

Forum:[Hack Forums](#)

Thread: [Check to see if char is unicode](#)

Matt AH - February 5, 2016, 4:02 pm

Hello,

I'm currently having some trouble detecting if a character is a specific Unicode.

My assignment is to code a Character Reader in which it reads an input file that contains a string and I'm supposed to print how many of the Unicode characters in the file are ASCII characters, Special characters, Characters that can be encoded in ISO 8859-1, Greek letters (?-?, ?-?), and Characters that are not in the basic multilingual plane

For example:

String: The circumference of a circle is $2\pi r^2$. H is the set of quaternions and $\mathbb{O}$ is the set of octonions.

Occurrences:

ASCII: 94

Special: 2

ISO 8859-1: 97

Greek: 1

notInBmp: 1

My problem is how do I detect a certain unicode occurrence between ASCII, special, greek letters, etc etc.

Would love some help on this.

Thanks

RedWaters - February 6, 2016, 1:41 am

```
Code String str = "2πr²?";
CharsetEncoder iso8859_1 = Charset.forName("ISO-8859-1").newEncoder();
for(int i=0; i < str.length(); i++) {
    char ch = str.charAt(i);
    if((int)ch < 128)
        System.out.println("ASCII at " + i);
    else if(iso8859_1.canEncode(ch))
        System.out.println("ISO-8859-1 at " + i);
    else if((int)ch >= 0x370 && (int)ch <= 0x3FF)
        System.out.println("Greek at " + i);
    else
        System.out.println("Unicode at " + i);
}
```

Output

Code ASCII at 0

ISO-8859-1 at 1

Greek at 2

ISO-8859-1 at 3
ASCII at 4
ISO-8859-1 at 5
Unicode at 6

To detect an ASCII char simply check if it is lower than 128, if you want the extended set check if it is lower than 256

To check a specific charset use `Charset.forName("CHARSET").newEncoder().canEncode(char)`

To check a specific unicode range see here

<http://www.unicode.org/Public/UCD/latest/ucd/Blocks.txt>

Else it is simply part of the unicode charset

Matt AH - February 6, 2016, 9:16 am

(Feb 6, 2016 02:41 AM)xSplit Wrote: [\[CITE\]](#) Code String `str = "2×?×r²?"`;

```
CharsetEncoder iso8859_1 = Charset.forName("ISO-8859-1").newEncoder();
```

```
for(int i=0; i < str.length(); i++) {
```

```
    char ch = str.charAt(i);
```

```
    if((int)ch < 128)
```

```
        System.out.println("ASCII at " + i);
```

```
    else if(iso8859_1.canEncode(ch))
```

```
        System.out.println("ISO-8859-1 at " + i);
```

```
    else if((int)ch >= 0x370 && (int)ch <= 0x3FF)
```

```
        System.out.println("Greek at " + i);
```

```
    else
```

```
        System.out.println("Unicode at " + i);
```

```
}
```

Output

Code ASCII at 0

ISO-8859-1 at 1

Greek at 2

ISO-8859-1 at 3

ASCII at 4

ISO-8859-1 at 5

Unicode at 6

To detect an ASCII char simply check if it is lower than 128, if you want the extended set check if it is lower than 256

To check a specific charset use `Charset.forName("CHARSET").newEncoder().canEncode(char)`

To check a specific unicode range see here

<http://www.unicode.org/Public/UCD/latest/ucd/Blocks.txt>

Else it is simply part of the unicode charset

`Charset.forName("CHARSET").newEncoder().canEncode(char)`

this changed my life

thanks