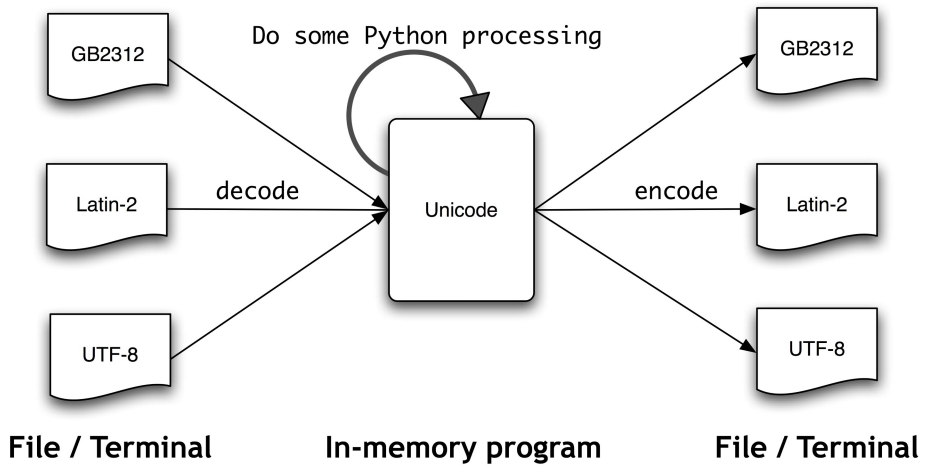
Unicode for korean characters

Wednesday, May 27, 2020 11:31 PM



https://www.nltk.org/book/ch03.html
3.3 Text Processing with Unicode
Unicode supports over a million characters including Korean, Chinese, Arabic, etc.

- Code point: Each character is assigned a number, called a **code point**.
- From a Unicode perspective, characters are abstract entities which can be realized (visualized) as one or more **glyphs**.
- Only glyphs appear on your screen or be printed on paper.
- A **font** is a mapping from characters to glyphs.

https://docs.python.org/3/howto/unicode.html
Introduction to Unicode

- A code point value is an integer in **[0, 0x10FFFF]**
- In the standard, a code point is written using the notation U+3C01 to mean the character with code point value 0x3C01.

UTF-8

- UTF: Unicode Transformation Format
- 8 means that 8-bit values are used in the encoding.
- One of the most commonly used encoding, and python 3 default.
- Rule:
    1. If the code point is < 128, it's represented by the corresponding byte value (ASCII)
    2. If the code point is >= 128, it's turned into a sequence of two, three, or four bytes, where each byte of the sequence is between 128 and 255.
- Properties:
    1. It can handle any Unicode code point
    2. NULL character is U+0000 only. It can be processed by C functions such as strcpy().
    3. ASCII is valid.

4. Fairly compact. The majority of commonly used characters can be represented with one or two bytes.
5. If bytes are corrupted or ost, it's possible to determine the start of the next UTF-8-encoded code point and resynchronize.

https://en.wikipedia.org/wiki/Korean_language_and_computers

https://en.wikipedia.org/wiki/Hangul_Syllables
Hangul in Unicode

- Hangul Syllables (완성형한글코드)
  - Hangul Syllables is a Unicode block containing precomposed Hangul syllable blocks for modern Korean.
  - Range: U+AC00 … U+D7AF
  - Source Standards: KS C 5601-1992

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+AC0x	가	각	갂	갃	간	갅	갆	갇	갈	갉	갊	갋	갌	갍	갎	갏
U+AC1x	감	갑	값	갓	갔	강	갖	갗	갘	같	갚	갛	개	객	갞	갟
U+AC2x	갠	갡	갢	갣	갤	갥	갦	갧	갨	갩	갪	갫	갬	갭	갮	갯
U+AC3x	갰	갱	갲	갳	갴	갵	갶	갷	갸	갹	갺	갻	갼	갽	갾	갿
U+AC4x	걀	걁	걂	걃	걄	걅	걆	걇	걈	걉	걊	걋	걌	걍	걎	걏
U+AC5x	걐	걑	걒	걓	걔	걕	걖	걗	걘	걙	걚	걛	걜	걝	걞	걟
U+AC6x	걠	걡	걢	걣	거	걥	걦	걧	걨	걩	걪	걫	걬	걭	걮	걯
U+D74x	퓕	퓖	퓗	퓘	퓜	퓝	퓞	퓟	퓠	퓡	퓢	퓣	퓤	퓥	퓦	퓧
U+D75x	흐	흑	흒	흓	흔	흕	흖	흗	흘	흙	흚	흛	흜	흝	흞	흟
U+D76x	흠	흡	흢	흣	흤	흥	흦	흧	흨	흩	흪	흫	희	흭	흮	흯
U+D77x	흰	흱	흲	흳	흴	흵	흶	흷	흸	흹	흺	흻	흼	흽	흾	흿
U+D78x	힀	힁	힂	힃	힄	힅	힆	힇	히	힉	힊	힋	힌	힍	힎	힏
U+D79x	힐	힑	힒	힓	힔	힕	힖	힗	힘	힙	힚	힛	힜	힝	힞	힟
U+D7Ax	힠	힡	힢	힣												

https://en.wikipedia.org/wiki/Hangul_Jamo_(Unicode_block)
조합형 한글을 위한 Unicode code points

Hangul Jamo (Korean: 한글 자모, Korean pronunciation: [ˈha:ngɯl tɕa:mo]) is a Unicode block containing positional (Choseong, Jungseong, and Jongseong) forms of the Hangul consonant and vowel clusters. They can be used to dynamically compose

syllables that are not available as precomposed Hangul syllables in Unicode, specifically archaic syllables containing sounds that have since merged phonetically with other sounds in modern pronunciation.

Hangul Jamo[1]

Official Unicode Consortium code chart (PDF)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+110x	ᄀ	ᄁ	ᄂ	ᄃ	ᄄ	ᄅ	ᄆ	ᄇ	ᄈ	ᄉ	ᄊ	ᄋ	ᄌ	ᄍ	ᄎ	ᄏ
U+111x	ᄐ	ᄑ	ᄒ	ᅡ	ᅢ	ᅣ	ᅤ	ᅥ	ᅦ	ᅧ	ᅨ	ᅩ	ᅪ	ᅫ	ᅬ	ᅭ
U+112x	ᅮ	ᅯ	ᅰ	ᅱ	ᅲ	ᅳ	ᅴ	ᅵ	ᅶ	ᅷ	ᅸ	ᅹ	ᅺ	ᅻ	ᅼ	ᅽ
U+113x	ᅾ	ᅿ	ᆀ	ᆁ	ᆂ	ᆃ	ᆄ	ᆅ	ᆆ	ᆇ	ᆈ	ᆉ	ᆊ	ᆋ	ᆌ	ᆍ
U+114x	ᆎ	ᆏ	ᆐ	ᆑ	ᆒ	ᆓ	ᆔ	ᆕ	ᆖ	ᆗ	ᆘ	ᆙ	ᆚ	ᆛ	ᆜ	ᆝ
U+115x	ᆞ	ᆟ	ᆠ	ᆡ	ᆢ	ᆣ	ᆤ	ᆥ	ᆦ	ᆧ	ᆨ	ᆩ	ᆪ	ᆫ	ᆬ	HCF
U+116x	HJF	ᆭ	ᆮ	ᆯ	ᆰ	ᆱ	ᆲ	ᆳ	ᆴ	ᆵ	ᆶ	ᆷ	ᆸ	ᆹ	ᆺ	ᆻ
U+117x	ᆼ	ᆽ	ᆾ	ᆿ	ᇀ	ᇁ	ᇂ	ᇃ	ᇄ	ᇅ	ᇆ	ᇇ	ᇈ	ᇉ	ᇊ	ᇋ
U+118x	ᇌ	ᇍ	ᇎ	ᇏ	ᇐ	ᇑ	ᇒ	ᇓ	ᇔ	ᇕ	ᇖ	ᇗ	ᇘ	ᇙ	ᇚ	ᇛ
U+119x	ᇜ	ᇝ	ᇞ	ᇟ	ᇠ	ᇡ	ᇢ	ᇣ	ᇤ	ᇥ	ᇦ	ᇧ	ᇨ	ᇩ	ᇪ	ᇫ
U+11Ax	ᇬ	ᇭ	ᇮ	ᇯ	ᇰ	ᇱ	ᇲ	ᇳ	ᇴ	ᇵ	ᇶ	ᇷ	ᇸ	ᇹ	ᇺ	ᇻ
U+11Bx	ᇼ	ᇽ	ᇾ	ᇿ	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ	ሇ	ለ	ሉ	ሊ	ላ
U+11Cx	ሌ	ል	ሎ	ሏ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ	ሗ	መ	ሙ	ሚ	ማ
U+11Dx	ሜ	ም	ሞ	ሟ	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ	ሧ	ረ	ሩ	ሪ	ራ
U+11Ex	ሬ	ር	ሮ	ሯ	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ	ሷ	ሸ	ሹ	ሺ	ሻ
U+11Fx	ሼ	ሽ	ሾ	ሿ	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቇ	ቈ	቉	ቊ	ቋ

Notes

1.^ As of Unicode version 13.0

2. ᄀ: Hangul *jamo* with a green background are modern-usage characters which can be converted into precomposed Hangul syllables under Unicode normalization form NFC.
Hangul *jamo* with a white background are used for archaic Korean only, and there are no corresponding precomposed Hangul syllables.
"Conjoining Jamo Behavior" (PDF). *The Unicode Standard*. March 2020.

Number	Lead	Jamo		Character reference
1	G	ㄱ	ㄱ	ᄀ
2	GG	ㄲ	ㄲ	ᄁ
3	N	ㄴ	ㄴ	ᄂ
4	D	ㄷ	ㄷ	ᄃ
5	DD	ㄸ	ㄸ	ᄄ
6	R	ㄹ	ㄹ	ᄅ
7	M	ㅁ	ㅁ	ᄆ
8	B	ㅂ	ㅂ	ᄇ
9	BB	ㅃ	ㅃ	ᄈ
10	S	ㅅ	ㅅ	ᄉ
11	SS	ㅆ	ㅆ	ᄊ
12		ㅇ	ㅇ	ᄋ
13	J	ㅈ	ㅈ	ᄌ
14	JJ	ㅉ	ㅉ	ᄍ
15	C	ㅊ	ㅊ	ᄎ
16	K	ㅋ	ㅋ	ᄏ
17	T	ㅌ	ㅌ	ᄐ
18	P	ㅍ	ㅍ	ᄑ
19	H	ㅎ	ㅎ	ᄒ

Number	Vowel	Jamo		Character reference
1	A	ㅏ	ㅏ	ᅡ
2	AE	ㅐ	ㅐ	ᅢ
3	YA	ㅑ	ㅑ	ᅣ
4	YAE	ㅒ	ㅒ	ᅤ
5	EO	ㅓ	ㅓ	ᅥ
6	E	ㅔ	ㅔ	ᅦ
7	YEO	ㅕ	ㅕ	ᅧ
8	YE	ㅖ	ㅖ	ᅨ
9	O	ㅗ	ㅗ	ᅩ
10	WA	ㅘ	ㅘ	ᅪ
11	WAE	ㅙ	ㅙ	ᅫ
12	OE	ㅚ	ㅚ	ᅬ
13	YO	ㅛ	ㅛ	ᅭ
14	U	ㅜ	ㅜ	ᅮ
15	WEO	ㅝ	ㅝ	ᅯ
16	WE	ㅞ	ㅞ	ᅰ
17	WI	ㅟ	ㅟ	ᅱ
18	YU	ㅠ	ㅠ	ᅲ
19	EU	ㅡ	ㅡ	ᅳ
20	YI	ㅢ	ㅢ	ᅴ
21	I	ㅣ	ㅣ	ᅵ

Number	Tail	Jamo		Character reference
1	G	ㄱ	ㄱ	ᆨ
2	GG	ㄲ	ㄲ	ᆩ
3	GS	ㄳ	ㄳ	ᆪ
4	N	ㄴ	ㄴ	ᆫ
5	NJ	ㄵ	ㄵ	ᆬ
6	NH	ㄶ	ㄶ	ᆭ
7	D	ㄷ	ㄷ	ᆮ
8	L	ㄹ	ㄹ	ᆯ
9	LG	ㄺ	ㄺ	ᆰ
10	LM	ㄻ	ㄻ	ᆱ
11	LB	ㄼ	ㄼ	ᆲ
12	LS	ㄽ	ㄽ	ᆳ
13	LT	ㄾ	ㄾ	ᆴ
14	LP	ㄿ	ㄿ	ᆵ
15	LH	ㅀ	ㅀ	ᆶ
16	M	ㅁ	ㅁ	ᆷ
17	B	ㅂ	ㅂ	ᆸ
18	BS	ㅄ	ㅄ	ᆹ
19	S	ㅅ	ㅅ	ᆺ
20	SS	ㅆ	ㅆ	ᆻ
21	NG	ㅇ	ㅇ	ᆼ
22	J	ㅈ	ㅈ	ᆽ
23	C	ㅊ	ㅊ	ᆾ
24	K	ㅋ	ㅋ	ᆿ
25	T	ㅌ	ㅌ	ᇀ
26	P	ㅍ	ㅍ	ᇁ
27	H	ㅎ	ㅎ	ᇂ

참고:

1. https://namu.wiki/w/%EC%99%84%EC%84%B1%ED%98%95 완성형 한글 코드
2. https://tapito.tistory.com/529 조합형/완성형/유니코드의 모든 것, 원제: 한글 및 한국어 정보 처리 코드 (한국어 정보처리의 과제), 전상훈, 1999. 01. 01 konan2@chollian.net, klipl@klipl.com
3. https://namu.wiki/w/%EC%98%9B%ED%95%9C%EA%B8%80 옛한글

```
cv = ord('각')

basevalue = ord('가')
print(basevalue, hex(basevalue), cv)
jong = (cv - basevalue) % 28
jung = ((cv - basevalue)/28) % 21 if jong >= 28 else 0
cho = (((cv - basevalue)/28)/21)%19 if jung >= 21 else 0

cho, jung, jong
```

```
44032 0xac00 44033
```

```
(0, 0, 1)
```

```
a, b, c = chr(cho+0x1100), chr(jung+0x1161), chr(jong+0x11A8 - 1)
a, b, c
```

```
('ㄱ', ' ㅏ', 'ㄱ')
```

```
a,b,c = ord(a) - 0x1100, ord(b) - 0x1161, ord(c) - 0x11A8 + 1
syllable = 0xAC00 + a*588 + b*28 + c
print(syllable, f'0x{syllable:X}', chr(syllable))
```

```
44033 0xAC01 각
```