과제02: 인코딩과 글꼴

언어 데이터 처리 2022-11-21 장태준 [컴퓨터 공학부 2017-17018]
https://github.com/jtjun/NLP/blob/main/hw02/hw02.md

1. 인코딩

A. 다양한 형식의 텍스트 파일

```
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src/hw02$ file *
hangul-utf16be-crlf.txt:  Big-endian UTF-16 Unicode text, with CRLF line terminators
hangul-utf16be-lf.txt:    Big-endian UTF-16 Unicode text
hangul-utf16le-crlf.txt:  Little-endian UTF-16 Unicode text, with CRLF line terminators
hangul-utf16le-lf.txt:    Little-endian UTF-16 Unicode text
hangul-utf8-bom-crlf.txt: UTF-8 Unicode (with BOM) text, with CRLF line terminators
hangul-utf8-bom-lf.txt:   UTF-8 Unicode (with BOM) text
hangul-utf8-crlf.txt:     UTF-8 Unicode text, with CRLF line terminators
hangul-utf8-lf.txt:       UTF-8 Unicode text
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src/hw02$
```

B. HEX 코드

hangul-utf8-lf.txt

```
0000000 95ed ea9c 80b8 000a
0000007
```

hangul-utf8-crlf.txt

```
0000000 95ed ea9c 80b8 0a0d
0000010
```

hangul-utf8-bom-lf.txt

```
0000000 bbef edbf 9c95 b8ea 0a80
0000012
```

hangul-utf8-bom-crlf.txt

```
0000000 bbef edbf 9c95 b8ea 0d80 000a
0000013
```

hangul-utf16le-lf.txt

```
0000000 feff d55c ae00 000a
0000010
```

hangul-utf16le-crlf.txt

```
0000000 feff d55c ae00 000d 000a
0000012
```

hangul-utf16be-lf.txt

```
0000000 fffe 5cd5 00ae 0a00
0000010
```

hangul-utf16be-crlf.txt

```
0000000 fffe 5cd5 00ae 0d00 0a00
0000012
```

한글:

- utf8: `ed 95 9c ea b8 80`
- utf16le: `d55c ae00`
- utf16be: `5cd5 00ae`

BOM:

- utf8: `ef bb bf`
- utf16le: `feff`
- utf16be: `fffe`

줄바꿈

- lf: `0a`
- crlf: `0d 0a`

`od` 명령어는 octal dump라는 의미로, 바이너리 파일을 8진수로 dump 하는 명령어다.
(https://linuxhint.com/linux-od-command/)
위 실습에선 `-x` 옵션을 통해서 16진수 HEX 코드로 출력했다. (`od -x {file_name}`)

C. 명령행 도구들

file

```
SYNOPSIS
     file [-bcdEhiklLNnprsSvzZ0] [--apple] [--extension] [--mime-encoding] [--
mime-type] [-e testname] [-F separator] [-f namefile] [-m magicfiles] [-P
name=value] file ...
     file -C [-m magicfiles]
     file [--help]

DESCRIPTION
     file tests each argument in an attempt to classify it.
     There are three sets of tests, performed in this order: filesystem tests,
magic tests, and language tests.
     The first test that succeeds causes the file type to be printed.
```

```
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP$ file hw02/hw02.md
hw02/hw02.md: Little-endian UTF-16 Unicode text, with CRLF line terminators
```

iconv

```
SYNOPSIS
     iconv [options] [-f from-encoding] [-t to-encoding] [inputfile]...

DESCRIPTION
     The  iconv  program  reads in text in one encoding and outputs the text in
another encoding.  If no input files are given, or if it is given as a dash (-),
iconv reads from standard input.
     If no output file is given, iconv writes to standard output.
     If no from-encoding is given, the default is derived from the current
locale's character encoding.
     If no to-encoding is given, the default is derived from the current locale's
character encod-ing.
```

```
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ file utf16be-crlf.txt
utf16be-crlf.txt: Big-endian UTF-16 Unicode text, with CRLF line terminators
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ iconv -f utf-16 -t utf-8 utf16be-crlf.txt > utf8-crlf.txt
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ file utf8-crlf.txt
utf8-crlf.txt: UTF-8 Unicode text, with CRLF line terminators
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$
```

iconv를 사용하여, UTF-16 BE 파일을 UTF-8로 변환해 저장하였다.

dos2unix, unix2dos

```
SYNOPSIS
          dos2unix [options] [FILE ...] [-n INFILE OUTFILE ...]
          unix2dos [options] [FILE ...] [-n INFILE OUTFILE ...]

DESCRIPTION
       The Dos2unix package includes utilities "dos2unix" and "unix2dos" to
convert plain text files in DOS or Mac format to Unix format and vice versa.

       In DOS/Windows text files a line break, also known as newline, is a
combination of two characters: a Carriage Return (CR) followed by a Line Feed
(LF). In Unix text files a line break is a single character: the Line Feed (LF).
       In Mac text files, prior to Mac OS X, a line break was single Carriage
Return (CR) character. Nowadays Mac OS uses Unix style (LF) line breaks.
```

```
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ file utf8-crlf.txt
utf8-crlf.txt: UTF-8 Unicode text, with CRLF line terminators
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ dos2unix utf8-crlf.txt
dos2unix: converting file utf8-crlf.txt to Unix format...
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ file utf8-crlf.txt
utf8-crlf.txt: UTF-8 Unicode text
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ unix2dos utf8-crlf.txt
unix2dos: converting file utf8-crlf.txt to DOS format...
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ file utf8-crlf.txt
utf8-crlf.txt: UTF-8 Unicode text, with CRLF line terminators
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$
```

실습에 앞서, apt를 통해 명령어를 설치하였다 `iconv`에서 생성한 utf-8 crlf 파일을 `dos`형식에서 `unix`형식으로 변환한 뒤, 다시 `dos`형식으로 변환하였다.

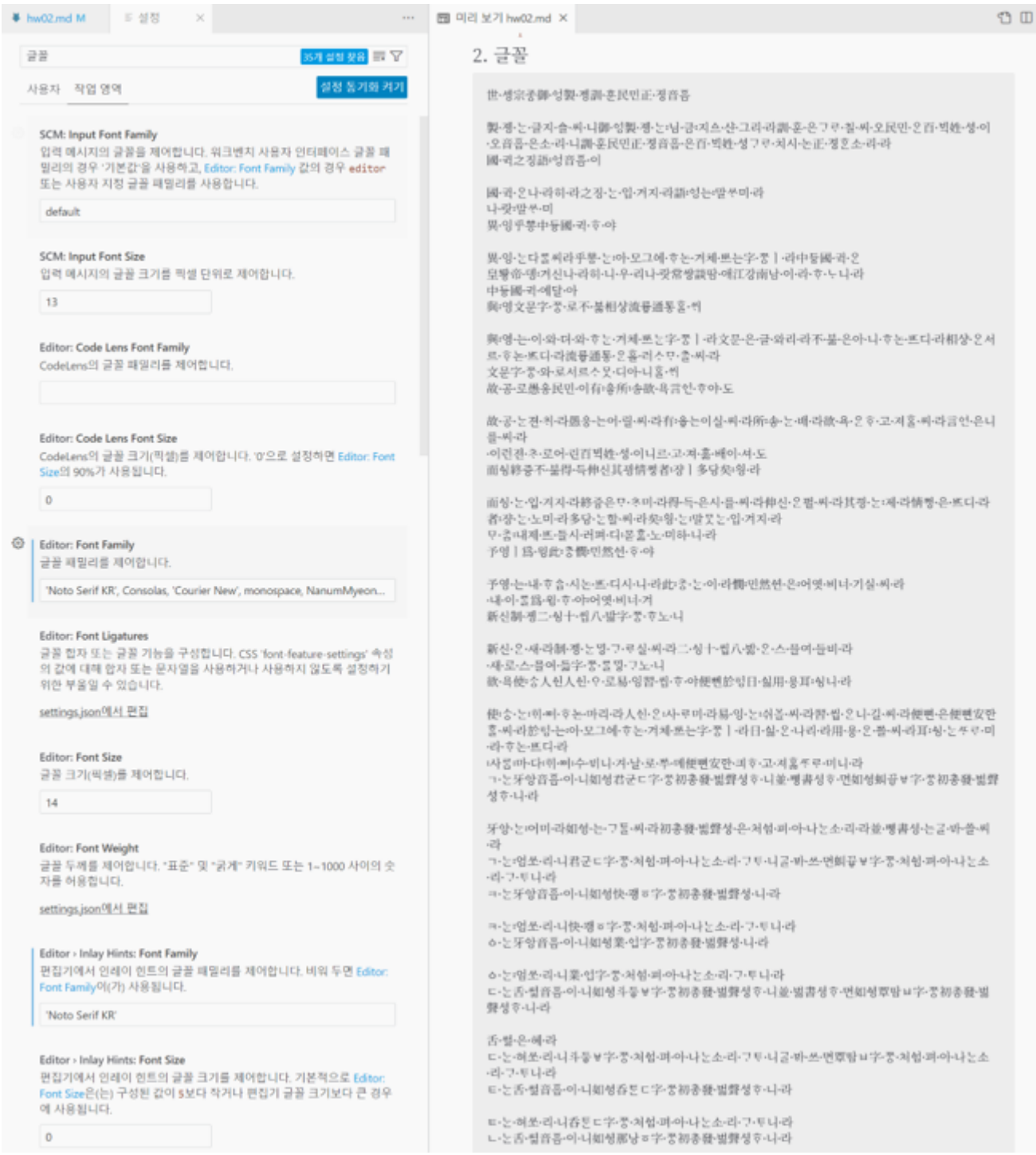
bomstrip

```
SYNOPSIS
     bomstrip
     bomstrip-files file ...

DESCRIPTION
     The bomstrip utility reads UTF-8 data from its standard input and copies it
to its standard output, stripping the BOM (byte-order mark) from the beginning of
the text if it is present. There are no command-line options and no parameters.

     The bomstrip-files utility removes the UTF-8 BOM from the specified files,
saving each file's original contents with a .bom extension.
     It uses the bomstrip utility, trying to execute it as “bomstrip”; if the
bomstrip utility is installed under another name, or if a more complex command is
desired, it may be supplied in the BOMSTRIP environment variable.
```

```
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ file utf8-bom-crlf.txt
utf8-bom-crlf.txt: UTF-8 Unicode (with BOM) text, with CRLF line terminators
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ bomstrip-files utf8-bom-crlf.txt
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$ file utf8-bom-crlf.txt
utf8-bom-crlf.txt: UTF-8 Unicode text, with CRLF line terminators
jtjun@JTJ-MacBook:/mnt/c/Users/JTJ/Desktop/2022/NLP/hw02/src$
```

실습에 앞서, apt를 통해 명령어를 설치하였다 `bomstrip-files` 명령어를 통해 bom을 제거하였다.
그 결과 `.bom`파일이 생성되었다.

2. 글꼴



Settings 에서 글꼴을 설정한 결과, VS code에서 옛 한글이 정상적으로 출력되는 것을 확인할 수 있다.