Text File Structure

Encodings and Character Sets Articles

The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!) by Joel Spolsky

An excerpt:

> **The Single Most Important Fact About Encodings**
>
> If you completely forget everything I just explained, please remember one extremely important fact. It does not make sense to have a string without knowing what encoding it uses. You can no longer stick your head in the sand and pretend that “plain” text is ASCII.
>
> **There Ain’t No Such Thing As Plain Text**

> Almost every stupid “my website looks like gibberish” or “she can’t read my emails when I use accents” problem comes down to one naive programmer who didn’t understand the simple fact that if you don’t tell me whether a particular string is encoded using UTF-8 or ASCII or ISO 8859-1 (Latin 1) or Windows 1252 (Western European), you simply cannot display it correctly or even figure out where it ends. There are over a hundred encodings and above code point 127, all bets are off.”

What Every Programmer Absolutely, Positively Needs To Know About Encodings And Character Sets To Work With Text

> An article by Joel Spolsky entitled The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!) is a nice introduction to the topic and I greatly enjoy reading it every once in a while. I hesitate to refer people to it who have trouble understanding encoding problems though since, while entertaining, it is pretty light on actual technical details. I hope this article can shed some more light on what exactly an encoding is and just why all your text screws up when you least need it.
>
> ...
>
> **TL;DR**
>
> Any character can be encoded in many different bit sequences and any particular bit sequence can represent many different characters, depending on which encoding is used to read or write them. The reason is simply because different encodings use different numbers of bits per characters and different values to represent different characters.”

Unicode and Python

In Python 3, there is:

- one text type: `str`, which holds Unicode data and
- two byte types: `bytes` and `bytearray`

More Information

- If you’re still confused about the difference between character sets and encoding, check out these articles:
    - The difference between UTF-8 and Unicode?
    - More About Unicode in Python 2 and 3

NEXT