

Kotlin static code analysis: Unicode-aware versions of character classes should be preferred

1 minute

When using POSIX classes like `\p{Alpha}` without the `(?U)` to include Unicode characters or when using hard-coded character classes like `"[a-zA-Z]"`, letters outside of the ASCII range, such as umlauts, accented letters or letter from non-Latin languages, won't be matched. This may cause code to incorrectly handle input containing such letters.

To correctly handle non-ASCII input, it is recommended to use Unicode classes like `\p{IsAlphabetic}`. When using POSIX classes, Unicode support should be enabled by using `(?U)` inside the regex.

Noncompliant Code Example

```
Regex("[a-zA-Z]")
Regex("\p{Alpha}")
Regex("""\p{Alpha}""")
```

Compliant Solution

```
Regex("""\p{IsAlphabetic}""") // matches all letters from all
languages
Regex("""\p{IsLatin}""") // matches latin letters, including umlauts
and other non-ASCII variations
Regex("""(?U)\p{Alpha}""")
Regex("(?U)\p{Alpha}")
```