Разработка программ-конвертеров кодировок UTF16-UTF8

Задание практикума (3 семестр)

Постановка задачи

Требуется разработать две программы-конвертера.

Первая программа читает из стандартного канала ввода текст в кодировке UTF-16, переводит его в кодировку UTF-8 и выводит перекодированный текст в стандартный канал вывода.

Вторая программа читает из стандартного канала ввода текст в кодировке UTF-8, переводит его в кодировку UTF-16 и выводит перекодированный текст в стандартный канал вывода.

Если при запуске программ задан аргумент командной строки — имя файла, то вместо стандартного ввода текст берется из этого файла. Если задано два аргумента, то первый рассматривается как имя файла ввода, а второй — как имя файлы вывода. Стандартные потоки ввода/вывода в этом случае не используются. В случае неправильного имени файла программа должна выдавать сообщение об ошибке и завершаться.

Обе программы должны корректно обрабатывать маркер порядка байтов (byte order mark – BOM) – символ с кодом 0xFEFF — в начале файла. Первая программа должна читать текст в UTF-16, а вторая программа — генерировать текст в UTF-16 в соответствии с прочитанным маркером. В случае отсутствия маркера принимается LE-порядок (то есть по умолчанию BOM представлен байтами 0xFF 0xFE в начале файла).

Также программы должны обрабатывать и случаи некорректного представления входного текста — нечетное количество байтов в UTF-16, некорректные последовательности в UTF-8. В этих случаях программы должны выдавать в стандартный канал вывода сообщений об ошибках (stderr) диагностику, включающую в себя значение некорректного символа (последовательности), а также его смещение относительно начала файла.

Заметим, что программы не должны сами генерировать маркер BOM — тот должен находиться в тексте (в начале файла).

К заданию приложены тестовые файлы в обеих кодировках для контроля правильности заданий, а также для отладки программ.

Методические указания

Задание нужно реализовать на языке Си в операционной системе UNIX (LINUX).

Для хранения UCS-символов следует использовать тип данных `unsigned short`, а для UTF-8 символов - `char`.

Для ввода-вывода текста можно использовать используются только библиотечные функции getchar() и putchar() для UTF-8 и fread() / fwrite() для UTF-16 (хотя последние функции можно использовать, при желании, везде). При тестировании и отладке программ должны использоваться данные из заранее подготовленных файлов. Эти файлы связываются со стандартными каналами ввода-вывода путем механизма перенаправления ввода-вывода, предоставляемого любой оболочкой в ОС UNIX или же задаются в командной строке:

```
utf2usc <le.ucs >le.utf
utf2usc le.ucs le.utf
```

Тестовые файлы

Для тестирования и отладки предлагается использовать следующий (минимальный) набор файлов.

UTF-16 файлы

- letext.ucs – текст в UTF-16 в перевернутом представлении (LE-порядок) с меткой BOM
- betext.ucs – текст в UTF-16 в прямом представлении (BE-порядок) с меткой BOM
- letextbad1.ucs – текст в UTF-16 в перевернутом представлении (LE-порядок) без метки BOM
- betextbad1.ucs – текст в UTF-16 в прямом представлении (BE-порядок) без метки BOM
- letextbad2.ucs – текст в UTF-16 в перевернутом представлении (LE-порядок) с меткой BOM, но с неверным символом (однобайтовым)
- betextbad2.ucs – текст в UTF-16 в прямом представлении (BE-порядок) с меткой BOM, но с неверным символом (однобайтовым)
- leempty.ucs – пустой текст в UTF-16 в перевернутом представлении (LE-порядок) с меткой BOM
- beempty.ucs – пустой текст в UTF-16 в прямом представлении (BE-порядок) с меткой BOM
- le30.ucs – односимвольный (код=0x30 – символ 0) текст в UTF-16 в перевернутом представлении (LE-порядок) с меткой BOM
- be30.ucs – односимвольный (код=0x30 – символ 0) текст в UTF-16 в прямом представлении (BE-порядок) с меткой BOM
- le42f.ucs – односимвольный (код=0x042F – символ Я) текст в UTF-16 в перевернутом представлении (LE-порядок) с меткой BOM
- be42f.ucs – односимвольный (код=0x042F – символ Я) текст в UTF-16 в прямом представлении (BE-порядок) с меткой BOM
- le263A.ucs – односимвольный (код=0x263A – символ ☺) текст в UTF-16 в перевернутом представлении (LE-порядок) с меткой BOM
- be262A.ucs – односимвольный (код=0x263A – символ ☺ текст в UTF-16 в прямом представлении (BE-порядок) с меткой BOM

UTF-8 файлы

- text.utf – текст в UTF-8 с меткой BOM (кодированной)
- text2.utf – текст в UTF-8 без метки BOM
- textbad1.utf – текст в UTF-8 с неверной последовательностью (начинается с байта продолжения) без метки BOM
- textbad2.utf – текст в UTF-8 с неверной последовательностью (отсутствует байт продолжения) без метки BOM

- empty.utf – пустой текст в UTF-8 с меткой BOM
- 30.utf – односимвольный (код=0x30 – символ 0) текст в UTF-8
- 42f.utf – односимвольный (код=0x042F – символ Я) текст в UTF-8
- 263A.utf – односимвольный (код=0x263A – символ ☺) текст в UTF-8