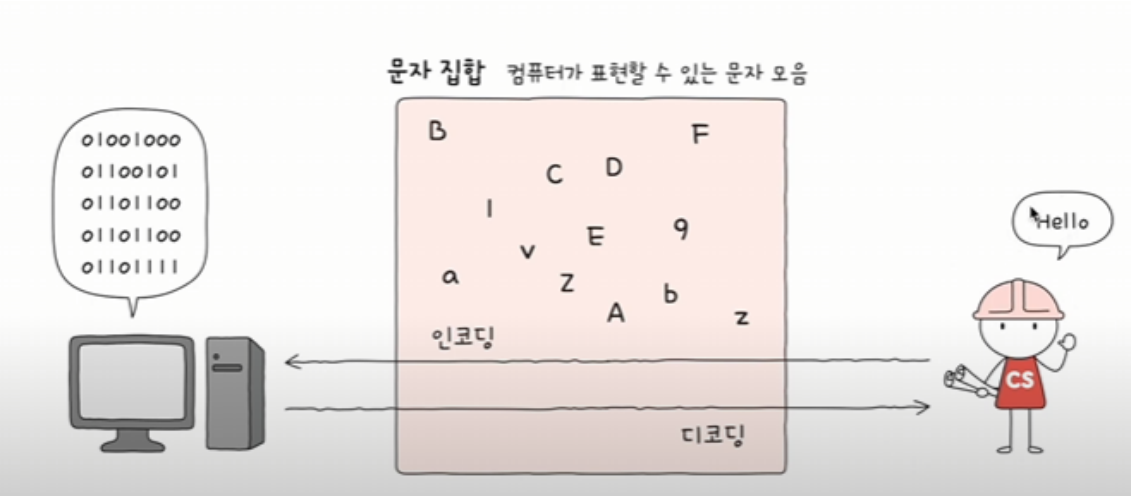
5강. 0과 1로 문자를 표현하는 방법

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	갇	갈	갉	갊	감	갑	값	갓	갔	강	갖	갗
b0b0	같	갚	갛	개	객	갠	갤	갬	갭	갯	갰	갱	갸	갹	갼	걀
b0c0	걋	걍	걔	걘	걜	거	걱	건	걷	걸	걺	검	겁	것	겄	겅
b0d0	겆	겉	겊	겋	게	겐	겔	겜	겝	겟	겠	겡	겨	격	겪	견
b0e0	겯	결	겸	겹	겻	겼	경	곁	계	곈	곌	곕	곗	고	곡	곤
b0f0	곧	골	곪	곬	곯	곰	곱	곳	공	곶	과	곽	관	괄	괆	
b1a0		괌	괍	괏	광	괘	괜	괠	괩	괬	괭	괴	괵	괸	괼	굄
b1b0	굅	굇	굉	교	굔	굘	굡	굣	구	국	군	굳	굴	굵	굶	굻
b1c0	굼	굽	굿	궁	궂	궈	궉	권	궐	궜	궝	궤	궷	귀	귁	귄
b1d0	귈	귐	귑	귓	규	균	귤	그	극	근	귿	글	긁	금	급	긋
b1e0	긍	긔	기	긱	긴	긷	길	긺	김	깁	깃	깅	깆	깊	까	깍

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

- http://dencode.com

약 2300개의 한글을 표현할 수 있지만, 뷁, 뙈 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

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

한: D55C (== 1101 0101 0101 1100)
글: AE00 (== 1010 1110 0000 0000)