

Arthur Reutenauer

UTF-8 support detection

When you need to detect if you are running an extension of $\text{T}_{\text{E}}\text{X}$ that supports UTF-8 input, you can use an extensive approach by making the list of engines that could be concerned, and check for particular control sequences like `\XeTeXversion` for $\text{X}_{\text{E}}\text{T}_{\text{E}}\text{X}$, or `\directlua` for $\text{LuaT}_{\text{E}}\text{X}$. But you can also simply check for UTF-8 directly, by counting the bytes:

Take τ , the letter Tau from the Greek alphabet, not the Latin T that looks similar. In UTF-8, its encoding form uses two bytes, which means it is read as two characters by 8-bit $\text{T}_{\text{E}}\text{X}$ engines, but only one by UTF-8 engines. Hence, the following lines detect UTF-8 engines:

```
\def\testengine#1#2!{\def\secondarg{#2}}
```

That's τ (as in Taco or $\text{T}_{\text{E}}\text{X}$),

```
\testengine  $\tau$ !\relax
```

```
UTF-8
```

```
\ifx\secondarg\empty
```

```
  is % We're UTF-8
```

```
\else
```

```
  not % We're 8-bit
```

```
\fi
```

```
supported.
```