











































# Unicode solutions in Python 2 and 3

|  | 00A       | 00B       | 00C       | 00D       | 00E       | 00F       |
|--|-----------|-----------|-----------|-----------|-----------|-----------|
|   | Ñ<br>00A0 | ◌<br>00B0 | À<br>00C0 | Ð<br>00D0 | à<br>00E0 | ð<br>00F0 |
|   | İ<br>00A1 | ±<br>00B1 | Á<br>00C1 | Ñ<br>00D1 | á<br>00E1 | ñ<br>00F1 |
|   | ç<br>00A2 | ²<br>00B2 | Â<br>00C2 | Ò<br>00D2 | â<br>00E2 | ò<br>00F2 |
|   | £<br>00A3 | ³<br>00B3 | Ã<br>00C3 | Ó<br>00D3 | ã<br>00E3 | ó<br>00F3 |
|   | □<br>00A4 | ´<br>00B4 | Ä<br>00C4 | Ô<br>00D4 | ä<br>00E4 | ô<br>00F4 |
|   | ¥<br>00A5 | μ<br>00B5 | Å<br>00C5 | Õ<br>00D5 | å<br>00E5 | õ<br>00F5 |
|   | ı<br>00A6 | ¶<br>00B6 | Æ<br>00C6 | Ö<br>00D6 | æ<br>00E6 | ö<br>00F6 |
|   | §<br>00A7 | ·<br>00B7 | Ç<br>00C7 | ×<br>00D7 | ç<br>00E7 | ÷<br>00F7 |
|   | ğ<br>00A8 | ¸<br>00B8 | È<br>00C8 | Ø<br>00D8 | è<br>00E8 | ø<br>00F8 |
|   | ©<br>00A9 | ¹<br>00B9 | É<br>00C9 | Ù<br>00D9 | é<br>00E9 | ù<br>00F9 |
|   | ä<br>00AA | º<br>00BA | Ê<br>00CA | Ú<br>00DA | ê<br>00EA | ú<br>00FA |
|   | «<br>00AB | »<br>00BB | Ë<br>00CB | Û<br>00DB | ë<br>00EB | û<br>00FB |
|   | ¬<br>00AC | ¼<br>00BC | Ì<br>00CC | Ü<br>00DC | ì<br>00EC | ü<br>00FC |
|   | ½<br>00AD | ½<br>00BD | Í<br>00CD | Ý<br>00DD | í<br>00ED | ý<br>00FD |
|   | ®<br>00AE | ¾<br>00BE | Î<br>00CE | Þ<br>00DE | î<br>00EE | þ<br>00FE |
|  | —<br>00AF | ¿<br>00BF | Ï<br>00CF | ß<br>00DF | ï<br>00EF | ÿ<br>00FF |

| 1F61  | 1F62  | 1F63  | 1F64  |
|---|---|---|---|
|    |    |    |    |
| 1F610   | 1F620   | 1F630   | 1F640   |
|    |    |    |    |
| 1F611   | 1F621   | 1F631   | 1F641   |
|    |    |    |    |
| 1F612   | 1F622   | 1F632   | 1F642   |
|    |    |    |    |
| 1F613   | 1F623   | 1F633   |   |
|   |   |   |   |
| 1F614   | 1F624   | 1F634   |   |
|  |  |  |  |
| 1F615   | 1F625   | 1F635   | 1F645   |
|  |  |  |  |
| 1F616   | 1F626   | 1F636   | 1F646   |
|  |  |  |  |
| 1F617   | 1F627   | 1F637   | 1F647   |
|  |  |  |  |
| 1F618   | 1F628   | 1F638   | 1F648   |
|  |  |  |  |
| 1F619   | 1F629   | 1F639   | 1F649   |
|  |  |  |  |
| 1F61A   | 1F62A   | 1F63A   | 1F64A   |
|  |  |  |  |
| 1F61B   | 1F62B   | 1F63B   | 1F64B   |
|  |  |  |  |
| 1F61C   | 1F62C   | 1F63C   | 1F64C   |
|  |  |  |  |
| 1F61D   | 1F62D   | 1F63D   | 1F64D   |
|  |  |  |  |

| 053   | 054         | 055   | 056   | 057         | 058   |
|---|-------------|---|---|-------------|---|
|  | Ƶ<br>(0540) | ƶ<br>(0550)   |  | Ʒ<br>(0570) | Ƹ<br>(0580)   |
| ƹ<br>(0531)   | ƺ<br>(0541) | ƻ<br>(0551)   | Ƽ<br>(0561)   | ƽ<br>(0571) | ƾ<br>(0581)   |
| ƿ<br>(0532)   | ƻ<br>(0542) | Ƽ<br>(0552)   | ƽ<br>(0562)   | ƾ<br>(0572) | ƿ<br>(0582)   |
| ƿ<br>(0533)   | ƻ<br>(0543) | Ƽ<br>(0553)   | ƽ<br>(0563)   | ƾ<br>(0573) | ƿ<br>(0583)   |
| ƿ<br>(0534)   | ƻ<br>(0544) | Ƽ<br>(0554)   | ƽ<br>(0564)   | ƾ<br>(0574) | ƿ<br>(0584)   |
| ƿ<br>(0535)   | ƻ<br>(0545) | Ƽ<br>(0555)   | ƽ<br>(0565)   | ƾ<br>(0575) | ƿ<br>(0585)   |
| ƿ<br>(0536)   | ƻ<br>(0546) | Ƽ<br>(0556)   | ƽ<br>(0566)   | ƾ<br>(0576) | ƿ<br>(0586)   |
| ƿ<br>(0537)   | ƻ<br>(0547) |  | ƽ<br>(0567)   | ƾ<br>(0577) | ƿ<br>(0587)   |
| ƿ<br>(0538)   | ƻ<br>(0548) |  | ƽ<br>(0568)   | ƾ<br>(0578) |  |
| ƿ<br>(0539)   | ƻ<br>(0549) | Ƽ<br>(0559)   | ƽ<br>(0569)   | ƾ<br>(0579) | ƿ<br>(0589)   |
| ƿ<br>(053A)   | ƻ<br>(054A) | Ƽ<br>(055A)   | ƽ<br>(056A)   | ƾ<br>(057A) | ƿ<br>(058A)   |
| ƿ<br>(053B)   | ƻ<br>(054B) | Ƽ<br>(055B)   | ƽ<br>(056B)   | ƾ<br>(057B) |  |
| ƿ<br>(053C)   | ƻ<br>(054C) | Ƽ<br>(055C)   | ƽ<br>(056C)   | ƾ<br>(057C) |  |
| ƿ<br>(053D)   | ƻ<br>(054D) | Ƽ<br>(055D)   | ƽ<br>(056D)   | ƾ<br>(057D) |  |
| ƿ<br>(053E)   | ƻ<br>(054E) | Ƽ<br>(055E)   | ƽ<br>(056E)   | ƾ<br>(057E) |  |

| 314   | 1315  | 1316  | 1317  | 1318  | 1319  | 131A   | 131B   |
|-------|-------|-------|-------|-------|-------|--------|--------|
|       |       |       |       |       |       |        |        |
| 13140 | 13150 | 13160 | 13170 | 13180 | 13190 | 131A0  | 131B0  |
|       |       |       |       |       |       |        |        |
| 13141 | 13151 | 13161 | 13171 | 13181 | 13191 | 131A1  | 131B1  |
|       |       |       |       |       |       |        |        |
| 13142 | 13152 | 13162 | 13172 | 13182 | 13192 | 131A2  | 131B2  |
|       |       |       |       |       |       |        |        |
| 13143 | 13153 | 13163 | 13173 | 13183 | 13193 | 131A3  | 131B3  |
|       |       |       |       |       |       |        |        |
| 13144 | 13154 | 13164 | 13174 | 13184 | 13194 | 131A4  | 131B4  |
|       |       |       |       |       |       |        |        |
| 13145 | 13155 | 13165 | 13175 | 13185 | 13195 | 131A5  | 131B5  |
|       |       |       |       |       |       |        |        |
| 13146 | 13156 | 13166 | 13176 | 13186 | 13196 | 131A6  | 131B6  |
|       |       |       |       |       |       |        |        |
| 13147 | 13157 | 13167 | 13177 | 13187 | 13197 | 131A7  | 131B7  |
|       |       |       |       |       |       |        |        |
| 13148 | 13158 | 13168 | 13178 | 13188 | 13198 | 131A8  | 131B8  |
|       |       |       |       |       |       |        |        |
| 13149 | 13159 | 13169 | 13179 | 13189 | 13199 | 131A9  | 131B9  |
|       |       |       |       |       |       |        |        |
| 1314A | 1315A | 1316A | 1317A | 1318A | 1319A | 131A1A | 131B1A |
|       |       |       |       |       |       |        |        |
| 1314B | 1315B | 1316B | 1317B | 1318B | 1319B | 131A1B | 131B1B |
|       |       |       |       |       |       |        |        |
| 1314C | 1315C | 1316C | 1317C | 1318C | 1319C | 131A1C | 131B1C |
|       |       |       |       |       |       |        |        |
| 1314D | 1315D | 1316D | 1317D | 1318D | 1319D | 131A1D | 131B1D |

|  | 0D1  | 0D2  | 0D3  | 0D4  | 0D5  | 0D6  |
|--|------|------|------|------|------|------|
|  | എ    | റ    | ര    | ീ    |      | ഃ    |
|  | 0D10 | 0D20 | 0D30 | 0D40 |      | 0D60 |
|  |      | ഡ    | റ    | ു    |      | ഊ    |
|  |      | 0D21 | 0D31 | 0D41 |      | 0D61 |
|  | ഒ    | ഘ    | ല    | ു    |      | ു    |
|  | 0D12 | 0D22 | 0D32 | 0D42 |      | 0D62 |
|  | ഓ    | ണ    | ള    | ു    |      | ു    |
|  | 0D13 | 0D23 | 0D33 | 0D43 |      | 0D63 |
|  | ഔ    | ത    | ഴ    | ു    |      |      |
|  | 0D14 | 0D24 | 0D34 | 0D44 |      |      |
|  | ക    | ഥ    | വ    |      |      |      |
|  | 0D15 | 0D25 | 0D35 |      |      |      |
|  | ഖ    | ദ    | ശ    | െ    |      | ഓ    |
|  | 0D16 | 0D26 | 0D36 | 0D46 |      | 0D66 |
|  | ഗ    | ധ    | ഷ    | േ    | ൌ    | ഘ    |
|  | 0D17 | 0D27 | 0D37 | 0D47 | 0D57 | 0D67 |
|  | ഘ    | ന    | സ    | ൈ    |      | ഊ    |
|  | 0D18 | 0D28 | 0D38 | 0D48 |      | 0D68 |
|  | ങ    | ഹ    |      |      |      | ന    |
|  | 0D19 | 0D29 | 0D39 |      |      | 0D69 |
|  | ച    | പ    |      | ൊ    |      | ർ    |
|  | 0D1A | 0D2A | 0D3A | 0D4A |      | 0D6A |
|  | ശ    | ഫ    |      | ോ    |      | ർ    |
|  | 0D1B | 0D2B |      | 0D4B |      | 0D6B |
|  | ജ    | ബ    |      | ൗ    |      | ന    |
|  | 0D1C | 0D2C |      | 0D4C |      | 0D6C |
|  | സ്വ  | ഭ    | ്    | ർ    |      | ഊ    |
|  | 0D1D | 0D2D | 0D3D | 0D4D |      | 0D6D |
|  | ണ    | മ    | ാ    |      |      | വു   |
|  | 0D1E | 0D2E | 0D3E | 0D4E |      | 0D6E |

| HEX            | C                         | J                       | K                       | V            |
|----------------|---------------------------|-------------------------|-------------------------|--------------|
| 50D0<br>人 9 12 | 僖<br>G5-3271<br>H4-9003   | 僖<br>T3-4588<br>J1-3238 | 僖<br>K2-234C<br>K2-234C |              |
| 50D1<br>人 9 12 | 僑<br>G1-4768<br>H81-8964  | 僑<br>J1-477A<br>J0-3623 | 僑<br>K0-4089            | 僑<br>V1-4C27 |
| 50D2<br>人 9 12 | 僭<br>G3-3238              | 僭<br>T4-422D            |                         |              |
| 50D3<br>人 9 12 | 債<br>G3-3165<br>H82-EDF9  | 債<br>T3-4021<br>J1-3239 | 債<br>K2-234D            | 債<br>V2-4E5D |
| 50D4<br>人 9 12 | 傳<br>G3-3237<br>H82-EDF1  | 傳<br>T3-4877<br>J1-323A | 傳<br>K2-234E            |              |
| 50D5<br>人 9 12 | 僕<br>G1-484D<br>H81-8982  | 僕<br>J1-6778<br>J0-494D | 僕<br>K0-5C52            | 僕<br>V1-4C28 |
| 50D6<br>人 9 12 | 僖<br>G0-5562<br>H81-894F  | 僖<br>J1-6775            | 僖<br>K0-703A            |              |
| 50D7<br>人 9 12 | 僨<br>G3-3137<br>H82-EDF2  | 僨<br>T3-4878            | 僨<br>K2-234F            |              |
| 50D8<br>人 9 12 | 倣<br>GE-2238              | 倣<br>T3-458E<br>J1-323B | 倣<br>K2-235D            | 倣<br>V2-8A42 |
| 50D9<br>人 9 12 | 僨<br>G5-3261<br>H4-906    | 僨<br>T3-459C<br>J4-2175 | 僨<br>K2-2351            |              |
| 50DA<br>人 9 12 | 僚<br>G0-4145<br>H81-8981  | 僚<br>J1-6777<br>J0-4E3D | 僚<br>K0-5E7E            | 僚<br>V1-4C29 |
| 50DB<br>人 9 12 | 倣<br>G3-3232<br>H82-EDF 5 | 倣<br>T2-487D            | 倣<br>K1-688A            |              |
| 50DC<br>人 9 12 | 僨<br>G3-3238              | 僨<br>T3-459D<br>J1-323C | 僨<br>K2-2352            |              |
| 50DD<br>人 9 12 | 倣<br>G3-3239<br>H82-EDF7  | 倣<br>T2-487D<br>J1-323D | 倣<br>K2-2353            | 倣<br>V2-8A43 |
| 50DE<br>人 9 12 | 僞<br>G1-4E31<br>G1-4E31   | 僞<br>J4-257E<br>J0-5128 | 僞<br>K0-6A8A            |              |
| 50DF<br>人 9 12 | 僣<br>GE-2239<br>H4-944B   | 僣<br>T4-423E<br>J1-323E | 僣<br>K2-235A            |              |
| 50E0<br>人 9 12 | 僖<br>GE-223A<br>H82-EDFE  | 僖<br>T3-405D            | 僖<br>K2-2355            |              |
| 50E1<br>人 9 12 | 僣<br>G5-3269<br>H4-9C72   | 僣<br>T3-457D            | 僣<br>J4-217C            | 僣<br>K2-2356 |
| 50E2<br>人 9 12 | 僣<br>G5-326F              | 僣<br>T3-457E            | 僣<br>J1-323F            | 僣<br>K1-6F32 |
| 50E3<br>人 9 12 | 僣                         | 僣                       | 僣                       | 僣            |

# Unicode solutions in Python 2 and 3

Latin-1

Armenian

Malayalam

emoticons

Egyptian  
Hieroglyphs

CJK  
Unified  
Ideographs

# Dance of the codepages

00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F

00 KOI8-R

10

20 ! " # \$ % & ' ( ) \* + , - . /

30 0 1 2 3 4 5 6 7 8 9 : ; < = > ?

40 @ A B C D E F G H I J K L M N O

50 P Q R S T U V W X Y Z [ \ ] ^ \_

60 ` a b c d e f g h i j k l m n o

70 p q r s t u v w x y z { | } ~

80

90

A0

B0

C0

D0

E0

F0

í „ ... † ‡ €

“ ” • — —

J α Γ ∫ §

i γ μ ¶



Video: <https://www.youtube.com/watch?v=J4qioAacrYo>

Source code: <http://bit.ly/10qt0MZ>



# Why Unicode

- Too many incompatible single-byte encodings
- Enough incompatible multi-byte encodings
- Separate concepts:
  - character identity: one **code point** for each abstract character
    - U+0041 → LATIN CAPITAL LETTER A
    - U+096C → DEVANAGARI DIGIT SIX
  - binary representation: multiple **encodings**

|   |  |  |
|---|--|--|
| • U+0041 → 0x41   |  | 0x41 0x00  |
| • U+096C → 0xE0 0xA5 0xAC   |  | 0x6C 0x09  |
|   |  |  |
|     |  |       |
| <div style="background-color: yellow; padding: 5px; display: inline-block;">UTF-8</div> |  | <div style="background-color: yellow; padding: 5px; display: inline-block;">UTF-16LE</div> |

# A sample of encodings

| char. | code point | ascii | latin1 | cp1252 | cp437 | gb2312 | utf-8       | utf-16le    |
|-------|------------|-------|--------|--------|-------|--------|-------------|-------------|
| A     | U+0041     | 41    | 41     | 41     | 41    | 41     | 41          | 41 00       |
| ¿     | U+00BF     | *     | BF     | BF     | A8    | *      | C2 BF       | BF 00       |
| Ã     | U+00C3     | *     | C3     | C3     | *     | *      | C3 83       | C3 00       |
| á     | U+00E1     | *     | E1     | E1     | A0    | A8 A2  | C3 A1       | E1 00       |
| Ω     | U+03A9     | *     | *      | *      | EA    | A6 B8  | CE A9       | A9 03       |
| ꣳ     | U+06BF     | *     | *      | *      | *     | *      | DA BF       | BF 06       |
| “     | U+201C     | *     | *      | 93     | *     | A1 B0  | E2 80 9C    | 1C 20       |
| €     | U+20AC     | *     | *      | 80     | *     | *      | E2 82 AC    | AC 20       |
| Г     | U+250C     | *     | *      | *      | DA    | A9 B0  | E2 94 8C    | 0C 25       |
| 气     | U+6C14     | *     | *      | *      | *     | C6 F8  | E6 B0 94    | 14 6C       |
| 氣     | U+6C23     | *     | *      | *      | *     | *      | E6 B0 A3    | 23 6C       |
| ♫     | U+1D11E    | *     | *      | *      | *     | *      | F0 9D 84 9E | 34 D8 1E DD |

Figure 4-1 of Fluent Python

# .encode() vs .decode()

- “Humans use text. Computers speak bytes.”
  - Esther Nam and Travis Fischer in *Character encoding and Unicode in Python (Pycon US 2014)*
- Use .encode() to convert **human** text to **bytes**
- Use .decode() to convert **bytes** to **human** text

2.7 gotcha:  
the methods  
.encode() and .decode()  
exist in **str** and **unicode**

# Unicode database

```
$ python3 numerics_demo.py
U+0031  1      re_dig isdig  isnum  1.00  DIGIT ONE
U+00bc  ¼      -      -      isnum  0.25  VULGAR FRACTION ONE QUARTER
U+00b2  ²      -      isdig  isnum  2.00  SUPERSCRIPT TWO
U+0969  ३      re_dig isdig  isnum  3.00  DEVANAGARI DIGIT THREE
U+136b  ፫      -      isdig  isnum  3.00  ETHIOPIC DIGIT THREE
U+216b  XII     -      -      isnum  12.00 ROMAN NUMERAL TWELVE
U+2466  ⑦      -      isdig  isnum  7.00  CIRCLED DIGIT SEVEN
U+2480  ⑬      -      -      isnum  13.00 PARENTHESESIZED NUMBER THIRTEEN
U+3285  ⑥      -      -      isnum  6.00  CIRCLED IDEOGRAPH SIX
$ █
```

# Unicode database

```
$ python3 numerics_demo.py
U+0031      1      re_dig isdig  isnum  1.00  DIGIT ONE
U+00bc      ¼      -      -      isnum  0.25  VULGAR FRACTION ONE QUARTER
U+00b2      ²      -      isdig  isnum  2.00  SUPERSCRIPT TWO
U+0969      ३      re_dig isdig  isnum  3.00  DEVANAGARI DIGIT THREE
U+136b      ፫      -      isdig  isnum  3.00  ETHIOPIC DIGIT THREE
U+216b      XII    -      -      isnum  12.00 ROMAN NUMERAL TWELVE
U+2466      ⑦      -      isdig  isnum  7.00  CIRCLED DIGIT SEVEN
U+2480      (13)   -      -      -
U+3285      ⑥      -      -      -
$
```

```
1 import unicodedata
2 import re
3
4 re_digit = re.compile(r'\d')
5
6 sample = '1\xbc\xb2\u0969\u136b\u216b\u2466\u2480\u3285'
7
8 for char in sample:
9     print('U+%04x' % ord(char),                # <A>
10         char.center(6),                        # <B>
11         're_digit' if re_digit.match(char) else '-', # <C>
12         'isdig' if char.isdigit() else '-',      # <D>
13         'isnum' if char.isnumeric() else '-',    # <E>
14         format(unicodedata.numeric(char), '5.2f'), # <F>
15         unicodedata.name(char),                  # <G>
16         sep='\t')
17
```

Characters: 578 - Words: 57



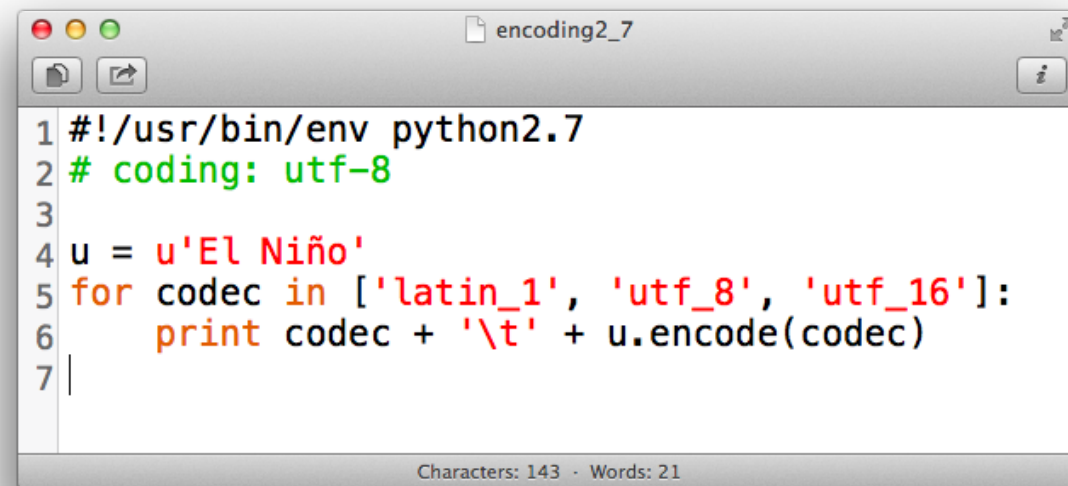
# Coping with Unicode Errors

- `SyntaxError`
  - A `.py` file is loaded with contents in an unexpected encoding
- `UnicodeDecodeError`
  - A binary sequence contains bytes that are not valid in the expected encoding
- `UnicodeEncodeError`
  - A Unicode string contains codepoints that have no representation in the desired encoding

# Coping with SyntaxError

- A .py file is loaded with contents in an unexpected encoding
  - The source file encoding is not the default, and no `# encoding` comment was found.
  - The source file encoding is not the one declared in the `# encoding` comment
- Defaults:
  - Python 2.7 == ASCII
  - Python 3.x == UTF-8

2.7 gotcha:  
default source  
encoding is ASCII



```
1#!/usr/bin/env python2.7
2# coding: utf-8
3
4u = u'El Niño'
5for codec in ['latin_1', 'utf_8', 'utf_16']:
6    print codec + '\t' + u.encode(codec)
7|
```

Characters: 143 · Words: 21

# Best practice

## The Unicode sandwich



bytes → str

100% str

str → bytes

Decode bytes on input,

process text only,

encode text on output.