BÀI 3

SỐ HOÁ VĂN BẢN

Học xong bài này, em sẽ:

- Biết một số bảng mã kí tự như ASCII, ASCII mở rộng, bảng mã chuẩn quốc tế Unicode là gì và chức năng của chúng.
- Biết được dữ liệu văn bản chứa thông tin về các kí tự kèm màu sắc, kiểu dáng, định dạng,...
- Biết vài khía cạnh lịch sử liên quan đến văn bản tiếng Việt trong máy tính.

Trang văn bản có thể có nhiều chữ số. Em hãy cho biết các kí tự là chữ số thập phân "0", "1", ... , "9" được số hoá, chuyển thành dãy bit như thế nào.

1 Bảng mã ASCII

Trong máy tính mỗi kí tự được biểu diễn bằng một dãy bit. Dãy bit này được gọi là mã nhị phân của nó. Để thống nhất cần có quy định chung.

Một trong số các quy định đầu tiên còn dùng đến ngày nay là bảng mã ASCII, (American Standard Code for Information Interchange). ASCII là bộ mã chuẩn của Mỹ để trao đổi thông tin. Bảng mã ASCII chứa mã nhị phân của bộ chữ cái dùng trong tiếng Anh và một số kí hiệu khác. Mã ASCII của một kí tự là dãy 7 bit, có thể biểu diễn 128 kí tự khác nhau. Ngoài những kí tự in ra màn hình được như ta vẫn hiểu, còn có những "kí tự" không in ra màn hình mà là một tín hiệu để điều khiển máy tính. Người ta gọi chúng là *kí tự điều khiển*.

Sự phát triển của máy tính và Internet vượt ra ngoài nước Mỹ làm xuất hiện nhu cầu mã hoá các kí tự trong nhiều ngôn ngữ khác chưa có trong bảng mã ASCII. Người ta mở rộng bảng mã ASCII bằng cách sử dụng mã nhị phân dài 8 bit, biểu diễn thêm được 128 kí tự nữa. Mã nhị phân của những kí tự đã có trong bảng mã ASCII được thêm bit 0 vào trước để đủ độ dài 8 bit. Các kí tự mới thêm đều có mã nhị phân bắt đầu với bit 1. Bảng mã ASCII mở rộng có thể biểu diễn 256 kí tự khác nhau. (Xem thêm tại *http://vi.wikipedia.org/wiki/ASCII_mở_rộng*).

2 Bảng mã Unicode

1 Em hãy tìm trong bảng mã ASCII mở rộng và cho biết các kí tự "ấ", "ẳ", "ế", "ệ",... có trong bảng mã này không.

Bảng mã ASCII dù đã mở rộng vẫn chưa có các kí tự của nhiều ngôn ngữ khác như Ả Rập, Hindi, Thái,... hay các kí tự tượng hình như chữ Hán, chữ Nhật hoặc chữ Nôm cổ của nước ta.

Bảng mã Unicode được thiết kế với mục đích thống nhất chung việc mã hoá các kí tự cho tất cả các ngôn ngữ khác nhau trên thế giới. Chữ Nôm cổ của nước ta cũng có trong bảng mã này. Với chức năng như vậy, bảng mã Unicode được sử dụng ngày càng phổ biến. (Có thể xem thêm về bảng mã Unicode tại web *https://vi.wikipedia.org/wiki/Unicode*).

3 Mã kí tự, bộ kí tự và mã nhị phân

Con đường đi từ các kí tự cho đến mã nhị phân của nó được chia làm hai bước:

Bước thứ nhất: Cho tương ứng mỗi kí tự với một mã kí tự duy nhất, là một dãy kí số, giống như số căn cước công dân là mã định danh duy nhất của mỗi người. Ý tưởng của Unicode là gán một điểm mã duy nhất (Unique code point) cho mỗi kí tự, kí hiệu, biểu tượng,... được dùng trong tất cả các ngôn ngữ khác nhau trên thế giới. Mỗi điểm mã có một tên gọi. Ví dụ, điểm mã U+1EC7 là của kí tự "ệ". Mỗi điểm mã được gắn một tên gọi duy nhất. Một khi đã gán tên thì không thể thay đổi nữa. Không gian mã Unicode được chia thành các khối, một khối mã sẽ được dành riêng cho một ngôn ngữ cụ thể.

Ví dụ: Với từ "Việt Nam" ta có các điểm mã Unicode như *Hình 1*.

U+0056	U+0069	U+1EC7	U+0074	U+0020	U+004E	U+0061	U+006D
V	i	ệ	t		N	a	m

Hình 1. Các điểm mã Unicode trong từ "Việt Nam"

Bước thứ hai: Chuyển từ mã kí tự thành dãy bit để máy tính xử lí được, gọi là mã hoá (tiếng Anh là *encoding*). Kết quả bước này là một dãy bit. Đây là *mã nhị phân* của kí tự. Bảng mã Unicode chỉ thực hiện bước thứ nhất, cho tương ứng mỗi kí tự với một mã kí tự. Sang bước thứ hai, chuyển thành mã nhị phân, có nhiều cách triển khai thực hiện khác nhau.

Các bộ kí tự UTF-8, UTF-16, UTF-32 được hiểu là các thực thi khác nhau chuyển mã kí tự Unicode thành mã nhị phân. UTF là viết tắt tên tiếng Anh của *Unicode Transformation Format*, số 8 nghĩa là dùng các khối 8 bit để biểu diễn một kí tự. UTF-8 có khả năng mã hoá tất cả 1 112 064 điểm mã kí tự hợp lệ trong Unicode bằng cách sử dụng từ một đến bốn đơn vị mã một byte (8 bit). Nó được thiết kế để tương thích lùi với ASCII: 128 kí tự đầu tiên của Unicode, tương ứng một – một với ASCII, được mã hoá bằng cách sử dụng một byte duy nhất có cùng giá trị nhị phân như ASCII. Văn bản hợp lệ ASCII cũng là hợp lệ UTF-8. UTF-8 an toàn để sử dụng trong hầu hết các ngôn ngữ lập trình.

4 Dữ liệu văn bản và số hoá văn bản

Trong bối cảnh phân biệt các loại dữ liệu trong máy tính thì dữ liệu văn bản được hiểu là văn bản chữ, không chứa hình ảnh, âm thanh. Khi muốn thêm hình ảnh vào một văn bản đang soạn thảo, phải thực hiện thao tác chèn, lấy một tệp hình ảnh đã có trước ở nơi khác.

Văn bản thuần chữ

2

Làm theo hướng dẫn và trả lời câu hỏi:

1) Mở trình soạn thảo văn bản *Notepad*, nhập vào đúng 30 kí tự Latinh đơn giản liền nhau thành một dòng. Không gõ kí tự có dấu trong tiếng Việt. Lưu tệp với tên *thuanchu.txt*.

a) Tệp có kích thước bao nhiêu byte?

b) Mỗi kí tự là mấy byte?

2) Đóng Notepad. Mở tệp *thuanchu.txt* bằng trình soạn thảo *WordPad*. Đổi màu chữ để có 3 dòng kí tự màu khác nhau. Lưu tệp thành dạng *.rtf*.

a) Tệp có kích thước bao nhiêu byte?

b) Tại sao kích thước tăng lên như vậy?

Văn bản thuần chữ (plain text), chỉ gồm các kí tự gõ nhập từ bàn phím khi soạn thảo văn bản. Văn bản thuần chữ là một dãy các kí tự xếp liên tiếp từ trái sang phải, từ trên xuống dưới. Mỗi kí tự là một dãy bit.

Dữ liệu văn bản

Khi mở một văn bản trong trình soạn thảo (không phải Notepad), ví dụ như trong Word, ta thấy các kí tự có thể có nhiều kiểu dáng, màu chữ,… khác nhau; các đoạn trong văn bản có thể được định dạng khác nhau. Ngoài mã nhị phân của các kí tự, trình soạn thảo văn bản phải ghi lưu các thông tin kiểu dáng, màu sắc, định dạng,…

Dữ liệu văn bản trong máy tính là một dãy bit biểu diễn các kí tự có kiểu dáng, màu sắc và các thông tin định dạng khác.

5 Kí tự tiếng Việt trong dữ liệu văn bản

Bộ kí tự của tiếng Việt dựa trên các chữ cái Latinh, phần lớn đã có trong bộ kí tự ASCII. Tuy nhiên, một số kí tự biến thể cộng thêm dấu thanh, ví dụ như "ấ", "ẳ", "ế", "ệ",… không có trong đó. Đây là một khó khăn cần giải quyết khi dùng tiếng Việt với máy tính thời kì trước đây. Đã có những cách làm mã kí tự khác nhau, không nhất quán, dẫn đến có sự lộn xộn, không tương thích, gây khó cho người dùng.

Hiện nay tiêu chuẩn Việt Nam đã thống nhất dùng bảng mã kí tự Unicode.

3 Nhấn **Ctrl + Shift + F6** để hiển thị bảng điều khiển của bộ gõ tiếng Việt **UniKey**; trong hộp **Bảng mã** nháy chuột vào nút mũi tên dấu trỏ xuống để mở ra danh sách các bảng mã có trong bộ gõ UniKey. Em hãy kể tên những bảng mã xuất hiện.

TCVN3 là bảng mã tiêu chuẩn cũ của Việt Nam. Nhiều văn bản cũ theo tiêu chuẩn này vẫn còn được lưu trữ và lấy ra sử dụng. Dễ nhận thấy là phải dùng các phông chữ có ".Vn" đứng đầu mới đọc được. Thậm chí có những văn bản gồm lẫn lộn cả các đoạn dùng phông chữ của bảng mã tiêu chuẩn cũ và cả các đoạn dùng phông chữ Unicode.

Bộ gõ tiếng Việt UniKey khá phổ biến hiện nay có công cụ dễ dàng chuyển đổi các văn bản theo tiêu chuẩn cũ sang dùng mã Unicode để phù hợp với tiêu chuẩn mới.

Hình 2. Công cụ chuyển đổi mã kí tự tiếng Việt trong bộ gõ UniKey

Lí do ra đời bảng mã chuẩn quốc tế Unicode là gì?

Em hãy tìm hiểu công cụ chuyển mã có trong bộ gõ tiếng Việt UniKey (*Hình 2*) và viết hướng dẫn để người khác biết cách sử dụng.

Câu 1. Bảng mã ASCII là gì?

Câu 2. Việc chuyển một kí tự thành mã nhị phân tương ứng gồm mấy bước? Bảng mã Unicode thực hiện bước nào?

Câu 3. Văn bản tiếng Việt hiện nay dùng bảng mã kí tự nào là đúng chuẩn quy định?

Tóm tắt bài học

- Bảng mã kí tự ASCII mở rộng gồm 256 kí tự; mã kí tự ASCII chính là số thứ tự của kí tự trong bảng.
- Bảng mã chuẩn quốc tế Unicode được thiết kế với mục đích thống nhất mã kí tự để máy tính có thể "viết chữ" của rất nhiều ngôn ngữ khác nhau trên thế giới.
- Dữ liệu văn bản trong máy tính là dãy bit biểu diễn các kí tự cùng các thông tin định dạng.