EE3980 Algorithms

Term Project. Encoding `utf-8` Files

Due: June 20, 2020

In Homework 9, we have developed an algorithm to encode an ASCII text file, and a significant saving in disk space can be obtained. For our daily life, however, a large number of files contain Chinese or other non-ASCII characters. Thus, the algorithm developed is not applicable to most of our text files. This project intends to fix that problem to develop a more advanced encoding scheme which can handle text files with utf-8 characters.

Unlike the ASCII character sets, in which each character needs only 7 bits, and thus, the most-significant-bit (MSB) of a byte is always set to 0, the utf-8 encoding is targeted to encode a very large number of characters and it can take more than 1 byte for a Chinese character, and characters from other language. It is also customary to call these characters as symbols. In utf-8, each symbol can take 1 to 4 bytes. For Chinese characters, it usually takes 3 bytes. A more detailed explanation for the utf-8 standard can be found in wikipedia: https://en.wikipedia.org/wiki/UTF-8.

The assignment for this project is to write a C program that takes an utf-8 encoded text file and find a set of encoding scheme to minimize the storage space, in the same way as homework 9 does to the ASCII text files. The output should also be similar to that of the homework 9. In case non-printable symbols (byte or bytes) need to be printed, they can be shown using hexadecimal escape sequence. For example, the new-line character '\n' can be printed as 0x0D or simply 0xD.

To test your program, 7 utf-8 files are provided: they are article1.txt, article2.txt, article3.txt, article4.txt, article5.txt, article6.txt, and article7.txt. Use these files to demonstrate the effectiveness of your encoding scheme.

As always, a pdf report file is also needed. Please describe your algorithm clearly in this report. Necessary complexity analyses need also be performed and results presented. Please use this report to convince me that your algorithm is the best possible.

Notes.

1. For this term project, you must do all the work by yourself.
   No discussions with anyone else is allowed.
2. One executable and error-free C source file should be turned in. This source file should be named as proj.c.
3. A pdf file is also needed. This report file should be named as proja.pdf.
4. Submit your proj.c and proja.pdf on EE workstations using the following command:

   ```
   $ ~ee3980/bin/submit proj proj.c proja.pdf
   ```

   where proj is necessary to indicate the term project.

5. Your report should be clearly written such that I can understand it. The writing, including English grammar, is part of the grading criteria.