Cây tiền tố Trie

Cây tiền tố (prefix tree), còn gọi là **Trie**, là một cấu trúc dữ liệu dạng cây được sử dụng để **lưu trữ tập hợp các chuỗi**, thường là **các từ hoặc dãy ký tự**, sao cho việc **tra cứu tiền tố** trở nên hiệu quả.

Tìm kiếm từ gợi ý (Autocomplete)

- **Ứng dụng**: Google Search, từ điển điện tử, ô tìm kiếm thông minh.
- **Mô tả**: Khi người dùng gõ "app", hệ thống dùng Trie để tìm tất cả từ bắt đầu bằng "app" như: *apple, application, appoint*.

Kiểm tra chính tả (Spell Checker)

- **Ứng dụng**: Trình xử lý văn bản (Word, Google Docs), IDE, trình duyệt.
- **Mô tả**: Trie giúp xác định nhanh chóng một từ có hợp lệ không và đưa ra gợi ý sửa lỗi (edit distance + trie traversal).

Tìm kiếm chuỗi theo từ khóa (Keyword matching)

- **Ứng dụng**: Hệ thống lọc nội dung, kiểm tra từ cấm, công cụ tìm kiếm.
- **Mô tả**: Trie giúp kiểm tra nhanh liệu có từ nhạy cảm hoặc từ khóa cụ thể nào trong đoạn văn bản hay không.

Lưu trữ từ điển hoặc tập hợp từ

- **Ứng dụng**: Từ điển ngôn ngữ, máy học (machine learning), NLP.
- **Mô tả**: Trie giúp tiết kiệm bộ nhớ và tăng tốc độ tìm kiếm so với lưu danh sách từ truyền thống.

Hãy cài đặt 2 ứng dụng sau của cây Trie

Ứng dụng 1. Kiểm tra chính tả

Đầu vào là 1 từ điển tiếng Anh tải về từ : https://raw.githubusercontent.com/dwyl/english-words/refs/heads/master/words.txt

Hãy đọc vào các từ và xây dựng cây prefix.

Tiếp đến đọc vào 1 văn bản tiếng Anh từ bàn phím (hoặc từ file). In ra màn hình (và file) những từ KHÔNG có trong từ điển.

Chú ý: Hãy để ý loại bỏ các chữ số!

Ứng dụng 2. Hệ thống gợi ý hoàn thiện từ (Autocomplete)

Đầu vào là văn bản (hoặc file văn bản) chứa các từ mà người dùng tự nhập. Hãy xây dựng cây prefix từ tập từ đó.

Tiếp đến người dùng sẽ nhập vào 1 vài ký tự đầu (ít nhất 2 ký tự trở lên) và hệ thống sẽ tự động in ra các gợi ý hoàn thiện từ cho người đó.

Hãy tìm cách lưu lại cây prefix (file cá nhân hóa của người dùng) dưới dạng file để lần chạy tiếp ta có thể có tùy chọn nạp dữ liệu cũ thay vì xây dựng cây prefix từ đầu.

Các tiêu chí khi in ra từ gợi ý hoàn thiện

- Phân biện Hoa/thường hoặc không phân biệt
- Duyệt cây bình thường (lấy theo chiều rộng) và đưa ra các từ theo đúng thứu tự duyệt
- Ưu tiên gợi ý các từ có tần số xuất hiện cao nhất trước.