System requirement:

PHP>=5.4

Mysql

Apache

Configuration:

Mysql dump file: jaccard.sql.gz

Config file: c-config.php

Cấu trúc chương trình:

Cơ sở dữ liệu:

1. Bảng docs: bảng lưu thông tin tài liệu(các trang web)
   1. Cột content: nội dung đã đc làm sạch, bỏ html tag, ảnh, chỉ giữ lại text
   2. Cột url: đường dẫn trang web lấy về
   3. Các cột i[number][number]: i21->i210, i31->i310, i41->i410 lần lượt là giá trị đại diện cho tài liệu đối với độ dài shingle là 2,3,4
2. Các bảng shingle[number] lưu thông tin shingle

Hoạt động:

1. Tải nội dung trang web và tách nội dung trang web
2. Tạo tập shingle và ánh xạ sang số
3. Tính tập đại diện
4. So sánh giá trị dại diện với csdl
   1. Thực ra là truy vấn tính tổng các giá trị đại diện trùng khớp
   2. Rồi chọn ra tài liệu có nhiều giá trị đại diện trùng khớp nhất

Thống kê:

CSDL: khoảng 10.000 tài liệu là các trang web tải xuống từ baomoi.com

Số lượng các shingle với độ dài 2 khoảng hơn 500.000

Thời gian chạy: ~0.2 giây

**Độ phức tạp thuật toán:**

Ánh xạ: O(n\*m), với m là số single của tài liệu hiện tại, n là tổng số single hiện có

Hàm băm: O(1)

Tìm kiếm tài liệu: O(n), sắp xếp kết quả O(n.log(n)), n là số tài liệu

Độ phức tạp cho toàn bộ quá trình O(m)+O(n.log(n)), m là số shingle, n là số tài liệu