Đề tài học phần mở rộng

Nén và phát hiện chuỗi bất thường bằng Trie và phân tích thống kê

Lớp Kỹ sư Tài năng – Cấu trúc dữ liệu và Giải thuật

1 Mục tiêu

Xây dựng một hệ thống sử dụng cấu trúc Trie để nén dữ liệu chuỗi (log, văn bản, chuỗi sinh học...) và phát hiện các chuỗi có hành vi bất thường. Hệ thống khai thác các kỹ thuật duyệt cây, phân tích nhánh Trie, và đo đếm thống kê (tần suất, độ bất thường nhánh, entropy cục bộ) để xác định các chuỗi có độ phổ biến thấp hoặc cấu trúc lạ thường.

2 Yêu cầu chính

1. **Nghiên cứu lý thuyết:**
   - Hiểu và cài đặt cấu trúc Trie.
   - Tìm hiểu cách sử dụng Trie để nén dữ liệu chuỗi.
   - Phân tích bất thường bằng các đặc trưng thống kê trên Trie.
2. **Xây dựng hệ thống:**
   - Cài đặt Trie để lưu trữ và nén chuỗi.
   - Tính toán các đặc trưng thống kê tại mỗi nút: tần suất, số nhánh con, chiều sâu.
   - Phát hiện các chuỗi bất thường dựa trên ngưỡng thống kê.
3. **Trực quan và đánh giá:**
   - Hiển thị cấu trúc Trie và các chuỗi bất thường.
   - So sánh mức độ nén trước và sau.
   - Phân tích độ chính xác và độ phủ của phát hiện bất thường.
4. **Ngôn ngữ lập trình:** Dùng C++ hoặc Python. Không sử dụng học sâu.

3 Kết quả đầu ra

1. **Chương trình phát hiện chuỗi bất thường:**

- **Input:** Một tập dữ liệu gồm nhiều chuỗi ký tự (ví dụ: dòng log hệ thống, câu văn trong văn bản, đoạn gen sinh học...), được lưu trong file văn bản hoặc danh sách chuỗi. Mỗi chuỗi sẽ được xem là một quan sát (sample) đưa vào cây Trie để phân tích.
- **Tiền xử lý:**
  - Chuẩn hóa chuỗi (loại bỏ ký tự đặc biệt, chuyển về chữ thường nếu cần).
  - Có thể cắt các chuỗi dài thành đoạn nhỏ hơn có độ dài cố định để phát hiện chi tiết hơn.
- **Xử lý chính:**
  - Lần lượt chèn từng chuỗi vào cây Trie và đếm số lần xuất hiện tại các nhánh.
  - Tính toán các thống kê tại mỗi nút: tần suất xuất hiện, số nhánh con, độ sâu...
- **Output:**
  - Danh sách các chuỗi bất thường, được xác định dựa trên một trong các tiêu chí:
    * Tần suất xuất hiện thấp (rare strings).
    * Cấu trúc khác biệt (nút có nhánh hiếm, độ sâu bất thường).
  - Bảng thống kê số lượng chuỗi bất thường phát hiện được, tỷ lệ trên tổng số.
  - Báo cáo văn bản tóm tắt đặc điểm các chuỗi bất thường.

2. **Thống kê hiệu năng:**
   - Mức nén đạt được (số node giảm).
   - Tỷ lệ chuỗi bất thường phát hiện.
3. **Trực quan kết quả:**
   - In cây Trie (dưới dạng text hoặc đồ họa đơn giản).
   - Hiển thị rõ vị trí các chuỗi bất thường.

4 Hình thức báo cáo

1. **Báo cáo kỹ thuật:** Trình bày rõ mục tiêu, thuật toán, cài đặt, kết quả. Viết bằng LaTeX, nộp PDF.
2. **Mã nguồn:**
   - Toàn bộ mã nguồn đưa lên GitHub, có `README.md` hướng dẫn chạy.
   - Yêu cầu chạy được trên Google Colab (dùng `%writefile` và `!g++` nếu dùng C++).
   - Chèn link GitHub và link Colab vào đầu báo cáo PDF.
3. **Video demo:** Giới thiệu kết quả thực nghiệm, khoảng 5–10 phút.
4. **Hạn nộp:** Trước ngày **XX/YY/2025** (sẽ thông báo sau).