

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

```
00000000 95ed ea9c 80b8 000a
00000007
```

hangul-utf8-crlf.txt

```
00000000 95ed ea9c 80b8 0a0d
00000010
```

hangul-utf8-bom-lf.txt

```
00000000 bbef edbf 9c95 b8ea 0a80
00000012
```

hangul-utf8-bom-crlf.txt

```
00000000 bbef edbf 9c95 b8ea 0d80 000a
00000013
```

hangul-utf16le-lf.txt

```
00000000 feff d55c ae00 000a
0000010
```

hangul-utf16le-crlf.txt

```
00000000 feff d55c ae00 000d 000a
0000012
```

hangul-utf16be-lf.txt

```
00000000 fffe 5cd5 00ae 0a00
0000010
```

hangul-utf16be-crlf.txt

```
00000000 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 encoding.

```
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

hw02.rmd M

설

...

글꼴

35개 설정 항목

사용자 작업 영역

설정 동기화 제거

SCM: Input Font Family

입력 메시지의 글꼴을 제어합니다. 위키백과 사용자 인터페이스 글꼴 패밀리(예: "기본")를 사용하고, Editor: Font Family 값의 경우 editor 또는 사용자 지정 글꼴 패밀리를 사용합니다.

default

SCM: Input Font Size

입력 메시지의 글꼴 크기를 픽셀 단위로 제어합니다.

13

Editor: Code Lens Font Family

CodeLens의 글꼴 패밀리를 제어합니다.

Editor: Code Lens Font Size

CodeLens의 글꼴 크기를 픽셀 단위로 제어합니다. 0으로 설정하면 Editor: Font Size의 90%가 사용됩니다.

0

Editor: Font Family

글꼴 패밀리를 제어합니다.

'Noto Serif KR', Consolas, 'Courier New', monospace, NanumMyeon...

Editor: Font Ligatures

글꼴 합자 또는 글꼴 기능을 구성합니다. CSS 'font-feature-settings' 속성의 값에 대해 합자 또는 문자열을 사용하거나 사용하지 않도록 설정하기 위한 부울일 수 있습니다.

settings.json에서 편집

Editor: Font Size

글꼴 크기(픽셀)를 제어합니다.

14

Editor: Font Weight

글꼴 두께를 제어합니다. '표준' 및 '굵게' 키워드 또는 1~1000 사이의 숫자를 허용합니다.

settings.json에서 편집

Editor > Inlay Hints: Font Family

편집기에서 인레이 힌트의 글꼴 패밀리를 제어합니다. 비워 두면 Editor: Font Family(가)가 사용됩니다.

'Noto Serif KR'

Editor > Inlay Hints: Font Size

편집기에서 인레이 힌트의 글꼴 크기를 제어합니다. 기본적으로 Editor: Font Size은(는) 구성된 값이 5보다 작거나 편집기 글꼴 크기보다 큰 경우에도 사용됩니다.

0

미리 보기 hw02.rmd X

...

2. 글꼴

世宗宗廟-성황-훈민정-정음용

製-정=글자술-비나-교-성황-정=남-글-지스-산-그라-리-훈=은-고-무-칠-씨-오-민-은-글-빅-姓-성-이-오-음-은-소-라-니-훈-훈민정-정음용=은-글-빅-姓-성-고-무-치-사-는-正-정-은-소-라-國-국-의-정-語-정음용-이

國-국-은-나-라-하-라-의-장-는-임-거-지-리-訓-임-는-말-부-미-라
 나-말-부-미
 萬-萬-무-봉-中-동-國-국-하-야

萬-萬-는-다-를-비-라-무-봉-는-아-모-그-에-하-는-거-제-는-字-宗-ㅣ-라-中-동-國-국-은-萬-萬-宗-정-이-신-나-라-하-나-우-리-나-뜻-常-常-訓-訓-에-江-강-南-남-아-라-하-노-나-라-中-동-國-국-에-달-아
 萬-萬-文-문-字-宗-宗-로-不-불-상-상-流-流-通-통-용-커

萬-萬-는-아-와-다-와-하-는-거-제-는-字-宗-ㅣ-라-文-문-은-글-와-리-라-所-不-불-은-아-니-하-는-正-다-라-相-상-은-서-프-하-는-正-다-라-流-流-通-통-은-용-리-스-무-출-씨-라
 文-문-字-宗-宗-와-로-서-스-모-트-다-아-니-를-커
 故-공-모-충-은-民-민-이-有-有-所-所-송-대-국-안-하-야-도

故-공-는-전-하-리-國-국-은-아-말-씨-라-有-有-는-이-상-씨-라-所-所-는-대-라-故-故-은-有-有-고-저-를-씨-라-言-언-은-니-를-씨-라
 이-리-전-주-로-아-리-고-빅-姓-성-이-니-리-고-저-를-배-이-사-도
 而-而-성-終-중-不-불-得-득-神-신-其-정-情-정-容-ㅣ-多-多-다-영-라

而-而-성-는-임-거-지-라-終-중-은-무-미-라-得-득-은-사-를-씨-라-神-신-은-필-씨-라-其-정-는-제-라-情-정-은-부-다-라-容-容-는-노-마-리-多-다-는-함-씨-라-容-容-는-말-못-는-임-거-지-라
 무-중-제-제-를-사-리-파-다-를-終-중-미-하-나-라
 주-영-ㅣ-爲-영-씨-終-終-민-然-然-하-야

주-영-는-내-후-후-시-는-正-다-사-나-라-此-중-는-아-라-國-國-민-然-然-은-어-몇-비-나-기-실-씨-라
 내-아-爲-爲-영-하-야-어-몇-비-나-거
 新-신-訓-訓-二-성-十-십-八-팔-字-宗-宗-하-노-니

新-신-은-새-라-訓-訓-는-말-고-무-실-씨-라-二-성-十-십-八-팔-은-스-를-어-들-씨-라
 새-로-스-를-어-들-字-宗-宗-말-고-노-니
 故-故-使-승-人-人-人-人-有-로-易-영-習-習-하-야-便-便-於-於-日-日-실-用-용-耳-성-나-라

使-승-는-하-야-하-는-마-리-라-人-人-은-사-무-미-라-易-易-는-하-를-씨-라-習-習-은-나-길-씨-라-便-便-은-便-便-安-안-함-씨-라-於-於-는-아-모-그-에-하-는-거-제-는-字-宗-ㅣ-라-日-日-은-나-라-리-用-용-은-를-씨-라-耳-성-는-무-미-라-하-는-부-다-라
 사-를-마-다-하-야-수-미-나-가-날-로-부-에-便-便-安-안-의-후-고-저-를-무-미-나-라
 가-는-牙-알-글-음-이-니-회-성-강-근-다-字-宗-初-初-發-發-발-聲-성-후-니-並-병-書-書-후-면-회-성-회-강-발-字-宗-初-初-發-發-발-聲-성-후-니-라

牙-알-는-어-미-라-회-성-는-그-를-씨-라-初-初-發-發-발-聲-성-은-처-형-씨-아-나-는-소-라-과-發-發-병-書-성-는-글-씨-를-씨-라
 가-는-牙-알-로-리-니-강-근-다-字-宗-初-初-發-發-발-聲-성-파-아-나-는-소-라-과-發-發-병-書-發-發-字-宗-初-初-發-發-발-聲-성-파-아-나-는-소-라-과-무-나-라
 가-는-牙-알-글-음-이-니-회-성-快-快-字-宗-初-初-發-發-발-聲-성-나-라

가-는-牙-알-로-리-니-快-快-字-宗-初-初-發-發-발-聲-성-파-아-나-는-소-라-과-무-나-라
 가-는-牙-알-글-음-이-니-회-성-集-集-字-宗-初-初-發-發-발-聲-성-나-라

가-는-牙-알-로-리-니-集-集-字-宗-初-初-發-發-발-聲-성-파-아-나-는-소-라-과-무-나-라
 나-는-舌-설-글-음-이-니-회-성-차-차-發-發-字-宗-初-初-發-發-발-聲-성-후-니-並-병-書-書-후-면-회-성-單-單-發-發-字-宗-初-初-發-發-발-聲-성-후-니-라

舌-설-은-혀-라
 나-는-혀-로-리-니-차-차-發-發-字-宗-初-初-發-發-발-聲-성-파-아-나-는-소-라-과-무-나-라
 나-는-舌-설-글-음-이-니-회-성-添-添-字-宗-初-初-發-發-발-聲-성-후-니-라

나-는-혀-로-리-니-添-添-字-宗-初-初-發-發-발-聲-성-파-아-나-는-소-라-과-무-나-라
 나-는-舌-설-글-음-이-니-회-성-部-部-字-宗-初-初-發-發-발-聲-성-후-니-라

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