Theo số liệu thống kê1, thế giới có 3,3 tỉ người sử dụng internet thường xuyên và hơn 750 triệu trang Web hiện tại đang hoạt động. Các trang Web ngày nay có dung lượng rất lớn vì các tập tin HTML, JavaScript, CSS, hình ảnh, các đoạn phim thường có kích thước nặng. Sự bùng nổ người dùng Web cùng dữ liệu các trang Web đặt ra những bài toán nâng cấp cơ sở hạ tầng và tìm ra các giải pháp khác nhằm tăng tốc độ tải các trang Web. Nén trang Web là một trong những giải pháp giúp cải thiện tốc độ tải trang Web, tiết kiệm dữ liệu lưu trữ với chi phí thấp và hiệu quả cao. Hiện nay, hầu hết các máy chủ và trình duyệt đều hỗ trợ chức năng nén và giải nén với các bộ nén Gzip hay Deflate thông qua HTTP nén. Hơn nữa, việc nén dữ liệu Web có thể cải thiện bằng việc thu nhỏ dữ liệu các tập tin HTML, CSS, JavaScript như xóa bỏ chú thích không quan trọng, những đoạn mã nguồn thừa, hay viết lại đoạn mã ngắn hơn, v.v... trước khi đưa vào các bộ nén Gzip hay Deflate. Hiện nay, các công cụ thu nhỏ dữ liệu các tập tin dạng văn bản được phát triển rất nhiều. Tuy nhiên, các công trình nghiên cứu so sánh các công cụ này rất ít. Vì vậy, khóa luận giới thiệu các kỹ thuật tiền xử lý dữ liệu Web với các tập tin HTML, CSS, JavaScript; so sánh thực nghiệm các kỹ thuật một cách khách quan; áp dụng một vài kỹ thuật tốt nhất từ so sánh thực nghiệm vào một dự án thực tế. Cuối cùng, khóa luận đưa ra những khuyến nghị cho lập trình viên có thể lựa chọn kỹ thuật phù hợp với dự án cần triển khai