TÓM LƯỢC BÀI GIẢNG
NHẬP MÔN LẬP TRÌNH

(Vũ Quốc Hoàng, vqhoang@fit.hcmus.edu.vn, FIT-HCMUS, 2020)

BÀI 9C
TẬP TIN VĂN BẢN CÓ CẤU TRÚC ĐƠN GIẢN - CSV

Chủ đề

- Tập tin CSV

Tài liệu

[1] Vũ Quốc Hoàng, *Bí kíp luyện Lập trình C (Quyển 1)*, hBook, 2017.

[2] Wikipedia, *Comma-separated values*, https://en.wikipedia.org/wiki/Comma-separated_values.

[3] Vũ Quốc Hoàng, *Bí kíp luyện Lập trình nhập môn với Python*, hBook, 2020.

Đọc tài liệu

- Đọc kĩ: Phần mở rộng Bài 4.7 [1]
- Đọc thêm: [2], Bài 17 [3]

Kiến thức

- **Dữ liệu dạng bảng** (tabular data) là dạng dữ liệu được tổ chức thành một hoặc nhiều **bảng** (table). Mỗi bảng gồm nhiều **dòng/bản ghi** (row, record) và **cột/trường** (column, field). Mỗi dòng tương ứng với dữ liệu của một **đối tượng** (object, unit, data point) với các **ô** (cell), ứng với các cột, mô tả dữ liệu ứng với từng **đặc trưng** (parameter, property, attribute, feature).
- Bảng thường có dòng đầu tiên mô tả tên các cột, khi đó, nó được gọi là **dòng tiêu đề** (header row) và không được tính vào số dòng của bảng. Một bảng dữ liệu gồm m dòng, n cột có thể được xem là mở rộng của mảng hai chiều (hay ma trận) kích thước $m \times n$, trong đó, các cột có thể khác kiểu và thường được đặt tên. Đôi khi các dòng cũng được đặt tên hoặc dùng số thứ tự.
- Dữ liệu dạng bảng được dùng rất phổ biến trong **phân tích dữ liệu** (data analysis), **cơ sở dữ liệu quan hệ** (relational database) và **bảng tính** (spreadsheet).
- Các bảng dữ liệu thường được lưu trữ bằng tập tin văn bản có cấu trúc đơn giản là **tập tin CSV** (Comma-Separated Values file), thường có phần mở rộng là `.csv`. Trong đó, mỗi dòng dữ liệu được để trên một dòng văn bản (dòng đầu tiên là dòng tiêu đề nếu có), dữ liệu của các ô được mô tả bằng chuỗi và phân cách bằng dấu phẩy (có thể chọn dùng các dấu phân cách khác). Nếu dữ liệu của ô phức tạp thì nó có thể được đặt trong cặp dấu nháy kép `"..."`.
- Khi xử lý dữ liệu dạng bảng, ta thường tổ chức cấu trúc tương ứng cho các dòng với các thành phần là các cột. (Đó là lý do mà kiểu cấu trúc thường được gọi là record và các thành phần của nó thường được gọi là field.) Ta cũng thường dùng mảng các cấu trúc này để chứa các dòng dữ liệu.
- Việc xử lý tập tin CSV cũng đơn giản như các tập tin văn bản khác nhưng cần để ý đến kí tự phân cách, dấu nháy kép (nếu có) và thứ tự của các trường.

Kĩ năng

- Biết cách tạo, mở, đọc, nhập, sửa, lưu bảng dữ liệu bằng tập tin CSV như một người dùng bình thường (chẳng hạn biết dùng Microsoft Excel, Notepad, Notepad++, …)
- Biết cách xử lý tập tin CSV
- Vận dụng kiểu dữ liệu cấu trúc, mảng cấu trúc, tổ chức chương trình theo module để làm việc với dữ liệu dạng bảng
- Vận dụng dữ liệu dạng bảng và tập tin CSV để lưu trữ, tổ chức, xử lý dữ liệu cho các tình huống hay gặp

Lưu ý

- Tập tin CSV được dùng phổ biến trong việc lưu trữ và trao đổi dữ liệu dạng bảng giữa nhiều phần mềm. Hơn nữa, các kĩ thuật xử lý trên tập tin CSV sẽ được mở rộng cho các dạng tập tin văn bản có cấu trúc phức tạp hơn. Do đó, sinh viên nên nắm vững và thành thạo.

Bài tập

1. Đọc và tìm hiểu thêm về tập tin CSV trong [2].
2. Viết chương trình "trộn thư" (Mail Merge) nhận:
   (1) tập tin văn bản chứa *nội dung* (template, form, main document) gồm *văn bản cố định* (fixed text) và các *biến* (variable, label).
   (2) tập tin CSV chứa *nguồn dữ liệu* (data source) gồm các *trường* (column, field) tương ứng với các biến trong tập tin nội dung.

   Chương trình cho phép người dùng chọn các *dòng* (row, record) trong nguồn dữ liệu và tạo ra các file văn bản *kết xuất* (output document, merged document) ứng với từng dòng.

   (Xem thêm: https://en.wikipedia.org/wiki/Mail_merge)

3. Viết chương trình "quản lý điểm sinh viên":
   - Xây dựng cấu trúc thích hợp cho các dạng dữ liệu phức tạp.
   - Tổ chức chương trình theo module.
   - Lưu trữ dữ liệu trên một (hoặc nhiều) tập tin CSV được thiết kế (và liên kết) phù hợp.
4. Tương tự câu 3, viết chương trình "quản lý nhân viên".