Thực hành
CHƯƠNG TRÌNH DỊCH

Bài 1: Lập bảng chỉ mục

Phạm Đăng Hải
haipd@soict.hut.edu.vn

Đề bài

- Đọc một tệp văn bản, hãy lập một bảng chỉ mục (*index table*) cho tệp văn bản đó.
- Bảng chỉ dẫn liệt kê tất cả các **từ** xuất hiện trong văn bản theo quy cách
  - Mỗi từ được liệt kê một lần cùng với số lần xuất hiện trong văn bản và dòng xuất hiện từ đó.
  - Các từ phải được sắp xếp theo thứ tự từ điển

Mô tả chi tiết

- Tệp văn bản
  - Đoạn văn bản tiếng Anh, định dạng ASCII
    - Tệp mẫu "***vanban.txt***"
- Từ là những dãy chữ cái phân biệt bởi
  - Khoảng trống/ Dấu phân cách
    - Các ký tự không phải chữ cái (a..z, A..Z)
- Không phân biệt chữ hoa, chữ thường
  - Khi đưa vào bảng chỉ mục phải chuyển tất cả các ký tự thành chữ thường

Mô tả chi tiết

- Không đưa vào bảng chỉ mục
  - Những từ không có ý nghĩa để tra cứu.
    - Những từ như vậy được lưu trong tệp: "***stopw.txt***", mỗi từ một dòng.
  - Những danh từ riêng.
    - Đó là những từ có chữ cái đầu là chữ hoa nhưng không đứng sau dấu chấm câu.
      - Ví dụ: "Will you visit **Hanoi** someday?"

Tình bày kết quả

Trình bày kết quả theo dòng:

- Đầu tiên là từ, sau đó là phần dãy số.
- Số đầu tiên là số lần xuất hiện của từ,
- Các số tiếp theo là dòng mà từ đó xuất hiện.

• Ví dụ

answer **7**,8,12,15

ant **2**,4,6

baby **7**,9,21

cruel **2**,4,5

Thiết kế khung cho chương trình

- Đọc một từ.
- Kiểm tra từ có nghĩa.
- Lưu từ vào danh sách được sắp xếp.
- Xác định các thông tin cho từ có nghĩa
  - Số lần xuất hiện, chỉ số dòng.

Đọc từ

- Đọc từng ký tự đến khi gặp ký tự kết thúc từ.
  - Ký tự kết thúc từ?
    - Chữ số, dấu cách, dấu chấm câu, dấu xuống dòng..
    - Không phải chữ cái a..zA..Z
      - Hàm int isalpha(char c);
- Kỹ thuật:
  - Xác định ký tự đầu tiên của từ?
  - Ghép các ký tự thành một từ?
    - Hàm ghép xâu?
      - **Xâu:** Mảng các ký tự, kết thúc bởi ký tự NUL
    - Chuyển ký tự về chữ thường?
      - Hàm int tolower(char c)

Kiểm tra từ có nghĩa

- Loại bỏ nếu từ trong danh sách "*stopw*",
- Loại bỏ nếu từ là danh từ riêng.
- Kỹ thuật:
  - Kiểm tra từ có trong danh sách
    - Hàm so sánh xâu.
      - Hàm int strcmp(const char * s1, const char * s2)
  - Kiểm tra từ có phải danh từ riêng
    - Vấn đề: Nếu danh từ riêng đứng ở đầu câu?

Chèn từ vào danh sách

- Nếu từ chưa có trong danh sách

  → thêm từ vào danh sách.
- Nếu từ đã có trong danh sách

  →Tăng số lần xuất hiện

  →Thêm chỉ số dòng (*nếu chưa có*)
- Kỹ thuật:
  - Sắp xếp danh sách theo thứ tự từ điển?
    - Hàm so sánh xâu
  - Xác định chỉ số dòng?
    - Ký tự xuống dòng: /n và /r
      - Khác biệt giữa linux và window? → sử dụng '\n'

Biến trong chương trình

- Danh sách từ cho bảng chỉ mục

  → Sử dụng mảng

  - Sắp xếp kiểu thêm dần
  - **Vấn đề:** kích thước mảng !?

- Dach sách từ không có nghĩa tra cứu:

  → Sử dụng mảng.

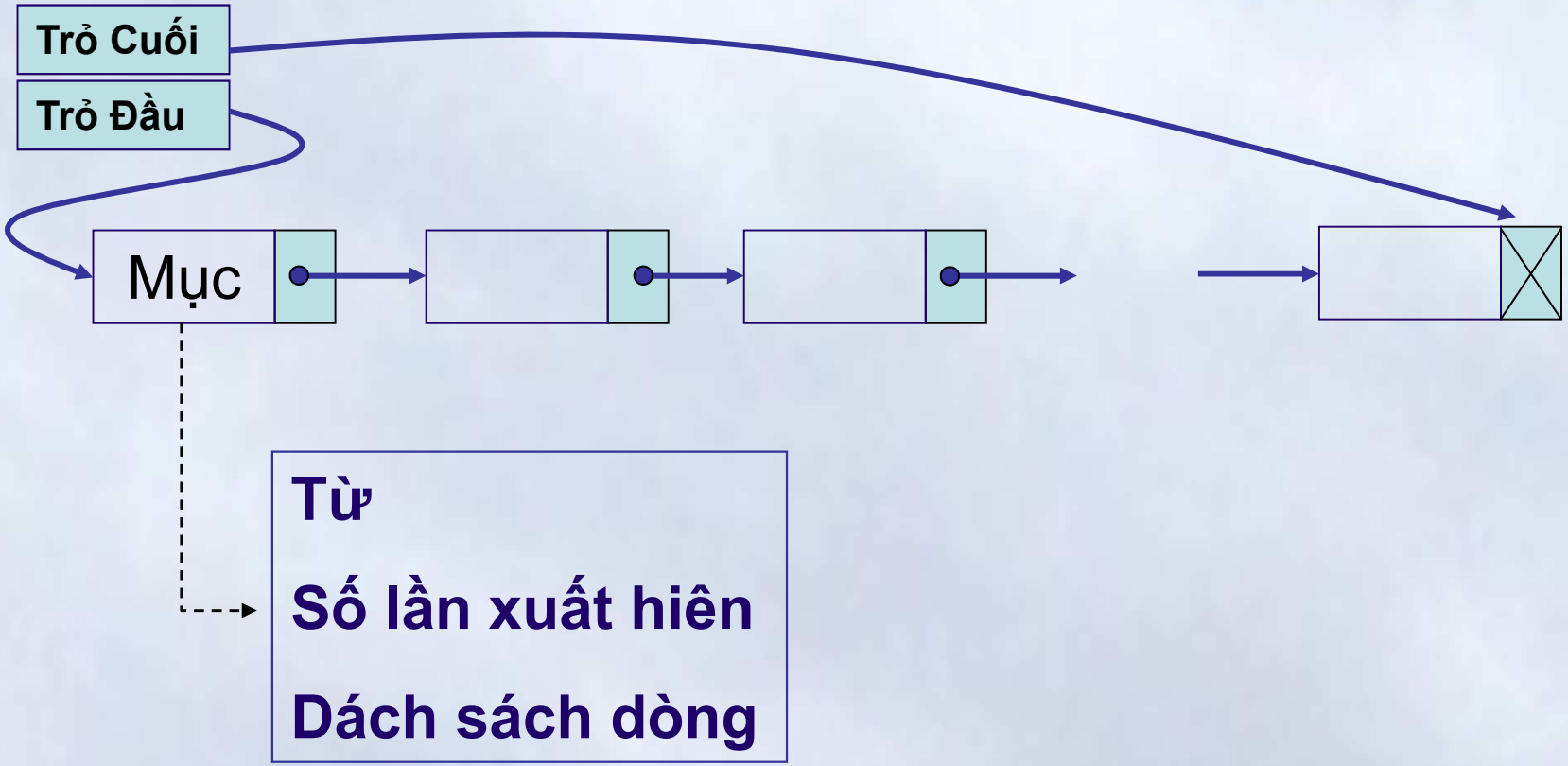
- Dãy các chỉ số dòng

  →Sử dụng xâu ký tự.

  Ví dụ "2, 5, 6, 7, 12"

Biến trong chương trình

Danh sách liên kết



Kiến thức lập trình cần chú ý

- Thao tác với tệp:
  - Mở, đóng tệp,
  - Đọc ký tự (*int fgetc(FILE * flot)*)
- Thao tác với xâu ký tự:
  - So sánh xâu ký tự,
  - Thêm ký tự vào xâu,
- Thao tác con trỏ
  - Xin, giải phóng vùng nhớ
    - Hàm malloc(int size), free(void *)..
  - Chèn phần tử vào danh sách