Internationalization and Issues with Non-ASCII Characters (saved)

English and ASCII

- **ASCII: American Standard Code for Information Interchange**
  - 7-bit character encoding standard: 128 valid codes
  - Range: 0x00 – 0x7F [$(0000\ 0000)_2$ to $(0111\ 1111)_2$]
  - Includes alphabets (upper and lower cases), digits, punctuations, common symbols, control characters
  - Worked (relatively) well for English typewriting

Written Scripts

- Latin: 36% (2.6B people)
- Chinese: 18% (1.3B)
- Devanagari: 14% (1B)
- Arabic: 14% (1B)
- Cyrillic: 4% (0.3B)
- Dravidian: 3.5% (0.25B)

Other Character Encodings

- IBM EBCDIC
- Latin-1
- JIS: Japanese Industrial Standards
- CCCII: Chinese Character Code for Information Interchange
- EUC: Extended Unix Code
- Numerous other national standards

- Unicode and UTF-8

Let's see an example: Résumé

Python 3

```
>>> text1="Résumé"
>>> len(text1)
6
>>> text1
'Résumé'
```

Python 2

```
>>> text1="Résumé"
>>> len(text1)
8
>>> text1
'R\xc3\xa9sum\xc3\xa9'
```

Take Home Concepts

- Diversity in Text
- ASCII and other character encodings
- Handling text in UTF-8