Special Topic 4.5

Strings and the `char` Type

Strings are sequences of Unicode characters (see Random Fact 4.2 on page 154). Character literals look like string literals, except that character literals are delimited by single quotes: `'H'` is a character, `"H"` is a string containing a single character.

Characters have numeric values. For example, if you look at Appendix A, you can see that the character `'H'` is actually encoded as the number 72.

You can use escape sequences (see Special Topic 4.4 on page 152) inside character literals. For example, `'\n'` is the newline character, and `'\u00E9'` is the character é.

When Java was first designed, each Unicode character was encoded as a two-byte quantity. The `char` type was intended to hold the code of a Unicode character. However, as of 2003, Unicode had grown so large that some characters needed to be encoded as pairs of `char` values. Thus, you can no longer think of a `char` value as a character. Technically speaking, a `char` value is a *code unit* in the UTF-16 encoding of Unicode. That encoding represents the most common characters as a single `char` value, and less common or *supplementary* characters as a pair of `char` values.

The `charAt` method of the `String` class returns a code unit from a string. As with the `substring` method, the positions in the string are counted starting at `0`. For example, the statement

```
String greeting = "Hello";
char ch = greeting.charAt(0);
```

sets `ch` to the value `'H'`.

However, if you use `char` variables, your programs may fail with some strings that contain international or symbolic characters. For example, the single character 𝕫 (the mathematical symbol for the set of integers) is encoded by the two code units `'\uD835'` and `'\uDD6B'`.

If you call `charAt(0)` on the string containing the single character 𝕫 (that is, the string `"\uD835\uDD6B"`), you only get the first half of a supplementary character.

Therefore, you should only use `char` values if you are absolutely sure that you won't need to encode supplementary characters.