Supplement: UTF in Java

For Introduction to Java Programming and Data Structures
By Y. Daniel Liang

The **`writeUTF(String s)`** method writes two bytes of length information of the string **`s`** to the output stream, followed by the modified UTF-8 representation of every character in the string **`s`**. UTF-8 is a coding scheme that allows systems to operate with both ASCII and Unicode. Most operating systems use ASCII. Java uses Unicode. The ASCII character set is a subset of the Unicode character set. Since most applications need only the ASCII character set, it is a waste to represent an 8-bit ASCII character as a 16-bit Unicode character. The modified UTF-8 scheme stores a character using one, two, or three bytes. Characters are coded in one byte if their code is less than or equal to **`0x7F`**, in two bytes if their code is greater than **`0x7F`** and less than or equal to **`0x7FF`**, or in three bytes if their code is greater than **`0x7FF`**.

The initial bits of a UTF-8 character indicate whether a character is stored in one byte, two bytes, or three bytes. If the first bit is **`0`**, it is a one-byte character. If the first bits are **`110`**, it is the first byte of a two-byte sequence. If the first bits are **`1110`**, it is the first byte of a three-byte sequence. The information that indicates the number of characters in a string is stored in the first two bytes preceding the UTF-8 characters. For example, **`writeUTF("ABCDEF")`** actually writes eight bytes (i.e., **`00 06 41 42 43 44 45 46`**) to the file because the first two bytes store the number of characters in the string.

<***margin note***>UTF-8 scheme

The **`writeUTF(String s)`** method converts a string into a series of bytes in the UTF-8 format and writes them into an output stream. The **`readUTF()`** method reads a string that has been written using the **`writeUTF`** method.

The UTF-8 format has the advantage of saving a byte for each ASCII character because a Unicode character takes up two bytes and an ASCII character in UTF-8 only one byte. If most of the characters in a long string are regular ASCII characters, using UTF-8 is more efficient.