



Thực hành Chương Trình Dịch

Bài thực hành số 1
THAO TÁC TRÊN GCC

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, mỗi từ một lần, kèm theo số lần xuất hiện và chỉ số của dòng xuất hiện từ đó.

Mô tả chi tiết


- Tập văn bản: Đoạn văn bản tiếng Anh, lưu theo định dạng ACSII.
Tập mẫu: “vanban.txt”
- Từ là những dãy chữ cái phân biệt nhờ **khoảng trống** hay **dấu phân cách**.
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 (tiếp)

- Không đưa vào bảng chỉ mục:
 - Những từ không có ý nghĩa để tra cứu. Những từ này 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?”
- Các từ phải được sắp xếp theo thứ tự từ điển.

Ví dụ kết quả

- Trình bày kết quả:

answer 7, 

ant 2, 4, 6

baby 7, 9, 21

cruel 2, 4, 5

- Trên mỗi 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 chỉ số dòng mà từ đó xuất hiện.



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 cho đến khi gặp ký tự kết thúc từ.
- Định nghĩa ký tự kết thúc từ:
 - Chữ số, Dấu cách, Dấu chấm (.) (?)(!), Dấu xuống dòng ... => không phải chữ cái a..zA..Z
- Kỹ thuật:
 - Xác định ký tự đầu tiên của từ?
 - Ghép các ký tự thành một từ?
 - Hàm ghép xâu.
 - Chuyển ký tự về chữ thường?

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.
 - ☐ Kiểm tra từ có phải danh từ riêng? 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 và thêm chỉ số dòng.
- Kỹ thuật:
 - ☐ Sắp xếp danh sách theo thứ tự từ điển?
 - ☐ Xác định chỉ số dòng? Ký tự xuống dòng?
 - /n và /r

Biến trong chương trình

- Từ: sử dụng xâu ký tự
- Danh sách từ cho bảng chỉ mục: sử dụng mảng.
- Danh sách từ không có nghĩa tra cứu: sử dụng mảng.
- Dãy các chỉ số dòng: sử dụng xâu ký tự.

Kiến thức lập trình cần chú ý

- Thao tác với tệp:
 - ☐ Mở, đóng tệp,
 - ☐ Đọc ký tự,
 - ☐ Ghi tệp.
- Thao tác với chuỗi ký tự:
 - ☐ So sánh chuỗi ký tự,
 - ☐ Thêm ký tự vào chuỗi,