTHER

Arguments

lo	A non-negative integer. Minimum number of repeats, when grouped.
hi	positive integer. Maximum number of repeats, when grouped.
char_class	TRUE or FALSE. Should the values be wrapped into a character class?

Format

An object of class `regex` (inherits from `character`) of length 1.

Value

A character vector representing part or all of a regular expression.

References

Table 12 of the Unicode Standard Annex #44 defines the Unicode General Categories. <http://www.unicode.org/reports/tr44>

You can see which characters are contained in a category by visiting, e.g., <http://www.fileformat.info/info/unicode/category/Nd/list.htm>

See Also

[unicode_property](#), [Unicode](#)

Examples

```
# Classes
ugc_lowercase_letter()
ugc_decimal_number()
ugc_paragraph_separator()
ugc_currency_symbol()

# With repetition
ugc_nonspacing_mark(3, 6)
ugc_separator(1, Inf)
ugc_dash_punctuation(0, Inf)

# Without a class wrapper
ugc_titlecase_letter(char_class = FALSE)

# Constants
UGC_UPPERCASE_LETTER
UGC_LETTER_NUMBER
UGC_MATH_SYMBOL
UGC_FORMAT_CONTROL

## Not run:
# All the Unicode general categories.
# Not run, since it generates lots of output
ls("package:rebus.unicode", pattern = "^ugc")

## End(Not run)

# Usage
library(rebus.base)
x <- "I exchanged $1000 for \u20ac665.41 and \u00a3243.13."
(rx <- capture(ugc_currency_symbol()) %R%
  capture(
    ugc_decimal_number(1, Inf) %R%
    optional(group(".", %R% ugc_decimal_number(2)))
  )
)
stringi::stri_match_all_regex(x, rx)
```

Unicode

Unicode classes

Description

Match ranges of unicode characters. In particular, you can match characters from a particular language.

Usage

```
armenian(lo, hi, char_class = TRUE)
armenian_ligatures(lo, hi, char_class = TRUE)
caucasian_albanian(lo, hi, char_class = TRUE)
cypriot_syllabary(lo, hi, char_class = TRUE)
cyrillic(lo, hi, char_class = TRUE)
cyrillic_supplement(lo, hi, char_class = TRUE)
cyrillic_extended_a(lo, hi, char_class = TRUE)
cyrillic_extended_b(lo, hi, char_class = TRUE)
elbasan(lo, hi, char_class = TRUE)
georgian(lo, hi, char_class = TRUE)
georgian_supplement(lo, hi, char_class = TRUE)
glagolitic(lo, hi, char_class = TRUE)
gothic(lo, hi, char_class = TRUE)
greek_and_coptic(lo, hi, char_class = TRUE)
greek_extended(lo, hi, char_class = TRUE)
latin(lo, hi, char_class = TRUE)
latin_1_supplement(lo, hi, char_class = TRUE)
latin_extended_a(lo, hi, char_class = TRUE)
latin_extended_b(lo, hi, char_class = TRUE)
latin_extended_c(lo, hi, char_class = TRUE)
latin_extended_d(lo, hi, char_class = TRUE)
latin_extended_e(lo, hi, char_class = TRUE)
latin_extended_additional(lo, hi, char_class = TRUE)
latin_ligatures(lo, hi, char_class = TRUE)
```

linear_a(lo, hi, char_class = TRUE)
linear_b_syllabary(lo, hi, char_class = TRUE)
linear_b_ideograms(lo, hi, char_class = TRUE)
ogham(lo, hi, char_class = TRUE)
old_italic(lo, hi, char_class = TRUE)
old_permic(lo, hi, char_class = TRUE)
phaistos_disc(lo, hi, char_class = TRUE)
runic(lo, hi, char_class = TRUE)
shavian(lo, hi, char_class = TRUE)
duployan(lo, hi, char_class = TRUE)
shorthand_format_controls(lo, hi, char_class = TRUE)
ipa_extensions(lo, hi, char_class = TRUE)
phonetic_extensions(lo, hi, char_class = TRUE)
phonetic_extensions_supplement(lo, hi, char_class = TRUE)
modifier_tone_letters(lo, hi, char_class = TRUE)
spacing_modifier_letters(lo, hi, char_class = TRUE)
superscripts_and_subscripts(lo, hi, char_class = TRUE)
combining_diacritic_marks(lo, hi, char_class = TRUE)
combining_diacritic_supplement(lo, hi, char_class = TRUE)
combining_diacritic_extended(lo, hi, char_class = TRUE)
combining_half_marks(lo, hi, char_class = TRUE)
bamun(lo, hi, char_class = TRUE)
bamun_supplement(lo, hi, char_class = TRUE)
bassa_vah(lo, hi, char_class = TRUE)

coptic(lo, hi, char_class = TRUE)
coptic_epact_numbers(lo, hi, char_class = TRUE)
egyptian_hieroglyphs(lo, hi, char_class = TRUE)
ethiopic(lo, hi, char_class = TRUE)
ethiopic_supplement(lo, hi, char_class = TRUE)
ethiopic_extended(lo, hi, char_class = TRUE)
ethiopic_extended_a(lo, hi, char_class = TRUE)
mende_kikakui(lo, hi, char_class = TRUE)
meroitic_cursive(lo, hi, char_class = TRUE)
meroitic_hieroglyphs(lo, hi, char_class = TRUE)
nko(lo, hi, char_class = TRUE)
osmanya(lo, hi, char_class = TRUE)
tifinagh(lo, hi, char_class = TRUE)
vai(lo, hi, char_class = TRUE)
arabic(lo, hi, char_class = TRUE)
arabic_supplement(lo, hi, char_class = TRUE)
arabic_extended_a(lo, hi, char_class = TRUE)
arabic_presentation_forms_a(lo, hi, char_class = TRUE)
arabic_presentation_forms_b(lo, hi, char_class = TRUE)
imperial_aramaic(lo, hi, char_class = TRUE)
avestan(lo, hi, char_class = TRUE)
carian(lo, hi, char_class = TRUE)
cuneiform(lo, hi, char_class = TRUE)
cuneiform_numbers_and_punctuation(lo, hi, char_class = TRUE)

old_persian(lo, hi, char_class = TRUE)
ugaritic(lo, hi, char_class = TRUE)
hebrew(lo, hi, char_class = TRUE)
lycian(lo, hi, char_class = TRUE)
lydian(lo, hi, char_class = TRUE)
mandaic(lo, hi, char_class = TRUE)
nabataean(lo, hi, char_class = TRUE)
old_north_arabian(lo, hi, char_class = TRUE)
old_south_arabian(lo, hi, char_class = TRUE)
pahlavi_inscriptional(lo, hi, char_class = TRUE)
pahlavi_psalter(lo, hi, char_class = TRUE)
palmyrene(lo, hi, char_class = TRUE)
phoenician(lo, hi, char_class = TRUE)
samaritan(lo, hi, char_class = TRUE)
syriac(lo, hi, char_class = TRUE)
manichaeian(lo, hi, char_class = TRUE)
mongolian(lo, hi, char_class = TRUE)
old_turkic(lo, hi, char_class = TRUE)
phags_pa(lo, hi, char_class = TRUE)
tibetan(lo, hi, char_class = TRUE)
bengali_and_assamese(lo, hi, char_class = TRUE)
brahmi(lo, hi, char_class = TRUE)
chakma(lo, hi, char_class = TRUE)
devanagari(lo, hi, char_class = TRUE)

devanagari_extended(lo, hi, char_class = TRUE)
grantha(lo, hi, char_class = TRUE)
gujarati(lo, hi, char_class = TRUE)
gurmukhi(lo, hi, char_class = TRUE)
kaithi(lo, hi, char_class = TRUE)
kannada(lo, hi, char_class = TRUE)
kharoshthi(lo, hi, char_class = TRUE)
khojki(lo, hi, char_class = TRUE)
khudawadi(lo, hi, char_class = TRUE)
lepcha(lo, hi, char_class = TRUE)
limbu(lo, hi, char_class = TRUE)
mahajani(lo, hi, char_class = TRUE)
malayalam(lo, hi, char_class = TRUE)
meetei_mayek(lo, hi, char_class = TRUE)
meetei_mayek_extensions(lo, hi, char_class = TRUE)
modi(lo, hi, char_class = TRUE)
mro(lo, hi, char_class = TRUE)
ol_chiki(lo, hi, char_class = TRUE)
oriya(lo, hi, char_class = TRUE)
saurashtra(lo, hi, char_class = TRUE)
sharada(lo, hi, char_class = TRUE)
siddham(lo, hi, char_class = TRUE)
sinhala(lo, hi, char_class = TRUE)
sinhala_archaic_numbers(lo, hi, char_class = TRUE)

sora_sompeng(lo, hi, char_class = TRUE)
syloiti_nagri(lo, hi, char_class = TRUE)
takri(lo, hi, char_class = TRUE)
tamil(lo, hi, char_class = TRUE)
telugu(lo, hi, char_class = TRUE)
thaana(lo, hi, char_class = TRUE)
tirhuta(lo, hi, char_class = TRUE)
vedic_extensions(lo, hi, char_class = TRUE)
warang_citi(lo, hi, char_class = TRUE)
cham(lo, hi, char_class = TRUE)
kayah_li(lo, hi, char_class = TRUE)
khmer(lo, hi, char_class = TRUE)
khmer_symbols(lo, hi, char_class = TRUE)
lao(lo, hi, char_class = TRUE)
myanmar(lo, hi, char_class = TRUE)
myanmar_extended_a(lo, hi, char_class = TRUE)
myanmar_extended_b(lo, hi, char_class = TRUE)
new_tai_lue(lo, hi, char_class = TRUE)
pahawh_hmong(lo, hi, char_class = TRUE)
pau_cin_hau(lo, hi, char_class = TRUE)
tai_le(lo, hi, char_class = TRUE)
tai_tham(lo, hi, char_class = TRUE)
tai_viet(lo, hi, char_class = TRUE)
thai(lo, hi, char_class = TRUE)

balinese(lo, hi, char_class = TRUE)
batak(lo, hi, char_class = TRUE)
buginese(lo, hi, char_class = TRUE)
buhid(lo, hi, char_class = TRUE)
hanunoo(lo, hi, char_class = TRUE)
javanese(lo, hi, char_class = TRUE)
rejang(lo, hi, char_class = TRUE)
sundanese(lo, hi, char_class = TRUE)
sundanese_supplement(lo, hi, char_class = TRUE)
tagalog(lo, hi, char_class = TRUE)
tagbanwa(lo, hi, char_class = TRUE)
bopomofo(lo, hi, char_class = TRUE)
bopomofo_extended(lo, hi, char_class = TRUE)
cjk_unified_ideographs(lo, hi, char_class = TRUE)
cjk_unified_ideographs_extension_a(lo, hi, char_class = TRUE)
cjk_unified_ideographs_extension_b(lo, hi, char_class = TRUE)
cjk_unified_ideographs_extension_c(lo, hi, char_class = TRUE)
cjk_unified_ideographs_extension_d(lo, hi, char_class = TRUE)
cjk_compatibility_ideographs(lo, hi, char_class = TRUE)
cjk_compatibility_ideographs_supplement(lo, hi, char_class = TRUE)
kangxi_radicals(lo, hi, char_class = TRUE)
kangxi_radicals_supplement(lo, hi, char_class = TRUE)
cjk_strokes(lo, hi, char_class = TRUE)
cjk_ideographic_description_characters(lo, hi, char_class = TRUE)

hangul_jamo(lo, hi, char_class = TRUE)
hangul_jamo_extended_a(lo, hi, char_class = TRUE)
hangul_jamo_extended_b(lo, hi, char_class = TRUE)
hangul_compatibility_jamo(lo, hi, char_class = TRUE)
hangul_syllables(lo, hi, char_class = TRUE)
hiragana(lo, hi, char_class = TRUE)
katakana(lo, hi, char_class = TRUE)
katakana_phonetic_extensions(lo, hi, char_class = TRUE)
kana_supplement(lo, hi, char_class = TRUE)
kanbun(lo, hi, char_class = TRUE)
lisu(lo, hi, char_class = TRUE)
miao(lo, hi, char_class = TRUE)
yi_syllables(lo, hi, char_class = TRUE)
yi_radicals(lo, hi, char_class = TRUE)
cherokee(lo, hi, char_class = TRUE)
deseret(lo, hi, char_class = TRUE)
unified_canadian_aboriginal_syllabics(lo, hi, char_class = TRUE)
unified_canadian_aboriginal_syllabics_extended(lo, hi, char_class = TRUE)
alphabetic_presentation_forms(lo, hi, char_class = TRUE)
halfwidth_and_fullwidth_forms(lo, hi, char_class = TRUE)
general_punctuation(lo, hi, char_class = TRUE)
latin_1_punctuation(lo, hi, char_class = TRUE)
small_form_variants(lo, hi, char_class = TRUE)
supplemental_punctuation(lo, hi, char_class = TRUE)

CJK_symbols_and_punctuation(lo, hi, char_class = TRUE)
CJK_compatibility_forms(lo, hi, char_class = TRUE)
fullwidth_ascii_punctuation(lo, hi, char_class = TRUE)
vertical_forms(lo, hi, char_class = TRUE)
letterlike_symbols(lo, hi, char_class = TRUE)
ancient_symbols(lo, hi, char_class = TRUE)
mathematical_alphanumeric_symbols(lo, hi, char_class = TRUE)
arabic_mathematical_alphanumeric_symbols(lo, hi, char_class = TRUE)
enclosed_alphanumerics(lo, hi, char_class = TRUE)
enclosed_alphanumeric_supplement(lo, hi, char_class = TRUE)
enclosed_CJK_letters_and_months(lo, hi, char_class = TRUE)
enclosed_ideographic_supplement(lo, hi, char_class = TRUE)
CJK_compatibility(lo, hi, char_class = TRUE)
miscellaneous_technical(lo, hi, char_class = TRUE)
control_pictures(lo, hi, char_class = TRUE)
optical_character_recognition(lo, hi, char_class = TRUE)
combining_diacritic_marks_for_symbols(lo, hi, char_class = TRUE)
aegean_numbers(lo, hi, char_class = TRUE)
ancient_greek_numbers(lo, hi, char_class = TRUE)
fullwidth_ascii_digits(lo, hi, char_class = TRUE)
common_indic_number_forms(lo, hi, char_class = TRUE)
coptic_epact_numbers(lo, hi, char_class = TRUE)
counting_rod_numerals(lo, hi, char_class = TRUE)
number_forms(lo, hi, char_class = TRUE)

```
rumi_numerals(lo, hi, char_class = TRUE)
sinhala_archaic_numbers(lo, hi, char_class = TRUE)
math_arrows(lo, hi, char_class = TRUE)
supplemental_arrows_a(lo, hi, char_class = TRUE)
supplemental_arrows_b(lo, hi, char_class = TRUE)
supplemental_arrows_c(lo, hi, char_class = TRUE)
additional_arrows(lo, hi, char_class = TRUE)
supplemental_mathematical_operators(lo, hi, char_class = TRUE)
miscellaneous_mathematical_symbols_a(lo, hi, char_class = TRUE)
miscellaneous_mathematical_symbols_b(lo, hi, char_class = TRUE)
floors_and_ceilings(lo, hi, char_class = TRUE)
invisible_operators(lo, hi, char_class = TRUE)
geometric_shapes(lo, hi, char_class = TRUE)
box_drawing(lo, hi, char_class = TRUE)
block_elements(lo, hi, char_class = TRUE)
geometric_shapes_extended(lo, hi, char_class = TRUE)
alchemical_symbols(lo, hi, char_class = TRUE)
braille_patterns(lo, hi, char_class = TRUE)
currency_symbols(lo, hi, char_class = TRUE)
dingbats(lo, hi, char_class = TRUE)
ornamental_dingbats(lo, hi, char_class = TRUE)
emoticons(lo, hi, char_class = TRUE)
chess_checkers draughts(lo, hi, char_class = TRUE)
domino_tiles(lo, hi, char_class = TRUE)
```



```
japanese_chess(lo, hi, char_class = TRUE)
mahjong_tiles(lo, hi, char_class = TRUE)
playing_cards(lo, hi, char_class = TRUE)
card_suits(lo, hi, char_class = TRUE)
miscellaneous_symbols_and_pictographs(lo, hi, char_class = TRUE)
musical_symbols(lo, hi, char_class = TRUE)
ancient_greek_musical_notation(lo, hi, char_class = TRUE)
byzantine_musical_symbols(lo, hi, char_class = TRUE)
transport_and_map_symbols(lo, hi, char_class = TRUE)
yijing_mono_di_and_trigrams(lo, hi, char_class = TRUE)
yijing_hexagram_symbols(lo, hi, char_class = TRUE)
tai_xuan_jing_symbols(lo, hi, char_class = TRUE)
specials(lo, hi, char_class = TRUE)
tags(lo, hi, char_class = TRUE)
variation_selectors(lo, hi, char_class = TRUE)
variation_selectors_supplement(lo, hi, char_class = TRUE)
private_use_area(lo, hi, char_class = TRUE)
supplementary_private_use_area_a(lo, hi, char_class = TRUE)
supplementary_private_use_area_b(lo, hi, char_class = TRUE)
ARMENIAN
ARMENIAN_LIGATURES
CAUCASIAN_ALBANIAN
CYPRIOT_SYLLABARY
CYRILLIC
```

CYRILLIC_SUPPLEMENT

CYRILLIC_EXTENDED_A

CYRILLIC_EXTENDED_B

ELBASAN

GEORGIAN

GEORGIAN_SUPPLEMENT

GLAGOLITIC

GOTHIC

GREEK_AND_COPTIC

GREEK_EXTENDED

LATIN

LATIN_1_SUPPLEMENT

LATIN_EXTENDED_A

LATIN_EXTENDED_B

LATIN_EXTENDED_C

LATIN_EXTENDED_D

LATIN_EXTENDED_E

LATIN_EXTENDED_ADDITIONAL

LATIN_LIGATURES

LINEAR_A

LINEAR_B_SYLLABARY

LINEAR_B_IDEOGRAMS

OGHAM

OLD_ITALIC

OLD_PERMIC
PHAISTOS_DISC
RUNIC
SHAVIAN
DUPLOYAN
SHORTHAND_FORMAT_CONTROLS
IPA_EXTENSIONS
PHONETIC_EXTENSIONS
PHONETIC_EXTENSIONS_SUPPLEMENT
MODIFIER_TONE_LETTERS
SPACING_MODIFIER_LETTERS
SUPERSCRIPTS_AND_SUBSCRIPTS
COMBINING_DIACRITIC_MARKS
COMBINING_DIACRITIC_SUPPLEMENT
COMBINING_DIACRITIC_EXTENDED
COMBINING_HALF_MARKS
BAMUN
BAMUN_SUPPLEMENT
BASSA_VAH
COPTIC
COPTIC_EPACT_NUMBERS
EGYPTIAN_HIEROGLYPHS
ETHIOPIA
ETHIOPIA_SUPPLEMENT

ETHIOPIC_EXTENDED
ETHIOPIC_EXTENDED_A
MENDE_KIKAKUI
MEROITIC_CURSIVE
MEROITIC_HIEROGLYPHS
NKO
OSMANYA
TIFINAGH
VAI
ARABIC
ARABIC_SUPPLEMENT
ARABIC_EXTENDED_A
ARABIC_PRESENTATION_FORMS_A
ARABIC_PRESENTATION_FORMS_B
IMPERIAL_ARAMAIC
AVESTAN
CARIAN
CUNEIFORM
CUNEIFORM_NUMBERS_AND_PUNCTUATION
OLD_PERSIAN
UGARITIC
HEBREW
LYCIAN
LYDIAN

MANDAIC
NABATAEAN
OLD_NORTH_ARABIAN
OLD_SOUTH_ARABIAN
PAHLAVI_INSCRIPTIONAL
PAHLAVI_PALTER
PALMYRENE
PHOENICIAN
SAMARITAN
SYRIAC
MANICHAEAN
MONGOLIAN
OLD_TURKIC
PHAGS_PA
TIBETAN
BENGLI_AND_ASSAMESE
BRAHMI
CHAKMA
DEVANAGARI
DEVANAGARI_EXTENDED
GRANTHA
GUJARATI
GURMUKHI
KAITHI

KANNADA

KHAROSHTHI

KHOJKI

KHUDAWADI

LEPCHA

LIMBU

MAHAJANI

MALAYALAM

MEETEI_MAYEK

MEETEI_MAYEK_EXTENSIONS

MODI

MRO

OL_CHIKI

ORIYA

SAURASHTRA

SHARADA

SIDDHAM

SINHALA

SINHALA_ARCHAIC_NUMBERS

SORA_SOMPENG

SYLOTI_NAGRI

TAKRI

TAMIL

TELUGU

THAANA

TIRHUTA

VEDIC_EXTENSIONS

WARANG_CITI

CHAM

KAYAH_LI

KHMER

KHMER_SYMBOLS

LAO

MYANMAR

MYANMAR_EXTENDED_A

MYANMAR_EXTENDED_B

NEW_TAI_LUE

PAHAWH_HMONG

PAU_CIN_HAU

TAI_LE

TAI_THAM

TAI_VIET

THAI

BALINESE

BATAK

BUGINESE

BUHID

HANUNOO

JAVANESE
REJANG
SUNDANESE
SUNDANESE_SUPPLEMENT
TAGALOG
TAGBANWA
BOPOMOFO
BOPOMOFO_EXTENDED
CJK_UNIFIED_IDEOGRAPHS
CJK_UNIFIED_IDEOGRAPHS_EXTENSION_A
CJK_UNIFIED_IDEOGRAPHS_EXTENSION_B
CJK_UNIFIED_IDEOGRAPHS_EXTENSION_C
CJK_UNIFIED_IDEOGRAPHS_EXTENSION_D
CJK_COMPATIBILITY_IDEOGRAPHS
CJK_COMPATIBILITY_IDEOGRAPHS_SUPPLEMENT
KANGXI_RADICALS
KANGXI_RADICALS_SUPPLEMENT
CJK_STROKES
CJK_IDEOGRAPHIC_DESCRIPTION_CHARACTERS
HANGUL_JAMO
HANGUL_JAMO_EXTENDED_A
HANGUL_JAMO_EXTENDED_B
HANGUL_COMPATIBILITY_JAMO
HANGUL_SYLLABLES

HIRAGANA
KATAKANA
KATAKANA_PHONETIC_EXTENSIONS
KANA_SUPPLEMENT
KANBUN
LISU
MIAO
YI_SYLLABLES
YI_RADICALS
CHEROKEE
DESERET
UNIFIED_CANADIAN_ABORIGINAL_SYLLABICS
UNIFIED_CANADIAN_ABORIGINAL_SYLLABICS_EXTENDED
ALPHABETIC_PRESENTATION_FORMS
HALFWIDTH_AND_FULLWIDTH_FORMS
GENERAL_PUNCTUATION
LATIN_1_PUNCTUATION
SMALL_FORM_VARIANTS
SUPPLEMENTAL_PUNCTUATION
CJK_SYMBOLS_AND_PUNCTUATION
CJK_COMPATIBILITY_FORMS
FULLWIDTH_ASCII_PUNCTUATION
VERTICAL_FORMS
LETTERLIKE_SYMBOLS

ANCIENT_SYMBOLS
MATHEMATICAL_ALPHANUMERIC_SYMBOLS
ARABIC_MATHEMATICAL_ALPHANUMERIC_SYMBOLS
ENCLOSED_ALPHANUMERICS
ENCLOSED_ALPHANUMERIC_SUPPLEMENT
ENCLOSED_CJK_LETTERS_AND_MONTHS
ENCLOSED_IDEOGRAPHIC_SUPPLEMENT
CJK_COMPATIBILITY
MISCELLANEOUS_TECHNICAL
CONTROL_PICTURES
OPTICAL_CHARACTER_RECOGNITION
COMBINING_DIACRITIC_MARKS_FOR_SYMBOLS
AEGEAN_NUMBERS
ANCIENT_GREEK_NUMBERS
FULLWIDTH_ASCII_DIGITS
COMMON_INDIC_NUMBER_FORMS
COPTIC_EPACT_NUMBERS
COUNTING_ROD_NUMERALS
NUMBER_FORMS
RUMI_NUMERAL_SYMBOLS
SINHALA_ARCHAIC_NUMBERS
MATH_ARROWS
SUPPLEMENTAL_ARROWS_A
SUPPLEMENTAL_ARROWS_A

SUPPLEMENTAL_ARROWS_A
ADDITIONAL_ARROWS
SUPPLEMENTAL_MATHEMATICAL_OPERATORS
MISCELLANEOUS_MATHEMATICAL_SYMBOLS_A
MISCELLANEOUS_MATHEMATICAL_SYMBOLS_B
FLOORS_AND_CEILINGS
INVISIBLE_OPERATORS
GEOMETRIC_SHAPES
BOX_DRAWING
BLOCK_ELEMENTS
GEOMETRIC_SHAPES_EXTENDED
ALCHEMICAL_SYMBOLS
BRAILLE_PATTERNS
CURRENCY_SYMBOLS
DINGBATS
ORNAMENTAL_DINGBATS
EMOTICONS
CHESS_CHECKERS_DRAUGHTS
DOMINO_TILES
JAPANESE_CHESS
MAHJONG_TILES
PLAYING_CARDS
CARD_SUITS
MISCELLANEOUS_SYMBOLS_AND_PICTOGRAPHS

MUSICAL_SYMBOLS
ANCIENT_GREEK_MUSICAL_NOTATION
BYZANTINE_MUSICAL_SYMBOLS
TRANSPORT_AND_MAP_SYMBOLS
YIJING_MONO_DI_AND_TRIGRAMS
YIJING_HEXAGRAM_SYMBOLS
TAI_XUAN_JING_SYMBOLS
SPECIALS
TAGS
VARIATION_SELECTORS
VARIATION_SELECTORS_SUPPLEMENT
PRIVATE_USE_AREA
SUPPLEMENTARY_PRIVATE_USE_AREA_A
SUPPLEMENTARY_PRIVATE_USE_AREA_B

Arguments

lo	A non-negative integer. Minimum number of repeats, when grouped.
hi	positive integer. Maximum number of repeats, when grouped.
char_class	TRUE or FALSE. Should the values be wrapped into a character class?

Format

An object of class `regex` (inherits from `character`) of length 1.

Value

A character vector representing part or all of a regular expression.

Note

Windows currently doesn't handle Unicode points with more than four digits correctly. See https://bugs.r-project.org/bugzilla3/show_bug.cgi?id=16098

References

<http://www.unicode.org/charts>

See Also

[ClassGroups](#)

Examples

```
# Classes
latin()
greek_and_coptic()
cyrillic()
arabic()

# With repetition
hebrew(3, 6)
hiragana(1, Inf)
katakana(0, Inf)

# Without a class wrapper
cjk_unified_ideographs(char_class = FALSE)

# Constants
ARMENIAN
LINEAR_B_IDEOGRAMS
DUPLOYAN
OSMANYA

## Not run:
# All the Unicode character classes
# Not run, since it generates lots of output
setdiff(
  ls("package:rebus.unicode", pattern = lower()),
  ls(
    "package:rebus.unicode",
    pattern = START %R% case_insensitive(or("up", "ugc", "unicode")))
)

## End(Not run)

# Usage
pythag <- "\u03b1^2 + \u03b2^2 = \u03b3^2"
stringi::stri_extract_all_regex(pythag, greek_and_coptic())
```

Description

Manipulate and combine Unicode Properties.

Usage

```
unicode_inverse(x, char_class = TRUE)
unicode_union(..., char_class = TRUE)
unicode_intersect(x, y, char_class = TRUE)
unicode_setdiff(x, y, char_class = TRUE)
```

Arguments

x	A character vector containing Unicode General Category or Unicode Properties. Use the functional forms (ugc_*) not the constants.
char_class	TRUE or FALSE. Should the values be wrapped into a character class?
...	Character vectors containing Unicode General Category or Unicode Properties. Use the functional forms (ugc_*) not the constants.
y	A character vector containing Unicode General Category or Unicode Properties. Use the functional forms (ugc_*) not the constants.

Note

Use these with ICU-based regular expression engines (stringi and stringr).

References

<http://userguide.icu-project.org/strings/unicodeset>

Examples

```
# POSIX [:punct:] is more or less equivalent to the union of
# Unicode punctuation and symbol general categories
unicode_union(ugc_punctuation(), ugc_symbol())

# Everything except "A" to "Z" (including punctuation, control chars etc.)
unicode_inverse("[A-Z]")

# Uppercase letters, except "A" to "Z"
unicode_setdiff(ugc_uppercase_letter(), "[A-Z]")

# "A" to "F" (in upper or lower case)
unicode_intersect(ugc_letter(), up_ascii_hex_digit())

# Usage
x <- c(letters, LETTERS)
rx <- unicode_intersect(ugc_letter(), up_ascii_hex_digit())
stringi::stri_extract_first_regex(x, rx)
```

up_alphabetic	<i>Unicode Properties</i>
---------------	---------------------------

Description

Match a Unicode Property.

Usage

```
up_alphabetic(lo, hi, char_class = TRUE)
up_ascii_hex_digit(lo, hi, char_class = TRUE)
up_bidi_control(lo, hi, char_class = TRUE)
up_bidi_mirrored(lo, hi, char_class = TRUE)
up_case_ignorable(lo, hi, char_class = TRUE)
up_case_sensitive(lo, hi, char_class = TRUE)
up_cased(lo, hi, char_class = TRUE)
up_changes_when_casefolded(lo, hi, char_class = TRUE)
up_changes_when_casemapped(lo, hi, char_class = TRUE)
up_changes_when_lowercased(lo, hi, char_class = TRUE)
up_changes_when_nfkc_casefolded(lo, hi, char_class = TRUE)
up_changes_when_titlecased(lo, hi, char_class = TRUE)
up_changes_when_uppercased(lo, hi, char_class = TRUE)
up_dash(lo, hi, char_class = TRUE)
up_default_ignorable_code_point(lo, hi, char_class = TRUE)
up_deprecated(lo, hi, char_class = TRUE)
up_diacritic(lo, hi, char_class = TRUE)
up_extender(lo, hi, char_class = TRUE)
up_hex_digit(lo, hi, char_class = TRUE)
```

```
up_hyphen(lo, hi, char_class = TRUE)
up_id_continue(lo, hi, char_class = TRUE)
up_id_start(lo, hi, char_class = TRUE)
up_ideographic(lo, hi, char_class = TRUE)
up_lowercase(lo, hi, char_class = TRUE)
up_math(lo, hi, char_class = TRUE)
up_noncharacter_code_point(lo, hi, char_class = TRUE)
up_posix_alnum(lo, hi, char_class = TRUE)
up_posix_blank(lo, hi, char_class = TRUE)
up_posix_graph(lo, hi, char_class = TRUE)
up_posix_print(lo, hi, char_class = TRUE)
up_posix_xdigit(lo, hi, char_class = TRUE)
up_quotation_mark(lo, hi, char_class = TRUE)
up_soft_dotted(lo, hi, char_class = TRUE)
up_terminal_punctuation(lo, hi, char_class = TRUE)
up_uppercase(lo, hi, char_class = TRUE)
up_white_space(lo, hi, char_class = TRUE)
UP_ALPHABETIC
UP_ASCII_HEX_DIGIT
UP_BIDI_CONTROL
UP_BIDI_MIRRORED
UP_DASH
UP_DEFAULT_IGNOREABLE_CODE_POINT
UP_DEPRECATED
```


UP_DIACRITIC
UP_EXTENDER
UP_HEX_DIGIT
UP_HYPHEN
UP_ID_CONTINUE
UP_ID_START
UP_IDEOGRAPHIC
UP_LOWERCASE
UP_MATH
UP_NONCHARACTER_CODE_POINT
UP_QUOTATION_MARK
UP_SOFT_DOTTED
UP_TERMINAL_PUNCTUATION
UP_UPPERCASE
UP_WHITE_SPACE
UP_CASE_SENSITIVE
UP_POSIX_ALNUM
UP_POSIX_BLANK
UP_POSIX_GRAPH
UP_POSIX_PRINT
UP_POSIX_XDIGIT
UP_CASED
UP_CASE_IGNOREABLE
UP_CHANGES_WHEN_LOWERCASED

```
UP_CHANGES_WHEN_UPPERCASED
UP_CHANGES_WHEN_TITLECASED
UP_CHANGES_WHEN_CASEFOLDED
UP_CHANGES_WHEN_CASEMAPPED
UP_CHANGES_WHEN_NFKC_CASEFOLDED
```

Arguments

lo	A non-negative integer. Minimum number of repeats, when grouped.
hi	positive integer. Maximum number of repeats, when grouped.
char_class	TRUE or FALSE. Should the values be wrapped into a character class?

Format

An object of class `regex` (inherits from `character`) of length 1.

Value

A character vector representing part or all of a regular expression.

References

Table 12 of the Unicode Standard Annex #44 defines the Unicode General Categories. <http://www.unicode.org/reports/tr44/>

You can see which characters are contained in a category by visiting, e.g., <http://www.fileformat.info/info/unicode/category/Nd/list.htm>

See Also

[unicode_general_category](#), [Unicode](#), [stringi-search-charclass](#)

Examples

```
# Classes
up_math()
up_posix_alnum()
up_changes_when_uppercased()
up_diacritic()

# With repetition
ugc_nonspacing_mark(3, 6)
up_quotation_mark(1, Inf)
up_posix_xdigit(0, Inf)

# Without a class wrapper
up_hyphen(char_class = FALSE)
```

```
# Constants
UP_ALPHABETIC
UP_DASH
UP_POSIX_ALNUM
UP_CHANGES_WHEN_LOWERCASED

## Not run:
# All the Unicode properties.
# Not run, since it generates lots of output
ls("package:rebus.unicode", pattern = "^up")

## End(Not run)

# Usage
# Hello in Samoan, Serbian, Persian, Simplified Chinese
hello <- "t\u101lofa, \u437\u434\u440\u430\u432\u43e, \u633\u644\u627\u645, \u4f60\u597d"
stringi::stri_extract_all_regex(hello, up_alphabetic(1, Inf))
stringi::stri_extract_all_regex(hello, up_case_sensitive(1, Inf))
```

Index

* datasets

- ugc_cased_letter, 2
 - Unicode, 6
 - up_alphabetic, 31
- ADDITIONAL_ARROWS (Unicode), 6
- additional_arrows (Unicode), 6
- AEGEAN_NUMBERS (Unicode), 6
- aegean_numbers (Unicode), 6
- ALCHEMICAL_SYMBOLS (Unicode), 6
- alchemical_symbols (Unicode), 6
- ALPHABETIC_PRESENTATION_FORMS (Unicode), 6
- alphabetic_presentation_forms (Unicode), 6
- ANCIENT_GREEK_MUSICAL_NOTATION (Unicode), 6
- ancient_greek_musical_notation (Unicode), 6
- ANCIENT_GREEK_NUMBERS (Unicode), 6
- ancient_greek_numbers (Unicode), 6
- ANCIENT_SYMBOLS (Unicode), 6
- ancient_symbols (Unicode), 6
- ARABIC (Unicode), 6
- arabic (Unicode), 6
- ARABIC_EXTENDED_A (Unicode), 6
- arabic_extended_a (Unicode), 6
- ARABIC_MATHEMATICAL_ALPHANUMERIC_SYMBOLS (Unicode), 6
- arabic_mathematical_alphanumeric_symbols (Unicode), 6
- ARABIC_PRESENTATION_FORMS_A (Unicode), 6
- arabic_presentation_forms_a (Unicode), 6
- ARABIC_PRESENTATION_FORMS_B (Unicode), 6
- arabic_presentation_forms_b (Unicode), 6
- ARABIC_SUPPLEMENT (Unicode), 6
- arabic_supplement (Unicode), 6
- ARMENIAN (Unicode), 6
- armenian (Unicode), 6
- ARMENIAN_LIGATURES (Unicode), 6
- armenian_ligatures (Unicode), 6
- AVESTAN (Unicode), 6
- avestan (Unicode), 6
- BALINESE (Unicode), 6
- balinese (Unicode), 6
- BAMUN (Unicode), 6
- bamun (Unicode), 6
- BAMUN_SUPPLEMENT (Unicode), 6
- bamun_supplement (Unicode), 6
- BASSA_VAH (Unicode), 6
- bassa_vah (Unicode), 6
- BATAK (Unicode), 6
- batak (Unicode), 6
- BENGALI_AND_ASSAMESE (Unicode), 6
- bengali_and_assamese (Unicode), 6
- BLOCK_ELEMENTS (Unicode), 6
- block_elements (Unicode), 6
- BOPOMOFO (Unicode), 6
- bopomofo (Unicode), 6
- BOPOMOFO_EXTENDED (Unicode), 6
- bopomofo_extended (Unicode), 6
- BOX_DRAWING (Unicode), 6
- box_drawing (Unicode), 6
- BRAHMI (Unicode), 6
- brahmi (Unicode), 6
- BRAILLE_PATTERNS (Unicode), 6
- braille_patterns (Unicode), 6
- BUGINESE (Unicode), 6
- buginese (Unicode), 6
- BUHID (Unicode), 6
- buhid (Unicode), 6
- BYZANTINE_MUSICAL_SYMBOLS (Unicode), 6
- byzantine_musical_symbols (Unicode), 6
- CARD_SUITS (Unicode), 6
- card_suits (Unicode), 6
- CARIAN (Unicode), 6
- carian (Unicode), 6
- CAUCASIAN_ALBANIAN (Unicode), 6

- caucasian_albanian (Unicode), 6
- CHAKMA (Unicode), 6
- chakma (Unicode), 6
- CHAM (Unicode), 6
- cham (Unicode), 6
- CHEROKEE (Unicode), 6
- cherokee (Unicode), 6
- CHESSE_CHECKERS_DRAUGHTS (Unicode), 6
- chess_checkers draughts (Unicode), 6
- CJK_COMPATIBILITY (Unicode), 6
- chk_compatibility (Unicode), 6
- CJK_COMPATIBILITY_FORMS (Unicode), 6
- chk_compatibility_forms (Unicode), 6
- CJK_COMPATIBILITY_IDEOGRAPHS (Unicode), 6
- chk_compatibility_ideographs (Unicode), 6
- CJK_COMPATIBILITY_IDEOGRAPHS_SUPPLEMENT (Unicode), 6
- chk_compatibility_ideographs_supplement (Unicode), 6
- CJK_IDEOGRAPHIC_DESCRIPTION_CHARACTERS (Unicode), 6
- chk_ideographic_description_characters (Unicode), 6
- CJK_STROKES (Unicode), 6
- chk_strokes (Unicode), 6
- CJK_SYMBOLS_AND_PUNCTUATION (Unicode), 6
- chk_symbols_and_punctuation (Unicode), 6
- CJK_UNIFIED_IDEOGRAPHS (Unicode), 6
- chk_unified_ideographs (Unicode), 6
- CJK_UNIFIED_IDEOGRAPHS_EXTENSION_A (Unicode), 6
- chk_unified_ideographs_extension_a (Unicode), 6
- CJK_UNIFIED_IDEOGRAPHS_EXTENSION_B (Unicode), 6
- chk_unified_ideographs_extension_b (Unicode), 6
- CJK_UNIFIED_IDEOGRAPHS_EXTENSION_C (Unicode), 6
- chk_unified_ideographs_extension_c (Unicode), 6
- CJK_UNIFIED_IDEOGRAPHS_EXTENSION_D (Unicode), 6
- chk_unified_ideographs_extension_d (Unicode), 6
- ClassGroups, 29
- COMBINING_DIACRITIC_EXTENDED (Unicode), 6
- combining_diacritic_extended (Unicode), 6
- COMBINING_DIACRITIC_MARKS (Unicode), 6
- combining_diacritic_marks (Unicode), 6
- COMBINING_DIACRITIC_MARKS_FOR_SYMBOLS (Unicode), 6
- combining_diacritic_marks_for_symbols (Unicode), 6
- COMBINING_DIACRITIC_SUPPLEMENT (Unicode), 6
- combining_diacritic_supplement (Unicode), 6
- COMBINING_HALF_MARKS (Unicode), 6
- combining_half_marks (Unicode), 6
- COMMON_INDIC_NUMBER_FORMS (Unicode), 6
- common_indic_number_forms (Unicode), 6
- CONTROL_PICTURES (Unicode), 6
- control_pictures (Unicode), 6
- COPTIC (Unicode), 6
- coptic (Unicode), 6
- COPTIC_EPACT_NUMBERS (Unicode), 6
- coptic_epact_numbers (Unicode), 6
- COUNTING_ROD_NUMERALS (Unicode), 6
- counting_rod_numerals (Unicode), 6
- CUNEIFORM (Unicode), 6
- cuneiform (Unicode), 6
- CUNEIFORM_NUMBERS_AND_PUNCTUATION (Unicode), 6
- cuneiform_numbers_and_punctuation (Unicode), 6
- CURRENCY_SYMBOLS (Unicode), 6
- currency_symbols (Unicode), 6
- CYPRIOT_SYLLABARY (Unicode), 6
- cyriot_syllabary (Unicode), 6
- CYRILLIC (Unicode), 6
- cyrillic (Unicode), 6
- CYRILLIC_EXTENDED_A (Unicode), 6
- cyrillic_extended_a (Unicode), 6
- CYRILLIC_EXTENDED_B (Unicode), 6
- cyrillic_extended_b (Unicode), 6
- CYRILLIC_SUPPLEMENT (Unicode), 6
- cyrillic_supplement (Unicode), 6
- DESERET (Unicode), 6
- deseret (Unicode), 6
- DEVANAGARI (Unicode), 6
- devanagari (Unicode), 6

- DEVANAGARI_EXTENDED (Unicode), 6
- devanagari_extended (Unicode), 6
- DINGBATS (Unicode), 6
- dingbats (Unicode), 6
- DOMINO_TILES (Unicode), 6
- domino_tiles (Unicode), 6
- DUPLOYAN (Unicode), 6
- duployan (Unicode), 6

- EGYPTIAN_HIEROGLYPHS (Unicode), 6
- egyptian_hieroglyphs (Unicode), 6
- ELBASAN (Unicode), 6
- elbasan (Unicode), 6
- EMOTICONS (Unicode), 6
- emoticons (Unicode), 6
- ENCLOSED_ALPHANUMERIC_SUPPLEMENT (Unicode), 6
- enclosed_alphanumeric_supplement (Unicode), 6
- ENCLOSED_ALPHANUMERICS (Unicode), 6
- enclosed_alphanumerics (Unicode), 6
- ENCLOSED_CJK_LETTERS_AND_MONTHS (Unicode), 6
- enclosed_cjk_letters_and_months (Unicode), 6
- ENCLOSED_IDEOGRAPHIC_SUPPLEMENT (Unicode), 6
- enclosed_ideographic_supplement (Unicode), 6
- ETHIOPIIC (Unicode), 6
- ethiopic (Unicode), 6
- ETHIOPIIC_EXTENDED (Unicode), 6
- ethiopic_extended (Unicode), 6
- ETHIOPIIC_EXTENDED_A (Unicode), 6
- ethiopic_extended_a (Unicode), 6
- ETHIOPIIC_SUPPLEMENT (Unicode), 6
- ethiopic_supplement (Unicode), 6

- FLOORS_AND_CEILINGS (Unicode), 6
- floors_and_ceilings (Unicode), 6
- FULLWIDTH_ASCII_DIGITS (Unicode), 6
- fullwidth_ascii_digits (Unicode), 6
- FULLWIDTH_ASCII_PUNCTUATION (Unicode), 6
- fullwidth_ascii_punctuation (Unicode), 6

- GENERAL_PUNCTUATION (Unicode), 6
- general_punctuation (Unicode), 6
- GEOMETRIC_SHAPES (Unicode), 6
- geometric_shapes (Unicode), 6

- GEOMETRIC_SHAPES_EXTENDED (Unicode), 6
- geometric_shapes_extended (Unicode), 6
- GEORGIAN (Unicode), 6
- georgian (Unicode), 6
- GEORGIAN_SUPPLEMENT (Unicode), 6
- georgian_supplement (Unicode), 6
- GLAGOLITIC (Unicode), 6
- glagolitic (Unicode), 6
- GOTHIC (Unicode), 6
- gothic (Unicode), 6
- GRANTHA (Unicode), 6
- grantha (Unicode), 6
- GREEK_AND_COPTIC (Unicode), 6
- greek_and_coptic (Unicode), 6
- GREEK_EXTENDED (Unicode), 6
- greek_extended (Unicode), 6
- GUJARATI (Unicode), 6
- gujarati (Unicode), 6
- GURMUKHI (Unicode), 6
- gurmukhi (Unicode), 6

- HALFWIDTH_AND_FULLWIDTH_FORMS (Unicode), 6
- halfwidth_and_fullwidth_forms (Unicode), 6
- HANGUL_COMPATIBILITY_JAMO (Unicode), 6
- hangul_compatibility_jamo (Unicode), 6
- HANGUL_JAMO (Unicode), 6
- hangul_jamo (Unicode), 6
- HANGUL_JAMO_EXTENDED_A (Unicode), 6
- hangul_jamo_extended_a (Unicode), 6
- HANGUL_JAMO_EXTENDED_B (Unicode), 6
- hangul_jamo_extended_b (Unicode), 6
- HANGUL_SYLLABLES (Unicode), 6
- hangul_syllables (Unicode), 6
- HANUNOO (Unicode), 6
- hanunoo (Unicode), 6
- HEBREW (Unicode), 6
- hebrew (Unicode), 6
- HIRAGANA (Unicode), 6
- hiragana (Unicode), 6

- IMPERIAL_ARAMAIC (Unicode), 6
- imperial_aramaic (Unicode), 6
- INVISIBLE_OPERATORS (Unicode), 6
- invisible_operators (Unicode), 6
- IPA_EXTENSIONS (Unicode), 6
- ipa_extensions (Unicode), 6

- JAPANESE_CHESS (Unicode), 6
 japanese_chess (Unicode), 6
 JAVANESE (Unicode), 6
 javanese (Unicode), 6

 KAITHI (Unicode), 6
 kaithi (Unicode), 6
 KANA_SUPPLEMENT (Unicode), 6
 kana_supplement (Unicode), 6
 KANBUN (Unicode), 6
 kanbun (Unicode), 6
 KANGXI_RADICALS (Unicode), 6
 kangxi_radicals (Unicode), 6
 KANGXI_RADICALS_SUPPLEMENT (Unicode), 6
 kangxi_radicals_supplement (Unicode), 6
 KANNADA (Unicode), 6
 kannada (Unicode), 6
 KATAKANA (Unicode), 6
 katakana (Unicode), 6
 KATAKANA_PHONETIC_EXTENSIONS (Unicode),
 6
 katakana_phonetic_extensions (Unicode),
 6
 KAYAH_LI (Unicode), 6
 kayah_li (Unicode), 6
 KHAROSHTHI (Unicode), 6
 kharoshthi (Unicode), 6
 KHMER (Unicode), 6
 khmer (Unicode), 6
 KHMER_SYMBOLS (Unicode), 6
 khmer_symbols (Unicode), 6
 KHOJKI (Unicode), 6
 khojki (Unicode), 6
 KHUDAWADI (Unicode), 6
 khudawadi (Unicode), 6

 LAO (Unicode), 6
 lao (Unicode), 6
 LATIN (Unicode), 6
 latin (Unicode), 6
 LATIN_1_PUNCTUATION (Unicode), 6
 latin_1_punctuation (Unicode), 6
 LATIN_1_SUPPLEMENT (Unicode), 6
 latin_1_supplement (Unicode), 6
 LATIN_EXTENDED_A (Unicode), 6
 latin_extended_a (Unicode), 6
 LATIN_EXTENDED_ADDITIONAL (Unicode), 6
 latin_extended_additional (Unicode), 6
 LATIN_EXTENDED_B (Unicode), 6
 latin_extended_b (Unicode), 6
 LATIN_EXTENDED_C (Unicode), 6
 latin_extended_c (Unicode), 6
 LATIN_EXTENDED_D (Unicode), 6
 latin_extended_d (Unicode), 6
 LATIN_EXTENDED_E (Unicode), 6
 latin_extended_e (Unicode), 6
 LATIN_LIGATURES (Unicode), 6
 latin_ligatures (Unicode), 6
 LEPCHA (Unicode), 6
 lepcha (Unicode), 6
 LETTERLIKE_SYMBOLS (Unicode), 6
 letterlike_symbols (Unicode), 6
 LIMBU (Unicode), 6
 limbu (Unicode), 6
 LINEAR_A (Unicode), 6
 linear_a (Unicode), 6
 LINEAR_B_IDEOGRAMS (Unicode), 6
 linear_b_ideograms (Unicode), 6
 LINEAR_B_SYLLABARY (Unicode), 6
 linear_b_syllabary (Unicode), 6
 LISU (Unicode), 6
 lisu (Unicode), 6
 LYCIAN (Unicode), 6
 lycian (Unicode), 6
 LYDIAN (Unicode), 6
 lydian (Unicode), 6

 MAHAJANI (Unicode), 6
 mahajani (Unicode), 6
 MAHJONG_TILES (Unicode), 6
 mahjong_tiles (Unicode), 6
 MALAYALAM (Unicode), 6
 malayalam (Unicode), 6
 MANDAIC (Unicode), 6
 mandaic (Unicode), 6
 MANICHAEAN (Unicode), 6
 manichaeian (Unicode), 6
 MATH_ARROWS (Unicode), 6
 math_arrows (Unicode), 6
 MATHEMATICAL_ALPHANUMERIC_SYMBOLS
 (Unicode), 6
 mathematical_alphanumeric_symbols
 (Unicode), 6
 MEETEI_MAYEK (Unicode), 6
 meetei_mayek (Unicode), 6
 MEETEI_MAYEK_EXTENSIONS (Unicode), 6
 meetei_mayek_extensions (Unicode), 6
 MENDE_KIKAKUI (Unicode), 6

- mende_kikakui (Unicode), 6
- MEROITIC_CURSIVE (Unicode), 6
- meroitic_cursive (Unicode), 6
- MEROITIC_HIEROGLYPHS (Unicode), 6
- meroitic_hieroglyphs (Unicode), 6
- MIAO (Unicode), 6
- miao (Unicode), 6
- MISCELLANEOUS_MATHEMATICAL_SYMBOLS_A (Unicode), 6
- miscellaneous_mathematical_symbols_a (Unicode), 6
- MISCELLANEOUS_MATHEMATICAL_SYMBOLS_B (Unicode), 6
- miscellaneous_mathematical_symbols_b (Unicode), 6
- MISCELLANEOUS_SYMBOLS_AND_PICTOGRAPHS (Unicode), 6
- miscellaneous_symbols_and_pictographs (Unicode), 6
- MISCELLANEOUS_TECHNICAL (Unicode), 6
- miscellaneous_technical (Unicode), 6
- MODI (Unicode), 6
- modi (Unicode), 6
- MODIFIER_TONE_LETTERS (Unicode), 6
- modifier_tone_letters (Unicode), 6
- MONGOLIAN (Unicode), 6
- mongolian (Unicode), 6
- MRO (Unicode), 6
- mro (Unicode), 6
- MUSICAL_SYMBOLS (Unicode), 6
- musical_symbols (Unicode), 6
- MYANMAR (Unicode), 6
- myanmar (Unicode), 6
- MYANMAR_EXTENDED_A (Unicode), 6
- myanmar_extended_a (Unicode), 6
- MYANMAR_EXTENDED_B (Unicode), 6
- myanmar_extended_b (Unicode), 6

- NABATAEAN (Unicode), 6
- nabataean (Unicode), 6
- NEW_TAI_LUE (Unicode), 6
- new_tai_lue (Unicode), 6
- NKO (Unicode), 6
- nko (Unicode), 6
- NUMBER_FORMS (Unicode), 6
- number_forms (Unicode), 6

- OGHAM (Unicode), 6
- ogham (Unicode), 6

- OL_CHIKI (Unicode), 6
- ol_chiki (Unicode), 6
- OLD_ITALIC (Unicode), 6
- old_italic (Unicode), 6
- OLD_NORTH_ARABIAN (Unicode), 6
- old_north_arabian (Unicode), 6
- OLD_PERMIC (Unicode), 6
- old_permic (Unicode), 6
- OLD_PERSIAN (Unicode), 6
- old_persian (Unicode), 6
- OLD_SOUTH_ARABIAN (Unicode), 6
- old_south_arabian (Unicode), 6
- OLD_TURKIC (Unicode), 6
- old_turkic (Unicode), 6
- OPTICAL_CHARACTER_RECOGNITION (Unicode), 6
- optical_character_recognition (Unicode), 6
- ORIYA (Unicode), 6
- oriya (Unicode), 6
- ORNAMENTAL_DINGBATS (Unicode), 6
- ornamental_dingbats (Unicode), 6
- OSMANYA (Unicode), 6
- osmanya (Unicode), 6

- PAHAHW_HMONG (Unicode), 6
- pahawh_hmong (Unicode), 6
- PAHLAVI_INSCRIPTIONAL (Unicode), 6
- pahlavi_inscriptional (Unicode), 6
- PAHLAVI_PSALTER (Unicode), 6
- pahlavi_psalter (Unicode), 6
- PALMYRENE (Unicode), 6
- palmyrene (Unicode), 6
- PAU_CIN_HAU (Unicode), 6
- pau_cin_hau (Unicode), 6
- PHAGS_PA (Unicode), 6
- phags_pa (Unicode), 6
- PHAISTOS_DISC (Unicode), 6
- phaistos_disc (Unicode), 6
- PHOENICIAN (Unicode), 6
- phoenician (Unicode), 6
- PHONETIC_EXTENSIONS (Unicode), 6
- phonetic_extensions (Unicode), 6
- PHONETIC_EXTENSIONS_SUPPLEMENT (Unicode), 6
- phonetic_extensions_supplement (Unicode), 6
- PLAYING_CARDS (Unicode), 6
- playing_cards (Unicode), 6

- PRIVATE_USE_AREA (Unicode), 6
private_use_area (Unicode), 6
- REJANG (Unicode), 6
rejang (Unicode), 6
- RUMI_NUMERAL_SYMBOLS (Unicode), 6
rumi_numeral_symbols (Unicode), 6
- RUNIC (Unicode), 6
runic (Unicode), 6
- SAMARITAN (Unicode), 6
samaritan (Unicode), 6
- SAURASHTRA (Unicode), 6
saurashtra (Unicode), 6
- SHARADA (Unicode), 6
sharada (Unicode), 6
- SHAVIAN (Unicode), 6
shavian (Unicode), 6
- SHORTHAND_FORMAT_CONTROLS (Unicode), 6
shorthand_format_controls (Unicode), 6
- SIDDHAM (Unicode), 6
siddham (Unicode), 6
- SINHALA (Unicode), 6
sinhala (Unicode), 6
- SINHALA_ARCHAIC_NUMBERS (Unicode), 6
sinhala_archaic_numbers (Unicode), 6
- SMALL_FORM_VARIANTS (Unicode), 6
small_form_variants (Unicode), 6
- SORA_SOMPENG (Unicode), 6
sora_sompeng (Unicode), 6
- SPACING_MODIFIER_LETTERS (Unicode), 6
spacing_modifier_letters (Unicode), 6
- SPECIALS (Unicode), 6
specials (Unicode), 6
- SUNDANESE (Unicode), 6
sundanese (Unicode), 6
- SUNDANESE_SUPPLEMENT (Unicode), 6
sundanese_supplement (Unicode), 6
- SUPERSCRIPTS_AND_SUBSCRIPTS (Unicode), 6
superscripts_and_subscripts (Unicode), 6
- SUPPLEMENTAL_ARROWS_A (Unicode), 6
supplemental_arrows_a (Unicode), 6
- SUPPLEMENTAL_MATHEMATICAL_OPERATORS (Unicode), 6
supplemental_mathematical_operators (Unicode), 6
- SUPPLEMENTAL_PUNCTUATION (Unicode), 6
supplemental_punctuation (Unicode), 6
- SUPPLEMENTARY_PRIVATE_USE_AREA_A (Unicode), 6
supplementary_private_use_area_a (Unicode), 6
- SUPPLEMENTARY_PRIVATE_USE_AREA_B (Unicode), 6
supplementary_private_use_area_b (Unicode), 6
- SYLOTI_NAGRI (Unicode), 6
syloiti_nagri (Unicode), 6
- SYRIAC (Unicode), 6
syriac (Unicode), 6
- TAGALOG (Unicode), 6
tagalog (Unicode), 6
- TAGBANWA (Unicode), 6
tagbanwa (Unicode), 6
- TAGS (Unicode), 6
tags (Unicode), 6
- TAI_LE (Unicode), 6
tai_le (Unicode), 6
- TAI_THAM (Unicode), 6
tai_tham (Unicode), 6
- TAI_VIET (Unicode), 6
tai_viet (Unicode), 6
- TAI_XUAN_JING_SYMBOLS (Unicode), 6
tai_xuan_jing_symbols (Unicode), 6
- TAKRI (Unicode), 6
takri (Unicode), 6
- TAMIL (Unicode), 6
tamil (Unicode), 6
- TELUGU (Unicode), 6
telugu (Unicode), 6
- THAANA (Unicode), 6
thaana (Unicode), 6
- THAI (Unicode), 6
thai (Unicode), 6
- TIBETAN (Unicode), 6
tibetan (Unicode), 6
- TIFINAGH (Unicode), 6
tifinagh (Unicode), 6
- TIRHUTA (Unicode), 6
tirhuta (Unicode), 6
- TRANSPORT_AND_MAP_SYMBOLS (Unicode), 6
transport_and_map_symbols (Unicode), 6
- UGARITIC (Unicode), 6
ugaritic (Unicode), 6
- UGC_CASED_LETTER (ugc_cased_letter), 2

- ugc_cased_letter, 2
- UGC_CLOSE_PUNCTUATION
(ugc_cased_letter), 2
- ugc_close_punctuation
(ugc_cased_letter), 2
- UGC_CONNECTOR_PUNCTUATION
(ugc_cased_letter), 2
- ugc_connector_punctuation
(ugc_cased_letter), 2
- UGC_CONTROL (ugc_cased_letter), 2
- ugc_control (ugc_cased_letter), 2
- UGC_CURRENCY_SYMBOL (ugc_cased_letter),
2
- ugc_currency_symbol (ugc_cased_letter),
2
- UGC_DASH_PUNCTUATION
(ugc_cased_letter), 2
- ugc_dash_punctuation
(ugc_cased_letter), 2
- UGC_DECIMAL_NUMBER (ugc_cased_letter), 2
- ugc_decimal_number (ugc_cased_letter), 2
- UGC_ENCLOSING_MARK (ugc_cased_letter), 2
- ugc_enclosing_mark (ugc_cased_letter), 2
- UGC_FINAL_PUNCTUATION
(ugc_cased_letter), 2
- ugc_final_punctuation
(ugc_cased_letter), 2
- UGC_FORMAT_CONTROL (ugc_cased_letter), 2
- ugc_format_control (ugc_cased_letter), 2
- UGC_INITIAL_PUNCTUATION
(ugc_cased_letter), 2
- ugc_initial_punctuation
(ugc_cased_letter), 2
- UGC_LETTER (ugc_cased_letter), 2
- ugc_letter (ugc_cased_letter), 2
- UGC_LETTER_NUMBER (ugc_cased_letter), 2
- ugc_letter_number (ugc_cased_letter), 2
- UGC_LINE_SEPARATOR (ugc_cased_letter), 2
- ugc_line_separator (ugc_cased_letter), 2
- UGC_LOWERCASE_LETTER
(ugc_cased_letter), 2
- ugc_lowercase_letter
(ugc_cased_letter), 2
- UGC_MARK (ugc_cased_letter), 2
- ugc_mark (ugc_cased_letter), 2
- UGC_MATH_SYMBOL (ugc_cased_letter), 2
- ugc_math_symbol (ugc_cased_letter), 2
- UGC_MODIFIER_LETTER (ugc_cased_letter),
2
- ugc_modifier_letter (ugc_cased_letter),
2
- UGC_MODIFIER_SYMBOL (ugc_cased_letter),
2
- ugc_modifier_symbol (ugc_cased_letter),
2
- UGC_NONSPACING_MARK (ugc_cased_letter),
2
- ugc_nonspacing_mark (ugc_cased_letter),
2
- UGC_NUMBER (ugc_cased_letter), 2
- ugc_number (ugc_cased_letter), 2
- UGC_OPEN_PUNCTUATION
(ugc_cased_letter), 2
- ugc_open_punctuation
(ugc_cased_letter), 2
- UGC_OTHER (ugc_cased_letter), 2
- ugc_other (ugc_cased_letter), 2
- UGC_OTHER_LETTER (ugc_cased_letter), 2
- ugc_other_letter (ugc_cased_letter), 2
- UGC_OTHER_NUMBER (ugc_cased_letter), 2
- ugc_other_number (ugc_cased_letter), 2
- UGC_OTHER_PUNCTUATION
(ugc_cased_letter), 2
- ugc_other_punctuation
(ugc_cased_letter), 2
- UGC_OTHER_SYMBOL (ugc_cased_letter), 2
- ugc_other_symbol (ugc_cased_letter), 2
- UGC_PARAGRAPH_SEPARATOR
(ugc_cased_letter), 2
- ugc_paragraph_separator
(ugc_cased_letter), 2
- UGC_PRIVATE_USE_CONTROL
(ugc_cased_letter), 2
- ugc_private_use_control
(ugc_cased_letter), 2
- UGC_PUNCTUATION (ugc_cased_letter), 2
- ugc_punctuation (ugc_cased_letter), 2
- UGC_SEPARATOR (ugc_cased_letter), 2
- ugc_separator (ugc_cased_letter), 2
- UGC_SPACE_SEPARATOR (ugc_cased_letter),
2
- ugc_space_separator (ugc_cased_letter),
2
- UGC_SPACING_MARK (ugc_cased_letter), 2
- ugc_spacing_mark (ugc_cased_letter), 2
- UGC_SURROGATE_CONTROL

- (ugc_cased_letter), 2
- ugc_surrogate_control
 - (ugc_cased_letter), 2
- UGC_SYMBOL (ugc_cased_letter), 2
- ugc_symbol (ugc_cased_letter), 2
- UGC_TITLECASE_LETTER
 - (ugc_cased_letter), 2
- ugc_titlecase_letter
 - (ugc_cased_letter), 2
- UGC_UNASSIGNED_CONTROL
 - (ugc_cased_letter), 2
- ugc_unassigned_control
 - (ugc_cased_letter), 2
- UGC_UPPERCASE_LETTER
 - (ugc_cased_letter), 2
- ugc_uppercase_letter
 - (ugc_cased_letter), 2
- Unicode, 6, 6, 34
- unicode_general_category, 34
- unicode_general_category
 - (ugc_cased_letter), 2
- unicode_intersect (UnicodeOperators), 29
- unicode_inverse (UnicodeOperators), 29
- unicode_operators (UnicodeOperators), 29
- unicode_property, 6
- unicode_property (up_alphabetic), 31
- unicode_setdiff (UnicodeOperators), 29
- unicode_union (UnicodeOperators), 29
- UnicodeGeneralCategory
 - (ugc_cased_letter), 2
- UnicodeOperators, 29
- UnicodeProperty (up_alphabetic), 31
- UNIFIED_CANADIAN_ABORIGINAL_SYLLABICS
 - (Unicode), 6
- unified_canadian_aboriginal_syllabics
 - (Unicode), 6
- UNIFIED_CANADIAN_ABORIGINAL_SYLLABICS_EXTENDED
 - (Unicode), 6
- unified_canadian_aboriginal_syllabics_extended
 - (Unicode), 6
- UP_ALPHABETIC (up_alphabetic), 31
- up_alphabetic, 31
- UP_ASCII_HEX_DIGIT (up_alphabetic), 31
- up_ascii_hex_digit (up_alphabetic), 31
- UP_BIDI_CONTROL (up_alphabetic), 31
- up_bidi_control (up_alphabetic), 31
- UP_BIDI_MIRRORED (up_alphabetic), 31
- up_bidi_mirrored (up_alphabetic), 31
- UP_CASE_IGNOREABLE (up_alphabetic), 31
- up_case_ignorable (up_alphabetic), 31
- UP_CASE_SENSITIVE (up_alphabetic), 31
- up_case_sensitive (up_alphabetic), 31
- UP_CASED (up_alphabetic), 31
- up_cased (up_alphabetic), 31
- UP_CHANGES_WHEN_CASEFOLDED
 - (up_alphabetic), 31
- up_changes_when_casefolded
 - (up_alphabetic), 31
- UP_CHANGES_WHEN_CASEMAPPED
 - (up_alphabetic), 31
- up_changes_when_casemapped
 - (up_alphabetic), 31
- UP_CHANGES_WHEN_LOWERCASED
 - (up_alphabetic), 31
- up_changes_when_lowercased
 - (up_alphabetic), 31
- UP_CHANGES_WHEN_NFKC_CASEFOLDED
 - (up_alphabetic), 31
- up_changes_when_nfkc_casefolded
 - (up_alphabetic), 31
- UP_CHANGES_WHEN_TITLECASED
 - (up_alphabetic), 31
- up_changes_when_titlecased
 - (up_alphabetic), 31
- UP_CHANGES_WHEN_UPPERCASED
 - (up_alphabetic), 31
- up_changes_when_uppercased
 - (up_alphabetic), 31
- UP_DASH (up_alphabetic), 31
- up_dash (up_alphabetic), 31
- UP_DEFAULT_IGNOREABLE_CODE_POINT
 - (up_alphabetic), 31
- up_default_ignorable_code_point
 - (up_alphabetic), 31
- UP_DEPRECATED (up_alphabetic), 31
- up_deprecated (up_alphabetic), 31
- UP_DIACRITIC (up_alphabetic), 31
- up_diacritic (up_alphabetic), 31
- UP_EXTENDER (up_alphabetic), 31
- up_extender (up_alphabetic), 31
- UP_HEX_DIGIT (up_alphabetic), 31
- up_hex_digit (up_alphabetic), 31
- UP_HYPHEN (up_alphabetic), 31
- up_hyphen (up_alphabetic), 31
- UP_ID_CONTINUE (up_alphabetic), 31
- up_id_continue (up_alphabetic), 31

- UP_ID_START (up_alphabetic), 31
- up_id_start (up_alphabetic), 31
- UP_IDEOGRAPHIC (up_alphabetic), 31
- up_ideographic (up_alphabetic), 31
- UP_LOWERCASE (up_alphabetic), 31
- up_lowercase (up_alphabetic), 31
- UP_MATH (up_alphabetic), 31
- up_math (up_alphabetic), 31
- UP_NONCHARACTER_CODE_POINT
 (up_alphabetic), 31
- up_noncharacter_code_point
 (up_alphabetic), 31
- UP_POSIX_ALNUM (up_alphabetic), 31
- up_posix_alnum (up_alphabetic), 31
- UP_POSIX_BLANK (up_alphabetic), 31
- up_posix_blank (up_alphabetic), 31
- UP_POSIX_GRAPH (up_alphabetic), 31
- up_posix_graph (up_alphabetic), 31
- UP_POSIX_PRINT (up_alphabetic), 31
- up_posix_print (up_alphabetic), 31
- UP_POSIX_XDIGIT (up_alphabetic), 31
- up_posix_xdigit (up_alphabetic), 31
- UP_QUOTATION_MARK (up_alphabetic), 31
- up_quotation_mark (up_alphabetic), 31
- UP_SOFT_DOTTED (up_alphabetic), 31
- up_soft_dotted (up_alphabetic), 31
- UP_TERMINAL_PUNCTUATION
 (up_alphabetic), 31
- up_terminal_punctuation
 (up_alphabetic), 31
- UP_UPPERCASE (up_alphabetic), 31
- up_uppercase (up_alphabetic), 31
- UP_WHITE_SPACE (up_alphabetic), 31
- up_white_space (up_alphabetic), 31

- VAI (Unicode), 6
- vai (Unicode), 6
- VARIATION_SELECTORS (Unicode), 6
- variation_selectors (Unicode), 6
- VARIATION_SELECTORS_SUPPLEMENT
 (Unicode), 6
- variation_selectors_supplement
 (Unicode), 6
- VEDIC_EXTENSIONS (Unicode), 6
- vedic_extensions (Unicode), 6
- VERTICAL_FORMS (Unicode), 6
- vertical_forms (Unicode), 6

- WARANG_CITI (Unicode), 6

- warang_citi (Unicode), 6

- YI_RADICALS (Unicode), 6
- yi_radicals (Unicode), 6
- YI_SYLLABLES (Unicode), 6
- yi_syllables (Unicode), 6
- YIJING_HEXAGRAM_SYMBOLS (Unicode), 6
- yijing_hexagram_symbols (Unicode), 6
- YIJING_MONO_DI_AND_TRIGRAMS (Unicode), 6
- yijing_mono_di_and_trigrams (Unicode), 6