Unicode and Python

School of Information Studies
Syracuse University

Unicode

Industry standard

Defined by Unicode Consortium

Smallest component of text

Language Characters

Standard

- A, B, C, D
- π, μ

Special

😠

♂

Definitions in Unicode

Code Point

- Integer value—base 16
- Assigned a standard name
- No implementation, fonts
- Represented by graphical elements—GLYPH
- Represented with U+0061—decimal number 97

Some Unicode Samples

U+0061	‘a’;	Latin small letter A
U+0394	‘Δ’;	Greek Capital Letter Delta
U+007B	‘{‘;	Left Curly Bracket

Unicode Code Points

Over a million code points

- From 0 to 10FFFF (largest hexadecimal number)
- Defined in layers

Character encoding

- Used to map to binary numbers
- ASCII—English characters
  - 7 bits
- Latin-1—additional characters for Western European languages
  - 8 bits

Unicode Code Points

UTF-8

- Most widely used

- Sequence of 8-bit bytes
  - Code point < 128—represented by byte value
  - Code point >=128—sequence of 2, 3, or 4 bytes

Unicode in Python

Every string in Unicode is using UTF-8

Problem is I/O

- Python interpreter to terminal output
- Python print function to terminal output
- Files in different OS
- Databases like MongoDB, Microsoft Word, browsers

Unicode in Python Interpreter

```
>>> 15                # the decimal number 15
15
>>>0xFF               # hexadecimal numbers
255
>>>0x7F
127
>>>’\u0394’           # using 4 hex digits
Δ
>>>’\U00000394’       # using 8 hex digits
Δ
>>>’\N[GREEK CAPITAL LETTER DELTA]’
Δ
```

Unicode Functions

bytes.decode ()

- Converts from bytes to Unicode strings

str.encode ()

- Converts from strings to bytes
- To output text to different devices