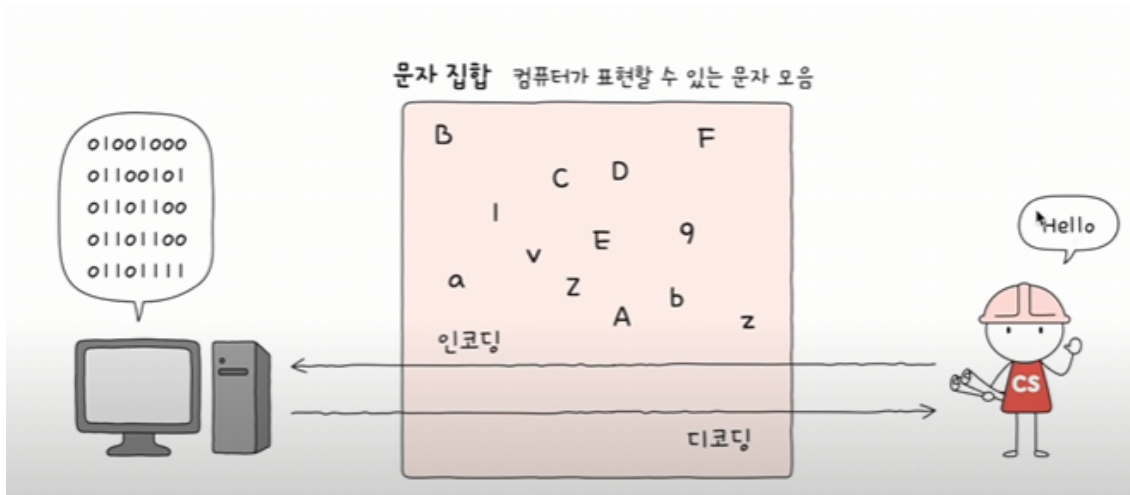


# 5강. 0과 1로 문자를 표현하는 방법

## 용어 정리



### 문자 집합

- 컴퓨터가 이해할 수 있는 문자의 모음

### 인코딩 (encoding)

- 코드화하는 과정
- 문자를 0과 1로 이루어진 문자 코드로 변환하는 과정

### 디코딩 (decoding)

- 코드를 해석하는 과정
- 0과 1로 표현된 문자 코드로 문자를 변환하는 과정

### 아스키 코드(ASCII)

가장 대중적인 문자 집합입니다.

아스키코드는 알파벳, 아라비아 숫자, 일부 특수 문자 및 제어 문자를 표현할 수 있습니다.

7비트로 하나의 문자 표현합니다. 8비트 중 1비트는 오류 검출을 위해 사용되는 패리티 비트(parity bit)

그래서  $2^7 \Rightarrow 128$ 개의 문자를 표현할 수 있습니다.

아스키 코드의 예시 및 자세한 설명은 해당 링크에서 볼 수 있습니다.

- <https://ko.wikipedia.org/wiki/ASCII>

아스키코드는 간단하게 인코딩할 수 있지만, 한글을 포함한 다른 언어 문자 및 다양한 특수 문자 표현을 못한다는 단점이 있습니다. 위와 같이 128개의 문자보다 더 많은 문자를 표현할 수 없다는 것입니다.

8비트 확장 아스키가 있지만, 여전히 부족합니다.

## 한글 인코딩 (완성 vs 조합)

한글 인코딩은 완성형과 조합형으로 나눌 수 있습니다.



## EUC-KR

KS X 1001, KS X 1003 문자집합 기반의 한글 인코딩 방식입니다.

EUC-KR은 완성형 인코딩으로 글자 하나 하나에 2바이트 크기의 코드를 부여합니다.

2byte = 16bit = 4자리 십육진수로 표현할 수 있습니다.

	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
b0a0	가	각	간	간	갈	갈	갈	감	갑	값	갓	갓	강	갓	갓	
b0b0	갈	값	갈	개	객	객	갸	갸	갸	갸	갸	갸	갸	갸	갸	
b0c0	갸	강	개	객	갸	거	거	건	건	걸	걸	검	검	것	것	
b0d0	것	걸	걸	겉	게	겐	겉	겉	겉	겉	겉	겉	겉	겉	겉	
b0e0	겉	겉	겉	겉	겉	겉	겉	겉	겉	겉	겉	겉	겉	겉	겉	
b0f0	곤	골	골	골	골	골	골	골	골	골	골	골	골	골	골	
b1a0	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	
b1b0	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	
b1c0	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	
b1d0	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	
b1e0	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	괘	

만약 직접 실습해보길 원한다면, 해당 링크를 통해 해보실 수 있습니다.

- <http://dencode.com>

약 2300개의 한글을 표현할 수 있지만, 뽕, 뽕 등과 같은 한글을 표현할 수 없습니다.

## 유니코드

- <https://ko.wikipedia.org/wiki/%EC%9C%A0%EB%8B%88%EC%BD%94%EB%93%9C>

통일된 문자 집합으로 전 세계의 모든 문자를 컴퓨터에서 일관되게 표현하고 다룰 수 있도록 설계된 산업 표준입니다.

한글, 영어, 화살표와 같은 특수 문자, 이모티콘 등을 표현할 수 있습니다.

실습 링크는 다음과 같습니다.

- <https://symbl.cc/>
- <https://onlinetools.com/utf8>

## 유니코드 인코딩 방식

- UTF-8, UTF-16, UTF-32이 있습니다.
- UTF (Unicode Transformation Format)로 유니코드 인코딩 방식입니다.
- 가변 길이 인코딩으로 인코딩 결과가 1 ~ 4 바이트입니다.

## UTF-8

## 유니코드 문자 집합

	6	7	8	9	A	B	C	D	E	F
6	도 D000	레 D001	리 D002	로 D003	라 D004	려 D005	려 D006	려 D007	려 D008	려 D009
7	뽕 D00A	푼 D00B	푼 D00C	푼 D00D	푼 D00E	푼 D00F	푼 D000	푼 D001	푼 D002	푼 D003
8	푼 D004	푼 D005	푼 D006	푼 D007	푼 D008	푼 D009	푼 D00A	푼 D00B	푼 D00C	푼 D00D
9	푼 D00E	푼 D00F	푼 D000	푼 D001	푼 D002	푼 D003	푼 D004	푼 D005	푼 D006	푼 D007
A	푼 D008	푼 D009	푼 D00A	푼 D00B	푼 D00C	푼 D00D	푼 D00E	푼 D00F	푼 D000	푼 D001
B	푼 D002	푼 D003	푼 D004	푼 D005	푼 D006	푼 D007	푼 D008	푼 D009	푼 D00A	푼 D00B
C	푼 D00C	푼 D00D	푼 D00E	푼 D00F	푼 D000	푼 D001	푼 D002	푼 D003	푼 D004	푼 D005
D	푼 D006	푼 D007	푼 D008	푼 D009	푼 D00A	푼 D00B	푼 D00C	푼 D00D	푼 D00E	푼 D00F
E	푼 D000	푼 D001	푼 D002	푼 D003	푼 D004	푼 D005	푼 D006	푼 D007	푼 D008	푼 D009
F	푼 D00A	푼 D00B	푼 D00C	푼 D00D	푼 D00E	푼 D00F	푼 D000	푼 D001	푼 D002	푼 D003

	AE0	AE1	AE2	AE3	AE4	AE5	AE6	AE7	AE8	AE9	AEA	AEB	AEC	AED	AEE	AEF
0	글 AE00	글 AE01	글 AE02	글 AE03	글 AE04	글 AE05	글 AE06	글 AE07	글 AE08	글 AE09	글 AE0A	글 AE0B	글 AE0C	글 AE0D	글 AE0E	글 AE0F
1	글 AE10	글 AE11	글 AE12	글 AE13	글 AE14	글 AE15	글 AE16	글 AE17	글 AE18	글 AE19	글 AE1A	글 AE1B	글 AE1C	글 AE1D	글 AE1E	글 AE1F
2	글 AE20	글 AE21	글 AE22	글 AE23	글 AE24	글 AE25	글 AE26	글 AE27	글 AE28	글 AE29	글 AE2A	글 AE2B	글 AE2C	글 AE2D	글 AE2E	글 AE2F
3	글 AE30	글 AE31	글 AE32	글 AE33	글 AE34	글 AE35	글 AE36	글 AE37	글 AE38	글 AE39	글 AE3A	글 AE3B	글 AE3C	글 AE3D	글 AE3E	글 AE3F
4	글 AE40	글 AE41	글 AE42	글 AE43	글 AE44	글 AE45	글 AE46	글 AE47	글 AE48	글 AE49	글 AE4A	글 AE4B	글 AE4C	글 AE4D	글 AE4E	글 AE4F
5	글 AE50	글 AE51	글 AE52	글 AE53	글 AE54	글 AE55	글 AE56	글 AE57	글 AE58	글 AE59	글 AE5A	글 AE5B	글 AE5C	글 AE5D	글 AE5E	글 AE5F
6	글 AE60	글 AE61	글 AE62	글 AE63	글 AE64	글 AE65	글 AE66	글 AE67	글 AE68	글 AE69	글 AE6A	글 AE6B	글 AE6C	글 AE6D	글 AE6E	글 AE6F
7	글 AE70	글 AE71	글 AE72	글 AE73	글 AE74	글 AE75	글 AE76	글 AE77	글 AE78	글 AE79	글 AE7A	글 AE7B	글 AE7C	글 AE7D	글 AE7E	글 AE7F
8	글 AE80	글 AE81	글 AE82	글 AE83	글 AE84	글 AE85	글 AE86	글 AE87	글 AE88	글 AE89	글 AE8A	글 AE8B	글 AE8C	글 AE8D	글 AE8E	글 AE8F

첫 코드 포인트	마지막 코드 포인트	1바이트	2바이트	3바이트	4바이트
0000	007F	0XXXXXX			
0080	07FF	110XXXX	10XXXXX		
0800	FFFF	1110XXXX	10XXXXX	10XXXXX	
10000	10FFFF	11110XXX	10XXXXX	10XXXXX	10XXXXX

한: D55C (== 1101 0101 0101 1100)

글: AE00 (== 1010 1110 0000 0000)